Non-linear Embeddings in Hilbert Simplex Geometry

Frank Nielsen [1] **Ke Sun** [2 3]

Abstract

A key technique of machine learning is to embed discrete weighted graphs into continuous spaces for further downstream analysis. Embedding discrete hierarchical structures in hyperbolic geometry has proven very successful since it was shown that any weighted tree can be embedded in that geometry with arbitrary low distortion. Various optimization methods for hyperbolic embeddings based on common models of hyperbolic geometry have been studied. In this paper, we consider the Hilbert geometry of the standard simplex which is isometric to a vector space equipped with a symmetric polytope norm. We study the representation power of this Hilbert simplex geometry by embedding distance matrices of graphs using a fast differentiable approximation of the Hilbert metric distance. Our findings demonstrate that Hilbert simplex geometry is competitive to alternative geometries such as the Poincaré hyperbolic ball or the Euclidean geometry for embedding tasks while being fast and numerically robust.

1. Introduction

Since Sarkar (Sarkar, 2011) proved that any weighted tree graph can be embedded as a Delaunay subgraph of points in hyperbolic geometry embedding nodes with arbitrary small distortions, hyperbolic embeddings have become widely popular in machine learning (Sala et al., 2018) and computer vision (Khrulkov et al., 2020) to represent various hierarchical structures (Surís et al., 2021). Various models of hyperbolic geometry have been considered from the viewpoint of time efficiency and numerical stability (Sonthalia & Gilbert, 2020) (e.g., Poincaré model (Nickel & Kiela, 2017), Minkowski hyperboloid model (Sun et al., 2015; Wang et al., 2020), Klein model (Feng et al., 2020), Lorentz model (Nickel & Kiela, 2018), etc.) and extensions to symmetric matrix spaces (Lopez et al., 2021) have also been considered recently.

In this work, we consider Hilbert geometry (Papadopoulos & Troyanov, 2014b) which can be seen as a generalization of Klein model of hyperbolic geometry where the unit ball domain is replaced by an arbitrary open bounded convex domain Ω. When the boundary $\partial\Omega$ is smooth, Hilbert geometry is of hyperbolic type (e.g., Cayley-Klein geometry (Richter-Gebert, 2011) when Ω is an ellipsoid). When the domain is a polytope, Hilbert geometry is bilipschitz quasi-isometric to a normed vector space (Vernicos, 2014), and isometric to a vector space with a polytope norm only when Ω is a simplex (de la Harpe, 1991). It is thus interesting to consider Hilbert simplex geometry for embeddings and compare its representation performance to hyperbolic embeddings. Hilbert simplex geometry has been used in machine learning for clustering histograms (Nielsen & Sun, 2019).

The paper is organized as follows: In §2, we present the Hilbert distance as a symmetrization of the oriented Funk weak distances, describe some properties of the Hilbert simplex distance, and illustrate qualitatively the ball shapes for the Funk and Hilbert distances (Nielsen & Shao, 2017). We first consider Hilbert simplex linear embeddings and prove that Hilbert simplex distance is a non-separable monotone distance (Theorem 1 in §2.3). Monotonicity of distances is an essential property: It states that the distance can only decrease by linear embeddings into smaller dimensional spaces. In information geometry, separable monotone divergences are exactly the class of f-divergences (Amari, 2016). Aitchison non-separable distance used in compositional data analysis was proven monotone (Erb & Ay, 2021) only recently. We explain a connection between Aitchison distance and Hilbert distance by using the variation seminorm in §2.4. Section 3 presents our experiments which demonstrates that in practice Funk and Hilbert non-linear embeddings outperforms or is competitive compared to various other distances (namely, Euclidean/Aitchison distance, ℓ_1-distance, hyperbolic Poincaré distance) while being fast and robust to compute. Section 4 concludes this work.

[1] Sony Computer Science Laboratories Inc., Japan [2] CSIRO Data61, Australia [3] The Australian National University. Correspondence to: Frank Nielsen <Frank.Nielsen@acm.org>, Ke Sun <sunk@ieee.org>.

Proceedings of the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

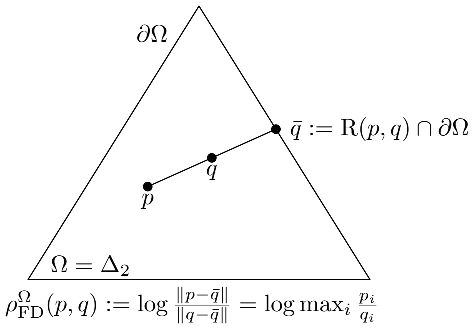

$\partial\Omega$
$\bar{q} := \mathrm{R}(p,q)\cap\partial\Omega$
q
p
$\Omega = \Delta_2$
$\rho_{\mathrm{FD}}^{\Omega}(p,q) := \log\frac{\|p-\bar{q}\|}{\|q-\bar{q}\|} = \log\max_i \frac{p_i}{q_i}$

Figure 1. Funk distance defined in the open standard simplex.

2. Hilbert simplex geometry

2.1. Definition

Let Ω be any open bounded convex set of $\mathbb{R}^d$. The Hilbert distance (Hilbert, 1895; Lemmens & Nussbaum, 2014; Papadopoulos & Troyanov, 2014b) $\rho_{\mathrm{HG}}^{\Omega}(p,q)$ between two points $p, q \in \Omega$ induced by Ω is defined as the symmetrization of the Funk distance $\rho_{\mathrm{FD}}^{\Omega}(p,q)$. The Funk distance (Papadopoulos & Troyanov, 2014a) is defined by

$$\rho_{\mathrm{FD}}^{\Omega}(p,q) := \begin{cases} \log\left(\frac{\|p-\bar{q}\|}{\|q-\bar{q}\|}\right) \geq 0, & p \neq q, \\ 0 & p = q. \end{cases}$$

where $\bar{q}$ denotes the intersection of the affine ray $R(p,q)$ emanating from p and passing through q with the domain boundary $\partial\Omega$. See Figure 1 for an illustration. The Funk distance is a weak metric distance since it satisfies the triangle inequality of metric distances but is an asymmetric dissimilarity measure: $\rho_{\mathrm{FD}}^{\Omega}(p,q) \neq \rho_{\mathrm{FD}}^{\Omega}(q,p)$.

Thus the Hilbert distance $\rho_{\mathrm{HG}}^{\Omega}(p,q)$ between any two points $p, q \in \Omega$ is:

$$\begin{aligned} \rho_{\mathrm{HG}}^{\Omega}(p,q) &:= \rho_{\mathrm{FD}}^{\Omega}(p,q) + \rho_{\mathrm{FD}}^{\Omega}(q,p), \\ &= \begin{cases} \log\frac{\|p-\bar{q}\|\,\|q-\bar{p}\|}{\|p-\bar{p}\|\,\|q-\bar{q}\|}, & p \neq q, \\ 0 & p = q. \end{cases} \end{aligned}$$

where $\bar{p}$ and $\bar{q}$ are the two intersection points of the line (pq) with $\partial\Omega$, and the four collinear points are arranged in the order $\bar{p}, p, q, \bar{q}$. The d-dimensional Hilbert distance can also be interpreted as a 1D Hilbert distance induced by the 1D interval domain $\Omega_{pq} := \Omega \cap (pq)$:

$$\rho_{\mathrm{HG}}^{\Omega}(p,q) = \rho_{\mathrm{HG}}^{\Omega_{pq}}(p,q).$$

This highlights that the quantity $\frac{\|p-\bar{q}\|\,\|q-\bar{p}\|}{\|p-\bar{p}\|\,\|q-\bar{q}\|}$ does not depend on the chosen norm $\|\cdot\|$ because we can consider the absolute value $|\cdot|$ on the domain Ω_{pq}. For any x, $y \in \mathbb{R}$, $\|x\| = c|x|$ and $\|x-y\| = c|x-y|$ where $c > 0$ is a constant. Thus we can express the Hilbert distance as the logarithm of the cross-ratio:

$$\rho_{\mathrm{HG}}^{\Omega}(p,q) = \begin{cases} \log \mathrm{CR}(\bar{p},p;q,\bar{q}), & p \neq q, \\ 0 & p = q, \end{cases}$$

where $\mathrm{CR}(\bar{p},p;q,\bar{q}) := \frac{\|p-\bar{q}\|\,\|q-\bar{p}\|}{\|p-\bar{p}\|\,\|q-\bar{q}\|}$ denotes the cross-ratio. The Hilbert distance is a metric distance, and it follows from the properties of the cross-ratio (Richter-Gebert, 2011) that straight lines are geodesics in Hilbert geometry:

$$\forall r \in [pq], \quad \rho_{\mathrm{HG}}^{\Omega}(p,q) = \rho_{\mathrm{HG}}^{\Omega}(p,r) + \rho_{\mathrm{HG}}^{\Omega}(r,q),$$

where $[pq]$ is the closed line segment connecting p and q.

Another property is that the Hilbert distance is invariant under homographies (Hartley & Zisserman, 2003) H (also called collineations):

$$\rho_{\mathrm{HG}}^{H\Omega}(Hp,Hq) = \rho_{\mathrm{HG}}^{\Omega}(p,q),$$

where $H\Omega := \{Hp \ : \ p \in \Omega\}$. The Hilbert geometry of the complex Siegel ball generalizing the Klein ball has been studied in (Nielsen, 2020).

2.2. Hilbert simplex distance

We shall consider $\Omega = \Delta_d$, the open $(d-1)$-dimensional simplex:

$$\Delta_d := \left\{ (x_1,\ldots,x_d) \in \mathbb{R}_{++}^d \ : \ \sum_{i=1}^{d} x_i = 1 \right\},$$

where $\mathbb{R}_{++} := (0,\infty)$.

In that case, we write $\rho_{\mathrm{FD}}(p,q) := \rho_{\mathrm{FD}}^{\Delta_d}(p,q)$, and we have

$$\rho_{\mathrm{FD}}(p,q) = \log\max_{i\in\{1,\ldots,d\}}\frac{p_i}{q_i}. \qquad (1)$$

Thus the Hilbert distance induced by the standard simplex is

$$\begin{aligned} \rho_{\mathrm{HG}}(p,q) &= \rho_{\mathrm{FD}}(p,q) + \rho_{\mathrm{FD}}(q,p) \\ &= \log\max_{i\in\{1,\ldots,d\}}\frac{p_i}{q_i}\max_{i\in\{1,\ldots,d\}}\frac{q_i}{p_i} \\ &= \log\frac{\max_{i\in\{1,\ldots,d\}}\frac{p_i}{q_i}}{\min_{i\in\{1,\ldots,d\}}\frac{p_i}{q_i}}. \qquad (2) \end{aligned}$$

Property 1. *We can compute efficiently the Hilbert simplex distance in Δ_d in optimal $O(d)$ time.*

As depicted in Figure 2, the geodesics are not unique in the Hilbert simplex geometry.

Figure 3 displays the shapes of balls with respect to the oriented Funk distances and the symmetrized Hilbert distance. Balls in the Hilbert simplex geometry have Euclidean polytope shapes of constant combinatorial complexity (e.g., hexagons in 2D). Since at infinitesimal scale, balls have polygonal shapes, it shows that the Hilbert simplex geometry is not Riemannian.

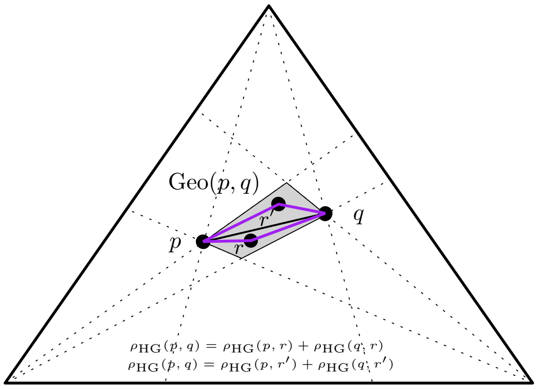

Geo(p, q)
p
q
r'
r
$\rho_{\mathrm{HG}}(p, q) = \rho_{\mathrm{HG}}(p, r) + \rho_{\mathrm{HG}}(q, r)$
$\rho_{\mathrm{HG}}(p, q) = \rho_{\mathrm{HG}}(p, r') + \rho_{\mathrm{HG}}(q, r')$

Figure 2. Non-uniqueness of geodesics in the Hilbert simplex geometry: The quadrilateral region Geo(p, q) denotes the set of points r satisfying the triangle equality with respect to p and q: $\rho_{\mathrm{HG}}(p, q) = \rho_{\mathrm{HG}}(p, r) + \rho_{\mathrm{HG}}(q, r)$. The purple paths connecting p and q are examples of geodesics.

2.3. Monotone distance

Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ be a partition of $\{1, \ldots, d\}$ into $m \leq d$ pairwise disjoint subsets X_i's. For $p \in \Delta_d$, let $p_{|\mathcal{X}} \in \Delta_m$ denote the reduced dimension point with $p_{|\mathcal{X}}[i] = \sum_{j \in X_i} p[i]$. A distance $D(p, q)$ is said *monotone* (Amari, 2016) iff

$$D\left(p_{|\mathcal{X}}, q_{|\mathcal{X}}\right) \leq D(p, q), \quad \forall \mathcal{X}, \forall p, q \in \Delta_d.$$

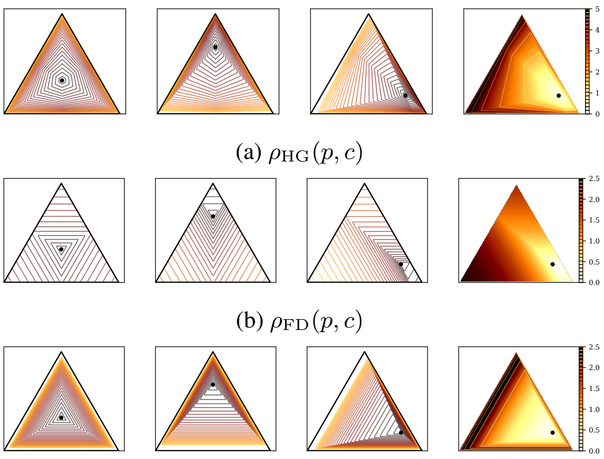
A distance is said *separable* iff it can be expressed as a sum of scalar distances. For example, the Euclidean distance is not separable but the squared Euclidean distance is separable. The only separable monotone distances are f-divergences (Amari, 2016) when $d > 2$. The special curious case $d = 2$ is dealt in (Jiao et al., 2014). We can interpret points in the simplex Δ_d as categorical distributions on a sample space of d outcomes. Hence, the Hilbert statistical distance can also be said information monotone (Amari, 2016).

We shall prove that the Funk oriented distance and the Hilbert distance are non-separable monotone distances.

Lemma 1. *Let $p, q \in \Delta_d$. Let $\tilde{p} = (p_1 + p_2, p_3, \cdots, p_d)$ and $\tilde{q} = (q_1 + q_2, q_3, \cdots, q_d)$ denote their coarse-grained points on Δ_{d-1}. We have $0 \leq \rho_{\mathrm{FD}}(\tilde{p}, \tilde{q}) \leq \rho_{\mathrm{FD}}(p, q)$.*

Proof. Denote $\iota = \max\{p_1/q_1, p_2/q_2\}$. As $q_1, q_2 > 0$, we have $p_1 \leq \iota q_1$ and $p_2 \leq \iota q_2$. Therefore

$$\frac{p_1 + p_2}{q_1 + q_2} \leq \frac{\iota q_1 + \iota q_2}{q_1 + q_2} = \iota.$$



(a) $\rho_{\mathrm{HG}}(p, c)$

(b) $\rho_{\mathrm{FD}}(p, c)$

(c) $\rho_{\mathrm{FD}}(c, p)$

Figure 3. Balls centered at $c \in \Delta_2$ with constant radius increment step. The last column also shows the distance color maps (dark color means long distance).

It follows that

$$\max\left\{\frac{p_1 + p_2}{q_1 + q_2}, \frac{p_3}{q_3}, \cdots, \frac{p_d}{q_d}\right\} \leq \max\left\{\frac{p_1}{q_1}, \frac{p_2}{q_2}, \frac{p_3}{q_3}, \cdots, \frac{p_d}{q_d}\right\}.$$

Hence

$$\log \max\left\{\frac{p_1 + p_2}{q_1 + q_2}, \frac{p_3}{q_3}, \cdots, \frac{p_d}{q_d}\right\} \leq \log \max_i \frac{p_i}{q_i}.$$

By the definition of the Funk distance, we get $\rho_{\mathrm{FD}}(\tilde{p}, \tilde{q}) \leq \rho_{\mathrm{FD}}(p, q)$. □

Theorem 1. *The Funk distance ρ_{FD} and the Hilbert distance ρ_{HG} in Δ_d satisfy the information monotonicity.*

The proof is straightforward from Lemma 1 by noting that any coarse-grained point can be recursively defined by merging two bins. Since the sum of two information monotone distances is monotone, we get the proof that Hilbert distance as the sum of the forward and reverse Funk (weak) metric is monotone.

In fact, we can also prove this result by using Birkhoff's contraction mapping theorem (Birkhoff, 1957). We can represent the coarse-graining mapping $p \mapsto p_{|\mathcal{X}}$ by a linear application with a $m \times d$ matrix A with columns summing up to one (i.e., positive column-stochastic matrix):

$$p_{|\mathcal{X}} = Ap.$$

Then we have (Birkhoff, 1957):

$$\rho_{\mathrm{HG}}(Ap, Aq) \leq \tanh\left(\frac{1}{4}\Delta(A)\right) \rho_{\mathrm{HG}}(p, q),$$

where $\Delta(A)$ is called the projective diameter of the positive mapping A: $\Delta(A) := \sup\{\rho_{\mathrm{HG}}(Ap, Aq) \ : \ p, q \in \mathbb{R}^d_{++}\}$.

Since $0 \leq \tanh(x) \leq 1$ for $x \geq 0$, we get the property that Hilbert distance on the probability simplex is a monotone non-separable distance: $\rho_{\mathrm{HG}}(p_{|\mathcal{X}}, q_{|\mathcal{X}}) \leq \rho_{\mathrm{HG}}(p, q)$. Note that Birkhoff's contraction theorem is also used to prove the convergence of Sinkhorn's algorithm (Peyré & Cuturi, 2019).

Hilbert distance of Eq. 2 can be extended to the positive orthant cone $\mathbb{R}^d_{++}$ which can be foliated by homothetic simplices $\lambda\Delta_d$: $\mathbb{R}^d_{++} = \cup_{\lambda>0}\lambda\Delta_d$:

$$\rho_{\mathrm{HG}}(\tilde{p}, \tilde{q}) = \log \frac{\max_{i\in\{1,\ldots,d\}} \frac{\tilde{p}_i}{\tilde{q}_i}}{\min_{i\in\{1,\ldots,d\}} \frac{\tilde{p}_i}{\tilde{q}_i}}, \quad \tilde{p}, \tilde{q} \in \mathbb{R}^d_{++}. \quad (3)$$

This extended Hilbert distance is projective because $\rho_{\mathrm{HG}}(\alpha\tilde{p}, \beta\tilde{q}) = \rho_{\mathrm{HG}}(\tilde{p}, \tilde{q})$. Thus the Hilbert distance is a projective distance between rays $\tilde{p}$ and $\tilde{q}$ and a metric distance on $\lambda\Delta_d$ for any prescribed value of $\lambda > 0$.

The Aitchison distance (Pawlowsky-Glahn & Buccianti, 2011) is also a non-separable distance in the standard simplex defined as follows:

$$\rho_{\mathrm{Aitchison}}(p, q) := \sqrt{\sum_{i=1}^{d} \left(\log \frac{p_i}{G(p)} - \log \frac{q_i}{G(q)} \right)^2}, \quad (4)$$

where $G(p)$ denotes the geometric mean of the coordinates of $p \in \Delta_d$:

$$G(p) = \left(\prod_{i=1}^{d} p_i \right)^{\frac{1}{d}} = \exp\left(\frac{1}{d} \sum_{i=1}^{d} \log p_i \right).$$

The Aitchison distance satisfies the monotonicity property (Erb & Ay, 2021).

2.4. Isometry to a normed vector space

Hilbert geometry is never a Hilbert space (i.e., complete metric space induced by the inner product of a vector space) because the convex domain Ω is bounded. It can be shown that the only domains Ω yielding an isometry of the Hilbert geometry to a normed vector space are simplices (Colbois & Verovic, 2008).

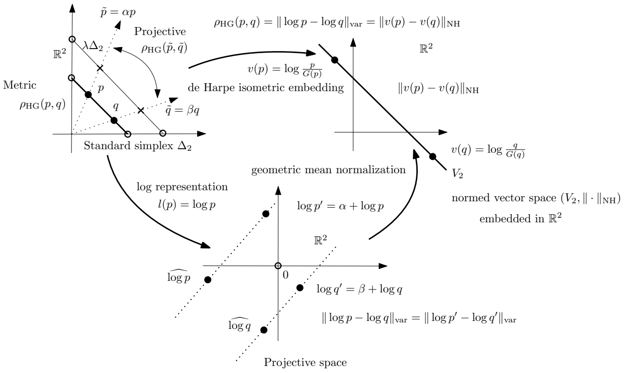

$\tilde{p} = \alpha p$
Projective
$\rho_{\mathrm{HG}}(\tilde{p}, \tilde{q})$
$\rho_{\mathrm{HG}}(p, q) = \| \log p - \log q\|_{\mathrm{var}} = \|v(p) - v(q)\|_{\mathrm{NH}}$
$\mathbb{R}^2$
$\lambda\Delta_2$
$\mathbb{R}^2$
$v(p) = \log \frac{p}{G(p)}$
Metric
p
de Harpe isometric embedding
$\|v(p) - v(q)\|_{\mathrm{NH}}$
$\rho_{\mathrm{HG}}(p, q)$
q
$\tilde{q} = \beta q$
Standard simplex Δ_2
$v(q) = \log \frac{q}{G(q)}$
geometric mean normalization
V_2
log representation
$l(p) = \log p$
$\log p' = \alpha + \log p$
normed vector space $(V_2, \| \cdot \|_{\mathrm{NH}})$
embedded in $\mathbb{R}^2$
$\mathbb{R}^2$
$\widehat{\log p}$
0
$\log q' = \beta + \log q$
$\| \log p - \log q\|_{\mathrm{var}} = \| \log p' - \log q'\|_{\mathrm{var}}$
$\widehat{\log q}$
Projective space

Figure 4. Different representations of the simplex and positive orthant cone.

We recall the isometry (de la Harpe, 1991) of the standard simplex to a normed vector space $(V_d, \| \cdot \|_{\mathrm{NH}})$. Let $V_d = \{v \in \mathbb{R}^d \ : \ \sum_{i=1}^d v_i = 1\}$ denote the $(d-1)$-dimensional vector space sitting in $\mathbb{R}^d$. Map a point $p = (p_1, \ldots, p_d) \in \Delta_d$ to a point $v(p) = (v_1, \ldots, v_d) \in V_d$ as follows:

$$\begin{aligned} v_i &= \frac{1}{d}\left((d-1)\log p_i - \sum_{j\neq i} \log p_j \right), \\ &= \log p_i - \frac{1}{d}\sum_{j=1}^{d} \log \ p_j. \end{aligned}$$

We define the corresponding norm $\| \cdot \|_{\mathrm{NH}}$ 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* $\| \cdot \|_{\mathrm{NH}}$ (Hilbert norm) with $2\binom{d}{2} = d(d-1)$ facets. Norms ℓ_1 and ℓ_∞ yield hypercube balls with $2d$ facets and 2^d vertices. 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:

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

where $A \oplus B = \{a + b \ : \ a \in A, b \in B\}$ is the Minkowski sum. Figure 5 illustrates the balls centered at the origin with respect to the polytope norm $\| \cdot \|_{\mathrm{NH}}$.

Let $l(p) = (\log p_1, \ldots, \log p_d) \in \mathbb{R}^d$ be the logarithmic mapping and $\mathbb{L}_d = \{l(p) \ : \ p \in \Delta_d\}$. We have

$$\rho_{\mathrm{HG}}(p, q) = \|l(p) - l(q)\|_{\mathrm{var}} = \|l(\tilde{p}) - l(\tilde{q})\|_{\mathrm{var}},$$

for any $\alpha, \beta \in \mathbb{R}$ with $\tilde{p} = \alpha p$, $\tilde{q} = \beta q$, and where $\|x\|_{\mathrm{var}} := \max_i x_i - \min_i x_i = \|x\|_{+\infty} - \|x\|_{-\infty}$ is the variation semi-norm. $\| \cdot \|_{\mathrm{var}}$ is only a *semi-norm* because $\|(\lambda, ..., \lambda)\|_{\mathrm{var}} = 0$ for any $\lambda \in \mathbb{R}$.

Thus to convert from $\mathbb{L}$ to Δ_d, we need to find the representative element of the equivalence class $\hat{l}$ of l: normalize $l \in \mathbb{L}$ by $p(l) = \hat{l} = \exp(l)/\sum_{i=1}^d e^{l_i}$. Then we convert $\hat{l}$ to $v(\hat{l})$ by choosing translation $-\log G(\hat{l})$, where G is the

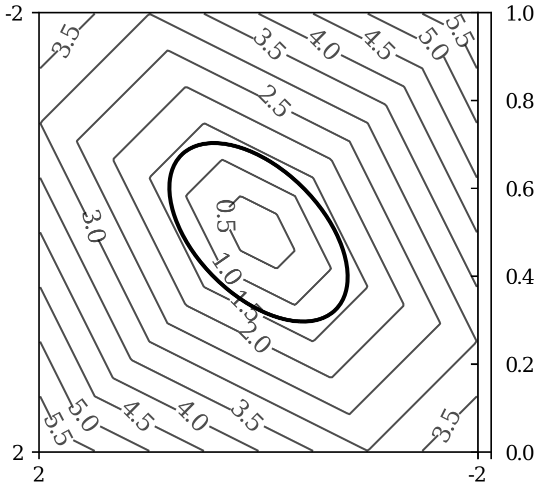


Figure 5. Polytope balls B_V and the Euclidean unit ball B_E shown on the 2D slanted plane $V^3 = \{v \in \mathbb{R}^3 \ : \ \sum_{i=1}^{3} v^i = 0\}$ of $\mathbb{R}^3$.

geometric mean:

$$v(l) = \\ \left(l_1 - \log G\left(\frac{e^l}{\sum_{i=1}^d e^{l_i}}\right), \ldots, l_d - \log G\left(\frac{e^l}{\sum_{i=1}^d e^{l_i}}\right)\right).$$

Thus we have

$$\rho_{\mathrm{HG}}(p,q) = \rho_{\mathrm{HG}}(\tilde{p},\tilde{q}) = \|\log \tilde{p} - \log \tilde{q}\|_{\mathrm{var}} \\ = \|v(p) - v(q)\|_{\mathrm{NH}} = \|l(\hat{p}) - l(\hat{q})\|_{\mathrm{NH}}.$$

Therefore

$$\|l - l'\|_{\mathrm{var}} = \|v(\hat{l}) - v(\hat{l'})\|_{\mathrm{NH}}.$$

Figure 4 illustrates different transformations of the simplex space included in the positive orthant cone.

The reverse map from the normed space V_d to the standard simplex Δ_d is given by the softmax function:

$$p_i = \frac{\exp(v_i)}{\sum_j \exp(v_j)}.$$

Thus we have $(\Delta_d, \rho_{\mathrm{HG}}) \cong (V_d, \|\cdot\|_{\mathrm{NH}})$. In 1D, $(V^2, \|\cdot\|_{\mathrm{NH}})$ is isometric to the Euclidean line.

Now, let us notice that coordinate v_i can be rewritten as

$$v_i = \log \frac{p_i}{G(p)},$$

where $G(p)$ is the coordinate geometric means. Recalling that the Hilbert simplex distance is a projective distance on the positive orthant cone domain:

$$\begin{aligned} \rho_{\mathrm{HG}}(p,q) &= \log \frac{\max_{i\in\{1,\ldots,d\}} \frac{p_i}{q_i}}{\min_{j\in\{1,\ldots,d\}} \frac{p_j}{q_j}}, \\ &= \rho_{\mathrm{HG}}(\lambda p, \lambda' q), \quad \forall \lambda > 0, \lambda' > 0. \end{aligned}$$

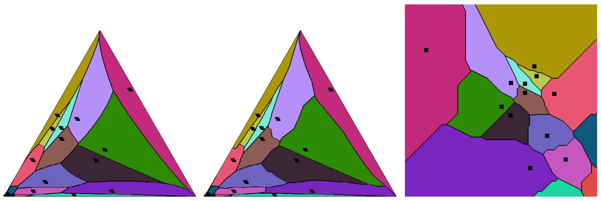


Figure 6. Voronoi diagram in the probability simplex with respect to the Aitchison distance (left), Hilbert simplex distance (middle) and equivalent variation-norm induced distance on normalized logarithmic representations.

Thus we have:

$$\begin{aligned} \rho_{\mathrm{HG}}(p,q) &= \|\log p - \log q\|_{\mathrm{var}}, \\ &= \|\log(\lambda p) - \log(\lambda' q)\|_{\mathrm{var}}, \quad \forall \lambda > 0, \lambda' > 0. \end{aligned}$$

Choose $\lambda = \frac{1}{G(p)}$ and $\lambda' = \frac{1}{G(q)}$ to get

$$\rho_{\mathrm{HG}}(p,q) = \left\|\log \frac{p}{G(p)} - \log \frac{q}{G(q)}\right\|_{\mathrm{var}}.$$

This highlights a nice connection with the Aitchison distance of Eq. 4:

$$\rho_{\mathrm{HG}}(p,q) = \left\|\log \frac{p}{G(p)} - \log \frac{q}{G(q)}\right\|_{\mathrm{var}}, \quad (5)$$

$$\rho_{\mathrm{Aitchison}}(p,q) = \left\|\log \frac{p}{G(p)} - \log \frac{q}{G(q)}\right\|_{2}. \quad (6)$$

Thus both the Aitchison distance and the Hilbert simplex distance are normed distances on the representation $p \mapsto \log \frac{p}{G(p)} = \left(\log \frac{p_1}{G(p)}, \ldots, \frac{p_d}{G(p)}\right)$.

Figure 6 displays the Voronoi diagram of $n = 16$ points in the probability simplex with respect to the Aitchison distance (Figure 6, left), and the Hilbert simplex distance (Figure 6, middle) and its equivalent variation norm space by logarithmic embedding (Figure 6, right). See also (Gezalyan & Mount, 2021).

The Hilbert simplex bisectors are piecewise linear in the normalized logarithmic representation since they are induced by the polyhedral semi-norm $\|.\|_{\mathrm{var}}$, and thus are more complex than the Aitchison bisectors which are linear/affine in the $\log x/G(x)$ representation: Aitchinson Voronoi diagram can be derived from an ordinary Euclidean Voronoi diagram (Boissonnat et al., 1998).

2.5. Differentiable approximation

The Hilbert simplex distance is not differentiable because of the max operations. However, since the logarithm function

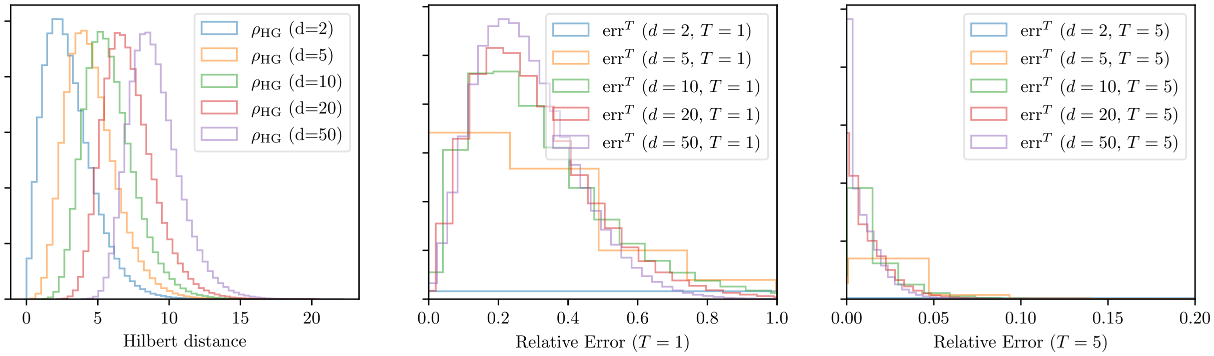


Figure 7. Histograms of Hilbert distances and the relative errors on 10^6 pairs of uniform random points in Δ_d.

is strictly increasing, we can rewrite the Funk distance as

$$\rho_{\mathrm{FD}}(p,q) = \log \max_i \frac{p_i}{q_i} = \max_i \log \frac{p_i}{q_i}.$$

In machine learning, the log-sum-exp function

$$\mathrm{LSE}(x_1,\ldots,x_d) := \log\left(\sum_{i=1}^d \exp(x_i)\right)$$

is commonly used to differentiably approximate the maximum operator. In fact, one can use the general approximation formula

$$\begin{aligned}\mathrm{LSE}^T(x_1,\ldots,x_d) &:= \frac{1}{T}\mathrm{LSE}(Tx_1,\ldots,Tx_d)\\ &= \frac{1}{T}\log\left(\sum_{i=1}^d \exp(Tx_i)\right),\end{aligned}$$

where $T>0$. For any $x\in\mathbb{R}^d$, we have the approximation bounds

$$\begin{aligned}\max_i x_i + \varepsilon_1(x,T) &\leq \mathrm{LSE}^T(x_1,\ldots,x_d)\\ &\leq \max_i x_i + \varepsilon_2(x,T), \qquad (7)\end{aligned}$$

where

$$\begin{aligned}\varepsilon_1(x,T) &:= \frac{1}{T}\log\left[1+(d-1)\exp(-T\|x\|_{\mathrm{var}})\right],\\ \varepsilon_2(x,T) &:= \frac{1}{T}\log\left[d-1+\exp(-T\|x\|_{\mathrm{var}})\right].\end{aligned}$$

Obviously, $0<\varepsilon_1(x,T)\leq\varepsilon_2(x,T)\leq\frac{1}{T}\log d$ and both $\varepsilon_1(x,T)$ and $\varepsilon_2(x,T)$ tend to 0 as T increases, making the approximation $\mathrm{LSE}^T(x_1,\ldots,x_d)$ accurate.

Because $\forall i$, $x_i \geq \max_i x_i - \|x\|_{\mathrm{var}}$, we have

$$\begin{aligned}\left(\sum_{i=1}^d \exp(Tx_i)\right) &\geq (d-1)\exp(T\max_i x_i - T\|x\|_{\mathrm{var}})\\ &\quad + \exp(T\max_i x_i)\\ &= ((d-1)\exp(-T\|x\|_{\mathrm{var}})+1)\exp(T\max_i x_i).\end{aligned}$$

Taking the logarithm on both sides gives the first "$\leq$" in Eq. (7). The proof of the second "$\leq$" is similar.

Thus we have

$$\begin{aligned}\rho_{\mathrm{FD}}(p,q)+\varepsilon_1(r,T) &\leq \frac{1}{T}\log\left(\sum_i\left(\frac{p_i}{q_i}\right)^T\right)\\ &\leq \rho_{\mathrm{FD}}(p,q)+\varepsilon_2(r,T), \qquad (8)\end{aligned}$$

where $r_i = \log p_i - \log q_i$.

Hence, we may define a differentiable pseudo-distance by symmetrizing the LSE^T function:

$$\begin{aligned}&\tilde{\rho}_{\mathrm{LSE}^T}(p,q) =\\ &\quad \tfrac{1}{T}\log\left(\sum_i\left(\tfrac{p_i}{q_i}\right)^T\right)\left(\sum_i\left(\tfrac{q_i}{p_i}\right)^T\right). \qquad (9)\end{aligned}$$

We have $\tilde{\rho}_{\mathrm{LSE}^T}(p,q)\geq 0$ and $\tilde{\rho}_{\mathrm{LSE}}(p,p)=\frac{2}{T}\log d$. Similar to Eq. (8), we have

$$\begin{aligned}\rho_{\mathrm{HG}}(p,q)+2\epsilon_1(r,T) &\leq \tilde{\rho}_{\mathrm{LSE}^T}(p,q)\\ &\leq \rho_{\mathrm{HG}}(p,q)+2\epsilon_2(r,T).\end{aligned}$$

The maximum deviation from the approximation $\tilde{\rho}_{\mathrm{LSE}^T}(p,q)$ to the true Hilbert distance $\rho_{\mathrm{HG}}(p,q)$ is bounded by $\frac{2}{T}\log d$.

Figure 7 shows the histogram of Hilbert distances on uniform random points drawn from Δ_d, and the relative error

$$\mathrm{err}^T(p,q) := \frac{\tilde{\rho}_{\mathrm{LSE}^T}(p,q)-\rho_{\mathrm{HG}}(p,q)}{\rho_{\mathrm{HG}}(p,q)}$$

from the differentiable approximation $\tilde{\rho}_{\mathrm{LSE}^T}(p,q)$ to the true Hilbert distance $\rho_{\mathrm{HG}}(p,q)$. We can verify empirically that $\tilde{\rho}_{\mathrm{LSE}^T}(p,q)$ is always larger than $\rho_{\mathrm{HG}}(p,q)$. We observe the approximation becomes more accurate when T increases from 1 to 5. The approximation error tends to be large at a small dimensionality d.

3. Comparing different geometries

We empirically compare the representation power of different geometries for embedding the input data as a set of points on the manifold. Our objective is not to build a full-fledged embedding method, but to have simple well-defined measurements to *compare different geometries*. Notice that if (M_1, ρ_1) is isometric to (M_2, ρ_2) then these geometries have the same representation power. We prefer to choose model geometries with unconstrained domains for optimization. Thus the Poincaré hyperbolic ball and the Aitchison simplex embeddings are considered via the equivalent Minkowski hyperboloid and Euclidean models, respectively.

We consider embedding two different types of data onto a manifold $\mathcal{M}^d$ of dimensionality d, which we also denote as $\mathcal{M}$. The first is given by a distance matrix $\mathcal{D}_{n\times n}$. The representation loss associated with $\mathcal{M}^d$ is

$$\ell(\mathcal{D}, \mathcal{M}^d) := \inf_{\boldsymbol{Y}\in(\mathcal{M}^d)^n} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathcal{D}_{ij} - \rho_{\mathcal{M}}(\boldsymbol{y}_i, \boldsymbol{y}_j)\right)^2,$$

where $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$ is a set of n free points on $\mathcal{M}^d$, and $\rho_{\mathcal{M}}$ is the distance on $\mathcal{M}$. The infimum means the error associated with the best representation of the given distance matrix. A smaller value of $\ell(\mathcal{D}, \mathcal{M}^d)$ means $\mathcal{M}^d$ can better represent the distance matrix $\mathcal{D}$. We set $\mathcal{D}$ to be ① the distance matrix between n random points in $\mathbb{R}^n$ ($n = 100$); or ② the pairwise shortest path between any two nodes on an Erdős–Rényi graph $G(n, p)$ ($n = 200$, $p = 0.2$); or ③ the node distance on a Barabási–Albert graph $G(n, m)$ ($n = 200$, $m = 2$).

On the other hand, we can evaluate the geometries based on a given probability matrix $\mathcal{P}_{n\times n}$, meaning some non-negative pair-wise similarities. $\mathcal{P}$ is row-normalized so that each row sums to 1. We consider the loss

$$\ell(\mathcal{P}, \mathcal{M}^d) := \inf_{\boldsymbol{Y}\in(\mathcal{M}^d)^n} \frac{1}{n}\sum_{i=1}^{n}\sum_{j:j\neq i}\mathcal{P}_{ij}\log\frac{\mathcal{P}_{ij}}{q_{ij}(\boldsymbol{Y})},$$

$$q_{ij}(\boldsymbol{Y}) := \frac{\exp(-\rho^2_{\mathcal{M}}(\boldsymbol{y}_i, \boldsymbol{y}_j))}{\sum_{j:j\neq i}\exp(-\rho^2_{\mathcal{M}}(\boldsymbol{y}_i, \boldsymbol{y}_j))},$$

where $\ell(\mathcal{P}, \mathcal{M}^d)$ is the empirical average of the KL divergence between the probability distributions $\mathcal{P}_{i\cdot}$ and $q_{i\cdot}$. Notice that ℓ is abused to denote both the loss associated with a distance matrix $\mathcal{D}$ and a probability matrix $\mathcal{P}$. Using the same datasets as in embedding $\mathcal{D}$, we set $\mathcal{P}$ to be ① pairwise similarities of n random points in $\mathbb{R}^n$ measured by the heat kernel after normalization; ② the random walk similarity starting from node i and ending at any other node j after 5 steps on an Erdős–Rényi graph, or ③ a Barabási–Albert graph.

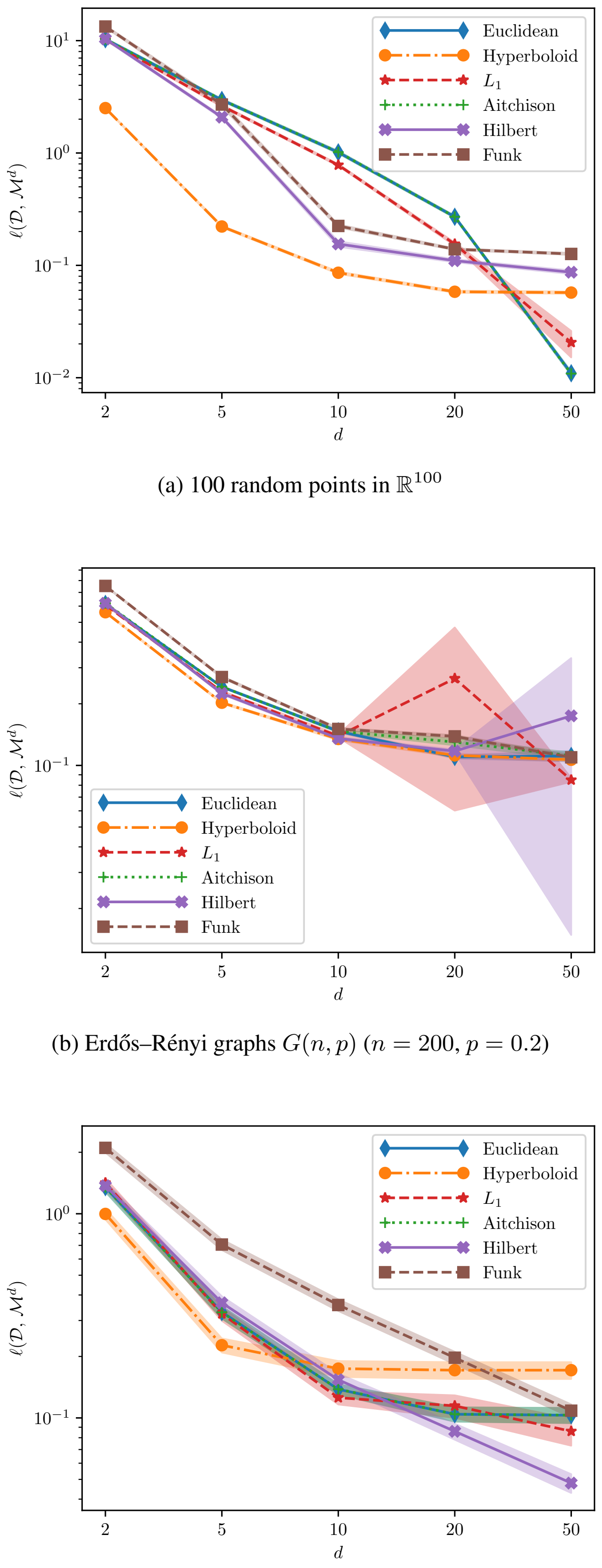
The embedding losses $\ell(\mathcal{D}, \mathcal{M}^d)$ and $\ell(\mathcal{P}, \mathcal{M}^d)$ are approximated based on the Adam optimizer (Kingma & Ba, 2015). We minimize the function whose infimum is to be taken with respect to $\boldsymbol{Y}$, starting from some randomly initialized points $\boldsymbol{Y}_0$, until converging to a local optimum. The loss $\ell(\mathcal{D}, \mathcal{M}^d)$ is similar to the stress function in multi-dimensional scaling (Borg & Groenen, 2005), while $\ell(\mathcal{P}, \mathcal{M}^d)$ is similar to the losses in manifold learning (Hinton & Roweis, 2003) or graph embedding (Perozzi et al., 2014). Our losses do not depend on many practical techniques such as negative sampling, and are helpful to measure the fitness of the manifold $\mathcal{M}^d$ to the input $\mathcal{D}$ or $\mathcal{P}$ regardless of these practical aspects. The detailed experimental protocols and more extensive results are in (Nielsen & Sun, 2022).

Figure 8 shows $\ell(\mathcal{D}, \mathcal{M}^d)$ (in log scale) against d for six different choices of $\mathcal{M}$: ① $\mathbb{R}^d$ with Euclidean norm; ② $\mathbb{R}^d$ with L_1 norm; ③ Poincaré/Minkowski hyperboloid; ④ Δ_d with Aitchison distance; ⑤ Δ_d with Hilbert distance; ⑥ Δ_d with Funk distance. For each configuration, we generate 10 different instances of the random points/graphs, and the standard deviation is shown as color bands. We observe that in general, as d grows, all manifolds have decreasing $\ell(\mathcal{D}, \mathcal{M}^d)$. The jitters and large deviations are due to that the optimizer stopped at a bad local optimum in some of the experiments. The Hilbert simplex and the Poincaré hyperboloid are observed as the best geometries which can preserve the input distance matrix. The Funk distance is asymmetric and is not as good as the other baselines.

Figure 9 shows $\ell(\mathcal{P}, \mathcal{M}^d)$ (in log scale) against d for the investigated geometries. On the random points dataset, the L_1 distance presents an increasing loss with d. This could be due to its mismatch with the geometry of the dataset and that the optimizer stopped at a local optimum. Overall, the proposed Hilbert simplex geometry can better represent pairwise similarities in $\mathbb{R}^d$ and graph random walk similarity matrix, as compared with the baselines. Funk geometry also achieves good score in representing the Erdős–Rényi graphs.

4. Conclusion

We presented the Hilbert simplex geometry with its closed form distance (Eq. 2) and its differentiable approximation (Eq. 9). We provided a simple proof that the Funk and Hilbert distances both satisfy the information monotonicity. We made use of an isometry between the Hilbert simplex and a normed vector space well-suited to carry optimization. We highlighted a connection between the Aitchinson distance and the Hilbert projective distance. By comparing with commonly-used geometries in machine learning, we showed experimentally that the Hilbert simplex geometry can better embed a given distance matrix or graph random walk similarities.

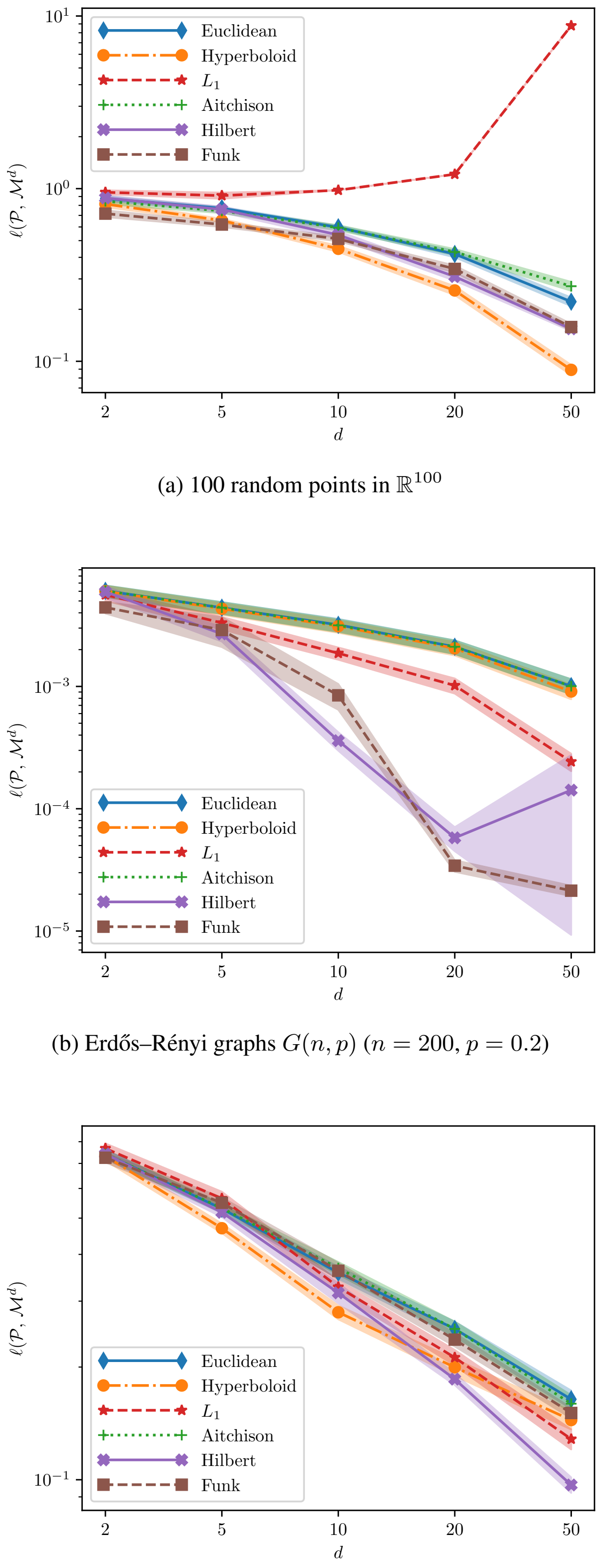


(a) 100 random points in $\mathbb{R}^{100}$

(b) Erdős–Rényi graphs $G(n, p)$ ($n = 200$, $p = 0.2$)

(c) Barabási–Albert graphs $G(n, m)$ ($n = 200$, $m = 2$)

Figure 8. Embedding loss $\ell(\mathcal{D}, \mathcal{M}^d)$ against the number of dimensions d on three random datasets.



(a) 100 random points in $\mathbb{R}^{100}$

(b) Erdős–Rényi graphs $G(n, p)$ ($n = 200$, $p = 0.2$)

(c) Barabási–Albert graphs $G(n, m)$ ($n = 200$, $m = 2$)

Figure 9. Embedding loss $\ell(\mathcal{P}, \mathcal{M}^d)$ against the number of dimensions d.

References

Amari, S.-i. *Information geometry and its applications*, volume 194. Springer, 2016.

Birkhoff, G. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

Boissonnat, J.-D., Sharir, M., Tagansky, B., and Yvinec, M. Voronoi diagrams in higher dimensions under certain polyhedral distance functions. *Discrete & Computational Geometry*, 19(4):485–519, 1998.

Borg, I. and Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer, 2nd edition, 2005.

Colbois, B. and Verovic, P. Hilbert domains quasi-isometric to normed vector spaces. *arXiv preprint arXiv:0804.1619*, 2008.

de la Harpe, P. On Hilbert's metric for simplices. In *Geometric Group Theory*, volume 1, pp. 97–118. Cambridge Univ. Press, 1991.

Erb, I. and Ay, N. The information-geometric perspective of compositional data analysis. In *Advances in Compositional Data Analysis*, pp. 21–43. Springer, 2021.

Feng, S., Tran, L. V., Cong, G., Chen, L., Li, J., and Li, F. HME: A hyperbolic metric embedding approach for next-POI recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1429–1438, 2020.

Gezalyan, A. H. and Mount, D. M. Voronoi Diagrams in the Hilbert Metric. *arXiv preprint arXiv:2112.03056*, 2021.

Hartley, R. and Zisserman, A. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

Hilbert, D. Über die gerade linie als kürzeste verbindung zweier punkte. *Mathematische Annalen*, 46(1):91–96, 1895.

Hinton, G. E. and Roweis, S. T. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, pp. 833–840. MIT Press, 2003.

Jiao, J., Courtade, T. A., No, A., Venkat, K., and Weissman, T. Information measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60 (12):7616–7626, 2014.

Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., and Lempitsky, V. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6418–6428, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Lemmens, B. and Nussbaum, R. Birkhoff's version of Hilbert's metric and its applications in analysis. *Handbook of Hilbert Geometry*, pp. 275–303, 2014.

Lopez, F., Pozzetti, B., Trettel, S., Strube, M., and Wienhard, A. Symmetric spaces for graph embeddings: A finsler-riemannian approach. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7090–7101. PMLR, 18–24 Jul 2021.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30:6338–6347, 2017.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788. PMLR, 2018.

Nielsen, F. The Siegel–Klein Disk: Hilbert Geometry of the Siegel Disk Domain. *Entropy*, 22(9):1019, 2020.

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

Nielsen, F. and Sun, K. Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices. In *Geometric Structures of Information*, pp. 297–331. Springer, 2019.

Nielsen, F. and Sun, K. Non-linear Embeddings in Hilbert Simplex Geometry. *arXiv preprint arXiv:2203.11434*, 2022.

Papadopoulos, A. and Troyanov, M. From Funk to Hilbert geometry, 2014a. arXiv:1406.6983 [math.MG].

Papadopoulos, A. and Troyanov, M. *Handbook of Hilbert Geometry*. IRMA lectures in mathematics and theoretical physics. European Mathematical Society, 2014b. ISBN 9783037191477.

Pawlowsky-Glahn, V. and Buccianti, A. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 701–710, 2014.

Peyré, G. and Cuturi, M. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Richter-Gebert, J. *Perspectives on projective geometry: A guided tour through real and complex geometry*. Springer-Verlag Berlin Heidelberg, 2011.

Sala, F., De Sa, C., Gu, A., and Ré, C. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pp. 4460–4469. PMLR, 2018.

Sarkar, R. Low distortion Delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

Sonthalia, R. and Gilbert, A. Tree! i am no tree! i am a low dimensional hyperbolic embedding. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 845–856. Curran Associates, Inc., 2020.

Sun, K., Wang, J., Kalousis, A., and Marchand-Maillet, S. Space-time local embeddings. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 100–108. Curran Associates, Inc., 2015.

Surís, D., Liu, R., and Vondrick, C. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12607–12617, 2021.

Vernicos, C. On the Hilbert geometry of convex polytopes. *Handbook of Hilbert Geometry*, pp. 111–125, 2014.

Wang, L., Lu, Y., Huang, C., and Vosoughi, S. Embedding node structural role identity into hyperbolic space. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2253–2256, 2020.

A. Experimental Configurations

We rewrite the loss as

$$\ell(\mathcal{D},\mathcal{M}^d) := \inf_{\boldsymbol{Y}\in(\mathcal{M}^d)^n} L(\mathcal{D},\mathcal{M}^d,\boldsymbol{Y})$$

$$L(\mathcal{D},\mathcal{M}^d,\boldsymbol{Y}) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left(\mathcal{D}_{ij} - \rho_{\mathcal{M}}(\boldsymbol{y}_i,\boldsymbol{y}_j)\right)^2.$$

$$\ell(\mathcal{P},\mathcal{M}^d) := \inf_{\boldsymbol{Y}\in(\mathcal{M}^d)^n} L(\mathcal{P},\mathcal{M}^d,\boldsymbol{Y})$$

$$L(\mathcal{P},\mathcal{M}^d,\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \mathcal{P}_{ij}\log\frac{\mathcal{P}_{ij}}{q_{ij}(\boldsymbol{Y})}.$$

The functions to be minimized, $L(\mathcal{D},\mathcal{M}^d,\boldsymbol{Y})$ and $L(\mathcal{P},\mathcal{M}^d,\boldsymbol{Y})$, are both expressed in the form of a sample average. Therefore they can be optimized based on stochastic gradient descent (SGD).

In the experiments, we use PyTorch to minimize $L(\mathcal{D},\mathcal{M}^d,\boldsymbol{Y})$ and $L(\mathcal{P},\mathcal{M}^d,\boldsymbol{Y})$ with respect to the coordinate matrix $\boldsymbol{Y}$. The initial $\boldsymbol{Y}_0$ is based on a multivariate Gaussian distribution so that the trace of the covariance matrix equals 1.

The optimizer is Adam in its default settings except the learning rate. The learning rate is based on a log-uniform distribution (logarithm of the learning rate is uniform) in the range $[10^{-3}, 1]$. The mini-batch size is simply set to 16. We observed that reducing the mini-batch size can generally achieve a smaller loss for all the methods. For each configuration of (dataset, manifold $\mathcal{M}^d$, dimensionality d), the optimal learning rate is selected based on a Tree Parzen estimator with 20 trials. The maximum number of epochs is 3000. We use early stopping to terminate the optimization process if convergence is detected.

Each dataset is generated independently for 10 times, based on different random seeds. The loss for each of these generated dataset is computed independently. The average and standard deviation are reported.

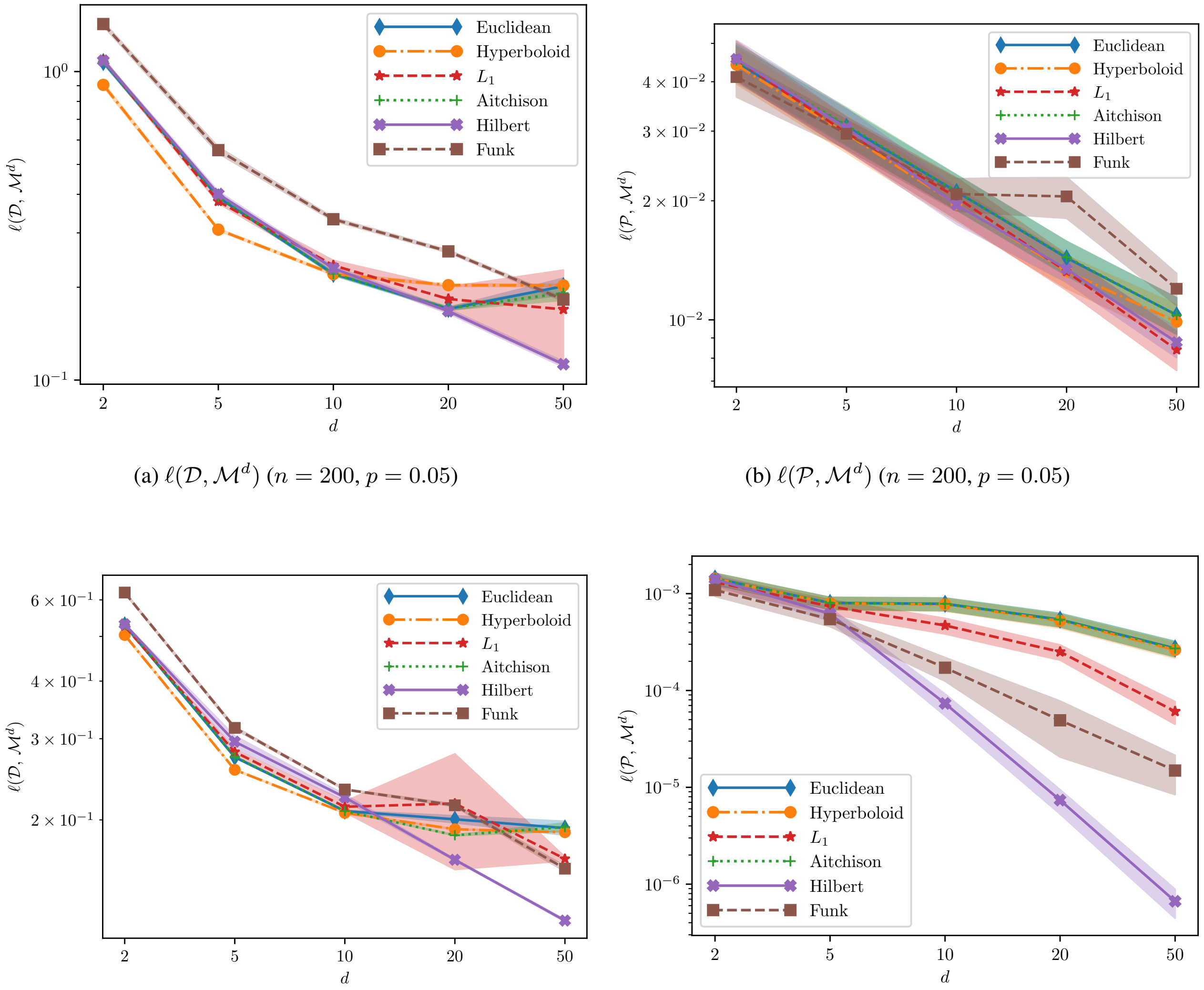
For all simplex embeddings (Hilbert, Funk, Aitchison), we represent the embedding in the log-coordinates $l(p) = (\log p_1, \ldots, \log p_d)$. Because $\rho_{\mathrm{HG}}(p,q) = \|l(p) - l(q)\|_{\mathrm{var}} = \|l(\tilde{p}) - l(\tilde{q})\|_{\mathrm{var}}$, we can directly optimize the Hilbert simplex embedding on the coordinates $l(\tilde{p})$ which are free points in $\mathbb{R}^d$. For Funk and Aitchison embeddings, we need to ensure that the embedding to be optimized can be mapped back into the simplex domain.

B. Experimental Results

Figure 10 shows $\ell(\mathcal{D},\mathcal{M}^d)$ (left) and $\ell(\mathcal{P},\mathcal{M}^d)$ (right) against d on the Erdős–Rényi random graph dataset with $p = 0.05$ and $p = 0.5$ (in the main text we studied the case when $p = 0.2$), where p is the probability for any pair of nodes i and j to be connected by an edge.

Figure 11 shows $\ell(\mathcal{D},\mathcal{M}^d)$ (left) and $\ell(\mathcal{P},\mathcal{M}^d)$ (right) against d on the Barabási–Albert graphs $G(n,m)$ with $m = 1$ and $m = 3$ (in the main text we only studied the case when $m = 2$), where m is the number of edges to attach when a new node is created.

In both figures, $\mathcal{P}$ is random walk similarities on these graph datasets. We do not simulate real random walks as in graph embedding methods. Instead, we use the graph adjacency matrix to construct the transition probability matrix, whose matrix power gives the random walk similarity matrix $\mathcal{P}$.



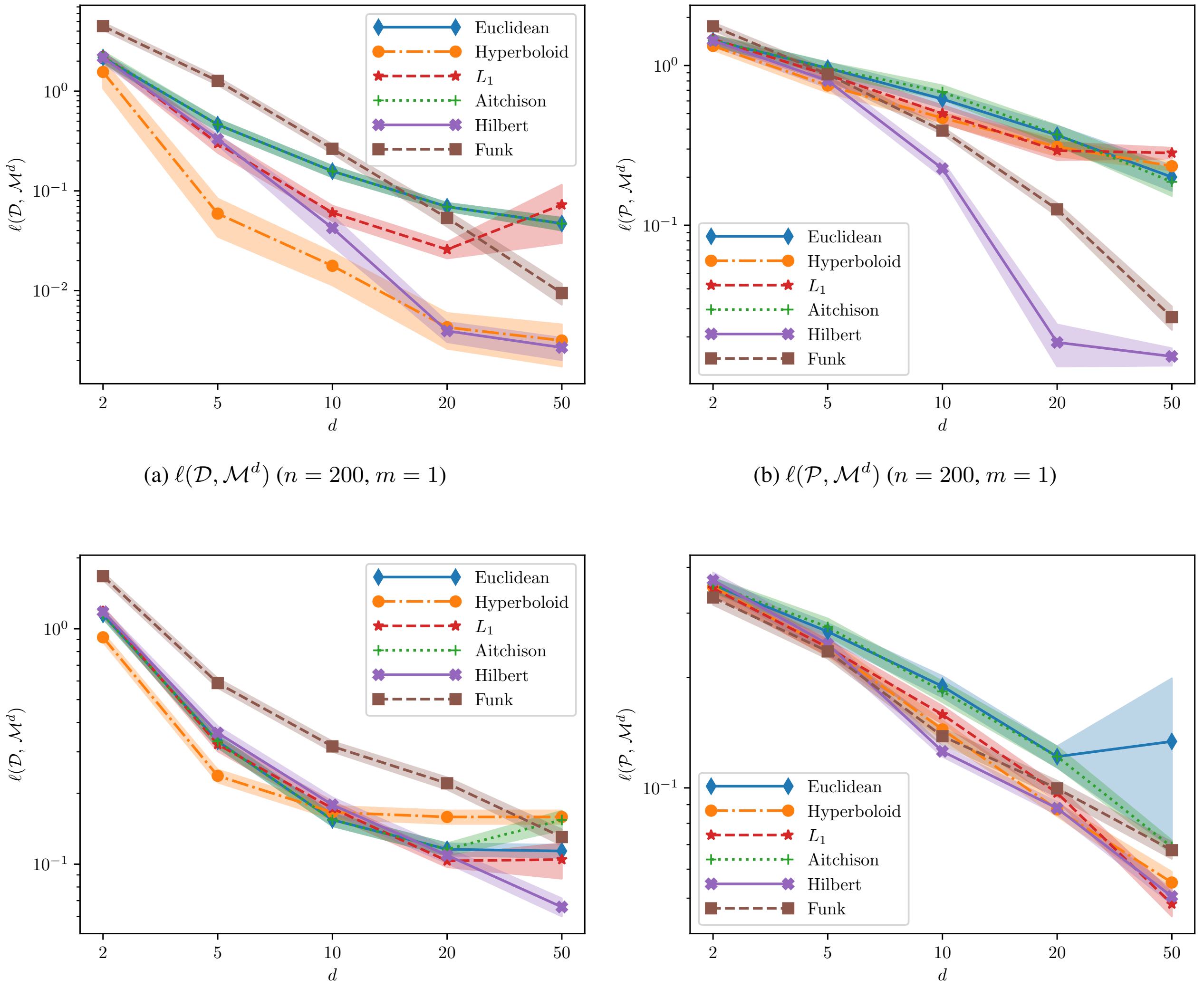
(a) $\ell(\mathcal{D}, \mathcal{M}^d)$ ($n = 200$, $p = 0.05$)

(b) $\ell(\mathcal{P}, \mathcal{M}^d)$ ($n = 200$, $p = 0.05$)

(c) $\ell(\mathcal{D}, \mathcal{M}^d)$ ($n = 200$, $p = 0.5$)

(d) $\ell(\mathcal{P}, \mathcal{M}^d)$ ($n = 200$, $p = 0.5$)

Figure 10. Embedding losses against d (Erdős–Rényi random graph $G(n, p)$).



(a) $\ell(\mathcal{D}, \mathcal{M}^d)$ ($n = 200$, $m = 1$)

(b) $\ell(\mathcal{P}, \mathcal{M}^d)$ ($n = 200$, $m = 1$)

(c) $\ell(\mathcal{D}, \mathcal{M}^d)$ ($n = 200$, $m = 3$)

(d) $\ell(\mathcal{P}, \mathcal{M}^d)$ ($n = 200$, $m = 3$)

Figure 11. Embedding losses against d (Barabási–Albert graphs $G(n, m)$).