Hyperbolic Sliced-Wasserstein via Geodesic and Horospherical Projections

Clément Bonet [1] **Laetitia Chapel** [2] **Lucas Drumetz** [3] **Nicolas Courty** [2]

Abstract

Hyperbolic space embeddings have been shown beneficial for many learning tasks where data have an underlying hierarchical structure. Consequently, many machine learning tools were extended to such spaces, but only few discrepancies to compare probability distributions defined over those spaces exist. Among the possible candidates, optimal transport distances are well defined on such Riemannian manifolds and enjoy strong theoretical properties, but suffer from high computational cost. On Euclidean spaces, sliced-Wasserstein distances, which leverage a closed-form solution of the Wasserstein distance in one dimension, are more computationally efficient, but are not readily available on hyperbolic spaces. In this work, we propose to derive novel hyperbolic sliced-Wasserstein discrepancies. These constructions use projections on the underlying geodesics either along horospheres or geodesics. We study and compare them on different tasks where hyperbolic representations are relevant, such as sampling or image classification.

1. Introduction

In recent years, hyperbolic spaces have received a lot of attention in machine learning (ML) as they allow efficiently processing data that present a hierarchical structure (Nickel & Kiela, 2017; 2018). This encompasses data such as graphs (Gupte et al., 2011), words (Tifrea et al., 2018) or images (Khrulkov et al., 2020). Embedding in hyperbolic spaces has been proposed for various applications such as drug embedding (Yu et al., 2020), image clustering (Park et al., 2021; Ghadimi Atigh et al., 2021), zero-shot recognition (Liu et al., 2020), remote sensing (Hamzaoui et al., 2021) or reinforcement learning (Cetin et al., 2022). Hence, many works proposed to develop tools to be used on such spaces, such as generalization of Gaussian distributions (Nagano et al., 2019; Galaz-Garcia et al., 2022), neural networks (Ganea et al., 2018b; Liu et al., 2019) or normalizing flows (Lou et al., 2020; Bose et al., 2020).

Optimal Transport (OT) (Villani, 2003; 2009) is a popular tool used in ML to compare probability distributions. Among others, it has been used for domain adaptation (Courty et al., 2016), learning generative models (Arjovsky et al., 2017) or document classification (Kusner et al., 2015). However, the main tool of OT is the Wasserstein distance which exhibits an expensive, super-cubical computational cost *w.r.t.* the number of samples of each distribution. Hence, many workarounds have been proposed to alleviate the computational burden such as entropic regularization (Cuturi, 2013), minibatch OT (Fatras et al., 2020) or the sliced-Wasserstein (SW) distance (Rabin et al., 2011). In particular, SW is a popular variant of the Wasserstein distance that computes the expected distance between one dimensional projections on some lines of the two distributions. Its computational advantages and theoretical properties make it an efficient and popular alternative to the Wasserstein distance. For example, it has been used for texture synthesis (Heitz et al., 2021) or for generative modeling with SW autoencoders (Kolouri et al., 2018), SW GANs (Deshpande et al., 2018), SW flows (Liutkus et al., 2019) or SW gradient flows (Bonet et al., 2022).

The theoretical study of the Wasserstein distance on Riemannian manifolds is well developed (McCann, 2001; Villani, 2009). When it comes to hyperbolic spaces, some optimal transport attempts aimed at aligning distributions of data which have been embedded in a hyperbolic space (Alvarez-Melis et al., 2020; Hoyos-Idrobo, 2020). Regarding SW, it is originally defined using Euclidean distances and projections, which are not well suited to other manifolds. Recently, Rustamov & Majumdar (2020) proposed to defined a SW distance on compact manifolds using the eigendecomposition of the Laplace-Beltrami operator while Bonet et al. (2023) proposed a SW distance to tackle this problem for measures supported on the sphere by using only objects intrinsically defined on this specific manifold. Contrary to the elliptical geometry of the sphere, the negative curvature of hyperbolic spaces calls for drastically different strategies to define geodesics and the associated

[1]Université Bretagne Sud, LMBA [2]Université Bretagne Sud, IRISA [3]IMT Atlantique, Lab-STICC. Correspondence to: Clément Bonet <clement.bonet@univ-ubs.fr>.

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

projection operators. This work proposes to close this gap by proposing new SW constructions on these spaces.

Contributions. We extend sliced-Wasserstein to data living in hyperbolic spaces. Analogously to Euclidean SW, we project the distributions on geodesics passing through the origin. Interestingly enough, different projections can be considered, leading to several new SW constructions that exhibit different theoretical properties and empirical benefits. We make connections with Radon transforms already defined in the literature and we show that hyperbolic SW are (pseudo-) distances. We provide the algorithmic procedure and discuss its complexity. We illustrate the benefits of these new hyperbolic SW distances on several tasks such as sampling or image classification.

2. Background

In this Section, we first provide some background on Optimal Transport with the Wasserstein and the sliced-Wasserstein distance. We then review two common hyperbolic models, namely the Lorentz and Poincaré ball models, on which we will define new OT discrepancies in the next section.

2.1. Optimal Transport

Optimal transport is a popular field which allows comparing distributions of probabilities by determining a transport plan minimizing some ground cost. The main tool of OT is the Wasserstein distance which we introduce now.

Wasserstein Distance on Riemannian Manifolds. Let M be a Riemannian manifold endowed with a Riemannian distance d. For $p \geq 1$, the p-Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{P}_p(M) = \{\mu \in \mathcal{P}(M),\ \int_M d(x, x_0)^p \,\mathrm{d}\mu(x) < \infty \text{ for any } x_0 \in M\}$ is defined as

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{M \times M} d(x, y)^p \,\mathrm{d}\gamma(x, y), \qquad (1)$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(M \times M),\ \pi^1_\# \gamma = \mu,\ \pi^2_\# \gamma = \nu\}$ is the set of couplings, $\pi^1(x, y) = x$, $\pi^2(x, y) = y$ and $\#$ is the pushforward operator defined as, for all borelian $A \subset M$, $T_\# \mu(A) = \mu(T^{-1}(A))$. For more details about OT, we refer to (Villani, 2009).

The main bottleneck of the Wasserstein distance is its computational complexity. Indeed, for two discrete probability measures with n samples, it can be solved using linear programs (Peyré et al., 2019) with a complexity of $O(n^3 \log n)$, which prevents its use when large amount of data are at stake. Hence, a whole literature consists at deriving alternative OT metrics with a smaller computational cost.

Sliced-Wasserstein Distance on Euclidean Space. On Euclidean spaces, a popular proxy of the Wasserstein distance is the so-called sliced-Wasserstein distance. On the real line, for $p \geq 1$, the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ admits the following closed-form (Peyré et al., 2019, Remark 2.30) :

$$W_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \,\mathrm{d}u \qquad (2)$$

where F_μ^{-1} and F_ν^{-1} denote the quantile functions of μ and ν. This can be approximated in practice very efficiently as it only requires to sort the samples, which has a complexity of $O(n \log n)$. Therefore, Rabin et al. (2011) defined the sliced-Wasserstein distance by projecting linearly the probabilities on all the possible directions. For a direction $\theta \in S^{d-1}$, denote, for all $x \in \mathbb{R}^d$, $P^\theta(x) = \langle x, \theta \rangle$ the projection in direction θ, and λ the uniform measure on S^{d-1}. Then, the SW distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

$$SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(P^\theta_\# \mu, P^\theta_\# \nu) \,\mathrm{d}\lambda(\theta). \qquad (3)$$

Using a Monte-Carlo approximation, this can be approximated in $O(Ln(d + \log n))$ where L is the number of projections and n the number of samples.

Moreover, the slicing process has many appealing properties, such as having a sample complexity independent of the dimension (Nadjahi et al., 2020), being topologically equivalent to Wasserstein (Bonnotte, 2013) and being an actual distance. For the latter point, it can be shown to be a pseudo-distance using that W_p is a distance. The indiscernible property relies on the link between the projection used in SW and the Radon transform (Bonneel et al., 2015; Kolouri et al., 2019) which is injective on the space of measures (Boman & Lindskog, 2009, Theorem A). More precisely, let $f \in L^1(\mathbb{R}^d)$, then its Radon transform $R : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times S^{d-1})$ is defined for $t \in \mathbb{R}$, $\theta \in S^{d-1}$ as,

$$Rf(t, \theta) = \int_{\mathbb{R}^d} f(x) \mathbb{1}_{\{\langle x, \theta \rangle = t\}} \,\mathrm{d}x. \qquad (4)$$

This transform admits a dual operator $R^* : C_0(\mathbb{R} \times S^{d-1}) \to C_0(\mathbb{R}^d)$, with $C_0(\mathbb{R} \times S^{d-1})$ the set of continuous functions that vanish at infinity, such that for all $g \in C_0(\mathbb{R} \times S^{d-1})$, $\langle Rf, g \rangle_{\mathbb{R} \times S^{d-1}} = \langle f, R^* g \rangle_{\mathbb{R}^d}$ (Bonneel et al., 2015). This allows defining the Radon transform of a measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ as the measure $R\mu \in \mathcal{M}(\mathbb{R} \times S^{d-1})$ satisfying for all $g \in C_0(\mathbb{R} \times S^{d-1})$, $\langle R\mu, g \rangle_{\mathbb{R} \times S^{d-1}} = \langle \mu, R^* g \rangle_{\mathbb{R}^d}$ (Boman & Lindskog, 2009). Then, it was shown in (Bonneel et al., 2015) that, by denoting by $(R\mu)^\theta$ the disintegration *w.r.t.* to the uniform distribution on S^{d-1},

$$SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p\big((R\mu)^\theta, (R\nu)^\theta\big) \,\mathrm{d}\lambda(\theta). \qquad (5)$$

Therefore, $SW_p^p(\mu,\nu) = 0$ implies that, for λ-ae θ, $(R\mu)^\theta = (R\nu)^\theta$, which implies that $\mu = \nu$ by injectivity of the Radon transform on measures.

Many variants of this distance were recently proposed. Most lines of work considered different subspaces for projecting the data: hypersurfaces (Kolouri et al., 2019), Hilbert curves (Li et al., 2022) or subspace of higher dimensions (Lin et al., 2020; 2021). When it comes to data living on Riemannian manifolds, Rustamov & Majumdar (2020) defined a variant on compact manifolds and Bonet et al. (2023) extended SW for spherical data.

2.2. Hyperbolic Spaces

Hyperbolic spaces are Riemannian manifolds of negative constant curvature (Lee, 2006). They have received recently a surge of interest in machine learning as they allow embedding efficiently data with a hierarchical structure (Nickel & Kiela, 2017; 2018). A thorough review of the recent use of hyperbolic spaces in machine learning can be found in (Peng et al., 2021).

There are five usual parameterizations of a hyperbolic manifold (Peng et al., 2021). They are equivalent (isometric) and one can easily switch from one formulation to the other. Hence, in practice, we use the one which is the most convenient, either given the formulae to derive or the numerical properties. In machine learning, the two most used models are the Poincaré ball and the Lorentz model (also known as the hyperboloid model). Each of these models has its own advantages compared to the other. For example, the Lorentz model has a distance which behaves better *w.r.t.* numerical issues compared to the distance of the Poincaré ball. However, the Lorentz model is unbounded, contrary to the Poincaré ball. We introduce in the following these two models as we will use both of them in our work.

Lorentz model. First, we introduce the Lorentz model $\mathbb{L}^d \subset \mathbb{R}^{d+1}$ of a d-dimensional hyperbolic space. It can be defined as

$$\mathbb{L}^d = \{(x_0,\dots,x_{d+1}) \in \mathbb{R}^d,\ \langle x,x\rangle_{\mathbb{L}} = -1, x_0 > 0\} \quad (6)$$

where

$$\forall x,y \in \mathbb{R}^{d+1},\ \langle x,y\rangle_{\mathbb{L}} = -x_0y_0 + \sum_{i=1}^d x_iy_i \quad (7)$$

is the Minkowski pseudo inner-product (Boumal, 2022, Chapter 7). The Lorentz model can be seen as the upper sheet of a two-sheet hyperboloid. In the following, we will denote $x^0 = (1,0,\dots,0) \in \mathbb{L}^d$ the origin of the hyperboloid. The geodesic distance in this manifold, which denotes the length of the shortest path between two points, can be defined as

$$\forall x,y \in \mathbb{L}^d,\ d_{\mathbb{L}}(x,y) = \operatorname{arccosh}(-\langle x,y\rangle_{\mathbb{L}}). \quad (8)$$

At any point $x \in \mathbb{L}^d$, we can associate a subspace of $\mathbb{R}^{d+1}$ orthogonal in the sense of the Minkowski inner product. These spaces are called tangent spaces and are described formally as $T_x\mathbb{L}^d = \{v \in \mathbb{R}^{d+1},\ \langle v,x\rangle_{\mathbb{L}} = 0\}$. Note that on tangent spaces, the Minkowski inner-product is a real inner product. In particular, on $T_{x^0}\mathbb{L}^d$, it is the usual Euclidean inner product, *i.e.* for $u,v \in T_{x^0}\mathbb{L}^d$, $\langle u,v\rangle_{\mathbb{L}} = \langle u,v\rangle$. Moreover, for all $v \in T_{x^0}\mathbb{L}^d$, $v_0 = 0$.

We can draw a connection with the sphere. Indeed, by endowing $\mathbb{R}^{d+1}$ with $\langle\cdot,\cdot\rangle_{\mathbb{L}}$, we obtain $\mathbb{R}^{1,d}$ the so-called Minkowski space. Then, $\mathbb{L}^d$ is the analog in the Minkowski space of the sphere S^d in the regular Euclidean space (Brisdson & Haefliger, 2013).

Poincaré ball. The second model of hyperbolic space we will be interested in is the Poincaré ball $\mathbb{B}^d \subset \mathbb{R}^d$. This space can be obtained as the stereographic projection of each point $x \in \mathbb{L}^d$ onto the hyperplane $\{x \in \mathbb{R}^{d+1},\ x_0 = 0\}$. More precisely, the Poincaré ball is defined as

$$\mathbb{B}^d = \{x \in \mathbb{R}^d,\ \|x\|_2 < 1\}, \quad (9)$$

with geodesic distance, for all $x,y \in \mathbb{B}^d$,

$$d_{\mathbb{B}}(x,y) = \operatorname{arccosh}\left(1 + 2\frac{\|x-y\|_2^2}{(1-\|x\|_2^2)(1-\|y\|_2^2)}\right). \quad (10)$$

We see on this formulation that the distance can be subject to numerical instabilities when one of the points is too close to the boundary of the ball.

We can switch from Lorentz to Poincaré using the following isometric projection (Nickel & Kiela, 2018):

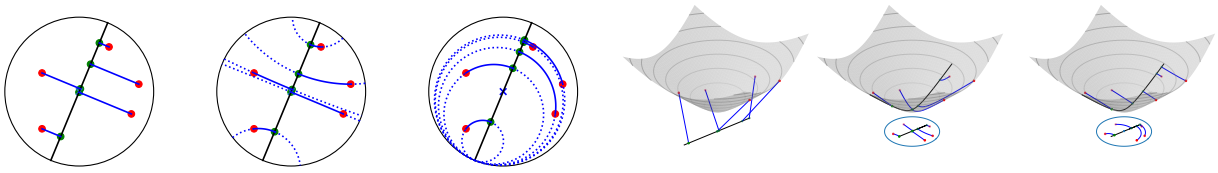
$$\forall x \in \mathbb{L}^d,\ P_{\mathbb{L}\to\mathbb{B}}(x) = \frac{1}{1+x_0}(x_1,\dots,x_d) \quad (11)$$

and from Poincaré to Lorentz by

$$\forall x \in \mathbb{B}^d,\ P_{\mathbb{B}\to\mathbb{L}}(x) = \frac{1}{1-\|x\|_2^2}(1+\|x\|_2^2, 2x_1,\dots,2x_d). \quad (12)$$

3. Hyperbolic Sliced-Wasserstein Distances

In this work, we aim at introducing sliced-Wasserstein type of distances on hyperbolic spaces. Interestingly enough, several constructions can be performed, depending on the projections that are involved. The first solution we consider is the extension of Euclidean SW between distributions whose support lies on hyperbolic spaces. We also provide variants that involve a geodesic cost. To do so, we first define the subspace on which the Wasserstein distance can be efficiently computed and then provide two different projection operators: geodesic and horospherical. We finally define the related hyperbolic sliced-Wasserstein distances and discuss some of their properties. All the proofs are reported in Appendix A.



(a) Euclidean. (b) Geodesics. (c) Horospheres. (d) Euclidean. (e) Geodesics. (f) Horospheres.

Figure 1: Projection of (red) points on a geodesic (black line) in the Poincaré ball and in the Lorentz model along Euclidean lines, geodesics or horospheres (in blue). Projected points on the geodesic are in green.

3.1. Euclidean Sliced-Wasserstein on Hyperbolic Spaces

The support of distributions lying on hyperbolic space are included in the ambient spaces $\mathbb{R}^d$ (Poincaré ball) or $\mathbb{R}^{d+1}$ (Lorentz model). As such, Euclidean SW can be used for such kind of data. On the Poincaré ball, the projections lie onto the manifold as geodesics passing through the origin are straight lines (see Section 3.2), but the initial geometry of the data might not be fully taken care of as the orthogonal projection does not respect the Poincaré geodesics. On the Lorentz model though, the projections lie out of the manifold. We will denote SWp and SWl the Poincaré ball and Lorentz model version. These formulations allow inheriting from the properties of SW, such as being a distance.

3.2. Projection Set and Wasserstein Distance

To generalize the sliced-Wasserstein distance on other spaces, we first define on which subspace to project. Euclidean spaces can be seen as Riemannian manifolds of null constant curvature whose geodesics are straight lines. Therefore, analogously to the Euclidean space , we project on geodesics passing through the origin. We now describe geodesics in the Lorentz model and in the Poincaré ball.

Geodesics. In the Lorentz model, geodesics passing through the origin x^0 can be obtained by taking the intersection between $\mathbb{L}^d$ and a 2-dimensional plane containing x^0 (Lee, 2006, Proposition 5.14). Any such plane can be obtained as $\text{span}(x^0, v)$ where $v \in T_{x^0}\mathbb{L}^d \cap S^d = \{v \in S^d,\ v_0 = 0\}$. The corresponding geodesic can be described by a geodesic line (Bridson & Haefliger, 2013, Corollary 2.8), *i.e.* a map $\gamma : \mathbb{R} \to \mathbb{L}^d$ satisfying for all $t, s \in \mathbb{R}$, $d_{\mathbb{L}}(\gamma(s), \gamma(t)) = |t - s|$, of the form

$$\forall t \in \mathbb{R},\ \gamma(t) = \exp_{x^0}(tv) = \cosh(t)x^0 + \sinh(t)v. \quad (13)$$

On the Poincaré ball, geodesics are circular arcs perpendicular to the boundary S^{d-1} (Lee, 2006, Proposition 5.14). In particular, geodesics passing through the origin are straight lines. Hence, they can be characterized by a point $\tilde{v}$ on the border S^{d-1}. Such points will be called ideal points.

Wasserstein distance on geodesics. In order to have an efficient way to compute the discrepancy, we need a practical way to compute the Wasserstein distance on geodesics. As the distance between any point on a geodesic line γ and the origin can take arbitrary values on $\mathbb{R}_+$, we project points from the geodesic to the real line $\mathbb{R}$. Indeed, on $\mathbb{R}$, there exists a well known closed-form (see Section 2.1) that can be efficiently computed in practice. In the Lorentz model, let $v \in T_{x^0}\mathbb{L}^d \cap S^d$ be a direction such that $\gamma(\mathbb{R}) = \mathbb{L}^d \cap \text{span}(x^0, v)$. Then, we propose to project a point $x \in \gamma(\mathbb{R})$ using

$$t^v_{\mathbb{L}}(x) = \text{sign}(\langle x, v\rangle) d_{\mathbb{L}}(x, x^0). \quad (14)$$

The scalar product with v gives an orientation to the geodesic, and the distance to the origin the coordinate of x. We can do the same on the Poincaré ball with $t^{\tilde{v}}_{\mathbb{B}}(x) = \text{sign}(\langle x, \tilde{v}\rangle) d_{\mathbb{B}}(x, 0)$, where $\tilde{v}$ is one of the ideal point to which the geodesic is perpendicular. In the remainder, we will remove the subscripts $\mathbb{L}$ and $\mathbb{B}$ when it is clear from the context. Finally, we need to check that this projection keeps the geodesic Wasserstein distance unchanged. We formulate the following proposition in the Lorentz model.

Proposition 3.1 (Wasserstein distance on geodesics.)**.** *Let $v \in T_{x^0}\mathbb{L}^d \cap S^d$ and $\mathcal{G} = \text{span}(x^0, v) \cap \mathbb{L}^d$ a geodesic passing through x^0. Then, for $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{G})$,*

$$\begin{aligned} W_p^p(\mu, \nu) &= W_p^p(t^v_{\#}\mu, t^v_{\#}\nu) \\ &= \int_0^1 |F^{-1}_{t^v_{\#}\mu}(u) - F^{-1}_{t^v_{\#}\nu}(u)|^p \, \mathrm{d}u. \end{aligned} \quad (15)$$

The last ingredient of hyperbolic SW is the way the points lying in the manifold are projected onto the geodesic. We introduce here two different projections that are illustrated on Figure 1.

3.3. Hyperbolic Sliced-Wasserstein

With geodesic projections. We discuss here the results in the Lorentz model, but we can also obtain all the results in the Poincaré ball. Let $v \in T_{x^0} \cap S^d$ and $\mathcal{G}^v = \{\exp_{x^0}(tv),\ t \in \mathbb{R}\}$ a geodesic passing through x^0. As a first generalization of the sliced-Wasserstein distance on hyperbolic spaces, we propose to use the geodesic projection

$\tilde{P}^v$, which projects points on $\mathcal{G}^v$ following the shortest path (geodesics), and which is defined as

$$\forall x \in \mathbb{L}^d,\ \tilde{P}^v(x) = \operatorname*{argmin}_{y \in \mathcal{G}^v} d(x, y). \tag{16}$$

We report in Appendix A.2 the closed-form formulas on both the Lorentz model and the Poincaré ball. Here, we are mostly interested into the coordinate on $\mathbb{R}$, which can be obtained either by computing $t^v \circ \tilde{P}^v$, or as

$$\forall x \in \mathbb{L}^d,\ P^v(x) = \operatorname*{argmin}_{t \in \mathbb{R}} d_{\mathbb{L}}\big(\exp_{x^0}(tv), x\big). \tag{17}$$

Regarding the implementation, we derive a closed-form in the following proposition.

Proposition 3.2 (Coordinate of the geodesic projection)**.**

1. *Let $\mathcal{G}^v = \operatorname{span}(x^0, v) \cap \mathbb{L}^d$ where $v \in T_{x^0}\mathbb{L}^d \cap S^d$. Then, the coordinate P^v of the geodesic projection on $\mathcal{G}^v$ of $x \in \mathbb{L}^d$ is*

$$P^v(x) = \operatorname{arctanh}\left(-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right). \tag{18}$$

2. *Let $\tilde{v} \in S^{d-1}$ be an ideal point. Then, the coordinate $P^{\tilde{v}}$ of the geodesic projection on the geodesic characterized by $\tilde{v}$ of $x \in \mathbb{B}^d$ is*

$$P^{\tilde{v}}(x) = 2\operatorname{arctanh}\big(s(x)\big), \tag{19}$$

where

$$s(x) = \begin{cases} \frac{1+\|x\|_2^2 - \sqrt{(1+\|x\|_2^2)^2 - 4\langle x, \tilde{v}\rangle^2}}{2\langle x, \tilde{v}\rangle} & \text{if } \langle x, \tilde{v}\rangle \neq 0 \\ 0 & \text{if } \langle x, \tilde{v}\rangle = 0. \end{cases} \tag{20}$$

Now, we have all the tools to define the geodesic hyperbolic sliced-Wasserstein discrepancy (GHSW) between $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$ as, for $p \geq 1$,

$$GHSW_p^p(\mu, \nu) = \int_{T_{x^0}\mathbb{L}^d \cap S^d} W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)\, \mathrm{d}\lambda(v). \tag{21}$$

Note that $T_{x^0}\mathbb{L}^d \cap S^d \cong S^{d-1}$ and that v can be drawn by first sampling $\tilde{v} \sim \mathrm{Unif}(S^{d-1})$ and then adding a 0 in the first coordinate, *i.e.* $v = (0, \tilde{v})$ with $\tilde{v} \in S^{d-1}$. Note also that $GHSW_p(\mu, \nu) < \infty$ for $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$. We also have the Poincaré formulation using $P^{\tilde{v}}$, and defined between $\mu, \nu \in \mathcal{P}(\mathbb{B}^d)$ as

$$GHSW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(P^{\tilde{v}}_{\#}\mu, P^{\tilde{v}}_{\#}\nu)\, \mathrm{d}\lambda(\tilde{v}). \tag{22}$$

With horospherical projections. As we saw in Section 2.1, the projection on geodesics in the Euclidean space is obtained by taking the inner product. A first viewpoint is to see it as the geodesic projection of $x \in \mathbb{R}^d$ on the geodesic $\operatorname{span}(\theta)$:

$$\langle x, \theta\rangle\theta = \operatorname*{argmin}_{y \in \operatorname{span}(\theta)} \|x - y\|_2. \tag{23}$$

In this case, using a similar projection as (14), the coordinates on the line are obtained as the inner product:

$$t^\theta(x) = \operatorname{sign}(\langle x, \theta\rangle)\|\langle x, \theta\rangle\theta - 0\|_2 = \langle x, \theta\rangle. \tag{24}$$

However, the inner product $\langle x, \theta\rangle$ can actually also be seen directly as a coordinate on the line $\operatorname{span}(\theta)$. This can be translated by the Busemann function on unit-speed geodesics, which can be generalized on certain Riemannian manifolds. More precisely, the Busemann function associated to the geodesic ray γ, *i.e.* a geodesic from $\mathbb{R}_+$ to the manifold satisfying $d(\gamma(t), \gamma(s)) = |t - s|$, is defined as (Bridson & Haefliger, 2013, Definition 8.17)

$$B^\gamma(x) = \lim_{t \to \infty} \big(d(x, \gamma(t)) - t\big), \tag{25}$$

where x belongs to the corresponding manifold and d is the geodesic distance. It can be checked that on Euclidean spaces, $B^{\operatorname{span}(\theta)}(x) = -\langle x, \theta\rangle$. While the Busemann function is not well defined on positively curved spaces such as the sphere (as geodesics are periodic), closed-form are available on hyperbolic spaces and provide different projections. We report them in the next proposition. As we only work with geodesics passing through the origin, we put as indices the directions which fully characterize them (either $v \in T_{x^0}\mathbb{L}^d$ in $\mathbb{L}^d$, or $\tilde{v} \in S^{d-1}$ in $\mathbb{B}^d$).

Proposition 3.3 (Busemann function on hyperbolic space)**.**

1. *On $\mathbb{L}^d$, for any direction $v \in T_{x^0}\mathbb{L}^d \cap S^d$,*

$$\forall x \in \mathbb{L}^d,\ B^v(x) = \log(-\langle x, x^0 + v\rangle_{\mathbb{L}}). \tag{26}$$

2. *On $\mathbb{B}^d$, for any ideal point $\tilde{v} \in S^{d-1}$,*

$$\forall x \in \mathbb{B}^d,\ B^{\tilde{v}}(x) = \log\left(\frac{\|\tilde{v} - x\|_2^2}{1 - \|x\|_2^2}\right). \tag{27}$$

To conserve Busemann coordinates, it has been proposed by Chami et al. (2021) to project points on a subset following the level sets of the Busemann function. Those level sets are known as horospheres, which can be seen as spheres of infinite radius (Izumiya, 2009). In the Poincaré ball, a horosphere is a Euclidean sphere tangent to an ideal point. Chami et al. (2021) argued that this projection is beneficial against the geodesic projection as it tends to better preserve the distances. This motivates us to project on geodesics following the level sets of the Busemann function in order to conserve the Busemann coordinates, *i.e.* we want to have $B^{\tilde{v}}(x) = B^{\tilde{v}}(P^{\tilde{v}}(x))$ (resp. $B^v(x) = B^v(P^v(x))$) on the Poincaré ball (resp. Lorentz model) where $\tilde{v} \in S^{d-1}$ (resp.

$v \in T_{x^0}\mathbb{L}^d \cap S^d$) is characterizing the geodesic. We report the closed-forms in Appendix A.5. In practice, noting that $B^\gamma(x) = B^\gamma(\gamma(t)) = -t$, we obtain that the coordinate is $t = -B^\gamma(x)$.

Using the projections along the horospheres, we can define a new hyperbolic sliced-Wasserstein discrepancy, called horospherical, between $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$ as, for $p \geq 1$,

$$HHSW_p^p(\mu, \nu) = \int_{T_{x^0}\mathbb{L}^d \cap S^d} W_p^p(B^v_\# \mu, B^v_\# \nu) \, \mathrm{d}\lambda(v). \tag{28}$$

Note that $HHSW_p(\mu, \nu) < \infty$ for $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$ (see Appendix B.1). We also provide a formulation on the Poincaré ball between $\mu, \nu \in \mathcal{P}_p(\mathbb{B}^d)$, using $B^{\tilde{v}}$, as

$$HHSW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(B^{\tilde{v}}_\# \mu, B^{\tilde{v}}_\# \nu) \, \mathrm{d}\lambda(\tilde{v}). \tag{29}$$

Using that the projections formula between $\mathbb{L}^d$ and $\mathbb{B}^d$ are isometries, we show in the next proposition that the two formulations are equivalent. Hence, we choose in practice the formulation which is the more suitable, either from the nature of data or from a numerical stability viewpoint.

Proposition 3.4. *For $p \geq 1$, let $\mu, \nu \in \mathcal{P}_p(\mathbb{B}^d)$ and denote $\tilde{\mu} = (P_{\mathbb{B}\to\mathbb{L}})_\# \mu$, $\tilde{\nu} = (P_{\mathbb{B}\to\mathbb{L}})_\# \nu$. Then,*

$$HHSW_p^p(\mu, \nu) = HHSW_p^p(\tilde{\mu}, \tilde{\nu}), \tag{30}$$
$$GHSW_p^p(\mu, \nu) = GHSW_p^p(\tilde{\mu}, \tilde{\nu}). \tag{31}$$

3.4. Properties

It can easily be showed that GHSW and HHSW are pseudo-distances as it only depends on the distance properties of the Wasserstein distance. Whether or not they satisfy the indiscernible property remains an open question. As described in the introduction for SW, we can derive the corresponding Radon transform. More precisely, we can show that

$$GHSW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p\big((\bar{R}\mu)^v, (\bar{R}\nu)^v\big) \, \mathrm{d}\lambda(v), \tag{32}$$

where $\bar{R}$ is the hyperbolical Radon transform, first introduced by Helgason (1959) and more recently studied *e.g.* in (Berenstein & Rubin, 1999; 2004; Rubin, 2002). We can also show a similar relation between HHSW and the horospherical Radon transform studied *e.g.* by Bray & Rubin (2019); Casadio Tarabusi & Picardello (2021). If these transforms are injective on the space of measures, then we would have that GHSW or HHSW are distances. However, to the best of our knowledge, the injectivity of such transforms on the space of measures has not been studied yet. We detail the derivations in Appendix B.2.

We also provide in Appendix B.3 the sample complexity and the projection complexity. We note that the results are similar as in the Euclidean case (Nadjahi et al., 2020), *i.e.* the sample complexity is independent of the dimension and the projection complexity converges in $O(1/\sqrt{L})$ with L the number of projections.

Algorithm 1 Guideline of GHSW

Input: $(x_i)_{i=1}^n \sim \mu$, $(y_j)_{j=1}^n \sim \nu$, $(\alpha_i)_{i=1}^n$, $(\beta_j)_{j=1}^n \in \Delta_n$, L the number of projections, p the order
for $\ell = 1$ **to** L **do**
  Draw $\tilde{v} \sim \mathrm{Unif}(S^{d-1})$, let $v = [0, \tilde{v}]$
  $\forall i, j,\ \hat{x}_i^\ell = P^v(x_i)$, $\hat{y}_j^\ell = P^v(y_j)$
  Compute $W_p^p(\sum_{i=1}^n \alpha_i \delta_{\hat{x}_i^\ell}, \sum_{j=1}^n \beta_j \delta_{\hat{y}_j^\ell})$
end for
Return $\frac{1}{L}\sum_{\ell=1}^L W_p^p(\sum_{i=1}^n \alpha_i \delta_{\hat{x}_i^\ell}, \sum_{j=1}^n \beta_j \delta_{\hat{y}_j^\ell})$

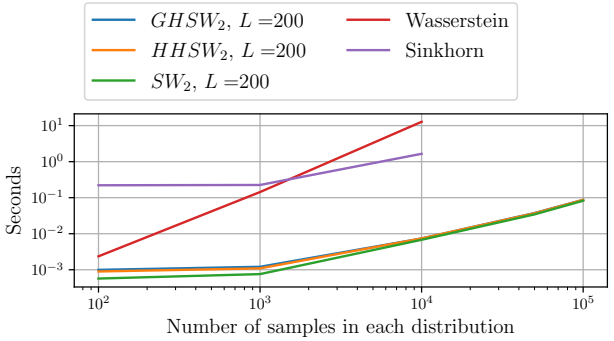


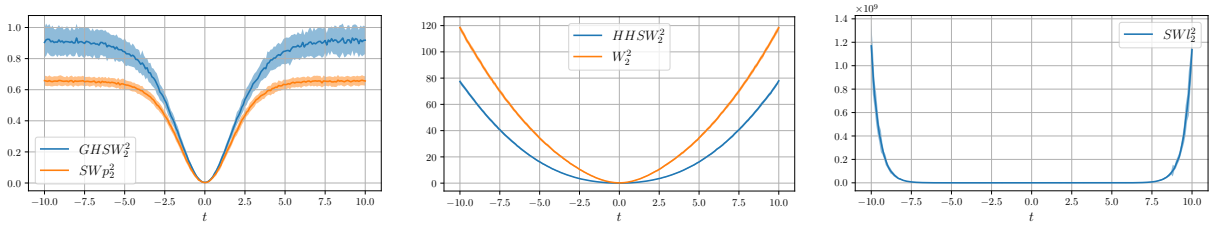
Figure 2: Runtime comparison in log-log scale between Wasserstein and Sinkhorn using the geodesic distance, SW_2, $GHSW_2$ and $HHSW_2$ with 200 projections, including the computation time of the cost matrices.

4. Implementation

In this Section, we discuss the implementation of GHSW and HHSW, as well as their complexity.

Implementation. In practice, we only have access to discrete distributions $\hat{\mu}_n = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\hat{\nu}_n = \sum_{i=1}^n \beta_i \delta_{y_i}$ where $(x_i)_i$ and $(y_i)_i$ are sample locations in hyperbolic space, and $(\alpha_i)_i$ and $(\beta_i)_i$ belong to the simplex $\Delta_n = \{\alpha \in [0,1]^n,\ \sum_{i=1}^n \alpha_i = 1\}$. We approximate the integral by a Monte-Carlo approximation by drawing a finite number L of projection directions $(v_\ell)_{\ell=1}^L$ in S^{d-1}. Then, computing GHSW and HHSW amount at first getting the coordinates on $\mathbb{R}$ by using the corresponding projections, and computing the 1D Wasserstein distance between them. We summarize the procedure in Algorithm 1 for GHSW.

Complexity. For both GHSW and HHSW, the projection procedure has a complexity of $O(nd)$. Hence, for L projections, the complexity is in $O(Ln(d + \log n))$ which is the same as for SW. In Figure 2, we compare the runtime between GHSW, HHSW, SW, Wasser-



(a) SW on Poincaré (SWp), GHSW (b) HHSW and Wasserstein (c) SW on Lorentz (SWl)

Figure 3: Comparison of the Wasserstein distance (with the geodesic distance as cost), GHSW, HHSW and SW between Wrapped Normal distributions. We gather the discrepancies together by scale of the values. SW on the Poincaré model has very small values as it operates on the unit ball, while on the Lorentz model, it can take very high values. GHSW returns small values as the geodesic projections tend to project the points close to the origin. HHSW has values which are closer to the geodesic Wasserstein distance as the horospherical projection tends to better keep the distance between points.

stein and Sinkhorn with geodesic distances in $\mathbb{L}^2$ for $n \in \{10^2, 10^3, 10^4, 5 \cdot 10^4, 10^5\}$ samples which are drawn from wrapped normal distributions (Nagano et al., 2019), and $L = 200$ projections. We used the POT library (Flamary et al., 2021) to compute SW, Wasserstein and Sinkhorn. We observe the quasi-linearity complexity of GHSW and HHSW. When we only have a few samples, the cost of the projection is higher than computing the 1D Wasserstein distance, and SW is the fastest.

5. Application

In this Section, we perform several experiments which aim at comparing GHSW, HHSW, SWp and SWl. First, we study the evolution of the different distances between wrapped normal distributions which move along geodesics. Then, we illustrate the ability to fit distributions on $\mathbb{L}^2$ using gradient flows. Finally, we use HHSW and GHSW for an image classification problem where they are used to fit a prior in the embedding space. We add more informations about distributions and optimization in hyperbolic spaces in Appendix C. Complete details of the experimental settings are reported in Appendix D. We also report in Appendix D.4 preliminary experiments on autoencoders with hierarchical latent priors.

Comparisons of the Different Hyperbolical SW Discrepancies. On Figure 3, we compare the evolutions of GHSW, HHSW, SW and Wasserstein with the geodesic distance between Wrapped Normal Distributions (WNDs), where one is centered and the other moves along a geodesic. More precisely, by denoting $\mathcal{G}(\mu, \Sigma)$ a WND, we plot the evolution of the distances between $\mathcal{G}(x^0, I_2)$ and $\mathcal{G}(x_t, I_2)$ where $x_t = \cosh(t)x^0 + \sinh(t)v$ for $t \in [-10, 10]$ and $v \in T_{x^0}\mathbb{L}^2 \cap S^2$. We observe first that SW on the Lorentz model explodes when the two distributions are getting far from each other. Then, we observe that $HHSW_2$ has values with a scale similar to W_2. We argue that it comes from the observation of Chami et al. (2021) which stated that

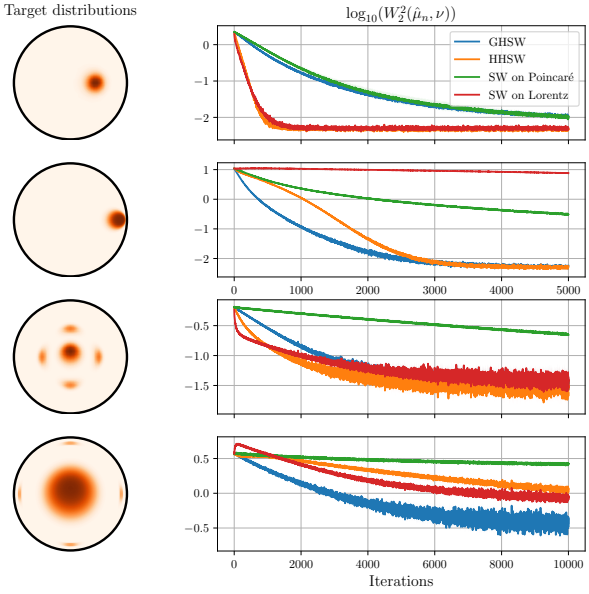


Figure 4: Log 2-Wasserstein between a target and the gradient flow of GHSW, HHSW and SW (averaged over 5 runs).

the horospherical projection better preserves the distance between points compared to the geodesic projection. As SWp operates on the unit ball using Euclidean distances, the distances are very small, even for distributions close to the border. Interestingly, as geodesic projections tend to project points close to the origin, GHSW tends also to squeeze the distance between distributions far from the origin. This might reduce numerical instabilities when getting far from the origin, especially in the Lorentz model. This experiments also allows to observe that, at least for WNDs, the indiscernible property is observed in practice as we only obtain one minimum when both measures coincide. Hence, it suggests that GHSW and HHSW are proper distances.

Gradient Flows. We now assess the ability to learn distributions by minimizing the hyperbolic SW discrepancies

(HSW). We suppose that we have a target distribution ν from which we have access to samples $(x_i)_{i=1}^n$. Therefore, we aim at learning ν by solving the following optimization problem: $\min_\mu \; HSW_2^2\big(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{x_i}\big)$. We model μ as a set of $n = 500$ particles and propose to perform a Riemannian gradient descent (Boumal, 2022) to learn the distribution.

To compare the dynamics of the different discrepancies, we plot on Figure 4 the evolution of the exact log 2-Wasserstein distance, with geodesic distance as ground cost, between the learned distribution at each iteration and the target, with the same learning rate. We use as targets wrapped normal distributions and mixtures of WNDs. For each type of target, we consider two settings, one in which the distribution is close to the origin and another in which the distribution lies closer to the border. We observe different behaviors in the two settings. When the target is lying close to the origin, SWl and HHSW, which present the biggest magnitude, are the fastest to converge. As for distant distributions however, GHSW converges the fastest. Moreover, SWl suffers from many numerical instabilities, as the projections of the gradients do not necessarily lie on the tangent space when points are too far of the origin. This requires to lower the learning rate, and hence to slow down the convergence. Interestingly, SWp is the slowest to converge in both settings.

Deep Classification with Prototypes. We now turn to a classification use case with real world data. Let $\{(x_i, y_i)_{i=1}^n\}$ be a training set where $x_i \in \mathbb{R}^m$ and $y_i \in \{1, \dots, C\}$ denotes a label. Ghadimi Atigh et al. (2021) perform classification on the Poincaré ball by assigning to each class $c \in \{1, \dots, C\}$ a prototype $p_c \in S^{d-1}$, and then by learning an embedding on the hyperbolic space using a neural network f_θ followed by the exponential map. Then, by denoting by $z = \exp_0\big(f_\theta(x)\big)$ the output, the loss to be minimized is, for a regularization parameter $s \geq 0$,

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^n \Big(B^{p_{y_i}}\big(z_i\big) - sd \cdot \log\big(1 - \|z_i\|_2^2\big)\Big). \quad (33)$$

The first term is the Busemann function which will draw the representations of x_i towards the prototype assigned to the class y_i, while the second term penalizes the overconfidence and pulls back the representation towards the origin. Ghadimi Atigh et al. (2021) showed that the second term can be decisive to improve the accuracy. Then, the classification of an input is done by solving $y^* = \operatorname{argmax}_c \langle \frac{z}{\|z\|}, p_c \rangle$.

We propose to replace the second term by a global prior on the distribution of the representations. More precisely, we add a discrepancy D between the distribution $(\exp_0 \circ f_\theta)_\# p_X$, where p_X denotes the distribution of the training set, and a mixture of C WNDs where the centers are chosen as $(\alpha p_c)_{c=1}^C$, with $(p_c)_c$ the prototypes and $0 < \alpha < 1$. In practice, we use $D = GHSW_2^2$, $D = HHSW_2^2$, $D = SWp_2^2$ and $D = SWl_2^2$ to assess their usability on a real problem and compared with W_2^2 and MMD with Laplacian kernel (Feragen et al., 2015). Let $(w_i)_{i=1}^n$ be a batch of points drawn from this mixture, then the loss we minimize is

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^n B^{p_i}(z_i) + \lambda D\left(\frac{1}{n}\sum_{i=1}^n \delta_{z_i}, \frac{1}{n}\sum_{i=1}^n \delta_{w_i}\right). \quad (34)$$

Table 1: Test Accuracy on deep classification with prototypes (best performance in bold)

	CIFAR10		CIFAR100		
Dimensions	2	4	3	5	10
PeBuse	$90.64_{\pm 0.06}$	$90.59_{\pm 0.11}$	$49.28_{\pm 1.95}$	$53.44_{\pm 0.76}$	$59.19_{\pm 0.39}$
GHSW	$91.39_{\pm 0.23}$	$91.66_{\pm 0.27}$	$\mathbf{53.97}_{\pm 1.35}$	$60.64_{\pm 0.87}$	$61.45_{\pm 0.41}$
HHSW	$91.28_{\pm 0.26}$	$\mathbf{91.98}_{\pm 0.05}$	$53.88_{\pm 0.06}$	$\mathbf{60.69}_{\pm 0.25}$	$\mathbf{62.80}_{\pm 0.09}$
SWp	$\mathbf{91.84}_{\pm 0.31}$	$91.68_{\pm 0.10}$	$53.25_{\pm 3.27}$	$59.77_{\pm 0.81}$	$60.36_{\pm 1.26}$
SWl	$91.13_{\pm 0.14}$	$91.74_{\pm 0.12}$	$53.88_{\pm 0.02}$	$60.62_{\pm 0.39}$	$62.30_{\pm 0.23}$
W	$91.67_{\pm 0.18}$	$91.83_{\pm 0.21}$	$50.07_{\pm 4.58}$	$57.49_{\pm 0.94}$	$58.82_{\pm 1.66}$
MMD	$91.47_{\pm 0.10}$	$91.68_{\pm 0.09}$	$50.59_{\pm 4.44}$	$58.10_{\pm 0.73}$	$58.91_{\pm 0.91}$

On Table 1, we report the classification accuracy on the test set for CIFAR10 and CIFAR100 (Krizhevsky, 2009), using the exact same setting as (Ghadimi Atigh et al., 2021). We rerun their method, called PeBuse here. We report results averaged over 3 runs. We observe that the proposed penalization outperforms the original method for all the different dimensions.

6. Conclusion and Discussion

In this work, we propose different sliced-Wasserstein discrepancies between distributions lying in hyperbolic spaces. In particular, we introduce two new SW discrepancies which are intrinsically defined on hyperbolic spaces. They are built by first identifying a closed-form for the Wasserstein distance on geodesics, and then by using different projections on the geodesics. We compare these metrics on multiple tasks such as sampling and image classification. We observe that, while Euclidean SW in the ambient space still works, it suffers from either slow convergence on the Poincaré ball or numerical instabilities on the Lorentz model when distributions are lying far from the origin. On the other hand, geodesic versions exhibit the same complexity and converge generally better for gradient flows. Further works will look into other tasks where hyperbolic embeddings and distributions have been showed to be beneficial, such as persistent diagrams (Carriere et al., 2017; Kyriakis et al., 2021). Besides further applications, proving that these discrepancies are indeed distances, and deriving statistical results are interesting directions of work. One might also consider different subspaces on which to project, such as horocycles which are circles of infinite radius and which can be seen as another analog object to lines in hyperbolic spaces (Casadio Tarabusi & Picardello, 2021).

Acknowledgements

This research was funded by project DynaLearn from Labex CominLabs and Region Bretagne ARED DLearnMe, and by the project OTTOPIA ANR-20-CHIA-0030 of the French National Research Agency (ANR).

References

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

Alvarez-Melis, D., Mroueh, Y., and Jaakkola, T. Unsupervised hierarchy matching with optimal transport over hyperbolic spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 1606–1617. PMLR, 2020.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Berenstein, C. A. and Rubin, B. Radon transform of lp-functions on the lobachevsky space and hyperbolic wavelet transforms. 1999.

Berenstein, C. A. and Rubin, B. Totally geodesic radon transform of l p-functions on real hyperbolic space. In *Fourier analysis and convexity*, pp. 37–58. Springer, 2004.

Boman, J. and Lindskog, F. Support theorems for the radon transform and cramér-wold theorems. *Journal of theoretical probability*, 22(3):683–710, 2009.

Bonet, C., Courty, N., Septier, F., and Drumetz, L. Efficient gradient flows in sliced-wasserstein space. *Transactions on Machine Learning Research*, 2022.

Bonet, C., Berg, P., Courty, N., Septier, F., Drumetz, L., and Pham, M.-T. Spherical sliced-wasserstein. In *International Conference on Learning Representations*, 2023.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Bose, J., Smofsky, A., Liao, R., Panangaden, P., and Hamilton, W. Latent variable modelling with hyperbolic normalizing flows. In *International Conference on Machine Learning*, pp. 1045–1055. PMLR, 2020.

Boumal, N. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Apr 2022. URL http://www.nicolasboumal.net/book.

Bray, W. and Rubin, B. Radon transforms over lower-dimensional horospheres in real hyperbolic space. *Transactions of the American Mathematical Society*, 372(2): 1091–1112, 2019.

Bray, W. O. and Rubin, B. Inversion of the horocycle transform on real hyperbolic spaces via a wavelet-like transform. In *Analysis of divergence*, pp. 87–105. Springer, 1999.

Bridson, M. R. and Haefliger, A. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Carriere, M., Cuturi, M., and Oudot, S. Sliced wasserstein kernel for persistence diagrams. In *International conference on machine learning*, pp. 664–673. PMLR, 2017.

Casadio Tarabusi, E. and Picardello, M. A. Radon transforms in hyperbolic spaces and their discrete counterparts. *Complex Analysis and Operator Theory*, 15(1): 1–40, 2021.

Cetin, E., Chamberlain, B., Bronstein, M., and Hunt, J. J. Hyperbolic deep reinforcement learning. *arXiv preprint arXiv:2210.01542*, 2022.

Chami, I., Gu, A., Nguyen, D. P., and Ré, C. Horopca: Hyperbolic dimensionality reduction via horospherical projections. In *International Conference on Machine Learning*, pp. 1419–1429. PMLR, 2021.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein : asymptotic and gradient properties. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2131–2141. PMLR, 26–28 Aug

2020. URL https://proceedings.mlr.press/v108/fatras20a.html.

Feragen, A., Lauze, F., and Hauberg, S. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.

Fournier, N. and Guillin, A. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Galaz-Garcia, F., Papamichalis, M., Turnbull, K., Lunagomez, S., and Airoldi, E. Wrapped distributions on homogeneous riemannian manifolds. *arXiv preprint arXiv:2204.09790*, 2022.

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

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018b.

Gelfand, I. M., Graev, M. I., and Vilenkin, N. I. *Generalized Functions-Volume 5. Integral Geometry and Representation Theory*. Academic Press, 1966.

Ghadimi Atigh, M., Keller-Ressel, M., and Mettes, P. Hyperbolic busemann learning with ideal prototypes. *Advances in Neural Information Processing Systems*, 34:103–115, 2021.

Gupte, M., Shankar, P., Li, J., Muthukrishnan, S., and Iftode, L. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World wide web*, pp. 557–566, 2011.

Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J. (eds.), *Proceedings of the 7th Python in Science Conference*, pp. 11 – 15, Pasadena, CA USA, 2008.

Hamzaoui, M., Chapel, L., Pham, M.-T., and Lefèvre, S. Hyperbolic variational auto-encoder for remote sensing scene classification. In *ORASIS 2021*, 2021.

Heitz, E., Vanhoey, K., Chambon, T., and Belcour, L. A sliced wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9412–9420, 2021.

Helgason, S. Differential operators on homogeneous spaces. *Acta mathematica*, 102(3-4):239–299, 1959.

Hoyos-Idrobo, A. Aligning hyperbolic representations: an optimal transport-based approach. *arXiv preprint arXiv:2012.01089*, 2020.

Izumiya, S. Horospherical geometry in the hyperbolic space. In *Noncommutativity and Singularities: Proceedings of French–Japanese symposia held at IHÉS in 2006*, volume 55, pp. 31–50. Mathematical Society of Japan, 2009.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images, 2009.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Kyriakis, P., Fostiropoulos, I., and Bogdan, P. Learning hyperbolic representations of topological features. *arXiv preprint arXiv:2103.09273*, 2021.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Lee, J. M. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

Li, T., Meng, C., Yu, J., and Xu, H. Hilbert curve projection distance for distribution comparison. *arXiv preprint arXiv:2205.15059*, 2022.

Lin, T., Fan, C., Ho, N., Cuturi, M., and Jordan, M. Projection robust wasserstein distance and riemannian optimization. *Advances in neural information processing systems*, 33:9383–9397, 2020.

Lin, T., Zheng, Z., Chen, E., Cuturi, M., and Jordan, M. I. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pp. 262–270. PMLR, 2021.

Liu, Q., Nickel, M., and Kiela, D. Hyperbolic graph neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Liu, S., Chen, J., Pan, L., Ngo, C.-W., Chua, T.-S., and Jiang, Y.-G. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9273–9281, 2020.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pp. 4104–4113. PMLR, 2019.

Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S. N., and De Sa, C. M. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems*, 33:17548–17558, 2020.

Mathieu, E., Le Lan, C., Maddison, C. J., Tomioka, R., and Teh, Y. W. Continuous hierarchical representations with poincaré variational auto-encoders. *Advances in neural information processing systems*, 32, 2019.

McCann, R. J. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11 (3):589–608, 2001.

Mettes, P., van der Pol, E., and Snoek, C. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pp. 4693–4702. PMLR, 2019.

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

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

Ovinnikov, I. Poincaré wasserstein autoencoder. *arXiv preprint arXiv:1901.01427*, 2019.

Park, J., Cho, J., Chang, H. J., and Choi, J. Y. Unsupervised hyperbolic representation learning via message passing auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5516–5526, 2021.

Paty, F.-P. and Cuturi, M. Subspace robust wasserstein distances. In *International conference on machine learning*, pp. 5072–5081. PMLR, 2019.

Peng, W., Varanka, T., Mostafa, A., Shi, H., and Zhao, G. Hyperbolic deep neural networks: A survey. *arXiv preprint arXiv:2101.04562*, 2021.

Pennec, X. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

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

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Rakotomamonjy, A., Alaya, M. Z., Berar, M., and Gasso, G. Statistical and topological properties of gaussian smoothed sliced probability divergences. *arXiv preprint arXiv:2110.10524*, 2021.

Rubin, B. Radon, cosine and sine transforms on real hyperbolic space. *Advances in Mathematics*, 170(2):206–223, 2002.

Rustamov, R. M. and Majumdar, S. Intrinsic sliced wasserstein distances for comparing collections of probability distributions on manifolds and graphs. *arXiv preprint arXiv:2010.15285*, 2020.

Said, S., Bombrun, L., and Berthoumieu, Y. New riemannian priors on the univariate normal model. *Entropy*, 16 (7):4015–4031, 2014.

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

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

Tifrea, A., Bécigneul, G., and Ganea, O.-E. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wilson, B. and Leimeister, M. Gradient descent in hyperbolic space. *arXiv preprint arXiv:1805.08207*, 2018.

Yu, K., Visweswaran, S., and Batmanghelich, K. Semi-supervised hierarchical drug embedding in hyperbolic space. *Journal of chemical information and modeling*, 60 (12):5647–5657, 2020.

A. Proofs

A.1. Proof of Proposition 3.1

Let γ a geodesic on $\mathbb{L}^d$ passing through x^0 and with direction $v \in T_{x^0}\mathbb{L}^d \cap S^d$, *i.e.* the geodesic is obtained as $\text{span}(x^0, v) \cap \mathbb{L}^d$. Let μ, ν probability measures on γ.

First, we need to show that for all $x, y \in \text{span}(x^0, v) \cap \mathbb{L}^d$,

$$d_{\mathbb{L}}(x, y) = |t^v(x) - t^v(y)|, \tag{35}$$

i.e. that t^v is an isometry from $\text{span}(x^0, v) \cap \mathbb{L}^d$ to $\mathbb{R}$.

As x and y belong to the geodesic γ, there exist $s, t \in \mathbb{R}$ such that

$$x = \gamma(s) = \cosh(s)x^0 + \sinh(s)v, \tag{36}$$

and

$$y = \gamma(t) = \cosh(t)x^0 + \sinh(t)v. \tag{37}$$

Then, on one hand, we have

$$\begin{aligned} d_{\mathbb{L}}(\gamma(s), \gamma(t)) &= \text{arccosh}(-\langle \gamma(s), \gamma(t) \rangle_{\mathbb{L}}) \\ &= \text{arccosh}\big(- \langle \cosh(t)x^0 + \sinh(t)v, \cosh(s)x^0 + \sinh(s)v \rangle \big) \\ &= \text{arccosh}\big(\cosh(t)\cosh(s) - \sinh(t)\sinh(s) \big) \\ &= \text{arccosh}\big(\cosh(t - s) \big) \\ &= |t - s|, \end{aligned} \tag{38}$$

where we used that $\langle x^0, x^0 \rangle_{\mathbb{L}} = -1$, $\langle x^0, v \rangle_{\mathbb{L}} = 0$, $\langle v, v \rangle_{\mathbb{L}} = \langle v, v \rangle = 1$ and $\cosh(t)\cosh(s) - \sinh(t)\sinh(s) = \cosh(t - s)$.

On the other hand, we have

$$\begin{aligned} |t^v(x) - t^v(y)| &= \big| \text{sign}(\langle x, v \rangle) d_{\mathbb{L}}(x, x^0) - \text{sign}(\langle y, v \rangle) d_{\mathbb{L}}(y, x^0) \big| \\ &= \big| \text{sign}(\langle x, v \rangle) d_{\mathbb{L}}(\gamma(s), \gamma(0)) - \text{sign}(\langle y, v \rangle) d_{\mathbb{L}}(\gamma(t), \gamma(0)) \big| \\ &= \big| \text{sign}(\langle x, v \rangle)|s| - \text{sign}(\langle y, v \rangle)|t| \big| \\ &= |t - s|, \end{aligned} \tag{39}$$

where we use at the last line that $\text{sign}(\langle x, v \rangle) = \text{sign}(s)$ (resp. $\text{sign}(\langle y, v \rangle) = \text{sign}(t)$) and $s = \text{sign}(s)|s|$ (resp. $t = \text{sign}(t)|t|$) supposing that v is oriented in the same sense of γ.

Therefore, we have

$$|t^v(x) - t^v(y)| = d_{\mathbb{L}}(x, y). \tag{40}$$

Now, we can show the equality for the Wasserstein distance:

$$\begin{aligned} W_p^p(\mu, \nu) &= \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{L}^d \times \mathbb{L}^d} d_{\mathbb{L}}(x, y)^p \, \mathrm{d}\gamma(x, y) \\ &= \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{L}^d \times \mathbb{L}^d} |t^v(x) - t^v(y)|^p \, \mathrm{d}\gamma(x, y) \\ &= \inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathbb{R} \times \mathbb{R}} |x - y|^p \, \mathrm{d}(t^v \otimes t^v)_{\#}\gamma(x, y) \\ &= \inf_{\tilde{\gamma} \in \Pi(t^v_{\#}\mu, t^v_{\#}\nu)} \int_{\mathbb{R} \times \mathbb{R}} |x - y|^p \, \mathrm{d}\tilde{\gamma}(x, y) \\ &= W_p^p(t^v_{\#}\mu, t^v_{\#}\nu) = \int_0^1 |F^{-1}_{t^v_{\#}\mu}(u) - F^{-1}_{t^v_{\#}\nu}(u)|^p \, \mathrm{d}u, \end{aligned} \tag{41}$$

where we apply (Paty & Cuturi, 2019, Lemma 6).

A.2. Geodesic Projection

Proposition A.1 (Geodesic projection)**.**

1. *Let $\mathcal{G}^v = \text{span}(x^0, v) \cap \mathbb{L}^d$ where $v \in T_{x^0}\mathbb{L}^d \cap S^d$. Then, the geodesic projection $\tilde{P}^v$ on $\mathcal{G}^v$ of $x \in \mathbb{L}^d$ is*

$$\begin{aligned}\tilde{P}^v(x) &= \frac{1}{\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}}\big(-\langle x, x^0\rangle_{\mathbb{L}} x^0 + \langle x, v\rangle_{\mathbb{L}} v\big) \\ &= \frac{P^{\text{span}(x^0,v)}(x)}{\sqrt{-\langle P^{\text{span}(x^0,v)}(x), P^{\text{span}(x^0,v)}(x)\rangle_{\mathbb{L}}}},\end{aligned} \tag{42}$$

   where $P^{\text{span}(x^0,v^0)}$ is the linear orthogonal projection on the subspace $\text{span}(x^0, v)$.

2. *Let $\tilde{v} \in S^{d-1}$ be an in ideal point. Then, the geodesic projection $\tilde{P}^{\tilde{v}}$ on the geodesic characterized by $\tilde{v}$ of $x \in \mathbb{B}^d$ is*

$$\tilde{P}^{\tilde{v}}(x) = s(x)\tilde{v}, \tag{43}$$

   where

$$s(x) = \begin{cases} \frac{1+\|x\|_2^2 - \sqrt{(1+\|x\|_2^2)^2 - 4\langle x, \tilde{v}\rangle^2}}{2\langle x, \tilde{v}\rangle} & \text{if } \langle x, \tilde{v}\rangle \neq 0 \\ 0 & \text{if } \langle x, \tilde{v}\rangle = 0. \end{cases} \tag{44}$$

Proof.

1. **Lorentz model.** Any point y on the geodesic obtained by the intersection between $E = \text{span}(x^0, v)$ and $\mathbb{L}^d$ can be written as

$$y = \cosh(t)x^0 + \sinh(t)v, \tag{45}$$

   where $t \in \mathbb{R}$. Moreover, as arccosh is an increasing function, we have

$$\begin{aligned}\tilde{P}^v(x) &= \underset{y \in E \cap \mathbb{L}^d}{\text{argmin}}\ d_{\mathbb{L}}(x, y) \\ &= \underset{y \in E \cap \mathbb{L}^d}{\text{argmin}}\ -\langle x, y\rangle_{\mathbb{L}}.\end{aligned} \tag{46}$$

   This problem is equivalent with solving

$$\underset{t \in \mathbb{R}}{\text{argmin}}\ -\cosh(t)\langle x, x^0\rangle_{\mathbb{L}} - \sinh(t)\langle x, v\rangle_{\mathbb{L}}. \tag{47}$$

   Let $g(t) = -\cosh(t)\langle x, x^0\rangle_{\mathbb{L}} - \sinh(t)\langle x, v\rangle_{\mathbb{L}}$, then

$$g'(t) = 0 \iff \tanh(t) = -\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}. \tag{48}$$

   Finally, using that $1 - \tanh^2(t) = \frac{1}{\cosh^2(t)}$ and $\cosh^2(t) - \sinh^2(t) = 1$, and observing that necessarily, $\langle x, x^0\rangle_{\mathbb{L}} \leq 0$, we obtain

$$\cosh(t) = \frac{1}{\sqrt{1 - \left(-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right)^2}} = \frac{-\langle x, x^0\rangle_{\mathbb{L}}}{\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}}, \tag{49}$$

   and

$$\sinh(t) = \frac{-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}}{\sqrt{1 - \left(-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right)^2}} = \frac{\langle x, v\rangle_{\mathbb{L}}}{\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}}. \tag{50}$$

2. **Poincaré ball.** A geodesic passing through the origin on the Poincaré ball is of the form $\gamma(t) = tp$ for an ideal point $p \in S^{d-1}$ and $t \in]-1, 1[$. Using that $\operatorname{arccosh}$ is an increasing function, we find

$$\begin{aligned}\tilde{P}^p(x) &= \operatorname*{argmin}_{y \in \operatorname{span}(\gamma)} d_{\mathbb{B}}(x, y) \\ &= \operatorname*{argmin}_{tp} \operatorname{arccosh}\left(1 + 2\frac{\|x - \gamma(t)\|_2^2}{(1 - \|x\|_2^2)(1 - \|\gamma(t)\|_2^2)}\right) \\ &= \operatorname*{argmin}_{tp} \log\left(\|x - \gamma(t)\|_2^2\right) - \log\left(1 - \|x\|_2^2\right) - \log\left(1 - \|\gamma(t)\|_2^2\right) \\ &= \operatorname*{argmin}_{tp} \log\left(\|x - tp\|_2^2\right) - \log\left(1 - t^2\right).\end{aligned} \tag{51}$$

Let $g(t) = \log\left(\|x - tp\|_2^2\right) - \log\left(1 - t^2\right)$. Then,

$$g'(t) = 0 \iff \begin{cases} t^2 - \frac{1+\|x\|_2^2}{\langle x, p\rangle} t + 1 = 0 & \text{if } \langle p, x\rangle \neq 0, \\ t = 0 & \text{if } \langle p, x\rangle = 0. \end{cases} \tag{52}$$

Finally, if $\langle x, p\rangle \neq 0$, the solution is

$$t = \frac{1 + \|x\|_2^2}{2\langle x, p\rangle} \pm \sqrt{\left(\frac{1 + \|x\|_2^2}{2\langle x, p\rangle}\right)^2 - 1}. \tag{53}$$

Now, let us suppose that $\langle x, p\rangle > 0$. Then,

$$\begin{aligned}\frac{1 + \|x\|_2^2}{2\langle x, p\rangle} + \sqrt{\left(\frac{1 + \|x\|_2^2}{2\langle x, p\rangle}\right)^2 - 1} &\geq \frac{1 + \|x\|_2^2}{2\langle x, p\rangle} \\ &\geq 1,\end{aligned} \tag{54}$$

because $\|x - p\|_2^2 \geq 0$ implies that $\frac{1+\|x\|_2^2}{2\langle x,p\rangle} \geq 1$, and therefor the solution is

$$t = \frac{1 + \|x\|_2^2}{2\langle x, p\rangle} - \sqrt{\left(\frac{1 + \|x\|_2^2}{2\langle x, p\rangle}\right)^2 - 1}. \tag{55}$$

Similarly, if $\langle x, p\rangle < 0$, then

$$\begin{aligned}\frac{1 + \|x\|_2^2}{2\langle x, p\rangle} - \sqrt{\left(\frac{1 + \|x\|_2^2}{2\langle x, p\rangle}\right)^2 - 1} &\leq \frac{1 + \|x\|_2^2}{2\langle x, p\rangle} \\ &\leq -1,\end{aligned} \tag{56}$$

because $\|x + p\|_2^2 \geq 0$ implies $\frac{1+\|x\|_2^2}{2\langle x,p\rangle} \leq -1$, and the solution is

$$\frac{1 + \|x\|_2^2}{2\langle x, p\rangle} + \sqrt{\left(\frac{1 + \|x\|_2^2}{2\langle x, p\rangle}\right)^2 - 1}. \tag{57}$$

Thus,

$$\begin{aligned}s(x) &= \begin{cases} \frac{1+\|x\|_2^2}{2\langle x,\tilde{v}\rangle} - \sqrt{\left(\frac{1+\|x\|_2^2}{2\langle x,\tilde{v}\rangle}\right)^2 - 1} & \text{if } \langle x, \tilde{v}\rangle > 0 \\ \frac{1+\|x\|_2^2}{2\langle x,\tilde{v}\rangle} + \sqrt{\left(\frac{1+\|x\|_2^2}{2\langle x,\tilde{v}\rangle}\right)^2 - 1} & \text{if } \langle x, \tilde{v}\rangle < 0. \end{cases} \\ &= \frac{1 + \|x\|_2^2}{2\langle x, \tilde{v}\rangle} - \operatorname{sign}(\langle x, \tilde{v}\rangle \sqrt{\left(\frac{1 + \|x\|_2^2}{2\langle x, \tilde{v}\rangle}\right)^2 - 1} \\ &= \frac{1 + \|x\|_2^2}{2\langle x, \tilde{v}\rangle} - \frac{\operatorname{sign}(\langle x, \tilde{v}\rangle}{2\operatorname{sign}(\langle x, \tilde{v}\rangle\langle x, \tilde{v}\rangle} \sqrt{(1 + \|x\|_2^2)^2 - 4\langle x, \tilde{v}\rangle^2} \\ &= \frac{1 + \|x\|_2^2 - \sqrt{(1 + \|x\|_2^2)^2 - 4\langle x, \tilde{v}\rangle^2}}{2\langle x, \tilde{v}\rangle}.\end{aligned} \tag{58}$$

□

We observe that the projection on the geodesic in the Lorentz model can be done by first projecting on the subspace $\mathrm{span}(x^0, v)$ and then by projecting on the hyperboloid by normalizing. This is analogous to the spherical case studied in (Bonet et al., 2023), the differences being that, in the hyperbolic case, we are on the Minkowski space and that the geodesics are not periodic, contrary to the sphere. Moreover, we only integrate *w.r.t.* geodesics passing through the origin when Bonet et al. (2023) integrate over all possible geodesics, as the sphere does not have a natural origin.

A.3. Proof of Proposition 3.2

1. **Lorentz model.** The coordinate on the geodesic can be obtained as

$$P^v(x) = \operatorname*{argmin}_{t\in\mathbb{R}} \; d_{\mathbb{L}}\big(\exp_{x^0}(tv), x\big). \tag{59}$$

Hence, by using (48), we obtain that the optimal t satisfies

$$\tanh(t) = -\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}} \iff t = \operatorname{arctanh}\left(-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right). \tag{60}$$

2. **Poincaré ball.** As a geodesic is of the form $\gamma(t) = \tanh\left(\frac{t}{2}\right) p$ for all $t \in \mathbb{R}$, we deduce from Proposition A.1 that

$$s(x) = \tanh\left(\frac{t}{2}\right) \iff t = 2\operatorname{arctanh}\big(s(x)\big). \tag{61}$$

A.4. Proof of Proposition 3.3

1. **Lorentz model.**

The geodesic in direction v can be characterized by

$$\forall t \in \mathbb{R},\ \gamma_v(t) = \cosh(t)x^0 + \sinh(t)v. \tag{62}$$

Hence, we have

$$\begin{aligned}
\forall x \in \mathbb{L}^d,\ d_{\mathbb{L}}(\gamma_v(t), x) &= \operatorname{arccosh}(-\cosh(t)\langle x, x^0\rangle_{\mathbb{L}} - \sinh(t)\langle x, v\rangle_{\mathbb{L}}) \\
&= \operatorname{arccosh}\left(-\frac{e^t + e^{-t}}{2}\langle x, x^0\rangle_{\mathbb{L}} - \frac{e^t - e^{-t}}{2}\langle x, v\rangle_{\mathbb{L}}\right) \\
&= \operatorname{arccosh}\left(\frac{e^t}{2}\big((-1 - e^{-2t})\langle x, x^0\rangle_{\mathbb{L}} + (-1 + e^{-2t})\langle x, v\rangle_{\mathbb{L}}\big)\right) \\
&= \operatorname{arccosh}\big(x(t)\big).
\end{aligned} \tag{63}$$

Then, on one hand, we have $x(t) \underset{t\to\infty}{\to} \pm\infty$, and using that $\operatorname{arccosh}(x) = \log\big(x + \sqrt{x^2 - 1}\big)$, we have

$$\begin{aligned}
d_{\mathbb{L}}(\gamma_v(y), x) - t &= \log\left(\big(x(t) + \sqrt{x(t)^2 - 1}\big)e^{-t}\right) \\
&= \log\left(e^{-t}x(t) + e^{-t}x(t)\sqrt{1 - \frac{1}{x(t)^2}}\right) \\
&\underset{\infty}{=} \log\left(e^{-t}x(t) + e^{-t}x(t)\left(1 - \frac{1}{2x(t)^2} + o\left(\frac{1}{x(t)^2}\right)\right)\right).
\end{aligned} \tag{64}$$

Moreover,

$$e^{-t}x(t) = \frac{1}{2}(-1 - e^{-2t})\langle x, x^0\rangle_{\mathbb{L}} + \frac{1}{2}(-1 + e^{-2t})\langle x, v\rangle_{\mathbb{L}} \underset{t\to\infty}{\to} -\frac{1}{2}\langle x, x^0 + v\rangle_{\mathbb{L}}. \tag{65}$$

Hence,

$$B^v(x) = \log(-\langle x, x^0 + v\rangle_{\mathbb{L}}). \tag{66}$$

2. **Poincaré ball.**

Note that this proof can be found *e.g.* in the Appendix of (Ghadimi Atigh et al., 2021). We report it for the sake of completeness.

Let $p \in S^{d-1}$, then the geodesic from 0 to p is of the form $\gamma_p(t) = \exp_0(tp) = \tanh(\frac{t}{2})p$. Moreover, recall that $\operatorname{arccosh}(x) = \log(x + \sqrt{x^2-1})$ and

$$d_{\mathbb{B}}(\gamma_p(t), x) = \operatorname{arccosh}\left(1 + 2\frac{\|\tanh(\frac{t}{2})p - x\|_2^2}{(1-\tanh^2(\frac{t}{2}))(1-\|x\|_2^2)}\right) = \operatorname{arccosh}(1+x(t)), \tag{67}$$

where

$$x(t) = 2\frac{\|\tanh(\frac{t}{2})p - x\|_2^2}{(1-\tanh^2(\frac{t}{2}))(1-\|x\|_2^2)}. \tag{68}$$

Now, on one hand, we have

$$\begin{aligned} B^p(x) &= \lim_{t\to\infty} \left(d_{\mathbb{B}}(\gamma_p(t), x) - t\right) \\ &= \lim_{t\to\infty} \log\left(1 + x(t) + \sqrt{x(t)^2 + 2x(t)}\right) - t \\ &= \lim_{t\to\infty} \log\left(e^{-t}(1 + x(t) + \sqrt{x(t)^2 + 2x(t)})\right). \end{aligned} \tag{69}$$

On the other hand, using that $\tanh(\frac{t}{2}) = \frac{e^t-1}{e^t+1}$,

$$\begin{aligned} e^{-t}x(t) &= 2e^{-t}\frac{\|\frac{e^t-1}{e^t+1}p - x\|_2^2}{(1-(\frac{e^t-1}{e^t+1})^2)(1-\|x\|_2^2)} \\ &= 2e^{-t}\frac{\|e^tp - p - e^tx - x\|_2^2}{4e^t(1-\|x\|_2^2)} \\ &= \frac{1}{2}\frac{\|p - e^{-t}p - x - e^{-t}x\|_2^2}{1-\|x\|_2^2} \\ &\underset{t\to\infty}{\to} \frac{1}{2}\frac{\|p-x\|_2^2}{1-\|x\|_2^2}. \end{aligned} \tag{70}$$

Hence,

$$B^p(x) = \lim_{t\to\infty} \log\left(e^{-t} + e^{-t}x(t) + e^{-t}x(t)\sqrt{1 + \frac{2}{x(t)}}\right) = \log\left(\frac{\|p-x\|_2^2}{1-\|x\|_2^2}\right), \tag{71}$$

using that $\sqrt{1+\frac{2}{x(t)}} = 1 + \frac{1}{x(t)} + o(\frac{1}{x(t)})$ and $\frac{1}{x(t)} \to_{t\to\infty} 0$.

A.5. Horospherical Projections

Proposition A.2 (Horospherical projection)**.**

1. *Let $v \in T_{x^0}\mathbb{L}^d \cap S^d$ be a direction and $\mathcal{G} = \operatorname{span}(x^0, v) \cap \mathbb{L}^d$ the corresponding geodesic passing through x^0. Then, for any $x \in \mathbb{L}^d$, the projection on $\mathcal{G}$ along the horosphere is given by*

$$\tilde{B}^v(x) = \frac{1+u^2}{1-u^2}x^0 + \frac{2u}{1-u^2}v, \tag{72}$$

where $u = \frac{1+\langle x, x^0+v\rangle_{\mathbb{L}}}{1-\langle x, x^0+v\rangle_{\mathbb{L}}}$.

2. *Let $\tilde{v} \in S^{d-1}$ be an ideal point. Then, for all $x \in \mathbb{B}^d$,*

$$\tilde{B}^{\tilde{v}}(x) = \left(\frac{1-\|x\|_2^2 - \|\tilde{v}-x\|_2^2}{1-\|x\|_2^2 + \|\tilde{v}-x\|_2^2}\right)\tilde{v}. \tag{73}$$

Proof. 1. **Lorentz model.**

First, a point on the geodesic γ_v is of the form

$$y(t) = \cosh(t)x^0 + \sinh(t)v, \tag{74}$$

with $t \in \mathbb{R}$.

The projection along the horosphere amounts at following the level sets of the Busemann function B^v. And we have

$$\begin{aligned} B^v(x) = B^v(y(t)) &\iff \log(-\langle x, x^0 + v\rangle_{\mathbb{L}}) = \log(-\langle \cosh(t)x^0 + \sinh(t)v, x^0 + v\rangle_{\mathbb{L}}) \\ &\iff \log(-\langle x, x^0 + v\rangle_{\mathbb{L}}) = \log(-\cosh(t)\|x^0\|_{\mathbb{L}}^2 - \sinh(t)\|v\|_{\mathbb{L}}^2) \\ &\iff \log(-\langle x, x^0 + v\rangle_{\mathbb{L}} = \log(\cosh(t) - \sinh(t)) \\ &\iff \langle x, x^0 + v\rangle_{\mathbb{L}} = \sinh(t) - \cosh(t). \end{aligned} \tag{75}$$

By noticing that $\cosh(t) = \frac{1+\tanh^2(\frac{t}{2})}{1-\tanh^2(\frac{t}{2})}$ and $\sinh(t) = \frac{2\tanh(\frac{t}{2})}{1-\tanh^2(\frac{t}{2})}$, let $u = \tanh(\frac{t}{2})$, then we have

$$\begin{aligned} B^v(x) = B^v(y(t)) &\iff \langle x, x^0 + v\rangle_{\mathbb{L}} = \frac{2u}{1-u^2} - \frac{1+u^2}{1-u^2} = \frac{-(u-1)^2}{(1-u)(1+u)} = \frac{u-1}{u+1} \\ &\iff u = \frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}}{1 - \langle x, x^0 + v\rangle_{\mathbb{L}}}. \end{aligned} \tag{76}$$

We can further continue the computation and obtain, by denoting $c = \langle x, x^0 + v\rangle_{\mathbb{L}}$,

$$\begin{aligned} \tilde{B}^v(x) &= \frac{1+u^2}{1-u^2}x^0 + \frac{2u}{1-u^2}v \\ &= \frac{1 + \left(\frac{1+c}{1-c}\right)^2}{1 - \left(\frac{1+c}{1-c}\right)^2}x^0 + 2\frac{\left(\frac{1+c}{1-c}\right)}{1 - \left(\frac{1+c}{1-c}\right)^2}v \\ &= \frac{(1-c)^2 + (1+c)^2}{(1-c)^2 - (1+c)^2}x^0 + 2\frac{(1+c)(1-c)}{(1-c)^2 - (1+c)^2}v \\ &= -\frac{1+c^2}{2c}x^0 - \frac{1-c^2}{2c}v \\ &= -\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}\big((1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)x^0 + (1 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2)v\big). \end{aligned} \tag{77}$$

2. **Poincaré ball.**

Let $p \in S^{d-1}$. First, we notice that points on the geodesic generated by p and passing through 0 are of the form $x(\lambda) = \lambda p$ where $\lambda \in]-1, 1[$.

Moreover, there is a unique horosphere $S(p, x)$ passing through x and starting from p. The points on this horosphere are of the form

$$\begin{aligned} y(\theta) &= \left(\frac{p + x(\lambda^*)}{2}\right) + \left\|\frac{p - x(\lambda^*)}{2}\right\|_2 \left(\cos(\theta)p + \sin(\theta)\frac{x - \langle x, p\rangle p}{\|x - \langle x, p\rangle p\|_2}\right) \\ &= \frac{1+\lambda^*}{2}p + \frac{1-\lambda^2}{2}\left(\cos(\theta)p + \sin(\theta)\frac{x - \langle x, p\rangle p}{\|x - \langle x, p\rangle p\|_2}\right), \end{aligned} \tag{78}$$

where λ^* characterizes the intersection between the geodesic and the horosphere.

Since the horosphere are the level sets of the Busemann function, we have $B^p(x) = B^p(\lambda^* p)$. Thus, we have

$$\begin{aligned} B^p(x) = B^p(\lambda^* p) &\iff \log\left(\frac{\|p-x\|_2^2}{1-\|x\|_2^2}\right) = \log\left(\frac{\|p-\lambda^* p\|_2^2}{1-\|\lambda^* p\|_2^2}\right) \\ &\iff \frac{\|p-x\|_2^2}{1-\|x\|_2^2} = \frac{(1-\lambda^*)^2}{1-(\lambda^*)^2} \\ &\iff \frac{\|p-x\|_2^2}{1-\|x\|_2^2} = \frac{1-\lambda^*}{1+\lambda^*} \\ &\iff \lambda^*\left(\frac{\|p-x\|_2^2}{1-\|x\|_2^2}+1\right) = 1 - \frac{\|p-x\|_2^2}{1-\|x\|_2^2} \\ &\iff \lambda^* = \frac{1-\|x\|_2^2-\|p-x\|_2^2}{1-\|x\|_2^2+\|p-x\|_2^2}. \end{aligned} \tag{79}$$

□

A.6. Proof of Proposition 3.4

First, we show some Lemma.

Lemma A.3 (Commutation of projections.)**.** *Let $v \in \operatorname{span}(x^0)^\perp \cap S^d$ of the form $v = (0, \tilde{v})$ where $\tilde{v} \in S^{d-1}$. Then, for all $x \in \mathbb{B}^d$, $y \in \mathbb{L}^d$*

$$P_{\mathbb{B}\to\mathbb{L}}\big(\tilde{B}^{\tilde{v}}(x)\big) = \tilde{B}^{v}\big(P_{\mathbb{B}\to\mathbb{L}}(x)\big), \tag{80}$$

$$\tilde{B}^{\tilde{v}}(P_{\mathbb{L}\to\mathbb{B}}(y)) = P_{\mathbb{L}\to\mathbb{B}}(\tilde{B}^{v}(y)) \tag{81}$$

$$P_{\mathbb{B}\to\mathbb{L}}\big(\tilde{P}^{\tilde{v}}(x)\big) = \tilde{P}^{v}\big(P_{\mathbb{B}\to\mathbb{L}}(x)\big), \tag{82}$$

$$\tilde{P}^{\tilde{v}}(P_{\mathbb{L}\to\mathbb{B}}(y)) = P_{\mathbb{L}\to\mathbb{B}}(\tilde{P}^{v}(y)). \tag{83}$$

Proof. We first show (80). Let's recall the formula of the different projections.

On one hand,

$$\forall x \in \mathbb{B}^d,\ \tilde{B}^{\tilde{v}}(x) = \left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)\tilde{v}, \tag{84}$$

$$\forall x \in \mathbb{L}^d,\ \tilde{B}^{v}(x) = -\frac{1}{2\langle x, x^0+v\rangle_{\mathbb{L}}}\big((1+\langle x, x^0+v\rangle_{\mathbb{L}}^2)x^0 + (1-\langle x, x^0+v\rangle_{\mathbb{L}}^2)v\big), \tag{85}$$

and

$$\forall x \in \mathbb{B}^d,\ P_{\mathbb{B}\to\mathbb{L}}(x) = \frac{1}{1-\|x\|_2^2}(1+\|x\|_2^2, 2x_1, \dots, 2x_d). \tag{86}$$

Let $x \in \mathbb{B}^d$. First, let's compute $P_{\mathbb{B}\to\mathbb{L}}\big(\tilde{B}^{\tilde{v}}(x)\big)$. We note that $\|\tilde{v}\|_2^2 = 1$ and therefore

$$\|\tilde{B}^{\tilde{v}}(v)\|_2^2 = \left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)^2. \tag{87}$$

Then,

$$
\begin{aligned}
P_{\mathbb{B}\to\mathbb{L}}\big(\tilde{B}^{\tilde{v}}(x)\big) &= \frac{1}{1-\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)^2}\left(1+\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)^2, 2\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)\tilde{v}\right)\\
&= \frac{1}{1-\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)^2}\left(\left(1+\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)^2\right)x^0 + 2\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)v\right)\\
&= \frac{\big(1-\|x\|_2^2+\|\tilde{v}-x\|_2^2\big)^2}{4\|\tilde{v}-x\|_2^2(1-\|x\|_2^2)}\left(\frac{2(1-\|x\|_2^2)^2+2\|\tilde{v}-x\|_2^4}{\big(1-\|x\|_2^2+\|\tilde{v}-x\|_2^2\big)^2}x^0 + 2\left(\frac{1-\|x\|_2^2-\|\tilde{v}-x\|_2^2}{1-\|x\|_2^2+\|\tilde{v}-x\|_2^2}\right)v\right)\\
&= \frac{1}{2\|\tilde{v}-x\|_2^2(1-\|x\|_2^2)}\left(((1-\|x\|_2^2)^2+\|\tilde{v}-x\|_2^4)x^0 + (1-\|x\|_2^2-\|\tilde{v}-x\|_2^2)(1-\|x\|_2^2+\|\tilde{v}-x\|_2^2)v\right)\\
&= \frac{1}{2\|\tilde{v}-x\|_2^2(1-\|x\|_2^2)}\left((1-\|x\|_2^2)^2+\|\tilde{v}-x\|_2^4)x^0 + \big((1-\|x\|_2^2)^2-\|\tilde{v}-x\|_2^4\big)v\right).
\end{aligned}
\tag{88}
$$

Now, let's compute $\tilde{B}^v\big(P_{\mathbb{B}\to\mathbb{L}}(x)\big)$. First, let's remark that for all $y\in\mathbb{L}^d$, $\langle y, x^0+v\rangle_{\mathbb{L}} = -y_0 + \langle y_{1:d},\tilde{v}\rangle$. Therefore, for all $x\in\mathbb{B}^d$,

$$
\begin{aligned}
\langle P_{\mathbb{B}\to\mathbb{L}}(x), x^0+v\rangle_{\mathbb{L}} &= \langle \frac{1}{1-\|x\|_2^2}(1+\|x\|_2^2, 2x_1,\dots,2x_d), x^0+v\rangle_{\mathbb{L}}\\
&= \frac{1}{1-\|x\|_2^2}\big(-1-\|x\|_2^2+2\langle x,\tilde{v}\rangle\big)\\
&= -\frac{1}{1-\|x\|_2^2}\|x-\tilde{v}\|_2^2.
\end{aligned}
\tag{89}
$$

Moreover,

$$
\langle P_{\mathbb{B}\to\mathbb{L}}(x), x^0+v\rangle_{\mathbb{L}}^2 = \frac{1}{(1-\|x\|_2^2)^2}\|\tilde{v}-x\|_2^4. \tag{90}
$$

Therefore, we have

$$
\begin{aligned}
\tilde{B}^v\big(P_{\mathbb{B}\to\mathbb{L}}(x)\big) &= \tilde{B}^v\left(\frac{1}{1-\|x\|_2^2}(1+\|x\|_2^2, 2x_1,\dots,2x_d)\right)\\
&= -\frac{1-\|x\|_2^2}{2\left(-1-\|x\|_2^2+2\langle x,\tilde{v}\rangle\right)}\left(\big(1+\langle P_{\mathbb{B}\to\mathbb{L}}(x), x^0+v\rangle_{\mathbb{L}}^2\big)x^0 + \big(1-\langle P_{\mathbb{B}\to\mathbb{L}}(x), x^0+v\rangle_{\mathbb{L}}^2\big)v\right)\\
&= \frac{1-\|x\|_2^2}{2\|x-\tilde{v}\|_2^2}\left(\frac{(1-\|x\|_2^2)^2+\|\tilde{v}-x\|^4}{(1-\|x\|_2^2)^2}x^0 + \frac{(1-\|x\|_2^2)^2-\|\tilde{v}-x\|^4}{(1-\|x\|_2^2)^2}v\right)\\
&= \frac{1}{2\|x-\tilde{v}\|_2^2(1-\|x\|_2^2)}\left((1-\|x\|_2^2)^2+\|\tilde{v}-x\|_2^4)x^0 + \big((1-\|x\|_2^2)^2-\|\tilde{v}-x\|_2^4\big)v\right)\\
&= P_{\mathbb{B}\to\mathbb{L}}\big(\tilde{B}^{\tilde{v}}(x)\big).
\end{aligned}
\tag{91}
$$

For (81), we use that $P_{\mathbb{B}\to\mathbb{L}}$ and $P_{\mathbb{L}\to\mathbb{B}}$ are inverse from each other. Hence, for all $x\in\mathbb{B}^d$, there exists $y\in\mathbb{L}^d$ such that $x = P_{\mathbb{L}\to\mathbb{B}}(y) \iff y = P_{\mathbb{B}\to\mathbb{L}}(x)$, and we obtain the second equality by plugging it into (80).

Now, let's show (82). The proof relies on the observation that $\{\exp_{x^0}(tv),\ t\in\mathbb{R}\} = P_{\mathbb{B}\to\mathbb{L}}\left(\{\exp_0(t\tilde{v}),\ t\in\mathbb{R}\}\right)$ (*i.e.* the images by $P_{\mathbb{B}\to\mathbb{L}}$ of geodesics in the Poincaré ball are geodesics in the Lorentz model). Thus,

$$
\begin{aligned}
\tilde{P}^v(P_{\mathbb{B}\to\mathbb{L}}(x)) &= \operatorname*{argmin}_{z\in\{\exp_{x^0}(tv),\ t\in\mathbb{R}\}} d_{\mathbb{L}}(P_{\mathbb{B}\to\mathbb{L}}(x), z)\\
&= P_{\mathbb{B}\to\mathbb{L}}\Big(\operatorname*{argmin}_{z\in\{\exp_0(t\tilde{v}),\ t\in\mathbb{R}\}} d_{\mathbb{B}}(P_{\mathbb{L}\to\mathbb{L}}(x), P_{\mathbb{B}\to\mathbb{L}}(z))\Big)\\
&= P_{\mathbb{B}\to\mathbb{L}}\Big(\operatorname*{argmin}_{z\in\{\exp_0(t\tilde{v}),\ t\in\mathbb{R}\}} d_{\mathbb{B}}(x,z)\Big)\\
&= P_{\mathbb{B}\to\mathbb{L}}\big(\tilde{P}^v(x)\big).
\end{aligned}
\tag{92}
$$

Similarly, we obtain (83). □

Lemma A.4. *Let $v = (0, \tilde{v}) \in \mathrm{span}(x^0)^\top$. For all $x \in \mathbb{L}^d$, $y \in \mathbb{B}^d$,*

$$B^v(x) = -t^v\big(\tilde{B}^v(x)\big), \tag{93}$$
$$B^{\tilde{v}}(y) = -t^{\tilde{v}}(\tilde{B}^{\tilde{v}}(y)). \tag{94}$$

Proof. First, let us show that

$$d_{\mathbb{L}}(\tilde{B}^v(x), x^0) = |B^v(x)|. \tag{95}$$

By recalling that $B^v\big(\tilde{P}^v(x)\big) = B^v(x) = \log\big(-\langle x, x^0 + v\rangle_{\mathbb{L}}\big)$ and that by (77),

$$\tilde{B}^v(x) = -\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}\left((1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)x^0 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2 v\right). \tag{96}$$

Now, by remarking that $\langle x, x^0 + v\rangle_{\mathbb{L}} \leq 0$, then we have,

$$\begin{aligned}
d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) &= \mathrm{arccosh}(-\langle \tilde{B}^v(x), x^0\rangle_{\mathbb{L}}) \\
&= \mathrm{arccosh}\left(\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}(1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)\langle x^0, x^0\rangle_{\mathbb{L}}\right) \\
&= \mathrm{arccosh}\left(-\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}(1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)\right) \\
&= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}} + \sqrt{\frac{(1 + \langle x, x^0 + v\rangle_{\mathbb{L}})^2}{4\langle x, x^0 + v\rangle_{\mathbb{L}}^2} - 1}\right) \\
&= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 + \sqrt{\big(\langle x, x^0 + v\rangle_{\mathbb{L}}^2 - 1\big)^2}}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}\right)
\end{aligned} \tag{97}$$

If $\langle x, x^0 + v\rangle_{\mathbb{L}}^2 \geq 1$, then

$$\begin{aligned}
d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) &= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 - 1}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) \\
&= \log(-\langle x, x^0 + v\rangle_{\mathbb{L}}) = B^v(x).
\end{aligned} \tag{98}$$

And if $\langle x, x^0 + v\rangle_{\mathbb{L}}^2 \leq 1$, then

$$\begin{aligned}
d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) &= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 + 1 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) \\
&= \log\left(\frac{1}{-\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) = -B^v(x).
\end{aligned} \tag{99}$$

Hence, we showed that for all $x \in \mathbb{L}^d$,

$$d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) = |B^v(x)|. \tag{100}$$

Then, using (77), we have

$$\langle \tilde{B}^v(x), v\rangle = \frac{1 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}. \tag{101}$$

Then, on one hand, we can show that $\langle x, x^0 + v\rangle_{\mathbb{L}} \leq 0$ since $x = \lambda_0 x^0 + \lambda_1 v + x_{\mathrm{span}(x^0, v)^\perp}$. Thus,

$$\langle x, x^0 + v\rangle_{\mathbb{L}} = -\lambda_0 + \lambda_1. \tag{102}$$

But, $\lambda_0 = \sqrt{1 + \sum_{i=1^d} \lambda_i^2} \geq \sqrt{\lambda_1^2} \geq \lambda_1$. Therefore, $\lambda_1 - \lambda_0 = \langle x, x^0 + v\rangle_{\mathbb{L}} \leq 0$.

Therefore, we have $-\langle x, x^0 + v\rangle_{\mathbb{L}} \geq 0$. And we have

$$\langle \tilde{B}^v(x), v\rangle \geq 0 \iff 1 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2 \geq 0 \iff 1 \geq \langle x, x^0 + v\rangle_{\mathbb{L}}^2 \iff B^v(x) \leq 0, \tag{103}$$

using that $B^v(x) = \log(-\langle x, x^0 + v\rangle_{\mathbb{L}})$.

Similarly,

$$\langle \tilde{B}^v(x), v\rangle \leq 0 \iff 1 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2 \leq 0 \iff 1 \leq \langle x, x^0 + v\rangle_{\mathbb{L}}^2 \iff B^v(x) \geq 0. \tag{104}$$

Hence,

$$\operatorname{sign}(\langle \tilde{B}^v(x), v\rangle) = -\operatorname{sign}(B^v(x)). \tag{105}$$

Finally, we deduce that

$$t^v\big(\tilde{B}^v(x)\big) = \operatorname{sign}(\langle \tilde{B}^v(x), v\rangle) d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) = -B^v(x). \tag{106}$$

For the second equality, let $y \in \mathbb{B}^d$, then,

$$\begin{aligned}
t^{\tilde{v}}(\tilde{B}^{\tilde{v}}(y)) &= \operatorname{sign}(\langle \tilde{B}^{\tilde{v}}(y), \tilde{v}\rangle)\, d_{\mathbb{B}}(\tilde{B}^{\tilde{v}}(y), 0) \\
&= \operatorname{sign}(\langle P_{\mathbb{B}\to\mathbb{L}}(\tilde{B}^{\tilde{v}}(y)), v\rangle) d_{\mathbb{L}}(P_{\mathbb{B}\to\mathbb{L}}(\tilde{B}^{\tilde{v}}(y)), x^0) \\
&= \operatorname{sign}(\langle \tilde{B}^v(P_{\mathbb{B}\to\mathbb{L}}(y)), v\rangle) d_{\mathbb{L}}(\tilde{B}^v(P_{\mathbb{B}\to\mathbb{L}}(y)), x^0) \quad \text{using Lemma A.3} \\
&= t^v(\tilde{B}^v(P_{\mathbb{B}\to\mathbb{L}}(y))) \quad \text{by definition of } t^v \\
&= -B^v(P_{\mathbb{B}\to\mathbb{L}}(y)) \quad \text{by (106)} \\
&= -B^{\tilde{v}}(y),
\end{aligned} \tag{107}$$

where the last line comes from

$$\begin{aligned}
B^v(P_{\mathbb{B}\to\mathbb{L}}(y)) &= \lim_{t\to\infty} \big(d_{\mathbb{L}}(\exp_{x^0}(tv), P_{\mathbb{B}\to\mathbb{L}}(y)) - t\big) \\
&= \lim_{t\to\infty} \big(d_{\mathbb{L}}(P_{\mathbb{B}\to\mathbb{L}}(P_{\mathbb{L}\to\mathbb{B}}(\exp_{x^0}(tv))), P_{\mathbb{B}\to\mathbb{L}}(y)) - t\big) \\
&= \lim_{t\to\infty} \big(d_{\mathbb{B}}(P_{\mathbb{L}\to\mathbb{B}}(\exp_{x^0}(tv)), y) - t\big) \\
&= \lim_{t\to\infty} \big(d_{\mathbb{B}}(\exp_0(t\tilde{v}), y) - t\big) \\
&= B^{\tilde{v}}(y).
\end{aligned} \tag{108}$$

□

Proof of Proposition 3.4. Let $\mu, \nu \in \mathcal{P}(\mathbb{B}^d)$, $\tilde{\mu} = (P_{\mathbb{B}\to\mathbb{L}})_{\#}\mu$, $\tilde{\nu} = (P_{\mathbb{B}\to\mathbb{L}})_{\#}\nu$, $\tilde{v} \in S^{d-1}$ an ideal point and $v = (0, \tilde{v}) \in \operatorname{span}(x^0)^{\perp}$.

First, by Lemma A.4, $B^v = -t^v \circ \tilde{B}^v$. Using the proof A.1, t^v is an isometry and we have that (by using the invariant of the Wasserstein distance),

$$W_p^p(B^v_{\#}\tilde{\mu}, B^v_{\#}\tilde{\nu}) = W_p^p(\tilde{B}^v_{\#}\tilde{\mu}, \tilde{B}^v_{\#}\tilde{\nu}). \tag{109}$$

Then,

$$\begin{aligned}
W_p^p(B^v_{\#}\tilde{\mu}, B^v_{\#}\tilde{\nu}) &= W_p^p(\tilde{B}^v_{\#}\tilde{\mu}, \tilde{B}^v_{\#}\tilde{\nu}) \\
&= W_p^p(\tilde{B}^v_{\#}(P_{\mathbb{B}\to\mathbb{L}})_{\#}\mu, \tilde{B}^v_{\#}(P_{\mathbb{B}\to\mathbb{L}})_{\#}\nu) \\
&= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{B}^d\times\mathbb{B}^d} d_{\mathbb{L}}\big(\tilde{B}^v(P_{\mathbb{B}\to\mathbb{L}}(x)), \tilde{B}^v(P_{\mathbb{B}\to\mathbb{L}}(y))\big)^p \,\mathrm{d}\gamma(x,y) \quad \text{by (Paty \& Cuturi, 2019, Lemma 6)} \\
&= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{B}^d\times\mathbb{B}^d} d_{\mathbb{L}}\big(P_{\mathbb{B}\to\mathbb{L}}(\tilde{B}^{\tilde{v}}(x)), P_{\mathbb{B}\to\mathbb{L}}(\tilde{B}^{\tilde{v}}(y))\big)^p \,\mathrm{d}\gamma(x,y) \quad \text{by Lemma A.3} \\
&= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{B}^d\times\mathbb{B}^d} d_{\mathbb{B}}\big(\tilde{B}^{\tilde{v}}(x), \tilde{B}^{\tilde{v}}(y)\big)^p \,\mathrm{d}\gamma(x,y) \quad \text{as } P_{\mathbb{B}\to\mathbb{L}} \text{ is an isometry} \\
&= W_p^p(\tilde{B}^{\tilde{v}}_{\#}\mu, \tilde{B}^{\tilde{v}}_{\#}\nu) \quad \text{by (Paty \& Cuturi, 2019, Lemma 6)} \\
&= W_p^p(B^{\tilde{v}}_{\#}\mu, B^{\tilde{v}}_{\#}\nu),
\end{aligned} \tag{110}$$

where for the last line, we use that by Lemma A.3, $B^{\tilde{v}} = -t^{\tilde{v}} \circ \tilde{B}^{\tilde{v}}$ and that $t^{\tilde{v}}$ is an isometry. Indeed,

$$\begin{aligned}\forall x, y \in \{\exp_0(t\tilde{v}),\ t \in \mathbb{R},\ |t^{\tilde{v}}(x) - t^{\tilde{v}}(y)| &= |\text{sign}(\langle x, \tilde{v}\rangle) d_{\mathbb{B}}(x, 0) - \text{sign}(\langle y, \tilde{v}\rangle) d_{\mathbb{B}}(y, 0)| \\ &= |\text{sign}(\langle P_{\mathbb{B}\to\mathbb{L}}(x), v\rangle) d_{\mathbb{L}}(P_{\mathbb{B}\to\mathbb{L}}(x), x^0) - \text{sign}(\langle P_{\mathbb{B}\to\mathbb{L}}(y), v\rangle) d_{\mathbb{L}}(P_{\mathbb{B}\to\mathbb{L}}(y), x^0)| \\ &= |t^v(P_{\mathbb{B}\to\mathbb{L}}(x)) - t^v(P_{\mathbb{B}\to\mathbb{L}}(y))| \\ &= d_{\mathbb{L}}(P_{\mathbb{B}\to\mathbb{L}}(x), P_{\mathbb{B}\to\mathbb{L}}(y)) \quad \text{by Proposition 3.1} \\ &= d_{\mathbb{B}}(x, y) \quad \text{as } P_{\mathbb{B}\to\mathbb{L}} \text{ is an isometry.}\end{aligned} \tag{111}$$

It is true for all $\tilde{v} \in S^{d-1}$, and hence for λ-almost all $\tilde{v} \in S^{d-1}$. Therefore, we have

$$HHSW_p^p(\mu, \nu) = HHSW_p^p(\tilde{\mu}, \tilde{\nu}). \tag{112}$$

Similarly, with the same reasonment, using that t^v and $t^{\tilde{v}}$ are isometries and Lemma A.3, we obtain

$$GHSW_p^p(\mu, \nu) = GHSW_p^p(\tilde{\mu}, \tilde{\nu}). \tag{113}$$

□

B. Properties

We derive in this section additional properties of HHSW and GHSW. First, we will start by showing that for $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$, we have well $GHSW_p(\mu, \nu) < \infty$ and $HHSW_p(\mu, \nu) < \infty$. We also show that the Busemann coordinates can directly be used in HHSW to compute the coordinates on $\mathbb{R}$. Then, we continue by showing that GHSW and HHSW are pseudo-distances. And finally, we make connections with Radon transforms known in the literature.

B.1. Finiteness of GHSW and HHSW

Proposition B.1. *Let $p \geq 1$, then for $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$, $GHSW_p(\mu, \nu) < \infty$ and $HHSW_p(\mu, \nu) < \infty$.*

Proof. First, we will deal with GHSW and then with HHSW. Let $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d) = \{\mu \in \mathcal{P}(\mathbb{L}^d),\ \int_{\mathbb{L}^d} d_{\mathbb{L}}(x, x_0)^p \mathrm{d}\mu(x) < \infty \text{ for some } x_0 \in \mathbb{L}^d\}$. Note that the choice of x_0 is arbitrary, since for any $x, y \in \mathbb{L}^d$, we have by the triangular inequality

$$d_{\mathbb{L}}(x, y) \leq d_{\mathbb{L}}(x, x_0) + d_{\mathbb{L}}(x_0, y). \tag{114}$$

Then, both proofs will follow from (Villani, 2009, Definition 6.4) using that

$$\forall x, y \in \mathbb{L}^d,\ d_{\mathbb{L}}(x, y)^p \leq 2^{p-1}\big(d_{\mathbb{L}}(x, x_0)^p + d_{\mathbb{L}}(x_0, y)^p\big). \tag{115}$$

GHSW. Let $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$. Then, using (115), we have, denoting $\gamma \in \Pi(\mu, \nu)$ an arbitrary coupling and using (Paty & Cuturi, 2019, Lemma 6),

$$\begin{aligned}W_p^p(P^v_{\#}\mu, P^v_{\#}\nu) = W_p^p(\tilde{P}^v_{\#}\mu, \tilde{P}^v_{\#}\nu) &= \inf_{\pi \in \Pi(\tilde{P}^v_{\#}\mu, \tilde{P}^v_{\#}\nu)} \int_{\mathbb{L}^d \times \mathbb{L}^d} d_{\mathbb{L}}(x, y)^p \,\mathrm{d}\pi(x, y) \\ &= \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{L}^d \times \mathbb{L}^d} d_{\mathbb{L}}\big(\tilde{P}^v(x), \tilde{P}^v(y)\big)^p \,\mathrm{d}\pi(x, y) \\ &\leq \int_{\mathbb{L}^d \times \mathbb{L}^d} d_{\mathbb{L}}\big(\tilde{P}^v(x), \tilde{P}^v(y)\big)^p \,\mathrm{d}\gamma(x, y) \\ &\leq 2^{p-1} \int_{\mathbb{L}^d \times \mathbb{L}^d} \Big(d_{\mathbb{L}}\big(\tilde{P}^v(x), x_0\big)^p + d_{\mathbb{L}}\big(\tilde{P}^v(y), x_0\big)^p\Big) \,\mathrm{d}\gamma(x, y) \\ &= 2^{p-1}\Big(\int_{\mathbb{L}^d} d_{\mathbb{L}}\big(\tilde{P}^v(x), x_0\big)^p \mathrm{d}\mu(x) + \int_{\mathbb{L}^d} d_{\mathbb{L}}\big(\tilde{P}^v(y), x_0\big)^p \mathrm{d}\nu(y)\Big).\end{aligned} \tag{116}$$

If we take x_0 belonging to the geodesic, then necessarily, $d_{\mathbb{L}}\big(\tilde{P}^v(x), x\big) \leq d_{\mathbb{L}}(x_0, x)$ using that $P^v(x) = \operatorname{argmin}_{y \in \operatorname{span}(x^0, v) \cap \mathbb{L}^d} d_{\mathbb{L}}(x, y)$. Hence, by using again (115), we have

$$\begin{aligned} W_p^p(P^v_{\#}\mu, P^v_{\#}\nu) &\leq 2^{2p-2}\Big(\int_{\mathbb{L}^d} d_{\mathbb{L}}(\tilde{P}^v(x), x)^p \mathrm{d}\mu(x) + \int_{\mathbb{L}^d} d_{\mathbb{L}}(x, x_0)^p \mathrm{d}\mu(x) \\ &\quad + \int_{\mathbb{L}^d} d_{\mathbb{L}}(\tilde{P}^v(y), y)^p \mathrm{d}\nu(y) + \int_{\mathbb{L}^d} d_{\mathbb{L}}(y, x_0)^p \mathrm{d}\nu(y)\Big) \\ &\leq \Big(\int_{\mathbb{L}^d} d_{\mathbb{L}}(x_0, x)^p \mathrm{d}\mu(x) + \int_{\mathbb{L}^d} d_{\mathbb{L}}(x, x_0)^p \mathrm{d}\mu(y) \\ &\quad + \int_{\mathbb{L}^d} d_{\mathbb{L}}(x_0, y)^p \mathrm{d}\nu(y) + \int_{\mathbb{L}^d} d_{\mathbb{L}}(y, x_0)^p \mathrm{d}\nu(y)\Big) \\ &< \infty. \end{aligned} \tag{117}$$

And hence $GHSW_p(\mu, \nu) < \infty$.

HHSW. Let's take first $x_0 = x^0$ as the base point. Then, by using again (115), we have:

$$W_p^p(B^v_{\#}\mu, B^v_{\#}\nu) \leq 2^{p-1}\Big(\int_{\mathbb{L}^d} d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big)^p \mathrm{d}\mu(x) + \int_{\mathbb{L}^d} d_{\mathbb{L}}\big(\tilde{B}^v(y), x^0\big)^p \mathrm{d}\nu(y)\Big). \tag{118}$$

Now, by recalling that $B^v\big(\tilde{B}^v(x)\big) = B^v(x) = \log\big(-\langle x, x^0 + v\rangle_{\mathbb{L}}\big)$ and

$$\tilde{B}^v(x) = -\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}\left((1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)x^0 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2 v\right). \tag{119}$$

Now, by remarking that $\langle x, x^0 + v\rangle_{\mathbb{L}} \leq 0$, then we have,

$$\begin{aligned} d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) &= \operatorname{arccosh}(-\langle \tilde{B}^v(x), x^0\rangle_{\mathbb{L}}) \\ &= \operatorname{arccosh}\left(\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}(1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)\langle x^0, x^0\rangle_{\mathbb{L}}\right) \\ &= \operatorname{arccosh}\left(-\frac{1}{2\langle x, x^0 + v\rangle_{\mathbb{L}}}(1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2)\right) \\ &= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}} + \sqrt{\frac{(1 + \langle x, x^0 + v\rangle_{\mathbb{L}})^2}{4\langle x, x^0 + v\rangle_{\mathbb{L}}^2} - 1}\right) \\ &= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 + \sqrt{\big(\langle x, x^0 + v\rangle_{\mathbb{L}}^2 - 1\big)^2}}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) \end{aligned} \tag{120}$$

If $\langle x, x^0 + v\rangle_{\mathbb{L}}^2 \geq 1$, then

$$\begin{aligned} d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) &= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 - 1}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) \\ &= \log(-\langle x, x^0 + v\rangle_{\mathbb{L}}) = B^v(x). \end{aligned} \tag{121}$$

And if $\langle x, x^0 + v\rangle_{\mathbb{L}}^2 \leq 1$, then

$$\begin{aligned} d_{\mathbb{L}}\big(\tilde{B}^v(x), x^0\big) &= \log\left(\frac{1 + \langle x, x^0 + v\rangle_{\mathbb{L}}^2 + 1 - \langle x, x^0 + v\rangle_{\mathbb{L}}^2}{-2\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) \\ &= \log\left(\frac{1}{-\langle x, x^0 + v\rangle_{\mathbb{L}}}\right) = -B^v(x). \end{aligned} \tag{122}$$

Then, using that B^v is 1-lipschitz, we have

$$|B^v(x) - B^v(x^0)| \leq d_{\mathbb{L}}(x, x^0), \tag{123}$$

and therefore

$$|B^v(x)| \leq d_{\mathbb{L}}(x, x^0), \tag{124}$$

since $B^v(x^0) = \log(-\langle x^0, x^0 + v\rangle_{\mathbb{L}}) = 0$. Hence we have well $HHSW_p(\mu, \nu) < \infty$. □

B.2. Pseudo-distance

Proposition B.2. *Let $p \geq 1$, then $GHSW_p$ and $HHSW_p$ are pseudo-distances.*

Proof. Let $p \geq 1$, then for all $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$, it is straightforward to see that $GHSW_p(\mu, \nu) \geq 0$, $HHSW_p(\mu, \nu)_p \geq 0$, $GHSW_p(\mu, \nu) = GHSW_p(\nu, \mu)$ and $HHSW_p(\mu, \nu) = HHSW_p(\nu, \mu)$. It is also easy to see that $\mu = \nu \implies GHSW_p(\mu, \nu) = 0$ and $HHSW_p(\mu, \nu) = 0$ using that W_p is a distance.

Now, we can also derive the triangular inequality using the triangular inequality for W_p and the Minkowski inequality:

$$\begin{aligned}
\forall \mu, \nu, \alpha \in \mathcal{P}(\mathbb{L}^d),\ GHSW_p(\mu, \nu) &= \Big(\int_{T_{x^0}\mathbb{L}^d \cap S^d} W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)\, \mathrm{d}\lambda(v) \Big)^{\frac{1}{p}} \\
&\leq \Big(\int_{T_{x^0}\mathbb{L}^d \cap S^d} \big(W_p(P^v_{\#}\mu, P^v_{\#}\alpha) + W_p(P^v_{\#}\alpha, P^v_{\#}\nu) \big)^p\, \mathrm{d}\lambda(v) \Big)^{\frac{1}{p}} \\
&\leq \Big(\int_{T_{x^0}\mathbb{L}^d \cap S^d} W_p^p(P^v_{\#}\mu, P^v_{\#}\alpha)\, \mathrm{d}\lambda(v) \Big)^{\frac{1}{p}} \\
&+ \Big(\int_{T_{x^0}\mathbb{L}^d \cap S^d} W_p^p(P^v_{\#}\alpha, P^v_{\#}\nu)\, \mathrm{d}\lambda(v) \Big)^{\frac{1}{p}} \\
&= GHSW_p(\mu, \alpha) + GHSW_p(\alpha, \nu).
\end{aligned} \tag{125}$$

The same holds for HHSW.

Therefore, $GHSW_p$ and $HHSW_p$ are pseudo-distances. □

To show that there are distances, we need additionally the positivity property, *i.e.* we need to show that $GHSW_p(\mu, \nu) = 0 \implies \mu = \nu$. As W_p is a distance, we have that $GHSW_p(\mu, \nu) = 0 \implies P^v_{\#}\mu = P^v_{\#}\nu$ for λ-ae v. But showing that this implies that $\mu = \nu$ is not straightforward. Following derivations obtained with SW, we can draw connections with known Radon transforms.

Radon transform for GHSW. Let $f \in L^1(\mathbb{L}^d)$. Then, let's define $\bar{R} : L^1(\mathbb{L}^d) \to L^1(\mathbb{R} \times S^{d-1})$ such that for all $t \in \mathbb{R}$ and $v \in S^{d-1}$,

$$\bar{R}f(t, v) = \int_{\mathbb{L}^d} f(x) \mathbb{1}_{\{P^v(x) = t\}}\, \mathrm{d}x. \tag{126}$$

Let's define a dual function $\bar{R}^* : C_0(\mathbb{R} \times S^{d-1}) \to C_0(\mathbb{L}^d)$ as

$$\bar{R}^* g(x) = \int_{S^{d-1}} g\big(P^v(x), v\big)\, \mathrm{d}\lambda(v), \tag{127}$$

where $g \in C_0(\mathbb{R} \times S^{d-1})$. Then, we can check that it well the dual.

Proposition B.3. *For all $f \in L^1(\mathbb{L}^d)$, $g \in C_0(\mathbb{R} \times S^{d-1})$,*

$$\langle \bar{R}f, g\rangle_{\mathbb{R}\times S^{d-1}} = \langle f, \bar{R}^* g\rangle_{\mathbb{L}^d}. \tag{128}$$

Proof. Let $f \in L^1(\mathbb{L}^d)$, $g \in C_0(\mathbb{R} \times S^{d-1})$, then,

$$\begin{aligned}
\langle \bar{R}f, g\rangle_{\mathbb{R}\times S^{d-1}} &= \int_{\mathbb{R}\times S^{d-1}} \bar{R}f(t,v)g(t,v)\mathrm{d}t\mathrm{d}\lambda(v) \\
&= \int_{\mathbb{R}}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{L}^d} f(x)\mathbb{1}_{\{P^v(x)=t\}}g(t,v)\,\mathrm{d}x\mathrm{d}t\mathrm{d}\lambda(v) \\
&= \int_{\mathbb{L}^d} f(x)\int_{S^{d-1}}\int_{\mathbb{R}} g(t,v)\mathbb{1}_{\{P^v(x)=t\}}\,\mathrm{d}t\mathrm{d}\lambda(v)\mathrm{d}x \\
&= \int_{\mathbb{L}^d} f(x)\int_{S^{d-1}} g\big(P^v(x),v\big)\,\mathrm{d}\lambda(v)\mathrm{d}x \\
&= \langle f, \bar{R}g\rangle_{\mathbb{L}^d}.
\end{aligned} \tag{129}$$

□

Then, we can as in (Boman & Lindskog, 2009), define the corresponding Radon transform of a measure $\mu \in \mathcal{M}(\mathbb{L}^d)$ as the measure $\bar{R}\mu \in \mathcal{M}(\mathbb{R}\times S^{d-1})$, such that for all $g \in C_0(\mathbb{R}\times S^{d-1})$, $\langle \bar{R}\mu, g\rangle_{\mathbb{R}\times S^{d-1}} = \langle \mu, \bar{R}^* g\rangle_{\mathbb{L}^d}$.

Next, denoting for $v \in S^{d-1}$, $(\bar{R}\mu)^v$ the disintegrated measure *w.r.t.* λ, *i.e.* the measure satisfying for all $\phi \in C(\mathbb{R}\times S^{d-1})$,

$$\int_{\mathbb{R}\times S^{d-1}} \phi(t,v)\mathrm{d}(\bar{R}\mu)(t,v) = \int_{S^{d-1}}\int_{\mathbb{R}} \phi(t,v)(R\mu)^v(\mathrm{d}t)\,\mathrm{d}\lambda(v), \tag{130}$$

we can show that $(\bar{R}\mu)^v = P^v_\#\mu$.

Proposition B.4. *Let $\mu \in \mathcal{M}(\mathbb{L}^d)$, then for λ-almost every $v \in S^{d-1}$, $(\bar{R}\mu)^v = P^v_\#\mu$.*

Proof. In the following, we will use that $S^{d-1} \cong T_{x^0}\mathbb{L}^d \cap S^d$. And therefore, P^v is well defined.

Let $g \in C_0(\mathbb{R}\times S^{d-1})$, then

$$\begin{aligned}
\int_{S^{d-1}}\int_{\mathbb{R}} g(t,v)(\bar{R}\mu)^v(\mathrm{d}t)\,\mathrm{d}\lambda(v) &= \int_{\mathbb{R}\times S^{d-1}} g(t,v)\,\mathrm{d}(\bar{R}\mu)(t,v) = \langle \bar{R}\mu, g\rangle_{\mathbb{R}\times S^{d-1}} \\
&= \int_{\mathbb{L}^d} \bar{R}^* g(x)\,\mathrm{d}\mu(x) \\
&= \int_{\mathbb{L}^d}\int_{S^{d-1}} g\big(P^v(x),v\big)\,\mathrm{d}\lambda(x)\mathrm{d}\mu(x) \\
&= \int_{S^{d-1}}\int_{\mathbb{L}^d} g\big(P^v(x),v\big)\mathrm{d}\mu(x)\mathrm{d}\lambda(v) \\
&= \int_{S^{d-1}}\int_{\mathbb{R}} g(t,v)\,\mathrm{d}(P^v_\#\mu)(x)\mathrm{d}\lambda(v),
\end{aligned} \tag{131}$$

where we use the duality properties and Fubini. □

From the previous proposition, we deduce that

$$\forall \mu,\nu \in \mathcal{P}(\mathbb{L}^d),\ GHSW_p^p(\mu,\nu) = \int_{S^{d-1}} W_p^p\big((\bar{R}\mu)^v, (\bar{R}\nu)^v\big)\,\mathrm{d}\lambda(v). \tag{132}$$

And $GHSW_p(\mu,\nu) = 0 \implies (\bar{R}\mu)^v = (\bar{R}\nu)^v$ for λ-almost every λ.

The transformation $\bar{R}$ is not really clear written like that. In the next proposition, we identify the integration set, which will give a connection to a known Radon transform.

Proposition B.5 (Set of integration)**.** *The integration set of $\bar{R}$ is, for $t \in \mathbb{R}$, $v \in S^{d-1}$,*

$$\{x \in \mathbb{L}^d,\ P^v(x) = t\} = \mathrm{span}(v_z)^\perp \cap \mathbb{L}^d, \tag{133}$$

where $v_z = R_z v$ with R_z a rotation matrix in the plan $\mathrm{span}(v, x^0)$ such that $\langle v_z, z\rangle = 0$.

Proof. We will prove this proposition directly by working on the geodesics. As t^v is an isometry (Proposition 3.1), for all $t \in \mathbb{R}$, there exists a unique z on the geodesic $\text{span}(x^0, v) \cap \mathbb{L}^d$ such that $t = t^v(z)$, and we can rewrite the set of integration as

$$\{x \in \mathbb{L}^d,\ P^v(x) = t\} = \{x \in \mathbb{L}^d,\ \tilde{P}^v(x) = z\}. \tag{134}$$

For the first inclusion, let $x \in \{x \in \mathbb{L}^d,\ \tilde{P}^v(x) = z\}$. By Proposition A.1 and hypothesis, we have that

$$\tilde{P}^v(x) = \frac{1}{\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}}\big(- \langle x, x^0\rangle_{\mathbb{L}} x^0 + \langle x, v\rangle_{\mathbb{L}} v\big) = z. \tag{135}$$

Let's denote $E = \text{span}(v, x^0)$ the plan generating the geodesic. Then, by denoting P^E the orthogonal projection on E, we have

$$\begin{aligned} P^E(x) &= \langle x, v\rangle v + \langle x, x^0\rangle x^0 \\ &= \langle x, v\rangle_{\mathbb{L}} v - \langle x, x^0\rangle_{\mathbb{L}} x^0 \\ &= \left(\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}\right) z, \end{aligned} \tag{136}$$

using that $v_0 = 0$ since $\langle x^0, v\rangle = v_0 = 0$, and hence $\langle x, v\rangle_{\mathbb{L}} = \langle x, v\rangle$, that $\langle x, x^0\rangle = x_0 = -\langle x, x^0\rangle_{\mathbb{L}}$ and (135). Then, since $v_z \in \text{span}(v, x^0)$ and $\langle z, v_z\rangle = 0$ (by construction of R_z), we have

$$\begin{aligned} \langle x, v_z\rangle &= \langle P^E(x), v_z\rangle \\ &= \langle \left(\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}\right) z, v_z\rangle = 0. \end{aligned} \tag{137}$$

Thus, $x \in \text{span}(v_z)^\perp \cap \mathbb{L}^d$.

For the second inclusion, let $x \in \text{span}(v_z)^\perp \cap \mathbb{L}^d$. Since $z \in \text{span}(v_z)^\perp$ (by construction of R_z), we can decompose $\text{span}(v_z)^\perp$ as $\text{span}(v_z)^\perp = \text{span}(z) \oplus (\text{span}(z)^\perp \setminus \text{span}(v_z))$. Hence, there exists $\lambda \in \mathbb{R}$ such that $x = \lambda z + x^\perp$. Moreover, as $z \in \text{span}(x^0, v)$, we have $\langle x, x^0\rangle_{\mathbb{L}} = \lambda\langle z, x^0\rangle_{\mathbb{L}}$ and $\langle x, v\rangle_{\mathbb{L}} = \langle x, v\rangle = \lambda\langle z, v\rangle = \lambda\langle z, v\rangle_{\mathbb{L}}$. Thus, the projection is

$$\begin{aligned} \tilde{P}^v(x) &= \frac{1}{\sqrt{\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}}\big(- \langle x, x^0\rangle_{\mathbb{L}} x^0 + \langle x, v\rangle_{\mathbb{L}} v\big) \\ &= \frac{\lambda}{|\lambda|}\frac{1}{\sqrt{\langle z, x^0\rangle_{\mathbb{L}}^2 - \langle z, v\rangle_{\mathbb{L}}^2}}\big(- \langle z, x^0\rangle_{\mathbb{L}} x^0 + \langle z, v\rangle_{\mathbb{L}} v\big) \\ &= \frac{\lambda}{|\lambda|} z = \text{sign}(\lambda) z. \end{aligned} \tag{138}$$

But, $-z \notin \mathbb{L}^d$, hence necessarily, $\tilde{P}^v(x) = z$.

Finally, we can conclude that $\{x \in \mathbb{L}^d,\ \tilde{P}^v(x) = z\} = \text{span}(v_z)^\perp \cap \mathbb{L}^d$. □

From the previous proposition, we see that the Radon transform $\bar{R}$ integrates over hyperplanes intersected with $\mathbb{L}^d$, which are totally geodesic submanifolds. This corresponds actually to the hyperbolic Radon transform first introduced by Helgason (1959) and studied more recently for example in (Berenstein & Rubin, 1999; Rubin, 2002; Berenstein & Rubin, 2004). However, to the best of our knowledge, its injectivity over the set of measures has not been studied yet.

Radon transform for HHSW. We can derive a Radon transform associated to HHSW in the same way. Moreover, the integration set can be intuitively derived as the level set of the Busemann function, since we project on the only point on the geodesic which has the same Busemann coordinates. Since the level sets of the Busemann functions correspond to horospheres, the associate Radon transform is the horospherical Radon transform. It has been for example studied by Bray & Rubin (1999; 2019) on the Lorentz model, and by Casadio Tarabusi & Picardello (2021) on the Poincaré ball. Note that it is also known as the Gelfand-Graev transform (Gelfand et al., 1966).

B.3. Statistical Properties

Sample Complexity. By adapting the proof of (Nadjahi et al., 2020, Corollary 2), we derive a sample complexity in Proposition B.6 for both $GHSW_p$ and $HHSW_p$. Interestingly, they are similar up to some constant. Moreover, similarly as the Euclidean SW distance, they are independent of the dimension.

Proposition B.6. *Let $p \geq 1$, $q > p$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$. Denote $\hat{\mu}_n$ and $\hat{\nu}_n$ their counterpart empirical measures and $M_q(\mu) = \int_{\mathbb{L}^d} d(x, x^0)^q \, \mathrm{d}\mu(x)$ the moments of order q. Then, there exists $C_{p,q}$ a constant depending only on p and q such that*

$$\mathbb{E}\left[|GHSW_p(\hat{\mu}_n, \hat{\nu}_n) - GHSW_p(\mu, \nu)|\right] \leq 2^{q/p} C_{p,q}^{1/p} (M_q(\mu)^{1/q} + M_q(\nu)^{1/q}) \begin{cases} n^{-1/(2p)} & \text{if } q > 2p \\ n^{-1/(2p)} \log(n)^{1/p} & \text{if } q = 2p \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases} \tag{139}$$

Similarly, with $C_{p,q}$ a possible another constant,

$$\mathbb{E}\left[|HHSW_p(\hat{\mu}_n, \hat{\nu}_n) - HHSW_p(\mu, \nu)|\right] \leq 2 C_{p,q}^{1/p} (M_q(\mu)^{1/q} + M_q(\nu)^{1/q}) \begin{cases} n^{-1/(2p)} & \text{if } q > 2p \\ n^{-1/(2p)} \log(n)^{1/p} & \text{if } q = 2p \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases} \tag{140}$$

Proof. For this proof, we first need to recall the following lemma adapted from (Fournier & Guillin, 2015, Theorem 2) and reported *e.g.* in (Rakotomamonjy et al., 2021).

Lemma B.7 (Lemma 1 in (Rakotomamonjy et al., 2021))**.** *Let $p \geq 1$ and $\eta \in \mathcal{P}_p(\mathbb{R})$. Denote $\tilde{M}_q(\eta) = \int |x|^q \, \mathrm{d}\eta(x)$ the moments of order q and assume that $M_q(\eta) < \infty$ for some $q > p$. Then, there exists a constant $C_{p,q}$ depending only on p, q such that for all $n \geq 1$,*

$$\mathbb{E}[W_p^p(\hat{\eta}_n, \eta)] \leq C_{p,q} \tilde{M}_q(\eta)^{p/q} \left(n^{-1/2} \mathbb{1}_{\{q > 2p\}} + n^{-1/2} \log(n) \mathbb{1}_{\{q = 2p\}} + n^{-(q-p)/q} \mathbb{1}_{\{q \in (p, 2p)\}} \right) . \tag{141}$$

Now, we will first deal with $GHSW_p$. First, let us observe that by the triangular and reverse triangular inequalities, as well as Jensen for $x \mapsto x^{1/p}$ (which is concave since $p \geq 1$),

$$\begin{aligned} \mathbb{E}\left[|\mathrm{GHSW}_p(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{GHSW}_p(\mu, \nu)|\right] &= \mathbb{E}[|\mathrm{GHSW}_p(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{GHSW}_p(\hat{\mu}_n, \nu) + \mathrm{GHSW}_p(\hat{\mu}_n, \nu) - \mathrm{GHSW}_p(\mu, \nu)|] \\ &\leq \mathbb{E}[|\mathrm{GHSW}_p(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{GHSW}_p(\hat{\mu}_n, \nu)|] + \mathbb{E}[|\mathrm{GHSW}_p(\hat{\mu}_n, \nu) - \mathrm{GHSW}_p(\mu, \nu)|] \\ &\leq \mathbb{E}[\mathrm{GHSW}_p(\hat{\nu}_n, \nu)] + \mathbb{E}[\mathrm{GHSW}_p(\hat{\mu}_n, \mu)] \\ &\leq \mathbb{E}[\mathrm{GHSW}_p^p(\hat{\nu}_n, \nu)]^{1/p} + \mathbb{E}[\mathrm{GHSW}_p^p(\hat{\mu}_n, \mu)]^{1/p}. \end{aligned} \tag{142}$$

Moreover, by Fubini-Tonelli,

$$\begin{aligned} \mathbb{E}[\mathrm{GHSW}_p^p(\hat{\mu}_n, \mu)] &= \mathbb{E}\left[\int_{T_{x^0} \mathbb{L}^d \cap S^d} W_p^p(P_\#^v \hat{\mu}_n, P_\#^v \mu) \, \mathrm{d}\lambda(v) \right] \\ &= \int_{T_{x^0} \mathbb{L}^d \cap S^d} \mathbb{E}[W_p^p(P_\#^v \hat{\mu}_n, P_\#^v \mu)] \, \mathrm{d}\lambda(v). \end{aligned} \tag{143}$$

By applying Lemma B.7, we get for $q > p$ that there exists a constant $C_{p,q}$ such that,

$$\mathbb{E}[W_p^p(P_\#^v \hat{\mu}_n, P_\#^v \mu)] \leq C_{p,q} \tilde{M}_q(P_\#^v \mu)^{p/q} \left(n^{-1/2} \mathbb{1}_{\{q > 2\}} + n^{-1/2} \log(n) \mathbb{1}_{\{q = 2p\}} + n^{-(q-p)/q} \mathbb{1}_{\{q \in (p, 2p)\}} \right) . \tag{144}$$

Furthermore, using (Villani, 2009, Defintion 6.4), *i.e.* that

$$\forall x, y, x_0 \in \mathbb{L}^d, \; d_{\mathbb{L}}(x, y)^p \leq 2^{p-1} \left(d_{\mathbb{L}}(x, x_0) + d_{\mathbb{L}}(x_0, y) \right) , \tag{145}$$

we obtain

$$d(\tilde{P}^v(x), x^0)^q \leq 2^{q-1} \left(d(\tilde{P}^v(x), x)^q + d(x, x^0)^q \right) , \tag{146}$$

and by definition of $\tilde{P}^v$, $d_{\mathbb{L}}(\tilde{P}^v(x), x) \leq d_{\mathbb{L}}(x^0, x)$. Hence, remembering that $t^v(x) = \text{sign}(\langle x, v\rangle) d_{\mathbb{L}}(x, x^0)$ and $P^v(x) = t^v(\tilde{P}^v(x))$, we have

$$\begin{aligned}
\tilde{M}_q(P^v_{\#}\mu) &= \int_{\mathbb{R}} |x|^q \, \mathrm{d}(P^v_{\#}\mu)(x) \\
&= \int_{\mathbb{L}^d} |P^v(x)|^q \, \mathrm{d}\mu(x) \\
&= \int_{\mathbb{L}^d} |t^v(\tilde{P}^v(x))|^q \, \mathrm{d}\mu(x) \\
&= \int_{\mathbb{L}^d} d_{\mathbb{L}}(\tilde{P}^v(x), x^0)^q \, \mathrm{d}\mu(x) \\
&\leq 2^{q-1} \left(\int_{\mathbb{L}^d} d_{\mathbb{L}}(\tilde{P}^v(x), x)^q \, \mathrm{d}\mu(x) + \int_{\mathbb{L}^d} d_{\mathbb{L}}(x, x^0)^q \, \mathrm{d}\mu(x) \right) \\
&\leq 2^q \int_{\mathbb{L}^q} d_{\mathbb{L}}(x, x^0)^q \, \mathrm{d}\mu(x) = 2^q M_q(\mu).
\end{aligned} \tag{147}$$

Therefore, we have that

$$\mathbb{E}[\text{GHSW}_p^p(\hat{\mu}_n, \mu)] \leq 2^q C_{p,q} M_q(\mu)^{p/q} \left(n^{-1/2} \mathbb{1}_{\{q>2p\}} + n^{-1/2} \log(n) \mathbb{1}_{\{q=2p\}} + n^{-(q-p)/q} \mathbb{1}_{\{q\in(p,2p)\}} \right), \tag{148}$$

and similarly

$$\mathbb{E}[\text{GHSW}_p^p(\hat{\nu}_n, \nu)] \leq C_{p,q} M_q(\nu)^{p/q} \left(n^{-1/2} \mathbb{1}_{\{q>2p\}} + n^{-1/2} \log(n) \mathbb{1}_{\{q=2p\}} + n^{-(q-p)/q} \mathbb{1}_{\{q\in(p,2p)\}} \right). \tag{149}$$

Hence, we conclude that the sample complexity is

$$\mathbb{E}\left[|\text{GHSW}_p(\hat{\mu}_n, \hat{\nu}_n) - \text{GHSW}_p(\mu, \nu)|\right] \leq 2^{q/p} C_{p,q}^{1/p} \left(M_q(\mu)^{1/q} + M_q(\nu)^{1/q} \right) \begin{cases} n^{-1/(2p)} & \text{if } q > 2p \\ n^{-1/(2p)} \log(n)^{1/p} & \text{if } q = 2p \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases} \tag{150}$$

Now, we can also do the same proof for $HHSW_p$. By using pseudo-distance properties, we also get

$$\mathbb{E}\left[|\text{HHSW}_p(\hat{\mu}_n, \hat{\nu}_n) - \text{HHSW}_p(\mu, \nu)|\right] \leq \mathbb{E}[\text{HHSW}_p^p(\hat{\nu}_n, \nu)]^{1/p} + \mathbb{E}[\text{HHSW}_p^p(\hat{\mu}_n, \mu)]^{1/p}, \tag{151}$$

and with Fubini-Tonelli,

$$\mathbb{E}[HHSW_p^p(\hat{\mu}_n, \mu)] = \int_{S^{d-1}} \mathbb{E}[W_p^p(t^v_{\#}\tilde{P}^v_{\#}\hat{\mu}_n, t^v_{\#}\tilde{P}^v_{\#}\mu)] \, \mathrm{d}\lambda(v). \tag{152}$$

Then, by Lemma B.7, we have that for $q > p$, there exists a constant $C_{p,q}$ such that,

$$\mathbb{E}[W_p^p(B^v_{\#}\hat{\mu}_n, B^v_{\#}\mu)] \leq C_{p,q} \tilde{M}_q(B^v_{\#}\mu)^{p/q} \left(n^{-1/2} \mathbb{1}_{\{q>2\}} + n^{-1/2} \log(n) \mathbb{1}_{\{q=2p\}} + n^{-(q-p)/q} \mathbb{1}_{\{q\in(p,2p)\}} \right). \tag{153}$$

But, as B^v is 1-Lipschitz, and $B^v(x^0) = 0$, we have that $|B^v(x) - B^v(x^0)| = |B^v(x)| \leq d_{\mathbb{L}}(x, x^0)$ for all $x \in \mathbb{L}^d$. Hence,

$$\tilde{M}_q(B^v_{\#}\mu) = \int_{\mathbb{L}^d} |B^v(x)|^q \, \mathrm{d}\mu(x) \leq \int_{\mathbb{L}^d} d_{\mathbb{L}}(x, x^0)^q \, \mathrm{d}\mu(x) = M_q(\mu), \tag{154}$$

and therefore

$$\mathbb{E}\left[|HHSW_p(\hat{\mu}_n, \hat{\nu}_n) - HHSW_p(\mu, \nu)|\right] \leq 2 C_{p,q}^{1/p} \left(M_q(\mu)^{1/q} + M_q(\nu)^{1/q} \right) \begin{cases} n^{-1/(2p)} & \text{if } q > 2p \\ n^{-1/(2p)} \log(n)^{1/p} & \text{if } q = 2p \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases} \tag{155}$$

□

Projection Complexity. The integral *w.r.t* the uniform measure on S^{d-1} is unfortunately intractable, and therefore is required to be approximated by a Monte-Carlo scheme. In Proposition B.8, we report the Monte-Carlo error of this approximation. We call this error the projection complexity. We recover here the same rate as (Nadjahi et al., 2020) in the Euclidean case. Since the proposition and the proof are the same for $GHSW_p$ and $HHSW_p$, we do it for both in the same time by denoting HSW_p in place of $GHSW_p$ or $HHSW_p$, and denoting by P^v the corresponding projection with $v \in T_{x^0}\mathbb{L}^d \cap S^d$.

Proposition B.8. *Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$. We denote HSW_p for both $HHSW_p$ and $GHSW_p$. Then, the error made by the Monte Carlo estimate of HSW_p with L projections can be bounded as follows*

$$\begin{aligned}\mathbb{E}_v\left[\left|\widehat{\mathrm{HSW}}_{p,L}^p(\mu,\nu) - \mathrm{HSW}_p^p(\mu,\nu)\right|\right]^2 &\leq \frac{1}{L}\int_{T_{x^0}\mathbb{L}^d\cap S^d}\left(W_p^p(P^v_{\#}\mu, P^v_{\#}\nu) - \mathrm{HSW}_p^p(\mu,\nu)\right)^2 \mathrm{d}\lambda(v)\\ &= \frac{1}{L}\mathrm{Var}_{v\sim\lambda}\left[W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)\right],\end{aligned} \tag{156}$$

where $\widehat{\mathrm{HSW}}_{p,L}^p(\mu,\nu) = \frac{1}{L}\sum_{i=1}^L W_p^p(P^{v_i}_{\#}\mu, P^{v_i}_{\#}\nu)$ with $(v_i)_{i=1}^L$ independent samples from λ.

Proof. Let $(v_i)_{i=1}^L$ be iid samples of λ. Then, by first using Jensen inequality and then remembering that $\mathbb{E}_v[W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)] = \mathrm{HSW}_p^p(\mu,\nu)$, we have

$$\begin{aligned}\mathbb{E}_v\left[\left|\widehat{\mathrm{HSW}}_{p,L}^p(\mu,\nu) - \mathrm{HSW}_p^p(\mu,\nu)\right|\right]^2 &\leq \mathbb{E}_v\left[\left|\widehat{\mathrm{HSW}}_{p,L}^p(\mu,\nu) - \mathrm{HSW}_p^p(\mu,\nu)\right|^2\right]\\ &= \mathbb{E}_v\left[\left|\frac{1}{L}\sum_{i=1}^L\left(W_p^p(P^{v_i}_{\#}\mu, P^{v_i}_{\#}\nu) - \mathrm{HSW}_p^p(\mu,\nu)\right)\right|^2\right]\\ &= \frac{1}{L^2}\mathrm{Var}_v\left[\sum_{i=1}^L W_p^p(P^{v_i}_{\#}\mu, P^{v_i}_{\#}\nu)\right]\\ &= \frac{1}{L}\mathrm{Var}_v\left[W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)\right]\\ &= \frac{1}{L}\int_{T_{x^0}\mathbb{L}^d\cap S^d}\left(W_p^p(P^v_{\#}\mu, P^v_{\#}\nu) - \mathrm{HSW}_p^p(\mu,\nu)\right)^2 \mathrm{d}\lambda(v).\end{aligned} \tag{157}$$

□

C. Hyperbolic Spaces

In this Section, we first recall different generalization of the Gaussian distribution on Hyperbolic spaces, with a particular focus on Wrapped normal distributions. Then, we recall how to perform Riemannian gradient descent in the Lorentz model and in the Poincaré ball.

C.1. Distributions on Hyperbolic Spaces

Let M be a manifold and denote G the corresponding Riemannian metric. For $x \in M$, $G(x)$ induces an infinitesimal change of volume on the tangent space T_xM, and thus a measure on the manifold,

$$\mathrm{dVol}(x) = \sqrt{|G(x)|}\,\mathrm{d}x.$$

We refer to (Pennec, 2006) for more details on distributions on manifolds. Now, we recap different generalizations of Gaussian distribution on Riemannian manifolds.

Riemannian normal. The first way of naturally generalizing Gaussian distributions to Riemannian manifolds is to use the geodesic distance in the density, which becomes

$$f(x) \propto \exp\left(-\frac{1}{2\sigma^2}d_M(x,\mu)^2\right).$$

It is actually the distribution maximizing the entropy (Pennec, 2006; Said et al., 2014). However, it is not straightforward to sample from such distribution. For example, Ovinnikov (2019) use a rejection sampling algorithm.

Wrapped normal distribution. A more convenient distribution, on which we can use the parameterization trick, is the Wrapped normal distribution (Nagano et al., 2019). This distribution can be sampled from by first drawing $v \sim \mathcal{N}(0, \Sigma)$ and then transforming it into $v \in T_{x^0}\mathbb{L}^d$ by concatenating a 0 in the first coordinate. Then, we perform parallel transport to transport v from the tangent space of x^0 to the tangent space of $\mu \in \mathbb{L}^d$. Finally, we can project the samples on the manifold using the exponential map. We recall the formula of parallel transport form x to y:

$$\forall v \in T_x\mathbb{L}^d,\ \mathrm{PT}_{x\to y}(v) = v + \frac{\langle y, v\rangle_{\mathbb{L}}}{1 - \langle x, y\rangle_{\mathbb{L}}}(x + y). \tag{158}$$

Since it only involves differentiable operations, we can perform the parameterization trick and *e.g.* optimize directly over the mean and the variance. Moreover, by the change of variable formula, we can also derive the density (Nagano et al., 2019; Bose et al., 2020). Let $\tilde{z} \sim \mathcal{N}(0, \Sigma)$, $z = (0, \tilde{z}) \in T_{x^0}\mathbb{L}^d$, $u = \mathrm{PT}_{x^0\to\mu}(z)$, then the density of $x = \exp_\mu(u)$ is:

$$\log p(x) = \log p(\tilde{z}) - (d-1)\log\left(\frac{\sinh(\|u\|_{\mathbb{L}})}{\|u\|_{\mathbb{L}}}\right). \tag{159}$$

In the paper, we write $x \sim \mathcal{G}(\mu, \Sigma)$.

C.2. Optimization on Hyperbolic Spaces

For gradient descent on hyperbolic space, we refer to (Boumal, 2022, Section 7.6) and (Wilson & Leimeister, 2018).

In general, for a functional $f : M \to \mathbb{R}$, Riemannian gradient descent is performed, analogously to the Euclidean space, by following the geodesics. Hence, the gradient descent reads as (Absil et al., 2009; Bonnabel, 2013)

$$\forall k \geq 0,\ x_{k+1} = \exp_{x_k}\big(-\gamma \mathrm{grad} f(x_k)\big). \tag{160}$$

Note that the exponential map can be replaced more generally by a retraction. We describe in the following paragraphs the different formulae in the Lorentz model and in the Poincaré ball.

Lorentz model. Let $f : \mathbb{L}^d \to \mathbb{R}$, then its Riemannian gradient is (Boumal, 2022, Proposition 7.7)

$$\mathrm{grad} f(x) = \mathrm{Proj}_x(J\nabla f(x)), \tag{161}$$

where $J = \mathrm{diag}(-1, 1, \dots, 1)$ and $\mathrm{Proj}_x(z) = z + \langle x, z\rangle_{\mathbb{L}} x$. Furthermore, the exponential map is

$$\forall v \in T_x\mathbb{L}^d,\ \exp_x(v) = \cosh(\|v\|_{\mathbb{L}})x + \sinh(\|v\|_{\mathbb{L}})\frac{v}{\|v\|_{\mathbb{L}}}. \tag{162}$$

Poincaré ball. On $\mathbb{B}^d$, the Riemannian gradient of $f : \mathbb{B}^d \to \mathbb{R}$ can be obtained as (Nickel & Kiela, 2017, Section 3)

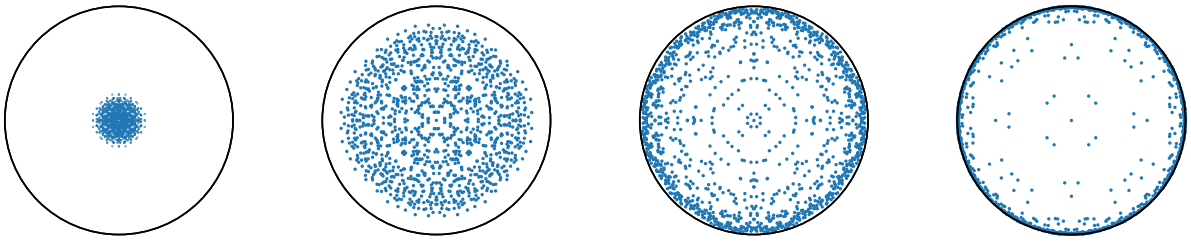
$$\mathrm{grad} f(x) = \frac{(1 - \|\theta\|_2^2)^2}{4}\nabla f(x). \tag{163}$$

Nickel & Kiela (2017) propose to use as retraction $R_x(v) = x + v$ instead of the exponential map, and add a projection, to constrain the value to remain within the Poincaré ball, of the form

$$\mathrm{proj}(x) = \begin{cases} \frac{x}{\|x\|_2} - \epsilon & \text{if } \|x\| \geq 1 \\ x & \text{otherwise,} \end{cases} \tag{164}$$

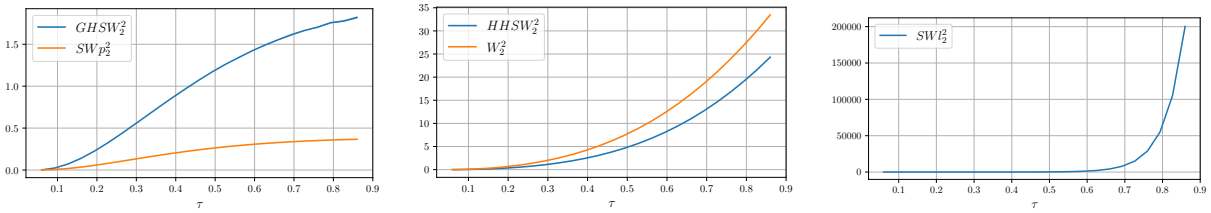
where $\epsilon = 10^{-5}$ is a small constant ensuring numerical stability. Hence, the algorithm becomes

$$x_{k+1} = \mathrm{proj}\left(x_k - \gamma_k \frac{(1 - \|x_k\|_2^2)^2}{4}\nabla f(x_k)\right). \tag{165}$$



(a) $\tau = 0.05$ (b) $\tau = 0.25$ (c) $\tau = 0.5$ (d) $\tau = 0.8$

Figure 5: Embeddings of trees using Sarkar's algorithm with different τ.



(a) SW on Poincaré (SWp), GHSW (b) HHSW and Wasserstein (c) SW on Lorentz (SWl)

Figure 6: Comparison of the Wasserstein distance (with the geodesic distance as cost), GHSW, HHSW and SW between embedded trees. We gather the discrepancies together by scale of the values.

A second solution is to compute directly the exponential map derived in (Ganea et al., 2018a, Corollary 1.1):

$$\exp_x(v) = \frac{\lambda_x\big(\cosh(\lambda_x\|v\|_2) + \langle x, \frac{v}{\|v\|_2}\rangle \sinh(\lambda_x\|v\|_2)\big)x + \frac{1}{\|v\|_2}\sinh(\lambda_x\|v\|_2)v}{1 + (\lambda_x - 1)\cosh(\lambda_x\|v\|_2) + \lambda_x\langle x, \frac{v}{\|v\|_2}\rangle \sinh(\lambda_x\|v\|_2)}, \tag{166}$$

where $\lambda_x = \frac{2}{1-\|x\|_2^2}$.

D. Additional Details and Experiments

D.1. Comparisons

In Section 5, we compare the evolution of GHSW, HHSW, SWl, SWp and the Wasserstein distance with geodesic cost between wrapped normal distributions. Here, we add a more "hyperbolical" setting in the sense that we compare trees embedded in hyperbolic space. Indeed, it is well known that hyperbolic spaces can be seen as a continuous analog of trees, and are therefore a natural embedding space for trees.

More precisely, we generate balanced trees using NetworkX (Hagberg et al., 2008) and embed them with Sarkar's algorithm (Sarkar, 2011; Sala et al., 2018). This algorithm takes as input a scaling factor τ which determines how close to the border will the leaves be. We illustrate such embeddings with different τ on Figure 5. We compare in Figure 6 the evolution of GHSW, HHSW, SWl and SWp between a tree embedded very close to the origin with $\tau = 0.05$ and τ growing towards 1. We observe here the same evolution than in Section 5.

Sample complexity. We showed in Proposition B.6 that the sample complexity of $HHSW_p$ and $GHSW_p$ does not depend on the dimension. We verify here on Figure 7 empirically this property for GHSW and HHSW between two set of samples drawn from $\mathcal{G}(x^0, I_2)$, and computed with 1000 projections. In dimension 3 and 50, HHSW and GHSW have the same convergence speed *w.r.t.* the number of samples, which is not the case for the Wasserstein distance which suffers from the curse of dimensionality.

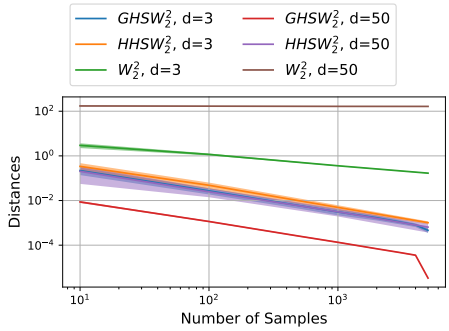


Figure 7: Sample complexity of GHSW, HHSW and Wasserstein with geodesic distance. GHSW and HHSW have the same convergence rate in dimension 3 and 50.

D.2. Gradient flows.

Denoting ν the target distribution from which we have access to samples $(y_i)_{i=1}^m$, we aim at learning ν by solving the following optimization problem:

$$\mu = \underset{\mu}{\operatorname{argmin}} \; HSW\left(\mu, \frac{1}{m}\sum_{i=1}^{m}\delta_{x_i}\right). \tag{167}$$

As we cannot directly learn μ, we model it as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and then learn the sample locations $(x_i)_{i=1}^n$ using a Riemannian gradient descent which we described in Appendix C.2. In practice, we take $n = 500$ and use batchs of 500 target samples at each iteration. To compute the sliced discrepancies, we always use 1000 projections. On Figure 4, we plot the log 2-Wasserstein with geodesic cost between the model measure $\hat{\mu}_k$ at each iteration k and ν. We average over 5 runs of each gradient descent. Now, we describe the specific setting for the different targets.

Wrapped normal distribution. For the first experiment, we choose as target a wrapped normal distribution $\mathcal{G}(m, \Sigma)$. In the fist setting, we use $m = (1.5, 1.25, 0) \in \mathbb{L}^2$ and $\Sigma = 0.1I_2$. In the second, we use $m = (8, \sqrt{63}, 0) \in \mathbb{L}^2$ and $\Sigma = 0.1I_2$. The learning rate is fixed as 5 for the different discrepancies, except for SWl on the second WND which lies far from origin, and for which we exhibit numerical instabilities with a learning rate too high. Hence, we reduced it to 0.1. We observed the same issue for HHSW on the Lorentz model. Fortunately, the Poincaré version, which is equal to the Lorentz version, did not suffer from these issues. It underlines the benefit of having both formulations.

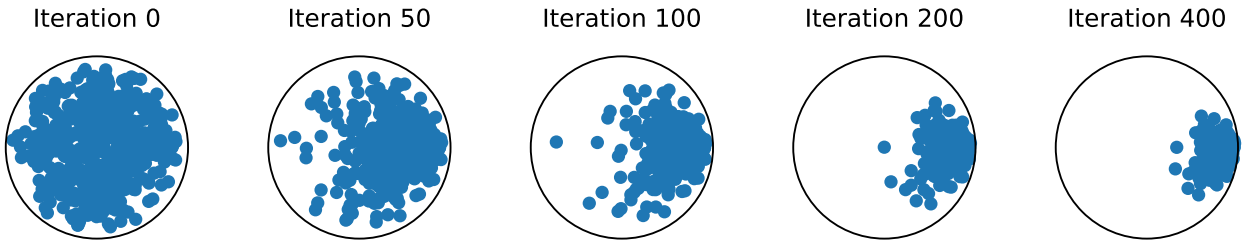
On Figure 8, we plotted the evolution of the particles for HHSW and GHSW with a target with mean $m = (8, \sqrt{63}, 0)$ and $\Sigma = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$. For GHSW, we use a learning rate of 10, and for HSHW a learning rate of 100. We observe that the trajectories are differents. With geodesic projections, the particles go towards the target by passing through the origin, while with horospherical projections, the tend first to leave the origin.

Mixture of wrapped normal distributions. For the second experiment, the target is a mixture of 5 WNDs. The covariance are all taken equal as $0.01I_2$. For the first setting, the outlying means are (on the Poincaré ball) $m_1 = (0, -0.5)$, $m_2 = (0, 0.5)$, $m_3 = (0.5, 0)$, $m_4 = (-0.5, 0)$ and the center mean is $m_5 = (0, 0.1)$. In the second setting, the outlying means are $m_1 = (0, -0.9)$, $m_2 = (0, 0.9)$, $m_3 = (0.9, 0)$ and $m_4 = (-0.9, 0)$. We use the same m_5. The learning rate in this experiment is fixed at 1 for all discrepancies.

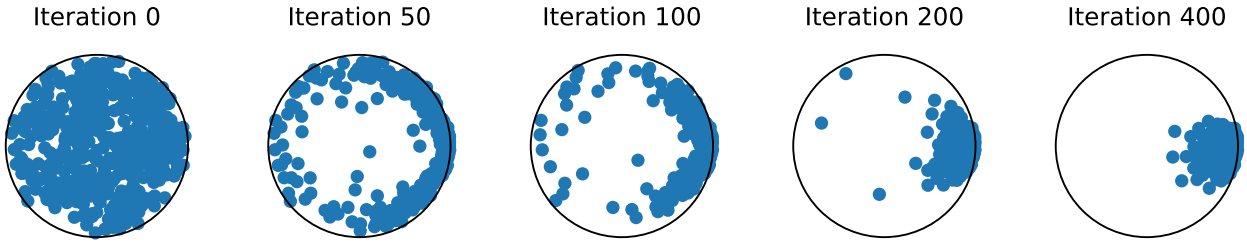
D.3. Classification of Images with Busemann

Denote $\{(x_i, y_i)_{i=1}^n\}$ the training set where $x_i \in \mathbb{R}^m$ and $y_i \in \{1, \dots, C\}$ is a label. The embedding is performed by using a neural network f_θ and the exponential map at the last layer, which projects the points on the Poincaré ball, *i.e.* for $i \in \{1, \dots, n\}$, the embedding of x_i is $z_i = \exp_0\left(f_\theta(z_i)\right)$, where $\exp_0$ is given by (166), or more simply by

$$\exp_0(x) = \tanh\left(\frac{\|x\|_2}{2}\right)\frac{x}{\|x\|_2}. \tag{168}$$



(a) With geodesic projection.



(b) With horospherical projection.

Figure 8: Evolution of the particles along the gradient flow of HSW (with geodesic or horospherical projection).

The experimental setting of this experiment is the same as (Ghadimi Atigh et al., 2021). That is, we use a Resnet-32 backbone and optimize it with Adam (Kingma & Ba, 2014), a learning rate of 5e-4, weight decay of 5e-5, batch size of 128 and without pre-training. The network is trained for all experiments for 1110 epochs with learning rate decay of 10 after 1000 and 1100 epochs. Moreover, the C prototypes are given by the algorithm of (Mettes et al., 2019) and are uniform on the sphere S^{d-1}.

For the additional hyperparameters in the loss (34), we use by default $\lambda = 1$, and a mixture of C wrapped normal distributions with means αp_c, where $p_c \in S^{d-1}$ is a prototype, $c \in \{1, \dots, C\}$ and $\alpha = 0.75$, and covariance matrix σI_d with $\sigma = 0.1$. The number of projection is by default set at L=1000.

D.4. Hyperbolic Sliced-Wassertein Autoencoder

As hyperbolic spaces allow to embed hierarchical data, it has been proposed in several works to put a prior on such space for autoencoder tasks (Ovinnikov, 2019; Nagano et al., 2019; Mathieu et al., 2019). Usually, an uninformative prior such as a Wrapped normal or a Riemannian normal distribution is used. For such distributions, the density is known and hence the Kullback-Leibler divergence can be approximated by a Monte-Carlo scheme. Moreover, we can also use the reparametrization trick. Then, a variational auto-encoder (Kingma & Welling, 2013) can be used. For more complicated distributions or deterministic prior with no density, we can use Wasserstein autoencoders (Tolstikhin et al., 2017). In this case, with a prior p_Z for which we have access to samples, an encoder f mapping the distribution data μ to the latent space, and a decoder g, we aim at minimizing the following loss:

$$\mathcal{L}(f, g) = \int c(x, g(f(x))) \, \mathrm{d}\mu(x) + D(f_{\#}\mu, p_Z), \tag{169}$$

with c some cost function and D some divergence. Several divergences D were proposed such as the MMD or SW (Kolouri et al., 2018). We propose here to study the latent space when using a tree prior, for which we cannot use a variational autoencoder. To learn the distribution in the latent space, we use a hyperbolic sliced discrepancy.

On Figure 9, we compare several priors on the Mnist dataset (LeCun & Cortes, 2010) with $D = HHSW_2^2$, which we denote HHSWAE. First, we use a Wrapped Normal distribution, and then a binary and a ternary tree as a prior. The trees are generated with NetworkX and embedded using Sarkar's algorithm, with $\tau = 0.6$ for the ternary tree and $\tau = 0.4$ for the binary tree. Moreover, we use a height of 3 for the ternary tree and of 4 for the binary one. For the HHSWAE, we used 200

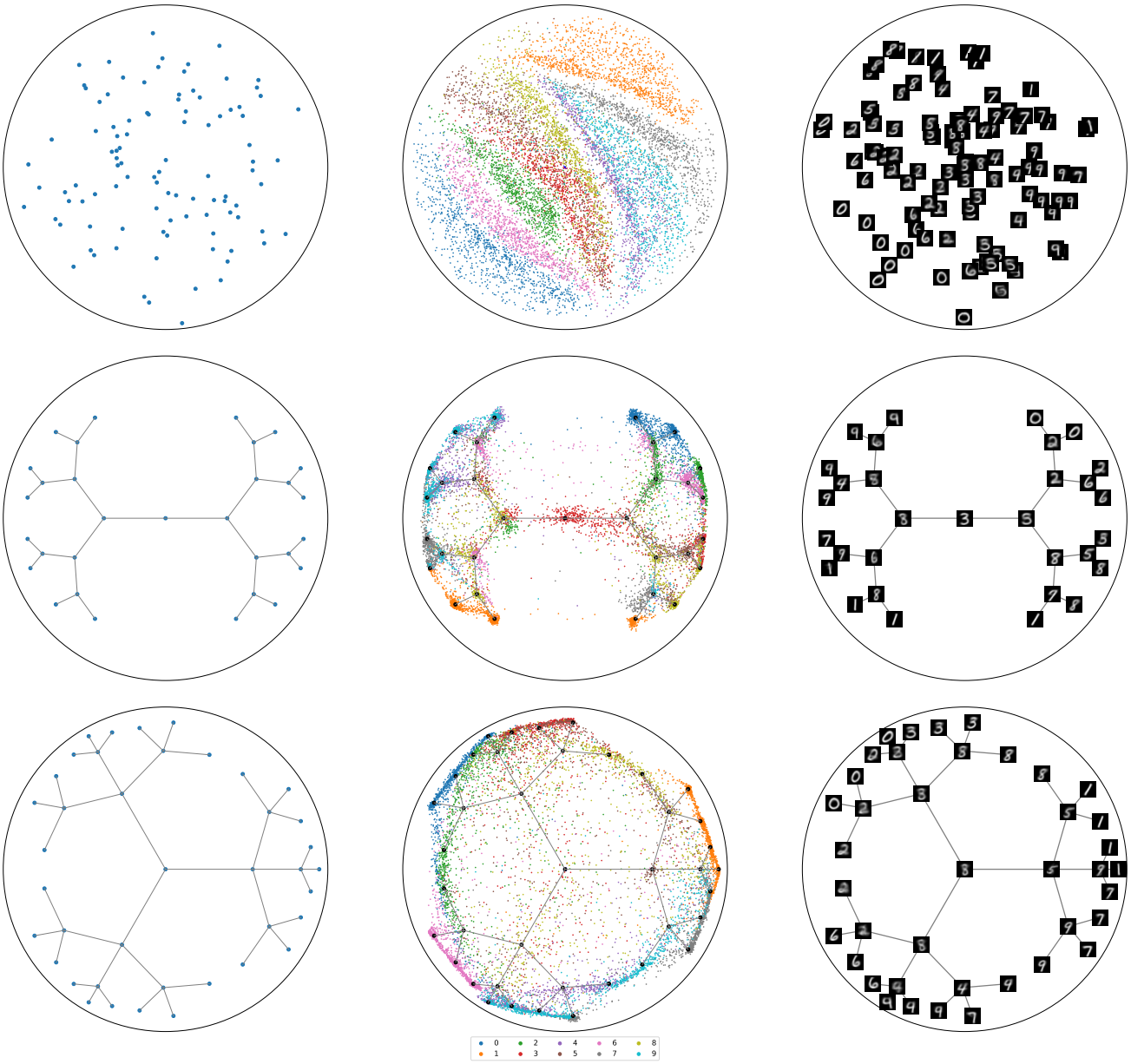


Figure 9: Embedding and reconstruction for HHSWAE. In the first column, we plot the prior. In the second column, we plot the embedding of MNIST and in the third column, we plot the reconstructed nodes of the tree or from samples of the wrapped normal distribution. In the first row, the prior is a Wrapped Normal Distribution. In the second row, the prior is a binary tree and in the third row a ternary tree.

epochs with the same architectures as (Kolouri et al., 2018) with an exp map before the output of the encoder, and a log map at the input of the decoder.

We observe that when using a tree prior, the points from the same class tend to be distributed around the same nodes. We believe that such an hierarchical prior can be beneficial in cases where one already has an assumption about the natural structure of the data. Next works will consider this question more thoroughly in different applicative settings.