Clustering in Hilbert Geometry
Two case studies: The Probability Simplex and the Correlation Elliptope

Frank Nielsen and Ke Sun

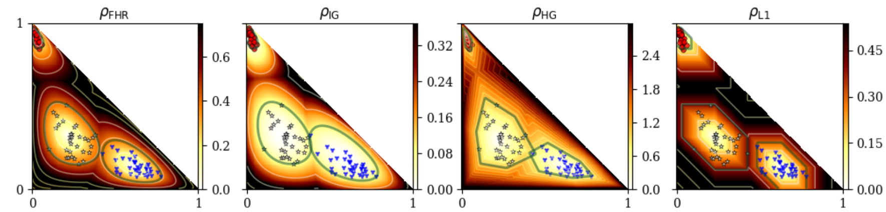
Abstract Clustering categorical distributions in the probability simplex is a fundamental primitive often met in applications dealing with histograms or mixtures of multinomials. Traditionally, the differential-geometric structure of the probability simplex has been used either by (i) setting the Riemannian metric tensor to the Fisher information matrix of the categorical distributions, or (ii) defining the information-geometric structure induced by a smooth dissimilarity measure, called a divergence. In this work, we introduce a novel computationally-friendly non-differential framework for modeling the probability simplex: Hilbert simplex geometry. We discuss the pros and cons of those three statistical modelings, and compare them experimentally for clustering tasks.
Keywords: Fisher-Riemannian geometry, information geometry, Hilbert simplex geometry, Finsler geometry, center-based clustering.

1 Introduction

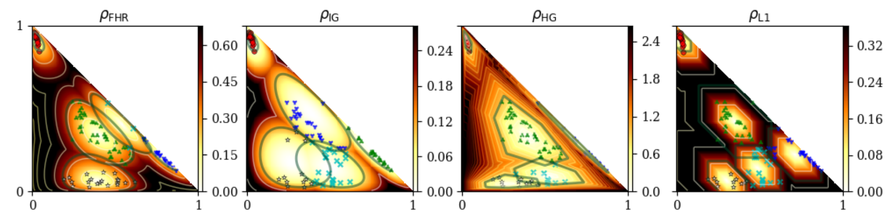
The multinomial distribution is an important representation in machine learning that is often met in applications [34, 19] as normalized histograms (with non-empty bins). A multinomial distribution (or categorical distribution) $p \in \Delta_d$ can be thought as a point lying in the probability simplex Δ_d (standard simplex) with coordinates $p = (\lambda_p^0, \dots, \lambda_p^d)$ such that $\lambda_p^i > 0$ and $\sum_{i=0}^d \lambda_p^i = 1$. The open probability simplex Δ_d sits in $\mathbb{R}^{d+1}$ on the hyperplane $H_{\Delta_d} : \sum_{i=0}^d x_i = 1$. We consider the task of clustering a set $\Lambda = \{p_1, \dots, p_n\}$ of n categorical distributions in Δ_d [19] using center-based k-means++ or k-center clustering algorithms [6, 25] that rely on a dissimilarity

Frank Nielsen
Sony Computer Science Laboratories, Tokyo, Japan, e-mail: `Frank.Nielsen@acm.org`

Ke Sun
Data61, Sydney, Australia, e-mail: `Ke.Sun@data61.csiro.au`



(a) $k = 3$ clusters



(b) $k = 5$ clusters

Fig. 1: k-Center clustering results on a toy dataset in the space of trinomials Δ_2. The color density maps indicate the distance from any point to its nearest cluster center.

measure (loosely called distance or divergence when smooth) between any two categorical distributions. In this work, we mainly consider three distances with their underlying geometries for clustering: (1) Fisher-Rao distance ρ_{FHR}, (2) Kullback-Leibler divergence ρ_{IG}, and (3) Hilbert distance ρ_{HG}. The geometric structures are necessary in algorithms, for example, to define midpoint distributions. Figure 1 displays the k-center clustering results obtained with these three geometries as well as the Euclidean L^1 distance ρ_{L1}. We shall now explain the Hilbert simplex geometry applied to the probability simplex, describe how to perform k-center clustering in Hilbert geometry, and report experimental results that demonstrate superiority of the Hilbert geometry when clustering multinomials.

The rest of this paper is organized as follows: Section 2 formally introduces the distance measures of Δ_d. Section 3 introduces how to efficiently compute the Hilbert distance. Section 4 presents algorithms for Hilbert minimax centers and Hilbert clustering. Section 5 performs an empirical study of clustering multinomial distributions, comparing Riemannian geometry, information geometry and Hilbert geometry. Section 6 presents a second case study of using Hilbert geometry in machine learning: clustering correlation matrices. Finally, section 7 concludes this work by summarizing the pros and cons of each geometry. Although some contents require prior knowledge on geometric structures, we will present the detailed algorithms so that general audience can still benefit from this work.

2 Three distances with their underlying geometries

2.1 Fisher-Hotelling-Rao geometry

The Rao distance between two multinomial distributions is [30, 34]:

$$\rho_{\mathrm{FHR}}(p,q) = 2\arccos\left(\sum_{i=0}^{d}\sqrt{\lambda_p^i\lambda_q^i}\right). \tag{1}$$

It is a Riemannian metric length distance (satisfying the symmetric and triangular inequality axioms) obtained by setting the metric tensor g to the *Fisher information matrix* (FIM) $\mathscr{I}(p) = (g_{ij}(p))_{d\times d}$ wrt the coordinate system $(\lambda_p^1,\cdots,\lambda_p^d)$, where

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

We term this geometry the Fisher-Hotelling-Rao (FHR) geometry [28, 61, 53, 54]. The metric tensor g allows to define an *inner product* on each tangent plane T_p of the probability simplex manifold: $\langle u,v\rangle_p = u^\top g(p)v$. When g is the identity matrix, we recover the Euclidean (Riemannian) geometry with the inner product being the scalar product: $\langle u,v\rangle = u^\top v$. The geodesics $\gamma(p,q;\alpha)$ are defined by the Levi-Civita metric connection [2, 15]. The FHR manifold can be embedded in the positive orthant of an Euclidean d-sphere in $\mathbb{R}^{d+1}$ by using the *square root representation* $p \mapsto \sqrt{p}$ [30]. Therefore the FHR manifold modeling of Δ_d has constant *positive* curvature: It is a spherical geometry restricted to the positive orthant with the metric distance measuring the arc length on a great circle.

2.2 Information geometry

A divergence D is a smooth C^3 differentiable dissimilarity measure [3] that allows to define a dual structure in Information Geometry (IG; [60, 15, 2]). A f-divergence is defined for a strictly convex function f with $f(1) = 0$ by:

$$I_f(p:q) = \sum_{i=0}^{d}\lambda_p^i f\left(\frac{\lambda_q^i}{\lambda_p^i}\right).$$

It is a *separable* divergence since the d-variate divergence can be written as a sum of d univariate divergences: $I_f(p:q) = \sum_{i=0}^{d} I_f(\lambda_p^i:\lambda_q^i)$. The class of f-divergences plays an essential role in information theory since they are provably the *only* separable divergences that satisfy the *information monotonicity* property [2, 37]. That is, by coarse-graining the histograms we obtain lower-dimensional multinomials, say p' and q', such that $0 \le I_f(p':q') \le I_f(p:q)$ [2]. The Kullback-Leibler (KL) divergence ρ_{IG} is a f-divergence obtained for $f(u) = -\log u$:

$$\rho_{\mathrm{IG}}(p,q) = \sum_{i=0}^{d} \lambda_p^i \log \frac{\lambda_p^i}{\lambda_q^i}. \tag{2}$$

It is an asymmetric non-metric distance: $\rho_{\mathrm{IG}}(p,q) \neq \rho_{\mathrm{IG}}(q,p)$. In differential geometry, the structure of a manifold is defined by two components:

1. A *metric tensor* g that allows to define an inner product $\langle \cdot, \cdot \rangle_p$ at each tangent space (for measuring vector lengths and angles between vectors);
2. A *connection* ∇ that defines *parallel transport* $\prod_{p,q}^{\nabla}$, *i.e.*, a way to move a tangent vector from one tangent plane T_p to any other one T_q.

For FHR geometry, the implicitly used connection is called the Levi-Civita connection that is induced by the metric g: $\nabla^{LC} = \nabla(g)$. It is a metric connection since it ensures that $\langle u, v \rangle_p = \langle \prod_{p,q}^{\nabla^{\mathrm{LC}}} u, \prod_{p,q}^{\nabla^{\mathrm{LC}}} v \rangle_q$. The underlying information-geometric structure of KL is characterized by a pair of dual connections [2] $\nabla = \nabla^{(-1)}$ (mixture connection) and $\nabla^* = \nabla^{(1)}$ (exponential connection) that induces two dual geodesics (technically, ± 1-autoparallel curves, [15]). Those connections are said *flat* as they define two dual affine coordinate systems θ and η on which the θ- and η-geodesics are straight line segments, respectively. For multinomials, the *expectation parameters* are: $\eta = (\lambda^1, \ldots, \lambda^d)$ and they 1-to-1 correspond to the *natural parameters*: $\theta = \left(\log \frac{\lambda^1}{\lambda^0}, \ldots, \log \frac{\lambda^d}{\lambda^0} \right)$. Thus in IG, we have two kinds of midpoint multinomials of p and q, depending on whether we perform the (linear) interpolation on the θ- or the η-geodesics. Informally speaking, the dual connections $\nabla^{(\pm 1)}$ are said coupled to the FIM since we have $\frac{\nabla + \nabla^*}{2} = \nabla(g) = \nabla^{\mathrm{LC}}$. Those dual connections are not metric connections but enjoy the following property: $\langle u, v \rangle_p = \langle \prod_{p,q} u, \prod^*{}_{p,q} v \rangle_q$, where $\prod = \prod^{\nabla}$ and $\prod^* = \prod^{\nabla^*}$ are the corresponding induced dual parallel transports. The geometry of f-divergences [3] is the α-geometry (for $\alpha = 3 + 2 f'''(1)$) with the dual $\pm\alpha$-connections, where $\nabla^{(\alpha)} = \frac{1+\alpha}{2} \nabla^* + \frac{1-\alpha}{2} \nabla$. The Levi-Civita metric connection is $\nabla^{\mathrm{LC}} = \nabla^{(0)}$. More generally, it was shown how to build a dual information-geometric structure for *any* divergence [3]. For example, we can build a dual structure from the symmetric Cauchy-Schwarz divergence [29]:

$$\rho_{\mathrm{CS}}(p,q) = -\log \frac{\langle \lambda_p, \lambda_q \rangle}{\sqrt{\langle \lambda_p, \lambda_p \rangle \langle \lambda_q, \lambda_q \rangle}}. \tag{3}$$

2.3 Hilbert simplex geometry

In Hilbert Geometry (HG; [27]), we are given a bounded convex domain $\mathscr{C}$ (here, $\mathscr{C} = \Delta_d$), and the distance between any two points M, M' of $\mathscr{C}$ is defined as follows: Consider the two intersection points AA' of the line (MM') with $\mathscr{C}$, and order them on the line so that we have A, M, M', A'. Then the Hilbert metric distance [14] is defined by:

$$\rho_{\mathrm{HG}}(M,M') = \begin{cases} \left|\log \frac{|A'M||AM'|}{|A'M'||AM|}\right|, & M \neq M', \\ 0 & M = M'. \end{cases} \tag{4}$$

It is also called the Hilbert cross-ratio metric distance [20, 35]. Notice that we take the absolute value of the logarithm since the Hilbert distance is a *signed distance* [56]. When $\mathscr{C}$ is the unit ball, HG lets us recover the Klein hyperbolic geometry [35]. When $\mathscr{C}$ is a quadric bounded convex domain, we obtain the Cayley-Klein hyperbolic geometry [12] which can be studied with the Riemannian structure and corresponding metric distance called the curved Mahalanobis distances [42, 41]. Cayley-Klein hyperbolic geometries have negative curvature.

In Hilbert geometry, the geodesics are *straight* Euclidean lines making it convenient for computation. Furthermore, the domain bounday $\partial\mathscr{C}$ need not to be smooth: One may also consider bounded polytopes [11]. This is particularly interesting for modeling Δ_d, the d-dimensional open standard simplex. We call this geometry the *Hilbert simplex geometry*. In Fig. (2), we show that the Hilbert distance between two multinomial distributions p (M) and q (M') can be computed by finding the two intersection points of the line $(1-t)p+tq$ with $\partial\Delta_d$, denoted as $t_0 \leq 0$ and $t_1 \geq 1$. Then

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

The shape of balls in polytope-domain HG is Euclidean polytopes [35], as depicted in Figure 3. Furthermore, the Euclidean shape of the balls do not change with the radius. Hilbert balls have hexagons shapes in 2D [48], rhombic dodecahedra shapes in 3D, and are polytopes [35] with $d(d+1)$ facets in dimension d. When the polytope domain is not a simplex, the combinatorial complexity of balls depends on the center location [48], see Figure 4. The HG of the probability simplex yields a non-Riemannian metric geometry, because at infinitesimal radius value, the balls are polytopes and not ellipsoidal balls (corresponding to squared Mahalanobis distance balls used to visualize metric tensors, [33]). The isometries in Hilbert polyhedral geometries are studied by [36]. In Appendix 9, we recall that any Hilbert geometry induces a Finslerian structure that becomes Riemannian iff the boundary is an ellipsoid (yielding the hyperbolic Cayley-Klein geometries, [56]). Let us notice that in Hilbert simplex/polytope geometry, the geodesics are not unique (see Figure 2 of [20]).

Table 1 summarizes the characteristics of the three introduced geometries: FHR, IG, and HG.

Let us conclude this introduction by mentioning the Cramér-Rao lower bound and its relationship with information geometry [39]: Consider an unbiased estimator $\hat{\theta} = T(X)$ of a parameter θ estimated from measurements distribution according to a smooth density $p(x;\theta)$ (i.e., $X \sim p(x;\theta)$). The Cramér-Rao Lower Bound (CRLB) states that the variance of $T(X)$ is greater or equal to the inverse of the Fisher Information Matrix (FIM) $I(\theta)$: $V_\theta[t(X)] \succ I^{-1}(\theta)$. For regular parametric families $\{p(x;\theta)\}_\theta$, the FIM is a positive-definite matrix and defines a metric ten-

Table 1: Comparing the three geometric modelings of the probability simplex Δ_d.

	Riemannian Geometry	**Information Rie. Geo.**	**Non-Rie. Hilbert Geo.**
Structure	$(\Delta_d, g, \nabla^{\mathrm{LC}} = \nabla(g))$	$(\Delta_d, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$	(Δ_d, ρ)
	Levi-Civita $\nabla^{\mathrm{LC}} = \nabla^{(0)}$	dual connections $\nabla^{(\pm\alpha)}$ so that $\frac{\nabla^{(\alpha)}+\nabla^{(-\alpha)}}{2} = \nabla^{(0)}$	connection of $\mathbb{R}^d$
Distance	Rao distance (metric)	α-divergence (non-metric) KL or reverse KL for $\alpha = \pm 1$	Hilbert distance (metric)
	invariant by reparameterization	information monotonicity	isometric to a normed space
Calculation	closed-form	closed-form	easy (Alg. 1)
Geodesic	minimizes length	straight either in θ/η	straight
Smoothness	manifold	manifold	non-manifold
Curvature	positive	dually flat	negative feature

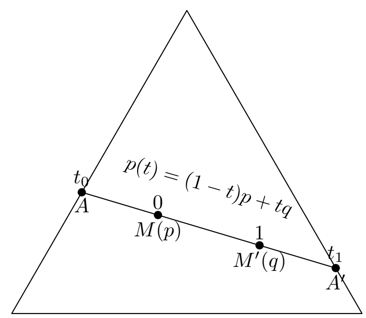


Fig. 2: Computing the Hilbert distance for trinomials on the 2D probability simplex.

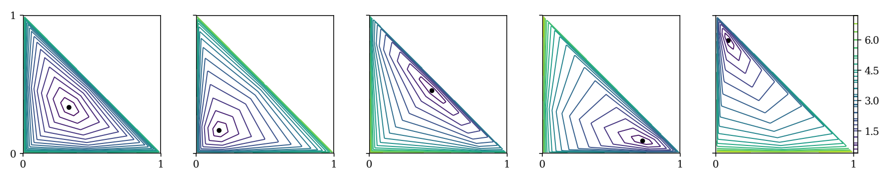


Fig. 3: Balls in the Hilbert simplex geometry Δ_2 have polygonal Euclidean shapes of constant combinatorial complexity. At infinitesimal scale, the balls have hexagonal shapes thus showing that the Hilbert geometry is not Riemannian.

sor, called the Fisher metric in Riemannian geometry. The FIM is the cornerstone of information geometry [2] but requires the differentiability of the probability density function (pdf).

A better lower bound that does not require the pdf differentiability is the Hammersley-Chapman-Robbins Lower Bound [26, 18] (HCRLB):

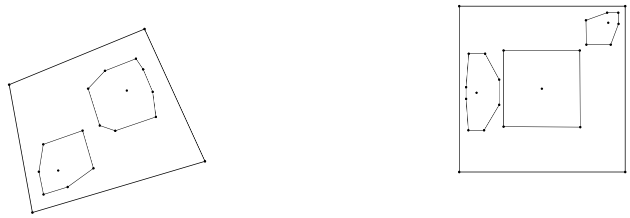


Fig. 4: Hilbert Balls in quadrangle domains have combinatorial complexity depending on the center location.

$$V_\theta[T(X)] \geq \sup_\Delta \frac{\Delta^2}{E_\theta\left[\left(\frac{p(x;\theta+\Delta)-p(x;\theta)}{p(x;\theta)}\right)^2\right]}. \quad (5)$$

Introducing the χ^2-divergence, $\chi^2(P:Q) = \int\left(\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)\mathrm{d}Q$, we rewrite the HCRLB using the χ^2-divergence in the denominator as follows:

$$V_\theta[T(X)] \geq \sup_\Delta \frac{\Delta^2}{\chi^2(P(x;\theta+\Delta):P(x;\theta))}. \quad (6)$$

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

3 Computing Hilbert distance in Δ_d

Let us first start by the simplest case: The 1D probability simplex Δ_1, the space of Bernoulli distributions. Any Bernoulli distribution can be represented by the activation probability of the random bit x: $\lambda = p(x=1) \in \Delta_1$ corresponding to a point in the interval $\Delta_1 = (0,1)$. We write the Bernoulli manifold as an exponential family as

$$p(x) = \exp\left(x\theta - F(\theta)\right), \quad x \in \{0,1\},$$

where $F(\theta) = \log(1+\exp(\theta))$. Therefore $\lambda = \frac{\exp(\theta)}{1+\exp(\theta)}$ and $\theta = \log\frac{\lambda}{1-\lambda}$.

3.1 1D probability simplex of Bernoulli distributions

By definition, the Hilbert distance has the closed form:

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

Note that $\theta_p=\log\frac{\lambda_p}{1-\lambda_p}$ is the canonical parameters of the Bernoulli distribution.

The FIM of the Bernoulli manifold in the λ-coordinates is given by: $g=\frac{1}{\lambda}+\frac{1}{1-\lambda}=\frac{1}{\lambda(1-\lambda)}$. The FHR distance is obtained by integration as:

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

Notice that $\rho_{\mathrm{FHR}}(p,q)$ has finite values on $\partial\Delta_1$.

The KL divergence of the ± 1-geometry is:

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

The KL divergence belongs to the family of α-divergences [2].

3.2 Arbitrary dimension case

Given $p,q\in\Delta_d$, we first need to compute the intersection of line (pq) with the border of the d-dimensional probability simplex to get the two intersection points p' and q' so that p',p,q,q' are ordered on (pq). Once this is done, we simply apply the formula in Eq. 4 to get the Hilbert distance.

A d-dimensional simplex consists of $d+1$ vertices with their corresponding $(d-1)$-dimensional facets. For the probability simplex Δ_d, let $e_i=(\underbrace{0,\ldots,0}_{i},1,0,\ldots,0)$ denote the $d+1$ vertices of the standard simplex embedded in the hyperplane $H_\Delta:\sum_{i=0}^{d}\lambda^i=1$ in $\mathbb{R}^{d+1}$. Let $f_{\setminus j}$ denote the simplex facets that is the convex hull of all points e_i except e_j: $f_{\setminus j}=\mathrm{hull}(e_0,\ldots,e_{j-1},e_{j+1},\ldots,e_d)$. Let $H_{\setminus j}$ denote the hyperplane supporting this facet, which is the affine hull $f_{\setminus j}=\mathrm{affine}(e_0,\ldots,e_{j-1},e_{j+1},\ldots,e_d)$.

To compute the two intersection points of (pq) with Δ_d, a naive algorithm consists in computing the unique intersection point r_j of the line (pq) with each hyperplane $H_{\setminus j}$ $(j=0,\cdots,d)$ and checking whether r_j belongs to $f_{\setminus j}$.

A more efficient implementation given by Alg. (1) calculates the intersection point of the line $x(t)=(1-t)p+tq$ with each $H_{\setminus j}$ $(j=0,\cdots,d)$. These intersection points are represented using the coordinate t. For example, $x(0)=p$, $x(1)=q$. Due to convexity, any intersection point with $H_{\setminus j}$ must satisfy either $t\le 0$ or $t\ge 1$. Then, the two intersection points with $\partial\Delta_d$ are obtained by $t_0=\max\{t:t\le 0\}$ and $t_1=\min\{t:t\ge 1\}$. This algorithm only requires $O(d)$ time.

Lemma 1. *The Hilbert distance in the probability simplex can be computed in optimal $\Theta(d)$ time.*

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
  $t_0 \leftarrow -\infty$; $t_1 \leftarrow +\infty$;
  for $i = 0 \cdots d$ **do**
    if $\lambda_p^i \neq \lambda_q^i$ **then**
      $t \leftarrow \lambda_p^i / (\lambda_p^i - \lambda_q^i)$;
      if $t_0 < t \leq 0$ **then**
        $t_0 \leftarrow t$;
      else if $1 \leq t < t_1$ **then**
        $t_1 \leftarrow t$;
  if $t_0 = -\infty$ *or* $t_1 = +\infty$ **then**
    Output $\rho_{\mathrm{HG}}(p,q) = 0$;
  else if $t_0 = 0$ *or* $t_1 = 1$ **then**
    Output $\rho_{\mathrm{HG}}(p,q) = \infty$;
  else
    Output $\rho_{\mathrm{HG}}(p,q) = \left|\log(1 - \frac{1}{t_0}) - \log(1 - \frac{1}{t_1})\right|$;

Once an arbitrary distance ρ is chosen, we can define a ball centered at c and of radius r as $B_\rho(c,r) = \{x \ : \ \rho(c,x) \leq r\}$. Figure 3 displays the hexagonal shapes of the Hilbert balls for various center locations in Δ_2.

Theorem 1 (Balls in a simplicial Hilbert geometry, [35]). *A ball in a Hilbert simplex geometry has a Euclidean polytope shape with $d(d+1)$ facets.*

Note that when the domain is not simplicial, the Hilbert balls can have varying combinatorial complexity depending on the center location. In 2D, the Hilbert ball can have $s \sim 2s$ edges, where s is the number of edges of the boundary of the Hilbert domain $\partial\mathscr{C}$.

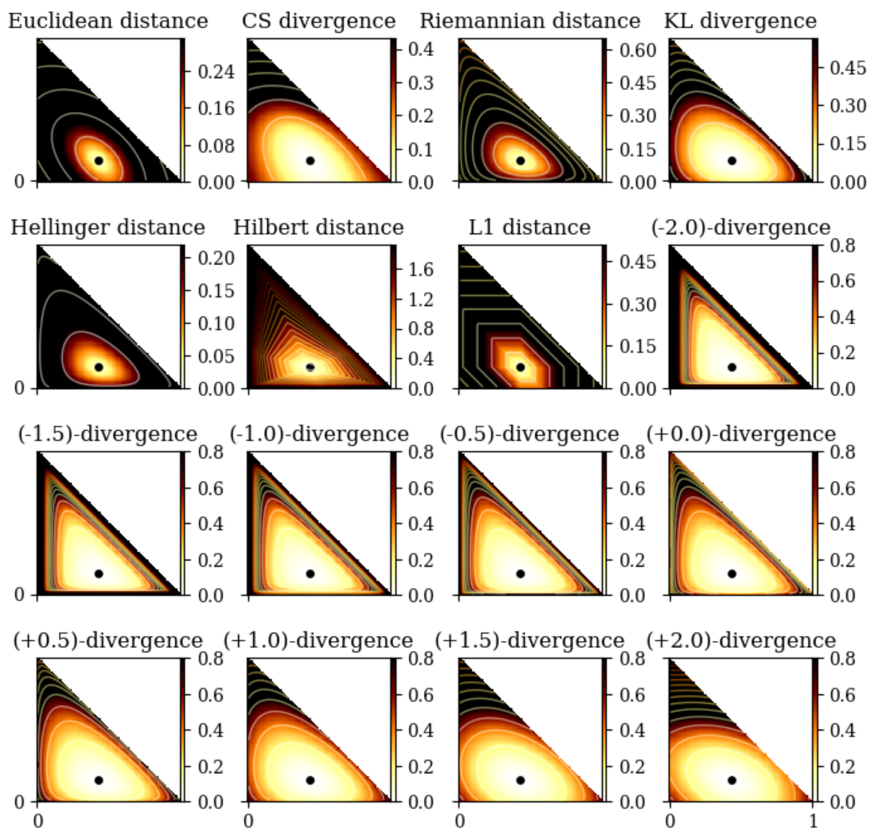
Since a Riemannian geometry is locally defined by a metric tensor, at infinitesimal scales, Riemannian balls have Mahalanobis smooth ellipsoidal shapes: $B_\rho(c,r) = \{x \ : \ (x-c)^\top g(c)(x-c) \leq r^2\}$. This property allows one to visualize Riemannian metric tensors [33]. Thus we conclude that:

Lemma 2 ([35]). *Hilbert simplex geometry is a non-manifold metric length space.*

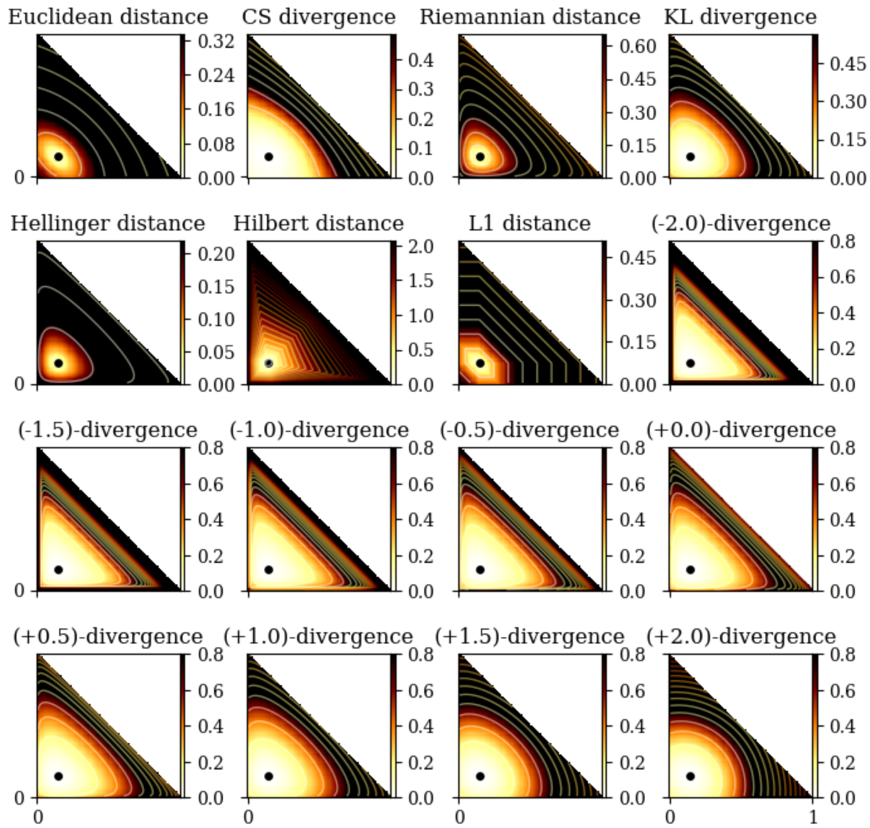
As a remark, let us notice that slicing a simplex with a hyperplane does not always produce a lower-dimensional simplex. For example, slicing a tetrahedron by a plane yields either a triangle or a quadrilateral. Thus the restriction of a d-dimensional ball B in a Hilbert simplex geometry Δ_d to a hyperplane H is a $(d-1)$-dimensional ball $B' = B \cap H$ of varying combinatorial complexity, corresponding to a ball in the induced Hilbert sub-geometry in the convex sub-domain $H \cap \Delta_d$.

3.3 Visualizing distance profiles

Figure 5 displays the distance profile from any point in the probability simplex to a fixed reference point (trinomial) for the following common distance measures [15]: Euclidean distance (metric), Cauchy-Schwarz (CS) divergence, Hellinger distance (metric), Fisher-Rao distance (metric), KL divergence and Hilbert simplicial distance (metric). The Euclidean and Cauchy-Schwarz divergence are clipped to Δ_2. The Cauchy-Schwarz distance is a projective distance so that $\rho_{\mathrm{CS}}(\lambda p, \lambda' q) = \rho_{\mathrm{CS}}(p,q)$ for any $\lambda, \lambda' > 0$. See [49].



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



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

Fig. 5: A comparison of different distance measures on Δ_2. The distance is measured from $\forall p \in \Delta_2$ to a fixed reference point (the black dot). Lighter color means shorter distance. Darker color means longer distance. The contours show equal distance curves with a precision step of 0.2.

4 Center-based clustering

We concentrate on comparing the efficiency of Hilbert simplex geometry for clustering multinomials. We shall compare the experimental results of k-means++ and k-center multinomial clustering for the three distances: Rao and Hilbert metric distances, and KL divergence. We describe how to implement those clustering algorithms when dealing with a Hilbert distance.

4.1 *k-means++ clustering*

The celebrated k-means clustering minimizes the sum of cluster variances, where each cluster has a center representative element. When dealing with $k = 1$ cluster, the center (also called centroid or cluster prototype) is the center of mass. For an arbitrary dissimilarity measure $D(\cdot : \cdot)$, the centroid c defined as the minimizer of

$$E_D(\Lambda, c) = \frac{1}{n} \sum_{i=1}^{n} D(\lambda_i : c)$$

may not be available in closed form. Nevertheless, using a generalization of the k-means initialization [6] (picking randomly seeds), one can bypass the centroid computation, and yet guarantee probabilistically a good clustering.

Let $C = \{c_1, \ldots, c_k\}$ denote the set of k cluster centers. Then the generalized k-means energy to minimize is defined by:

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

By defining the distance $D(\lambda, C) = \min_{j \in \{1,\ldots,k\}} D(\lambda : c_j)$ of a point to a set, we can rewrite the objective function as $E_D(\Lambda, C) = \frac{1}{n} \sum_{i=1}^{n} D(\lambda_i, C)$. Let $E_D^*(\Lambda, k) = \min_{C \,:\, |C|=k} E_D(\Lambda, C)$ denote the global minimum of $E_D(\Lambda, C)$ wrt some given Λ and k.

The k-means++ seeding proceeds for an arbitrary divergence D as follows: Pick uniformly at random at first seed c_1, and then iteratively choose the $(k-1)$ remaining seeds according to the following probability distribution:

$$\Pr(c_i = x) = \frac{D(x, \{c_1, \ldots, c_{i-1}\})}{\sum_{y \in \mathscr{X}} D(y, \{c_1, \ldots, c_{i-1}\})}.$$

Since its inception (2007), this k-means++ seeding has been extensively studied [7]. We state the general theorem established by [46]:

Theorem 2 (Generalized k-means++ performance, [46]). *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:

$$D(x : y) \leq \kappa_2 D(y : x), \quad \forall x, y \in \Delta_d.$$

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

$$E_D(\Lambda, C) \leq 2\kappa_1^2 (1 + \kappa_2)(2 + \log k) E_D^*(\Lambda, k). \tag{7}$$

The ratio $\frac{E_D(\Lambda,C)}{E_D^*(\Lambda,k)}$ is called the *competitive factor*. The seminal result of ordinary k-means++ was shown [6] to be $8(2+\log k)$-competitive. While evaluating κ_1, one has to note that squared metric distances are not metric because they do not satisfy the triangular inequality. For example, the squared Euclidean distance is not a metric but it satisfies the 2-quasi triangular inequality ($\kappa_1 = 2$).

We state the following general performance theorem:

Theorem 3 (k-means++ performance in a metric space). *In any metric space $(\mathscr{X}, d)$, the k-means++ with respect to d^2 is $16(2+\log k)$-competitive.*

Proof. Since a metric distance is symmetric, it follows that $\kappa_2 = 1$. Consider the quasi-triangular inequality property for the squared non-metric dissimilarity d^2:

$$d(p,q) \leq d(p,q) + d(q,r), \tag{8}$$

$$d^2(p,q) \leq (d(p,q) + d(q,r))^2, \tag{9}$$

$$d^2(p,q) \leq d^2(p,q) + d^2(q,r) + 2d(p,q)d(q,r). \tag{10}$$

Let us use the inequality of arithmetic and geometric means[1] to bound:

$$\sqrt{d^2(p,q)d^2(q,r)} \leq \frac{d^2(p,q)+d^2(q,r)}{2} \tag{11}$$

Thus we have

$$d^2(p,q) \leq d^2(p,q) + d^2(q,r) + 2d(p,q)d(q,r) \leq 2(d^2(p,q) + d^2(q,r)). \tag{12}$$

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

Theorem 4 (k-means++ performance in a normed space). *In any normed space $(\mathscr{X}, \|\cdot\|)$, the k-means++ with $D(x:y) = \|x-y\|^2$ is $16(2+\log k)$-competitive.*

Proof. In any normed space $(\mathscr{X}, \|\cdot\|)$, due to the triangle inequality, we have

$$\|x-z\| \leq \|x-y\| + \|y-z\|.$$

Squaring this inequality, we get

$$\|x-z\|^2 \leq (\|x-y\| + \|y-z\|)^2 = \|x-y\|^2 + \|y-z\|^2 + 2\|x-y\|\|y-z\|$$

Using the arithmetic-geometric inequality, we have

$$\|x-y\|\|y-z\| \leq \frac{\|x-y^2 + \|y-z^2}{2}$$

Therefore it comes that

[1] For positive values a and b, the arithmetic-geometric mean inequality states that $\sqrt{ab} \leq \frac{a+b}{2}$.

$$\|x-z\|^2 \leq (\|x-y\|+\|y-z\|)^2 \leq 2(\|x-y\|^2+\|y-z\|^2).$$

Therefore $\|\cdot\|^2$ satisfies the 2-quasi triangular inequality ($\kappa_1 = 2$). Furthermore, we have $\|x-y\|^2 = \|y-x\|^2$ ($\kappa_2 = 1$). Plugging $\kappa_1 = 2$ and $\kappa_2 = 1$ into Eq. 7, we get the $16(2+\log k)$-competitive factor.

Since any inner product space $(\mathscr{X}, \langle\cdot,\cdot\rangle)$ has an induced norm $\|x\| = \sqrt{\langle x,x\rangle}$, we have the following corollary.

Corollary 1. *In any inner product space* $(\mathscr{X}, \langle\cdot,\cdot\rangle)$*, the k-means++ with* $D(x:y) = \langle x-y, x-y\rangle$ *is* $16(2+\log k)$*-competitive.*

We need to report a bound for the squared Hilbert symmetric distance ($\kappa_2 = 1$). In ([35], Theorem 3.3), it was shown that Hilbert geometry of a bounded convex domain $\mathscr{C}$ is isometric to a normed vector space iff $\mathscr{C}$ is an open simplex: $(\Delta_d, \rho_{\mathrm{HG}}) \simeq (V_d, \|\cdot\|_H)$, where $\|\cdot\|_H$ is the corresponding norm. Therefore $\kappa_1 = 2$. We write NH for short for this equivalent normed Hilbert geometry. Appendix 8 recalls the construction due to [20], and shows the squared Hilbert distance fails the triangle inequality and it is not a distance induced by an inner product.

As an empirical study, we randomly generate $n = 10^5$ tuples (x,y,z) based on the uniform distribution in Δ_d. For each tuple (x,y,z), we evaluate the ratio

$$\kappa_1 = \frac{D(x:z)}{D(x:y)+D(y:z)}.$$

Figure 6 shows the statistics for three different choices of D: (1) $D(x:y) = \rho_{\mathrm{FHR}}^2(x,y)$; (2) $D(x:y) = \frac{1}{2}\mathrm{KL}(x:y) + \frac{1}{2}\mathrm{KL}(y:x)$; (3) $D(x:y) = \rho_{\mathrm{HG}}^2(x,y)$. We find experimentally that κ_1 is upper bounded by 2 for both ρ_{FHR} and ρ_{HG}, while the average κ_1 value is smaller than 0.5. For all the compared distances, $\kappa_2 = 1$. Therefore ρ_{FHR} and ρ_{HG} have better k-means++ performance guarantee as compared to ρ_{IG}.

We get by applying Theorem 4:

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

Figure 7 displays the results of a k-means++ clustering in Hilbert simplex geometry for $k \in \{3,5\}$.

The KL divergence can be interpreted as a separable Bregman divergence [1]. The Bregman k-means++ performance has been studied in [1, 38], and a competitive factor of $O(\frac{1}{\mu})$ is reported using the notion of Bregman μ-similarity (that is suited for data-sets on a compact domain).

In [23], spherical k-means++ is studied with respect to the distance $d_S(x,y) = 1-\langle x,y\rangle$ for any pair of points x,y on the unit sphere. Since $\langle x,y\rangle = \|x\|_2\|y\|_2\cos(\theta_{x,y}) = \cos(\theta_{x,y})$, we have $d_S(x,y) = 1-\cos(\theta_{x,y})$, where $\theta_{x,y}$ denotes the angle between a pair of unit vectors x and y. This distance is called the cosine distance since it amounts to one minus the cosine similarity. Notice that the cosine distance is related to the squared Euclidean distance via the identity:

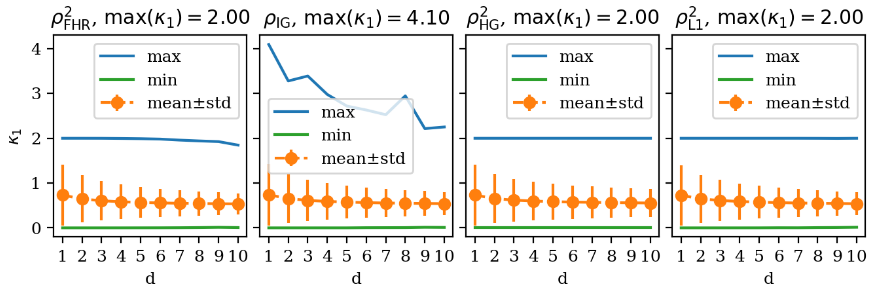
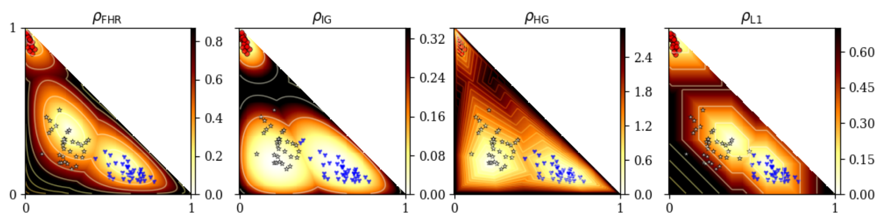
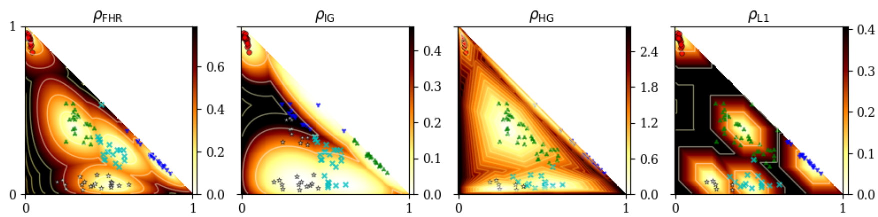


Fig. 6: The maximum, mean, standard deviation, and minimum of κ_1 on 10^6 randomly generated tuples (x,y,z) in Δ_d for $d = 1,\ldots,10$.



(a) $k = 3$ clusters



(b) $k = 5$ clusters

Fig. 7: k-Means++ clustering results on a toy dataset in the space of trinomials Δ_2. The color density maps indicate the distance from any point to its nearest cluster center.

$d_S(x,y) = \frac{1}{2}\|x-y\|^2$. The cosine distance is different from the spherical distance that relies on the arccos function.

Since divergences may be asymmetric, one can further consider mixed divergence $M(p:q:r) = \lambda D(p:q) + (1-\lambda)D(q:r)$ for $\lambda \in [0,1]$, and extend the k-means++ seeding procedure and analysis [47].

For a given data set, we can compute κ_1 or κ_2 by inspecting triples and pairs of points, and get data-dependent competitive factor improving the bounds mentioned above.

4.2 *k*-Center clustering

Let $\mathscr{X}$ be a finite point set. The cost function for a k-center clustering with centers $\mathscr{C}$ ($|\mathscr{C}| = k$) is:

$$f_D(\mathscr{X}, \mathscr{C}) = \max_{x \in \mathscr{X}} \min_{y \in \mathscr{C}} D(x : y).$$

The farthest first traversal heuristic [25] has a guaranteed approximation factor of 2 for any metric distance (see Algorithm 2).

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

Data: A set $\mathscr{X}$, a number k of clusters and a metric ρ
Result: A 2-approximation of the k-center clustering

begin
- $c_1 \leftarrow \mathrm{RandomPointOf}(\mathscr{X})$;
- $\mathscr{C} \leftarrow \{c_1\}$;
- **for** $i = 2, \cdots, k$ **do**
  - $c_i \leftarrow \arg\max_{x \in \mathscr{X}} \rho(x, \mathscr{C})$;
  - $\mathscr{C} \leftarrow \mathscr{C} \cup \{c_i\}$;

Output $\mathscr{C}$;

In order to use the k-center clustering algorithm described in Algorithm 3, we need to be able to compute a 1-center (or minimax center) for the Hilbert simplex geometry, that is the Minimum Enclosing Ball (MEB, also called the Smallest Enclosing Ball, SEB).

We may consider the SEB equivalently either in Δ_d or in the normed space V_d. In both spaces, the shapes of the balls are convex. Let $\mathscr{X} = \{x_1, \ldots, x_n\}$ denote the point set in Δ_d, and $\mathscr{V} = \{v_1, \ldots, v_n\}$ the equivalent point set in the normed vector space (following the mapping explained in Appendix 8). Then the SEBs $B_{\mathrm{HG}}(\mathscr{X})$ in Δ_d and $B_{\mathrm{NH}}(\mathscr{V})$ in V_d have radii r^*_{HG} and r^*_H defined by:

$$r^*_{\mathrm{HG}} = \min_{c \in \Delta_d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(x_i, c), \tag{13}$$

$$r^*_H = \min_{v \in V_d} \max_{i \in \{1,\ldots,n\}} \|v_i - v\|_H. \tag{14}$$

The SEB in the normed vector space $(V_d, \|\cdot\|_H)$ amounts to find the minimum covering norm polytope of a finite point set. This problem has been well-studied in computational geometry [57, 13, 51]. By considering the equivalent Hilbert norm polytope with $d(d+1)$ facets, we state the result of [57]:

Theorem 5 (SEB in Hilbert polytope normed space, [57]). *A $(1+\varepsilon)$-approximation of the SEB in V_d can be computed in $O(d^3 \frac{n}{\varepsilon})$.*

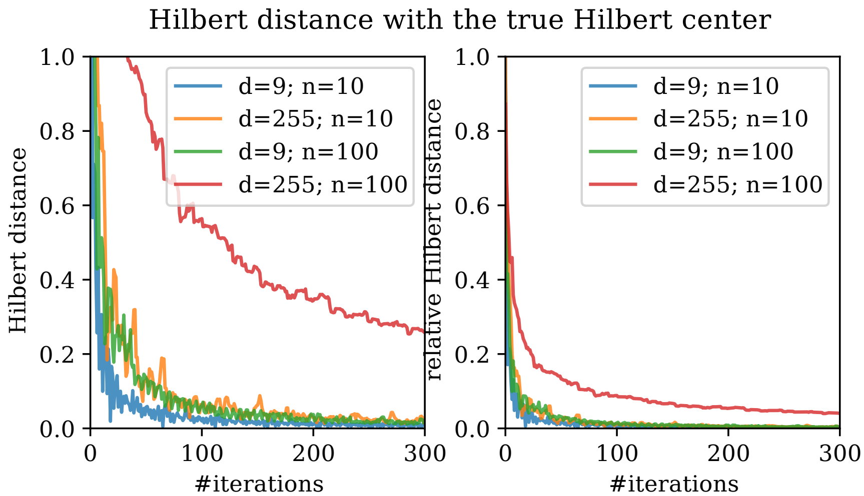


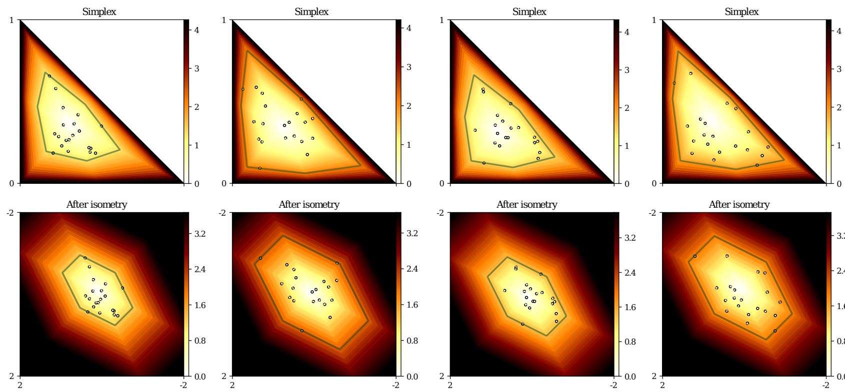
Fig. 8: Convergence rate of Alg. (4) measured by the Hilbert distance between the current minimax center and the true center (left) or their Hilbert distance divided by the Hilbert radius of the dataset (right). The plot is based on 100 random points in Δ_9/Δ_{255}.

Algorithm 3: k-center clustering

Data: A set of points $p_1, \cdots, p_n \in \Delta_d$. A distance measure ρ on Δ_d. The maximum number k of clusters. The maximum number T of iterations.
Result: A clustering scheme assigning each p_i a label $l_i \in \{1, \ldots, k\}$

```
begin
    Randomly pick k cluster centers c_1, ..., c_k using the kmeans++ heuristic;
    for t = 1, ..., T do
        for i = 1, ..., n do
            l_i ← argmin_{l=1}^{k} ρ(p_i, c_l);
        for l = 1, ..., k do
            c_l ← argmin_c max_{i: l_i = l} ρ(p_i, c);
    Output {l_i}_{i=1}^{n};
```

We shall now report two algorithms for computing the SEBs: One exact algorithm in V_d that do not scale well in high dimensions, and one approximation algorithm in Δ_d that works well for large dimensions.



(a) 3 points on the border

(b) 2 points on the border

Fig. 9: Computing the smallest enclosing ball in Hilbert simplex geometry amounts to compute the smallest enclosing ball in the corresponding normed vector space.

4.2.1 Exact smallest enclosing ball in a Hilbert simplex geometry

Given a finite point set $\{x_1,\ldots,x_n\} \in \Delta_d$, we define the Smallest Enclosing Ball (SEB) in Hilbert simplex geometry as:

$$r^* = \min_{c\in\Delta_d} \max_{i\in\{1,\ldots,n\}} \rho_{\mathrm{HG}}(c,x_i). \tag{15}$$

The radius of the SEB is r^*.

Consider the equivalent problem of finding the SEB in the isometric normed vector space via the mapping reported in Appendix 8. To each simplex point x_i corresponds a point v_i in the normed vector space V_d.

Figure 9 displays some examples of the exact smallest enclosing balls in the Hilbert simplex geometry and the corresponding normed vector space.

To compute the SEB, one may also consider the generic LP-type randomized algorithm [44]. We notice that an enclosing ball for a point set in general position as a number k of points on the border of the ball, with $2 \leq k \leq \frac{d(d+1)}{2}$. Let $D = \frac{d(d+1)}{2}$ denote the varying size of the combinatorial basis: Then we can apply the LP-type framework (we check the axioms of locality and monotonicity, [58]) to solve efficiently for the SEBs.

Theorem 6 (Smallest Enclosing Hilbert ball is LP-type, [63, 58]). *The smallest enclosing Hilbert ball amounts to find the smallest enclosing ball in a vector space with respect to a polytope norm that can be solved using a LP-type randomized algorithm.*

The Enclosing Ball Decision Problem (EBDP, [45]) asks for a given value r, whether $r \geq r^*$ or not. The decision problem amounts to find whether a set $\{rB_V +$

$v_i\}$ of translates can be stabbed by a point [45]: That is, whether $\cap_{i=1}^n (rB_V + v_i)$ is empty or not. Since the translates are polytopes with $d(d+1)$ facets, this can be solved in linear time using *Linear Programming*.

Theorem 7 (Enclosing Hilbert Ball Decision Problem). *The decision problem to test whether $r \geq r^*$ or not can be solved by Linear Programming.*

This yields a simple scheme to approximate the optimal value r^*: Let $r_0 = \max_{i\in\{1,\dots,n\}} \|v_i - v_1\|_H$. Then $r^* \in [\frac{r_0}{2}, r_0] = [a_0, b_0]$. At stage i, perform a dichotomic search on $[a_i, b_i]$ by answering the decision problem for $r_i = \frac{b_i - a_i}{2}$, and update the radius range accordingly [45].

However, the LP-type randomized algorithm or the decision problem-based algorithm do not scale well with dimensions. Next, we introduce a simple approximation algorithm that relies on the fact that the line segment $[pq]$ is a geodesic in Hilbert simplex geometry. (Geodesics are not unique, see Figure 2 of [20])

4.2.2 Geodesic bisection approximation heuristic

Algorithm 4: Geodesic walk for approximating the Hilbert minimax center, generalizing [9]

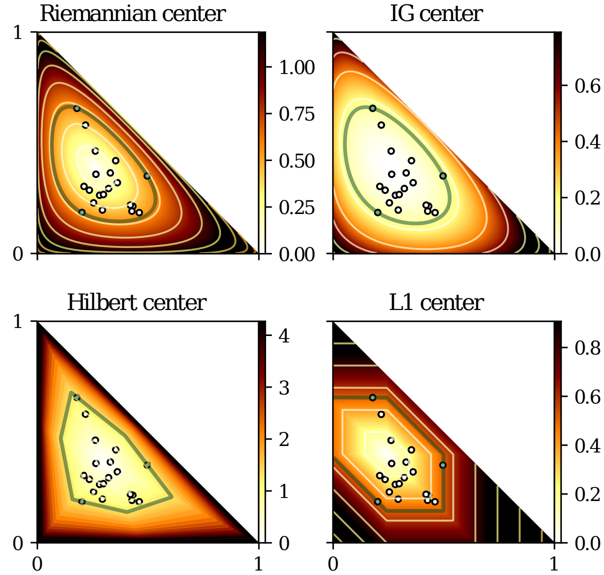
```
Data: A set of points p_1, ..., p_n ∈ Δ_d. The maximum number T of iterations.
Result: c = argmin_c max_i ρ_HG(p_i, c)
begin
    c_0 ← RandomPointOf({p_1, ..., p_n});
    for t = 1, ..., T do
        p ← argmax_{p_i} ρ_HG(p_i, c_{t-1});
        c_t ← c_{t-1} #^ρ_{t/(t+1)} p;
    Output c_T;
```

In Riemannian geometry, the 1-center can be arbitrarily finely approximated by a simple geodesic bisection algorithm [9, 5] This algorithm can be extended to HG straightforwardly as detailed in Algorithm 4:

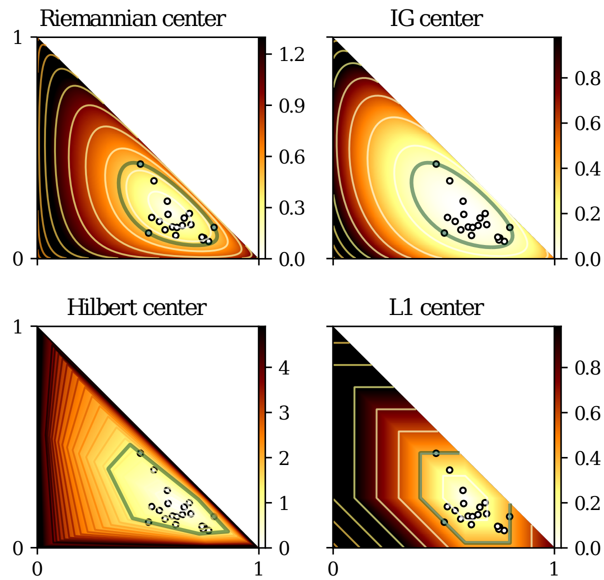
The algorithm first picks up a point c_0 at random from $\mathcal{X}$ for the initial center, computes the farthest point f_i (with respect to the distance ρ), and walk on the geodesic from c_0 to f_i by a certain amount to define c_1, etc. For an arbitrary distance ρ, we define the operator $\#_\alpha^\rho$ as follows:

$$p\#_\alpha^\rho q = v = \gamma(p,q,\alpha), \quad \rho(p:v) = \alpha\rho(p:q),$$

where $\gamma(p,q,\alpha)$ is the geodesic passing through p and q, and parameterized by α ($0 \leq \alpha \leq 1$). When the equations of the geodesics are explicitly known, we can

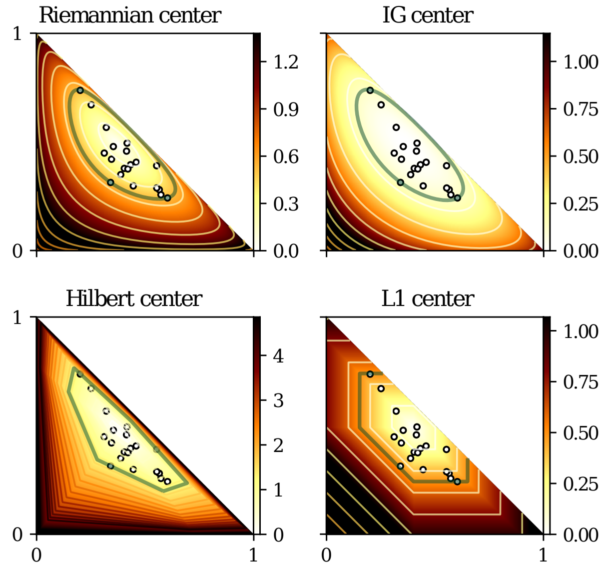


(a) Point Cloud 1



(b) Point Cloud 2

either get a closed form solution for $\#_\alpha^\rho$ or perform a bisection search on α to approximately compute α' such that $\rho(p:\gamma(p,q,\alpha'))=\alpha\rho(p:q)$. See [40] for an



(c) Point Cloud 3

Fig. 10: The Riemannian/IG/Hilbert (from left to right) minimax centers of three point clouds in Δ_2 based on Alg. (4). The color maps show the distance from $\forall p \in \Delta_2$ to the corresponding center.

extension and analysis in hyperbolic geometry. See Fig (8) to get an intuitive idea on the *experimental* convergence rate of Algorithm 4.

Furthermore, this iterative algorithm implies a core-set [10] (namely, the set of farthest points visited when iterating the geodesic walks) that is useful for clustering large data-sets [8]. See [13] for core-set results concerning containment problems with respect to a convex homothetic object (the equivalent Hilbert polytope norm in our case).

[51] described a simple algorithm dubbed MINCON for finding an approximation of the Minimum Enclosing Polytope. The algorithm induces a core-set of size $O(\frac{1}{\varepsilon^2})$ although the theorem is challenged in [13].

Thus by combining the k-center seeding of [25] with the iteration Lloyd-like batched iterations, we get an efficient k-center clustering algorithm for the FHR and Hilbert metric geometries. When dealing with the Kullback-Leibler divergence, we use the fact that KL is a Bregman divergence, and use the 1-center algorithm ([50, 43], approximation in any dimension, and [44], exact but limited to small dimensions).

Since Hilbert simplex geometry is isomorphic to a normed vector space [35] with a polytope norm with $d(d+1)$ facets, the Voronoi diagram in Hilbert geometry of Δ_d amounts to compute a Voronoi diagram with respect to a polytope norm [32, 55, 21].

5 Experiments

We generate a dataset consisting of a set of clusters in a high dimensional statistical simplex Δ_d. Each cluster is generated independently as follows. We first pick a random $c = (\lambda_c^0, \ldots, \lambda_c^d)$ based on the uniform distribution on Δ_d. Then we generate a random sample $p = (\lambda^0, \ldots, \lambda^d)$ based on

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

where $\sigma > 0$ is a noise level parameter, and each ε^i follows independently a standard Gaussian distribution (generator 1) or the Student's t-distribution with 5 degrees of freedom (generator 2). Let $\sigma = 0$, we get $\lambda^i = \lambda_c^i$. Therefore p is randomly distributed around c. We repeat generating random samples for each cluster center, and make sure that different clusters have almost the same number of samples. Then we run k-center clustering in Alg. (3) 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 true number of clusters to avoid model selection. For each configuration, we repeat the clustering experiment based on 300 different random datasets. The performance is measured by the normalized mutual information (NMI).

The results of k-menas++ and k-centers are shown in Table 2 and Table 3, respectively. The large variance of NMI is because that each experiment is performed on different datasets given by the same generator based on different random seeds. Generally, the performance deteriorates as we increase the number of clusters, increase the noise level or decrease the dimensionality, which have the same effect to reduce the gap among the clusters.

The key comparison is the three columns ρ_{FHR}, ρ_{HG} and ρ_{IG}, as they are based on exactly the same algorithm (k-center) with the only difference being the underlying geometry. We see clearly that the performance of the three compared geometries presents the order HG $>$ FHR $>$ IG. The performance of HG is superior to the other two geometries, especially when the noise level is large. Intuitively, the Hilbert balls are more compact and therefore can better capture the clustering structure (see Fig. (1)).

The column ρ_{EUC} represents k-center based on the Euclidean enclosing ball. It shows the worst scores because the intrinsic geometry of the probability simplex is far from being Euclidean.

6 Hilbert geometry of the space of correlation matrices

In this section, we present the Hilbert geometry to the space of correlation matrices

$$\mathscr{C}_d = \{C_{d\times d} : C \succ 0; C_{ii} = 1, \forall i\}.$$

Table 2: k-means++ clustering accuracy in percentage on randomly generated datasets based on different geometries. The table shows the mean and standard deviation after 300 independent runs for each configuration. ρ is the distance measure. n is sample size of multinomial distributions. d is the dimension of the statistical simplex. σ is noise level.

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	$\mathbf{0.81\pm0.22}$	0.64 ± 0.23	0.70 ± 0.22
			0.9	0.44 ± 0.20	0.44 ± 0.20	$\mathbf{0.57\pm0.22}$	0.31 ± 0.17	0.38 ± 0.18
		255	0.5	0.80 ± 0.24	0.81 ± 0.24	$\mathbf{0.88\pm0.21}$	0.74 ± 0.25	0.79 ± 0.24
			0.9	0.65 ± 0.27	0.66 ± 0.28	$\mathbf{0.72\pm0.27}$	0.46 ± 0.24	0.63 ± 0.27
	100	9	0.5	0.76 ± 0.22	0.76 ± 0.21	$\mathbf{0.82\pm0.22}$	0.60 ± 0.21	0.69 ± 0.23
			0.9	0.42 ± 0.19	0.41 ± 0.18	$\mathbf{0.54\pm0.22}$	0.27 ± 0.14	0.34 ± 0.16
		255	0.5	0.82 ± 0.23	0.82 ± 0.24	$\mathbf{0.89\pm0.20}$	0.74 ± 0.24	0.80 ± 0.25
			0.9	0.66 ± 0.26	0.66 ± 0.28	$\mathbf{0.72\pm0.26}$	0.45 ± 0.25	0.64 ± 0.27
5	50	9	0.5	0.75 ± 0.14	0.74 ± 0.15	$\mathbf{0.81\pm0.13}$	0.61 ± 0.13	0.68 ± 0.13
			0.9	0.44 ± 0.13	0.42 ± 0.13	$\mathbf{0.55\pm0.15}$	0.31 ± 0.11	0.36 ± 0.12
		255	0.5	0.83 ± 0.15	0.83 ± 0.15	$\mathbf{0.88\pm0.14}$	0.77 ± 0.16	0.82 ± 0.15
			0.9	0.71 ± 0.17	0.70 ± 0.19	$\mathbf{0.75\pm0.17}$	0.50 ± 0.17	0.68 ± 0.18
	100	9	0.5	0.74 ± 0.13	0.74 ± 0.14	$\mathbf{0.80\pm0.14}$	0.60 ± 0.13	0.67 ± 0.13
			0.9	0.42 ± 0.11	0.40 ± 0.12	$\mathbf{0.55\pm0.15}$	0.29 ± 0.09	0.35 ± 0.11
		255	0.5	0.83 ± 0.14	0.83 ± 0.15	$\mathbf{0.88\pm0.13}$	0.77 ± 0.15	0.81 ± 0.15
			0.9	0.69 ± 0.18	0.69 ± 0.18	$\mathbf{0.73\pm0.17}$	0.48 ± 0.17	0.67 ± 0.18

(a) generator 1

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	$\mathbf{0.71\pm0.23}$	0.45 ± 0.20	0.54 ± 0.22
			0.9	0.29 ± 0.17	0.27 ± 0.16	$\mathbf{0.39\pm0.19}$	0.17 ± 0.13	0.25 ± 0.15
		255	0.5	0.70 ± 0.25	0.69 ± 0.26	$\mathbf{0.74\pm0.25}$	0.37 ± 0.29	0.70 ± 0.26
			0.9	$\mathbf{0.42\pm0.25}$	0.35 ± 0.20	0.40 ± 0.19	0.03 ± 0.08	$\mathbf{0.44\pm0.26}$
	100	9	0.5	0.63 ± 0.22	0.61 ± 0.22	$\mathbf{0.71\pm0.22}$	0.46 ± 0.19	0.56 ± 0.20
			0.9	0.29 ± 0.15	0.26 ± 0.14	$\mathbf{0.38\pm0.20}$	0.18 ± 0.12	0.24 ± 0.14
		255	0.5	0.71 ± 0.26	0.69 ± 0.27	$\mathbf{0.75\pm0.25}$	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	$\mathbf{0.43\pm0.26}$
5	50	9	0.5	0.64 ± 0.15	0.61 ± 0.14	$\mathbf{0.70\pm0.14}$	0.48 ± 0.14	0.57 ± 0.15
			0.9	0.31 ± 0.12	0.29 ± 0.12	$\mathbf{0.41\pm0.15}$	0.20 ± 0.09	0.26 ± 0.10
		255	0.5	0.74 ± 0.17	0.72 ± 0.17	$\mathbf{0.77\pm0.16}$	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	$\mathbf{0.47\pm0.17}$
	100	9	0.5	0.62 ± 0.14	0.61 ± 0.14	$\mathbf{0.71\pm0.14}$	0.46 ± 0.13	0.54 ± 0.14
			0.9	0.30 ± 0.10	0.27 ± 0.11	$\mathbf{0.40\pm0.13}$	0.19 ± 0.08	0.25 ± 0.09
		255	0.5	0.73 ± 0.18	0.70 ± 0.18	$\mathbf{0.75\pm0.16}$	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	$\mathbf{0.46\pm0.18}$

(b) generator 2

Table 3: k-center clustering accuracy in percentage on randomly generated datasets based on different geometries. The table shows the mean and standard deviation after 300 independent runs for each configuration. ρ is the distance measure. n is sample size of multinomial distributions. d is the dimension of the statistical simplex. σ is noise level.

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

(a)

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

(b)

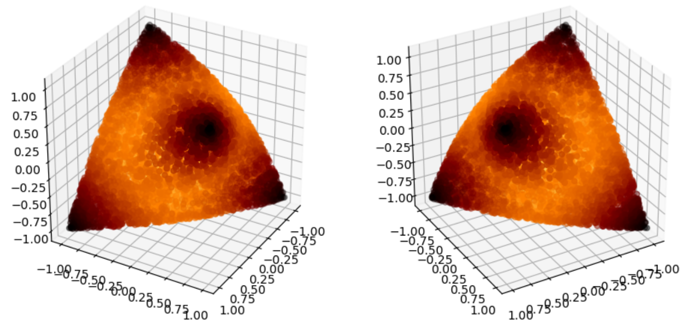


Fig. 11: The Elliptope $\mathscr{C}_3$ (two different perspectives).

If $C_1, C_2 \in \mathscr{C}$, then $(1-\lambda)C_1 + \lambda C_2 \in \mathscr{C}$ for $0 < \lambda < 1$. Therefore $\mathscr{C}$ is a convex set, known as an *elliptope* embedded in the p.s.d. cone. See fig. 11 for an intuitive view of $\mathscr{C}_3$, where the coordinate system (x, y, z) is the off-diagonal entries of $C \in \mathscr{C}_3$.

In order to compute the Hilbert distance $\rho_{\mathrm{HG}}(C_1, C_2)$, we need to compute the intersection of the line (C_1, C_2) with $\partial\mathscr{C}$, denoted as C_1' and C_2', then we have

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

Unfortunately there is no closed form solution of C_1' and C_2'. Instead, we apply a binary searching algorithm. Note a necessary condition for $C \in \mathscr{C}$ is that C has a positive spectrum. If C has at least one non-positive eigenvalue, then $C \notin \mathscr{C}$.

We compare the Hilbert elliptope geometry with commonly used distance measures including the L_2 distance ρ_{EUC}, L_1 distance ρ_{L1}, and the square root of the log-det divergence

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

Due to the high computational complexity, we only investigate k-means++ clustering. The investigated dataset consists of 100 matrices forming 3 clusters in $\mathscr{C}_3$ with almost identical size. Each cluster is independently generated according to

$$\begin{aligned} P &\sim \mathscr{W}^{-1}(I_{3\times 3}, \nu_1), \\ C_i &\sim \mathscr{W}^{-1}(P, \nu_2), \end{aligned}$$

where $\mathscr{W}^{-1}(A, \nu)$ denotes the inverse Wishart distribution with scale matrix A and ν degrees of freedom, and C_i is a point in this cluster. Table 4 shows the k-means++ clustering performance in terms of NMI. Again Hilbert geometry is favorable as compared to alternative measures, showing that the good performance of Hilbert clustering is not limited to the simplex case.

Table 4: NMI (mean±std) of k-means++ clustering based on different distance measures in the elliptope (500 independent runs)

v_1	v_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	$\mathbf{0.84 \pm 0.18}$

7 Conclusion

We introduced the Hilbert metric distance and its underlying non-Riemannian geometry for modeling the space of multinomials of the open probability simplex, and compared experimentally this geometry with the traditional differential-geometric modelings (either FHR metric connection or dually coupled non-metric affine connection of information geometry, [2]) for clustering tasks. The main feature of HG is that it is a metric non-manifold geometry where geodesics are straight (Euclidean) line segments. For simplex domains, the Hilbert balls have fixed combinatorial (Euclidean) polytope structures, and HG is known to be isometric to a normed space [20, 24]. This latter isometry allows one to generalize easily the standard proofs of clustering (*e.g.*, k-means or k-center). We demonstrated it for the k-means++ competitive performance analysis, and for the convergence of the 1-center heuristic [9] (smallest enclosing Hilbert ball allows one to implement efficiently the k-center clustering). Our experimental k-means++/k-center comparisons of HG algorithms with the manifold modeling approach yield striking superior performance: This may be explained by the sharpness of Hilbert balls with respect to the FHR/IG ball profiles.

Chentsov [17] defined statistical invariance on a probability manifold under Markov morphisms, and proved that the Fisher Information Metric (FIM) is the unique Riemannian metric (up to rescaling) for multinomials. However, this does not rule out that other distances (with underlying geometric structures) may be used to model statistical manifolds (*eg.*, Finsler statistical manifolds, [16, 59], or the total variation distance — the only metric f-divergence, [31]). Defining statistical invariance related to geometry is the cornerstone problem of information geometry that can be tackled from many directions (see [22] and references therein for a short review).

In this paper, we introduced Hilbert geometries in machine learning by considering clustering tasks in the probability simplex and in the elliptope. Hilbert geometries proved computationally handy since geodesics are straight lines. One future direction is to consider the Hilbert metric for regularization and sparsity in machine learning (due to its equivalence with a polytope normed distance).

Our Python codes are freely available online for reproducible research:
`https://www.lix.polytechnique.fr/~nielsen/HSG/`

References

1. Ackermann, M.R., Blömer, J.: Bregman clustering for separable instances. In: Scandinavian Workshop on Algorithm Theory, pp. 212–223. Springer (2010)
2. Amari, S.i.: Information Geometry and Its Applications, *Applied Mathematical Sciences*, vol. 194. Springer Japan (2016)
3. Amari, S.i., Cichocki, A.: Information geometry of divergence functions. Bulletin of the Polish Academy of Sciences: Technical Sciences **58**(1), 183–195 (2010)
4. Arnaudon, M., Nielsen, F.: Medians and means in Finsler geometry. LMS Journal of Computation and Mathematics **15**, 23–37 (2012)
5. Arnaudon, M., Nielsen, F.: On approximating the Riemannian 1-center. Computational Geometry **46**(1), 93–104 (2013)
6. Arthur, D., Vassilvitskii, S.: *k*-means++: The advantages of careful seeding. In: ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1027–1035 (2007)
7. Bachem, O., Lucic, M., Hassani, S.H., Krause, A.: Approximate *k*-means++ in sublinear time. In: AAAI, pp. 1459–1467 (2016)
8. Bachem, O., Lucic, M., Krause, A.: Scalable and distributed clustering via lightweight coresets **arXiv:1702.08248 [stat.ML]** (2017)
9. Bâdoiu, M., Clarkson, K.L.: Smaller core-sets for balls. In: ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 801–802 (2003)
10. Bădoiu, M., Clarkson, K.L.: Optimal core-sets for balls. Computational Geometry **40**(1), 14–22 (2008)
11. Bernig, A.: Hilbert geometry of polytopes. Archiv der Mathematik **92**(4), 314–324 (2009)
12. Bi, Y., Fan, B., Wu, F.: Beyond Mahalanobis metric: Cayley-Klein metric learning. In: CVPR, pp. 2339–2347 (2015)
13. Brandenberg, R., König, S.: No dimension-independent core-sets for containment under homothetics. Discrete & Computational Geometry **49**(1), 3–21 (2013)
14. Busemann, H.: The Geometry of Geodesics, *Pure and Applied Mathematics*, vol. 6. Elsevier Science (1955)
15. Calin, O., Udriste, C.: Geometric Modeling in Probability and Statistics. Mathematics and Statistics. Springer International Publishing (2014)
16. Cena, A.: Geometric structures on the non-parametric statistical manifold. Ph.D. thesis, University of Milano (2002)
17. Cencov, N.N.: Statistical Decision Rules and Optimal Inference, *Translations of mathematical monographs*, vol. 53. American Mathematical Society (2000)
18. Chapman, D.G., Robbins, H.: Minimum variance estimation without regularity assumptions. The Annals of Mathematical Statistics pp. 581–586 (1951)
19. Chaudhuri, K., McGregor, A.: Finding metric structure in information theoretic clustering. In: COLT, pp. 391–402 (2008)
20. de la Harpe, P.: On Hilbert's metric for simplices. In: Geometric Group Theory, vol. 1, pp. 97–118. Cambridge Univ. Press (1991)
21. Deza, M., Sikirić, M.D.: Voronoi polytopes for polyhedral norms on lattices. Discrete Applied Mathematics **197**, 42–52 (2015)
22. Dowty, J.G.: Chentsov's theorem for exponential families **arXiv:1701.08895 [math.ST]** (2017)
23. Endo, Y., Miyamoto, S.: Spherical *k*-means++ clustering. In: Modeling Decisions for Artificial Intelligence, pp. 103–114. Springer (2015)
24. Foertsch, T., Karlsson, A.: Hilbert metrics and Minkowski norms. Journal of Geometry **83**(1-2), 22–31 (2005)

25. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoretical Computer Science **38**, 293–306 (1985)
26. Hammersley, J.: On estimating restricted parameters. Journal of the Royal Statistical Society. Series B (Methodological) **12**(2), 192–240 (1950)
27. Hilbert, D.: Über die gerade linie als kürzeste verbindung zweier punkte. Mathematische Annalen **46**(1), 91–96 (1895)
28. Hotelling, H.: Spaces of statistical parameters. In: Bulletin AMS, vol. 36, p. 191 (1930)
29. Jenssen, R., Principe, J.C., Erdogmus, D., Eltoft, T.: The Cauchy-Schwarz divergence and Parzen windowing: Connections to graph theory and mercer kernels. Journal of the Franklin Institute **343**(6), 614–629 (2006)
30. Kass, R.E., Vos, P.W.: Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics. Wiley-Interscience (1997)
31. Khosravifard, M., Fooladivanda, D., Gulliver, T.A.: Confliction of the convexity and metric properties in f-divergences. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **90**(9), 1848–1853 (2007)
32. Körner, M.C.: Minisum hyperspheres, *Springer Optimization and Its Applications*, vol. 51. Springer New York (2011)
33. Laidlaw, D.H., Weickert, J.: Visualization and Processing of Tensor Fields: Advances and Perspectives. Mathematics and Visualization. Springer-Verlag Berlin Heidelberg (2009)
34. Lebanon, G.: Learning Riemannian metrics. In: UAI, pp. 362–369 (2002)
35. Lemmens, B., Nussbaum, R.: Birkhoffs version of Hilberts metric and its applications in analysis. Handbook of Hilbert Geometry pp. 275–303 (2014)
36. Lemmens, B., Walsh, C.: Isometries of polyhedral hilbert geometries. Journal of Topology and Analysis **3**(02), 213–241 (2011)
37. Liang, X.: A note on divergences. Neural Computation **28**(10), 2045–2062 (2016)
38. Manthey, B., Röglin, H.: Worst-case and smoothed analysis of k-means clustering with Bregman divergences. Journal of Computational Geometry **4**(1), 94–132 (2013)
39. Nielsen, F.: Cramér-Rao lower bound and information geometry. In: Connected at Infinity II, pp. 18–37. Springer (2013)
40. Nielsen, F., Hadjeres, G.: Approximating covering and minimum enclosing balls in hyperbolic geometry. In: International Conference on Networked Geometric Science of Information, pp. 586–594. Springer (2015)
41. Nielsen, F., Muzellec, B., Nock, R.: Classification with mixtures of curved Mahalanobis metrics. In: IEEE International Conference on Image Processing (ICIP), pp. 241–245 (2016)
42. Nielsen, F., Muzellec, B., Nock, R.: Large margin nearest neighbor classification using curved Mahalanobis distances **arXiv:1609.07082 [cs.LG]** (2016)
43. Nielsen, F., Nock, R.: On approximating the smallest enclosing Bregman balls. In: Proceedings of the twenty-second annual symposium on Computational geometry, pp. 485–486. ACM (2006)
44. Nielsen, F., Nock, R.: On the smallest enclosing information disk. Information Processing Letters **105**(3), 93–97 (2008)
45. Nielsen, F., Nock, R.: Approximating smallest enclosing balls with applications to machine learning. International Journal of Computational Geometry & Applications **19**(05), 389–414 (2009)
46. Nielsen, F., Nock, R.: Total Jensen divergences: Definition, properties and k-means++ clustering **arXiv:1309.7109 [cs.IT]** (2013)
47. Nielsen, F., Nock, R., Amari, S.i.: On clustering histograms with k-means by using mixed α-divergences. Entropy **16**(6), 3273–3301 (2014)
48. Nielsen, F., Shao, L.: On balls in a polygonal Hilbert geometry. In: 33st International Symposium on Computational Geometry (SoCG 2017). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2017). URL `https://www.youtube.com/watch?v=XE5x5rAK8Hk`
49. Nielsen, F., Sun, K., Marchand-Maillet, S.: On Hölder projective divergences. Entropy **19**(3) (2017)

50. Nock, R., Nielsen, F.: Fitting the smallest enclosing Bregman ball. In: ECML, pp. 649–656. Springer (2005)
51. Panigrahy, R.: Minimum enclosing polytope in high dimensions **arXiv:cs/0407020 [cs.CG]** (2004)
52. Papadopoulos, A., Troyanov, M.: From Funk to Hilbert geometry **arXiv:1406.6983 [math.MG]** (2014)
53. Rao, C.R.: Information and accuracy attainable in the estimation of statistical parameters. Bull. Cal. Math. Soc. **37**(3), 81–91 (1945)
54. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. In: Breakthroughs in statistics, pp. 235–247. Springer (1992)
55. Reem, D.: The geometric stability of Voronoi diagrams in normed spaces which are not uniformly convex **arXiv:1212.1094 [cs.CG]** (2012)
56. Richter-Gebert, J.: Perspectives on projective geometry: A guided tour through real and complex geometry. Springer-Verlag Berlin Heidelberg (2011)
57. Saha, A., Vishwanathan, S., Zhang, X.: New approximation algorithms for minimum enclosing convex shapes. In: ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1146–1160 (2011)
58. Sharir, M., Welzl, E.: A combinatorial bound for linear programming and related problems. STACS 92 pp. 567–579 (1992)
59. Shen, Z.: Riemann-Finsler geometry with applications to information geometry. Chinese Annals of Mathematics-Series B **27**(1), 73–94 (2006)
60. Shima, H.: The Geometry of Hessian Structures. World Scientific (2007)
61. Stigler, S.M., et al.: The epic story of maximum likelihood. Statistical Science **22**(4), 598–620 (2007)
62. Vernicos, C.: Introduction aux géométries de hilbert. Séminaire de théorie spectrale et géométrie **23**, 145–168 (2004)
63. Welzl, E.: Smallest enclosing disks (balls and ellipsoids). New results and new trends in computer science pp. 359–370 (1991)

8 Isometry of Hilbert simplex geometry to a normed vector space

Consider the Hilbert simplex metric space $(\Delta_d, \rho_{\mathrm{HG}})$ where Δ_d denotes the d-dimensional open probability simplex and ρ_{HG} the Hilbert cross-ratio metric. Let us recall the [20] isometry (1991) of the open standard simplex to a normed vector space $(V_d, \|\cdot\|_H)$. Let $V_d = \{v \in \mathbb{R}^d \ : \ \sum_i v^i = 0\}$ denote the d-dimensional vector space sitting in $\mathbb{R}^{d+1}$. Map a point $x = (x^1, \ldots, x^{d+1}) \in \Delta_d$ to a point $v(x) = (v^1, \ldots, v^{d+1}) \in V_d$ as follows:

$$v^i = \frac{1}{d+1}\left(d \log x^i - \sum_{j \neq i} \log x^j\right). \tag{16}$$

We define the corresponding norm $\|\cdot\|_H$ in V_d by considering the shape of its unit ball $B_V = \{v \in V_d \ : \ |v^i - v^j| \leq 1, \forall i \neq j\}$. The unit ball B_V is a symmetric convex set containing the origin in its interior, and thus yields a *polytope norm* $\|\cdot\|_H$ (Hilbert norm) with $2\binom{d+1}{2} = d(d+1)$ facets. Reciprocally, let us notice that a norm induces a unit ball centered at the origin that is convex and symmetric around the origin.

The distance in the normed vector space between $v \in V_d$ and $v' \in V_d$ is defined by:

$$\rho_V(v,v') = \|v-v'\|_H = \min_{\lambda}\{v' \in \lambda(B_V \oplus v)\}, \tag{17}$$

where $A \oplus B = \{a+b \ : \ a \in A, b \in B\}$ is the Minkowski sum.

The reverse map from the normed space V_d to the probability simplex Δ_d is given by:

$$x^i = \frac{1}{\sum_j \exp(v^j)} \exp(v^i). \tag{18}$$

Thus we have $(\Delta_d, \rho_{\mathrm{HG}}) \cong (V_d, \|\cdot\|_H)$. In 1D, $(V_1, \|\cdot\|_H)$ is isometric to the Euclidean line.

Note that computing the distance in the normed vector space will require naively $O(d^2)$ time.

Unfortunately, the norm $\|\cdot\|_H$ does not satisfy the parallelogram law.[2] Notice that a norm satisfying the parallelogram law can be associated an inner product via the polarization identity. Thus the isometry of the Hilbert geometry to a normed vector space is not equipped with an inner product. However, all norms in a finite dimensional space are equivalent. This implies that in finite dimension, $(\Delta_d, \rho_{\mathrm{HG}})$ is *quasi-isometric* to the Euclidean space $\mathbb{R}^d$. An example of Hilbert geometry in infinite dimension is reported in [20]. Hilbert spaces are not CAT spaces except when $\mathscr{C}$ is an ellipsoid [62].

9 Hilbert geometry with Finslerian/Riemannian structures

In a Riemannian geometry, each tangent plane T_pM of the d-dimensional manifold M is equivalent to $\mathbb{R}^d$: $T_pM \simeq \mathbb{R}^d$. The inner product at each tangent plane T_pM can be visualized by an ellipsoid shape, a convex symmetric object centered at point p. In a *Finslerian geometry*, a norm $\|\cdot\|_p$ is defined in each tangent plane T_pM, and this norm is visualized as a symmetric convex object with non-empty interior. Finslerian geometry thus generalizes Riemannian geometry by taking into account generic symmetric convex objects instead of ellipsoids for inducing norms at each tangent plane. Any Hilbert geometry induced by a compact convex domain $\mathscr{C}$ can be expressed by en equivalent Finslerian geometry by defining the norm in T_p at p as follows [62]:

$$\|v\|_p = F_{\mathscr{C}}(p,v) = \frac{\|v\|}{2}\left(\frac{1}{pp^+} + \frac{1}{pp^-}\right), \tag{19}$$

where $\|\cdot\|$ is an *arbitrary norm* on $\mathbb{R}^d$, and p^+ and p^- are the intersection points of the line passing through p with direction v:

---

[2] Consider $A = (1/3,1/3,1/3)$, $B = (1/6,1/2,1/3)$, $C = (1/6,2/3,1/6)$ and $D = (1/3,1/2,1/6)$. Then $2AB^2 + 2BC^2 = 4.34$ but $AC^2 + BD^2 = 3.84362411135$.

$$p^+ = p + t^+ v, \quad p^- = p + t^- v$$

$F_{\mathscr{C}}$ is the *Finsler metric*.

A geodesic γ in a Finslerian geometry satisfies:

$$d_{\mathscr{C}}(\gamma(t_1), \gamma(t_2)) = \int_{t_1}^{t_2} F_{\mathscr{C}}(\gamma(t), \dot{\gamma}(t))\mathrm{d}t. \tag{20}$$

In T_pM, a ball of center c and radius r is defined by:

$$B(c,r) = \{v \ : \ F_{\mathscr{C}}(c,v) \leq r\}. \tag{21}$$

Thus any Hilbert geometry induces an equivalent Finslerian geometry, and since Finslerian geometries include Riemannian geometries, one may wonder which Hilbert geometries induce Riemannian structures? The only Riemannian geometries induced by Hilbert geometries are the *hyperbolic Cayley-Klein geometries* [56, 42, 41] with the domain $\mathscr{C}$ being an ellipsoid. The Finslerian modeling of information geometry has been studied in [16, 59].

There is not a canonical way of defining measures in a Hilbert geometry since Hilbert geometries are Finslerian but not necessary Riemannian geometries [62]. The Busemann measure is defined according to the Lebesgue measure λ of $\mathbb{R}^d$: Let B_p denote the unit ball wrt. to the Finsler norm at point $p \in \mathscr{C}$, and B_e the Euclidean unit ball. Then the Busemann measure for a Borel set $\mathscr{B}$ is defined by [62]:

$$\mu_{\mathscr{C}}(\mathscr{B}) = \int_{\mathscr{B}} \frac{\lambda(B_e)}{\lambda(B_p)} \mathrm{d}\lambda(p).$$

The existence and uniqueness of center points of a probability measure in Finsler geometry have been investigated in [4].

10 Bounding Hilbert norm with other norms

Let us show that $\|v\|_H \leq \beta_{d,c}\|v\|_c$, where $\|\cdot\|_c$ is any norm. Let $v = \sum_{i=1}^{d+1} e_i x_i$ where $\{e_i\}$ is a basis of $\mathbb{R}^{d+1}$.

For any norm $\|\cdot\|_c$, we have:

$$\|v\|_c \leq \sum_{i=1}^{d+1} |x_i| \|e_i\|_c \leq \|x\|_2 \underbrace{\sqrt{\sum_{i=1}^{d+1} \|e_i\|_c}}_{\beta_d},$$

where the first inequality comes from the triangle inequality, and the second inequality from the Cauchy-Schwarz inequality. Thus we have:

$$\|v\|_H \leq \beta_d \|x\|_2,$$

with $\beta_d = \sqrt{d+1}$ since $\|e_i\|_H \leq 1$.

Let $\alpha_{d,c} = \min_{\{v\,:\,\|v\|_c=1\}} \|v\|_c$. Consider $u = \frac{v}{\|v\|_c}$. Then $\|u\|_c = 1$ so that $\|v\|_H \geq \alpha_{d,c}\|v\|_c$.

To find α_d, we consider the unit ℓ_2 ball in V_d, and find the smallest $\lambda > 0$ so that λB_V fully contains the Euclidean ball.

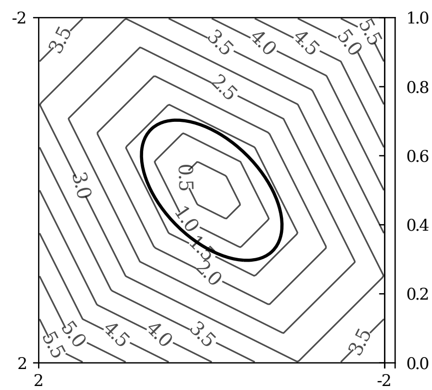


Fig. 12: Polytope balls B_V and the Euclidean unit ball B_E. From the figure smallest polytope ball has radius ≈ 1.5.

Therefore, we have overall:

$$\alpha_d \|x\|_2 \leq \|v\|_H \leq \sqrt{d+1}\|x\|_2$$

In general, note that we may consider two arbitrary norms $\|\cdot\|_l$ and $\|\cdot\|_u$ so that:

$$\alpha_{d,l}\|x\|_l \leq \|v\|_H \leq \beta_{d,u}\|x\|_u.$$

11 Funk directed metrics and Funk balls

$$F_{\mathscr{C}}(x,y) = \log\left(\frac{\|x-a\|}{\|y-a\|}\right),$$

the reverse Funk metric by:

$$F_{\mathscr{C}}(y,x) = \log\left(\frac{\|y-b\|}{\|x-b\|}\right),$$

with $b = r_{\mathscr{C}}(y,x)$, and the Hilbert metric by the arithmetic symmetrization:

$$H_{\mathscr{C}}(x,y) = \frac{F_{\mathscr{C}}(x,y) + F_{\mathscr{C}}(y,x)}{2}$$

Forward ball is Euclidean a homothet of the domain of center r and dilation factor $(1 - e^{-\rho})$ (cf. proposition 5.1, [52])