SPARSE HYPERBOLIC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Minimizing the space complexity of entity representations without the loss of information makes data science procedures computationally efficient and effective. For the entities with the tree structure, hyperbolic-space-based representation learning (HSBRL) has successfully reduced the space complexity of representations by using low-dimensional space. Nevertheless, it has not minimized the space complexity of each representation since it has used the same dimension for all representations and has not selected the best dimension for each representation. This paper, for the first time, constructs a sparse learning scheme to minimize the dimension for each representation in HSBRL. The most significant difficulty is that we cannot construct a well-defined sparse learning scheme for HSBRL based on a coordinate system since there is no canonical coordinate system that reflects geometric structure perfectly, unlike in linear space. Forcibly applying a linear sparse learning method on a coordinate system of hyperbolic space causes a non-uniform sparsity. Another difficulty is that existing Riemannian gradient descent cannot reach a sparse solution since the algorithm oscillates on a non-smooth function, which is essential in sparse learning. To overcome the above issue, for the first time, we geometrically define the sparseness and sparse regularization in hyperbolic space, to achieve geometrically uniform sparsity. Also, we propose the first optimization algorithm that can avoid the oscillation problem and obtain sparse representations in hyperbolic space by geometric shrinkage-thresholding.

1 INTRODUCTION

Data science applies a composition of mathematical operations, called algorithms, to solve real-life issues. For mathematical operations to handle real-world entities, we need their mathematical representations. Representation learning (RL) aims to obtain those entities' mathematical representations that reflect their semantic meaning. As an interface between the real world and data science, RL has been applied to various areas, e.g., machine translation and sentiment analysis for natural language (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Ganea et al., 2018a), and community detection and link prediction for social network data (Hoff et al., 2002; Perozzi et al., 2014; Tang et al., 2015b;a; Grover & Leskovec, 2016), pathway prediction of biochemical network (Dale et al., 2010; MA Basher & Hallam, 2021), and link prediction and triplet classification for knowledge base (Nickel et al., 2011; Bordes et al., 2013; Riedel et al., 2013; Nickel et al., 2016; Trouillon et al., 2016; Ebisu & Ichise, 2018). The fact that data science handles real entities throughout mathematical representations implies that reducing the space complexity of representations will benefit the whole of data science. Recent studies have shown that hyperbolic-space-based RL (HSBRL) can obtain effective low-dimensional representations, typical low space complexity representations, if the entities have the tree structure (Nickel & Kiela, 2017; Ganea et al., 2018b; Nickel & Kiela, 2018; Tay et al., 2018; Sala et al., 2018; Gu et al., 2019; Suzuki et al., 2019; Tifrea et al., 2019; Yu & De Sa, 2019; Law et al., 2019; Mathieu et al., 2019; Balazevic et al., 2019; Tabaghi & Dokmanic, 2020; Chami et al., 2020; Sonthalia & Gilbert, 2020; Kyriakis et al., 2021; Suzuki et al., 2021a;b). Still, existing HSBRL methods have not minimized the space complexity of each representation since they have used the same dimension for all representations and have not selected the best space complexity for each entity. Specifically, it would be more efficient to allocate high space complexity only for representations in non-tree-like parts of the data, if the data structure has tree-like parts and non-tree-like parts. For minimizing each representation's space complexity, sparse learning has been effective for linear-space-based machine learning. Sparse learning aims to obtain representations filled with the most zero elements to the extent that it does not lose the essential information. Typically, we have implemented sparse learning by the 1-norm regularization, which optimizes the sum of the original loss function of the problem and the 1-norm regularization term. The 1-norm regularization has mainly been developed in many application areas, such as time-frequency data analysis (Donoho & Johnstone, 1995; Chen et al., 2001), natural image processing (Olshausen & Field, 1996), and visual tracking (Liu et al., 2010). In the context of the linear model, the 1-norm regularization's statistical property has also been well studied (Tibshirani, 1996; Chen et al., 2001; Friedman et al., 2010) as well as geometric property (Donoho & Tanner, 2009a;b; 2010). Also, in statistics and machine learning communities, it has been intensively studied from theoretical perspectives (Ng, 2004; Shalev-Shwartz & Tewari, 2009; Tomioka et al., 2011; Zhang et al., 2016b) and applications for precision matrix learning (Friedman et al., 2008), neural-network-based learning (Liu et al., 2017b; Alizadeh et al., 2020), Ising model learning (Kuang et al., 2017), and transfer learning (Takada & Fujisawa, 2020). However, the above work is all for linear space; sparse learning has not been proposed for HSBRL. This paper aims to establish a sparse learning scheme for HSBRL. We have the following three challenges in establishing the scheme.

Firstly, defining sparsity in hyperbolic space is non-trivial. In linear space, we can define a point's sparsity as the number of zero elements in the coordinate representing the point. This definition's advantage is that it directly indicates representations' space complexity since a computer uses a coordinate system to represent points. Hence, we want to define the sparsity of a point in hyperbolic space so that it inherits this property. The issue is that the straightforward application of the definition in hyperbolic space is ill-defined since we cannot determine a unique coordinate system for a general manifold. To avoid this issue, we must have a geometric definition, i.e., one independent of the coordinate system, and we expect it to be interpretable as computational complexity at the same time.

Secondly, we need to design a continuous regularization term to induce the sparsity of representations. If the objective function includes the 0-norm, the number of non-zero elements, it involves combinatorial optimization since the 0-norm is not a continuous function. Hence, direct use of the 0-norm does not scale to big data. For this reason, we relax the 0-norm to the 1-norm in optimizing a function in linear space. However, defining a hyperbolic counterpart of the 1-norm is, again, non-trivial. For example, applying the 1-norm in a coordinate system of hyperbolic space naïvely has two drawbacks below. First, we will get different results depending on which coordinate system we use. Second, the regularization's strength cannot be uniform because a coordinate does not always reflect the distance structure in hyperbolic space.

Thirdly, we need an effective optimization algorithm to obtain sparse representations. We can apply the Riemannian gradient descent (RGD) algorithm in theory to the objective function with our 1-norm regularization term since it is a continuous function on the Riemannian manifold. However, the RGD fails to get sparse representations since it oscillates around the true sparse solution.

This paper proposes a novel sparse learning scheme for HSBRL to solve these issues. It is regularization-based, so applicable to any HSBRL using a continuous loss function. The **core contributions** of our scheme are the following three, each of which addresses one of the above issues:

1. The ***Hyperbolic sparsity***, a novel geometric definition of a point's sparsity. We can also interpret the value as the space complexity of a point in hyperbolic space.
2. The ***Hyperbolic 1-norm***, a novel sparse regularization term for continuous optimization. It achieves a uniform sparseness strength since it measures the geometric distance from the point to the subspaces with higher hyperbolic sparsity.
3. The ***Hyperbolic iterative shrinkage-thresholding algorithm (HISTA)***, a novel optimization algorithm free from the oscillating issue, realizing the shrinkage-thresholding idea (Bruck Jr, 1977; Chambolle et al., 1998; Figueiredo & Nowak, 2003; Daubechies et al., 2004; Vonesch & Unser, 2007; Hale et al., 2007; Wright et al., 2009; Beck & Teboulle, 2009) in hyperbolic space.

The above definitions are all geometric, i.e., free from the coordinate system selection problem. Also, we give closed-form formulae for them in hyperbolic space. We can calculate them in linear time/space complexity for the space dimension, the same as in the linear case. The above concepts are well-defined for a general Cartan-Hadamard manifold (CHM), i.e., a simply-connected complete Riemannian manifold with non-positive curvature, including Euclidean space, hyperbolic space, and the products of Euclidean and multiple hyperbolic spaces.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 is the preliminary for the core discussion of the paper. Section 4 geometrically defines the sparsity on a

CHM, including hyperbolic space. Section 5 proposes our sparse learning scheme on a Riemannian manifold. Section 6 specializes the scheme for hyperbolic space and gives specific formulae, Section briefly explains numerical experiments in Section F while the details are in the Appendix. Section 8 discusses the limitation of the framework. Section 9 concludes this paper. Note that we provide the table of acronyms in Section H in Supplementary materials.

2 Related work

This paper proposes multiple operations in hyperbolic space. There has been much work on defining operations for hyperbolic space in the mathematical context e.g., (Gromov, 1987; Ungar, 1994; 1996; Sabinin et al., 1998), and machine learning context e.g., the work cited above and general Riemannian optimization work (Absil et al., 2009; Qi et al., 2010; Zhang & Sra, 2016; Zhang et al., 2016a; Liu et al., 2017a; Becigneul & Ganea, 2019; Kasai et al., 2019; Zhou et al., 2019) and neural hyperbolic network papers (Ganea et al., 2018c; Chami et al., 2019; Liu et al., 2019; Tan et al., 2021; Shimizu et al., 2021; Takeuchi et al., 2022). Still, none of them have proposed the operations to get sparse representations. The work closest to our motivation is the Riemannian proximal gradient (RPG) descent (Huang & Wei, 2022) since the iterative shrinkage-thresholding algorithm (ISTA) can be regarded as a special case of the proximal gradient descent method. Despite its solid theoretical background, however, the closed form solution of the RPG operation for the sparse regularization has not been known except for the linear space case, where the RPG is reduced to the ISTA. Hence, the RPG cannot get sparse representations. Conversely, our algorithm can give sparse representations and computationally cheap since the closed form algorithm is given for hyperbolic space. Still, both the RPG and our algorithm can be regarded as generalizations of the ISTA since they are reduced to the ISTA for optimizing the 1-norm regularization problem in linear space.

3 Preliminary

Linear space We denote the set of real values, nonnegative real values, positive real values, and nonnegative integers by $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{>0}$, and $\mathbb{Z}_{\geq 0}$, respectively. For $D \in \mathbb{Z}_{\geq 0}$, we denote the D-dimensional real coordinate space (RCS) by $\mathbb{R}^D$. Also, for $M, N \in \mathbb{Z}_{\geq 0}$ we denote the set of $M \times N$-real matrices by $\mathbb{R}^{M \times N}$. In this paper, an element in RCS is denoted by a bold lower letter, e.g., $\boldsymbol{p}, \boldsymbol{q}$, and a real matrix is denoted by a bold upper letter, e.g., $\boldsymbol{G}$. We denote the transpose of $\boldsymbol{p} \in \mathbb{R}^D$ by $\boldsymbol{p}^\top$ and define $|\boldsymbol{p}| := \sqrt{\boldsymbol{p}^\top \boldsymbol{p}}$. We denote the D-dimensional zero vector and one vector by $\mathbf{0}_D$ and $\mathbf{1}_D$, respectively, and $D \times D$ identity matrix by $\mathbf{I}_D$. In this paper, a mathematical constant is denoted by an upright letter, e.g., $\mathbf{0}_D$ and $\mathbf{I}_D$, whereas a variable is denoted by an italic letter, e.g., $D, \boldsymbol{p}, \boldsymbol{q}$. We call the inner product space $\left(\mathbb{R}^D, \langle \cdot, \cdot \rangle\right)$ the D-dimensional Euclidean vector space (EVS), where $\langle \cdot, \cdot \rangle : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ by $\langle \boldsymbol{p}, \boldsymbol{q} \rangle = \boldsymbol{p}^\top \boldsymbol{q}$. For a scalar-valued function f of a scalar, typically a hyperbolic function in this paper, we extend it to a vector-valued function of a vector, defined by $f\left(\begin{bmatrix} p_1 & p_2 & \cdots & p_D \end{bmatrix}^\top\right) := \begin{bmatrix} f(p_1) & f(p_2) & \cdots & f(p_D) \end{bmatrix}^\top$.

Riemannian manifold Let $\mathcal{M}$ and $\mathcal{M}'$ be C^∞-Riemannian manifolds. We denote by $C^\infty(\mathcal{M})$ the set of real C^∞-functions on $\mathcal{M}$, by $\mathscr{T}_p\mathcal{M}$ the tangent space at point p, and by $\langle \cdot, \cdot \rangle_p^{\mathcal{M}} : \mathscr{T}_p\mathcal{M} \times \mathscr{T}_p\mathcal{M} \to \mathbb{R}$ the metric tensor at $\mathscr{T}_p\mathcal{M}$. For C^∞-map $\varphi : \mathcal{M} \to \mathcal{M}'$, the differential $\mathrm{d}\varphi_p : \mathscr{T}_p\mathcal{M} \to \mathscr{T}_{\varphi(p)}\mathcal{M}'$ of φ at p is defined by $(\mathrm{d}\varphi_p(v))(f) = v(f \circ \varphi)$. We call a diffeomorphism $\varphi : \mathcal{M} \to \mathcal{M}'$ an ***isometry*** if $\langle u, v \rangle_p^{\mathcal{M}} = \langle \mathrm{d}\varphi_p(u), \mathrm{d}\varphi_p(v) \rangle_{\varphi(p)}^{\mathcal{M}}$ for any point p and tangent vectors $u, v \in \mathscr{T}_p\mathcal{M}$. We denote by $\Delta_{\mathcal{M}}(p, q)$ the Riemannian distance between $p \in \mathcal{M}$ and $q \in \mathcal{M}$, which is defined as the length of the shortest piecewise smooth curve connecting p and q. We denote by $\exp_p : \mathscr{T}_p\mathcal{M} \to \mathcal{M}$ the ***exponential map*** at point $p \in \mathcal{M}$. Here, we define the exponential map based on the Levi-Civita connection. Intuitively, $\exp_p(v)$ indicates the point reached in one unit time by the motion with a zero acceleration (called a ***geodesic***) starting at the point p with the initial velocity v. If the exponential map $\exp_p$ is bijective, then we can define its inverse map, called ***logarithmic map*** $\log_p : \mathcal{M} \to \mathscr{T}_p\mathcal{M}$. A complete, simply connected Riemannian manifold with nonpositive sectional curvature is called a ***Cartan-Hadamard manifold (CHM)***. The exponential map at any point in a CHM is diffeomorphic, in particular, bijective, according to Cartan-Hadamard theorem (See, e.g., Kobayashi & Nomizu, 1996). For definitions of these terms of the Riemannian geometry, we refer readers to (e.g., Kobayashi & Nomizu, 1996; Lee, 2018).

Coordinate system This paper deals with CHMs. Since they are always diffeomorphic to $\mathbb{R}^D$ for some $D \in \mathbb{Z}_{\geq 0}$, we can represent a CHM by a pair $(\mathbb{U}, \boldsymbol{G}_\cdot)$ of an open set $\mathbb{U} \subset \mathbb{R}^D$ and the coordinate representation matrix $\boldsymbol{G}_p \in \mathbb{R}^{D \times D}$ of the Riemannian metric at point p for all $p \in \mathbb{U}$. We denote by $(\mathbb{U}, \boldsymbol{G}_\cdot)$ the CHM determined this way. For the open set $\mathbb{U} \subset \mathbb{R}^D$, the coordinate system $\left(x^1, x^2, \dots, x^D\right)$ for $\mathbb{U}$, and $d = 1, 2, \dots, D$, we define the partial derivative operator $\partial_d|_{\boldsymbol{p}} : C^\infty(\mathbb{U}) \to \mathbb{R}$ by $\partial_d|_{\boldsymbol{p}} f := \frac{\partial}{\partial x^d} f(\boldsymbol{p})$, where $C^\infty(\mathbb{U})$ is the set of real C^∞-functions on $\mathbb{U}$. For $\boldsymbol{v} := \begin{bmatrix} v^1 & v^2 & \dots & v^D \end{bmatrix}^\top \in \mathbb{R}^D$, we denote $v^1 \partial_1|_{\boldsymbol{p}} + v^2 \partial_2|_{\boldsymbol{p}} + \dots + v^D \partial_D|_{\boldsymbol{p}}$ by $\boldsymbol{v}^\top \boldsymbol{\partial}|_{\boldsymbol{p}}$. Here, we have that $\mathscr{T}_{\boldsymbol{p}}\mathcal{M} = \left\{ \boldsymbol{v}^\top \boldsymbol{\partial}|_{\boldsymbol{p}} \mid \boldsymbol{v} \in \mathbb{R}^D \right\}$. Using the coordinate representation $\boldsymbol{G}_{\boldsymbol{p}}$ at $\boldsymbol{p} \in \mathbb{U}$ of the Riemannian metric $\langle \cdot, \cdot \rangle_{\cdot, (\mathbb{U}, \boldsymbol{G}_\cdot)}$, we can calculate the inner product $\langle \boldsymbol{u}^\top \boldsymbol{\partial}|_{\boldsymbol{p}}, \boldsymbol{v}^\top \boldsymbol{\partial}|_{\boldsymbol{p}} \rangle_{\cdot, (\mathbb{U}, \boldsymbol{G}_\cdot)}$ of two tangent vectors $\boldsymbol{u}^\top \boldsymbol{\partial}|_{\boldsymbol{p}}, \boldsymbol{v}^\top \boldsymbol{\partial}|_{\boldsymbol{p}} \in \mathscr{T}_{\boldsymbol{p}}(\mathbb{U}, \boldsymbol{G}_\cdot)$ by $\langle \boldsymbol{u}^\top \boldsymbol{\partial}|_{\boldsymbol{p}}, \boldsymbol{v}^\top \boldsymbol{\partial}|_{\boldsymbol{p}} \rangle_{\cdot, (\mathbb{U}, \boldsymbol{G}_\cdot)} = \boldsymbol{u}^\top \boldsymbol{G}_{\boldsymbol{p}} \boldsymbol{v}$.

The following examples are the CHMs we mainly discuss in the paper.

Example 1. (a) The CHM $\left(\mathbb{R}^D, \mathbf{I}_{D,\cdot}\right)$ is the D-dimensional Euclidean space, where $\mathbf{I}_{D,\boldsymbol{p}} = \mathbf{I}_D$ for all $\boldsymbol{p} \in \mathbb{R}^D$.

(b) Let $\mathbb{D}^D$ be the D-dimensional open ball $\mathbb{D}^D := \left\{ \boldsymbol{p} \in \mathbb{R}^D \mid \boldsymbol{p}^\top \boldsymbol{p} < 1 \right\}$. For $\boldsymbol{p} \in \mathbb{D}^D$, we define $\boldsymbol{G}^{\mathrm{P}}_{\boldsymbol{p}}, \boldsymbol{G}^{\mathrm{K}}_{\boldsymbol{p}} \in \mathbb{R}^{D \times D}$ by $\boldsymbol{G}^{\mathrm{P}}_{\boldsymbol{p}} = \frac{4}{1 - \boldsymbol{p}^\top \boldsymbol{p}} \mathbf{I}_D$ and $\boldsymbol{G}^{\mathrm{K}}_{\boldsymbol{p}} = \left(1 - \boldsymbol{p}^\top \boldsymbol{p}\right)\left(\mathbf{I}_D - \boldsymbol{p}\boldsymbol{p}^\top\right)$. The CHMs $\left(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_\cdot\right)$ and $\left(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{K}}_\cdot\right)$ are isometric to each other. Specifically, there is an isometry $\varphi_{\mathrm{P,K}} : \mathbb{D}^D \to \mathbb{D}^D$ defined by $\varphi_{\mathrm{P,K}}(\boldsymbol{p}) := \tanh\left(\operatorname{atanh}\left(\boldsymbol{p}\right)\right) \frac{\boldsymbol{p}}{|\boldsymbol{p}|}$. The CHM determined by these two is called the ***D-dimensional hyperbolic space***. When we distinguish these two coordinate systems of the D-dimensional hyperbolic space, we call $\left(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_\cdot\right)$ and $\left(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{K}}_\cdot\right)$ the ***Poincaré model*** and ***Klein model***, respectively.

(c) The product manifold of Euclidean space and hyperbolic spaces is also a CHM. In the representation learning context, we consider optimization of multiple points in a CHM. This can be regarded as the optimization in the product manifold. In general, the product manifold of CHMs is also a CHM.

4 The Cartan-Hadamard Sparseness

In this section, we define the sparsity of a point in CHM geometrically. Our sparsity also indicates the space computational complexity to represent a point in hyperbolic space. In the definition of the sparsity of a point in RCS, the origin and canonical bases play an essential role as a "reference point" and "reference direction." Hence, we need to begin with the definition of the origin and the bases to discuss the sparsity of a point in a manifold.

Definition 1 (CHM with an origin and orthonormal bases (CHMOO))**.** Let $\mathcal{M}$ be a D-dimensional CHM. For a point $\mathrm{o} \in \mathcal{M}$ and orthonormal basis (ONB) $e_1, e_2, \dots, e_D \in \mathscr{T}_{\mathrm{o}}\mathcal{M}$, the triple $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ is called a D-dimensional ***CHM with an origin and orthonormal bases (CHMOO)***. Here, the point o and set $(e_1, e_2, \dots, e_D)$ are called the origin and ONB of the CHMOO, respectively.

Definition 2 (Isometric CHMOOs)**.** Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ and $(\mathcal{M}', \mathrm{o}', (e'_1, e'_2, \dots, e'_D))$ be CHMOOs and $\varphi : \mathcal{M} \to \mathcal{M}'$ is an isometry. If $\varphi(\mathrm{o}) = \mathrm{o}'$ and $d\varphi_{\mathrm{o}}(e_d) = e'{}_d$ for $d = 1, 2, \dots, D$, then φ is called an ***isometry*** from $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ and $(\mathcal{M}', \mathrm{o}', (e'_1, e'_2, \dots, e'_D))$. If there exists an isometry between two CHMOOs, we say that the CHMOOs are ***isometric*** to each other.

Example 2. (a) The triple $\left((\mathbb{R}^D, \mathbf{I}_{D,\cdot}), \mathbf{0}, (\partial_1|_{\mathbf{0}}, \partial_2|_{\mathbf{0}}, \dots, \partial_D|_{\mathbf{0}})\right)$ is a CHMOO. We call it the D-dimensional ***Euclidean vector CHMOO (EVCHMOO)***. Here, we can identify $\boldsymbol{v}^\top \boldsymbol{\partial}|_{\mathbf{0}} \in \mathscr{T}_{\mathbf{0}}\mathbb{R}^D$ with $\boldsymbol{v} \in \mathbb{R}^D$ since $\exp_{\mathbf{0}}\left(\boldsymbol{v}^\top \boldsymbol{\partial}|_{\mathbf{0}}\right) = \boldsymbol{v}$ holds. Under these identifications, we can say that the canonical bases $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_D$ of the D-dimensional RCS are the ONB of the EVCHMOO. Although the D-dimensional EVCHMOO is no more than Riemannian reformulation of the D-dimensional RCS, it helps us to see the correspondence between the D-dimensional RCS and hyperbolic space.

(b) $\left((\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_\cdot), \mathbf{0}, \left(\frac{1}{2}\partial_1|_{\mathbf{0}}, \frac{1}{2}\partial_2|_{\mathbf{0}}, \dots, \frac{1}{2}\partial_D|_{\mathbf{0}}\right)\right)$ and $\left((\mathbb{D}^D, \boldsymbol{G}^{\mathrm{K}}_\cdot), \mathbf{0}, (\partial_1|_{\mathbf{0}}, \partial_2|_{\mathbf{0}}, \dots, \partial_D|_{\mathbf{0}})\right)$ are CHMOOs isometric to each other. We call it the D-dimensional ***hyperbolic CHMOO***. Here, the map $\varphi_{\mathrm{P,K}}$ is an isometry.

(c) If we have multiple CHMOOs, we can consider the product of them. Let $\left(\mathcal{M}^{[m]}, \mathrm{o}^{[m]}, \left(\tilde{e}_1^{[m]}, \tilde{e}_2^{[m]}, \dots, \tilde{e}_{D^{[m]}}^{[m]}\right)\right)_{m=1}^{M}$ be the series of CHMOOs and let $\mathcal{M} = \mathcal{M}^{[1]} \times$

$\mathcal{M}^{[2]} \times \dots, \times \mathcal{M}^{[M]}$ be the product CHM and define $\mathrm{o} = \left(\mathrm{o}^{[1]}, \mathrm{o}^{[2]}, \dots, \mathrm{o}^{[M]}\right) \in \mathcal{M}$. For any $m = 1, 2, \dots, M$, $d = 1, 2, \dots, D^{[m]}$, and any basis $\tilde{e}_d^{[m]} \in \mathscr{T}_{\mathrm{o}^{[m]}} \mathcal{M}^{[m]}$, there is a corresponding tangent vector $e_d^{[m]} \in \mathscr{T}_{\mathrm{o}} \mathcal{M}$ (See Appendix for the specific construction).

We can see that $\left(e_1^{[1]}, e_2^{[1]}, \dots, e_{D^{[1]}}^{[1]}, e_1^{[2]}, e_2^{[2]}, \dots, e_{D^{[2]}}^{[2]}, \dots, e_1^{[M]}, e_2^{[M]}, \dots, e_{D^{[M]}}^{[M]}\right)$ is an ONB in $\mathscr{T}_p \mathcal{M}$. Hence, we can define the product CHMOO $\left(\mathcal{M}, \mathrm{o}, \left(e_1^{[1]}, e_2^{[1]}, \dots, e_{D^{[1]}}^{[1]}, e_1^{[2]}, e_2^{[2]}, \dots, e_{D^{[2]}}^{[2]}, \dots, e_1^{[M]}, e_2^{[M]}, \dots, e_{D^{[M]}}^{[M]}\right)\right)$.

Other important examples are the direct product of Euclidean space and (multiple) hyperbolic space, which we discuss in Appendix.

In EVS, a vector has sparsity with respect to a basis if the point is orthogonal to the basis. We can generalize this definition to CHMOOs as follows.

Definition 3 (Sparse hyperplane (SHP))**.** Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ be a D-dimensional CHMOO. The d-th ***sparse hyperplane (SHP)***, denoted by $\mathit{\Pi}_d^{\mathcal{M}}$, is defined by $\mathit{\Pi}_d^{\mathcal{M}} := \{p \in \mathcal{M} \mid \langle e_d, \log_{\mathrm{o}}(p) \rangle_{\mathrm{o}} = 0\}$.

Remark 1. (i) Since $\exp_{\mathrm{o}} : \mathscr{T}_{\mathrm{o}} \mathcal{M} \to \mathcal{M}$ is bijective in a CHM, $\log_{\mathrm{o}}$ is uniquely defined.

(ii) An SHP is a $D-1$-dimensional submanifold of $\mathcal{M}$.

Once we have defined the sparseness of a point for a basis, we can define that for all the bases as follows.

Definition 4 (Cartan-Hadamard sparseness and 0 norm)**.** Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ be a D-dimensional CHMOO. We define the ***Cartan-Hadamard sparseness (CH sparseness)*** $\mathrm{sp}^{\mathcal{M}}(p)$ of a point $p \in \mathcal{M}$ as the number of SHPs that includes p, i.e., $\mathrm{sp}^{\mathcal{M}}(p) := \left|\left\{d = 1, 2, \dots, D \mid p \in \mathit{\Pi}_{\hat{d}}\right\}\right|$. Also, we define the ***Cartan-Hadamard 0-norm (CH 0-norm)*** $\|p\|_0^{\mathcal{M}}$ of a point $p \in \mathcal{M}$ by $\|p\|_0^{\mathcal{M}} = D - \mathrm{sp}^{\mathcal{M}}(p)$.

Remark 2. (i) By definition, the CH 0-norm does not depend on the coordinate system. Specifically, for two isometric CHMOOs $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ and $(\mathcal{M}', \mathrm{o}', (e'_1, e'_2, \dots, e'_D))$ and an isometry $\varphi : \mathcal{M} \to \mathcal{M}'$, and we have that $\|p\|_0^{\mathcal{M}} = \|\varphi(p)\|_0^{\mathcal{M}'}$ for any point $p \in \mathcal{M}$.

(ii) Since the SHPs, CH sparseness, and CH 0-norm depend on the origin o and ONB $(e_1, e_2, \dots, e_D)$, we could clarify these by, e.g., $\mathit{\Pi}_d^{((\mathcal{M}, \langle \cdot, \cdot \rangle_\cdot), \mathrm{o}, (e_1, e_2, \dots, e_D))}$. Nevertheless, we omit these from notation since they are clear for each CHM and coordinate system in this paper.

Example 3. For the D-dimensional EVCHMOO $\left((\mathbb{R}^D, \mathbf{I}_{D,\cdot}), \mathbf{0}, (\partial_1|_{\mathbf{0}}, \partial_2|_{\mathbf{0}}, \dots, \partial_D|_{\mathbf{0}})\right)$, the d-th SHP is given by $\mathit{\Pi}_d^{\mathcal{M}} = \left\{\boldsymbol{p} \in \mathbb{R}^D \mid \boldsymbol{p}^\top \boldsymbol{e}_D = 0\right\} = \left\{\boldsymbol{p} \in \mathbb{R}^D \mid [\boldsymbol{p}]_d = 0\right\}$, where $[\boldsymbol{p}]_d$ indicates the d-th element of $\boldsymbol{p}$. Hence, we have $\|\boldsymbol{p}\|_0^{(\mathbb{R}^D, \mathbf{I}_{D,\cdot})} = \|\boldsymbol{p}\|_0$. The above results directly follow the fact that $\langle \log_{\mathbf{0}}(\boldsymbol{p}), \log_{\mathbf{0}}(\boldsymbol{q}) \rangle_{\mathbf{0}}^{(\mathbb{R}^D, \mathbf{I}_{D,\cdot})} = \boldsymbol{p}^\top \boldsymbol{q}$.

Example 4. In D-dimensional hyperbolic space, the specific form of the 0-norm is given by $\|\boldsymbol{p}\|_0^{\left(\mathbb{D}^D, \boldsymbol{G}_\cdot^{\mathrm{P}}\right)} = \|\boldsymbol{p}\|_0$, and $\|\boldsymbol{p}\|_0^{\left(\mathbb{D}^D, \boldsymbol{G}_\cdot^{\mathrm{K}}\right)} = \|\boldsymbol{p}\|_0$.

We call $\|\boldsymbol{p}\|_0^{\left(\mathbb{D}^D, \boldsymbol{G}_\cdot^{\mathrm{P}}\right)}$ and $\|\boldsymbol{p}\|_0^{\left(\mathbb{D}^D, \boldsymbol{G}_\cdot^{\mathrm{P}}\right)}$ the ***hyperbolic 0-norm*** of $\boldsymbol{p} \in \mathbb{D}^D$ in the Poincaré model and in the Klein model, respectively.

Example 5. The SHP $\mathit{\Pi}_{[m],d}^{\mathcal{M}}$ of the product CHMOO corresponding to $e_d^{[m]} \in \mathscr{T}_{\mathrm{o}}^{[m]} \mathcal{M} \subset \mathscr{T}_{\mathrm{o}} \mathcal{M}$ is given by

$$\mathit{\Pi}_{[m],d}^{\mathcal{M}} = \left\{ \left(p^{[1]}, p^{[2]}, \dots, p^{[M]}\right) \middle| p^{[m']} \in \mathit{\Pi}_d^{\mathcal{M}^{[m]}} \text{ if } m' = m, \quad p^{[m']} \in \mathcal{M}^{[m]} \text{ if } m' \neq m, \right\}. \quad (1)$$

Remark 3. (i) Example 4 implies that the Riemannian 0-norm, which is defined geometrically in Definition 4, also indicates the number of non-zero elements to represent the point in each coordinate system on a computer. This fact justifies our Riemannian 0-norm definition for HSBRL in both geometric and computational senses.

(ii) Results similar to Example 4 hold for the proper velocity model, hyperboloid model, and hemisphere model. Note that it does NOT hold for the upper half space model. In the first place, the upper plane model does not use the coordinate whose 0-norm is larger than two; hence the model is not suitable for discussing the sparsity of a point. This is not our definitions' issue but the coordinate system's issue. Even in RCS, if we introduce a new coordinate system by $\boldsymbol{p}' = \exp(\boldsymbol{p})$, where $\boldsymbol{p}$ is a canonical coordinate, then we cannot discuss the sparsity on the new coordinate system since all the elements of the coordinate is positive everywhere.

5 Sparse learning scheme on a Cartan-Hadamard manifold (CHM)

We have defined the CH sparseness and 0-norm. We are ready to discuss the scheme to obtain sparse representations in the sense of the CH sparseness. This section discusses the general CHMOO case. We will discuss the hyperbolic CHMOO case in the next section as a special case of this section's discussion.

5.1 The Cartan-Hadamard 1-norm (CH 1-norm)

The 0-norm in RCS is not continuous as a function of a point. Hence, we cannot optimize the 0-norm regularized function by a gradient method. To solve this issue, we have often relaxed the 0-norm to the 1-norm, which is a continous function. Likewise, we define the CH 1-norm of a point on a CHMOO as a relaxation of the CH 0-norm so that we can use it as a regularization term for gradient methods. To define an appropriate counterpart of the 1-norm on CHMOO, we discuss D-dimensional RCS again. The 1-norm $\|\boldsymbol{p}\|_1$ of a point $\boldsymbol{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_D \end{bmatrix}^\top \in \mathbb{R}^D$ is given by the sum of the absolute element values $|p_1|, |p_2|, \ldots, |p_D|$. The absolute value $|p_d|$ of each element of a vector is natural as a relaxed regularization term to induce sparsity with respect to the d-th axis since the element indicates the distance $\Delta_{\mathcal{M}}(\boldsymbol{p}, \mathit{\Pi}_{\hat{d}}) := \inf_{\boldsymbol{p}' \in \mathit{\Pi}_{\hat{d}}} \Delta_{\mathcal{M}}(\boldsymbol{p}, \boldsymbol{p}')$ between the point and the d-th SHP. The above observation inspires our definition below of the 1-norm in a CHMOO. We begin with defining the (signed) distance between a point and an SHP, since it plays an essential role to define the 1-norm as we have seen in RCS.

Definition 5 (Signed SHP distance map (SSDM))**.** Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \ldots, e_D))$ be a D-dimensional CHMOO, and assume that $\exp_{\mathrm{o}} : \mathscr{T}_{\mathrm{o}}\mathcal{M} \to \mathcal{M}$ is bijective. We define the ***signed distance*** $\delta_d^{\mathcal{M}}(p) \in \mathbb{R}$ of a point $p \in \mathcal{M}$ from the d-th SHP by $\delta_d^{\mathcal{M}}(p) := \operatorname{sgn}(\langle e_d, \log_{\mathrm{o}}(p)\rangle_{\mathrm{o}})\Delta_{\mathcal{M}}(p, \mathit{\Pi}_{\hat{d}})$, where $\Delta_{\mathcal{M}}(p, \mathit{\Pi}_{\hat{d}}) := \inf_{p' \in \mathit{\Pi}_{\hat{d}}} \Delta_{\mathcal{M}}(p, p')$. We also define the ***signed SHP distance map (SSDM)*** $\boldsymbol{\delta}^{\mathcal{M}} : \mathcal{M} \to \mathbb{R}^D$ by $\boldsymbol{\delta}^{\mathcal{M}}(p) := \begin{bmatrix} \delta_1^{\mathcal{M}}(p) & \delta_2^{\mathcal{M}}(p) & \cdots & \delta_D^{\mathcal{M}}(p) \end{bmatrix}^\top$.

Remark 4. We consider the signed version in Definition 5 since we often have a bijective SSDM as a result of considering the signed version. We can specify a point by the signed distances if the SSDM is bijective. Hence, a bijective SSDM plays an essential role to develop our optimization algorithm controlling the signed distances, as we see later.

Appendix A has the visualization of the SSDM. Based on the SSDM, we define the Cartan-Hadamard 1-norm (CH 1-norm), which is this subsection's objective.

Definition 6 (Cartan-Hadamard 1-norm (CH 1-norm))**.** Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \ldots, e_D))$ be a D-dimensional CHMOO, and assume that $\exp_{\mathrm{o}} : \mathscr{T}_{\mathrm{o}}\mathcal{M} \to \mathcal{M}$ is bijective. We define the ***Cartan-Hadamard 1-norm (CH 1-norm)*** $\|p\|_1^{\mathcal{M}}$ of a point $p \in \mathcal{M}$ by $\|p\|_1^{\mathcal{M}} := \left\|\boldsymbol{\delta}^{\mathcal{M}}(p)\right\|_1 = \sum_{d=1}^{D} \left|\delta_d^{\mathcal{M}}(p)\right|$. We call it the ***CH 1-norm regularization*** to add the CH 1-norms of the points multiplied by some factor in optimizing a loss function of points in a CHMOO.

Example 6. Consider the D-dimensional EVCHMOO $\left((\mathbb{R}^D, \boldsymbol{G}_{\cdot}^{\mathbb{R}^D}), \mathbf{0}, (\partial_1|_{\mathbf{0}}, \partial_2|_{\mathbf{0}}, \ldots, \partial_D|_{\mathbf{0}})\right)$. Then, the SSDM is given by $\boldsymbol{\delta}^{(\mathbb{R}^D, \boldsymbol{G}_{\cdot}^{\mathbb{R}^D})}(\boldsymbol{p}) = \boldsymbol{p}$. The CH 1-norm is given by $\|\boldsymbol{p}\|_1 = \sum_{d=1}^{D} |p_d|$, which is equivalent to the 1-norm as a real numeric vector.

Remark 5. Since the SHP does not depend on the coordinate system, the SSDM does not depend on it, either. Hence, the CH 1-norm does not depend on the coordinate system, either.

5.2 The Cartan-Hadamard iterative shrinkage-thresholding algorithm

To optimize a function differentiable almost everywhere, one might use (sub)gradient methods. However, the direct application of gradient methods in the 1-norm regularization causes residuals to the sparse solution. See the following simplest example. Details are also in Appendix B

Example 7. Suppose that we optimize $f(p) = |p|$ by the gradient descent method with learning rate $\alpha > 0$ and initial point $p^{(0)} \neq 0$. Then the algorithm ends up oscillating between $p^{(0)} - \alpha n$ and $p^{(0)} - \alpha(n+1)$ and cannot achieve the true sparse solution $p^{(0)} = 0$, unless $p^{(0)}$ is an integral multiple of α. Here, $n = \left[\frac{p^{(0)}}{\alpha}\right]$ is the maximum integer that is no greater than $\frac{p^{(0)}}{\alpha}$.

Example 7 is also an example of an oscillation in a CHMOO optimization problem since it is equivalent to optimization of the 1-norm function of 1-dimensional Euclidean space or hyperbolic space. Note that 1-dimensional Euclidean space and 1-dimensional hyperbolic space are isometric.

To address the above oscillation issue to obtain the true sparse solution, the ***iterative shrinkage-thresholding algorithm (ISTA)*** (e.g., Bruck Jr, 1977) has been used. The ISTA is designed to optimize a function in RCS in the following form:

$$J(\boldsymbol{p}) \coloneqq L(\boldsymbol{p}) + \lambda \|\boldsymbol{p}\|_1, \tag{2}$$

where $L : \mathbb{R}^D \to \mathbb{R}$ is an almost everywhere differentiable function and $\lambda \in \mathbb{R}_{\geq 0}$ is the regularization weight. The ISTA iterates the following two steps: (i) $\boldsymbol{q} \leftarrow \boldsymbol{p} - \alpha \frac{\partial}{\partial \boldsymbol{p}} L(\boldsymbol{p})$, (ii) $\boldsymbol{p} \leftarrow \mathcal{T}_{\alpha\lambda}(\boldsymbol{q})$. Here, $\mathcal{T}_\beta : \mathbb{R}^D \to \mathbb{R}^D$ is the ***soft-thresholding operator (STO)*** defined by

$$\mathcal{T}_\beta \left(\begin{bmatrix} p_1 & p_2 & \cdots & p_D \end{bmatrix}^\top \right) \coloneqq \begin{bmatrix} \tau_\beta(p_1) & \tau_\beta(p_2) & \cdots & \tau_\beta(p_D) \end{bmatrix}^\top, \tag{3}$$

where $\tau_\beta : \mathbb{R}^D \to \mathbb{R}^D$ is defined by $\tau_\beta(p) \coloneqq \operatorname{sgn}(p) \max\{|p| - \beta, 0\}$.

This section aims to generalize the ISTA for functions in a CHMOO. We begin with discussing why in $\mathbb{R}^D$ the ISTA, and in particular, the STO, works well to obtain the sparse solution in the 1-norm regularization. The STO decreases the absolute value of each coordinate element by β, leaving the sign unchanged. If the absolute value is no larger than β, then the element will shrink to zero. This is why the ISTA causes sparsity of the solution. This is in contrast to the gradient descent method. Since each SHP's volume is zero, we cannot expect gradient descent method to induce sparsity unless the function has a special property. Let us evaluate the strength of the shrinkage effect geometrically. Recall that each element of a vector indicates the signed distance from an SHP. Hence, the following holds.

Proposition 1. *For $\boldsymbol{p} \in \mathbb{R}^D$, $\|\mathcal{T}_\beta(\boldsymbol{p})\|_0 < \|\boldsymbol{p}\|_0$ if and only if $\min_{d=1,2,\dots,D} \Delta_{\mathbb{R}^D}\left(\boldsymbol{p}, \mathit{\Pi}_{\hat{d}}\right) \leq \beta$.*

Remark 6. Proposition 1 clarifies when the STO increases the sparsity of a point. It only depends on the distance of a point from the SHPs and β, and does not directly depend on the position of the point. This means that the strength of STO's sparsity induction effect is "uniform" everywhere in $\mathbb{R}^D$ and only β determines the strength. Since $\beta = \alpha\lambda$ in the ISTA, the strength of the STO is proportional to λ. This is compatible to the aim of the objective function J in (2), where we expect λ to control the regularization strength.

Algorithm 1 Cartan-Hadamard ISTA (CHISTA)

```
Require: $p_{\mathrm{init}} \in \mathcal{M}$: initial point,
    $\alpha \in \mathbb{R}_{>0}$: learning rate,
    $T \in \mathbb{Z}_{\geq 0}$: # iterations.
Ensure: $p_{\mathrm{output}} \in \mathcal{M}$
  $p^{(0)} \leftarrow p_{\mathrm{init}}$
  for $t \leftarrow 1, 2, \dots, T$ do
    $q^{(t)} \leftarrow \exp_{p^{(t-1)}}\left(-\alpha \operatorname{grad}_{p^{(t-1)}} L\right)$
    $p^{(t)} \leftarrow \mathcal{T}_{\alpha\lambda}\left(q^{(t)}\right)$        ▷Difference from RGD.
  end for
  $p_{\mathrm{output}} \leftarrow p^{(T)}$
```

As explained in Remark 6, Proposition 1 is the core property of the STO for the ISTA. This motivates us to keep the property in designing the STO for a CHMOO and the Cartan-Hadamard ISTA based on that. Our idea is to shrink the signed distance from each SHP the same way as the STO in RCS does. We can formulate this idea using the SSDM and its inverse if the SSDM is bijective, as follows:

Definition 7 (Cartan-Hadamard STO (CHSTO))**.** Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ be a D-dimensional CHMOO, and assume that $\exp_{\mathrm{o}} : \mathscr{T}_{\mathrm{o}}\mathcal{M} \to \mathcal{M}$ and the SSDM $\boldsymbol{\delta} : \mathcal{M} \to \mathbb{R}^D$ are bijective. We define the ***Cartan-Hadamard STO (CHSTO)*** $\mathcal{T}_{\beta}^{\mathcal{M}} : \mathcal{M} \to \mathcal{M}$ on the CHMOO by $\mathcal{T}_{\beta}^{\mathcal{M}}(p) := \boldsymbol{\delta}^{-1}(\mathcal{T}_{\beta}(\boldsymbol{\delta}(p)))$.

The CHSTO applies the STO on the signed distances from SHPs. By this definition, the CHSTO has the uniform strength property corresponding to Proposition 1.

Theorem 1. *Let $(\mathcal{M}, \mathrm{o}, (e_1, e_2, \dots, e_D))$ be a D-dimensional CHMOO, and assume that $\exp_{\mathrm{o}} : \mathscr{T}_{\mathrm{o}}\mathcal{M} \to \mathcal{M}$ is bijective. Also, we assume that the SSDM $\boldsymbol{\delta}^{\mathcal{M}} : \mathcal{M} \to \mathbb{R}^D$ is bijective. For all $p \in \mathcal{M}$, $\left\|\mathcal{T}_{\beta}^{\mathcal{M}}(p)\right\|_0 < \|p\|_0$ if and only if $\min_{d=1,2,\dots,D} \Delta_{\mathbb{R}^D}\left(p, \mathit{\Pi}_{\hat{d}}\right) \leq \beta$.*

Remark 7. Similar to Proposition 1, Theorem 1 states that the CHSTO has a "uniform" sparsity inducing strength determined only by β everywhere in $\mathcal{M}$. See also Remark 6.

We can define the ***Cartan-Hadamard ISTA (CHISTA)***, the CHMOO version of the ISTA, based on Definition 7. Recall that the Riemannian gradient descent methods, e.g., (Zhang & Sra, 2016), update the point by the exponential map: $p^{(t+1)} \leftarrow \exp_{p^{(t)}}\left(-\alpha \operatorname{grad}_{p^{(t)}} L\right)$, where $J : \mathcal{M} \to \mathbb{R}$ is the objective function, $\operatorname{grad}_{p^{(t)}} J$ is its Riemannian gradient at $p^{(t)}$, and $\alpha \in \mathbb{R}_{>0}$ is the learning rate. This operation is the Riemannian counterpart of the "negative gradient addition" in the gradient descent method in linear space. Replacing the "negative gradient addition" and the STO by the update by the exponential map and the CHSTO, we obtain the CHISTA as in Algorithm 1. As in the ISTA for linear space, the STO's parameter is proportional to $\alpha\lambda$. We have established the sparse learning scheme for a general CHMOO. In the next section, we will derive the specific formulae of the 1-norm and STO for hyperbolic space, which is this paper's primary interest.

6 Sparse learning scheme for hyperbolic space

This section derives specific formulae of the core operations of our sparse learning scheme, the CH 1-norm and CHSTO, for hyperbolic space. We also briefly discuss the product manifold in this section.

As we can see in the previous sections, we can immediately calculate the CH 1-norm and CHSTO if we have the formula of the SSDM. We provide the SSDM formula for hyperbolic space and a product manifold below (The proof is in Appendix C).

Theorem 2. *The SSDM $\boldsymbol{\delta}^{\left(\mathbb{D}^D, \boldsymbol{G}^{P}_{\cdot}\right)} : \mathbb{D}^D \to \mathbb{R}^D$ is bijective. We can calculate the SSDM $\boldsymbol{\delta}^{\left(\mathbb{D}^D, \boldsymbol{G}^{P}_{\cdot}\right)}$ and its inverse maps $\left(\boldsymbol{\delta}^{\left(\mathbb{D}^D, \boldsymbol{G}^{P}_{\cdot}\right)}\right)^{-1} : \mathbb{R}^D \to \mathbb{D}^D$ by*

$$\boldsymbol{\delta}^{\left(\mathbb{D}^D, \boldsymbol{G}^{P}_{\cdot}\right)}(\boldsymbol{p}) = \operatorname{asinh}\left(\frac{2\boldsymbol{p}}{1 - \boldsymbol{p}^\top \boldsymbol{p}}\right), \quad \left(\boldsymbol{\delta}^{\left(\mathbb{D}^D, \boldsymbol{G}^{P}_{\cdot}\right)}\right)^{-1}(\boldsymbol{\sigma}) = \frac{\sinh(\boldsymbol{\sigma})}{\sqrt{1 + (\sinh \boldsymbol{\sigma})^\top (\sinh \boldsymbol{\sigma})} + 1}, \tag{4}$$

whose time and space complexities are $O(D)$.

Proposition 2. *The SSDM of $\boldsymbol{\delta}^{\mathcal{M}}(p)$ of the product CHMOO corresponding to $e_d^{[m]} \in \mathscr{T}_{\mathrm{o}}^{[m]}\mathcal{M} \subset \mathscr{T}_{\mathrm{o}}\mathcal{M}$ is given by $\boldsymbol{\delta}^{\mathcal{M}}(p) = \left[\boldsymbol{\delta}^{\mathcal{M}^{[1]}}\left(p^{[1]}\right) \quad \boldsymbol{\delta}^{\mathcal{M}^{[2]}}\left(p^{[2]}\right) \quad \cdots \quad \boldsymbol{\delta}^{\mathcal{M}^{[M]}}\left(p^{[M]}\right)\right]^\top$.*

Remark 8. The SSDM formula in Theorem 2 is not surprising since point-hyperplane distance formulae have been given (e.g., Cho et al., 2019; Chien et al., 2021). Theorem 2's significance is that it calculates the distance to D SHPs at once in $O(D)$, while the straightforward application of the existing formula requires time complexity $O(D^2)$. The inverse SSDM formula is also novel.

Corollary 1. *The CH 1-norm* $\|\boldsymbol{p}\|_1^{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}$ *of any* $\boldsymbol{p} \in \mathbb{D}^D$*, which we call the* ***hyperbolic 1-norm*** *of* $\boldsymbol{p}$*, and the CHSTO* $\mathcal{T}_{\beta}^{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}$*, which we call the* ***hyperbolic STO (HSTO)****, are given by*

$$\|\boldsymbol{p}\|_1^{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})} = \left\| \operatorname{asinh}\left(\frac{2\boldsymbol{p}}{1 - \boldsymbol{p}^\top \boldsymbol{p}} \right) \right\|_1, \quad \mathcal{T}_{\beta}^{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{p}) = \frac{\sinh(\boldsymbol{\sigma})}{\sqrt{1 + (\sinh \boldsymbol{\sigma})^\top (\sinh \boldsymbol{\sigma})} + 1}, \tag{5}$$

where $\boldsymbol{\sigma} \in \mathbb{R}^D$ *is given by* $\boldsymbol{\sigma} := \operatorname{asinh}\left(\frac{2\boldsymbol{p}}{1 - \boldsymbol{p}^\top \boldsymbol{p}} \right) - \beta \mathbf{1}_D$.

By substituting the CHSTO in Algorithm 1 by HSTO, we obtain the CHISTA for hyperbolic space, which we call the ***hyperbolic ISTA (HISTA)***. Appendix D shows a pseudocode of HISTA.

Remark 9. Since HISTA is a CHISTA, HSTO in HISTA satisfies Theorem 1. It means that we have provided a sparse learning scheme with a uniform sparseness strength in hyperbolic space, which is the goal of our paper.

Lastly, we evaluate the non-uniform strength of the STO applied to the Poincaré model. Let $\epsilon = 10^{-3}$ and $\boldsymbol{p}, \boldsymbol{q} = [0, \epsilon]^\top, [1 - \epsilon, \epsilon]^\top$. The distances from these two points to the 2nd SHP are significantly different: $\Delta_{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{p}, \mathit{\Pi}_2) \approx 0.002, \Delta_{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{q}, \mathit{\Pi}_2) \approx 0.882$. Nevertheless, both points are on the border of the area where the sparsity is increased by $\mathcal{T}_{\epsilon}$. This example shows the non-uniformness of the STO. In contrast, Theorem 1 guarantees that the CHSTO does not cause this issue.

7 Numerical Experiments in Supplementary Materials

We have formulated our sparse learning scheme in hyperbolic space and shown its theoretical advantages, which achieve our motivation. Nevertheless, for readers interested in the empirical behaviors of our sparse learning scheme, we provide numerical experiment results in Supplementary materials. Section F empirically show our HISTA can avoid the oscillation issue as expected. Section G empirically show that our sparse learning scheme can improve both space complexity and representation quality of existing HSBRL. They also empirically clarify when our sparse learning scheme is practically effective and when not.

8 Limitation

One limitation of our sparse learning scheme is that it is limited to Cartan-Hadamard manifolds. Although the discussion on Cartan-Hadamard manifolds is sufficient to deal with Euclidean spaces and hyperbolic spaces, other types of Riemannian manifolds are also used in the representation learning context, such as a torus (Ebisu & Ichise, 2018) and a sphere (Gu et al., 2019). Since a torus and a sphere are symmetric, defining sparsity for each manifold itself is not difficult. Still, defining sparsity for all types of manifolds in an integrated way is possible future work.

Another limitation is the convergence analysis of CHISTA and HISTA. In Euclidean space, the convergence of ISTA has been intensively analyzed as a special case of the proximal gradient method. However, we cannot analyze the convergence of CHISTA and HISTA in the same technique. The convergence analysis of CHISTA and HISTA can be interesting future work.

9 Conclusion and Future Work

This paper has established a novel sparse learning scheme for HSBRL for the first time, based on geometric definitions of the sparsity, regularization term, and the optimization algorithm, with effective closed-form formulae. The HISTA is the fundamental optimization algorithm for sparse learning with uniform sparseness strength and avoiding the oscillation issue. Its extension to algorithms in general Riemannian manifolds and accelerated methods could be interesting and realistic future work. Also, its convergence analyses have not been given, which is a limitation of this paper and can be interesting future work. Although this paper has focused on the theoretical concept, application of HISTA in general machine learning settings, such as hyperbolic neural networks, would be interesting future work.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Milad Alizadeh, Arash Behboodi, Mart van Baalen, Christos Louizos, Tijmen Blankevoort, and Max Welling. Gradient ℓ_1 regularization for quantization robustness. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxK0JBtPr`.

Ivana Balazevic, Carl Allen, and Timothy Hospedales. Multi-relational poincaré graph embeddings. *Advances in Neural Information Processing Systems*, 32, 2019.

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1eiqi09K7`.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. URL `https://transacl.org/ojs/index.php/tacl/article/view/999`.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2013.

Ronald E Bruck Jr. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977.

Antonin Chambolle, Ronald A De Vore, Nam-Yong Lee, and Bradley J Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–335, 1998.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076, 2020.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Eli Chien, Chao Pan, Puoya Tabaghi, and Olgica Milenkovic. Highly scalable and provably accurate classification in poincaré balls. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 61–70. IEEE, 2021.

Hyunghoon Cho, Benjamin DeMeo, Jian Peng, and Bonnie Berger. Large-margin classification in hyperbolic space. In *the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1832–1840. PMLR, 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Joseph M Dale, Liviu Popescu, and Peter D Karp. Machine learning methods for metabolic pathway prediction. *BMC bioinformatics*, 11(1):1–14, 2010.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

David Donoho and Jared Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009a.

David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906): 4273–4293, 2009b.

David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

David L Donoho and Jared Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete & Computational Geometry*, 43(3):522–541, 2010.

Takuma Ebisu and Ryutaro Ichise. TorusE: Knowledge graph embedding on a Lie group. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *the 32nd AAAI Conference on Artificial Intelligence,*, pp. 1819–1826. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16227.

Mário AT Figueiredo and Robert D Nowak. An em algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pp. 1646–1655. PMLR, 2018a.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In Jennifer G. Dy and Andreas Krause (eds.), *the 35th International Conference on Machine Learning*, volume 80 of *Machine Learning Research*, pp. 1632–1641. PMLR, 2018b. URL http://proceedings.mlr.press/v80/ganea18a.html.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *the 32nd Conference on Neural Information Processing Systems*, pp. 5350–5360, 2018c. URL https://proceedings.neurips.cc/paper/2018/hash/dbab2adc8f9d078009ee3fa810bea142-Abstract.html.

Mikhael Gromov. Hyperbolic groups. In *Essays in Group Theory*, pp. 75–263. Springer, 1987.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864. ACM, 2016. doi: 10.1145/2939672.2939754. URL https://doi.org/10.1145/2939672.2939754.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJxeWnCcF7.

Elaine T Hale, Wotao Yin, and Yin Zhang. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. *CAAM TR07-07, Rice University*, 43:44, 2007.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194 (1):371–413, 2022.

Hiroyuki Kasai, Pratik Jawanpuria, and Bamdev Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *International Conference on Machine Learning*, pp. 3262–3271. PMLR, 2019.

S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry, Volume 2*. A Wiley Publication in Applied Statistics. Wiley, 1996. ISBN 9780471157328. URL https://books.google.co.jp/books?id=FjX7vQEACAAJ.

Zhaobin Kuang, Sinong Geng, and David Page. A screening rule for l1-regularized ising model estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

Panagiotis Kyriakis, Iordanis Fostiropoulos, and Paul Bogdan. Learning hyperbolic representations of topological features. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=yqPnIRhHtZv.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *International Conference on Machine Learning*, pp. 3672–3681. PMLR, 2019.

John M Lee. *Introduction to Riemannian manifolds*, volume 176. Springer, 2018.

Baiyang Liu, Lin Yang, Junzhou Huang, Peter Meer, Leiguang Gong, and Casimir Kulikowski. Robust and fast collaborative tracking with two stage sparse optimization. In *European Conference on Computer Vision*, pp. 624–637. Springer, 2010.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 30, 2017a.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *the IEEE International Conference on Computer Vision*, pp. 2736–2744, 2017b.

Abdur Rahman MA Basher and Steven J Hallam. Leveraging heterogeneous network embedding for metabolic pathway prediction. *Bioinformatics*, 37(6):822–829, 2021.

Emile Mathieu, Charline Le Lan, Chris J Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with poincaré variational auto-encoders. *Advances in neural Information Processing Systems*, 32, 2019.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *the 27th Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *the 21st international conference on Machine learning*, pp. 78, 2004.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *the 31st Conference on Neural Information Processing Systems*, pp. 6338–6347, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In Lise Getoor and Tobias Scheffer (eds.), *the 28th International Conference on Machine Learning*, pp. 809–816. Omnipress, 2011. URL https://icml.cc/2011/papers/438_icmlpaper.pdf.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In Dale Schuurmans and Michael P. Wellman (eds.), *the 30th AAAI Conference on Artificial Intelligence*, pp. 1955–1961. AAAI Press, 2016. URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484`.

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

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. ACL, 2014. doi: 10.3115/v1/d14-1162. URL `https://doi.org/10.3115/v1/d14-1162`.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: online learning of social representations. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani (eds.), *the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pp. 701–710. ACM, 2014. doi: 10.1145/2623330.2623732. URL `https://doi.org/10.1145/2623330.2623732`.

Chunhong Qi, Kyle A Gallivan, and P-A Absil. Riemannian bfgs algorithm with applications. In *Recent advances in optimization and its applications in engineering*, pp. 183–192. Springer, 2010.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation extraction with matrix factorization and universal schemas. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 74–84. the Association for Computational Linguistics, 2013. URL `https://www.aclweb.org/anthology/N13-1008/`.

Lev V Sabinin, Ludmila L Sabinina, and Larissa V Sbitneva. On the notion of gyrogroup. *aequationes mathematicae*, 56(1):11–17, 1998.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In Jennifer G. Dy and Andreas Krause (eds.), *the 35th International Conference on Machine Learning*, volume 80 of *Machine Learning Research*, pp. 4457–4466. PMLR, 2018. URL `http://proceedings.mlr.press/v80/sala18a.html`.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l 1 regularized loss minimization. In *the 26th Annual International Conference on Machine Learning*, pp. 929–936, 2009.

Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Ec85b0tUwbA`.

Rishi Sonthalia and Anna Gilbert. Tree! i am no tree! i am a low dimensional hyperbolic embedding. *Advances in Neural Information Processing Systems*, 33:845–856, 2020.

Atsushi Suzuki, Jing Wang, Feng Tian, Atsushi Nitanda, and Kenji Yamanishi. Hyperbolic ordinal embedding. In Wee Sun Lee and Taiji Suzuki (eds.), *the 11th Asian Conference on Machine Learning*, volume 101 of *Machine Learning Research*, pp. 1065–1080. PMLR, 2019. URL `http://proceedings.mlr.press/v101/suzuki19a.html`.

Atsushi Suzuki, Atsushi Nitanda, Jing Wang, Linchuan Xu, Kenji Yamanishi, and Marc Cavazza. Generalization error bound for hyperbolic ordinal embedding. *arXiv preprint arXiv:2105.10475*, 2021a.

Atsushi Suzuki, Atsushi Nitanda, Linchuan Xu, Kenji Yamanishi, Marc Cavazza, et al. Generalization bounds for graph embedding using negative sampling: Linear vs hyperbolic. *Advances in Neural Information Processing Systems*, 34:1243–1255, 2021b.

Puoya Tabaghi and Ivan Dokmanic. Hyperbolic distance matrices. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1728–1738. ACM, 2020. URL `https://dl.acm.org/doi/10.1145/3394486.3403224`.

Masaaki Takada and Hironori Fujisawa. Transfer learning via ℓ_1 regularization. *Advances in Neural Information Processing Systems*, 33:14266–14277, 2020.

Jun Takeuchi, Noriki Nishida, and Hideki Nakayama. Neural networks in a product of hyperbolic spaces. In *the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pp. 211–221, 2022.

Xingwei Tan, Gabriele Pergola, and Yulan He. Extracting event temporal relations via hyperbolic geometry. In *the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8065–8077, 2021.

Jian Tang, Meng Qu, and Qiaozhu Mei. PTE: predictive text embedding through large-scale heterogeneous text networks. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams (eds.), *the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174. ACM, 2015a. doi: 10.1145/2783258.2783307. URL `https://doi.org/10.1145/2783258.2783307`.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: large-scale information network embedding. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (eds.), *the 24th International Conference on World Wide Web, WWW 2015*, pp. 1067–1077. ACM, 2015b. doi: 10.1145/2736277.2741093. URL `https://doi.org/10.1145/2736277.2741093`.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Hyperbolic representation learning for fast and efficient neural question answering. In *the 11th ACM International Conference on Web Search and Data Mining*, pp. 583–591, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Ske5r3AqK7`.

Ryota Tomioka, Taiji Suzuki, and Masashi Sugiyama. Super-linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation. *Journal of Machine Learning Research*, 12(5), 2011.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *the 33nd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2071–2080. JMLR.org, 2016. URL `http://proceedings.mlr.press/v48/trouillon16.html`.

Abraham A Ungar. the holomorphic automorphism group of the complex disk. *aequationes mathematicae*, 47(2):240–254, 1994.

Abraham A Ungar. Extension of the unit disk gyrogroup into the unit ball of any real inner product space. *Journal of Mathematical Analysis and Applications*, 202(3):1040–1057, 1996.

Cédric Vonesch and Michael Unser. A fast iterative thresholding algorithm for wavelet-regularized deconvolution. In *Wavelets XII*, volume 6701, pp. 135–139. SPIE, 2007.

Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Tao Yu and Christopher M De Sa. Numerically accurate hyperbolic embeddings using tiling-based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pp. 1617–1638. PMLR, 2016.

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 29, 2016a.

Yuchen Zhang, Jason D Lee, and Michael I Jordan. l1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pp. 993–1001. PMLR, 2016b.

Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on riemannian manifolds. In *the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 138–147. PMLR, 2019.

Supplementary materials

A Visualization of the SSDM

This section gives a visualization of the SSDM. See Figure 1. Recall that the d-th ***sparse hyperplane*** (SHP), denoted by $\mathit{\Pi}_d^{\mathcal{M}}$, is defined by $\mathit{\Pi}_d^{\mathcal{M}} := \{p \in \mathcal{M} \mid \langle e_d, \log_{\mathrm{o}}(p) \rangle_{\mathrm{o}} = 0\}$. This means that $\mathit{\Pi}_d^{\mathcal{M}} = \exp_{\mathrm{o}}(\operatorname{span}\{e_1, e_2, \dots, e_{d-1}, e_{d+1}, e_{d+2}, \dots, e_D\})$. For example, $\mathit{\Pi}_1^{\mathcal{M}} = \mathit{\Pi}_1$ in Figure 1 is the image under the exponential map $\exp_{\mathrm{o}}$ of a linear subspace spanned by e_2 and e_3. The d-th element $\delta_d^{\mathcal{M}}(p) = \delta_d(p)$ of SSDM measures the signed distance from the d-th SHP to the point p. For example, $\delta_1(p)$ is the signed distance from $\mathit{\Pi}_1$ to p in Figure 1.

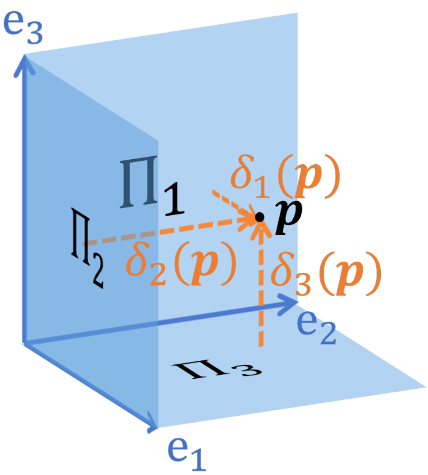

Figure 1: The SSDM's visualization for a 3-dimensional CHMOO case. The hyperplane $\mathit{\Pi}_d^{\mathcal{M}} = \mathit{\Pi}_d$ is the d-th sparse hyperplane (SHP). The d-th element $\delta_d^{\mathcal{M}}(p) = \delta_d(p)$ of SSDM measures the signed distance from the d-th SHP to the point p.

B Detailed explanation of Example 7

The function $f(p) = |p|$ is differentiable at $p \neq 0$ and the derivative is given by $\frac{\mathrm{d}}{\mathrm{d}p} f(p) = \operatorname{sgn}(p)$. Suppose that the learning rate is $\alpha > 0$ and the initial point is $p^{(0)} \neq 0$. By the symmetry about the origin, we can assume that $p^{(0)} > 0$ without loss of generality. Then the gradient descent generates the series $p^{(0)}, p^{(1)}, \dots$ of points according to the following recursion:

$$p^{(t+1)} \leftarrow p^{(t)} - \alpha \frac{\mathrm{d}}{\mathrm{d}p} f(p^{(t)}) = \begin{cases} p^{(t)} - \alpha & \text{if } p^{(t)} > 0, \\ p^{(t)} + \alpha & \text{if } p^{(t)} < 0. \end{cases} \tag{6}$$

Here, we usually set $p^{(t+1)} \leftarrow p^{(t)}$ if $p^{(t)} = 0$, which we can justify as a subgradient method. We can see from (6) that the algorithm ends up oscillating between $p^{(0)} - \alpha n$ and $p^{(0)} - \alpha(n+1)$ unless $p^{(0)}$ is an integral multiple of α, where $n = \left\lfloor \frac{p^{(0)}}{\alpha} \right\rfloor$ is the maximum integer that is no greater than $\frac{p^{(0)}}{\alpha}$.

C Proof of Theorem 2

Proof. We prove for the SSDM $\boldsymbol{\delta}^{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}$. It suffices to prove that the absolute values are correct since the logarithmic map at the origin of the Poincaré model does not change the sign of each element. Let $\boldsymbol{h} \in \mathbb{D}^2$ be the foot of the geodesic pass through $\boldsymbol{p}$ on $\mathit{\Pi}_d$. Note that $\boldsymbol{h}$ is unique according to Gauss-Bonnet theorem. We have that $\boldsymbol{h} = \operatorname{argmin}_{\boldsymbol{q}} \Delta_{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{p}, \boldsymbol{q})$ from hyperbolic Pythagorean theorem. In the following, we regard the ball of the Poincaré model as a unit ball in

Euclidean space and discuss using elementary geometry. A geodesic in hyperbolic space is now an arc orthogonal to the unit ball and $\mathit{\Pi}_d$ and passing through $\boldsymbol{p}$. Define $\boldsymbol{p}' = \frac{\boldsymbol{p}}{\boldsymbol{p}^\top \boldsymbol{p}}$ and $\boldsymbol{h}' = \frac{\boldsymbol{h}}{\boldsymbol{h}^\top \boldsymbol{h}}$. Also denote by $\boldsymbol{m}$ the midpoint of $\boldsymbol{p}$ and $\boldsymbol{p}'$ and by $\boldsymbol{j}$ the midpoint of $\boldsymbol{h}$ and $\boldsymbol{h}'$. Note that $\boldsymbol{j}$ is the center of the arc drawn by the geodesic. Since the arc is orthogonal to the unit ball, it also passes through $\boldsymbol{p}'$ and $\boldsymbol{h}'$ according to the power of a point theorem. The subplane including the arc also contains $\boldsymbol{p}$, $\boldsymbol{h}$, and $\boldsymbol{p}'$. Hence, the following discussion is on the subplane. We regard the axis in the subplane on the intersection of the subplane and $\mathit{\Pi}_d$ as x-axis, and the other axis orthogonal to $\mathit{\Pi}_d$ to y-axis. We indicate the coordinate of the $\boldsymbol{p}$ in the subplane by $\begin{bmatrix} x & y \end{bmatrix}^\top$ and that of $\boldsymbol{h}$ by $\begin{bmatrix} h & 0 \end{bmatrix}^\top$. The coordinates of $\boldsymbol{p}'$ and $\boldsymbol{h}'$ are $\frac{1}{x^2+y^2}\begin{bmatrix} x & y \end{bmatrix}^\top$ and $\begin{bmatrix} 1/h & 0 \end{bmatrix}^\top$, respectively. See also Figure 2. We have that $|\boldsymbol{m}| = \frac{1}{2}\left(\sqrt{x^2+y^2} + \frac{1}{\sqrt{x^2+y^2}}\right)$ and $|\boldsymbol{j}| = \frac{1}{2}\left(h + \frac{1}{h}\right)$. By similarity of two right triangles, we have that $\frac{\sqrt{x^2+y^2}}{x} = \frac{|\boldsymbol{j}|}{|\boldsymbol{m}|}$. Hence, $|\boldsymbol{j}| = \frac{x^2+y^2+1}{2x}$. Noting that $h < 1 < \frac{1}{h}$, we have that $h = \frac{x^2+y^2-\sqrt{(x^2+y^2)^2-4x^2}}{2x}$. We get the expected result by substituting this to the distance formula of the Poincaré model: $\Delta_{(\mathbb{D}^2, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{p}, \boldsymbol{q}) = \operatorname{acosh}\left(1 + \frac{2|\boldsymbol{p}-\boldsymbol{q}|^2}{(1-|\boldsymbol{p}|^2)(1-|\boldsymbol{q}|^2)}\right)$. Specifically,

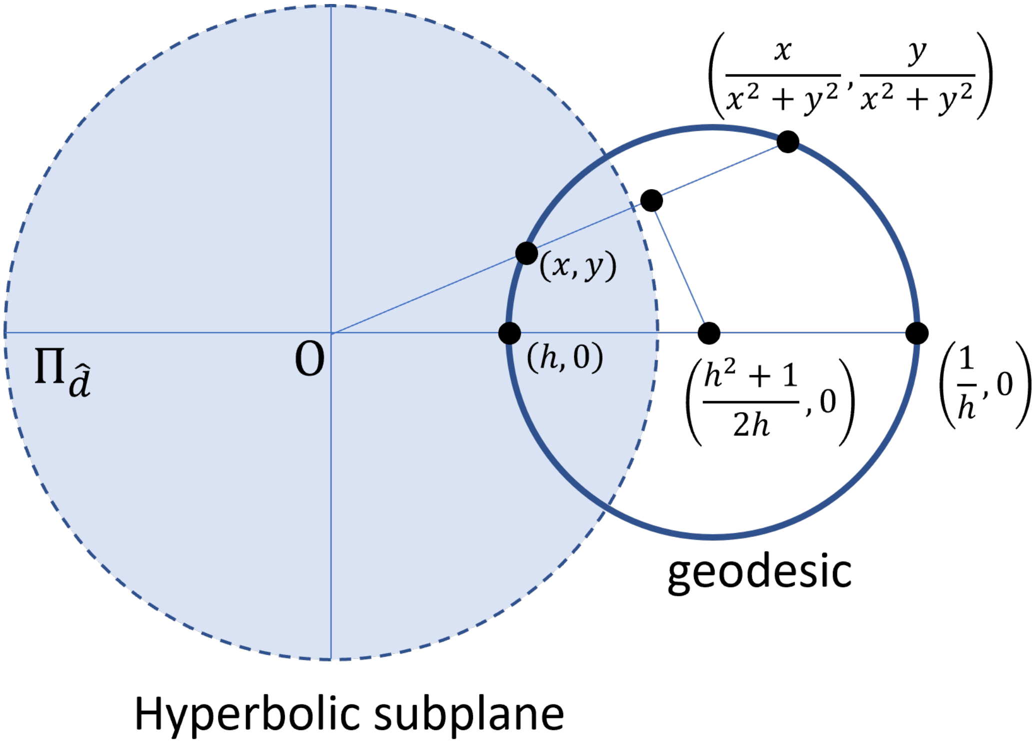
$\left(\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2}\right)$
(x, y)
$\Pi_{\hat{d}}$
O
$(h, 0)$
$\left(\frac{h^2+1}{2h}, 0\right)$
$\left(\frac{1}{h}, 0\right)$
geodesic
Hyperbolic subplane

Figure 2: Hyperbolic subdisk.

$$
\begin{aligned}
&\Delta_{(\mathbb{D}^2, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\begin{bmatrix} x & y \end{bmatrix}, \begin{bmatrix} h & 0 \end{bmatrix}) \\
&= \operatorname{acosh}\left(1 + \frac{2\left((x-h)^2 + y^2\right)}{(1-(x^2+y^2))(1-h^2)}\right) \\
&= \operatorname{acosh}\left(\sqrt{1 + \frac{4y^2}{(1-(x^2+y^2))^2}}\right) \\
&= \operatorname{asinh}\left(\frac{2y}{(1-(x^2+y^2))^2}\right).
\end{aligned}
\tag{7}
$$

We complete the proof by recalling that $y = p_d$ and $(x^2 + y^2) = \boldsymbol{p}^\top \boldsymbol{p}$. □

D Explicit pseudocode of the HISTA

Algorithm 2 shows the explicit form of HISTA on the Poincaré model. Here, sinhc is defined by

$$\operatorname{sinhc}(x) := \begin{cases} 1 & \text{if } x = 0, \\ \frac{\sinh x}{x} & \text{if } x \neq 0. \end{cases} \tag{8}$$

Algorithm 2 HISTA (Explicit form)

```
Require: $\boldsymbol{p}_{\mathrm{init}} \in \mathbb{D}^D$: initial point,
    $\alpha \in \mathbb{R}_{>0}$: learning rate,
    $t \in \mathbb{Z}_{\geq 0}$: # iterations.
Ensure: $\boldsymbol{p}_{\mathrm{output}} \in \mathbb{D}^D$
  $\boldsymbol{p}^{(0)} \leftarrow \boldsymbol{p}_{\mathrm{init}}$
  for $t \leftarrow 1, 2, \ldots, T$ do
    $\boldsymbol{\gamma}^{(t)} \leftarrow \boldsymbol{\partial}\big|_{\boldsymbol{p}^{(t-1)}} J$
    $\rho^{(t)} \leftarrow \frac{4}{\left(1 - \left|\boldsymbol{p}^{(t-1)}\right|^2\right)^2}$
    $\boldsymbol{g}^{(t)} \leftarrow \left(\rho^{(t)}\right)^2 \boldsymbol{\gamma}^{(t)}$
    $\boldsymbol{q}^{(t-1)} \leftarrow \rho^{(t)} \cdot \frac{\left[\cosh\left(\left|-\alpha \boldsymbol{g}^{(t)}\right|\right) - \rho^{(t)} \alpha \left(\boldsymbol{g}^{(t)}\right)^\top \left(\boldsymbol{p}^{(t-1)}\right)\right] \boldsymbol{p}^{(t-1)} + \operatorname{sinhc}\left(\left|\alpha \boldsymbol{g}^{(t)}\right|\right) \boldsymbol{g}^{(t)}}{1 + \left(\rho^{(t)} - 1\right) \cosh\left(\left|-\alpha \boldsymbol{g}^{(t)}\right|\right) - \left(\rho^{(t)}\right)^2 \alpha \left(\boldsymbol{g}^{(t)}\right)^\top \left(\boldsymbol{p}^{(t-1)}\right) \operatorname{sinhc}\left(\left|\alpha \boldsymbol{g}^{(t)}\right|\right)}$
    $\boldsymbol{\sigma}^{(t)} \leftarrow \operatorname{asinh}\left(\frac{2\boldsymbol{p}}{1 - \boldsymbol{p}^\top \boldsymbol{p}}\right) - \alpha \lambda \mathbf{1}_D$
    $\boldsymbol{p}^{(t)} \leftarrow \frac{\sinh\left(\boldsymbol{\sigma}^{(t)}\right)}{\sqrt{1 + \left(\sinh \boldsymbol{\sigma}^{(t)}\right)^\top \left(\sinh \boldsymbol{\sigma}^{(t)}\right)} + 1}$
  end for
  $p_{\mathrm{output}} \leftarrow p^{(T)}$
```

E Formulae for a product manifold.

We discuss the formulae of our sparse representation learning scheme for product manifolds. We do not give detailed proofs, but they can easily be proved using a specific coordinate space. Let $\mathcal{M}^{[1]}, \mathcal{M}^{[2]}, \ldots, \mathcal{M}^{[M]}$ are Riemannian manifolds. The Riemannian product manifold $\mathcal{M} = \mathcal{M}^{[1]} \times \mathcal{M}^{[2]} \times \cdots \times \mathcal{M}^{[M]}$ is given as the topological product manifold of $\mathcal{M}^{[1]}, \mathcal{M}^{[2]}, \ldots, \mathcal{M}^{[M]}$ equipped with the metric tensor defined as follows. For $p = \left(p^{[1]}, p^{[2]}, \ldots p^{[M]}\right) \in \mathcal{M}$, the metric tensor and a direct some decomposition of the tangent space $\mathscr{T}_p\mathcal{M}$ is given as follows. For tangent vectors $v, v' \in \mathscr{T}_p\mathcal{M}$, let $c : I \to \mathcal{M}$ and $c' : I' \to \mathcal{M}$ be C^∞ curves on $\mathcal{M}$ tangent to v and v' respectively, where $I, I' \subset \mathbb{R}$ is open intervals such that $0 \in I \cap I'$. Here, c being tangent to u means that $c(0) = p$ and $v = \dot{c}\big|_0$, where $\dot{c}\big|_t : C^\infty(\mathcal{M}) \to \mathbb{R}$ for $t \in I$ is defined by $\dot{c}\big|_{t'} f := \frac{\mathrm{d}}{\mathrm{d}t}(c(t))\big|_{t'}$. There exist C^∞ curves $c^{[m]} : I \to \mathcal{M}$ and $c'^{[m]} : I' \to \mathcal{M}$ for $m = 1, 2, \ldots, M$ such that $c^{[m]}(t) = c'^{[m]}(t) = p^{[m]}$ for $m = 1, 2, \ldots, M$ and $c(t) = \left(c^{[1]}(t), c^{[2]}(t), \ldots, c^{[M]}(t)\right)$ and $c'(t) = \left(c'^{[1]}(t), c'^{[2]}(t), \ldots, c'^{[M]}(t)\right)$. Then we can define $\tilde{v}^{[m]} := c\big|_0 \in \mathscr{T}_p\mathcal{M}$ and $\tilde{v}'^{[m]} := c'\big|_0 \in \mathscr{T}_p\mathcal{M}$ for $m = 1, 2, \ldots, M$. We can prove that it does not depend on the choice of c. In the following, we denote the linear operation to obtain $\tilde{v}^{[m]} \in \mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]}$ from $v \in \mathscr{T}_p\mathcal{M}$ by $\pi_p^{[m]} : \mathscr{T}_p\mathcal{M} \to \mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]}$. We define the metric tensor on $\mathscr{T}_p\mathcal{M}$ by

$$\langle v, v' \rangle_p^{\mathcal{M}} := \sum_{m=1}^{M} \langle \tilde{v}^{[m]}, \tilde{v}'^{[m]} \rangle_p^{\mathcal{M}}. \tag{9}$$

We can prove that the above metric is symmetric and positive-definite.

A tangent vector $\tilde{v}^{[m]} \in \mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]}$ can be identified with $v^{[m]} \in \mathscr{T}_p\mathcal{M}$ that satisfies

$$\begin{cases} \pi_p^{[m']}\left(v^{[m]}\right) = \tilde{v}^{[m]} & \text{if} \quad m' = m, \\ \pi_p^{[m']}\left(v^{[m]}\right) = 0 & \text{if} \quad m' \neq m. \end{cases} \tag{10}$$

We can prove that the above $v^{[m]}$ is uniquely defined. By the above identification, we can identify each $\mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]}$ with $\mathscr{T}_p^{[m]}\mathcal{M} := \left(\pi_p^{[m]}\right)^{-1}\left(\mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]}\right) \subset \mathscr{T}_p\mathcal{M}$, which is a linear subspace of $\mathscr{T}_p\mathcal{M}$ with the same dimension as $\mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]}$. Also, we can prove that $\mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]} = \bigoplus_{m=1}^{M} \mathscr{T}_p^{[m]}\mathcal{M}$ is the orthogonal direct sum decomposition. Since the restriction $\pi_p^{[m]}\Big|_{\mathscr{T}_p^{[m]}\mathcal{M}}$ is one-to-one, we can define its inverse map $\left(\pi_p^{[m]}\Big|_{\mathscr{T}_p^{[m]}\mathcal{M}}\right)^{-1} : \mathscr{T}_{p^{[m]}}\mathcal{M}^{[m]} \to \mathscr{T}_p^{[m]}\mathcal{M}$.

Based on the identification, if we have multiple CHMOOs $\left(\mathcal{M}^{[m]}, \mathrm{o}^{[m]}, \left(\tilde{e}_1^{[m]}, \tilde{e}_2^{[m]}, \dots, \tilde{e}_{D^{[m]}}^{[m]}\right)\right)_{m=1}^{M}$, we can see that $\left(e_1^{[1]}, e_2^{[1]}, \dots, e_{D^{[1]}}^{[1]}, e_1^{[2]}, e_2^{[2]}, \dots, e_{D^{[2]}}^{[2]}, \dots, e_1^{[M]}, e_2^{[M]}, \dots, e_{D^{[M]}}^{[M]}\right)$ is an ONB in $\mathscr{T}_p\mathcal{M}$, where $e_d^{[m]} = \left(\pi_\mathrm{o}^{[m]}\Big|_{\mathscr{T}_\mathrm{o}^{[m]}\mathcal{M}}\right)^{-1}\left(\tilde{e}_d^{[m]}\right)$ for $m = 1, 2, \dots, M$ and $d = 1, 2, \dots, D^{[m]}$, $\mathcal{M} = \mathcal{M}^{[1]} \times \mathcal{M}^{[2]} \times \dots, \times \mathcal{M}^{[M]}$, and $\mathrm{o} = \left(\mathrm{o}^{[1]}, \mathrm{o}^{[2]}, \dots, \mathrm{o}^{[M]}\right) \in \mathcal{M}$. Hence, we can define the product CHMOO $\left(\mathcal{M}, \mathrm{o}, \left(e_1^{[1]}, e_2^{[1]}, \dots, e_{D^{[1]}}^{[1]}, e_1^{[2]}, e_2^{[2]}, \dots, e_{D^{[2]}}^{[2]}, \dots, e_1^{[M]}, e_2^{[M]}, \dots, e_{D^{[M]}}^{[M]}\right)\right)$.

We are interested in the SHP, Riemannian 0-norm, SSDM, its inverse, CH 1-norm, and CHSTO of the product CHMOO. In the following, we derive the formula for the SSDM, which enable us to calculate the others. To achieve that, we review the basic property of the product Riemannian manifold.

Suppose that $p = \left(p^{[1]}, p^{[2]}, \dots, p^{[M]}\right) \in \mathcal{M}$ and $v \in \mathscr{T}_p\mathcal{M}$ and we consider the unique decomposition $v = \sum_m^M v^{[m]}$, where $v^{[m]} \in \mathscr{T}_p^{[m]}\mathcal{M}$ for $m = 1, 2, \dots, M$. Define $\tilde{v}^{[m]} = \pi_p^{[m]}\Big|_{\mathscr{T}_p^{[m]}\mathcal{M}}\left(v^{[m]}\right)$ for $m = 1, 2, \dots, M$. We can see that if we can define $\exp_{p^{[m]}}\left(\tilde{v}^{[m]}\right) \in \mathcal{M}^{[m]}$ for $m = 1, 2, \dots, M$, it follows that $\exp_p(v) = \left(\exp_{p^{[1]}}\left(\tilde{v}^{[1]}\right), \exp_{p^{[2]}}\left(\tilde{v}^{[2]}\right), \dots, \exp_{p^{[M]}}\left(\tilde{v}^{[M]}\right)\right)$. We can calculate the logarithmic map similarly.

The distance between two points $p = \left(p^{[1]}, p^{[2]}, \dots, p^{[M]}\right)$ and $q = \left(q^{[1]}, q^{[2]}, \dots, q^{[M]}\right)$ is given by

$$\Delta_\mathcal{M}(p, q) = \sqrt{\sum_{m=1}^{M} \left[\Delta_{\mathcal{M}^m}\left(p^{[m]}, q^{[m]}\right)\right]^2}. \tag{11}$$

From the above property, we can confirm that the SHP $\mathit{\Pi}_{[m],d}^{\mathcal{M}}$ of the product CHMOO corresponding to $e_d^{[m]} \in \mathscr{T}_\mathrm{o}^{[m]}\mathcal{M} \subset \mathscr{T}_\mathrm{o}\mathcal{M}$ is given by

$$\mathit{\Pi}_{[m],d}^{\mathcal{M}} = \left\{\left(p^{[1]}, p^{[2]}, \dots, p^{[M]}\right) \middle| p^{[m']} \in \mathit{\Pi}_d^{\mathcal{M}^{[m]}} \text{ if } m' = m, \quad p^{[m']} \in \mathcal{M}^{[m]} \text{ if } m' \neq m, \right\}. \tag{12}$$

Hence, we immediately get the SSDM formula.

$$\boldsymbol{\delta}^\mathcal{M}(p) = \begin{bmatrix} \boldsymbol{\delta}^{\mathcal{M}^{[1]}}\left(p^{[1]}\right) \\ \boldsymbol{\delta}^{\mathcal{M}^{[2]}}\left(p^{[2]}\right) \\ \dots \\ \boldsymbol{\delta}^{\mathcal{M}^{[M]}}\left(p^{[M]}\right) \end{bmatrix}. \tag{13}$$

The above formula enables us to calculate the SHP, Riemannian 0-norm, SSDM, its inverse, CH 1-norm, and CHSTO for the product CHMOO if we can calculate them for each component CHMOO.

For example, we can calculate them for the product of the EVCHMOO and hyperbolic CHMOOs. Note that the product of EVCHMOOs is again a higher dimensional EVCHMOO, while the product of hyperbolic CHMOOs is not a hyperbolic CHMOO since the product of hyperbolic space is not a hyperbolic space.

F Numerical experiments: HISTA avoids the oscillation

Remark 9 states that our motivation in this paper has almost been achieved. The last thing we need to do is to confirm by numerical experiments that our HISTA avoids the oscillation issue. Hence, we compare the HISTA and RGD for sparse solution and non-sparse solution cases.

We consider minimizing the square distance with the hyperbolic 1-norm regularization: $L(\boldsymbol{z}) = \left[\Delta_{(\mathbb{D}^2,\boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{z},\boldsymbol{z}')\right]^2 + \lambda\|\boldsymbol{z}\|_{1,(\mathbb{D}^2,\boldsymbol{G}^{\mathrm{P}}_{\cdot})}$. Here, we set $\boldsymbol{z}' = \begin{bmatrix}0.0 & 0.0\end{bmatrix}^\top, \begin{bmatrix}0.0 & 0.8\end{bmatrix}^\top, \begin{bmatrix}0.4 & 0.8\end{bmatrix}^\top$. We expect that the true solution is sparse for the first two cases and non-sparse for the last case, though we do not know the analytic solution for the latter two. We set $\lambda = 1.0$ and $\alpha = 0.1$ for all cases.

Figure 3 shows that the HISTA outperforms the RGD for $\boldsymbol{z}' = \begin{bmatrix}0.0 & 0.0\end{bmatrix}^\top, \begin{bmatrix}0.0 & 0.4\end{bmatrix}^\top$ in terms of the objective function's value as well as obtaining a sparse solution. For $\boldsymbol{z}' = \begin{bmatrix}0.4 & 0.8\end{bmatrix}^\top$, the RGD can outperform the HISTA. We also observe a "bounce back" effect by the HISTA, which could be a drawback. Still, the HISTA is stable for all the cases, while the oscillation of the RGD is significant for $\boldsymbol{z}' = \begin{bmatrix}0.0 & 0.0\end{bmatrix}^\top$. See also Figure 4. Our results confirm that the superiority of the HISTA over the RGD in the function value for sparse solution cases. Still, detecting the cause of the bounce back effect by the HISTA would be interesting future work.

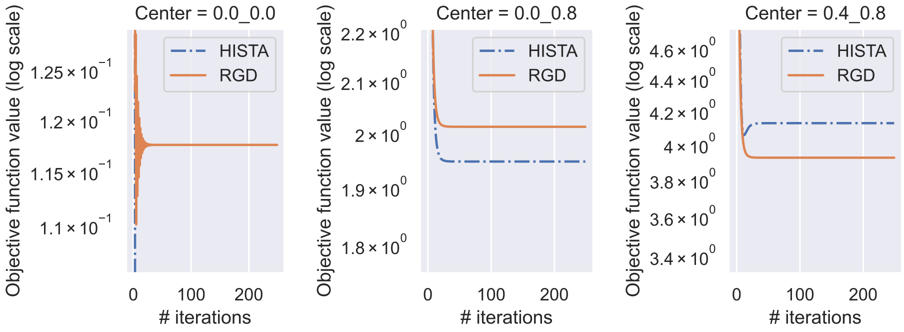

Figure 3: The optimization performances of HISTA and RGD in minimizing the square distance from a fixed point $\boldsymbol{z}'$ with the hyperbolic 1-norm, where $\boldsymbol{z}' = [0.0, 0.0]^\top$ (**Left**), $\boldsymbol{z}' = [0.0, 0.8]^\top$ (**Center**), and $\boldsymbol{z}' = [0.4, 0.8]^\top$ (**Right**). Note that in (**Left**), the blue dashed line indicating HISTA goes infinitely downward because the objective function value reaches 0. See Figure 4 for details.

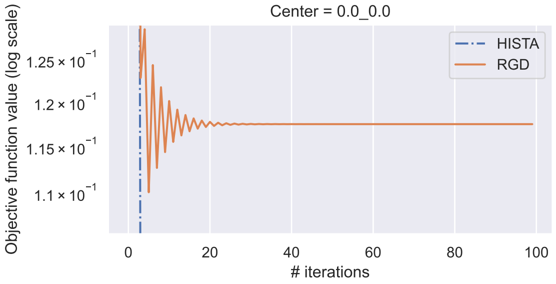

Figure 4: The optimization performances of HISTA and RGD in minimizing the square distance from a fixed point $\boldsymbol{z}'$ with the hyperbolic 1-norm, where $\boldsymbol{z}' = [0.0, 0.0]^\top$. Here we focus on the first 100 iterations to see the oscillation issue of the RGD.

G NUMERICAL EXPERIMENTS ON GRAPH EMBEDDING SETTING

This section gives numerical-experimental results. Note that the objective is not to achieve state-of-the-art representations, but to show how our sparse learning scheme influences HSBRL.

We denote a graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E}$ is the edge set. Since edges are the most fundamental form of information about entity relations, graph embedding has wide applications. Hence, we choose graph embedding as the problem on which we evaluate the performance of our sparse learning scheme. Our objective here is NOT to maximize the quality of representations, but to compare our sparse learning scheme with possible alternatives. Hence, we use the following simplest graph embedding setting. Let $\mathrm{C}(\mathcal{V}, 2)$ be the set of subsets of $\mathcal{V}$, whose size is two. That is, $\mathrm{C}(\mathcal{V}, 2)$ is the set of unordered vertex pairs. Define the label of a vertex pair $y_{u,v} \in \{-1, +1\}$ and the sample weight $w_{u,v} \in \mathbb{R}_{>0}$ for $u, v \in \mathcal{V}$ such that $u \notin v$ by

$$\begin{aligned} y_{u,v} &\coloneqq \begin{cases} +1 & \text{if } \{u, v\} \in \mathcal{E}, \\ -1 & \text{if } \{u, v\} \notin \mathcal{E}, \end{cases} \\ w_{u,v} &\coloneqq 2 \cdot \frac{\left|\{\{u', v'\} \in \mathrm{C}(\mathcal{V}, 2) \mid y_{u',v'} = y_{u,v}\}\right|}{\left|\mathrm{C}(\mathcal{V}, 2)\right|}. \end{aligned} \tag{14}$$

Also, define the 0-1 loss function $l_{\text{0-1}} : \mathbb{R} \times \{-1, +1\} \to \{0, +1\}$ by

$$l_{\text{0-1}}(\hat{y}, y) \coloneqq \begin{cases} 0 & \text{if } \operatorname{sgn}(\hat{y}) = y, \\ +1 & \text{otherwise.} \end{cases} \tag{15}$$

The objective of our experimental setting is to minimize the following balanced 0-1 loss:

$$L_{\text{0-1}}\big((\boldsymbol{z}_v)_{v \in \mathcal{V}}; \mathcal{G}\big) \coloneqq \sum_{\{u,v\} \in \mathrm{C}(\mathcal{V},2)} w_{u,v} l_{0,1}\Big(\big[\Delta_{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{z}_u, \boldsymbol{z}_v)\big]^2 - \theta, y_{u,v}\Big), \tag{16}$$

where the hyperparameter $\theta \in \mathbb{R}_{>0}$ determines the threshold in labeling the pair to be positive or negative.

The above function $L_{\text{0-1}}$ is not easy to optimize since it is not continuous. Hence, in the optimization step, we replace $l_{\text{0-1}}$ by the hinge loss function $l_{\text{hinge}} : \mathbb{R} \times \{-1, +1\} \to \mathbb{R}_{\leq 0}$ defined by $l_{\text{hinge}}(\hat{y}, y) = \max -\hat{y}y + 1, 0$, widely used in machine learning area, e.g., support vector machines (Cortes & Vapnik, 1995). That is, the loss function in the optimization step is

$$L\big((\boldsymbol{z}_v)_{v \in \mathcal{V}}; \mathcal{G}\big) \coloneqq \sum_{\{u,v\} \in \mathrm{C}(\mathcal{V},2)} w_{u,v} l\Big(\big[\Delta_{(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})}(\boldsymbol{z}_u, \boldsymbol{z}_v)\big]^2 - \theta, y_{u,v}\Big). \tag{17}$$

Also, we add the regularization term $\lambda \sum_{v \in \mathcal{V}} r(\boldsymbol{z}_v)$ to the objective function, where the regularization function $r : \mathbb{D}^D \to \mathbb{R}_{\geq 0}$ is the object of the comparison in the experiments and varies for each method. Note that $\lambda \mathbb{R}_{\geq 0}$ determines the regularization strength. To wrap up, we optimize the function $J\big((\boldsymbol{z}_v)_{v \in \mathcal{V}}; \mathcal{G}\big) \coloneqq L\big((\boldsymbol{z}_v)_{v \in \mathcal{V}}; \mathcal{G}\big) + \lambda \sum_{v \in \mathcal{V}} r(\boldsymbol{z}_v)$.

As a regularization function r, we compare the following three:

$$r(\boldsymbol{z}) = \begin{cases} \|\boldsymbol{z}\|_{1,(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})} & \textbf{hyperbolic 1-norm (H 1-norm)}, \\ \|\boldsymbol{z}\|_1 & \textbf{linear 1-norm}, \\ 0 & \textbf{no regularization}. \end{cases} \tag{18}$$

We use the HISTA for the hyperbolic 1-norm. For the linear-norm, we apply Riemannian gradient descent and traditional shrinkage-thresholding operator. Note that this is also what we propose for a baseline. For the no regularization case, these two are the same. Strictly speaking, we need to regard the problem as the optimization of a function of the product of $|\mathcal{V}|$ hyperbolic spaces since we consider $|\mathcal{V}|$ points in hyperbolic space. The rigorous discussion for the product manifold is given in Appendix E. Still, it shows that what we need to do is to calculate a partial derivative for each point, convert it into a Riemannian gradient, and apply the HISTA or RGD for each point, like existing papers do.

We evaluate the balanced accuracy $1 - L_{\text{0-1}}\big((\boldsymbol{z}_v)_{v \in \mathcal{V}}; \mathcal{G}\big)$ and the sum $\sum_{v \in \mathcal{V}} \|\boldsymbol{z}_v\|_{0,(\mathbb{D}^D, \boldsymbol{G}^{\mathrm{P}}_{\cdot})} = \sum_{v \in \mathcal{V}} \|\boldsymbol{z}_v\|_0$ of the 0-norms of the representations. The higher accuracy and lower 0-norm, the

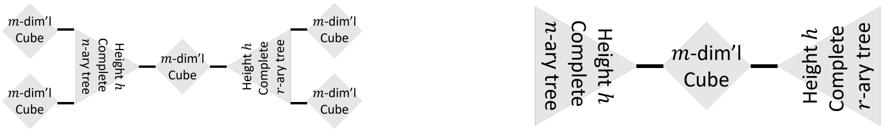
m-dim'l Cube
m-dim'l Cube
Height h Complete n-ary tree
m-dim'l Cube
Height h Complete r-ary tree
m-dim'l Cube
m-dim'l Cube
Height h Complete n-ary tree
m-dim'l Cube
Height h Complete r-ary tree

Figure 5: The datasets' structure. Left: **TRLC**, right: **TRC**.

better, but there is a trade-off between the accuracy and the 0-norm. Specifically, the stronger the regularization is, the lower accuracy and lower 0-norm it gets, and vice versa. Hence, we vary the regularization weight λ and observe how the accuracy and the 0-norm changes. For the no regularization method, we vary D instead of λ. In this case, the lower D is, the lower accuracy and lower 0-norm it gets, and vice versa.

As a graph, we consider tree-like structures that are not completely tree, which are our main focus. To see the difference between the CH 1-norm regularization and Linear 1-norm regularization, we experiment on synthetic datasets defined below.

- TREE-ROOTLEAFCUBES (TRLC) consisting of two complete n-ary trees with height h and five m-dimensional cubes. One cube is in between the roots of the two trees, where each vertex of a hyperbody diagonal pair (a most distant pair) in the cube has an edge to the root of a tree. The other four cubes are connected to a leaf of a tree. Two cubes are connected to one tree and the other two cubes are connected to the other tree. Here, each of the former two cubes have one edge to a leaf of the tree, where the two leaves connected to a cube are most distant to each other. The same holds true for the other tree and the latter two cubes.
- TREE-ROOTCUBES (TRC) is similar to TRLC but without the cube connected to the leaves.

Uniform regularization is needed for TRLC since it has cubes both at the root and around the leaves. Conversely, strong regularization around the boundary and weak regularization could work well for TRC since it has a cube only at the root. Hence, our natural expectation is that CH 1-norm regularization works better for TRLC than Linear 1-norm regularization, but the tendency is not clear for TRC. Figure 5 visualizes these graph structures.

To show our sparse learning scheme's behavior in real applications, we conduct the same experiment on the following real datasets.

- ENRON-EMAIL is an email network reflecting the hierarchical tree structure of the company. At the same time, it also contains edges corresponding to cross-departmental communications, which might be an omen of Enron's bankruptcy in 2001. Since it is a mixture of a tree-like structure and a non-tree-like structure, we expect that our sparse learning scheme works better than the **no regularization** method in the ENRON-EMAIL.
- CORA is a citation network, which shows a highly tree-like structure. Since it is highly tree-like, we expect it to be more effective to apply a (non-sparse) hyperbolic embedding method in low-dimensional space than to apply our sparse learning scheme in high-dimensional space. Hence, we do NOT expect that our sparse learning scheme works so effectively in CORA as in ENRON-EMAIL.

Other experimental settings are as follows. We set $n = 2$, $h = 3$, and $m = 3$. The dimension $D = 6$ is fixed for the two regularization methods, while $D = 2, 3, 4, 5, 6$ for the no regularization method. The regularization strength varies among $\lambda = \{1.0, 2.0, 5.0\} \times 10^{\{-3,-2,-1\}}$ for the two regularization methods. The learning rate that achieved the best accuracy is selected from $\lambda = \{1.0, 2.0, 5.0\} \times 10^{\{1,2,3\}}$. The threshold hyperparameter θ is set to 1.0. The number of iterations is set to $T = 10000$.

Figure 6 shows how the accuracy and the 0-norm changes by varying λ or D. The **closer to the left upper corner**, the better. Here, we show the range where the sum of the 0-norms is no smaller than $2|\mathcal{V}|$; otherwise the mean 0-norm would be lower than two, which would be meaningless as representations. We have also plotted the results of the **H 1-norm** regularization optimized by RGD, which shows that RGD fails to get sparse representations, while shrinkage-thresholding operators succeeded. As we have expected, the **H 1-norm** regularization outperforms the others in TRLC. It shows that our **H 1-norm** regularization can select the dimension of each representations efficiently. In TRC, the linear 1-norm regularization outperforms others around where the sum of the 0-norm is

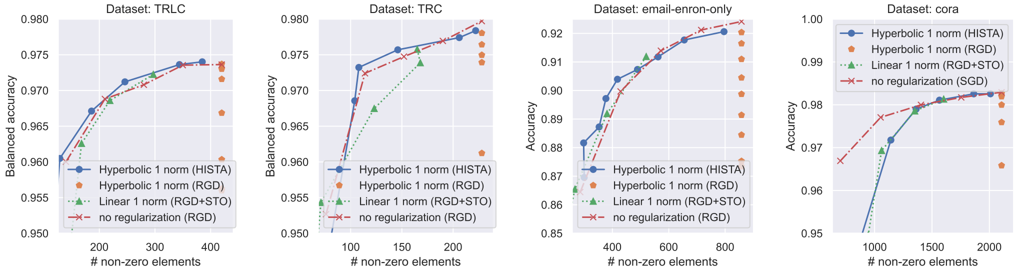

Figure 6: The trade-off between the representation quality (balanced accuracy) and the space complexity (the 0-norm). From left to right: TRLC, TRC, EMAIL-ENRON, **Cora**. The closer to the left upper corner are the graphs, the better.

75, as we have expected. Interestingly, our **H 1-norm** regularization outperforms the **linear 1-norm** regularizations where the sum of the 0-norm is larger. One possible reason is that the **linear 1-norm** regularizations tend to be unstable since the STO changes the representations dramatically around the ball boundary. Although comparing the optimization process is not trivial since they optimize different functions, clarifying the reason for the low performance of the linear regularization would be interesting future work. In ENRON-EMAIL, our **H 1-norm** outperformed other methods both in the balanced accuracy and the sum of 0-norm, when the sum of 0-norm is around 400. While our method's advantage is clear where the sum of 0-norm is small, no method outperforms the remaining in both balanced accuracy and the sum of 0-norm where the sum of 0-norm is large. The investigation of this phenomenon is interesting future work. In CORA, **no regularization** method outperformed both the regularization methods. The reason is that CORA is highly tree-like, so it is most suitable for low-dimensional hyperbolic space. Hence, the result is consistent with our expectations.

H ACRONYM TABLE

To increase readability, we provide the table of acronyms used in this paper in Table 1

RL	representation learning
HSBRL	hyperbolic-space-based representation learning
RGD	Riemannian gradient descent
HISTA	hyperbolic iterative shrinkage-thresholding algorithm
RCS	real coordinate space
CHM	Cartan-Hadamard manifold
CHMOO	CHM with an origin and orthonormal bases
ONB	orthonormal basis
EVCHMOO	Euclidean vector CHMOO
SHP	sparse hyperplane
SSDM	signed SHP distance map
CH	Cartan-Hadamard
H	hyperbolic
ISTA	iterative shrinkage-thresholding algorithm
STO	soft-thresholding operator
CHSTO	Cartan-Hadamard STO
CHISTA	Cartan-Hadamard ISTA

Table 1: Acronyms