Ultrahyperbolic Representation Learning

Marc T. Law   **Jos Stam**

NVIDIA

Abstract

In machine learning, data is usually represented in a (flat) Euclidean space where distances between points are along straight lines. Researchers have recently considered more exotic (non-Euclidean) Riemannian manifolds such as hyperbolic space which is well suited for tree-like data. In this paper, we propose a representation living on a pseudo-Riemannian manifold of constant nonzero curvature. It is a generalization of hyperbolic and spherical geometries where the nondegenerate metric tensor need not be positive definite. We provide the necessary learning tools in this geometry and extend gradient-based optimization techniques. More specifically, we provide closed-form expressions for distances via geodesics and define a descent direction to minimize some objective function. Our novel framework is applied to graph representations.

1 Introduction

In most machine learning applications, data representations lie on a smooth manifold [16] and the training procedure is optimized with an iterative algorithm such as line search or trust region methods [20]. In most cases, the smooth manifold is Riemannian, which means that it is equipped with a positive definite metric. Due to the positive definiteness of the metric, the negative of the (Riemannian) gradient is a descent direction that can be exploited to iteratively minimize some objective function [1].

The choice of metric on the Riemannian manifold determines how relations between points are quantified. The most common Riemannian manifold is the flat Euclidean space, which has constant zero curvature and the distances between points are measured by straight lines. An intuitive example of non-Euclidean Riemannian manifold is the spherical model (*i.e.* representations lie on a sphere) that has constant positive curvature and is used for instance in face recognition [25, 26]. On the sphere, geodesic distances are a function of angles. Similarly, Riemannian spaces of constant negative curvature are called hyperbolic [23]. Such spaces were shown by Gromov to be well suited to represent tree-like structures [10]. The machine learning community has adopted these spaces to learn tree-like graphs [5] and hierarchical data structures [11, 18, 19], and also to compute means in tree-like shapes [6, 7].

In this paper, we consider a class of pseudo-Riemannian manifolds of constant nonzero curvature [28] not previously considered in machine learning. These manifolds not only generalize the hyperbolic and spherical geometries mentioned above, but also contain hyperbolic and spherical submanifolds and can therefore describe relationships specific to those geometries. The difference is that we consider the larger class of pseudo-Riemannian manifolds where the considered nondegenerate metric tensor need not be positive definite. Optimizing a cost function on our non-flat ultrahyperbolic space requires a descent direction method that follows a path along the curved manifold. We achieve this by employing tools from differential geometry such as geodesics and exponential maps. The theoretical contributions in this paper are two-fold: (1) explicit methods to calculate dissimilarities and (2) general optimization tools on pseudo-Riemannian manifolds of constant nonzero curvature.
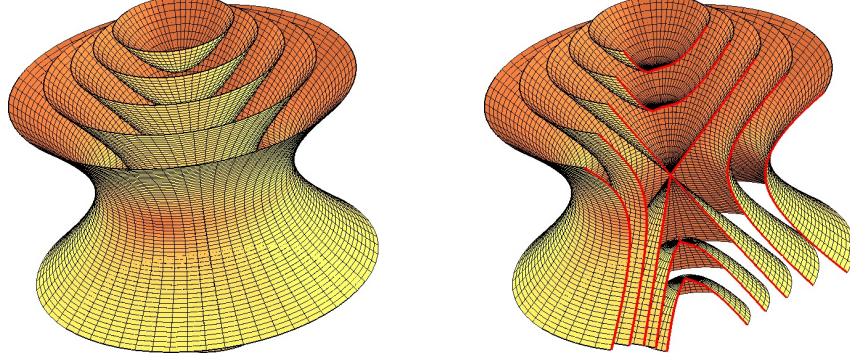
Figure 1: Two iso-surface depictions of an orthographic projection of the same pseudo-hyperboloid $\mathcal{Q}_{-1}^{2,1}$ into $\mathbb{R}^3$ along one time dimension. It contains the entire family of hyperboloids as submanifolds.

2 Pseudo-Hyperboloids

Notation: We denote points on a smooth manifold $\mathcal{M}$ [16] by boldface Roman characters $\mathbf{x} \in \mathcal{M}$. $T_{\mathbf{x}}\mathcal{M}$ is the tangent space of $\mathcal{M}$ at $\mathbf{x}$ and we write tangent vectors $\boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{M}$ in boldface Greek fonts. $\mathbb{R}^d$ is the (flat) d-dimensional Euclidean space, it is equipped with the (positive definite) dot product denoted by $\langle \cdot, \cdot \rangle$ and defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$. The ℓ_2-norm of $\mathbf{x}$ is $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. $\mathbb{R}_*^d = \mathbb{R}^d \backslash \{\mathbf{0}\}$ is the Euclidean space with the origin removed.

Pseudo-Riemannian manifolds: A smooth manifold $\mathcal{M}$ is pseudo-Riemannian (also called semi-Riemannian [21]) if it is equipped with a pseudo-Riemannian metric tensor (named "metric" for short in differential geometry). The pseudo-Riemannian metric $g_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{M} \times T_{\mathbf{x}}\mathcal{M} \to \mathbb{R}$ at some point $\mathbf{x} \in \mathcal{M}$ is a nondegenerate symmetric bilinear form. Nondegeneracy means that if for a given $\boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{M}$ and for all $\boldsymbol{\zeta} \in T_{\mathbf{x}}\mathcal{M}$ we have $g_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\zeta}) = 0$, then $\boldsymbol{\xi} = \mathbf{0}$. If the metric is also positive definite (*i.e.* $\forall \boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{M}$, $g_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\xi}) > 0$ iff $\boldsymbol{\xi} \neq \mathbf{0}$), then it is Riemannian. Riemannian geometry is a special case of pseudo-Riemannian geometry where the metric is positive definite. In general, this is not the case and non-Riemannian manifolds distinguish themselves by having some non-vanishing tangent vectors $\boldsymbol{\xi} \neq \mathbf{0}$ that satisfy $g_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\xi}) \leq 0$. We refer the reader to [2, 21, 28] for details.

Pseudo-hyperboloids generalize spherical and hyperbolic manifolds to the class of pseudo-Riemannian manifolds. Let us note $d = p + q + 1 \in \mathbb{N}$ the dimensionality of some pseudo-Euclidean space where each vector is written $\mathbf{x} = (x_0, x_1, \cdots, x_{q+p})^\top$. That space is denoted by $\mathbb{R}^{p,q+1}$ when it is equipped with the following scalar product (*i.e.* nondegenerate symmetric bilinear form [21]):

$$\forall \mathbf{a} = (a_0, \cdots, a_{q+p})^\top, \ \mathbf{b} = (b_0, \cdots, b_{q+p})^\top, \ \langle \mathbf{a}, \mathbf{b} \rangle_q = -\sum_{i=0}^{q} a_i b_i + \sum_{j=q+1}^{p+q} a_j b_j = \mathbf{a}^\top \mathbf{G} \mathbf{b}, \quad (1)$$

where $\mathbf{G} = \mathbf{G}^{-1} = \mathbf{I}_{q+1,p}$ is the $d \times d$ diagonal matrix with the first $q + 1$ diagonal elements equal to -1 and the remaining p equal to 1. Since $\mathbb{R}^{p,q+1}$ is a vector space, we can identify the tangent space to the space itself by means of the natural isomorphism $T_{\mathbf{x}}\mathbb{R}^{p,q+1} \approx \mathbb{R}^{p,q+1}$. Using the terminology of special relativity, $\mathbb{R}^{p,q+1}$ has $q + 1$ time dimensions and p space dimensions.

A pseudo-hyperboloid is the following submanifold of codimension one (*i.e.* hypersurface) in $\mathbb{R}^{p,q+1}$:

$$\mathcal{Q}_{\beta}^{p,q} = \left\{ \mathbf{x} = (x_0, x_1, \cdots, x_{p+q})^\top \in \mathbb{R}^{p,q+1} : \|\mathbf{x}\|_q^2 = \beta \right\}, \quad (2)$$

where $\beta \in \mathbb{R}_*$ is a nonzero real number and the function $\| \cdot \|_q^2$ given by $\|\mathbf{x}\|_q^2 = \langle \mathbf{x}, \mathbf{x} \rangle_q$ is the associated quadratic form of the scalar product. It is equivalent to work with either $\mathcal{Q}_{\beta}^{p,q}$ or $\mathcal{Q}_{-\beta}^{q+1,p-1}$ as they are interchangeable via an anti-isometry (see supp. material). For instance, the unit q-sphere $\mathcal{S}^q = \{\mathbf{x} \in \mathbb{R}^{q+1} : \|\mathbf{x}\| = 1\}$ is anti-isometric to $\mathcal{Q}_{-1}^{0,q}$ which is then spherical. In the literature, the set $\mathcal{Q}_{\beta}^{p,q}$ is called a "pseudo-sphere" when $\beta > 0$ and a "pseudo-hyperboloid" when $\beta < 0$. In the rest of the paper, we only consider the pseudo-hyperbolic case (*i.e.* $\beta < 0$). Moreover, for any $\beta < 0$, $\mathcal{Q}_{\beta}^{p,q}$ is homothetic to $\mathcal{Q}_{-1}^{p,q}$, the value of β can then be considered to be -1. We can obtain the

2

spherical and hyperbolic geometries by constraining all the elements of the space dimensions of a pseudo-hyperboloid to be zero or constraining all the elements of the time dimensions except one to be zero, respectively. Pseudo-hyperboloids then generalize spheres and hyperboloids.

The pseudo-hyperboloids that we consider in this paper are hard to visualize as they live in ambient spaces with dimension higher than 3. In Fig. 1, we show iso-surfaces of a projection of the 3-dimensional pseudo-hyperboloid $\mathcal{Q}_{-1}^{2,1}$ (embedded in $\mathbb{R}^{2,2}$) into $\mathbb{R}^3$ along its first time dimension.

Metric tensor and tangent space: The metric tensor at $\mathbf{x} \in \mathcal{Q}_{\beta}^{p,q}$ is $g_{\mathbf{x}}(\cdot,\cdot) = \langle \cdot, \cdot \rangle_q$ where $g_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{Q}_{\beta}^{p,q} \times T_{\mathbf{x}}\mathcal{Q}_{\beta}^{p,q} \to \mathbb{R}$. By using the isomorphism $T_{\mathbf{x}}\mathbb{R}^{p,q+1} \approx \mathbb{R}^{p,q+1}$ mentioned above, the tangent space of $\mathcal{Q}_{\beta}^{p,q}$ at $\mathbf{x}$ can be defined as $T_{\mathbf{x}}\mathcal{Q}_{\beta}^{p,q} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^{p,q+1} : \langle \mathbf{x}, \boldsymbol{\xi} \rangle_q = 0 \right\}$ for all $\beta \neq 0$. Finally, the orthogonal projection of an arbitrary d-dimensional vector $\mathbf{z}$ onto $T_{\mathbf{x}}\mathcal{Q}_{\beta}^{p,q}$ is:

$$\Pi_{\mathbf{x}}(\mathbf{z}) = \mathbf{z} - \frac{\langle \mathbf{z}, \mathbf{x} \rangle_q}{\langle \mathbf{x}, \mathbf{x} \rangle_q} \mathbf{x}. \tag{3}$$

3 Measuring Dissimilarity on Pseudo-Hyperboloids

This section introduces the differential geometry tools necessary to quantify dissimilarities/distances between points on $\mathcal{Q}_{\beta}^{p,q}$. Measuring dissimilarity is an important task in machine learning and has many applications (*e.g.* in metric learning [29]).

Intrinsic geometry: The *intrinsic geometry* of the hypersurface $\mathcal{Q}_{\beta}^{p,q}$ embedded in $\mathbb{R}^{p,q+1}$ (*i.e.* the geometry perceived by the inhabitants of $\mathcal{Q}_{\beta}^{p,q}$ [21]) derives solely from its metric tensor applied to tangent vectors to $\mathcal{Q}_{\beta}^{p,q}$. For instance, it can be used to measure the arc length of a tangent vector joining two points along a geodesic and define their geodesic distance. Before considering geodesic distances, we consider extrinsic distances (*i.e.* distances in the ambient space $\mathbb{R}^{p,q+1}$). Since $\mathbb{R}^{p,q+1}$ is isomorphic to its tangent space, tangent vectors to $\mathbb{R}^{p,q+1}$ are naturally identified with points. Using the quadratic form of Eq. (1), the extrinsic distance between two points $\mathbf{a}, \mathbf{b} \in \mathcal{Q}_{\beta}^{p,q}$ is:

$$\mathsf{d}_q(\mathbf{a}, \mathbf{b}) = \sqrt{|\|\mathbf{a} - \mathbf{b}\|_q^2|} = \sqrt{|\|\mathbf{a}\|_q^2 + \|\mathbf{b}\|_q^2 - 2\langle \mathbf{a}, \mathbf{b} \rangle_q|} = \sqrt{|2\beta - 2\langle \mathbf{a}, \mathbf{b} \rangle_q|}. \tag{4}$$

This distance is a good proxy for the geodesic distance $\mathsf{d}_\gamma(\cdot,\cdot)$, that we introduce below, if it preserves *distance relations*: $\mathsf{d}_\gamma(\mathbf{a}, \mathbf{b}) < \mathsf{d}_\gamma(\mathbf{c}, \mathbf{d})$ iff $\mathsf{d}_q(\mathbf{a}, \mathbf{b}) < \mathsf{d}_q(\mathbf{c}, \mathbf{d})$. This relation is satisfied for two special cases of pseudo-hyperboloids for which the geodesic distance is well known:

• **Spherical manifold ($\mathcal{Q}_{\beta}^{0,q}$):** If $p = 0$, the geodesic distance $\mathsf{d}_\gamma(\mathbf{a}, \mathbf{b}) = \sqrt{|\beta|} \cos^{-1}\left(\frac{\langle \mathbf{a}, \mathbf{b} \rangle_q}{\beta}\right)$ is called spherical distance. In practice, the cosine similarity $\frac{\langle \cdot, \cdot \rangle_q}{\beta}$ is often considered instead of $\mathsf{d}_\gamma(\cdot,\cdot)$ since it satisfies $\mathsf{d}_\gamma(\mathbf{a}, \mathbf{b}) < \mathsf{d}_\gamma(\mathbf{c}, \mathbf{d})$ iff $\langle \mathbf{a}, \mathbf{b} \rangle_q < \langle \mathbf{c}, \mathbf{d} \rangle_q$ iff $\mathsf{d}_q(\mathbf{a}, \mathbf{b}) < \mathsf{d}_q(\mathbf{c}, \mathbf{d})$.

• **Hyperbolic manifold (upper sheet of the two-sheet hyperboloid $\mathcal{Q}_{\beta}^{p,0}$):** If $q = 0$, the geodesic distance $\mathsf{d}_\gamma(\mathbf{a}, \mathbf{b}) = \sqrt{|\beta|} \cosh^{-1}\left(\frac{\langle \mathbf{a}, \mathbf{b} \rangle_q}{\beta}\right)$ with $a_0 > 0$ and $b_0 > 0$ is called Poincaré distance [19]. The (extrinsic) Lorentzian distance was shown to be a good proxy in hyperbolic geometry [11].

For the **ultrahyperbolic** case (*i.e.* $q \geq 1$ and $p \geq 2$), the distance relations are not preserved: $\mathsf{d}_\gamma(\mathbf{a}, \mathbf{b}) < \mathsf{d}_\gamma(\mathbf{c}, \mathbf{d}) \not\Longleftrightarrow \mathsf{d}_q(\mathbf{a}, \mathbf{b}) < \mathsf{d}_q(\mathbf{c}, \mathbf{d})$. We then need to consider only geodesic distances. This section introduces closed-form expressions for geodesic distances on ultrahyperbolic manifolds.

Geodesics: Informally, a geodesic is a curve joining points on a manifold $\mathcal{M}$ that minimizes some "effort" depending on the metric. More precisely, let $I \subseteq \mathbb{R}$ be a (maximal) interval containing 0. A geodesic $\gamma : I \to \mathcal{M}$ maps a real value $t \in I$ to a point on the manifold $\mathcal{M}$. It is a curve on $\mathcal{M}$ defined by its initial point $\gamma(0) = \mathbf{x} \in \mathcal{M}$ and initial tangent vector $\gamma'(0) = \boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{M}$ where $\gamma'(t)$ is the derivative of γ at t. By analogy with physics, t is considered as a time value. Intuitively, one can think of the curve as the trajectory over time of a ball being pushed from a point $\mathbf{x}$ at $t = 0$ with initial velocity $\boldsymbol{\xi}$ and constrained to roll on the manifold. We denote this curve explicitly by $\gamma_{\mathbf{x} \to \boldsymbol{\xi}}(t)$ unless the dependence is obvious from the context. For this curve to be a geodesic, its acceleration has to be zero: $\forall t \in I, \gamma''(t) = \mathbf{0}$. This condition is a second-order ordinary differential equation that has a unique solution for a given set of initial conditions [17]. The interval I is said to be maximal if it cannot be extended to a larger interval. In the case of $\mathcal{Q}_{\beta}^{p,q}$, we have $I = \mathbb{R}$ and I is then maximal.
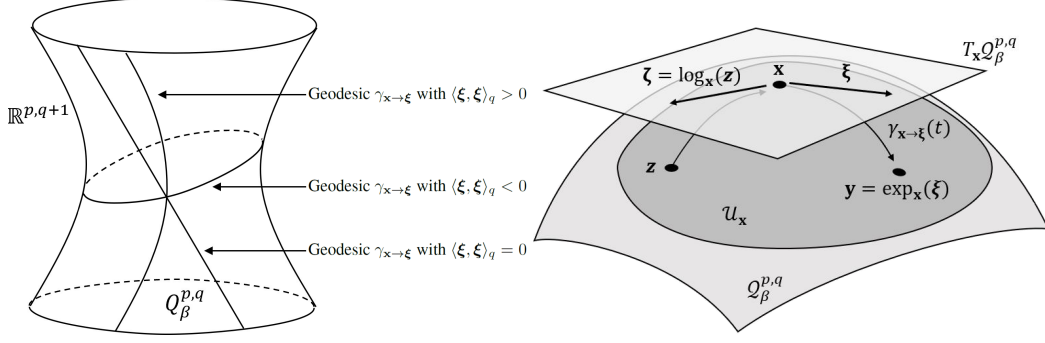
Figure 2: (left) Illustration of the three types of geodesics of a pseudo-hyperboloid defined in Eq. (5). (right) Exponential map and logarithm map.

Geodesic of $\mathcal{Q}_\beta^{p,q}$: As we show in the supp. material, the geodesics of $\mathcal{Q}_\beta^{p,q}$ are a combination of the hyperbolic, flat and spherical cases. The nature of the geodesic $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}$ depends on the sign of $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q$. For all $t \in \mathbb{R}$, the geodesic $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}$ of $\mathcal{Q}_\beta^{p,q}$ with $\beta < 0$ is written:

$$
\gamma_{\mathbf{x}\to\boldsymbol{\xi}}(t) = \begin{cases} \cosh\left(\frac{t\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right)\mathbf{x} + \frac{\sqrt{|\beta|}}{\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}\sinh\left(\frac{t\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right)\boldsymbol{\xi} & \text{if } \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q > 0 \\ \mathbf{x} + t\boldsymbol{\xi} & \text{if } \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q = 0 \\ \cos\left(\frac{t\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right)\mathbf{x} + \frac{\sqrt{|\beta|}}{\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}\sin\left(\frac{t\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right)\boldsymbol{\xi} & \text{if } \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q < 0 \end{cases} \quad (5)
$$

We recall that $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q = 0$ does not imply $\boldsymbol{\xi} = \mathbf{0}$. The geodesics are an essential ingredient to define a mapping known as the exponential map. See Fig. 2 (left) for a depiction of these three types of geodesics, and Fig. 2 (right) for a depiction of the other quantities introduced in this section.

Exponential map: Exponential maps are a way of collecting all of the geodesics of a pseudo-Riemannian manifold $\mathcal{M}$ into a unique differentiable mapping. Let $\mathcal{D}_\mathbf{x} \subseteq T_\mathbf{x}\mathcal{M}$ be the set of tangent vectors $\boldsymbol{\xi}$ such that $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}$ is defined at least on the interval $[0,1]$. This allows us to uniquely define the exponential map $\exp_\mathbf{x} : \mathcal{D}_\mathbf{x} \to \mathcal{M}$ such that $\exp_\mathbf{x}(\boldsymbol{\xi}) = \gamma_{\mathbf{x}\to\boldsymbol{\xi}}(1)$.

The manifold $\mathcal{Q}_\beta^{p,q}$ is geodesically complete, the domain of its exponential map is then $\mathcal{D}_\mathbf{x} = T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$. Using Eq. (5) with $t = 1$, we obtain an exponential map of the entire tangent space to the manifold:

$$
\forall\boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}, \ \exp_\mathbf{x}(\boldsymbol{\xi}) = \gamma_{\mathbf{x}\to\boldsymbol{\xi}}(1). \quad (6)
$$

We make the important observation that the image of the exponential map does not necessarily cover the entire manifold: not all points on a manifold are connected by a geodesic. This is the case for our pseudo-hyperboloids. Namely, for a given point $\mathbf{x} \in \mathcal{Q}_\beta^{p,q}$ there exist points $\mathbf{y}$ that are not in the image of the exponential map (*i.e.* there does not exist a tangent vector $\boldsymbol{\xi}$ such that $\mathbf{y} = \exp_\mathbf{x}(\boldsymbol{\xi})$).

Logarithm map: We provide a closed-form expression of the logarithm map for pseudo-hyperboloids. Let $\mathcal{U}_\mathbf{x} \subseteq \mathcal{Q}_\beta^{p,q}$ be some neighborhood of $\mathbf{x}$. The logarithm map $\log_\mathbf{x} : \mathcal{U}_\mathbf{x} \to T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ is defined as the inverse of the exponential map on $\mathcal{U}_\mathbf{x}$ (*i.e.* $\log_\mathbf{x} = \exp_\mathbf{x}^{-1}$). We propose:

$$
\forall\mathbf{y} \in \mathcal{U}_\mathbf{x}, \ \log_\mathbf{x}(\mathbf{y}) = \begin{cases} \frac{\cosh^{-1}(\frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{\beta})}{\sqrt{(\frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{\beta})^2-1}}\left(\mathbf{y} - \frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{\beta}\mathbf{x}\right) & \text{if } \frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{|\beta|} < -1 \\ \mathbf{y} - \mathbf{x} & \text{if } \frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{|\beta|} = -1 \\ \frac{\cos^{-1}(\frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{\beta})}{\sqrt{1-(\frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{\beta})^2}}\left(\mathbf{y} - \frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{\beta}\mathbf{x}\right) & \text{if } \frac{\langle\mathbf{x},\mathbf{y}\rangle_q}{|\beta|} \in (-1,1) \end{cases} \quad (7)
$$

By substituting $\boldsymbol{\xi} = \log_\mathbf{x}(\mathbf{y})$ into Eq. (6), one can verify that our formulas are the inverse of the exponential map. The set $\mathcal{U}_\mathbf{x} = \left\{\mathbf{y} \in \mathcal{Q}_\beta^{p,q} : \langle\mathbf{x},\mathbf{y}\rangle_q < |\beta|\right\}$ is called a normal neighborhood of $\mathbf{x} \in \mathcal{Q}_\beta^{p,q}$ since for all $\mathbf{y} \in \mathcal{U}_\mathbf{x}$, there exists a geodesic from $\mathbf{x}$ to $\mathbf{y}$ such that $\log_\mathbf{x}(\mathbf{y}) = \gamma'_{\mathbf{x}\to\log_\mathbf{x}(\mathbf{y})}(0)$. We show in the supp. material that the logarithm map is not defined if $\langle\mathbf{x},\mathbf{y}\rangle_q \geq |\beta|$.

Proposed dissmilarity: We define our dissimilarity function based on the general notion of arc length and radius function on pseudo-Riemannian manifolds that we recall in the next paragraph (see details in Chapter 5 of [21]). This corresponds to the geodesic distance in the Riemannian case.

Let $\mathcal{U}_{\mathbf{x}}$ be a normal neighborhood of $\mathbf{x} \in \mathcal{M}$ with $\mathcal{M}$ pseudo-Riemannian. The *radius function* $r_{\mathbf{x}} : \mathcal{U}_{\mathbf{x}} \to \mathbb{R}$ is defined as $r_{\mathbf{x}}(\mathbf{y}) = \sqrt{|g_{\mathbf{x}}(\log_{\mathbf{x}}(\mathbf{y}), \log_{\mathbf{x}}(\mathbf{y}))|}$ where $g_{\mathbf{x}}$ is the metric at $\mathbf{x}$. If $\sigma_{\mathbf{x} \to \boldsymbol{\xi}}$ is the radial geodesic from $\mathbf{x}$ to $\mathbf{y} \in \mathcal{U}_{\mathbf{x}}$ (*i.e.* $\boldsymbol{\xi} = \log_{\mathbf{x}}(\mathbf{y})$), then the arc length of $\sigma_{\mathbf{x} \to \boldsymbol{\xi}}$ equals $r_{\mathbf{x}}(\mathbf{y})$.

We then define the geodesic "distance" between $\mathbf{x} \in \mathcal{Q}_{\beta}^{p,q}$ and $\mathbf{y} \in \mathcal{U}_{\mathbf{x}}$ as the arc length of $\sigma_{\mathbf{x} \to \log_{\mathbf{x}}(\mathbf{y})}$:

$$\mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{y}) = \sqrt{|\|\log_{\mathbf{x}}(\mathbf{y})\|_q^2|} = \begin{cases} \sqrt{|\beta|} \cosh^{-1}\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle_q}{\beta}\right) & \text{if } \frac{\langle \mathbf{x}, \mathbf{y} \rangle_q}{|\beta|} < -1 \\ 0 & \text{if } \frac{\langle \mathbf{x}, \mathbf{y} \rangle_q}{|\beta|} = -1 \\ \sqrt{|\beta|} \cos^{-1}\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle_q}{\beta}\right) & \text{if } \frac{\langle \mathbf{x}, \mathbf{y} \rangle_q}{|\beta|} \in (-1, 1) \end{cases} \quad (8)$$

It is important to note that our "distance" is **not** a distance metric. However, it satisfies the axioms of a *symmetric premetric*: (i) $\mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{y}) = \mathsf{d}_{\gamma}(\mathbf{y}, \mathbf{x}) \geq 0$ and (ii) $\mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{x}) = 0$. These conditions are sufficient to quantify the notion of nearness via a ρ-ball centered at $\mathbf{x}$: $B_{\mathbf{x}}^{\rho} = \{\mathbf{y} : \mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{y}) < \rho\}$.

In general, topological spaces provide a qualitative (not necessarily quantitative) way to detect "nearness" through the concept of a neighborhood at a point [15]. Something is true "near $\mathbf{x}$" if it is true in the neighborhood of $\mathbf{x}$ (*e.g.* in $B_{\mathbf{x}}^{\rho}$). Our premetric is similar to metric learning methods [13, 14, 29] that learn a Mahalanobis-like distance pseudo-metric parameterized by a positive semi-definite matrix. Pairs of distinct points can have zero "distance" if the matrix is not positive definite. However, unlike classic metric learning, we can have triplets $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ that satisfy $\mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{y}) = \mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{z}) = 0$ but $\mathsf{d}_{\gamma}(\mathbf{y}, \mathbf{z}) > 0$ (*e.g.* $\mathbf{x} = (1, 0, 0, 0)^{\top}, \mathbf{y} = (1, 1, 1, 0)^{\top}, \mathbf{z} = (1, 1, 0, 1)^{\top}$ in $\mathcal{Q}_{-1}^{2,1}$).

Since the logarithm map is not defined if $\langle \mathbf{x}, \mathbf{y} \rangle_q \geq |\beta|$, we propose to use the following continuous approximation defined on the whole manifold instead:

$$\forall \mathbf{x} \in \mathcal{Q}_{\beta}^{p,q}, \mathbf{y} \in \mathcal{Q}_{\beta}^{p,q}, \; \mathsf{D}_{\gamma}(\mathbf{x}, \mathbf{y}) = \begin{cases} \mathsf{d}_{\gamma}(\mathbf{x}, \mathbf{y}) & \text{if } \langle \mathbf{x}, \mathbf{y} \rangle_q \leq 0 \\ \sqrt{|\beta|}\left(\frac{\pi}{2} + \frac{\langle \mathbf{x}, \mathbf{y} \rangle_q}{|\beta|}\right) & \text{otherwise} \end{cases} \quad (9)$$

To the best of our knowledge, the explicit formulation of the logarithm map for $\mathcal{Q}_{\beta}^{p,q}$ in Eq. (7) and its corresponding radius function in Eq. (8) to define a dissimilarity function are novel. We have also proposed some linear approximation to evaluate dissimilarity when the logarithm map is not defined but other choices are possible. For instance, when a geodesic does not exist, a standard way in differential geometry to calculate curves is to consider broken geodesics. One might consider instead the dissimilarity $\mathsf{d}_{\gamma}(\mathbf{x}, -\mathbf{x}) + \mathsf{d}_{\gamma}(-\mathbf{x}, \mathbf{y}) = \pi\sqrt{|\beta|} + \mathsf{d}_{\gamma}(-\mathbf{x}, \mathbf{y})$ if $\log_{\mathbf{x}}(\mathbf{y})$ is not defined since $-\mathbf{x} \in \mathcal{Q}_{\beta}^{p,q}$ and $\log_{-\mathbf{x}}(\mathbf{y})$ is defined. This interesting problem is left for future research.

4 Ultrahyperbolic Optimization

In this section we present optimization frameworks to optimize any differentiable function defined on $\mathcal{Q}_{\beta}^{p,q}$. Our goal is to compute descent directions on the ultrahyperbolic manifold. We consider two approaches. In the first approach, we map our representation from Euclidean space to ultrahyperbolic space. This is similar to the approach taken by [11] in hyperbolic space. In the second approach, we optimize using gradients defined directly in pseudo-Riemannian tangent space. We propose a novel descent direction which guarantees the minimization of some cost function.

4.1 Euclidean optimization via a differentiable mapping onto $\mathcal{Q}_{\beta}^{p,q}$

Our first method maps Euclidean representations that lie in $\mathbb{R}^d$ to the pseudo-hyperboloid $\mathcal{Q}_{\beta}^{p,q}$, and the chain rule is exploited to perform standard gradient descent. To this end, we construct a differentiable mapping $\varphi : \mathbb{R}_*^{q+1} \times \mathbb{R}^p \to \mathcal{Q}_{\beta}^{p,q}$. The image of a point already on $\mathcal{Q}_{\beta}^{p,q}$ under the mapping φ is itself: $\forall \mathbf{x} \in \mathcal{Q}_{\beta}^{p,q}, \varphi(\mathbf{x}) = \mathbf{x}$. Let $\mathcal{S}^q = \{\mathbf{x} \in \mathbb{R}^{q+1} : \|\mathbf{x}\| = 1\}$ denote the unit q-sphere. We first introduce the following diffeomorphisms:

Theorem 4.1 (Diffeomorphisms). *For any $\beta < 0$, there is a diffeomorphism $\psi : \mathcal{Q}_\beta^{p,q} \to \mathcal{S}^q \times \mathbb{R}^p$. Let us note $\mathbf{x} = \begin{pmatrix} \mathbf{t} \\ \mathbf{s} \end{pmatrix} \in \mathcal{Q}_\beta^{p,q}$ with $\mathbf{t} \in \mathbb{R}_*^{q+1}$ and $\mathbf{s} \in \mathbb{R}^p$, let us note $\mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \in \mathcal{S}^q \times \mathbb{R}^p$ where $\mathbf{u} \in \mathcal{S}^q$ and $\mathbf{v} \in \mathbb{R}^p$. The mapping ψ and its inverse ψ^{-1} are formulated (see proofs in supp. material):*

$$\psi(\mathbf{x}) = \begin{pmatrix} \frac{1}{\|\mathbf{t}\|}\mathbf{t} \\ \frac{1}{\sqrt{|\beta|}}\mathbf{s} \end{pmatrix} \qquad and \qquad \psi^{-1}(\mathbf{z}) = \sqrt{|\beta|} \begin{pmatrix} \sqrt{1 + \|\mathbf{v}\|^2}\mathbf{u} \\ \mathbf{v} \end{pmatrix}. \tag{10}$$

With these mappings, any vector $\mathbf{x} \in \mathbb{R}_*^{q+1} \times \mathbb{R}^p$ can be mapped to $\mathcal{Q}_\beta^{p,q}$ via $\varphi = \psi^{-1} \circ \psi$. φ is differentiable everywhere except when $x_0 = \cdots = x_q = 0$, which should never occur in practice. It can therefore be optimized using standard gradient methods.

4.2 Pseudo-Riemannian optimization

We now introduce a novel method to optimize any differentiable function $f : \mathcal{Q}_\beta^{p,q} \to \mathbb{R}$ defined on the pseudo-hyperboloid. As we show below, the (negative of the) pseudo-Riemannian gradient is not a descent direction. We propose a simple and efficient way to calculate a descent direction.

Pseudo-Riemannian gradient: Since $\mathbf{x} \in \mathcal{Q}_\beta^{p,q}$ also lies in the Euclidean ambient space $\mathbb{R}^d$, the function f has a well defined Euclidean gradient $\nabla f(\mathbf{x}) = (\partial f(\mathbf{x})/\partial x_0, \cdots, \partial f(\mathbf{x})/\partial x_{p+q})^\top \in \mathbb{R}^d$. The gradient of f in the pseudo-Euclidean ambient space $\mathbb{R}^{p,q+1}$ is $(\mathbf{G}^{-1}\nabla f(\mathbf{x})) = (\mathbf{G}\nabla f(\mathbf{x})) \in \mathbb{R}^{p,q+1}$. Since $\mathcal{Q}_\beta^{p,q}$ is a submanifold of $\mathbb{R}^{p,q+1}$, the pseudo-Riemannian gradient $Df(\mathbf{x}) \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ of f on $\mathcal{Q}_\beta^{p,q}$ is the orthogonal projection of $(\mathbf{G}\nabla f(\mathbf{x}))$ onto $T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ (see Chapter 4 of [21]):

$$Df(\mathbf{x}) = \Pi_\mathbf{x}\left(\mathbf{G}\nabla f(\mathbf{x})\right) = \mathbf{G}\nabla f(\mathbf{x}) - \frac{\langle \mathbf{G}\nabla f(\mathbf{x}), \mathbf{x}\rangle_q}{\langle \mathbf{x}, \mathbf{x}\rangle_q}\mathbf{x} = \mathbf{G}\nabla f(\mathbf{x}) - \frac{\langle \nabla f(\mathbf{x}), \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle_q}\mathbf{x}. \tag{11}$$

This gradient forms the foundation of our descent method optimizer as will be shown in Eq. (13).

Iterative optimization: Our goal is to iteratively decrease the value of the function f by following some descent direction. Since $\mathcal{Q}_\beta^{p,q}$ is not a vector space, we do not "follow the descent direction" by adding the descent direction multiplied by a step size as this would result in a new point that does not necessarily lie on $\mathcal{Q}_\beta^{p,q}$. Instead, to remain on the manifold, we use our exponential map defined in Eq. (6). This is a standard way to optimize on Riemannian manifolds [1]. Given a step size $t > 0$, one step of descent along a tangent vector $\boldsymbol{\zeta} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ is given by:

$$\mathbf{y} = \exp_\mathbf{x}(t\boldsymbol{\zeta}) \in \mathcal{Q}_\beta^{p,q}. \tag{12}$$

Descent direction: We now explain why the negative of the pseudo-Riemannian gradient is not a descent direction. Our explanation extends Chapter 3 of [20] that gives the criteria for a tangent vector $\boldsymbol{\zeta}$ to be a descent direction when the domain of the optimized function is a Euclidean space. By using the properties described in Section 3, we know that for all $t \in \mathbb{R}$ and all $\boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$, we have the equalities: $\exp_\mathbf{x}(t\boldsymbol{\xi}) = \gamma_{\mathbf{x} \to t\boldsymbol{\xi}}(1) = \gamma_{\mathbf{x} \to \boldsymbol{\xi}}(t)$ so we can equivalently fix t to 1 and choose the scale of $\boldsymbol{\xi}$ appropriately. By exploiting Taylor's first-order approximation, there exists some small enough tangent vector $\boldsymbol{\zeta} \neq \mathbf{0}$ (*i.e.* with $\exp_\mathbf{x}(\boldsymbol{\zeta})$ belonging to a convex neighborhood of $\mathbf{x}$ [4, 8]) that satisfies the following conditions: $\gamma_{\mathbf{x} \to \boldsymbol{\zeta}}(0) = \mathbf{x} \in \mathcal{Q}_\beta^{p,q}$, $\gamma'_{\mathbf{x} \to \boldsymbol{\zeta}}(0) = \boldsymbol{\zeta} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$, $\gamma_{\mathbf{x} \to \boldsymbol{\zeta}}(1) = \mathbf{y} \in \mathcal{Q}_\beta^{p,q}$, and the function $f \circ \gamma_{\mathbf{x} \to \boldsymbol{\zeta}} : \mathbb{R} \to \mathbb{R}$ can be approximated at $t = 1$ by:

$$f(\mathbf{y}) = f \circ \gamma_{\mathbf{x} \to \boldsymbol{\zeta}}(1) \simeq f \circ \gamma_{\mathbf{x} \to \boldsymbol{\zeta}}(0) + (f \circ \gamma_{\mathbf{x} \to \boldsymbol{\zeta}})'(0) = f(\mathbf{x}) + \langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q. \tag{13}$$

where we use the following properties: $\forall t, (f \circ \gamma)'(t) = \mathrm{d}f(\gamma'(t)) = g_{\gamma(t)}(Df(\gamma(t)), \gamma'(t))$ (see details in pages 11, 15 and 85 of [21]), $\mathrm{d}f$ is the differential of f and γ is a geodesic.

To be a descent direction at $\mathbf{x}$ (*i.e.* so that $f(\mathbf{y}) < f(\mathbf{x})$), the search direction $\boldsymbol{\zeta}$ has to satisfy $\langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q < 0$. However, choosing $\boldsymbol{\zeta} = -\eta Df(\mathbf{x})$, where $\eta > 0$ is a step size, might increase the function value if the scalar product $\langle \cdot, \cdot\rangle_q$ is not positive definite. If $p + q \geq 1$, then $\langle \cdot, \cdot\rangle_q$ is positive definite only if $q = 0$ (see details in supp. material), and it is negative definite iff $p = 0$ since $\langle \cdot, \cdot\rangle_q = -\langle \cdot, \cdot\rangle$ in this case. A simple solution would be to choose $\boldsymbol{\zeta} = \pm\eta Df(\mathbf{x})$ depending on the sign of $\langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q$, but $\langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q$ might be equal to 0 even if $Df(\mathbf{x}) \neq \mathbf{0}$ if $\langle \cdot, \cdot\rangle_q$ is indefinite. The optimization algorithm might then be stuck to a level set of f, which is problematic.

Algorithm 1 Pseudo-Riemannian optimization on $\mathcal{Q}_\beta^{p,q}$

---

input: differentiable function $f : \mathcal{Q}_\beta^{p,q} \to \mathbb{R}$ to be minimized, some initial value of $\mathbf{x} \in \mathcal{Q}_\beta^{p,q}$

1: **while** not converge **do**
2:      Calculate $\nabla f(\mathbf{x})$ ▷ *i.e.* the Euclidean gradient of f at $\mathbf{x}$ in the Euclidean ambient space
3:      $\boldsymbol{\chi} \leftarrow \Pi_\mathbf{x}(\mathbf{G}\Pi_\mathbf{x}(\mathbf{G}\nabla f(\mathbf{x})))$ ▷ see Eq. (14)
4:      $\mathbf{x} \leftarrow \exp_\mathbf{x}(-\eta\boldsymbol{\chi})$ ▷ where $\eta > 0$ is a step size (*e.g.* determined with line search)
5: **end while**

Proposed solution: To ensure that $\boldsymbol{\zeta} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ is a descent direction, we propose a simple expression that satisfies $\langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q < 0$ if $Df(\mathbf{x}) \neq \mathbf{0}$ and $\langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q = 0$ otherwise. We propose to formulate $\boldsymbol{\zeta} = -\eta\Pi_\mathbf{x}(\mathbf{G}Df(\mathbf{x})) \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$, and we define the following tangent vector $\boldsymbol{\chi} = -\frac{1}{\eta}\boldsymbol{\zeta}$:

$$\boldsymbol{\chi} = \Pi_\mathbf{x}(\mathbf{G}Df(\mathbf{x})) = \nabla f(\mathbf{x}) - \frac{\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle}{\langle\mathbf{x}, \mathbf{x}\rangle_q}\mathbf{G}\mathbf{x} - \frac{\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle_q}{\langle\mathbf{x}, \mathbf{x}\rangle_q}\mathbf{x} + \frac{\|\mathbf{x}\|^2\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle}{\langle\mathbf{x}, \mathbf{x}\rangle_q^2}\mathbf{x}. \quad (14)$$

The tangent vector $\boldsymbol{\zeta}$ is a descent direction because $\langle Df(\mathbf{x}), \boldsymbol{\zeta}\rangle_q = -\eta\langle Df(\mathbf{x}), \boldsymbol{\chi}\rangle_q$ is nonpositive:

$$\langle Df(\mathbf{x}), \boldsymbol{\chi}\rangle_q = \|\nabla f(\mathbf{x})\|^2 - 2\frac{\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle_q}{\langle\mathbf{x}, \mathbf{x}\rangle_q} + \frac{\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle^2\|\mathbf{x}\|^2}{\langle\mathbf{x}, \mathbf{x}\rangle_q^2} \quad (15)$$

$$= \|\mathbf{G}\nabla f(\mathbf{x}) - \frac{\langle\nabla f(\mathbf{x}), \mathbf{x}\rangle}{\langle\mathbf{x}, \mathbf{x}\rangle_q}\mathbf{x}\|^2 = \|Df(\mathbf{x})\|^2 \geq 0. \quad (16)$$

We also have $\langle Df(\mathbf{x}), \boldsymbol{\chi}\rangle_q = \|Df(\mathbf{x})\|^2 = 0$ iff $Df(\mathbf{x}) = \mathbf{0}$ (*i.e.* $\mathbf{x}$ is a stationary point). It is worth noting that $Df(\mathbf{x}) = \mathbf{0}$ implies $\boldsymbol{\chi} = \Pi_\mathbf{x}(\mathbf{G0}) = \mathbf{0}$. Moreover, $\boldsymbol{\chi} = \mathbf{0}$ implies that $\|Df(\mathbf{x})\|^2 = \langle Df(\mathbf{x}), \mathbf{0}\rangle_q = 0$. We then have $\boldsymbol{\chi} = \mathbf{0}$ iff $Df(\mathbf{x}) = \mathbf{0}$.

Our proposed algorithm to the minimization problem $\min_{\mathbf{x}\in\mathcal{Q}_\beta^{p,q}} f(\mathbf{x})$ is illustrated in Algorithm 1. Following generic Riemannian optimization algorithms [1], at each iteration, it first computes the descent direction $-\boldsymbol{\chi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$, then decreases the function by applying the exponential map defined in Eq. (6). It is worth noting that our proposed descent method can be applied to any differentiable function $f : \mathcal{Q}_\beta^{p,q} \to \mathbb{R}$, not only to those that exploit the distance introduced in Section 3.

Interestingly, our method can also be seen as a preconditioning technique [20] where the descent direction is obtained by preconditioning the pseudo-Riemannian gradient $Df(\mathbf{x})$ with the matrix $\mathbf{P}_\mathbf{x} = \left[\mathbf{G} - \frac{1}{\langle\mathbf{x},\mathbf{x}\rangle_q}\mathbf{x}\mathbf{x}^\top\right] \in \mathbb{R}^{d\times d}$. In other words, we have $\boldsymbol{\chi} = \mathbf{P}_\mathbf{x}Df(\mathbf{x}) = \Pi_\mathbf{x}(\mathbf{G}Df(\mathbf{x}))$.

In the more general setting of pseudo-Riemannian manifolds, another preconditioning technique was proposed in [8]. The method in [8] requires performing a Gram-Schmidt process at each iteration to obtain an (*ordered* [28]) orthonormal basis of the tangent space at $\mathbf{x}$ *w.r.t.* the induced quadratic form of the manifold. However, the Gram-Schmidt process is unstable and has algorithmic complexity that is cubic in the dimensionality of the tangent space. On the other hand, our method is more stable and its algorithmic complexity is linear in the dimensionality of the tangent space.

5 Experiments

We now experimentally validate our proposed optimization methods and the effectiveness of our dissimilarity function. Our main experimental results can be summarized as follows:

• Both optimizers introduced in Section 4 decrease some objective function $f : \mathcal{Q}_\beta^{p,q} \to \mathbb{R}$. While both optimizers manage to learn high-dimensional representations that satisfy the problem-dependent training constraints, only the pseudo-Riemannian optimizer satisfies all the constraints in lower-dimensional spaces. This is because it exploits the underlying metric of the manifold.

• Hyperbolic representations are popular in machine learning as they are well suited to represent hierarchical trees [10, 18, 19]. On the other hand, hierarchical datasets whose graph contains cycles cannot be represented using trees. Therefore, we propose to represent such graphs using our ultrahyperbolic representations. An important example are community graphs such as Zachary's karate club [30] that contain leaders. Because our ultrahyperbolic representations are more flexible
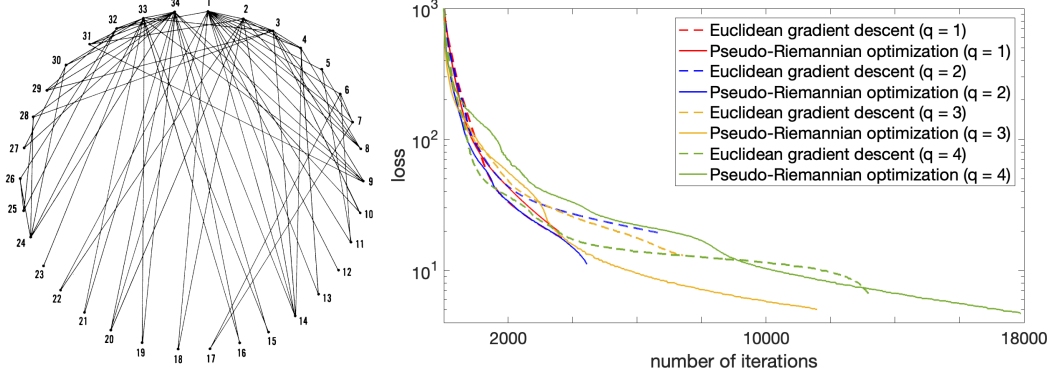
Figure 3: (left) graphic representation of Zachary's karate club (figure courtesy of [30]). (right) Loss values of Eq. (17) for different numbers of time dimensions and optimizers.

than hyperbolic representations, we believe that our representations are better suited for these non tree-like hierarchical structures.

Graph: Our ultrahyperbolic representations describe graph-structured datasets. Each dataset is an undirected weighted graph $G = (V, E)$ which has node-set $V = \{v_i\}_{i=1}^n$ and edge-set $E = \{e_k\}_{k=1}^m$. Each edge e_k is weighted by an arbitrary capacity $c_k \in \mathbb{R}_+$ that models the strength of the relationship between nodes. The higher the capacity c_k, the stronger the relationship between the nodes connected by e_k.

Learned representations: Our problem formulation is inspired by hyperbolic representation learning approaches [18, 19] where the nodes of a tree (*i.e.* graph without cycles) are represented in hyperbolic space. The hierarchical structure of the tree is then reflected by the order of distances between its nodes. More precisely, a node representation is learned so that each node is closer to its descendants and ancestors in the tree (*w.r.t.* the hyperbolic distance) than to any other node. For example, in a hierarchy of words, ancestors and descendants are hypernyms and hyponyms, respectively.

Our goal is to learn a set of n points $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{Q}_\beta^{p,q}$ (embeddings) from a given graph G. The presence of cycles in the graph makes it difficult to determine ancestors and descendants. For this reason, we introduce for each pair of nodes $(v_i, v_j) = e_k \in E$, the set of "*weaker*" pairs that have lower capacity: $\mathcal{W}(e_k) = \{e_l : c_k > c_l\} \cup \{(v_a, v_b) : (v_a, v_b) \notin E\}$. Our goal is to learn representations such that pairs (v_i, v_j) with higher capacity have their representations $(\mathbf{x}_i, \mathbf{x}_j)$ closer to each other than weaker pairs. Following [18], we formulate our problem as:

$$\min_{\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{Q}_\beta^{p,q}} \sum_{(v_i, v_j) = e_k \in E} -\log \frac{\exp\left(-\mathsf{d}(\mathbf{x}_i, \mathbf{x}_j)/\tau\right)}{\displaystyle\sum_{(v_a, v_b) \in \mathcal{W}(e_k) \cup \{e_k\}} \exp\left(-\mathsf{d}(\mathbf{x}_a, \mathbf{x}_b)/\tau\right)} \tag{17}$$

where d is the chosen dissimilarity function (*e.g.* $\mathsf{D}_\gamma(\cdot, \cdot)$ defined in Eq. (9)) and $\tau > 0$ is a fixed temperature parameter. The formulation of Eq. (17) is classic in the metric learning literature [3, 12, 27] and corresponds to optimizing some order on the learned distances via a softmax function.

Implementation details: We coded our approach in PyTorch [22] that automatically calculates the Euclidean gradient $\nabla f(\mathbf{x}_i)$. Initially, a random set of vectors $\{\mathbf{z}_i\}_{i=1}^n$ is generated close to the positive pole $(\sqrt{|\beta|}, 0, \cdots, 0) \in \mathcal{Q}_\beta^{p,q}$ with every coordinate perturbed uniformly with a random value in the interval $[-\varepsilon, \varepsilon]$ where $\varepsilon > 0$ is chosen small enough so that $\|\mathbf{z}_i\|_q^2 < 0$. We set $\beta = -1$, $\varepsilon = 0.1$ and $\tau = 10^{-2}$. Initial embeddings are generated as follows: $\forall i, \mathbf{x}_i = \sqrt{|\beta|} \frac{\mathbf{z}_i}{\sqrt{|\|\mathbf{z}_i\|_q^2|}} \in \mathcal{Q}_\beta^{p,q}$.

Zachary's karate club dataset [30] is a social network graph of a karate club comprised of $n = 34$ nodes, each representing a member of the karate club. The club was split due to a conflict between instructor "Mr. Hi" (node v_1) and administrator "John A" (node v_n). The remaining members now have to decide whether to join the new club created by v_1 or not. In [30], Zachary defines a matrix of relative strengths of the friendships in the karate club called $\mathbf{C} \in \{0, 1, \cdots, 7\}^{n \times n}$ and that depends on various criteria. We note that the matrix is not symmetric and has 7 different pairs (v_i, v_j) for which $\mathbf{C}_{ij} \neq \mathbf{C}_{ji}$. Since our dissimilarity function is symmetric, we consider the symmetric matrix

Table 1: Evaluation scores for the different learned representations (mean $\pm$ standard deviation)

Evaluation metric	$\mathbb{R}^4$ (flat)	$\mathcal{Q}_{-1}^{4,0}$ (hyperbolic)	$\mathcal{Q}_{-1}^{3,1}$ (ours)	$\mathcal{Q}_{-1}^{2,2}$ (ours)	$\mathcal{Q}_{-1}^{1,3}$ (ours)	$\mathcal{Q}_{-1}^{0,4}$ (spherical)
Rank of the first leader	5.4 ± 1.1	2.8 ± 0.4	$\mathbf{1.2 \pm 0.4}$	$\mathbf{1.2 \pm 0.4}$	1.8 ± 0.8	2.0 ± 0.7
Rank of the second leader	6.6 ± 0.9	4.2 ± 0.7	$\mathbf{2.4 \pm 0.9}$	2.6 ± 0.5	4.0 ± 1.2	4.0 ± 1.4
top 5 Spearman's ρ	-0.44 ± 0.19	0.20 ± 0.48	$\mathbf{0.76 \pm 0.21}$	0.66 ± 0.30	0.36 ± 0.40	0.18 ± 0.37
top 10 Spearman's ρ	0.00 ± 0.14	0.38 ± 0.06	0.74 ± 0.11	$\mathbf{0.79 \pm 0.12}$	0.71 ± 0.08	0.55 ± 0.20

$\mathbf{S} = \mathbf{C} + \mathbf{C}^\top$ instead. The value of $\mathbf{S}_{ij}$ is the capacity/weight assigned to the edge joining v_i and v_j, and there is no edge between v_i and v_j if $\mathbf{S}_{ij} = 0$. Fig. 3 (left) illustrates the 34 nodes of the dataset, an edge joining the nodes v_i and v_j is drawn iff $\mathbf{S}_{ij} \neq 0$. The level of a node in the hierarchy corresponds approximately to its height in the figure.

Optimizers: We validate that our optimizers introduced in Section 4 decrease the cost function. First, we consider the simple unweighted case where every edge weight is 1. For each edge $e_k \in E$, $\mathcal{W}(e_k)$ is then the set of pairs of nodes that are not connected. In other words, Eq. (17) learns node representations that have the property that every connected pair of nodes has smaller distance than non-connected pairs. We use this condition as a stopping criterion of our algorithm.

Fig. 3 (right) illustrates the loss values of Eq. (17) as a function of the number of iterations with the Euclidean gradient descent (Section 4.1) and our pseudo-Riemannian optimizer (introduced in Section 4.2). In each test, we vary the number of time dimensions $q + 1$ while the ambient space is of fixed dimensionality $d = p + q + 1 = 10$. We omit the case $q = 0$ since it corresponds to the (hyperbolic) Riemannian case already considered in [11, 19]. Both optimizers decrease the function and manage to satisfy all the expected distance relations. We note that when we use $-Df(\mathbf{x})$ instead of $-\chi$ as a search direction, the algorithm does not converge. Moreover, our pseudo-Riemannian optimizer manages to learn representations that satisfy all the constraints for low-dimensional manifolds such as $\mathcal{Q}_{-1}^{4,1}$ and $\mathcal{Q}_{-1}^{4,2}$, while the optimizer introduced in Section 4.1 does not. Consequently, we only use the pseudo-Riemannian optimizer in the following results.

Hierarchy extraction: To quantitatively evaluate our approach, we apply it to the problem of predicting the high-level nodes in the hierarchy from the weighted matrix $\mathbf{S}$ given as supervision. We consider the challenging low-dimensional setting where all the learned representations lie on a 4-dimensional manifold (*i.e.* $p + q + 1 = 5$). Hyperbolic distances are known to grow exponentially as we get further from the origin. Therefore, the sum of distances $\delta_i = \sum_{j=1}^n \mathsf{d}(\mathbf{x}_i, \mathbf{x}_j)$ of a node v_i with all other nodes is a good indication of importance. Intuitively, high-level nodes will be closer to most nodes than low-level nodes. We then sort the scores $\delta_1, \cdots, \delta_n$ in ascending order and report the ranks of the two leaders v_1 or v_n (in no particular order) in the first two rows of Table 1 averaged over 5 different initializations/runs. Leaders tend to have a smaller δ_i score with ultrahyperbolic distances than with Euclidean, hyperbolic or spherical distances.

Instead of using δ_i for hyperbolic representations, the importance of a node v_i can be evaluated by using the Euclidean norm of its embedding $\mathbf{x}_i$ as proxy [11, 18, 19]. This is because high-level nodes of a tree in hyperbolic space are usually closer to the origin than low-level nodes. Not surprisingly, this proxy leads to worse performance (8.6 ± 2.3 and 18.6 ± 4.9) as the relationships are not that of a tree. Since hierarchy levels are hard to compare for low-level nodes, we select the 10 (or 5) most influential members based on the score $s_i = \sum_{j=1}^n \mathbf{S}_{ij}$. The corresponding nodes are 34, 1, 33, 3, 2, 32, 24, 4, 9, 14 (in that order). Spearman's rank correlation coefficient [24] between the selected scores s_i and corresponding δ_i is reported in Table 1 and shows the relevance of our representations.

Due to lack of space, we also report in the supp. material similar experiments on a larger hierarchical dataset [9] that describes co-authorship from papers published at NIPS from 1988 to 2003.

6 Conclusion

We have introduced ultrahyperbolic representations. Our representations lie on a pseudo-Riemannian manifold of constant nonzero curvature which generalizes hyperbolic and spherical geometries and includes them as submanifolds. Any relationship described in those geometries can then be described with our representations that are more flexible. We have introduced new optimization tools and experimentally shown that our representations can extract hierarchies in graphs that contain cycles.

Broader Impact

We introduce a novel way of representing relationships between data points by considering the geometry of non-Riemannian manifolds of constant nonzero curvature. The relationships between data points are described by a dissimilarity function that we introduce and exploits the structure of the manifold. It is more flexible than the distance metric used in hyperbolic and spherical geometries often used in machine learning and computer vision. Nonetheless, since the problems involving our representations are not straightforward to optimize, we propose novel optimization algorithms that can potentially benefit the machine learning, computer vision and natural language processing communities. Indeed, our method is application agnostic and could extend existing frameworks.

Our contribution is mainly theoretical but we have included one practical application. Similarly to hyperbolic representations that are popular for representing tree-like data, we have shown that our representations are well adapted to the more general case of hierarchical graphs with cycles. These graphs appear in many different fields of research such as medicine, molecular biology and the social sciences. For example, an ultrahyperbolic representation of proteins might assist in understanding their complicated folding mechanisms. Moreover, these representations could assist in analyzing features of social media such as discovering new trends and leading "connectors". The impact of community detection for commercial or political advertising is already known in social networking services. We foresee that our method will have many more graph-based practical applications.

We know of very few applications outside of general relativity that use pseudo-Riemannian geometry. We hope that our research will stimulate other applications in machine learning and related fields. Finally, although we have introduced a novel descent direction for our optimization algorithm, future research could study and improve its rate of convergence.

Acknowledgments and Disclosure of Funding

References

[1] Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Henri Anciaux. *Minimal submanifolds in pseudo-Riemannian geometry*. World Scientific, 2011.

[3] Tianshi Cao, Marc T. Law, and Sanja Fidler. A theoretical analysis of the number of shots in few-shot learning. In *International Conference on Learning Representations*, 2020.

[4] Manfredo Perdigao do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

[5] Wei Chen, Wenjie Fang, Guangda Hu, and Michael W. Mahoney. On the hyperbolicity of small-world and treelike random graphs. *Internet Mathematics*, 9(4):434–491, 2013.

[6] Aasa Feragen, Søren Hauberg, Mads Nielsen, and François Lauze. Means in spaces of tree-like shapes. In *2011 International Conference on Computer Vision*, pages 736–746. IEEE, 2011.

[7] Aasa Feragen, Pechin Lo, Marleen de Bruijne, Mads Nielsen, and François Lauze. Toward a theory of statistical tree-shape analysis. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):2008–2021, 2012.

[8] Tingran Gao, Lek-Heng Lim, and Ke Ye. Semi-riemannian manifold optimization. *arXiv preprint arXiv:1812.07643*, 2018.

[9] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean Embedding of Co-occurrence Data. *The Journal of Machine Learning Research*, 8:2265–2295, 2007.

[10] Mikhael Gromov. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer, 1987.

[11] Marc T. Law, Renjie Liao, Jake Snell, and Richard S. Zemel. Lorentzian distance learning for hyperbolic representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3672–3681, 2019.

[12] Marc T. Law, Jake Snell, Amir massoud Farahmand, Raquel Urtasun, and Richard S Zemel. Dimensionality reduction for representing the knowledge of probabilistic models. In *International Conference on Learning Representations*, 2019.

[13] Marc T. Law, Yaoliang Yu, Matthieu Cord, and Eric P. Xing. Closed-form training of mahalanobis distance for supervised clustering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3909–3917, 2016.

[14] Marc T. Law, Yaoliang Yu, Raquel Urtasun, Richard S. Zemel, and Eric P. Xing. Efficient multiple instance metric learning using weakly supervised data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5948–5956, 2017.

[15] John M. Lee. *Introduction to topological manifolds*, volume 202. Springer Science & Business Media, 2010.

[16] John M. Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2013.

[17] Ernest Lindelöf. Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 116(3):454–457, 1894.

[18] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017.

[19] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788, 2018.

[20] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[21] Barrett O'neill. *Semi-Riemannian geometry with applications to relativity*, volume 103. Academic press, 1983.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[23] Peter Petersen. *Riemannian geometry*, volume 171. Springer, 2006.

[24] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 1904.

[25] Makarand Tapaswi, Marc T. Law, and Sanja Fidler. Video face clustering with unknown number of clusters. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[26] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[27] Jixuan Wang, Kuan-Chieh Wang, Marc T. Law, Frank Rudzicz, and Michael Brudno. Centroid-based deep metric learning for speaker recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3652–3656, 2019.

[28] Joseph Albert Wolf. *Spaces of constant curvature*, volume 372. American Mathematical Soc., 1972.

[29] Eric P. Xing, Michael I. Jordan, Stuart J. Russell, and Andrew Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

[30] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

A Supplementary material of "Ultrahyperbolic Representation Learning"

The supplementary material is structured as follows:

• In Section B, we provide a short discussion on the choice of geometry to represent graphs.

• In Section C.1, we study the formulation of the geodesics in Eq. (5). It is combination of hyperbolic, spherical and flat space due to its formulation.

• In Section C.2, we explain why $\log_{\mathbf{x}}(\mathbf{y})$ (see Eq. (7)) is not defined if $\langle \mathbf{x}, \mathbf{y} \rangle_q \geq |\beta|$.

• In Section C.3, we explain the anti-isometry between $\mathcal{Q}_{\beta}^{p,q}$ or $\mathcal{Q}_{-\beta}^{q+1,p-1}$.

• In Section C.4, we study the curvature of $\mathcal{Q}_{\beta}^{p,q}$.

• In Section C.5, we give the proof of Theorem 4.1.

• In Section C.6, we explain why the upper sheet of the two-sheet hyperboloid $\mathcal{Q}_{\beta}^{p,0}$ (*i.e.* the case where $q = 0$) is a Riemannian manifold even if its metric matrix $\mathbf{G} = \mathbf{I}_{1,p}$ is not positive definite.

• In Section D, we report additional experiments.

B Choice of Geometry

The choice of geometry to represent graphs is still an open problem in general. It depends on the topology of the graph and the kind of relationship between nodes. For instance, hyperbolic geometry was mathematically shown to be appropriate for tree-like graphs [10], but not for other types of graphs. Ultrahyperbolic geometry has the advantage of generalizing both hyperbolic and spherical geometries and can describe relationships specific to those geometries. In particular, the geodesic *distance* can be written in the same way as the Poincaré and spherical distances as shown in Eq. (8), Some parts of the manifold are hyperbolic or spherical as explained in the paper. The converse is not true. The framework might then automatically learn representations to be part of a same hyperbolic or spherical part of the manifold depending on the context.

Those reasons led us to consider hierarchical graphs that were similar to trees, but where the presence of cycles in the graph limited the relevance of hyperbolic geometry. We experimentally validated our intuition. The choice of manifold of constant nonzero curvature (i.e. the optimal number of time and space dimensions $q + 1$ and p) then seems to depend on how much the graph is similar to a tree or to a graph where spherical geometry is appropriate, such as cycle graphs (or a mix of both).

It is also worth noting that ultrahyperbolic geometry can describe some graph concepts differently, if not better, than hyperbolic and spherical geometries. For instance, it is known that *triadic closure* is a concept in social network theory that is not valid for most large and complex networks. In other words, if $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}, \mathbf{z})$ are strongly tied, triadic closure would imply that $(\mathbf{y}, \mathbf{z})$ are strongly tied as well. As explained in Section 3, the fact that we can find triplets that satisfy $d_\gamma(\mathbf{x}, \mathbf{y}) = d_\gamma(\mathbf{x}, \mathbf{z}) = 0$ but $d_\gamma(\mathbf{y}, \mathbf{z}) > 0$ allows to avoid triadic closure. Moreover, although we have applied our proposed representations to some type of graph, they can be applied to other applications that do not involve graphs. The possible applications are left for future research.

C Some properties about geodesics and logarithm map

C.1 Geodesics of $\mathcal{Q}_{\beta}^{p,q}$

Tangent space of pseudo-Riemannian submanifolds: In the paper, we exploit the fact that our considered manifold $\mathcal{M}$ (here $\mathcal{Q}_{\beta}^{p,q}$) is a pseudo-Riemannian submanifold of $\overline{\mathcal{M}}$ (here $\mathbb{R}^{p,q+1}$). Since $\overline{\mathcal{M}}$ is chosen to be a vector space, we have a natural isomorphism between $\overline{\mathcal{M}}$ and its tangent space.

If $\mathcal{M}$ is a pseudo-Riemannian submanifold of $\overline{\mathcal{M}}$, we have the following direct sum decomposition:

$$T_{\mathbf{x}}(\overline{\mathcal{M}}) = T_{\mathbf{x}}(\mathcal{M}) + T_{\mathbf{x}}(\mathcal{M})^{\perp}, \tag{18}$$

where $T_{\mathbf{x}}(\mathcal{M})^{\perp} = \{\boldsymbol{\zeta} \in T_{\mathbf{x}}(\overline{\mathcal{M}}) : \forall \boldsymbol{\xi} \in T_{\mathbf{x}}(\mathcal{M}), \overline{g_{\mathbf{x}}}(\boldsymbol{\zeta}, \boldsymbol{\xi}) = 0\}$ is the orthogonal complement of $T_{\mathbf{x}}(\mathcal{M})$ and called the *normal space* of $\mathcal{M}$ at $\mathbf{x}$. It is a nondegenerate subspace of $T_{\mathbf{x}}(\overline{\mathcal{M}})$, and $\overline{g_{\mathbf{x}}}$ is

the metric at $\mathbf{x} \in \overline{\mathcal{M}}$. In the case of $\mathcal{Q}_\beta^{p,q}$, $T_\mathbf{x}(\mathcal{Q}_\beta^{p,q})^\perp$ is defined as:

$$T_\mathbf{x}(\mathcal{Q}_\beta^{p,q})^\perp = \{\boldsymbol{\zeta} \in T_\mathbf{x}\mathbb{R}^{p,q+1} : \forall \boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}, \ \langle \boldsymbol{\zeta}, \boldsymbol{\xi} \rangle_q = 0\} \quad (19)$$

$$\approx N_\mathbf{x}(\mathcal{Q}_\beta^{p,q}, \mathbb{R}^{p,q+1}) = \{\lambda\mathbf{x} \in \mathbb{R}^{p,q+1} : \lambda \in \mathbb{R}\}. \quad (20)$$

where $\approx$ denotes isomorphism.

Geodesic of a submanifold: As mentioned in the main paper, a curve γ is a geodesic if its acceleration is zero. However, the acceleration depends on the choice of the affine connection while the velocity does not (*i.e.* the velocity does not depend on the Christoffel symbols whereas the acceleration does, and different connections produce different geodesics, see details in page 66 of [21] or Chapter 5.4 of [1]). Let us note $\frac{\overline{\mathrm{D}}}{dt}(\gamma'(t))$ (resp. $\frac{\mathrm{D}}{dt}(\gamma'(t))$) the covariant derivative of $\gamma'(t)$ along $\gamma(t)$ in $\mathbb{R}^{p,q+1}$ (resp. in $\mathcal{Q}_\beta^{p,q}$). By using the *induced connection* (see page 98 of [21]) and the fact that $\mathcal{Q}_\beta^{p,q}$ is isometrically embedded in $\mathbb{R}^{p,q+1}$, the second-order ordinary differential equation about the zero acceleration of the geodesic is equivalent to (see page 103 of [21]):

$$\frac{\overline{\mathrm{D}}}{dt}(\gamma'(t)) \in N_{\gamma(t)}(\mathcal{Q}_\beta^{p,q}, \mathbb{R}^{p,q+1}) \iff \gamma''(t) = \frac{\mathrm{D}}{dt}(\gamma'(t)) = \Pi_{\gamma(t)}\left(\frac{\overline{\mathrm{D}}}{dt}(\gamma'(t))\right) = \mathbf{0}. \quad (21)$$

where $\Pi_{\gamma(t)}$ is defined in Eq. (3) and orthogonally projects onto $T_{\gamma(t)}\mathcal{Q}_\beta^{p,q}$. In other words, γ is straight in $\mathcal{M}$ but its curving in $\overline{\mathcal{M}}$ is the one forced by the curving of $\mathcal{M}$ itself in $\overline{\mathcal{M}}$ [21].

The initial conditions $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}(0) = \mathbf{x} \in \mathcal{Q}_\beta^{p,q}$ and $\gamma'_{\mathbf{x}\to\boldsymbol{\xi}}(0) = \boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ are straightforward. With the formulation of $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}$ in Eq. (5), one can first verify that for all t, $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}(t)$ lies on $\mathcal{Q}_\beta^{p,q}$. Indeed, since $\langle \mathbf{x}, \boldsymbol{\xi} \rangle_q = 0$, we have:

$$\forall t \in \mathbb{R}, \ \forall \mathbf{x} \in \mathcal{Q}_\beta^{p,q}, \ \forall \boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}, \ \langle \gamma_{\mathbf{x}\to\boldsymbol{\xi}}(t), \gamma_{\mathbf{x}\to\boldsymbol{\xi}}(t) \rangle_q = \beta. \quad (22)$$

The acceleration of γ at t (in the ambient space $\mathbb{R}^{p,q+1}$) is in $N_{\gamma(t)}(\mathcal{Q}_\beta^{p,q}, \mathbb{R}^{p,q+1})$ since we have $\forall \mathbf{x} \in \mathcal{Q}_\beta^{p,q}, \ \forall \boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$:

$$\forall t \in \mathbb{R}, \ \frac{\overline{\mathrm{D}}}{dt}(\gamma'_{\mathbf{x}\to\boldsymbol{\xi}}(t)) = \frac{\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q}{|\beta|}\gamma_{\mathbf{x}\to\boldsymbol{\xi}}(t) \in N_{\gamma_{\mathbf{x}\to\boldsymbol{\xi}}(t)}(\mathcal{Q}_\beta^{p,q}, \mathbb{R}^{p,q+1}) \quad (23)$$

$$\implies \forall t \in \mathbb{R}, \ \gamma''_{\mathbf{x}\to\boldsymbol{\xi}}(t) = \frac{\mathrm{D}}{dt}(\gamma'_{\mathbf{x}\to\boldsymbol{\xi}}(t)) = \mathbf{0}. \quad (24)$$

We have found a solution for our second-order ordinary differential equation, the solution is unique by definition. $\qquad\square$

From its formulation in Eq. (5), the nonconstant geodesic $\gamma_{\mathbf{x}\to\boldsymbol{\xi}}$ is similar to the hyperbolic, flat or spherical case if $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q$ is positive, zero or negative, respectively.

C.2 On the non existence of the logarithm map of some points

We explain here why $\log_\mathbf{x}(\mathbf{y})$ is not defined if $\langle \mathbf{x}, \mathbf{y} \rangle_q \geq |\beta|$. Our proof relies on the fact the domain of the exponential map is the whole tangent space. We also use the fact that $\forall \boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$, we have $\langle \mathbf{x}, \boldsymbol{\xi} \rangle_q = 0$ by definition of tangent vectors to $\mathcal{Q}_\beta^{p,q}$.

Assume that there exists a tangent vector $\boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ such that $\mathbf{y} = \exp_\mathbf{x}(\boldsymbol{\xi})$ and $\langle \mathbf{y}, \mathbf{x} \rangle_q \geq |\beta|$. We consider the three possible cases of sign of $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q$.

● Let us first assume that $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q > 0$. For all $\boldsymbol{\xi} \in T_\mathbf{x}\mathcal{Q}_\beta^{p,q}$ that satisfies $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q > 0$, we have by definition of the corresponding exponential map:

$$\langle \mathbf{x}, \exp_\mathbf{x}(\boldsymbol{\xi}) \rangle_q = \|\mathbf{x}\|_q^2 \cosh\left(\frac{\sqrt{|\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q|}}{\sqrt{|\beta|}}\right) + \sqrt{|\beta|}\frac{\langle \mathbf{x}, \boldsymbol{\xi} \rangle_q}{\sqrt{|\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q|}} \sinh\left(\frac{\sqrt{|\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q|}}{\sqrt{|\beta|}}\right) \quad (25)$$

$$= \|\mathbf{x}\|_q^2 \cosh\left(\frac{\sqrt{|\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q|}}{\sqrt{|\beta|}}\right) < 0. \quad (26)$$

Therefore, there exists no tangent vector $\boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{Q}_\beta^{p,q}$ that satisfies $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q > 0$ and $\langle\mathbf{x}, \exp_{\mathbf{x}}(\boldsymbol{\xi})\rangle_q > |\beta| > 0$.

• Let us now assume that $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q = 0$. For all $\boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{Q}_\beta^{p,q}$ that satisfies $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q = 0$, we have:

$$\langle\mathbf{x}, \exp_{\mathbf{x}}(\boldsymbol{\xi})\rangle_q = \langle\mathbf{x}, \mathbf{x} + \boldsymbol{\xi}\rangle_q = \langle\mathbf{x}, \mathbf{x}\rangle_q + \langle\mathbf{x}, \boldsymbol{\xi}\rangle_q = \langle\mathbf{x}, \mathbf{x}\rangle_q = \beta < 0. \tag{27}$$

• Let us now assume that $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q < 0$. For all $\boldsymbol{\xi} \in T_{\mathbf{x}}\mathcal{Q}_\beta^{p,q}$ that satisfies $\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q < 0$, we have:

$$\langle\mathbf{x}, \exp_{\mathbf{x}}(\boldsymbol{\xi})\rangle_q = \|\mathbf{x}\|_q^2 \cos\left(\frac{\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right) + \sqrt{|\beta|}\frac{\langle\mathbf{x}, \boldsymbol{\xi}\rangle_q}{\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}} \sin\left(\frac{\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right) \tag{28}$$

$$= \|\mathbf{x}\|_q^2 \cos\left(\frac{\sqrt{|\langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_q|}}{\sqrt{|\beta|}}\right) \in [\beta, |\beta|]. \tag{29}$$

Given the formulation of the geodesic in Eq. (5), $\langle\mathbf{x}, \mathbf{y}\rangle_q = \langle\mathbf{x}, \exp_{\mathbf{x}}(\boldsymbol{\xi})\rangle_q \geq |\beta|$ iff $\mathbf{y} = -\mathbf{x}$.

From our study above, there exists no geodesic (hence no logarithm map) joining $\mathbf{x}$ and $\mathbf{y}$ if $\langle\mathbf{x}, \mathbf{y}\rangle_q \geq |\beta|$ except if $\mathbf{y} = -\mathbf{x}$. The antipodal point $\mathbf{y} = -\mathbf{x}$ is also a special case for which there does not exist a logarithm map since $\mathbf{x}$ and $-\mathbf{x}$ are joined by infinitely many minimizing geodesics of equal length (a similar case is the q-sphere). Nonetheless, the geodesic "distance" between $\mathbf{x}$ and $-\mathbf{x}$ can be defined as $\pi\sqrt{|\beta|}$. $\qquad\square$

C.3 Anti-isometry between $\mathcal{Q}_\beta^{p,q}$ and $\mathcal{Q}_{-\beta}^{q+1,p-1}$

In the main paper, we state that there is an anti-isometry between $\mathcal{Q}_\beta^{p,q}$ and $\mathcal{Q}_{-\beta}^{q+1,p-1}$. We recall that $\mathcal{Q}_{-\beta}^{q+1,p-1}$ is embedded in $\mathbb{R}^{q+1,p}$. Let us note $\sigma : \mathbb{R}^{p,q+1} \to \mathbb{R}^{q+1,p}$ the mapping defined as:

$$\forall\mathbf{x} = (x_0, \cdots, x_{p+q})^\top \in \mathbb{R}^{p,q+1}, \ \sigma(\mathbf{x}) = (x_{p+q}, x_{p+q-1}, \cdots, x_1, x_0)^\top. \tag{30}$$

We have:
$$\forall\mathbf{x} \in \mathbb{R}^{p,q+1}, \mathbf{y} \in \mathbb{R}^{p,q+1}, \ \langle\mathbf{x}, \mathbf{y}\rangle_q = -\langle\sigma(\mathbf{x}), \sigma(\mathbf{y})\rangle_{p-1}. \tag{31}$$

C.4 Curvature of $\mathcal{Q}_\beta^{p,q}$

The manifold $\mathcal{Q}_\beta^{p,q}$ is a total umbilic hypersurface of $\mathbb{R}^{p,q+1}$ and has constant sectional curvature $1/\beta$ and constant mean curvature $\kappa = |\beta|^{-1/2}$ with respect to the unit normal vector field $\mathcal{N}(\mathbf{x}) = -\kappa\mathbf{x}$. More details can be found in Chapter 3 of [2].

C.5 Proof of Theorem 4.1

We give the proofs for Theorem 4.1 that we recall below:

Theorem C.1 (Diffeomorphisms). *For any $\beta < 0$, there is a diffeomorphism $\psi : \mathcal{Q}_\beta^{p,q} \to \mathcal{S}^q \times \mathbb{R}^p$. Let us note $\mathbf{x} = \begin{pmatrix}\mathbf{t} \\ \mathbf{s}\end{pmatrix} \in \mathcal{Q}_\beta^{p,q}$ with $\mathbf{t} \in \mathbb{R}_*^{q+1}$ and $\mathbf{s} \in \mathbb{R}^p$, let us note $\mathbf{z} = \begin{pmatrix}\mathbf{u} \\ \mathbf{v}\end{pmatrix} \in \mathcal{S}^q \times \mathbb{R}^p$ where $\mathbf{u} \in \mathcal{S}^q$ and $\mathbf{v} \in \mathbb{R}^p$. The mapping ψ and its inverse ψ^{-1} are formulated:*

$$\psi(\mathbf{x}) = \begin{pmatrix}\frac{1}{\|\mathbf{t}\|}\mathbf{t} \\ \frac{1}{\sqrt{|\beta|}}\mathbf{s}\end{pmatrix} \qquad and \qquad \psi^{-1}(\mathbf{z}) = \sqrt{|\beta|}\begin{pmatrix}\sqrt{1 + \|\mathbf{v}\|^2}\mathbf{u} \\ \mathbf{v}\end{pmatrix}. \tag{32}$$

We need to show that $\psi(\psi^{-1}(\mathbf{z})) = \mathbf{z}$ and $\psi^{-1}(\psi(\mathbf{x})) = \mathbf{x}$.

Space dimensions: The mapping of the space dimensions of $\mathbf{x}$ to the space dimensions of $\mathbf{z}$ simply involves a scaling factor of $\sqrt{|\beta|}$ (*i.e.* $\mathbf{v} = \frac{1}{\sqrt{|\beta|}}\mathbf{s}$, and $\mathbf{s} = \sqrt{|\beta|}\mathbf{v}$).

Time dimensions: We show the invertibility of the mappings taking time dimensions as input.

• Let us first show $\psi^{-1}(\psi(\mathbf{x})) = \mathbf{x}$. We recall that $\mathbf{x} = \begin{pmatrix}\mathbf{t} \\ \mathbf{s}\end{pmatrix} \in \mathcal{Q}_\beta^{p,q}$ with $\mathbf{t} \in \mathbb{R}_*^{q+1}$ and $\mathbf{s} \in \mathbb{R}^p$.

We need to show that:

$$\mathbf{t} = \sqrt{|\beta|}\sqrt{1 + \|\mathbf{v}\|^2}\frac{\mathbf{t}}{\|\mathbf{t}\|} = \sqrt{|\beta|}\sqrt{1 + \|\frac{1}{\sqrt{|\beta|}}\mathbf{s}\|^2}\frac{\mathbf{t}}{\|\mathbf{t}\|} = \sqrt{|\beta|}\sqrt{1 + \frac{1}{|\beta|}\|\mathbf{s}\|^2}\frac{\mathbf{t}}{\|\mathbf{t}\|}. \tag{33}$$

To show Eq. (33), it is sufficient to prove that the following property is satisfied:

$$\|\mathbf{t}\| = \sqrt{|\beta|}\sqrt{1 + \frac{1}{|\beta|}\|\mathbf{s}\|^2}, \text{which is satisfied if } \|\mathbf{t}\|^2 = |\beta|(1 + \frac{1}{|\beta|}\|\mathbf{s}\|^2) = |\beta| + \|\mathbf{s}\|^2. \tag{34}$$

Since $\mathbf{x} \in \mathcal{Q}_\beta^{p,q}$, we have by definition $\|\mathbf{x}\|_q^2 = \|\mathbf{s}\|^2 - \|\mathbf{t}\|^2 = \beta < 0$. Therefore, we have:

$$\|\mathbf{t}\|^2 = \|\mathbf{s}\|^2 - \beta = \|\mathbf{s}\|^2 + |\beta|. \tag{35}$$

• Let us now show $\psi(\psi^{-1}(\mathbf{z})) = \mathbf{z}$. We recall that $\mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \in \mathcal{S}^q \times \mathbb{R}^p$ where $\mathbf{u} \in \mathcal{S}^q$ and $\mathbf{s} \in \mathbb{R}^p$.

We need to show:

$$\mathbf{u} = \frac{\sqrt{|\beta|}\sqrt{1 + \|\mathbf{v}\|^2}\mathbf{u}}{\|\sqrt{|\beta|}\sqrt{1 + \|\mathbf{v}\|^2}\mathbf{u}\|} = \frac{\mathbf{u}}{\|\mathbf{u}\|}. \tag{36}$$

By definition of $\mathcal{S}^q$, $\mathbf{u}$ satisfies Eq. (36). □

C.6 The metric tensor is positive definite only if $q = 0$ and indefinite in the general case

We show here that the hyperboloid (*i.e.* upper sheet of the two-sheet hyperboloid $\mathcal{Q}_\beta^{p,0}$ with $p \geq 1$) is a Riemannian manifold even if its symmetric metric matrix $\mathbf{G} = \mathbf{I}_{1,p}$ is not positive definite. That result is already known in the literature and a proof can be found in [23]. We still give it because we think it is important.

We recall that a Riemannian manifold is a pseudo-Riemannian manifold with positive definite metric tensor. In other words, for all $\mathbf{x} \in \mathcal{M}$ where $\mathcal{M}$ is a pseudo-Riemannian manifold, $\mathcal{M}$ is Riemannian iff $\forall \boldsymbol{\xi} \in T_\mathbf{x}\mathcal{M}$, $g_\mathbf{x}(\boldsymbol{\xi}, \boldsymbol{\xi}) > 0$ iff $\boldsymbol{\xi} \neq 0$.

We recall that we consider that $\beta < 0$, the usual definition of the hyperboloid $\mathcal{H}_\beta^p$ in the literature is:

$$\mathcal{H}_\beta^p = \left\{ \mathbf{x} = (x_0, x_1, \cdots, x_p)^\top \in \mathcal{Q}_\beta^{p,0} : x_0 \geq 0 \right\}, \tag{37}$$

Let us note $\mathbf{x} = (x_0, x_1, \cdots, x_p)^\top \in \mathcal{H}_\beta^p$ and $\boldsymbol{\xi} = (\xi_0, \cdots, \xi_p)^\top$ a tangent vector at $\mathbf{x}$. For simplicity, we will denote the vectors $\mathbf{u} = (x_1, \cdots, x_p)^\top \in \mathbb{R}^p$ and $\mathbf{v} = (\xi_1, \cdots, \xi_p)^\top \in \mathbb{R}^p$ so that $\mathbf{x} = \begin{pmatrix} x_0 \\ \mathbf{u} \end{pmatrix} \in \mathcal{H}_\beta^p$ and $\boldsymbol{\xi} = \begin{pmatrix} \xi_0 \\ \mathbf{v} \end{pmatrix}$. By definition of $\mathcal{H}_\beta^p$, we have $x_0 = \sqrt{-\beta + \|\mathbf{u}\|^2} > 0$.

Moreover, by definition of the tangent space, and since $\mathcal{H}_\beta^p$ is a subset of $\mathcal{Q}_\beta^{p,0}$, any tangent vector $\boldsymbol{\xi}$ satisfies $0 = \langle \mathbf{x}, \boldsymbol{\xi} \rangle_q = -x_0\xi_0 + \langle \mathbf{u}, \mathbf{v} \rangle$ where $q = 0$. This implies:

$$\xi_0 = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{x_0} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\sqrt{-\beta + \|\mathbf{u}\|^2}} \implies \xi_0^2 = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{-\beta + \|\mathbf{u}\|^2}. \tag{38}$$

It is obvious that if $\boldsymbol{\xi} = 0$, then $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q = 0$. To show that $\mathcal{H}_\beta^p$ is a Riemannian manifold, we need to show that every non-vanishing tangent vector $\boldsymbol{\xi}$ satisfies $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q > 0$. By definition of $\langle \cdot, \cdot \rangle_q$, we have

$$\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q = -\xi_0^2 + \|\mathbf{v}\|^2. \tag{39}$$

We then need to show that $\xi_0^2 < \|\mathbf{v}\|^2$. By the Cauchy-Schwarz inequality, we have $\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \|\mathbf{v}\|^2\|\mathbf{u}\|^2 \leq \|\mathbf{v}\|^2(-\beta + \|\mathbf{u}\|^2)$. This implies from Eq. (38):

$$\xi_0^2 = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{-\beta + \|\mathbf{u}\|^2} \leq \|\mathbf{v}\|^2. \tag{40}$$

By using the Cauchy-Schwarz inequality, Eq. (40) can be an equality iff $\boldsymbol{\xi} = \mathbf{0}$ since $\|\mathbf{u}\|^2 < -\beta + \|\mathbf{u}\|^2$. Since every non-vanishing tangent vector $\boldsymbol{\xi}$ to $\mathcal{H}_\beta^p$ satisfies $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q > 0$, the metric tensor of $\mathcal{H}_\beta^p$ is positive definite. This shows that $\mathcal{H}_\beta^p$ is Riemannian. $\hfill\square$

The fact that we consider $\beta < 0$ is important. For instance, if $\beta = 1$, then let us note $\mathbf{x} = (0, 1, 0)^\top \in \mathcal{Q}_1^{2,0}$. If $\boldsymbol{\xi} = (1, 0, 0)^\top \in T_\mathbf{x} \mathcal{Q}_1^{2,0}$, then we have $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q < 0$. If $\boldsymbol{\xi} = (0, 0, 1)^\top \in T_\mathbf{x} \mathcal{Q}_1^{2,0}$, then we have $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q > 0$. Finally, if $\boldsymbol{\xi} = (1, 0, 1)^\top \in T_\mathbf{x} \mathcal{Q}_1^{2,0}$, then we have $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q = 0$. The metric tensor is then indefinite if $\beta = 1$, and $\mathcal{Q}_1^{2,0}$ is non-Riemannian. That result can be extended to any positive value of β by considering instead $\mathbf{x} = (0, \sqrt{\beta}, 0)^\top$.

Indefiniteness: If $\beta < 0$, one can show in a similar way that $\langle \cdot, \cdot \rangle_q$ is indefinite if $q \geq 1$ and $p \geq 1$. Let us consider $q \geq 1, p \geq 1$ and $\mathbf{x} = (0, \sqrt{|\beta|}, \cdots, 0)^\top \in \mathcal{Q}_\beta^{p,q}$ where only the second element of $\mathbf{x}$ is nonzero. If $\boldsymbol{\xi} \in T_\mathbf{x} \mathcal{Q}_\beta^{p,q}$ is defined such that all its elements except the first one are equal to zero, then $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q < 0$. If $\boldsymbol{\xi} \in T_\mathbf{x} \mathcal{Q}_\beta^{p,q}$ is defined such that all its elements except the last one are equal to zero, then $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_q > 0$. The scalar product $\langle \cdot, \cdot \rangle_q$ is then indefinite if $q \geq 1$ and $p \geq 1$.

D Experimental Results

We complete here the experimental result section. First, we provide additional details about Zachary's karate club dataset experiments in Section D.1. We provide in Section D.2 similar experimental results on a larger social network dataset about co-authorship at NIPS conferences.

D.1 Additional details about Zachary's karate club experiments

Complexity: We implemented our approach in PyTorch 1.0 [22] which automatically calculates the Euclidean gradient $\nabla f(\mathbf{x}_i)$ for each node $v_i \in V$. Once $\nabla f(\mathbf{x}_i)$ is computed, the computation of $\boldsymbol{\chi}_i = \Pi_\mathbf{x}(\mathbf{G}\Pi_\mathbf{x}(\mathbf{G}\nabla f(\mathbf{x}_i)))$ is linear in the dimensionality d and is very efficient to compute. The exponential map is also efficient to calculate (*i.e.* linear algorithmic complexity).

Running time: The step size/learning rate for all our optimizers is fixed to $\eta = 10^{-6}$ (without momentum). Step size strategies could have been used to improve convergence rate but the goal of our experiment was only to verify that our solvers decrease the optimized function.

In the weighted case, we stop after 10,000 iterations of descent method. The training process takes 182 seconds ($\sim$3 minutes) on an Intel i7-7700k CPU, and 254 seconds ($\sim$4 minutes) on an Intel i7-8650U CPU when using the pseudo-Riemannian optimizer introduced in Section 4.2. The optimizer introduced in Section 4.1 is 10% faster (165 seconds on an Intel i7-7700k CPU) since it requires less computations.

D.2 Co-authorship from papers published at NIPS

We also quantitatively evaluate our approach on a dataset that describes co-authorship information from papers published at NIPS from 1988 to 2003 [9].

Description of the dataset: The graph $G = (V, E)$ is constructed by considering each author as a node, and an edge $e_k = (v_i, v_j)$ is created iff the authors i and j are co-authors of at least one paper. The capacity c_k is the number of papers co-authored by the pair of authors $e_k = (v_i, v_j)$. The original number of authors in the dataset is 2865, but only $n = |V| = 2715$ authors have at least one co-author. The number of edges is $m = |E| = 4733$, which means that the number of pairs of nodes that have no edge joining them is 3,679,522. The capacity c_k of each edge e_k is a natural number in $\{1, \cdots, 9\}$. The "leaders" of this dataset are not the authors with highest number of papers, but those with highest number of co-authors.

Implementation details and running times: We ran all our experiments for this dataset on a 12 GB NVIDIA TITAN V GPU. To fit into memory, when constructing each weaker set $\mathcal{W}(e_k)$, we take into account all the edges with capacity lower than c_k and randomly sample 42,000 pairs of nodes without edge joining them. Our fixed hyperparameters are: $\beta = -1$, the temperature is $\tau = 10^{-5}$, and the step size is $\eta = 10^{-8}$. We stop the training of representations lying on 4-dimensional manifolds (see Table 2) after 10,000 iterations, which takes 10 hours to train ($\sim 1,000$ iterations per hour). We stop

Table 2: Evaluation scores for the different learned representations lying on 4-dimensional manifolds on the NIPS dataset

Evaluation metric	$\mathbb{R}^4$ (flat)	$\mathcal{Q}_{-1}^{4,0}$ (hyperbolic)	$\mathcal{Q}_{-1}^{3,1}$ (ours)	$\mathcal{Q}_{-1}^{2,2}$ (ours)	$\mathcal{Q}_{-1}^{1,3}$ (ours)	$\mathcal{Q}_{-1}^{0,4}$ (spherical)
Spearman's ρ for the whole dataset	0.469	0.460	0.629	**0.667**	0.625	0.437
Spearman's ρ for $s_i \geq 10$	0.512	0.490	**0.552**	0.493	0.441	0.493
Spearman's ρ for $s_i \geq 20$	0.217	0.292	0.316	0.307	0.227	**0.387**
Recall@1 (in %)	70.7	70.2	**83.5**	82.6	78.1	69.7

Table 3: Evaluation scores for the different learned representations lying on 6-dimensional manifolds on the NIPS dataset

Evaluation metric	$\mathbb{R}^6$ (flat)	$\mathcal{Q}_{-1}^{6,0}$ (hyperbolic)	$\mathcal{Q}_{-1}^{5,1}$ (ours)	$\mathcal{Q}_{-1}^{4,2}$ (ours)	$\mathcal{Q}_{-1}^{3,3}$ (ours)	$\mathcal{Q}_{-1}^{2,4}$ (ours)	$\mathcal{Q}_{-1}^{1,5}$ (ours)	$\mathcal{Q}_{-1}^{0,6}$ (spherical)
Spearman's ρ for the whole dataset	0.554	0.455	0.617	0.666	**0.688**	0.675	0.610	0.456
Spearman's ρ for $s_i \geq 10$	0.459	0.515	0.507	0.501	0.514	0.470	0.494	**0.532**
Spearman's ρ for $s_i \geq 20$	0.372	0.321	0.211	0.317	**0.383**	0.370	0.213	0.257
Recall@1 (in %)	88.6	**99.9**	93.1	96.4	96.7	95.7	93.5	**99.9**

the training of 6-dimensional manifolds (see Table 3) after 25,000 iterations because they take longer to converge.

Quantitative evaluation: The evaluation process is the same as the evaluation for Zachary's karate club dataset. However, unlike for Zachary's karate club dataset, the factions and their leaders are unknown. We then only consider Spearman's rank correlation coefficient scores. We also consider the average recall at 1 as explained below.

The capacity matrix $\mathbf{C} \in \{0, 1, \cdots, 8, 9\}^{n \times n}$ is symmetric, so our score matrix is $\mathbf{S} = \mathbf{C}$. We select the most influential members based on the score $s_i = \sum_{j=1}^n \mathbf{S}_{ij} \in \{1, \cdots, 97\}$. Spearman's rank correlation coefficient [24] between the selected ordered scores s_i and corresponding $\delta_i = \sum_{j=1}^n \mathsf{d}(\mathbf{x}_i, \mathbf{x}_j)$ is reported in Table 2 and shows the relevance of our representations. The most influential members are selected if their score s_i is at least 1 (*i.e.* top n members where $n = 2715$ is the number of nodes), at least 10 (*i.e.* top 232 authors) and at least 20 (*i.e.* top 57 authors). We also report the average recall at 1 (in percentage): for each node v_i, we find the nearest neighbor v_j *w.r.t.* the chosen metric, the recall at 1 is 0 if $\mathbf{C}_{ij} = 0$, and 1 otherwise. Both $Q_{-1}^{3,1}$ $Q_{-1}^{2,2}$ outperform the hyperbolic model.

If we use the Euclidean norm of hyperbolic representations (in $\mathcal{Q}_{-1}^{4,0}$) as proxy, we obtain the following scores: 0.044, -0.113, 0.264 for the top 2715, to 232 and top 57 members respectively. These scores are worse than using δ_i as proxy due to the presence of cycles in the graph.

In the 6-dimensional manifold case, the Recall@1 score is better for the hyperbolic and spherical representations. On the other hand, the ultrahyperbolic representations perform better in terms of Spearman's rank correlation coefficient.

In conclusion, our proposed ultrahyperbolic representations extract hierarchy better than hyperbolic representations when the hierarchy graph contains cycles.