Journal of Machine Learning Research 26 (2025) 1-76 Submitted 3/24; Revised 12/24; Published 1/25

Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds

Clément Bonet CLEMENT.BONET@ENSAE.FR
ENSAE, CREST, Institut Polytechnique de Paris

Lucas Drumetz LUCAS.DRUMETZ@IMT-ATLANTIQUE.FR
IMT Atlantique, Lab-STICC

Nicolas Courty NICOLAS.COURTY@IRISA.FR
Université Bretagne Sud, IRISA

Editor: Shiqian Ma

Abstract

While many Machine Learning methods have been developed or transposed on Riemannian manifolds to tackle data with known non-Euclidean geometry, Optimal Transport (OT) methods on such spaces have not received much attention. The main OT tool on these spaces is the Wasserstein distance, which suffers from a heavy computational burden. On Euclidean spaces, a popular alternative is the Sliced-Wasserstein distance, which leverages a closed-form solution of the Wasserstein distance in one dimension, but which is not readily available on manifolds. In this work, we derive general constructions of Sliced-Wasserstein distances on Cartan-Hadamard manifolds, Riemannian manifolds with non-positive curvature, which include among others Hyperbolic spaces or the space of Symmetric Positive Definite matrices. Then, we propose different applications such as classification of documents with a suitably learned ground cost on a manifold, and data set comparison on a product manifold. Additionally, we derive non-parametric schemes to minimize these new distances by approximating their Wasserstein gradient flows.

Keywords: Optimal Transport, Sliced-Wasserstein, Riemannian Manifolds, Cartan-Hadamard manifolds, Wasserstein Gradient Flows

1. Introduction

It is widely accepted that data have an underlying structure on a low-dimensional manifold (Bengio et al., 2013). However, it can be challenging to work directly on such data manifolds because of the absence of an analytical model. Therefore, most works only focus on Euclidean space and do not take advantage of this representation. In some cases though, the data naturally and explicitly lie on a manifold, or can be embedded on some known manifolds, allowing leveraging their intrinsic structure. In such cases, it has been shown to be beneficial to exploit such a structure by leveraging the metric of the manifold rather than relying on a Euclidean embedding. To name a few examples, directional or geophysical data — data for which only the direction provides information — naturally lie on the sphere (Mardia et al., 2000) and hence their structure can be exploited by using methods suited to the sphere. Another popular example is given by data that have a known hierarchical structure. Then, such data benefit from being embedded in hyperbolic spaces (Nickel and Kiela, 2017).

Motivated by these examples, many works have proposed new tools to handle data lying on Riemannian manifolds. To cite a few, Fletcher et al. (2004) and Huckemann and Ziezold (2006) developed PCA to perform dimension reduction on manifolds, while Le Brigant and Puechmorel (2019) studied density approximation, Feragen et al. (2015); Jayasumana et al. (2015); Fang et al. (2021) studied kernel methods and Azangulov et al. (2024a,b) developed Gaussian processes on (homogeneous) manifolds. More recently, there has been much interest in developing new neural networks with architectures that take into account the geometry of the ambient manifold (Bronstein et al., 2017), such as Residual Neural Networks (Katsman et al., 2024), discrete Normalizing Flows (Bose et al., 2020; Rezende et al., 2020; Rezende and Racanière, 2021) or Continuous Normalizing Flows (Mathieu and Nickel, 2020; Lou et al., 2020; Rozen et al., 2021; Yataka et al., 2023). In the generative model literature, we can also mention the recent work by Chen and Lipman (2023), which extended the flow matching training of Continuous Normalizing Flows to Riemannian manifolds, as well as the works by Bortoli et al. (2022) and Huang et al. (2022), who performed score based generative modeling, and Thornton et al. (2022), who studied Schrödinger bridges on manifolds.

To compare probability distributions or perform generative modeling tasks, one usually needs suitable discrepancies or distances. In Machine Learning, classical divergences include, for example, the Kullback-Leibler divergence and the Maximum Mean Discrepancy (MMD). While these distances are well defined for distributions lying on Riemannian manifolds, generally by crafting dedicated kernels for the MMD (Feragen et al., 2015), other choices that take more into account the geometry of the underlying space are Optimal Transport based distances, whose most prominent example is the Wasserstein distance.

While the Wasserstein distance is well defined on Riemannian manifolds and has been studied in many works theoretically, see *e.g.* (McCann, 2001; Villani et al., 2009), it suffers from a significant computational burden. In the Euclidean case, various approaches have been proposed to alleviate this computational cost, such as adding an entropic regularization and leveraging the Sinkhorn algorithm (Cuturi, 2013), approximating the Wasserstein distance using minibatchs (Fatras et al., 2020), using low-rank couplings (Scetbon and Cuturi, 2022), or tree metrics (Le et al., 2019). These approximations can be easily extended to Riemannian manifolds by using the right ground costs. For example, Alvarez-Melis et al. (2020) and Hoyos-Idrobo (2020) used the entropic regularized formulation on Hyperbolic spaces. Another popular alternative to the Wasserstein distance is the Sliced-Wasserstein distance (SW). While on Euclidean spaces, the Sliced-Wasserstein distance is a tractable alternative that allows working in large-scale settings, it cannot be directly extended to Riemannian manifolds since it relies on orthogonal projections of the measures onto straight lines. Hence, deriving new SW based distances on manifolds could be of much interest. This question has led to several works in this direction, first on compact manifolds in (Rustamov and Majumdar, 2023) and then on the sphere (Bonet et al., 2023b; Quellmalz et al., 2023). Here, we focus on the particular case of Cartan-Hadamard manifolds, which encompass, in particular, Euclidean spaces, Hyperbolic spaces and Symmetric Positive Definite matrices endowed with appropriate metrics. More precisely, we develop a theoretically grounded way to define Sliced-Wasserstein distances on any Cartan-Hadamard manifold. We discuss their implementation in various specific cases, including Pullback-Euclidean manifolds, Hyperbolic spaces, Symmetric Positive Definite matrices, and product manifolds. Furthermore, we propose applications to different machine learning tasks, such as document classification

and data set comparison, and we discuss the minimization of these discrepancies using the framework of Wasserstein gradient flows.

Contributions. In this article, we start in Section 2 by providing some background on Optimal Transport and on Riemannian manifolds. Then, in Section 3, we introduce different ways to construct intrinsically Sliced-Wasserstein discrepancies on geodesically complete Riemannian manifolds with non-positive curvature (Cartan-Hadamard manifolds) by either using geodesic projections or horospherical projections. In Section 4, we specify the framework to different Cartan-Hadamard manifolds, including manifolds endowed with a pullback Euclidean metric, Hyperbolic spaces, Symmetric positive Definite matrices with specific metrics and product of Cartan-Hadamard manifolds. Then, in Section 5, we derive some theoretical properties common to any sliced discrepancy on these Riemannian manifolds, as well as properties specific to the pullback Euclidean case. We also propose in Section 6 applications of the Sliced-Wasserstein distance on the Euclidean space endowed with the Mahalanobis distance on a document classification task, and of the Sliced-Wasserstein distance on product manifolds for comparing data sets represented on the product space of the samples and of the labels. Finally, we propose in Section 7 non-parametric schemes to minimize these different distances using Wasserstein gradient flows, and hence allowing to derive new sampling algorithms on manifolds.[1]

2. Background

In this section, we first introduce background on Optimal Transport through the Wasserstein distance and the Sliced-Wasserstein distance on Euclidean spaces. Then, we introduce general Riemannian manifolds. For more details about Optimal Transport, we refer to (Villani et al., 2009; Santambrogio, 2015; Peyré et al., 2019). And for more details about Riemannian manifolds, we refer to (Gallot et al., 1990; Lee, 2006, 2012).

2.1 Optimal Transport on Euclidean Spaces

Wasserstein Distance. Optimal transport provides a principled way to compare probability distributions through the Wasserstein distance. Let $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}),\ \int \|x\|_2^p \, \mathrm{d}\mu(x) < \infty\}$ two probability distributions with p finite moments. Then, the Wasserstein distance is defined as

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x - y\|_2^p \, \mathrm{d}\gamma(x, y),$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d),\ \pi^1_\# \gamma = \mu,\ \pi^2_\# \gamma = \nu\}$ denotes the set of couplings between μ and ν, $\pi^1(x, y) = x$, $\pi^2(x, y) = y$ and $\#$ is the push-forward operator, defined as $T_\# \mu(A) = \mu(T^{-1}(A))$ for any Borel set $A \subset \mathbb{R}^d$ and map $T : \mathbb{R}^d \to \mathbb{R}^d$.

For discrete probability distributions with n samples, *e.g.*, for $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$ with $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}^d$, computing W_p^p requires solving a linear program, which has a $O(n^3 \log n)$ worst-case complexity (Pele and Werman, 2009). Thus, it becomes intractable in large scale settings.

1. Code available at `https://github.com/clbonet/Sliced-Wasserstein_Distances_and_Flows_on_Cartan-Hadamard_Manifolds`

For unknown probability distributions μ and ν, from which we have access to samples $x_1, \dots, x_n \sim \mu$ and $y_1, \dots, y_n \sim \nu$, a common practice to estimate $W_p^p(\mu, \nu)$ is to compute the plug-in estimator

$$\widehat{W}_p^p(\mu, \nu) = W_p^p \left(\frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \frac{1}{n} \sum_{i=1}^n \delta_{y_i} \right).$$

However, the approximation error, known as the sample complexity, is quantified in $O(n^{-\frac{1}{d}})$ (Boissard and Le Gouic, 2014). Thus, estimating the Wasserstein distance becomes less accurate in higher dimensions with the same sample size or computationally expensive if larger samples are used to maintain accuracy.

To alleviate the computational burden and the curse of dimensionality, different variants were proposed. We focus in this work on the Sliced-Wasserstein distance.

Sliced-Wasserstein Distance. For $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, it is well known that the Wasserstein distance can be computed in closed-form (Peyré et al., 2019, Remark 2.30). More precisely, let $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, then

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

where F_μ^{-1} and F_ν^{-1} denote the quantile functions of μ and ν. For discrete distributions with n samples, quantiles can be computed in $O(n \log n)$ since they only require sorting the samples. Thus, for $x_1 < \cdots < x_n$, $y_1 < \cdots < y_n$, $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$,

$$W_p^p(\mu_n, \nu_n) = \frac{1}{n} \sum_{i=1}^n |x_i - y_i|^p.$$

Motivated by this closed-form, Rabin et al. (2012) introduced the Sliced-Wasserstein distance, which is defined by first projecting linearly the distributions on every possible direction, and then by taking the average of the one dimensional Wasserstein distances on each line. More precisely, for a direction $\theta \in S^{d-1}$, the coordinate of the orthogonal projection of $x \in \mathbb{R}^d$ on the line $\mathrm{span}(\theta)$ is defined by $P^\theta(x) = \langle x, \theta \rangle$. Then, by denoting by λ the uniform measure on the sphere $S^{d-1} = \{\theta \in \mathbb{R}^d, \ \|\theta\|_2 = 1\}$, the p-Sliced-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

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

The projection process is illustrated in Figure 1.

Since the outer integral is intractable, a common practice to estimate this integral is to rely on a Monte-Carlo approximation by first sampling L directions $\theta_1, \dots, \theta_L$ and then taking the average of the L Wasserstein distances:

$$\widehat{\mathrm{SW}}_p^p(\mu, \nu) = \frac{1}{L} \sum_{\ell=1}^L W_p^p(P^{\theta_\ell}_\# \mu, P^{\theta_\ell}_\# \nu).$$

Thus, approximating SW requires to compute L Wasserstein distances, and Ln projections, resulting in a computational complexity of $O\big(Ln(\log n + d)\big)$. Note that other integral

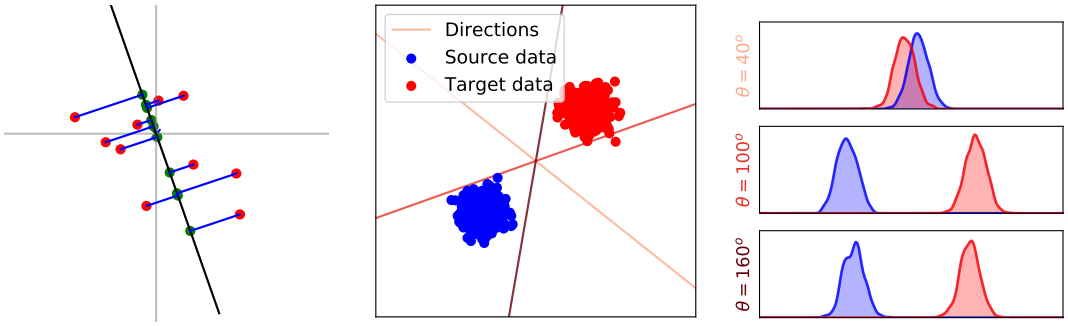


Figure 1: (**Left**) Orthogonal projection of points on a line passing through the origin 0. (**Middle** and **Right**) Illustration of the projection of 2d distributions on 3 different lines.

approximations have been recently proposed. For example, Nguyen et al. (2024) proposed to use quasi Monte-Carlo samples, and Leluc et al. (2023, 2024); Nguyen and Ho (2024a) used control variates to reduce the variance of the approximation.

We are now interested in transposing this method to Riemannian manifolds, for which we give a short introduction in the following section.

2.2 Riemannian Manifolds

Definition. A Riemannian manifold $(\mathcal{M}, g)$ of dimension d is a space that behaves locally as a linear space diffeomorphic to $\mathbb{R}^d$, called a tangent space. To any $x \in \mathcal{M}$, one can associate a tangent space $T_x\mathcal{M}$ endowed with an inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ that varies smoothly with x. This inner product is defined by the metric g_x associated to the Riemannian manifold as $g_x(u, v) = \langle u, v \rangle_x$ for any $x \in \mathcal{M}$, $u, v \in T_x\mathcal{M}$. We denote $G(x)$ the matrix representation of g_x defined such that

$$\forall u, v \in T_x\mathcal{M}, \ \langle u, v \rangle_x = g_x(u, v) = u^T G(x) v.$$

In some spaces, different metrics can give very different geometries. We call tangent bundle the disjoint union of all tangent spaces $T\mathcal{M} = \{(x, v), \ x \in \mathcal{M} \text{ and } v \in T_x\mathcal{M}\}$, and we call a vector field a map $V : \mathcal{M} \to T\mathcal{M}$ such that $V(x) \in T_x\mathcal{M}$ for all $x \in \mathcal{M}$.

Geodesics. A generalization of straight lines in Euclidean spaces to Riemannian manifolds is given by geodesics, which are smooth curves connecting two points $x, y \in \mathcal{M}$ with minimal length, *i.e.*, continuous mappings $\gamma : [0, 1] \to \mathcal{M}$ such that $\gamma(0) = x$, $\gamma(1) = y$, and which minimize the length $\mathcal{L}$ defined as

$$\mathcal{L}(\gamma) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)} \, \mathrm{d}t,$$

where $\|\gamma'(t)\|_{\gamma(t)} = \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}}$. In this work, we will focus on geodesically complete Riemannian manifolds, in which case there is always a geodesic between two points $x, y \in$

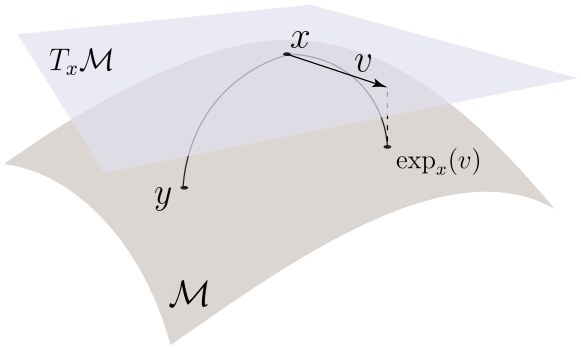


Figure 2: Illustration of geodesics, of the tangent space and the exponential map on a Riemannian manifold.

$\mathcal{M}$. Furthermore, in this specific case, all geodesics are actually geodesic lines, *i.e.*, they can be extended to $\mathbb{R}$. Let $x, y \in \mathcal{M}$, $\gamma : [0,1] \to \mathcal{M}$ a geodesic between x and y such that $\gamma(0) = x$ and $\gamma(1) = y$, then the value of the length defines actually a distance $(x, y) \mapsto d(x, y)$ between x and y, which we call the geodesic distance:

$$d(x,y) = \inf_{\gamma,\ \gamma(0)=x,\ \gamma(1)=y} \mathcal{L}(\gamma).$$

Exponential Map. Let $x \in \mathcal{M}$, then for any $v \in T_x\mathcal{M}$, there exists a unique geodesic $\gamma_{(x,v)}$ starting at x with velocity v, *i.e.*, such that $\gamma_{(x,v)}(0) = x$ and $\gamma'_{(x,v)}(0) = v$ (Sommer et al., 2020). We can now define the exponential map as $\exp : T\mathcal{M} \to \mathcal{M}$ which for any $x \in \mathcal{M}$, maps tangent vectors $v \in T_x\mathcal{M}$ back to the manifold at the point reached by the geodesic $\gamma_{(x,v)}$ at time $t = 1$:

$$\forall (x,v) \in T\mathcal{M},\ \exp_x(v) = \gamma_{(x,v)}(1).$$

On geodesically complete manifolds, the exponential map is defined on the entire tangent space, but is not necessarily a bijection. If it is the case, we note $\log_x$ the inverse of $\exp_x$, which allows mapping elements from the manifold to the tangent space. We illustrate these different notions on Figure 2.

Let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable map. We now define its Riemannian gradient, which is notably very important in order to generalize first-order optimization algorithms to Riemannian manifolds (Bonnabel, 2013; Boumal, 2023).

Definition 1 (Gradient) *We define the Riemannian gradient of f as the unique vector field* $\operatorname{grad}_{\mathcal{M}} f : \mathcal{M} \to T\mathcal{M}$ *satisfying*

$$\forall (x,v) \in T\mathcal{M},\ \frac{\mathrm{d}}{\mathrm{d}t} f\big(\exp_x(tv)\big)\Big|_{t=0} = \langle v, \operatorname{grad}_{\mathcal{M}} f(x)\rangle_x.$$

Sectional Curvature. A notion that allows studying the geometry as well as the topology of a given Riemannian manifold is the sectional curvature. Consider $x \in \mathcal{M}$ and two linearly independent vectors $u, v \in T_x\mathcal{M}$. The sectional curvature $\kappa_x(u, v)$ is defined

geometrically as the Gaussian curvature of the plane $E = \text{span}(u, v)$ (Zhang et al., 2016), *i.e.,*

$$\kappa_x(u, v) = \frac{\langle R(u, v)u, v\rangle_x}{\langle u, u\rangle_x\langle v, v\rangle_x - \langle u, v\rangle_x^2},$$

where R denotes the Riemannian curvature tensor. We refer to (Lee, 2006) for more details. The behavior of geodesics changes given the curvature of the manifold. For instance, they usually diverge on manifolds of negative sectional curvature and converge on manifolds of positive sectional curvature (Hu et al., 2023). Important examples of Riemannian manifolds include Euclidean spaces, which have constant null curvature, the sphere, which has positive constant curvature and Hyperbolic spaces, which have negative constant curvature (*i.e.*, have the same value at any point $x \in \mathcal{M}$ and for any 2-planes E) with their standard metrics. Another example is the torus endowed with the ambient metric which has some points of positive curvature, some points of negative curvature and some points of null curvature (de Ocáriz Borde et al., 2023b). In this paper, we will mostly focus on Cartan-Hadamard manifolds which are complete connected Riemannian manifolds of non-positive sectional curvature.

2.3 Probability Distributions on Riemannian Manifolds

Probability Distributions. Let $(\mathcal{M}, g)$ be a Riemannian manifold. For $x \in \mathcal{M}$, $G(x)$ induces an infinitesimal change of volume on the tangent space $T_x\mathcal{M}$, resulting in a measure on the manifold,

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

Here, we denote by $\text{d}x$ the Lebesgue measure. We refer to (Pennec, 2006) for more details on distributions on manifolds.

Particularly interesting examples of probability distributions are wrapped distributions (Chevallier and Guigui, 2020; Chevallier et al., 2022; Galaz-Garcia et al., 2022), which are defined as the push-forward of a distribution $\mu \in \mathcal{P}(T_x\mathcal{M})$ onto $\mathcal{P}(\mathcal{M})$ using, *e.g.*, the exponential map when it is invertible over the whole tangent space. Since it provides a very convenient way to sample on manifolds, this has received much attention notably on hyperbolic spaces with the wrapped normal distribution (Nagano et al., 2019; Cho et al., 2022b), for which the distribution in the tangent space is a Gaussian, and for which all transformations are differentiable, and can be used *e.g.* for variational autoencoders since they are amenable to the reparametrization trick.

Optimal Transport. Optimal Transport is also well defined on Riemannian manifolds using appropriate ground costs into the Kantorovich problem. Using the geodesic distance at the power $p \geq 1$, we recover the p-Wasserstein distance (McCann, 2001; Villani et al., 2009)

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

where $\mu, \nu \in \mathcal{P}_p(\mathcal{M}) = \{\mu \in \mathcal{P}(\mathcal{M}),\ \int_{\mathcal{M}} d(x, o)^p\ \text{d}\mu(x) < \infty\}$, with $o \in \mathcal{M}$ some origin which can be arbitrarily chosen (because of the triangular inequality).

3. Riemannian Sliced-Wasserstein

In this section, we introduce natural generalizations of the Sliced-Wasserstein distance for probability distributions supported on Riemannian manifolds, leveraging tools that are intrinsically defined on these spaces. To achieve this, we begin by examining the Euclidean space from a Riemannian manifold perspective. Doing so, we naturally extend the Sliced-Wasserstein distance to Riemannian manifolds of non-positive curvature. The proofs of this section are postponed to Appendix B.

3.1 Euclidean Sliced-Wasserstein as a Riemannian Sliced-Wasserstein Distance

It is well known that the Euclidean space can be viewed as a Riemannian manifold of null constant curvature (Lee, 2006). From this perspective, we can translate the elements used to build the Sliced-Wasserstein distance as Riemannian elements, and identify how to generalize it to more general Riemannian manifolds.

First, let us recall that the p-Sliced-Wasserstein distance for $p \geq 1$ between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as

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

where $P^\theta(x) = \langle x, \theta \rangle$ and λ is the uniform distribution S^{d-1}. Geometrically, it amounts to projecting the distributions on every possible line passing through the origin 0. Hence, we see that we first need to generalize lines passing through the origin. Next, we need to find suitable projections onto these subsets. Finally, we need to ensure that we can still compute the Wasserstein distance efficiently between distributions supported on these subsets to maintain a computational advantage over solving the linear program.

Lines. From a Riemannian manifold point of view, straight lines can be seen as geodesics, which are, as we saw in Section 2.2, curves minimizing the distance between any two points on it. For any direction $\theta \in S^{d-1}$, the geodesic passing through 0 in direction θ is described by the curve $\gamma_\theta : \mathbb{R} \to \mathbb{R}^d$ defined as $\gamma_\theta(t) = t\theta = \exp_0(t\theta)$ for any $t \in \mathbb{R}$, and the corresponding geodesic is $\mathcal{G}^\theta = \mathrm{span}(\theta)$. Hence, when it makes sense, a natural generalization of projections onto straight lines would be projections on geodesics passing through an origin.

Projections. The projection $P^\theta(x)$ of $x \in \mathbb{R}^d$ can be seen as the coordinate of the orthogonal projection on the geodesic $\mathcal{G}^\theta$. Indeed, the orthogonal projection $\tilde{P}$ is formally defined as

$$\tilde{P}^\theta(x) = \operatorname*{argmin}_{y \in \mathcal{G}^\theta} \|x - y\|_2 = \langle x, \theta \rangle \theta.$$

From this formulation, we see that $\tilde{P}^\theta$ is a metric projection, which can also be called a geodesic projection on Riemannian manifolds as the metric is a geodesic distance. Then, we see that its coordinate on $\mathcal{G}^\theta$ is $t = \langle x, \theta \rangle = P^\theta(x)$, which can be also obtained by first giving a direction to the geodesic, and then computing the distance between $\tilde{P}^\theta(x)$ and the origin 0, as

$$P^\theta(x) = \mathrm{sign}(\langle x, \theta \rangle) \|\langle x, \theta \rangle \theta - 0\|_2 = \langle x, \theta \rangle.$$

Note that this can also be recovered by solving

$$P^\theta(x) = \operatorname*{argmin}_{t \in \mathbb{R}} \|\exp_0(t\theta) - x\|_2.$$

This formulation will be useful to generalize it to more general manifolds by replacing the Euclidean distance by the right geodesic distance.

Note also that the geodesic projection can be seen as a projection along hyperplanes, *i.e.*, the level sets of the projection function $g(x, \theta) = \langle x, \theta \rangle$ are (affine) hyperplanes. This observation will come useful in generalizing SW to manifolds of non-positive curvature.

Wasserstein Distance. The Wasserstein distance between measures lying on the real line has a closed-form which can be computed very efficiently (see Section 2.1). On more general Riemannian manifolds, as the geodesics will not necessarily be lines, we will need to check how to compute the Wasserstein distance between the projected measures.

3.2 On Manifolds of Non-Positive Curvature

In this section, we focus on complete connected Riemannian manifolds of non-positive curvature, also known as Hadamard manifolds or Cartan-Hadamard manifolds (Lee, 2006; Robbin and Salamon, 2011; Lang, 2012). These spaces include Euclidean spaces, but also spaces with constant negative curvature such as Hyperbolic spaces, or with variable non-positive curvatures such as the space of Symmetric Positive Definite matrices and product of manifolds with constant negative curvature (Gu et al., 2019, Lemma 1). We refer to (Ballmann et al., 2006) or (Bridson and Haefliger, 2013) for more details. These spaces share many properties with Euclidean spaces (Bertrand and Kloeckner, 2012) which make it possible to extend the Sliced-Wasserstein distance on them. We will denote $(\mathcal{M}, g)$ a Hadamard manifold in the following. Particular cases, such as Hyperbolic spaces and the space of Symmetric Positive Definite matrices among others, will be further studied in Section 4.

Properties of Hadamard Manifolds. First, since a Hadamard manifold is a complete connected Riemannian manifold, by the Hopf-Rinow theorem (Lee, 2006, Theorem 6.13), it is also geodesically complete. Therefore, any geodesic curve $\gamma : [0, 1] \to \mathcal{M}$ connecting $x \in \mathcal{M}$ to $y \in \mathcal{M}$ can be extended to $\mathbb{R}$ as a geodesic line. Furthermore, by the Cartan-Hadamard theorem (Lee, 2006, Theorem 11.5), Hadamard manifolds are diffeomorphic to the Euclidean space $\mathbb{R}^d$, and the exponential map at any $x \in \mathcal{M}$ is bijective from $T_x\mathcal{M}$ to $\mathcal{M}$ with the logarithm map as its inverse. Moreover, their injectivity radius is infinite, and thus their geodesics are aperiodic and can be mapped to the real line, allowing us to find coordinates on the real line, and hence to compute the Wasserstein distance between the projected measures efficiently. The SW discrepancy on such spaces is therefore analogous to the Euclidean case. Note that Hadamard manifolds belong to the more general class of CAT(0) metric spaces, and hence inherit their properties described in (Bridson and Haefliger, 2013). We now discuss two different possible projections, which both generalize the Euclidean orthogonal projection.

Geodesic Projections. As we saw in Section 3.1, a natural projection onto geodesics is the geodesic projection. Let $\mathcal{G}$ be a geodesic passing through an origin point $o \in \mathcal{M}$. This origin is often naturally chosen on the space and corresponds to the analog of 0 in $\mathbb{R}^d$. Then, the geodesic projection onto $\mathcal{G}$ is obtained naturally as

$$\forall x \in \mathcal{M},\ \tilde{P}^{\mathcal{G}}(x) = \operatorname*{argmin}_{y \in \mathcal{G}}\ d(x, y).$$

From the projection, we can obtain a coordinate on the geodesic by first assigning it a direction and then computing the distance to the origin. By noting $v \in T_o\mathcal{M}$ a vector in

the tangent space at the origin, such that $\mathcal{G} = \mathcal{G}^v = \{\exp_o(tv),\ t \in \mathbb{R}\}$, we can give a direction to the geodesic by computing the sign of the inner product in the tangent space of o between v and the log of $\tilde{P}^{\mathcal{G}}$. Analogously to the Euclidean case, we can restrict v to be of unit norm, *i.e.*, $\|v\|_o = 1$. We now denote the projection and coordinate projection on $\mathcal{G}^v$ as $\tilde{P}^v$ and P^v, respectively. Hence, we obtain the coordinates using

$$P^v(x) = \operatorname{sign}\big(\langle \log_o\big(\tilde{P}^v(x)\big), v\rangle_o\big)\ d\big(\tilde{P}^v(x), o\big).$$

We show in Proposition 2 that the map $t^v : \mathcal{G}^v \to \mathbb{R}$ defined as

$$\forall x \in \mathcal{G}^v,\ t^v(x) = \operatorname{sign}\big(\langle \log_o(x), v\rangle_o\big)\ d(x, o), \tag{2}$$

is an isometry, *i.e.*, it satisfies $|t^v(x) - t^v(y)| = d(x, y)$ for all $x, y \in \mathcal{G}^v$.

Proposition 2 *Let $(\mathcal{M}, g)$ be a Hadamard manifold with origin o. Let $v \in T_o\mathcal{M}$. The map t^v defined in Equation (2) is an isometry from $\mathcal{G}^v = \{\exp_o(tv),\ t \in \mathbb{R}\}$ to $\mathbb{R}$.*

Note that to obtain the coordinate directly from $x \in \mathcal{M}$, we can also solve the following problem:

$$P^v(x) = \operatorname*{argmin}_{t\in\mathbb{R}}\ d\big(\exp_o(tv), x\big). \tag{3}$$

Since Hadamard manifolds belong to the more general class of CAT(0) metric spaces, by (Bridson and Haefliger, 2013, II. Proposition 2.2), the geodesic distance is geodesically convex. Hence, $t \mapsto d\big(\exp_o(tv), x\big)^2$ is a coercive strictly convex function, and thus it admits a unique minimizer. Therefore, Equation (3) is well defined. Moreover, we have the following characterization for the optimum:

Proposition 3 *Let $(\mathcal{M}, g)$ be a Hadamard manifold with origin o. Let $v \in T_o\mathcal{M}$, and note $\gamma(t) = \exp_o(tv)$ for all $t \in \mathbb{R}$. Then, for any $x \in \mathcal{M}$,*

$$P^v(x) = \operatorname*{argmin}_{t\in\mathbb{R}}\ d(\gamma(t), x)^2 \iff \big\langle \gamma'\big(P^v(x)\big), \log_{\gamma(P^v(x))}(x)\big\rangle_{\gamma(P^v(x))} = 0.$$

In the Euclidean case $\mathbb{R}^d$, geodesics are of the form $\gamma(t) = t\theta$ for any $t \in \mathbb{R}$ and for a direction $\theta \in S^{d-1}$. Since $\log_x(y) = y - x$ for $x, y \in \mathbb{R}^d$, we recover the projection formula:

$$\big\langle \gamma'\big(P^\theta(x)\big), \log_{\gamma(P^\theta(x))}(x)\big\rangle_{\gamma(P^\theta(x))} = 0 \iff \langle \theta, x - P^\theta(x)\theta\rangle = 0 \iff P^\theta(x) = \langle \theta, x\rangle.$$

Busemann Projections. The level sets of the geodesic projections are geodesic subspaces. It has been shown that projecting along geodesics is not always the best solution, as it might not preserve distances between the original points (Chami et al., 2021). Indeed, on Euclidean spaces, as mentioned earlier, the projections are done along hyperplanes, which preserve the distances between points belonging to another geodesic with the same direction (see Figure 3). On Hadamard manifolds, hyperplane analogs can be obtained through the level sets of the Busemann function, which we now introduce.

Let $\gamma : \mathbb{R} \to \mathcal{M}$ be a geodesic line, then the Busemann function associated to γ is defined as (Bridson and Haefliger, 2013, II. Definition 8.17)

$$\forall x \in \mathcal{M},\ B^\gamma(x) = \lim_{t\to\infty}\ \big(d\big(x, \gamma(t)\big) - t\big).$$

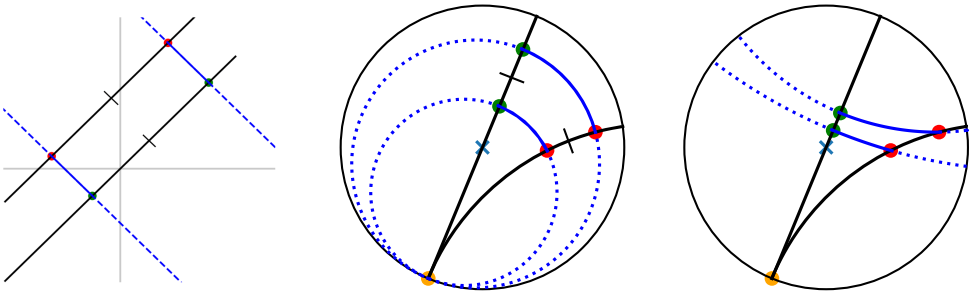


Figure 3: (**Left**) On Euclidean spaces, the distance between the projections of two points belonging to a geodesic with the same direction is conserved. (**Middle**) On Hyperbolic spaces, this is also the case using the horospherical projection as demonstrated in (Chami et al., 2021, Proposition 3.4), but not for geodesic projections (**Right**).

On Hadamard manifolds, and more generally on CAT(0) spaces, the limit exists (Bridson and Haefliger, 2013, II. Lemma 8.18). This function returns a coordinate on the geodesic γ, which can be understood as a normalized distance to infinity towards the direction given by γ (Chami et al., 2021). The level sets of this function are called horospheres. On spaces of constant curvature (*i.e.*, Euclidean or Hyperbolic spaces), horospheres are of constant null curvature and hence very similar to hyperplanes. We illustrate horospheres in Hyperbolic spaces in the middle of Figure 3 and in Figure 5.

For example, in the Euclidean case, we can show that the Busemann function associated with $\mathcal{G}^\theta = \text{span}(\theta)$ for $\theta \in S^{d-1}$ is given by

$$\forall x \in \mathbb{R}^d,\ B^\theta(x) = -\langle x, \theta \rangle.$$

It actually coincides, up to a sign, with the inner product, which can be seen as a coordinate on the geodesic $\mathcal{G}^\theta$. Moreover, its level sets in this case are (affine) hyperplanes orthogonal to θ.

Hence, the Busemann function offers a principled way to project elements $x \in \mathcal{M}$ from a Hadamard manifold onto $\mathbb{R}$, provided its closed-form can be computed. To find the projection onto the geodesic γ, we can solve the equation in $s \in \mathbb{R}$, $B^\gamma(x) = B^\gamma\big(\gamma(s)\big) = -s$, and we find that the projection onto the geodesic γ characterized by $v \in T_o\mathcal{M}$ such that $\|v\|_o = 1$ and $\gamma(t) = \exp_o(tv)$ for all $t \in \mathbb{R}$ is

$$\tilde{B}^v(x) = \exp_o\big(-B^v(x)v\big).$$

Wasserstein Distance on Geodesics. In Proposition 4, we verify that the Wasserstein distance between the coordinates (on $\mathcal{P}_p(\mathbb{R})$) is equal to the Wasserstein distance between the measures projected onto geodesics (on $\mathcal{P}_p(\mathcal{M})$). This relies on the isometry property of t^v derived in Proposition 2.

Proposition 4 *Let $(\mathcal{M}, g)$ a Hadamard manifold, $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$. Let $v \in T_o\mathcal{M}$ such that $\|v\|_o = 1$ and $\mathcal{G}^v = \{\exp_o(tv),\ t \in \mathbb{R}\}$ the geodesic on which the measures are projected. Then,*

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

where the Wasserstein distances are defined with the corresponding geodesic distance given the space, i.e.*, with $d(x, y)$ the geodesic distance on $\mathcal{M}$ for the W_p on the left, and $|t - s|$ for W_p on the right.*

From these properties, we can work equivalently in $\mathbb{R}$ and on the geodesics when using the Busemann projection (also called horospherical projection) or the geodesic projection of measures. In practice, analogously to the Euclidean case, we use the projections on $\mathbb{R}$ and the closed-form of the Wasserstein distance in $\mathcal{P}_p(\mathbb{R})$ given by Equation (1).

Sliced-Wasserstein on Hadamard Manifolds. We are now ready to define the Sliced-Wasserstein distance on Hadamard manifolds. For directions, we will sample from the uniform measure on $S_o = \{v \in T_o\mathcal{M},\ \|v\|_o = 1\}$. Note that other distributions could be used, such as a Dirac in the maximum direction, similarly to max-SW (Deshpande et al., 2019), or any variant using different slicing distributions, as in (Nguyen et al., 2021a,b; Ohana et al., 2023; Nguyen and Ho, 2024b). However, to define a strict generalization of SW, we choose to focus on the uniform one in this work.

Definition 5 (Cartan-Hadamard Sliced-Wasserstein) *Let $(\mathcal{M}, g)$ a Hadamard manifold with o as its origin. Denote λ_o as the uniform distribution on $S_o = \{v \in T_o\mathcal{M},\ \|v\|_o = 1\}$. Let $p \geq 1$, then we define the p-Geodesic Cartan-Hadamard Sliced-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ as*

$$\mathrm{GCHSW}_p^p(\mu, \nu) = \int_{S_o} W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)\ \mathrm{d}\lambda_o(v).$$

Likewise, we define the p-Horospherical Cartan-Hadamard Sliced-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ as

$$\mathrm{HCHSW}_p^p(\mu, \nu) = \int_{S_o} W_p^p(B^v_{\#}\mu, B^v_{\#}\nu)\ \mathrm{d}\lambda_o(v).$$

In the following, when we want to mention both GCHSW and HCHSW, for example for properties satisfied by both, we will use the term Cartan-Hadamard Sliced-Wasserstein, abbreviated as CHSW. Then, without loss of generality, we will write

$$\mathrm{CHSW}_p^p(\mu, \nu) = \int_{S_o} W_p^p(P^v_{\#}\mu, P^v_{\#}\nu)\ \mathrm{d}\lambda_o(v),$$

with P^v denoting either the geodesic or the horospherical projection. We illustrate the projection process in Figure 4.

Guidelines between Geodesic and Horospherical CHSW. A natural question to ask is which projection we should choose. As we will see in Section 4, both projections

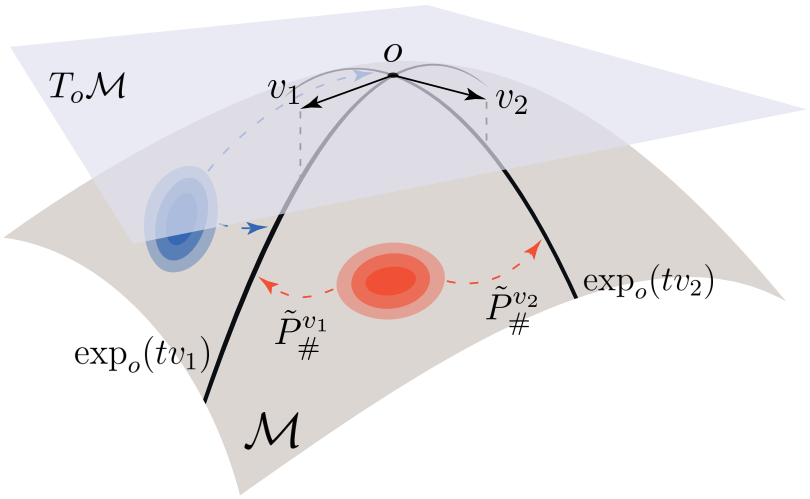


Figure 4: Illustration of the projection process of measures on geodesics $t \mapsto \exp_o(tv_1)$ and $t \mapsto \exp_o(tv_2)$.

coincide for any pullback Euclidean metric, which includes many manifolds of interest. For negatively curved spaces, they do not coincide. Nonetheless, there are cases where we can compute only the horospherical projection in closed-form. For the few cases where we can compute both (*e.g.*, the hyperbolic case), we refer to (Bonet et al., 2023a, Figure 3) for the behavior of the two distances between distributions. We observe that the horospherical HSW more closely aligns with the behavior of the Wasserstein distance. However, in applications, both variants perform well.

3.3 Related Works

Intrinsic Sliced-Wasserstein. To the best of our knowledge, the first attempt to define a generalization of the Sliced-Wasserstein distance on Riemannian manifolds was made by Rustamov and Majumdar (2023). In this work, they restricted their analysis to compact spaces and proposed to use the eigendecomposition of the Laplace-Beltrami operator (see (Gallot et al., 1990, Definition 4.7)). Let $(\mathcal{M}, g)$ be a compact Riemannian manifold. For $\ell \in \mathbb{N}$, denote λ_ℓ the eigenvalues and ϕ_ℓ the eigenfunctions of the Laplace-Beltrami operator sorted by increasing eigenvalues. Then, we can define spectral distances as

$$\forall x, y \in \mathcal{M},\ d_\alpha(x, y) = \sum_{\ell \geq 0} \alpha(\lambda_\ell)\big(\phi_\ell(x) - \phi_\ell(y)\big)^2,$$

where $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is a monotonically decreasing function. Then, they define the Intrinsic Sliced-Wasserstein (ISW) distance between $\mu, \nu \in \mathcal{P}_2(\mathcal{M})$ as

$$\mathrm{ISW}_2^2(\mu, \nu) = \sum_{\ell \geq 0} \alpha(\lambda_\ell) W_2^2\big((\phi_\ell)_\# \mu, (\phi_\ell)_\# \nu\big).$$

The eigenfunctions are used to map the measures to the real line, which makes it very efficient to compute in practice. The eigenvalues are sorted in increasing order, and the

series is often truncated by keeping only the L smallest eigenvalues. This distance cannot be applied to Hadamard manifolds as these spaces are not compact.

Sliced-Wasserstein on the Sphere. Bonet et al. (2023b) then proposed a Spherical Sliced-Wasserstein distance by integrating and projecting over all geodesics using the geodesic projection in an attempt to generalize the Sliced-Wasserstein distance intrinsically to the sphere S^{d-1}. We note that ISW is more in the spirit of a max-K Sliced-Wasserstein distance (Dai and Seljak, 2021), which projects over the K maximal directions, rather than the Sliced-Wasserstein distance. More recently, Quellmalz et al. (2023, 2024) studied different Sliced-Wasserstein distances on S^2 by using spherical Radon transforms, while Tran et al. (2024) proposed to use the stereographic projection along the Generalized Sliced-Wasserstein distance (Kolouri et al., 2019), and Garrett et al. (2024) proposed Sliced-Wasserstein distances over the space of functions on the sphere using a convolution slicer *w.r.t* a kernel for the projection. Moreover, Genest et al. (2024) leveraged the Sliced-Wasserstein distance on manifolds to sample noise on non-Euclidean spaces such as meshes.

Generalized Sliced-Wasserstein. A somewhat related distance is the Generalized Sliced-Wasserstein distance (GSW) introduced by Kolouri et al. (2019), and which uses nonlinear projections onto the real lines. The main difference is that GSW focuses on probability distributions lying in Euclidean space by projecting the measures along nonlinear hypersurfaces. That said, adapting the definition of GSW to handle probability measures on Riemannian manifolds, and the properties that need to be satisfied by the defining function g such as the homogeneity, then we can write the CHSW in the framework of GSW using $g : (x, v) \mapsto P^v(x)$.

4. Examples of Cartan-Hadamard Sliced-Wasserstein

In this section, we specify the framework derived in full generality in Section 3 for particular Hadamard manifolds. More precisely, we first focus on manifolds endowed with a Pullback Euclidean metric, which are Hadamard manifolds with null curvature. Then, we look at Hyperbolic spaces which are manifolds of constant negative curvature. We also study the space of Symmetric Positive Definite matrices (SPD) endowed with metrics for which it is a Hadamard manifold. Finally, we discuss the case of the product manifold of Hadamard manifolds, which is itself a Hadamard manifold, as products of manifolds of non-positive curvature are still of non-positive curvature (Gu et al., 2019, Lemma 1). We defer the proofs of this section to Appendix C.

4.1 Pullback Euclidean Manifold

Cartan-Hadamard manifolds include, among others, spaces of null curvature. As the curvature is preserved by the pullback operator, pullback Euclidean metrics are such spaces. We formally recall the definition of a pullback Euclidean metric along with its geodesic distance and exponential map, following (Chen et al., 2024b, Theorem 3.3).

Theorem 6 (Pullback Euclidean Metric) *Let $\mathcal{N}$ be a Euclidean space and denote $\langle \cdot, \cdot \rangle$ its inner product and $\| \cdot \|$ the associated norm. Let $\mathcal{M}$ be some space and $\phi : \mathcal{M} \to \mathcal{N}$ be a diffeomorphism. Then, defining for any $x \in \mathcal{M}$ and $u, v \in T_x\mathcal{M}$ the metric $g^{\phi}_x(u, v) = \langle \phi_{*,x}(u), \phi_{*,x}(v) \rangle$ where $\phi_{*,x} : T_x\mathcal{M} \to T_{\phi(x)}\mathcal{N}$ is the differential of ϕ at x, $(\mathcal{M}, g^{\phi})$ is a*

Riemannian manifold with geodesic distance

$$d_{\mathcal{M}}(x, y) = \|\phi(x) - \phi(y)\|.$$

Moreover, the exponential map is

$$\forall x \in \mathcal{M}, v \in T_x\mathcal{M},\ \exp_x(v) = \phi^{-1}\big(\phi(x) + \phi_{*,x}(v)\big).$$

Let $(\mathcal{M}, g^{\phi})$ be such a space. Denote o the origin of $\mathcal{M}$. Geodesics passing through o in direction $v \in T_o\mathcal{M}$ have the form

$$\forall t \in \mathbb{R},\ \gamma_v(t) = \phi^{-1}\big(\phi(o) + t\phi_{*,o}(v)\big).$$

Moreover, tangent vectors $v \in T_o\mathcal{M}$ belong to the sphere S_o if and only if $\|v\|_o^2 = \|\phi_{*,o}(v)\|^2 = 1$. Thus, using this formula, we can obtain both the geodesic and horospherical coordinates, which actually coincide (up to a sign), as in the Euclidean case.

Proposition 7 *Let $v \in S_o$, then the projection coordinate on $\mathcal{G}^v = \{\gamma_v(t),\ t \in \mathbb{R}\}$ is*

$$\forall x \in \mathcal{M},\ P^v(x) = -B^v(x) = \langle \phi(x) - \phi(o), \phi_{*,o}(v) \rangle.$$

For instance, the Euclidean space endowed with the Mahalanobis distance fits this framework for $\phi(x) = A^{\frac{1}{2}}x$ with $A \in S_d^{++}(\mathbb{R})$ a positive definite matrix, since in this case, for any $x, y \in \mathbb{R}^d$,

$$d(x, y)^2 = (x - y)^T A(x - y) = \|A^{\frac{1}{2}}x - A^{\frac{1}{2}}y\|_2^2.$$

In this case, we have $\phi(0) = 0$ and $\phi_{*,0}(v) = A^{\frac{1}{2}}v$. Thus, the projection is obtained by $P^v(x) = \langle A^{\frac{1}{2}}x, A^{\frac{1}{2}}v \rangle = x^T Av$ for $v \in S_0$, *i.e.*, which satisfies $\|v\|_0^2 = \|A^{\frac{1}{2}}v\|_2^2 = 1$. In this situation, as expected, the directions and the data points are first mapped by the linear projection $x \mapsto A^{\frac{1}{2}}x$, and then the usual orthogonal projections are performed as for the Euclidean Sliced-Wasserstein distance.

Definition 8 (Mahalanobis Sliced-Wasserstein) *Let $p \geq 1$ and $A \in S_d^{++}(\mathbb{R})$. The p-Mahalanobis Sliced-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined as*

$$\mathrm{SW}_{p,A}^p(\mu, \nu) = \int_{S_0} W_p^p(P_{\#}^v\mu, P_{\#}^v\nu)\ \mathrm{d}\lambda_0(v),$$

with $P^v(x) = x^T Av$ for $v \in S_0 = \{v \in \mathbb{R}^d,\ v^T Av = 1\}$, $x \in \mathbb{R}^d$ and λ_0 the uniform distribution on S_0.

The Mahalanobis distance is often learned in metric learning, which has been used for different applications in, *e.g.*, computer vision, information retrieval or bioinformatic (Bellet et al., 2013). In Section 6.1, we use the Mahalanobis Sliced-Wasserstein distance for a document classification task (Kusner et al., 2015), where the underlying metric A is previously learned using metric learning methods (Huang et al., 2016).

More generally, this Pullback Euclidean framework includes any squared geodesic distance for which the metric is of the form $\langle u, v \rangle_x = u^T A(x)v$ with $A(x) \in S_d^{++}(\mathbb{R})$ for any $x \in \mathbb{R}^d$ (Scarvelis and Solomon, 2023; Pooladian et al., 2023). For such a metric, we have $\phi_{*,x}(v) = A(x)^{\frac{1}{2}}v$, and computing $\phi(x)$ in closed-form may not be straightforward. It also includes many useful metrics used on the space of SPD matrices, which we describe more thoroughly in Section 4.3.2.

4.2 Hyperbolic Spaces

Hyperbolic spaces are Riemannian manifolds of negative constant curvature $K < 0$ (Lee, 2006) and are thus particular cases of Hadamard manifolds. They have recently received a surge of interest in machine learning as they allow embedding data with a hierarchical structure efficiently (Nickel and Kiela, 2017, 2018). A thorough review of the recent use of hyperbolic spaces in machine learning can be found in (Peng et al., 2021; Mettes et al., 2024).

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 that is the most convenient, either given the formulae to derive or the numerical properties. In machine learning, the two most commonly 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 that behaves better *w.r.t.* numerical issues compared to the distance of the Poincaré ball. However, the Lorentz model is unbounded, unlike the Poincaré ball. We introduce these two models in the following.

Lorentz Model. The Lorentz model of curvature $K < 0$ is defined as

$$\mathbb{L}_K^d = \left\{ (x_0, \dots, x_d) \in \mathbb{R}^{d+1},\ \langle x, x \rangle_{\mathbb{L}} = \frac{1}{K},\ x_0 > 0 \right\},$$

where for $x, y \in \mathbb{R}^{d+1}$, $\langle x, y \rangle_{\mathbb{L}} = -x_0 y_0 + \sum_{i=1}^d x_i y_i$ is the Minkowski pseudo inner-product. The Lorentz model can be seen as the upper sheet of a two-sheet hyperboloid. In the following, we will denote $x^0 = (\frac{1}{\sqrt{-K}}, 0, \dots, 0) \in \mathbb{L}_K^d$ the origin of the hyperboloid. The geodesic distance in this manifold is defined as

$$\forall x, y \in \mathbb{L}_K^d,\ d_{\mathbb{L}}(x, y) = \frac{1}{\sqrt{-K}} \operatorname{arccosh}\left(K \langle x, y \rangle_{\mathbb{L}}\right).$$

At any $x \in \mathbb{L}_K^d$, the tangent space is $T_x \mathbb{L}_K^d = \{v \in \mathbb{R}^{d+1},\ \langle x, v \rangle_{\mathbb{L}} = 0\}$. Note that on $T_{x^0} \mathbb{L}_K^d$, the Minkowski inner product equals the usual Euclidean inner product. Moreover, geodesics passing through x in direction $v \in T_x \mathbb{L}_K^d$ are obtained as the intersection between the plane $\operatorname{span}(x, v)$ and the hyperboloid $\mathbb{L}_K^d$, and are of the form

$$\forall t \in \mathbb{R},\ \exp_x(tv) = \cosh(\sqrt{-K} t \|v\|_{\mathbb{L}}) x + \frac{\sinh(\sqrt{-K} t \|v\|_{\mathbb{L}})}{\sqrt{-K}} \frac{v}{\|v\|_{\mathbb{L}}}.$$

In particular, geodesics passing through the origin x^0 in direction $v \in S_{x^0}$ are

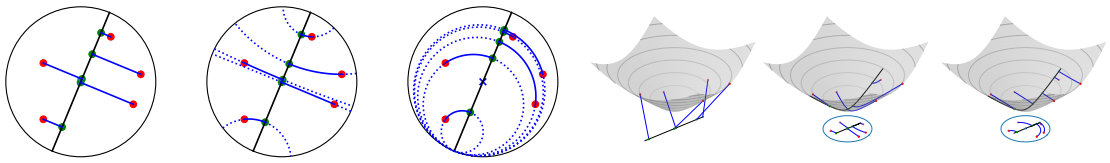
$$\forall t \in \mathbb{R},\ \gamma_v(t) = \exp_{x^0}(tv) = \cosh\left(\sqrt{-K} t\right) x^0 + \frac{\sinh\left(\sqrt{-K} t\right)}{\sqrt{-K}} v.$$

Poincaré Ball. The Poincaré ball of curvature $K < 0$ is defined as

$$\mathbb{B}_K^d = \left\{ x \in \mathbb{R}^d,\ \|x\|_2^2 < -\frac{1}{K} \right\}.$$

It can be seen as the stereographic projection of each point of $\mathbb{L}_K^d$ on the hyperplane $\{x \in \mathbb{R}^{d+1},\ x_0 = 0\}$. The origin of $\mathbb{B}_K^d$ is 0 and the geodesic distance is defined as

$$\forall x, y \in \mathbb{B}_K^d,\ d_{\mathbb{B}}(x, y) = \frac{1}{\sqrt{-K}} \operatorname{arccosh}\left(1 - 2K \frac{\|x - y\|_2^2}{(1 + K\|x\|_2^2)(1 + K\|y\|_2^2)}\right).$$



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

Figure 5: Projection of (red) points on a geodesic (black line) in the Poincaré ball (**Left**) and in the Lorentz model (**Right**) along Euclidean lines, geodesics or horospheres (in blue). Projected points on the geodesic are shown in green.

The tangent space is $\mathbb{R}^d$ and for any $\tilde{v} \in S^{d-1}$, the geodesic passing through the origin is defined as

$$\forall t \in \mathbb{R},\ \gamma_{\tilde{v}}(t) = \exp_0(t\tilde{v}) = \frac{1}{\sqrt{-K}} \tanh\left(\frac{\sqrt{-K}t}{2}\right)\tilde{v}.$$

Hyperbolic Sliced-Wasserstein. To define Hyperbolic Sliced-Wasserstein distances, we first need to sample geodesics, which can be done in both models by simply sampling from a uniform measure on the sphere. Indeed, let $\tilde{v} \in S^{d-1}$, then the direction of the geodesic in $\mathbb{L}_K^d$ is obtained as $v = (0, \tilde{v}) \in T_{x^0}\mathbb{L}_K^d \cap S^d = S_{x^0}$ by concatenating 0 to $\tilde{v}$. On the Poincaré ball, $\tilde{v}$ gives directly the direction to the geodesic, and is called an ideal point.

Thus, we only need to compute the projection coordinates on the geodesics in order to build the corresponding Geodesic and Horospherical Sliced-Wasserstein distances. We provide the closed-form of the geodesic projection and the Busemann function for both models in the following propositions. Additionally, we illustrate the projection process in Figure 5.

Proposition 9 (Coordinate projections on Hyperbolic spaces)

1. *Let $v \in S_{x^0} = T_{x^0}\mathbb{L}_K^d \cap S^d$, the geodesic and horospherical projection coordinates on $\mathcal{G}^v = \operatorname{span}(x^0, v) \cap \mathbb{L}_K^d$ are for all $x \in \mathbb{L}_K^d$,*

$$P^v(x) = \frac{1}{\sqrt{-K}} \operatorname{arctanh}\left(-\frac{1}{\sqrt{-K}}\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right),$$
$$B^v(x) = \frac{1}{\sqrt{-K}} \log\left(-\sqrt{-K}\left\langle x, \sqrt{-K}x^0 + v\right\rangle_{\mathbb{L}}\right).$$

2. *Let $\tilde{v} \in S^{d-1}$ an ideal point. Then the geodesic and horospherical projections coordinates on $\mathcal{G}^{\tilde{v}} = \{\gamma_{\tilde{v}}(t),\ t \in \mathbb{R}\}$ are for all $x \in \mathbb{B}_K^d$,*

$$P^{\tilde{v}}(x) = \frac{2}{\sqrt{-K}} \operatorname{arctanh}\left(\sqrt{-K}s(x)\right),$$
$$B^{\tilde{v}}(x) = \frac{1}{\sqrt{-K}} \log\left(\frac{\|\tilde{v} - \sqrt{-K}x\|_2^2}{1 + K\|x\|_2^2}\right),$$

where s is defined as

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

This proposition allows to define hyperbolic Sliced-Wasserstein distances by specifying CHSW with the right formulas.

Definition 10 (Hyperbolic Sliced-Wasserstein)

1. *Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(\mathbb{L}_K^d)$. Then, the p-Geodesic Hyperbolic Sliced-Wasserstein distance and the p-Horospherical Hyperbolic Sliced-Wasserstein distance on the Lorentz model $\mathbb{L}_K^d$ are defined as*

$$\mathrm{GHSW}_p^p(\mu, \nu) = \int_{T_{x^0}\mathbb{L}_K^d \cap S^d} W_p^p(P_\#^v \mu, P_\#^v \nu)\ \mathrm{d}\lambda(v)$$
$$\mathrm{HHSW}_p^p(\mu, \nu) = \int_{T_{x^0}\mathbb{L}_K^d \cap S^d} W_p^p(B_\#^v \mu, B_\#^v \nu)\ \mathrm{d}\lambda(v).$$

2. *Let $p \geq 1$, $\tilde{\mu}, \tilde{\nu} \in \mathcal{P}_p(\mathbb{B}_K^d)$. Then, the p-Geodesic Hyperbolic Sliced-Wasserstein distance and the p-Horospherical Hyperbolic Sliced-Wasserstein distance on the Poincaré ball $\mathbb{B}_K^d$ are defined as*

$$\mathrm{GHSW}_p^p(\tilde{\mu}, \tilde{\nu}) = \int_{S^{d-1}} W_p^p(P_\#^{\tilde{v}} \tilde{\mu}, P_\#^{\tilde{v}} \tilde{\nu})\ \mathrm{d}\lambda(\tilde{v})$$
$$\mathrm{HHSW}_p^p(\tilde{\mu}, \tilde{\nu}) = \int_{S^{d-1}} W_p^p(B_\#^{\tilde{v}} \tilde{\mu}, B_\#^{\tilde{v}} \tilde{\nu})\ \mathrm{d}\lambda(\tilde{v}).$$

Note that we could also work on other models such as the Klein model, the Poincaré half-plane model or the hemisphere model (see *e.g.* (Cannon et al., 1997; Loustau, 2020)) and derive the corresponding projections in order to define the Hyperbolic Sliced-Wasserstein distances in these models. Note also that these different Sliced-Wasserstein distances are actually equal from one model to the other when using the isometry mappings, which is a particular case of Proposition 11.

Proposition 11 *Let $(\mathcal{M}, g^{\mathcal{M}})$ and $(\mathcal{N}, g^{\mathcal{N}})$ be two isometric Cartan-Hadamard manifolds, $\phi : \mathcal{M} \to \mathcal{N}$ an isometry, and assume that $\lambda_{\phi(o)} = (\phi_{*,o})_\# \lambda_o$.*[2] *Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ and $\tilde{\mu} = \phi_\# \mu$, $\tilde{\nu} = \phi_\# \nu$. Then,*

$$\mathrm{CHSW}_p^p(\mu, \nu; \lambda_o) = \mathrm{CHSW}_p^p(\tilde{\mu}, \tilde{\nu}; \lambda_{\phi(o)}),$$

where we denote $\mathrm{CHSW}_p^p(\mu, \nu; \lambda)$ the Cartan-Hadamard Sliced-Wasserstein distance with slicing distribution λ.

2. We expect it to be true in general as ϕ is an isometry, but we did not find in the literature a formal proof. In practice, this fact was verified for each tested case.

Proposition 11 includes as a particular case the Hyperbolic Sliced-Wasserstein distances (and in particular is more general than (Bonet et al., 2023a, Proposition 3.4)). This demonstrates that the Hyperbolic Sliced-Wasserstein distances are independent from the chosen model. Thus, we can work in the model which is the most convenient for us. Moreover, if we work on a model for which we do not have necessarily a closed-form, we can project the distributions on a model where we have the closed-forms such as the Lorentz model or the Poincaré ball.

Bonet et al. (2023a) compared GHSW and HHSW on different tasks such as gradient flows or as regularizers for deep classification with prototypes. Moreover, they also verified empirically that GHSW and HHSW are independent with respect to the model while comparing evolutions of the distances between Wrapped Normal distributions. In particular, they observed that HHSW had values closer to the Wasserstein distance compared to GHSW.

4.3 Symmetric Positive Definite Matrices

Let $S_d(\mathbb{R})$ be the set of symmetric matrices of $\mathbb{R}^{d\times d}$, and let $S_d^{++}(\mathbb{R})$ be the set of SPD matrices of $\mathbb{R}^{d\times d}$, *i.e.*, matrices $M \in S_d(\mathbb{R})$ satisfying for all $x \in \mathbb{R}^d \setminus \{0\}$, $x^T M x > 0$. $S_d^{++}(\mathbb{R})$ is a Riemannian manifold (Bhatia, 2009) which can be endowed with different metrics. At each $M \in S_d^{++}(\mathbb{R})$, we can associate a tangent space $T_M S_d^{++}(\mathbb{R})$ which can be identified with the space of symmetric matrices $S_d(\mathbb{R})$.

SPD matrices have received a lot of attention in Machine Learning. On one hand, this is the natural space to deal with invertible covariance matrices, which are often used to represent M/EEG data (Blankertz et al., 2007; Sabbagh et al., 2019) or images (Tuzel et al., 2006; Pennec, 2020). Moreover, this space is more expressive than Euclidean spaces, and endowed with specific metrics such as the Affine-Invariant metric, it enjoys a non-constant non-positive curvature. This property was leveraged to embed different type of data (Harandi et al., 2014; Brooks et al., 2019b). This motivated the development of different machine learning algorithms (Chevallier et al., 2017; Yair et al., 2019; Zhuang et al., 2020; Lei et al., 2021; Ju and Guan, 2022) and of neural networks architectures (Huang and Van Gool, 2017; Brooks et al., 2019a).

We now introduce the Sliced-Wasserstein distance on the space of SPD matrices first endowed with the Affine-Invariant metric, and then endowed with different pullback Euclidean metrics.

4.3.1 Symmetric Positive Definite Matrices with Affine-Invariant Metric.

A classical metric widely used with SPDs is the geometric Affine-Invariant metric (Pennec et al., 2006), where the inner product is defined as

$$\forall M \in S_d^{++}(\mathbb{R}),\ A, B \in T_M S_d^{++}(\mathbb{R}),\ \langle A, B\rangle_M = \mathrm{Tr}(M^{-1}AM^{-1}B).$$

Denoting by Tr the Trace operator, the corresponding geodesic distance $d_{AI}(\cdot,\cdot)$ is given by

$$\forall X, Y \in S_d^{++}(\mathbb{R}),\ d_{AI}(X,Y) = \sqrt{\mathrm{Tr}\big(\log(X^{-1}Y)^2\big)}.$$

An interesting property justifying the use of the Affine-Invariant metric is that d_{AI} satisfies the affine-invariant property: for any $g \in GL_d(\mathbb{R})$, where $GL_d(\mathbb{R})$ denotes the set of invertible

matrices in $\mathbb{R}^{d\times d}$,

$$\forall X, Y \in S_d^{++}(\mathbb{R}),\ d_{AI}(g\cdot X, g\cdot Y) = d_{AI}(X,Y),$$

where $g\cdot X = gXg^T$. With this metric, $S_d^{++}(\mathbb{R})$ is of (non-constant) non-positive curvature and hence a Hadamard manifold.

The natural origin is the identity matrix I_d and geodesics passing through I_d, in direction $A \in S_d(\mathbb{R})$ are of the form (Pennec, 2020, Section 3.6.1)

$$\forall t \in \mathbb{R},\ \gamma_A(t) = \exp_{I_d}(tA) = \exp(tA),$$

where exp denotes the matrix exponential.

For the Affine-Invariant case, to the best of our knowledge, there is no closed-form for the geodesic projection on $\mathcal{G}^A$, the difficulty being that the matrices do not necessarily commute. Hence, we will discuss here the horospherical projection which can be obtained with the Busemann function. For $A \in S_d(\mathbb{R})$ such that $\|A\|_F = 1$, denoting $\gamma_A : t \mapsto \exp(tA)$ the geodesic line passing through I_d with direction A, the Busemann function B^A associated to γ_A writes as

$$\forall M \in S_d^{++}(\mathbb{R}),\ B^A(M) = \lim_{t\to\infty} \big(d_{AI}\big(\exp(tA), M\big) - t\big).$$

We cannot directly compute this quantity by expanding the distance since $\exp(-tA)$ and M are not necessarily commuting. The main idea to solve this issue is to first find a group $G \subset GL_d(\mathbb{R})$ that will leave the Busemann function invariant. Then, we can find an element of this group which will project M on the space of matrices commuting with $\exp(A)$. This part of the space is of null curvature, *i.e.*, it is isometric to a Euclidean space. In this case, we can compute the Busemann function since the matrices are commuting. Hence, the Busemann function take the form

$$B^A(M) = -\left\langle A, \log\left(\pi_A(M)\right)\right\rangle_F,$$

where π_A is a projection on the space of commuting matrices which can be obtained in practice through a UDU or LDL decomposition. We detail more precisely in Appendix F how to obtain π^A. For more details about the Busemann function on the Affine-invariant space, we refer to Bridson and Haefliger (2013, Section II.10) and Fletcher et al. (2009, 2011).

We note that computing the Busemann function on this space induces a heavy computational cost. Thus, we advocate for using in practice Sliced-Wasserstein distances obtained using Pullback-Euclidean metrics on SPDs as described in the next section.

4.3.2 Symmetric Positive Definite Matrices with Pullback Euclidean Metrics.

We study here metrics endowing the space of SPD matrices which are pullback Euclidean metrics (Chen et al., 2024a,b), *i.e.*, metrics which are obtained through a diffeomorphism from $S_d^{++}(\mathbb{R})$ to $(S_d(\mathbb{R}), \langle\cdot,\cdot\rangle_F)$. Pullback Euclidean metrics and more generally, pullback metrics allow inheriting properties from the mapped space (Chen et al., 2024b). The pullback Euclidean metrics studied here belong to the framework presented in Section 4.1,

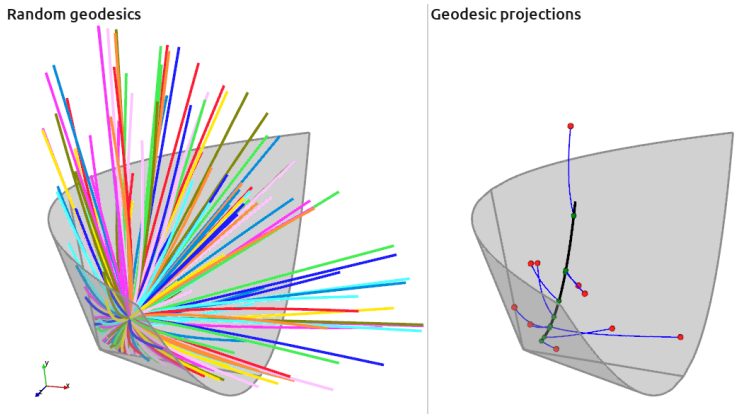


Figure 6: (**Left**) Random geodesics drawn in $S_2^{++}(\mathbb{R})$. (**Right**) Projections (green points) of covariance matrices (depicted as red points) over one geodesic (in black) passing through I_2 along the Log-Euclidean geodesics (blue lines).

with $\mathcal{M} = S_d^{++}(\mathbb{R})$ and $\mathcal{N} = S_d(\mathbb{R})$. This framework includes many interesting metrics, such as the Log-Euclidean metric with $\phi = \log$ (Arsigny et al., 2005, 2006) which is a good first-order approximation of the Affine-Invariant metric (Arsigny et al., 2005; Pennec, 2020), the Log-Cholesky metric (Lin, 2019) or the recently proposed $O(n)$-invariant Log-Euclidean metric (Thanwerdas and Pennec, 2023; Chen et al., 2024a) and Adaptive Riemannian metric (Chen et al., 2024b).

Log-Euclidean Metric. We first focus on the Log-Euclidean metric, for which $\phi = \log$. To apply Proposition 7, we first need to compute its differential in the origin I_d. For completeness, we recall here the differential form of the matrix logarithm derived *e.g.* in (Pennec, 2020).

Lemma 12 (Section 3.2.2 in (Pennec, 2020)) *Let $\phi : X \mapsto \log(X)$ and $X = UDU^T \in S_d^{++}(\mathbb{R})$ where $D = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. The differential operator of ϕ at X is given by*

$$\forall V \in T_X S_d^{++}(\mathbb{R}),\ \phi_{*,X}(V) = U\Sigma(V)U^T,$$

where $\Sigma(V) = U^T V U \odot \Gamma$ and Γ is the Loewner's matrix defined for all $i, j \in \{1, \dots, d\}$ as

$$\Gamma_{ij} = \begin{cases} \frac{\log \lambda_i - \log \lambda_j}{\lambda_i - \lambda_j} & \text{if } i \neq j \\ \frac{1}{\lambda_i} & \text{if } i = j. \end{cases}$$

Proof Apply the Daleckii-Krein formula, see *e.g.*, (Noferini, 2017, Theorem 2.11). ■

We note that for close eigenvalues (Pennec, 2020),

$$\frac{\log \lambda_i - \log \lambda_j}{\lambda_i - \lambda_j} = \frac{1}{\lambda_j}\left(1 - \frac{\lambda_i - \lambda_j}{2\lambda_j} + \frac{(\lambda_i - \lambda_j)^2}{3\lambda_j^2} + O\big((\lambda_i - \lambda_j)^3\big)\right).$$

Furthermore, for $X = D = U = I_d$, since $[U^T V U]_{ij} = V_{ij}$, we find $\phi_{*,I_d}(V) = V$ for any V. Thus, as $\log(I_d) = 0$, we obtain the following projections.

Proposition 13 *Let $\phi = \log$. Then, for any $A \in S_d(\mathbb{R})$ such that $\|A\|_F = 1$, the coordinate projection is*

$$\forall X \in S_d^{++}(\mathbb{R}),\ P^A(X) = -B^A(X) = \langle \log(X), A \rangle_F.$$

Proof Apply Proposition 7 with $\phi(X) = \log(X)$ observing that $\phi(I_d) = 0$ and $\phi_{*,I_d} = \mathrm{Id}$. ■

We illustrate on Figure 6 the projection of matrices $M \in S_2^{++}(\mathbb{R})$ embedded as vectors $(m_{11}, m_{22}, m_{12}) \in \mathbb{R}^3$ on geodesics passing through I_2. This projection was first derived by Bonet et al. (2023c) which introduced the Sliced-Wasserstein distance on the space of SPDs endowed with the Log-Euclidean metric, named SPDSW, and applied it to M/EEG data to perform brain-age prediction and domain adaptation for brain computer interfaces.

O(n)-Invariant Log-Euclidean Metric. The $O(n)$-invariant Log-Euclidean metric was introduced by Thanwerdas and Pennec (2023) and further studied by Chen et al. (2024a). It is a pullback Euclidean metric with, for $X \in S_d^{++}(\mathbb{R})$ and $p, q \geq 0$, $\phi^{p,q}(X) = F^{p,q}\big(\log(X)\big)$ where $F^{p,q}(A) = qA + \frac{p-q}{d}\mathrm{Tr}(A)I_d$ for $A \in S_d(\mathbb{R})$. It can be seen as a generalization of the Log-Euclidean metric since for $p = q = 1$, $F^{1,1}(A) = A$. Since $F^{p,q}$ is a linear function, the differential of $\phi^{p,q}$ at $X \in S_d^{++}(\mathbb{R})$ reads $\phi^{p,q}_{*,X}(V) = F^{p,q}\big(\log_{*,X}(V)\big)$ for any $V \in S_d(\mathbb{R})$. Thus, we have $\phi^{p,q}(I_d) = 0$, $\phi^{p,q}_{*,I_d} = F^{p,q}$, and we can apply Proposition 7.

Proposition 14 *Let $p, q \geq 0$, $\phi^{p,q} = F^{p,q} \circ \log$ with $F^{p,q}(A) = qA + \frac{p-q}{d}\mathrm{Tr}(A)I_d$ for $A \in S_d(\mathbb{R})$. Then, for any $A \in S_d(\mathbb{R})$ such that $\|A\|^2_{I_d} = \langle F^{p,q}(A), F^{p,q}(A)\rangle_F = 1$, the coordinate projection is*

$$\forall X \in S_d^{++}(\mathbb{R}),\ P^A(X) = \big\langle F^{p,q}\big(\log(X)\big), F^{p,q}(A)\big\rangle_F.$$

Proof Apply Proposition 7 with $\phi(X) = F^{p,q}\big(\log(X)\big)$ observing that $\phi(I_d) = 0$ and $\phi_{*,I_d} = F^{p,q}$. ■

Log-Cholesky Metric. Lin (2019) introduced the Log-Cholesky metric which is obtained as a pullback Euclidean metric with respect to $L_d(\mathbb{R})$, the space of lower triangular matrices, endowed with the Frobenius inner product. The diffeomorphism between $S_d^{++}(\mathbb{R})$ and $L_d(\mathbb{R})$ is of the form $\phi : X \mapsto \varphi(\mathcal{L}(X))$ with $\mathcal{L} : S_d^{++}(\mathbb{R}) \to L_d^{++}(\mathbb{R})$ which returns the lower triangular matrix obtained by the Cholesky decomposition, *i.e.*, for $X = LL^T \in S_d^{++}(\mathbb{R})$, $\mathcal{L}(X) = L$, and $\varphi : L_d^{++}(\mathbb{R}) \to L_d(\mathbb{R})$ defined as $\varphi(L) = \lfloor L \rfloor + \log\big(\mathrm{diag}(L)\big)$ with $\lfloor \cdot \rfloor$ the strictly lower triangular part of the matrix and diag its diagonal part.

It is easy to see that $\phi(I_d) = 0$. We compute the differential of ϕ in Lemma 48 using the chain rule and (Lin, 2019, Proposition 4) which gives the differential of $\mathcal{L}$. Then, applying Proposition 7, we can compute the projection.

Proposition 15 *Let $\phi = \varphi \circ \mathcal{L}$. Then, for any $A \in S_d(\mathbb{R})$ such that $\|A\|^2_{I_d} = 1$, the coordinate projection is*

$$\forall X = LL^T \in S_d^{++}(\mathbb{R}),\ P^A(X) = \big\langle \lfloor L \rfloor, \lfloor A \rfloor \big\rangle_F + \big\langle \log(\mathrm{diag}(L)), \frac{1}{2}\mathrm{diag}(A)\big\rangle_F.$$

4.4 Product of Hadamard Manifolds.

In recent attempts to embed data into more flexible spaces, it has been proposed to use products of manifolds (Gu et al., 2019; Skopek et al., 2020; de Ocáriz Borde et al., 2023a,b) instead of constant curvature spaces, as real-world data may not be uniformly curved. Since products of constant curvature spaces do not necessarily have constant curvature, they offer greater flexibility for data embedding and they better capture the curvature of the underlying manifold. Since the product of Hadamard manifolds is still a Hadamard manifold (Gu et al., 2019), the product of hyperbolic spaces is a Hadamard manifold, and can be used to obtain flexible spaces *e.g.* by learning the curvature of the different spaces. Another example of a product of Hadamard manifolds is the Poincaré polydisk (Cabanes, 2022) which is the product manifold of $\mathbb{R}_+^*$ with the distance $d(x,y) = |\log(y/x)|$ and the Poincaré disk, and which has received attention for radar applications (Le Brigant, 2017). Note also that Gaussian distributions with diagonal covariance matrices endowed with the Fisher information matrix form a product of hyperbolic spaces (Cho et al., 2022a). Therefore, it is of interest to provide tools to compare probability distributions on products of Hadamard manifolds.

Let $\big((\mathcal{M}_i, g_i)\big)_{i=1}^n$ be n Hadamard manifolds and define the product manifold $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_n$. Then, at $x = (x_1, \dots, x_n) \in \mathcal{M}$, the tangent space is simply the inner product $T_x\mathcal{M} = T_{x_1}\mathcal{M}_1 \times \cdots \times T_{x_n}\mathcal{M}_n$, and $\mathcal{M}$ is equipped with the metric tensor $g = \sum_{i=1}^n g_i$. Moreover, for $v = (v_1, \dots, v_n) \in T_o\mathcal{M}$, the geodesic passing through the origin $o = (o_1, \dots, o_n)$ in direction v reads

$$\forall t \in \mathbb{R},\ \gamma_o(t) = \big(\gamma_{o_1}(t), \dots, \gamma_{o_n}(t)\big),$$

where γ_{o_i} is a geodesic in $\mathcal{M}_i$ passing through o_i in direction v_i. Moreover, the squared geodesic distance can be simply obtained as (Gu et al., 2019)

$$\forall x, y \in \mathcal{M},\ d_{\mathcal{M}}(x,y)^2 = \sum_{i=1}^n d_{\mathcal{M}_i}(x_i, y_i)^2.$$

Deriving the closed-form of the geodesic projection

$$t^* = \operatorname*{argmin}_{t \in \mathbb{R}} \sum_{i=1}^n d_{\mathcal{M}_i}\big(\gamma_{o_i}(t), y_i\big)^2$$

might depend on the context and may not be straightforward. Nonetheless, the Busemann function on a product of Hadamard manifolds is simply the weighted sum of the Busemann function on each geodesic line, and is thus easy to compute provided we know in closed-form the Busemann function on each manifold $\mathcal{M}_i$. This was first observed in (Bridson and Haefliger, 2013, Section II. 8.24) in the case of two manifolds, and we generalize the result to an arbitrary number of manifolds.

Proposition 16 (Busemann function on product Hadamard manifold) *Let $(\mathcal{M}_i)_{i=1}^n$ be n Hadamard manifolds and let $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_n$ be the product manifold. Let $\lambda_1, \dots, \lambda_n$ be such that $\sum_{i=1}^n \lambda_i^2 = 1$. For any $i \in \{1, \dots, n\}$, let γ_i be a geodesic line on $\mathcal{M}_i$ and define $\gamma : t \mapsto \big(\gamma_1(\lambda_1 t), \dots, \gamma_n(\lambda_n t)\big)$ a geodesic line on $\mathcal{M}$. Then,*

$$\forall x = (x_1, \dots, x_n) \in \mathcal{M},\ B^\gamma(x) = \sum_{i=1}^n \lambda_i B^{\gamma_i}(x_i).$$

In Section 6.2, we leverage this projection and the corresponding Sliced-Wasserstein distance to compare data sets viewed as distributions on $\mathbb{R}^{d_x} \times \mathbb{H}^{d_y}$.

5. Properties

In this section, we derive theoretical properties of the Cartan-Hyperbolic Sliced-Wasserstein distance. First, we will study its topology and the conditions required to have that CHSW is a true distance. In particular, we will first focus on the general case, and then on the specific case of pullback Euclidean metrics. Then, we will study some of its statistical properties. The proofs of this section are postponed to Appendix D.

5.1 Topology

Distance Property. First, we are interested in the distance properties of CHSW. From the properties of the Wasserstein distance and of the slicing process, we can show that it is a pseudo-distance, *i.e.*, that it satisfies the positivity, the symmetry and the triangular inequality.

Proposition 17 *Let $p \geq 1$, then* CHSW_p *is a finite pseudo-distance on $\mathcal{P}_p(\mathcal{M})$.*

For now, the lacking property is the one of indiscernibility, *i.e.*, that $\mathrm{CHSW}_p(\mu, \nu) = 0$ implies that $\mu = \nu$. We conjecture that it holds but we have not been able to prove it yet in full generality. In the following, we derive a sufficient condition on a related Radon transform for this property to hold. These derivations are inspired by (Boman and Lindskog, 2009; Bonneel et al., 2015).

Let $f \in L^1(\mathcal{M})$, and let us define, analogously to the Euclidean Radon transform, the Cartan-Hadamard Radon transform $\mathrm{CHR} : L^1(\mathcal{M}) \to L^1(\mathbb{R} \times S_o)$ which integrates the function f over a level set of the projection P^v:

$$\forall t \in \mathbb{R},\ \forall v \in S_o,\ \mathrm{CHR}f(t, v) = \int_{\mathcal{M}} f(x) \mathbb{1}_{\{t = P^v(x)\}}\, \mathrm{dVol}(x).$$

Then, we can also define its dual operator $\mathrm{CHR}^* : C_0(\mathbb{R} \times S_o) \to C_b(\mathcal{M})$ for $g \in C_0(\mathbb{R} \times S_o)$, where $C_0(\mathbb{R} \times S_o)$ is the space of continuous functions from $\mathbb{R} \times S_o$ to $\mathbb{R}$ that vanish at infinity and $C_b(\mathcal{M})$ is the space of continuous bounded functions from $\mathcal{M}$ to $\mathbb{R}$, as

$$\forall x \in \mathcal{M},\ \mathrm{CHR}^* g(x) = \int_{S_o} g(P^v(x), v)\, \mathrm{d}\lambda_o(v).$$

Proposition 18 CHR^* *is the dual operator of* CHR, i.e., *for all $f \in L^1(\mathcal{M})$, $g \in C_0(\mathbb{R} \times S_o)$,*

$$\langle \mathrm{CHR}f, g \rangle_{\mathbb{R} \times S_o} = \langle f, \mathrm{CHR}^* g \rangle_{\mathcal{M}}.$$

CHR^* maps $C_0(\mathbb{R} \times S_o)$ to $C_b(\mathcal{M})$ because g is necessarily bounded as a continuous function which vanishes at infinity. Note that CHR^* actually maps $C_0(\mathbb{R} \times S_o)$ to $C_0(\mathcal{M})$.

Proposition 19 *Let $g \in C_0(\mathbb{R} \times S_o)$, then* $\mathrm{CHR}^* g \in C_0(\mathcal{M})$.

Let us now recall the disintegration theorem.

Definition 20 (Disintegration of a measure) *Let $(Y, \mathcal{Y})$ and $(Z, \mathcal{Z})$ be measurable spaces, and $(X, \mathcal{X}) = (Y \times Z, \mathcal{Y} \otimes \mathcal{Z})$ the product measurable space. Then, for $\mu \in \mathcal{P}(X)$, we denote the marginals as $\mu_Y = \pi^Y_\# \mu$ and $\mu_Z = \pi^Z_\# \mu$, where π^Y (respectively π^Z) is the projection on Y (respectively Z). Then, a family $\big(K(y, \cdot)\big)_{y \in \mathcal{Y}}$ is a disintegration of μ if for all $y \in Y$, $K(y, \cdot)$ is a measure on Z, for all $A \in \mathcal{Z}$, $K(\cdot, A)$ is measurable and:*

$$\forall g \in C(X), \ \int_{Y \times Z} g(y, z) \ \mathrm{d}\mu(y, z) = \int_Y \int_Z g(y, z) K(y, \mathrm{d}z) \ \mathrm{d}\mu_Y(y),$$

where $C(X)$ is the set of continuous functions on X. We can denote $\mu = \mu_Y \otimes K$. K is a probability kernel if for all $y \in Y$, $K(y, Z) = 1$.

The disintegration of a measure actually corresponds to conditional laws in the context of probabilities. In the case where $X = \mathbb{R}^d$, we have existence and uniqueness of the disintegration (see (Santambrogio, 2015, Box 2.2) or (Ambrosio et al., 2008, Chapter 5) for the more general case).

Using the dual operator, we can define the Radon transform of a measure μ in $\mathcal{M}$ as the measure $\mathrm{CHR}\mu$ satisfying

$$\forall g \in C_0(\mathbb{R} \times S_o), \ \int_{\mathbb{R} \times S_o} g(t, v) \ \mathrm{d}(\mathrm{CHR}\mu)(t, v) = \int_{\mathcal{M}} \mathrm{CHR}^* g(x) \ \mathrm{d}\mu(x).$$

$\mathrm{CHR}\mu$ being a measure on $\mathbb{R} \times S_o$, we can disintegrate it *w.r.t.* the uniform measure on S_o as $\mathrm{CHR}\mu = \lambda_o \otimes K_\mu$ where K_μ is a probability kernel on $S_o \times \mathcal{B}(\mathbb{R})$. In the following proposition, we show that for λ_o-almost every $v \in S_o$, $K(v, \cdot)$ coincides with $P^v_\# \mu$.

Proposition 21 *Let μ be a measure on $\mathcal{M}$, and K_μ a probability kernel on $S_o \times \mathcal{B}(\mathbb{R})$ such that $\mathrm{CHR}\mu = \lambda_o \otimes K_\mu$. Then for λ_o-almost every $v \in S_o$, $K_\mu(v, \cdot) = P^v_\# \mu$.*

All these derivations allow to link the Cartan-Hadamard Sliced-Wasserstein distance with the Radon transform defined with the corresponding projection (geodesic or horospherical). Then, $\mathrm{CHSW}_p(\mu, \nu) = 0$ implies that for λ_o-almost every $v \in S_o$, $P^v_\# \mu = P^v_\# \nu$. Showing that the Radon transform is injective would allow us to conclude that $\mu = \nu$.

Actually, we derived here two different Cartan-Hadamard Radon transforms. Using P^v as the geodesic projection, the Radon transform integrates over geodesic subspaces of dimension $\dim(\mathcal{M}) - 1$. Such spaces are totally geodesic subspaces, and are related to the more general geodesic Radon transform (Rubin, 2003). In the case where the geodesic subspace is of dimension one, *i.e.*, it integrates only over geodesics, this coincides with the X-ray transform and it has been studied, *e.g.*, in (Lehtonen et al., 2018). Here, we are interested in the case of dimension $\dim(\mathcal{M}) - 1$, which, to the best of our knowledge, has only been studied in (Lehtonen, 2016) in the case where $\dim(\mathcal{M}) = 2$ and hence when the geodesic Radon transform and the X-ray transform coincide. However, no results on the injectivity over the sets of measures is yet available. In the case where P^v is the Busemann projection, the set of integration is a horosphere. To the best of our knowledge, general horospherical Radon transforms on Cartan-Hadamard manifolds have not yet been studied.

Link with the Wasserstein Distance. An important property of the Sliced-Wasserstein distance on Euclidean spaces is that it is topologically equivalent to the Wasserstein distance, *i.e.*, it metrizes the weak convergence. Such results rely on properties of the Fourier transform which do not translate straightforwardly to manifolds. Hence, deriving such results will require further investigations. We note that a possible lead for the horospherical case is the connection between the Busemann function and the Fourier-Helgason transform (Biswas, 2018; Sonoda et al., 2022). Using that the projections are Lipschitz functions, we can still show that CHSW is a lower bound of the geodesic Wasserstein distance.

Proposition 22 *Let $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$, then*

$$\mathrm{CHSW}_p^p(\mu, \nu) \leq W_p^p(\mu, \nu).$$

This property means that it induces a weaker topology compared to the Wasserstein distance, which can be computationally beneficial but which also comes with less discriminative power (Nadjahi et al., 2020).

Hilbert Embedding. CHSW also comes with the interesting property that it can be embedded in Hilbert spaces. This is in contrast with the Wasserstein distance which is known to not be Hilbertian (Peyré et al., 2019, Section 8.3) except in one dimension where it coincides with its sliced counterpart.

Proposition 23 *Let $p \geq 1$ and $\mathcal{H} = L^p([0,1] \times S_o, \mathrm{Leb} \otimes \lambda_o)$. We define Φ as*

$$\begin{array}{rccc} \Phi: & \mathcal{P}_p(\mathcal{M}) & \to & \mathcal{H} \\ & \mu & \mapsto & \big((q,v) \mapsto F^{-1}_{P^v_\# \mu}(q)\big), \end{array}$$

where $F^{-1}_{P^v_\# \mu}$ is the quantile function of $P^v_\# \mu$. Then CHSW_p is Hilbertian and for all $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$,

$$\mathrm{CHSW}_p^p(\mu, \nu) = \|\Phi(\mu) - \Phi(\nu)\|^p_{\mathcal{H}}.$$

This is a nice property which allows us to define a valid positive definite kernel for measures, such as the Gaussian kernel (Jayasumana et al., 2015, Theorem 6.1), and hence to use kernel methods (Hofmann et al., 2008). This can allow, for example, to perform distribution clustering, classification (Kolouri et al., 2016; Carriere et al., 2017) or regression (Meunier et al., 2022).

Proposition 24 *Define the kernel $K : \mathcal{P}_2(\mathcal{M}) \times \mathcal{P}_2(\mathcal{M}) \to \mathbb{R}$ as $K(\mu, \nu) = \exp\big(-\gamma \mathrm{CHSW}_2^2(\mu, \nu)\big)$ for $\gamma > 0$. Then K is a positive definite kernel.*

Proof Apply (Jayasumana et al., 2015, Theorem 6.1). ■

Bonet et al. (2023c) notably used this property to perform brain-age regression by first representing M/EEG data as a probability distribution of SPD matrices. And then by plugging the Gaussian kernel constructed with the Cartan-Hadamard Sliced-Wasserstein on the space of SPDs endowed with the Log-Euclidean metric, that we presented in Section 4.3.2, into the kernel Ridge regression method.

Note that to show that the Gaussian kernel is universal, *i.e.*, that the resulting Reproducing Kernel Hilbert Space (RKHS) is powerful enough to approximate any continuous function (Meunier et al., 2022), we would need additional results, such as that it metrizes the weak convergence and that CHSW_2 is a distance, as shown in (Meunier et al., 2022, Proposition 7).

5.2 Topology for Pullback Euclidean Manifolds

In this section, we focus on particular Hadamard manifolds for which the metric is a pullback Euclidean metric, which allows inheriting the properties of Euclidean spaces, and deriving additional properties of the corresponding Sliced-Wasserstein distance. This covers for example the space of SPD matrices with Pullback Euclidean metrics studied in Section 4.3.2 as well as the Mahalanobis manifold introduced in Section 4.1.

Let $\mathcal{N}$ be a Euclidean space with $\langle \cdot, \cdot \rangle$ its inner product and $\|\cdot\|$ the associated norm. Let $\phi : \mathcal{M} \to \mathcal{N}$ be a diffeomorphism and denote $(\mathcal{M}, g^\phi)$ the resulting Riemannian manifold (see Theorem 6 for more details). We recall that, by Proposition 7, the projection of $x \in \mathcal{M}$ on the geodesic characterized by the direction $v \in S_o$ is of the form

$$P^v(x) = \langle \phi(x) - \phi(o), \phi_{*,o}(v) \rangle.$$

In this case, given the formula, we can link CHSW with the Euclidean SW, with the integration made on $S_{\phi(o)} = \{v \in T_{\phi(o)}\mathcal{N},\ \|v\|_{\phi(o)} = 1\}$ with respect to the measure $\lambda_{\phi(o)}$.

Lemma 25 *Let $(\mathcal{M}, g^\phi)$ a pullback Euclidean Riemannian manifold and assume that $\lambda_{\phi(o)} = (\phi_{*,o})_\# \lambda_o$. Let $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$. Then,*

$$\mathrm{CHSW}_p^p(\mu, \nu) = \int_{S_{\phi(o)}} W_p^p(Q^v_\# \phi_\# \mu, Q^v_\# \phi_\# \nu)\ \mathrm{d}\lambda_{\phi(o)}(v) = \mathrm{SW}_p^p(\phi_\# \mu, \phi_\# \nu),$$

with $Q^v(x) = \langle x, v \rangle$ and SW_p the Euclidean Sliced-Wasserstein distance.

Using this simple lemma, we can leverage results known for the Euclidean Sliced-Wasserstein distance to CHSW on these particular spaces. First, we show that we recover the distance property by additionally showing the indiscernible property.

Proposition 26 *Let $(\mathcal{M}, g^\phi)$ a pullback Euclidean Riemannian manifold. Let $p \geq 1$, then CHSW_p is a finite distance on $\mathcal{P}_p(\mathcal{M})$.*

We can also obtain the important property that CHSW metrizes the weak convergence, as does the Wasserstein distance (Villani et al., 2009). This property was first shown for arbitrary measures in (Nadjahi et al., 2019) for the regular Euclidean SW.

Proposition 27 *Let $(\mathcal{M}, g^\phi)$ a pullback Euclidean Riemannian manifold of dimension d. Let $p \geq 1$, $(\mu_n)_n$ a sequence in $\mathcal{P}_p(\mathcal{M})$ and $\mu \in \mathcal{P}_p(\mathcal{M})$. Then, $\lim_{n\to\infty} \mathrm{CHSW}_p(\mu_n, \mu) = 0$ if and only if $(\mu_n)_n$ converges weakly towards μ.*

With these additional properties, we can also show that the corresponding Gaussian kernel is universal by applying (Meunier et al., 2022, Theorem 4). In addition to Proposition 22, we show that we can lower bound CHSW with the Wasserstein distance when the measures are compactly supported.

Proposition 28 *Let $(\mathcal{M}, g^\phi)$ a pullback Euclidean Riemannian manifold of dimension d. Let $p \geq 1$, $r > 0$ a radius and $B(o,r) = \{x \in \mathcal{M},\ d_{\mathcal{M}}(x,o) \leq r\}$ a closed ball. Then there exists a constant $C_{d,p,r}$ such that for all $\mu, \nu \in \mathcal{P}_p\big(B(o,r)\big)$,*

$$W_p^p(\mu,\nu) \leq C_{d,p,r}\mathrm{CHSW}_p(\mu,\nu)^{\frac{1}{d+1}}.$$

5.3 Statistical Properties

Sample Complexity. In practical settings, we usually cannot directly compute the closed-form between $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$, but we have access to samples $x_1, \dots, x_n \sim \mu$ and $y_1, \dots, y_n \sim \nu$. Then, it is common practice to estimate the discrepancy with the plug-in estimator $\mathrm{CHSW}_p^p(\hat{\mu}_n, \hat{\nu}_n)$ (Manole et al., 2021, 2022; Niles-Weed and Rigollet, 2022) where $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ are empirical estimations of the measures. We are interested in characterizing the speed of convergence of the plug-in estimator towards the true distance. Relying on the proof of Nadjahi et al. (2020, Corollary 2), we derive in Proposition 29 the sample complexity of CHSW. As in the Euclidean case, we find that the sample complexity does not depend on the dimension, which is an important and appealing property of sliced divergences (Nadjahi et al., 2020) compared to the Wasserstein distance, which has a sample complexity in $O(n^{-1/d})$ (Niles-Weed and Rigollet, 2022).

Proposition 29 *Let $p \geq 1$, $q > p$ and $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$. Denote $\hat{\mu}_n$ and $\hat{\nu}_n$ their counterpart empirical measures and $M_q(\mu) = \int_{\mathcal{M}} d(x,o)^q \, \mathrm{d}\mu(x)$ their moments of order q. Then, there exists $C_{p,q}$ a constant depending only on p and q such that*

$$\mathbb{E}\big[|\mathrm{CHSW}_p(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{CHSW}_p(\mu,\nu)|\big] \leq 2\alpha_{n,p,q} C_{p,q}^{1/p}\big(M_q(\mu)^{1/q} + M_q(\nu)^{1/q}\big),$$

where

$$\alpha_{n,p,q} = \begin{cases} n^{-1/(2p)} & \text{if } q > 2p, \\ n^{-1/(2p)}\log(n)^{1/p} & \text{if } q = 2p, \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases}$$

This property is very appealing in practical settings as it allows to use the same number of samples while having the same convergence rate in any dimension. In practice though, we cannot compute exactly $\mathrm{CHSW}_p(\hat{\mu}_n, \hat{\nu}_n)$ as the integral on S_o *w.r.t.* the uniform measure λ_o is intractable.

Projection Complexity. Thus, to compute it in practice, we usually rely on a Monte-Carlo approximation, by drawing $L \geq 1$ directions $v_1, \dots, v_L$ and approximating the distance by $\widehat{\mathrm{CHSW}}_{p,L}$ defined between $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ as

$$\widehat{\mathrm{CHSW}}_{p,L}^p(\mu,\nu) = \frac{1}{L}\sum_{\ell=1}^L W_p^p(P_\#^{v_\ell}\mu, P_\#^{v_\ell}\nu).$$

In the following proposition, we derive the Monte-Carlo error of this approximation, and we show that we recover the classical rate of $O(1/\sqrt{L})$.

Proposition 30 *Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$. Then, the error made by the Monte-Carlo estimate of* CHSW_p *with L projections can be bounded as follows*

$$\mathbb{E}_v\left[|\widehat{\mathrm{CHSW}}_{p,L}^p(\mu,\nu) - \mathrm{CHSW}_p^p(\mu,\nu)|\right]^2 \leq \frac{1}{L}\mathrm{Var}_v\left[W_p^p(P_\#^v\mu, P_\#^v\nu)\right].$$

We note that here the dimension actually intervenes in the term of variance.

Computational Complexity. As we project onto the real line, the complexity of computing the Wasserstein distances between each projected distribution is in $O(Ln \log n)$. Then, we add the complexity of computing the projections, which depends on the specific spaces and whether or not we have access to a closed-form, and the complexity of sampling the directions. For instance, in the hyperbolic case, using the closed-forms derived in Proposition 9, the projection procedure has a complexity of $O(nd)$ and thus the full complexity is in $O(Ln(\log n + d))$. For SPDs with the Log-Euclidean metric, the closed-form derived in Theorem 13 requires computing n matrix logarithm which has a complexity of $O(d^3)$ and Ln inner products with complexity $O(d^2)$. Moreover, sampling the directions can be done in $O(d^3)$ as detailed in (Bonet et al., 2023c, Section 2.4). Thus, the full complexity is in $O(Ln(\log n + d^2) + (L + n)d^3)$. We refer to (Bonet et al., 2023a, Figure 2) and (Bonet et al., 2023c, Figure 2) for runtime comparisons with different numbers of samples against the Sinkhorn algorithm and the Wasserstein distance.

6. Application of Cartan-Hadamard Sliced-Wasserstein Distances

In this section, we provide some illustrations of Cartan-Hadamard Sliced-Wasserstein distances on manifolds which were not yet studied in previous works. We note that Bonet et al. (2023a) used HSW to perform deep classification with prototypes on Hyperbolic spaces, while Bonet et al. (2023c) used SPDSW to perform domain adaptation for Brain Computer Interface and to perform Brain-Age prediction by leveraging the Gaussian kernel from Proposition 24 and plugging it into Kernel Ridge regression. Here, we first provide an experiment using the Mahalanobis Sliced-Wasserstein distance to classify documents, and then an experiment on a product of Cartan-Hadamard manifolds to compare data sets.

6.1 Document Classification with Mahalanobis Sliced-Wasserstein

We propose here to perform an experiment of document classification. Suppose that we have N documents $D_1, \dots, D_N$. Following the work of Kusner et al. (2015), we represent each document D_k as a distribution over words. More precisely, denote $x_1 \dots, x_n \in \mathbb{R}^d$ the set of words, embedded using `word2vec` (Mikolov et al., 2013) in dimension $d = 300$. Then, D_k is represented by the probability distribution $D_k = \sum_{i=1}^n w_i^k \delta_{x_i}$, where w_i^k represents the frequency of the word x_i in D_k normalized such that $\sum_{i=1}^n w_i^k = 1$.

Then, following (Huang et al., 2016), we learn a matrix $A \in S_d^{++}(\mathbb{R})$ using the Neighborhood Component Analysis (NCA) method (Goldberger et al., 2004) combined with the Word Centroid Distance (WCD), defined as $\mathrm{WCD}_A(D_k, D_\ell)^2 = (Xw^k - Xw^\ell)^T A(Xw^k - Xw^\ell)$ with $X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$. We use the `pytorch-metric-learning` library (Musgrave et al., 2020) to learn A.

Once A is learned, we compute the distance between documents using the Wasserstein distance or the Sliced-Wasserstein distance with Mahalanobis ground cost distance, *i.e.*, $d_A(x, y)^2 = (x - y)^T A(x - y)$. Once we compute the distance between each documents $\big(d(D_k, D_\ell)\big)_{k,\ell}$, we use a k-nearest neighbor classifier. On Table 1, we report the results for the BBCSport data set (Kusner et al., 2015), the Movies reviews data set (Pang et al., 2002) and the Goodread data set (Maharjan et al., 2017). All the data sets are split in 5 different

	BBCSport	**Movies**	**Goodreads genre**	**Goodreads like**
W_2	94.55	74.44	56.18	71.00
W_A	98.36	76.04	56.81	68.37
SW_2	$89.42_{\pm 0.89}$	$67.27_{\pm 0.69}$	$50.01_{\pm 1.21}$	$65.90_{\pm 0.17}$
$\mathrm{SW}_{2,A}$	$97.58_{\pm 0.04}$	$76.55_{\pm 0.11}$	$57.03_{\pm 0.68}$	$67.54_{\pm 0.14}$

Table 1: Accuracy for the document classification task.

train/test sets. The number of neighbors is found using a cross validation. We compare the results when using the regular Wasserstein and Sliced-Wasserstein distances, *i.e.*, with $A = I_d$, and when learning A using NCA with the WCD metric. The Wasserstein distance is computed using the Python Optimal Transport (`POT`) library (Flamary et al., 2021). The results for SW are averaged over 3 runs and SW is approximated with $L = 500$ projections.

With this simple initialization, we observe that the results obtained with the Mahalanobis Sliced-Wasserstein distance become very competitive with the ones obtained using the Wasserstein distance with the Mahalanobis ground cost. We note that the results might be further improved by performing then a NCA with W_A or $\mathrm{SW}_{2,A}$ as distances in the same spirit of (Huang et al., 2016). Here, we just use an initialization through WCD as a proof of concept to demonstrate how much it can already improve the results when using SW with a carefully chosen groundcost distance.

We showcase the computational benefits of using the Sliced-Wasserstein distance compared to the Wasserstein distance on Figure 7 by plotting the runtime for comparing each pair of documents, and on Table 2 with the full runtimes. We note that the Wasserstein distance is computed on CPU while the Sliced-Wasserstein distance is implemented in Pytorch and uses GPU. We used as CPU an Intel Xeon 4214 and as GPU a Titan RTX. We observe a computational gain even on small scale data sets where the documents contain few words, and therefore for which the underlying representative distributions contain few samples. For data sets with distributions with a larger number of samples such as goodreads, the computational benefits are pretty big. We sum up the statistics of the different data sets in Table 3.

6.2 Data Sets Comparisons with Sliced-Wasserstein on a Product Manifold

Assume we have data sets defined as sets of feature-label pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (Alvarez-Melis and Fusi, 2020), where the samples are in $\mathbb{R}^{d_x}$ and the labels are embedded in a Hyperbolic space $\mathbb{H}^{d_y}$. Then, a data set D_i can then be seen as a probability distribution on $\mathbb{R}^{d_x} \times \mathbb{H}^{d_y}$ which we can compare using CHSW on product manifolds.

We assume that the data sets are already embedded in such spaces. In practice, such embedding could come up for instance when we are given image-text pairs, which could be embedded both in Hyperbolic spaces *e.g.* using (Desai et al., 2023), or for more classical data sets using label embeddings methods (Akata et al., 2015).

Here, to get a data set represented in $\mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{H}^{d_y})$, we follow (Liu et al., 2025) and use a multidimensional scaling (MDS) method in hyperbolic spaces (Walter, 2004; Cvetkovski and Crovella, 2011) to get an embedding $\psi : \mathcal{P}(\mathbb{R}^{d_x}) \to \mathbb{H}^{d_y}$ into the hyperbolic space such that, for ν_y denoting the conditional probability distribution of samples in $\mathbb{R}^{d_x}$ with labels $y \in \mathcal{Y}$,

$$W_2^2(\nu_y, \nu_{y'}) \approx \alpha \cdot d_{\mathbb{H}}\big(\psi(\nu_y), \psi(\nu_{y'})\big)^2,$$

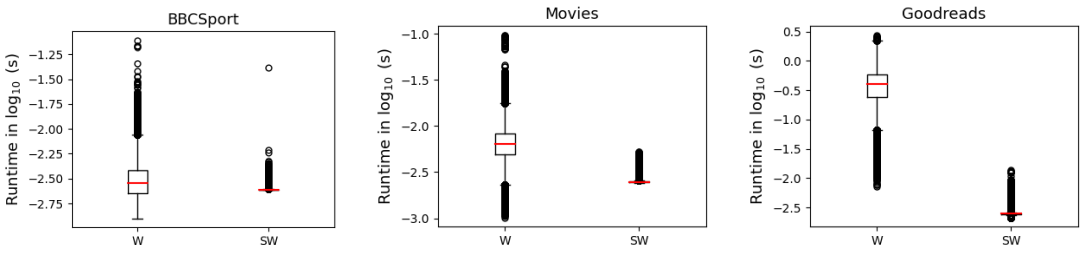


Figure 7: Runtime between each pair of documents.

		BBCSport	Movies	Goodreads
W_A	Average ($\cdot 10^{-3}$ s)	$3.29_{\pm 1.61}$	$6.78_{\pm 2.74}$	$440.30_{\pm 259}$
	Full (s)	891	13544	221252
$\mathrm{SW}_{2,A}$	Average ($\cdot 10^{-3}$s)	$2.45_{\pm 0.008}$	$2.47_{\pm 0.04}$	$2.5_{\pm 0.12}$
	Full (s)	665	4931	1256

Table 2: Runtimes for the document classification task (in averaged between the pairs of documents, or for the full computation of all the pairs).

with α some scaling parameter. To find this embedding, we minimize the absolute different squared loss (Cvetkovski and Crovella, 2011) defined as, for an original distance matrix $\Delta = (\delta_{i,j})_{i,j}$ and a scaling factor $\sqrt{\alpha}$,

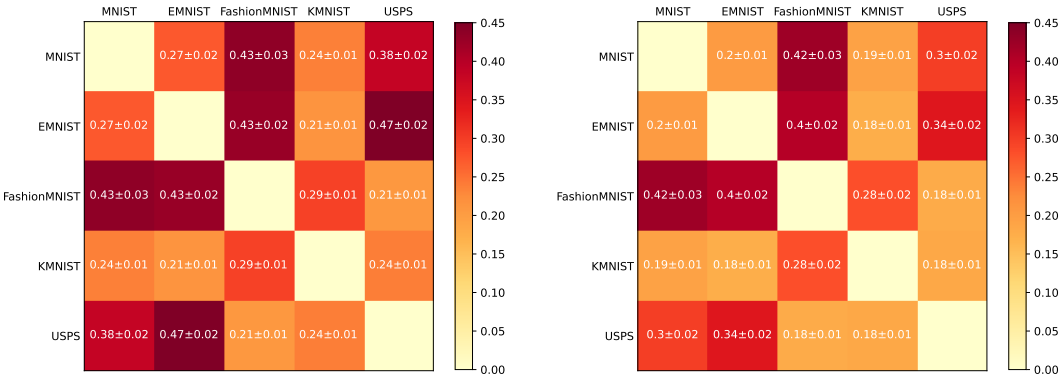
$$\forall z_1, \dots, z_n \in \mathbb{L}^{d_y},\ \mathcal{L}(z) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{\mathbb{L}}(z_i, z_j) - \sqrt{\alpha}\delta_{ij} \right)^2.$$

To improve the numerical stability, we perform the optimization in the tangent space following (Mishne et al., 2023) using the parametrization

$$z_i = \exp_{x^0}\left((0, \tilde{z}_i)\right) = \left(\cosh(\|\tilde{z}_i\|), \sinh(\|\tilde{z}_i\|) \frac{\tilde{z}_i}{\|\tilde{z}_i\|} \right)$$

for $\tilde{z}_i \in \mathbb{R}^{d_y - 1}$, and then performing the optimization in the Euclidean space.

We focus here on **NIST* data sets, which include MNIST (LeCun and Cortes, 2010), EMNIST (Cohen et al., 2017), FashionMNIST (Xiao et al., 2017), KMNIST (Clanuwat et al., 2018), and USPS (Hull, 1994). We plot on Figure 8 the matrix distance obtained between the **NIST* data sets either using SW between the data sets seen only through their features, *i.e.*, with $D_i \in \mathcal{P}(\mathbb{R}^{d_x})$, and using HCHSW on the space $\mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{L}^{d_y})$ where the labels were embedded in $\mathbb{L}^{d_y}$ using the method described in the previous paragraph with a scaling of $\sqrt{\alpha} = 0.1$ and $d_y = 10$. We observe that when the labels are not taken into account, the USPS and MNIST data sets have a huge discrepancy between each other. However, when taking into account the labels, we recover that these two data sets are in fact more similar as they both represent numbers. Thus, we argue that using the sliced distance on the product data set to take into account the labels provides better comparisons of the data sets. Furthermore, from a computation point of view, CHSW on the product manifold



(a) SW. (b) Product HCHSW.

Figure 8: Comparison between SW between the data sets and HCHSW between the data sets embedded on $\mathbb{R}^{d_x} \times \mathbb{L}^{d_y}$. Results are averaged over 100 draws of projections.

is much cheaper compared to *e.g.* computing the Wasserstein distance. On our experiments, computing the full distance matrix with CHSW took in average 0.05s against 120s to compute the Wasserstein distance, where we used here only 10000 samples of the data sets.

7. Cartan-Hadamard Sliced-Wasserstein Flows

We now propose to derive the Wasserstein gradient flows of the CHSW distances, along with non-parametric particle schemes that approximate the flows. We provide first the results on general Hadamard manifolds and then we specify them to Mahalanobis manifolds, Hyperbolic spaces and SPDs endowed with the Log-Euclidean metric. The proofs of this section are postponed to Appendix E.

7.1 Wasserstein Gradient Flows

First Variations. On Hadamard manifolds, CHSW discrepancies can be used to learn parametric or empirical distributions by minimizing them. One possible solution is to leverage Wasserstein gradient flows (Ambrosio et al., 2008; Santambrogio, 2017) of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{CHSW}_2^2(\mu, \nu)$, where ν is some target distribution. Approximating this flow would then allow providing new samples from ν. Computing such a flow requires first computing the first variations of the given functional. As a first step towards computing Wasserstein gradient flows of CHSW on Hadamard spaces, and analyzing them, we derive in Proposition 31 the differential of $\mathcal{F}$ in the Wasserstein space.

Proposition 31 *Let K be a compact subset of $\mathcal{M}$, $\mu, \nu \in \mathcal{P}_2(K)$ with $\mu \ll \mathrm{Vol}$. Let $v \in S_o$, denote ψ_v the Kantorovich potential between $P^v_\# \mu$ and $P^v_\# \nu$ for the cost $c(x,y) = \frac{1}{2}|x-y|^2$ for $x, y \in \mathbb{R}$. Let ξ be a diffeomorphic vector field on K and denote for all $\epsilon \geq 0$, $T_\epsilon : K \to \mathcal{M}$*

defined as $T_\epsilon(x) = \exp_x\big(\epsilon\xi(x)\big)$ *for all* $x \in K$. *Then,*

$$\lim_{\epsilon\to 0^+} \frac{\mathrm{CHSW}_2^2\big((T_\epsilon)_\#\mu,\nu\big) - \mathrm{CHSW}_2^2(\mu,\nu)}{2\epsilon} = \int_{S_o}\int_{\mathcal{M}} \psi_v'\big(P^v(x)\big)\langle \mathrm{grad}_{\mathcal{M}} P^v(x), \xi(x)\rangle_x \,\mathrm{d}\mu(x)\,\mathrm{d}\lambda_o(v).$$

In the Euclidean case, we recover well the formula of the differential of SW first derived in (Bonnotte, 2013, Proposition 5.1.7). Indeed, for $x \in \mathbb{R}^d$, $T_\epsilon(x) = x + \epsilon\xi(x)$, and for $\theta \in S^{d-1}$, $P^\theta(x) = \langle x, \theta\rangle$. Thus $\mathrm{grad}P^\theta(x) = \nabla P^\theta(x) = \theta$, and we recover

$$\lim_{\epsilon\to 0^+} \frac{\mathrm{SW}_2^2\big((\mathrm{Id}+\epsilon\xi)_\#\mu,\nu\big) - \mathrm{SW}_2^2(\mu,\nu)}{2\epsilon} = \int_{S^{d-1}}\int_{\mathbb{R}^d} \psi_\theta'\big(P^\theta(x)\big)\big\langle \theta, \xi(x)\big\rangle \,\mathrm{d}\mu(x)\,\mathrm{d}\lambda(\theta).$$

Cartan-Hadamard Sliced-Wasserstein Flow. Given the differential, we can derive the Wasserstein gradient flow of $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{CHSW}_2^2(\mu,\nu)$ as the continuity equation governed by the vector field v_t obtained through the Wasserstein gradient

$$\forall x \in \mathcal{M},\ v_t(x) = -\nabla_{W_2}\mathcal{F}(\mu_t)(x) = -\int_{S_o} \psi_{t,v}'\big(P^v(x)\big)\mathrm{grad}_{\mathcal{M}} P^v(x)\,\mathrm{d}\lambda_o(v),$$

with $\psi_{t,v}$ the Kantorovich potential between $P^v_\#\mu_t$ and $P^v_\#\nu$ such that $\psi_{t,v}'(x) = x - F^{-1}_{P^v_\#\mu_t}\big(F_{P^v_\#\nu}(x)\big)$, *i.e.*, the Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ of $\mathcal{F}$ is a solution (in the distributional sense) of

$$\partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0.$$

Forward Euler Scheme. To provide an algorithm to sample from ν by minimizing $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{CHSW}_2^2(\mu,\nu)$ while following its Wasserstein gradient flow, there are several possible strategies of discretization of the flow. For instance, a solution could be to compute the backward Euler scheme, also known as the Jordan-Kinderlehrer-Otto (JKO) scheme from the seminal work of Jordan et al. (1998). This strategy has for example been used to minimize the Sliced-Wasserstein distance in (Bonet et al., 2022). Here, we propose instead to use the forward Euler scheme, which allows defining a particle scheme approximating the trajectory of the Wasserstein gradient flow. Such a strategy has been used to minimize different functionals such as the MMD (Arbel et al., 2019), the Kernel Stein Discrepancy (Korba et al., 2021) or the KL divergence (Fang et al., 2021; Wang et al., 2022). For SW, Liutkus et al. (2019) proposed to minimize SW with an entropy term, which required to use a McKean Vlasov SDE.

Let $\mu_0 \in \mathcal{P}_p(\mathcal{M})$ and $\tau > 0$. On a Riemannian manifold, analogously to the Riemannian gradient descent (Bonnabel, 2013), the forward Euler scheme becomes

$$\forall k \geq 0,\ \mu_{k+1} = \exp_{\mathrm{Id}}\big(-\tau\nabla_{W_2}\mathcal{F}(\mu_k)\big)_\#\mu_k,$$

where $\nabla_{W_2}\mathcal{F}(\mu_k)$ is the Wasserstein gradient, and is defined as $\nabla_{W_2}\mathcal{F}(\mu_k)(x) = -v_k(x) = \int_{S_o}\psi_{k,v}'\big(P^v(x)\big)\mathrm{grad}_{\mathcal{M}}P^v(x)\,\mathrm{d}\lambda_o(v)$ for $x \in \mathcal{M}$. In the Euclidean case, we recover the usual forward Euler scheme $\mu_{k+1} = \big(\mathrm{Id} - \tau\nabla_{W_2}\mathcal{F}(\mu_k)\big)_\#\mu_k$.

Algorithm 1 Wasserstein gradient flows of CHSW

Input: $(y_j)_{j=1}^n \sim \nu$, μ_0, L the number of projections, N the number of steps
Sample $(x_i^0)_{i=1}^n \sim \mu_0$
for $k = 0$ **to** $N-1$ **do**
  Draw $v_1, \dots, v_L \sim \lambda_o$
  Compute $\hat{x}_{i,\ell}^k = P^{v_\ell}(x_i^k)$, $\hat{y}_{j,\ell} = P^{v_\ell}(y_j)$ for all $\ell \in \{1, \dots, L\}$
  Define $P^{v_\ell}_{\#}\hat{\nu} = \frac{1}{n}\sum_{j=1}^n \delta_{\hat{y}_{j,\ell}}$, $P^{v_\ell}_{\#}\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{x}_{i,\ell}^k}$
  Compute $\hat{z}_{i,\ell}^k = \hat{x}_{i,\ell}^k - F^{-1}_{P^{v_\ell}_{\#}\hat{\nu}}(F_{P^{v_\ell}_{\#}\hat{\mu}_k}(\hat{x}_{i,\ell}^k))$
  Compute $g_\ell(x_i^k) = \mathrm{grad}_{\mathcal{M}} P^{v_\ell}(x_i^k)$
  Compute $\hat{v}_k(x_i^k) = \frac{1}{L}\sum_{\ell=1}^L (\hat{x}_{i,\ell}^k - \hat{z}_{i,\ell}^k) g_\ell(x_i^k)$
  For all $i \in \{1, \dots, n\}$, $x_i^{k+1} = \exp_{x_i^k}\left(\tau \hat{v}_k(x_i^k)\right)$
end for

In practice, we approximate the Wasserstein gradient by first sampling $v_1, \dots, v_L \sim \lambda_o$ and then using

$$\forall x \in \mathcal{M},\ \hat{v}_k(x) = -\frac{1}{L}\sum_{\ell=1}^L \psi'_{v_\ell,k}\big(P^{v_\ell}(x)\big)\mathrm{grad}_{\mathcal{M}} P^{v_\ell}(x), \tag{4}$$

where

$$\psi'_{v,k}\big(P^v(x)\big) = P^v(x) - F^{-1}_{P^v_{\#}\nu}\big(F_{P^v_{\#}\mu_k}(P^v(x))\big).$$

Following (Liutkus et al., 2019), the cumulative distribution functions and the quantiles are approximated using linear interpolations between the true points.[3] Finally, the particle scheme is given by,

$$\forall k \geq 0, i \in \{1, \dots, n\},\ x_i^{k+1} = \exp_{x_i^k}\left(\tau \hat{v}_k(x_i^k)\right).$$

We sum up the procedure in Algorithm 1.

7.2 Application to the Mahalanobis Manifold

For pullback Euclidean metrics, the Riemannian gradient can be obtained by using the inverse of the differential operator as stated in the following lemma.

Lemma 32 (Lemma 4 in (Chen et al., 2024a)) *Let $(\mathcal{M}, g^\phi)$ be a Pullback Euclidean Riemannian manifold. For $f : \mathcal{M} \to \mathbb{R}$ a smooth map, the gradient is of the form*

$$\forall x \in \mathcal{M},\ \mathrm{grad}_{\mathcal{M}} f(x) = \phi^{-1}_{*,x}\big(\phi^{-*}_{*,x}\big(\nabla f(x)\big)\big).$$

For the Mahalanobis distance, *i.e.*, for $\phi(x) = A^{\frac{1}{2}}x$ for any $x \in \mathbb{R}^d$ with $A \in S_d^{++}(\mathbb{R})$, the inverse of the differential is simply $\phi^{-1}_{*,x}(v) = A^{-\frac{1}{2}}v$, and we recall that the projection is $P^v(x) = x^T Av$ for $v \in S_o$. Thus the Riemannian gradient of the projection P^v for $v \in S_o$ is

$$\mathrm{grad}_{\mathcal{M}} P^v(x) = A^{-\frac{1}{2}}\big(A^{-\frac{1}{2}}(Av)\big) = v.$$

3. using `https://github.com/aliutkus/torchinterp1d`

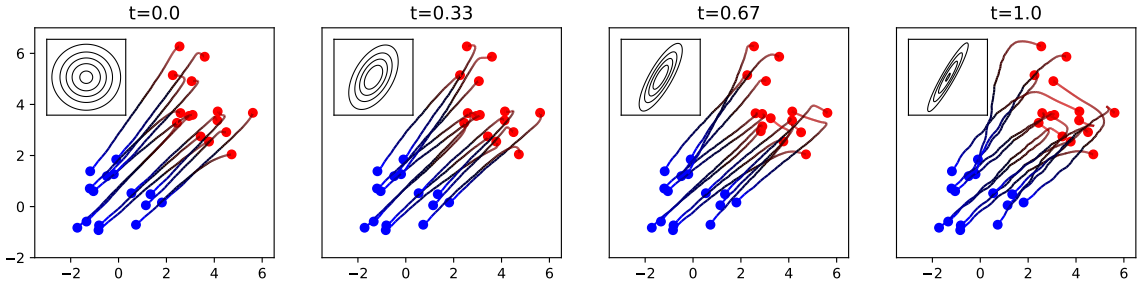


Figure 9: Trajectories of Mahalanobis Sliced-Wasserstein flows using four SPD matrices A_t along the geodesic between I_2 and a randomly chosen $A \in S_d^{++}(\mathbb{R})$. Ellipses represent the matrices A_t.

We recover the same gradient. But the matrix A is still involved in the formula of the projection, which can change the trajectory of the particles. Choosing well the matrix A can help improving the convergence of flows for ill conditioned problems, see *e.g.* (Duchi et al., 2011; Dong et al., 2023).

We illustrate, on Figure 9, the effect on the trajectory when using a randomly sampled SPD matrix A to specify the Mahalanobis distance compared to the classical Euclidean metric. We plot the trajectories for different SPDs obtained on the geodesic between I_2 and A, which is of the form $A_t = \exp\big(t \log(A)\big)$ for $t \in [0, 1]$ when using the Affine-Invariant metric.

7.3 Application to Hyperbolic Spaces

Here, we propose to minimize the Hyperbolic Sliced-Wasserstein distances in order to derive new non-parametric schemes allowing to learn a distribution given its samples. We first recall how to compute the gradient on the Lorentz model.

Proposition 33 *Let $f : \mathbb{L}_K^d \to \mathbb{R}$ and note $\bar{f} : \mathbb{R}^{d+1} \to \mathbb{R}$ a smooth extension on $\mathbb{R}^{d+1}$. Then, the gradient of f at $x \in \mathbb{L}_K^d$ is*

$$\operatorname{grad}_{\mathbb{L}_K^d} f(x) = \operatorname{Proj}_x^K\big(-KJ\nabla \bar{f}(x)\big),$$

where $J = \operatorname{diag}(-1, 1, \dots, 1)$ and

$$\operatorname{Proj}_x^K(z) = z - K\langle x, z\rangle_{\mathbb{L}} x.$$

Proof We extend (Boumal, 2023, Proposition 7.7) to $\mathbb{L}_K^d$. ■

Then, leveraging Proposition 33, we derive the closed-forms of the gradients of the geodesic and horospherical projections, which allows deriving the forward Euler scheme of this functional, by plugging the different formulas in Equation (4).

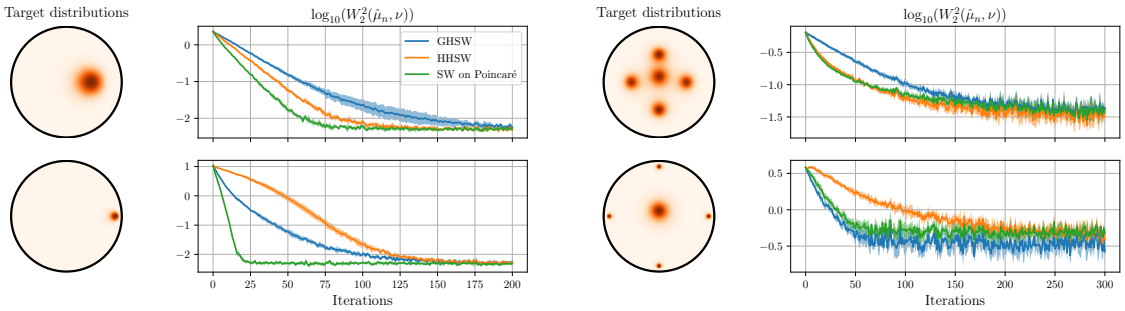


Figure 10: Log 2-Wasserstein between the target distribution and particles obtained from HSWFs (averaged over 5 runs).

Proposition 34 *Let $v \in T_{x^0}\mathbb{L}_K^d \cap S^d$ and $x \in \mathbb{L}_K^d$, then*

$$\operatorname{grad}_{\mathbb{L}_K^d} B^v(x) = K\sqrt{-K}\left(Kx - \frac{\sqrt{-K}x^0 + v}{\langle x, \sqrt{-K}x^0 + v\rangle_{\mathbb{L}}}\right),$$
$$\operatorname{grad}_{\mathbb{L}_K^d} P^v(x) = \frac{K^2\left(\langle x, x^0\rangle_{\mathbb{L}} v - \langle x, v\rangle_{\mathbb{L}} x^0\right)}{\langle x, v\rangle_{\mathbb{L}}^2 + K\langle x, x^0\rangle_{\mathbb{L}}^2}.$$

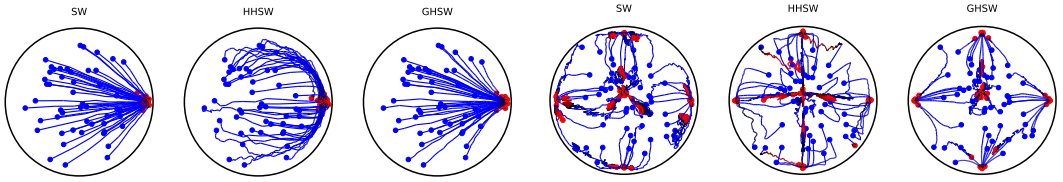
On $\mathbb{B}^d$, the gradient can be obtained by rescaling the Euclidean gradient with the inverse Poincaré ball metric (Nickel and Kiela, 2017) which is $\left(\frac{1+K\|x\|_2^2}{2}\right)^2$ (Park et al., 2021). Thus, we can also derive the corresponding formulas on the Poincaré ball. For example, for the Busemann function, we have

$$\nabla B^{\tilde{v}}(x) = 2\left(\frac{x}{1 - \|x\|_2^2} - \frac{\tilde{v} - x}{\|\tilde{v} - x\|_2^2}\right),$$

and therefore its Riemannian gradient is

$$\operatorname{grad}_{\mathbb{B}_K^d} B^{\tilde{v}}(x) = \left(\frac{1 + K\|x\|_2^2}{2}\right)^2 \nabla B^{\tilde{v}}(x).$$

In Figure 10, we plot the 2-Wasserstein distance between the target distribution and samples from the Hyperbolic Sliced-Wasserstein Flows on Hyperbolic space of curvature $K = -1$. We compare the evolution among GHSW, HHSW and SW (on the Poincaré ball for SW) across 4 different scenarios. The two first ones involve a target distribution which is a Wrapped Normal Distribution (WND) located either close to the center or to the border of the disk. The second ones involve a mixture of WNDs, with some modes either close to the border or to the center. HHSW and GHSW can be done both on the Lorentz model or the Poincaré ball. Using either model give similar results. As hyperparameters, we chose $n = 500$ particles, a learning rate of $\tau = 0.1$ with $N = 200$ epochs for centered targets, and $\tau = 0.5$ with $N = 300$ epochs for bordered targets. We note that the three gradient flows perform similarly, with an advantage of speed for SW. This might be due to the fact that



(a) WND. (b) Mixture of WNDs.

Figure 11: Trajectories of 50 particles when the target is the WND on the border or the Mixture of WNDs on the border.

the minimization is done in the space of Euclidean probabilities, and thus does not take into account that the modes are actually on the border.

We add on Figure 11a and Figure 11b trajectories for the border scenarios. When minimizing GHSW, particles tend to go to the modes by following the shortest path, while when minimizing HHSW, they tend to first go to the border before converging to the modes. As the distance on the border of the Poincaré disk are bigger than to the center, this may explain the observed slower convergence of HHSW in Figure 10.

7.4 Application to SPD matrices with the Log-Euclidean Metric

For SPDSW with the Log-Euclidean metric, the formula of the gradient can be derived using the inverse of the differential as stated in Lemma 32. We report the inverse of the differential of the log in Lemma 35.

Lemma 35 *Let $\phi : X \mapsto \log(X)$ and $X = UDU^T \in S_d^{++}(\mathbb{R})$ where $D = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. Then, we have*

$$\forall W \in T_{\phi(X)} S_d^{++}(\mathbb{R}),\ \phi_{*,X}^{-1}(W) = U\tilde{\Sigma}(W)U^T,$$

where $\tilde{\Sigma}(W) = U^T W U \oslash \Gamma$ with Γ defined as in Lemma 12.

Finally, in Lemma 36, we report the gradient of the projection obtained with the Log-Euclidean metric, which can be obtained using that the differential of the matrix log satisfies $\langle A, \log_{*,X}(V)\rangle_F = \langle \log_{*,X}(A), V\rangle_F$ for any $A, V \in S_d(\mathbb{R})$.

Lemma 36 *Let $A \in S_d(\mathbb{R})$ and $X = UDU^T \in S_d^{++}(\mathbb{R})$ with $U = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. Then,*

$$\nabla P^A(X) = U\Sigma(A)U^T.$$

We now have all the tools to apply Algorithm 1 for the particular case of SPDSW. In Figure 12, trajectories plotted inside the $S_2^{++}(\mathbb{R})$ cone depict the evolution of the matrices along the gradient flow. The noisy behavior of some of them can be mostly explained by numerical instabilities arising from the different matrix operators used in the process, which require to use small step sizes.

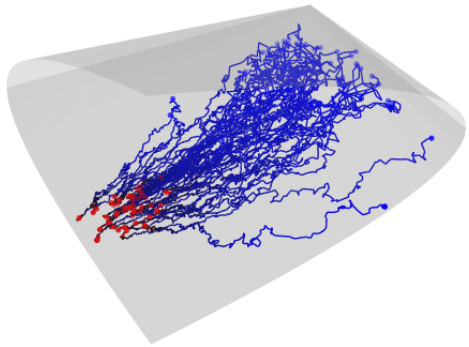


Figure 12: Trajectories of particles following the Wasserstein gradient flow of SPDSW.

8. Future Works and Discussions

In this article, we formally introduced a way to generalize the Sliced-Wasserstein distance on Riemannian manifolds of non-positive curvature and specified this construction for different particular cases: pullback Euclidean metrics, Hyperbolic spaces, the space of Symmetric Positive-Definite matrices, and product of Hadamard manifolds. These new discrepancies can be computed very efficiently and scale to distributions composed of a large number of samples in contrast to the computation of the Wasserstein distance. We also analyzed these constructions theoretically while providing new applications and non-parametric schemes to minimize them using Wasserstein gradient flows.

Further works might include studying other Hadamard manifolds for which we do not necessarily have a closed-form for the projections such as Siegel spaces (Cabanes, 2022), or extending this construction to more general manifolds, such as Riemannian manifolds of non-negative curvature, Finsler manifolds (Shen, 2001), which have recently received some attention in Machine Learning (López et al., 2021; Pouplin et al., 2023; Lin et al., 2023), or more generally, metric spaces.

For projections, we studied two natural generalizations of the projection used in Euclidean spaces. We could also study other projections which do not follow geodesics subspaces or horospheres, but are well suited to Riemannian manifolds, in the same spirit of the Generalized Sliced-Wasserstein. Other subspaces could also be used, such as Hilbert curves (Bernton et al., 2019; Li et al., 2024) adapted to manifolds, or higher dimensional subspaces (Paty and Cuturi, 2019; Chami et al., 2021). Finally, we could also define other variations of CHSW such as max-CHSW for instance, and more generally adapt many of the variants which have been proposed for SW to the case of Riemannian manifolds. Note also that we could plug these constructions into the framework introduced by Bonet et al. (2024) to compare positive measures on Hadamard manifolds.

On the theoretical side, we still need to show that these Sliced-Wasserstein discrepancies are proper distances by showing the indiscernible property. It might also be interesting to study whether statistical properties for the Euclidean SW distance, derived in *e.g.*, (Nietert et al., 2022; Manole et al., 2022; Goldfeld et al., 2022; Xu and Huang, 2022; Xi and Niles-Weed, 2022) still hold more generally for CHSW on any Cartan-Hadamard manifold, or to study the properties of the space of probabilities endowed with these distances, such as

geodesic properties or the gradient flows in this space, as it was recently done in (Candau-Tilh, 2020; Bonet et al., 2022; Park and Slepčev, 2023; Kitagawa and Takatsu, 2023) for the Euclidean Sliced-Wasserstein distance.

Acknowledgments

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). Clément Bonet's research was partially supported by the center Hi! PARIS.

Appendix A. Useful Lemmas

We derive here some lemmas which will be useful for the proofs.

Lemma 37 (Lemma 6 in (Paty and Cuturi, 2019)) *Let $\mathcal{M}$, $\mathcal{N}$ be two Riemannian manifolds. Let $f : \mathcal{M} \to \mathcal{N}$ be a measurable map and $\mu, \nu \in \mathcal{P}(\mathcal{M})$. Then,*

$$\Pi(f_{\#}\mu, f_{\#}\nu) = \{(f \otimes f)_{\#}\gamma,\ \gamma \in \Pi(\mu, \nu)\}.$$

Proof This is a straightforward extension of (Paty and Cuturi, 2019, Lemma 6). ■

Lemma 38 *Let $(\mathcal{M}, g)$ be a Hadamard manifold with origin o. Let $v \in T_o\mathcal{M}$, then*

1. *the geodesic projection P^v is 1-Lipschitz.*
2. *the Busemann function B^v is 1-Lipschitz.*

Proof

1. By Proposition 2, we know that

$$\forall x, y \in \mathcal{M},\ |P^v(x) - P^v(y)| = d\big(\tilde{P}^v(x), \tilde{P}^v(y)\big).$$

   Moreover, by (Ballmann et al., 2006, Page 9), $\tilde{P}^v$ is 1-Lipschitz, so is P^v.

2. The Busemann function is 1-Lipschitz, see *e.g.* (Bridson and Haefliger, 2013, II. Proposition 8.22).

■

Lemma 39 *Let d be a metric on $\mathcal{M}$. Then, for any $p \geq 1$,*

$$\forall x, y \in \mathcal{M},\ d(x, y)^p \leq 2^{p-1}\big(d(x, o)^p + d(o, y)^p\big).$$

Lemma 40 (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 $n \geq 1$,*

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

Lemma 41 *Let $y \in \mathcal{M}$ and denote for all $x \in \mathcal{M}$, $f(x) = d(x, y)^2$. Then,* $\mathrm{grad}_{\mathcal{M}} f(x) = -2\log_x(y)$.

For references about Lemma 41, see *e.g.* (Chewi et al., 2020, Appendix A) or (Goto and Sato, 2021).

Appendix B. Proofs of Section 3

B.1 Proof of Proposition 2

Proof of Proposition 2 Let $x, y \in \mathcal{G}^v$. Then, there exists $s, t \in \mathbb{R}$ such that $x = \exp_o(sv)$ and $y = \exp_o(tv)$. By a simple calculation, we have on one hand that

$$\begin{aligned}\operatorname{sign}(\langle \log_o(x), v\rangle_o) &= \operatorname{sign}(\langle \log_o(\exp_o(sv)), v\rangle_o)\\ &= \operatorname{sign}(s\|v\|_o^2)\\ &= \operatorname{sign}(s),\end{aligned}$$

using that $\log_o \circ \exp_o = \mathrm{Id}$. And similarly, $\operatorname{sign}(\langle \log_o(y), v\rangle_o) = \operatorname{sign}(t)$.

Then, by noting that $o = \exp_o(0)$, and recalling that $d(x, y) = d(\exp_o(tv), \exp_o(sv)) = |t - s|\|v\|_o$,

$$\begin{aligned}|t^v(x) - t^v(y)| &= |\operatorname{sign}(\langle \log_o(x), v\rangle_o) d(x, o) - \operatorname{sign}(\langle \log_o(y), v\rangle_o d(y, o)|\\ &= \big|\operatorname{sign}(s) d(\exp_o(sv), \exp_o(0)) - \operatorname{sign}(t) d(\exp_o(tv), \exp_o(0))\big|\\ &= \big|\operatorname{sign}(s)|s| - \operatorname{sign}(t)|t|\big| \cdot \|v\|_o\\ &= |s - t|\|v\|_o\\ &= d(x, y).\end{aligned}$$

■

B.2 Proof of Proposition 3

Proof of Proposition 3 We want to solve:

$$P^v(x) = \operatorname*{argmin}_{t \in \mathbb{R}} \; d\big(\gamma(t), x\big)^2,$$

where $\gamma(t) = \exp_o(tv)$. For $t \in \mathbb{R}$, let $g(t) = d\big(\gamma(t), x\big)^2 = f\big(\gamma(t)\big)$ where $f(x) = d(x, y)^2$ for $x, y \in \mathcal{M}$. Then, by Lemma 41, we have for any $t \in \mathbb{R}$,

$$\begin{aligned}g'(t) = 0 &\iff \langle \gamma'(t), \operatorname{grad}_{\mathcal{M}} f\big(\gamma(t)\big)\rangle_{\gamma(t)} = 0\\ &\iff \langle \gamma'(t), -2 \log_{\gamma(t)}(x)\rangle_{\gamma(t)} = 0.\end{aligned}$$

■

B.3 Proof of Proposition 4

Proof of Proposition 4 First, we note that $P^v = t^v \circ \tilde{P}^v$. Then, by using Lemma 37 which states that $\Pi(f_{\#}\mu, f_{\#}\nu) = \{(f \otimes f)_{\#}\gamma,\ \gamma \in \Pi(\mu, \nu)\}$ for any f measurable, as well

as that by Proposition 2, $|t^v(x) - t^v(y)| = d(x, y)$, we have:

$$\begin{aligned}
W_p^p(P^v_\# \mu, P^v_\# \nu) &= \inf_{\gamma \in \Pi(P^v_\# \mu, P^v_\# \nu)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p \, \mathrm{d}\gamma(x,y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p \, \mathrm{d}(P^v \otimes P^v)_\# \gamma(x,y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} |P^v(x) - P^v(y)|^p \, \mathrm{d}\gamma(x,y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} |t^v(\tilde{P}^v(x)) - t^v(\tilde{P}^v(y))|^p \, \mathrm{d}\gamma(x,y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d\big(\tilde{P}^v(x), \tilde{P}^v(y)\big)^p \, \mathrm{d}\gamma(x,y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d(x,y)^p \, \mathrm{d}(\tilde{P}^v \otimes \tilde{P}^v)_\# \gamma(x,y) \\
&= \inf_{\gamma \in \Pi(\tilde{P}^v_\# \mu, \tilde{P}^v_\# \nu)} \int_{\mathcal{G}^v\times\mathcal{G}^v} d(x,y)^p \, \mathrm{d}\gamma(x,y) \\
&= W_p^p(\tilde{P}^v_\# \mu, \tilde{P}^v_\# \nu).
\end{aligned}$$

Now, let us show the results when using the Busemann projection. Let $v \in T_o\mathcal{M}$ such that $\|v\|_o = 1$, and recall that $\tilde{B}^v(x) = \exp_o(-B^v(x)v)$. First, let us compute $t^v \circ \tilde{B}^v$:

$$\begin{aligned}
\forall x \in \mathcal{M},\ t^v\big(\tilde{B}^v(x)\big) &= \mathrm{sign}(\langle \log_o(\tilde{B}^v(x)), v\rangle_o)\ d(\tilde{B}^v(x), o) \\
&= \mathrm{sign}\big(\langle \log_o(\exp_o(-B^v(x)v)), v\rangle_o\big)\ d\big(\exp_o(-B^v(x)v), \exp_o(0)\big) \\
&= \mathrm{sign}(-B^v(x)\|v\|_o^2)\ d(\exp_o(-B^v(x)v), \exp_o(0)) \\
&= \mathrm{sign}(-B^v(x))|-B^v(x)| \\
&= -B^v(x).
\end{aligned}$$

Then, using the same computation as before, we get

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

■

Appendix C. Proofs of Section 4

C.1 Proof of Proposition 7

Proof of Proposition 7

Geodesic projection. Let $x \in \mathcal{M}$. Denote $f : \mathbb{R} \to \mathbb{R}$ such that

$$\begin{aligned}
f(t) &= d_{\mathcal{M}}\big(\gamma(t), x\big)^2 \\
&= d_{\mathcal{M}}\big(\phi^{-1}(\phi(o) + t\phi_{*,o}(v)), x\big)^2 \\
&= \big\|\phi\big(\phi^{-1}(\phi(o) + t\phi_{*,o}(v))\big) - \phi(x)\big\|^2 \\
&= t^2\|\phi_{*,o}(v)\|^2 - 2t\langle \phi(x) - \phi(o), \phi_{*,o}(v)\rangle + \|\phi(o) - \phi(x)\|^2 \\
&= t^2 - 2t\langle \phi(x) - \phi(o), \phi_{*,o}(v)\rangle + \|\phi(o) - \phi(x)\|^2,
\end{aligned}$$

using in the last line that $\|\phi_{*,o}(v)\|^2 = 1$ since $v \in S_o$. Then,

$$f'(t) = 0 \iff t = \langle \phi(x) - \phi(o), \phi_{*,o}(v) \rangle.$$

Therefore,

$$P^v(x) = \operatorname*{argmin}_{t \in \mathbb{R}} \; f(t) = \langle \phi(x) - \phi(o), \phi_{*,o}(v) \rangle.$$

Busemann function. First, following (Bridson and Haefliger, 2013), we have for all $x \in \mathcal{M}$,

$$B^v(x) = \lim_{t \to \infty} \big(d_{\mathcal{M}}(\gamma_v(t), x) - t \big) = \lim_{t \to \infty} \frac{d_{\mathcal{M}}(\gamma_v(t), x)^2 - t^2}{2t},$$

denoting $\gamma_v : t \mapsto \phi^{-1}\big(\phi(o) + t\phi_{*,o}(v)\big)$ the geodesic line associated to $\mathcal{G}^v$. Then, we get

$$\begin{aligned}
\frac{d_{\mathcal{M}}(\gamma_v(t), x)^2 - t^2}{2t} &= \frac{1}{2t}\big(\|\phi(\gamma_v(t))^2 - \phi(x)\|^2 - t^2 \big) \\
&= \frac{1}{2t}\big(\|\phi(o) + t\phi_{*,o}(v) - \phi(x)\|^2 - t^2 \big) \\
&= \frac{1}{2t}\big(t^2 \|\phi_{*,o}(v)\|^2 - 2t\langle \phi_{*,o}(v), \phi(x) - \phi(o) \rangle + \|\phi(x) - \phi(o)\|^2 - t^2 \big) \\
&= -\langle \phi_{*,o}(v), \phi(x) - \phi(o) \rangle + \frac{1}{2t} \|\phi(x) - \phi(o)\|^2,
\end{aligned}$$

using that $\|v\|_o = \|\phi_{*,o}(v)\| = 1$. Then, by passing to the limit $t \to \infty$, we find

$$B^v(x) = -\langle \phi_{*,o}(v), \phi(x) - \phi(o) \rangle.$$

■

C.2 Proof of Proposition 9

We start by giving the proof of the coordinate geodesic projection which we recall in Proposition 42.

Proposition 42 (Coordinate of the geodesic projection on Hyperbolic space)

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

$$P^v(x) = \frac{1}{\sqrt{-K}} \operatorname{arctanh}\left(-\frac{1}{\sqrt{-K}} \frac{\langle x, v \rangle_{\mathbb{L}}}{\langle x, x^0 \rangle_{\mathbb{L}}} \right).$$

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

$$P^{\tilde{v}}(x) = \frac{2}{\sqrt{-K}} \operatorname{arctanh}\big(\sqrt{-K} s(x) \big),$$

where s is defined as

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

First, we will compute in Proposition 43 the geodesic projections.

Proposition 43 (Geodesic projection)

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

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

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 $x \in \mathbb{B}_K^d$ is*

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

where

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

Proof of Proposition 43

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

$$y = \cosh(\sqrt{-K}t)x^0 + \sinh(\sqrt{-K}t)\frac{v}{\sqrt{-K}},$$

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

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

This problem is equivalent with solving

$$\underset{t \in \mathbb{R}}{\text{argmin}} \; K\cosh(\sqrt{-K}t)\langle x, x^0\rangle_{\mathbb{L}} + K\frac{\sinh(\sqrt{-K}t)}{\sqrt{-K}}\langle x, v\rangle_{\mathbb{L}}.$$

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

$$g'(t) = 0 \iff \tanh(\sqrt{-K}t) = -\frac{1}{\sqrt{-K}}\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}. \tag{5}$$

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

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

and

$$\sinh(\sqrt{-K}t) = \frac{\langle x, v\rangle_{\mathbb{L}}}{\sqrt{-K\langle x, x^0\rangle_{\mathbb{L}}^2 - \langle x, v\rangle_{\mathbb{L}}^2}}.$$

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 $t \in]-\frac{1}{\sqrt{-K}}, \frac{1}{\sqrt{-K}}[$. Using that arccosh is an increasing function, we find

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

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

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

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

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

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

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

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

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

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

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

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

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

Thus,

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

■

Proof of Proposition 42

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

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

Hence, by using Equation (5), we obtain that the optimal t satisfies

$$\tanh(\sqrt{-K}t) = -\frac{1}{\sqrt{-K}}\frac{\langle x,v\rangle_{\mathbb{L}}}{\langle x,x^0\rangle_{\mathbb{L}}} \iff t = \frac{1}{\sqrt{-K}}\operatorname{arctanh}\left(-\frac{1}{\sqrt{-K}}\frac{\langle x,v\rangle_{\mathbb{L}}}{\langle x,x^0\rangle_{\mathbb{L}}}\right).$$

2. *Poincaré ball.* As a geodesic is of the form $\gamma(t) = \tanh\left(\frac{\sqrt{-K}t}{2}\right)\frac{p}{\sqrt{-K}}$ for all $t \in \mathbb{R}$, we deduce from Proposition 43 that

$$s(x) = \frac{1}{\sqrt{-K}}\tanh\left(\frac{\sqrt{-K}t}{2}\right) \iff t = \frac{2}{\sqrt{-K}}\operatorname{arctanh}\big(\sqrt{-K}s(x)\big).$$

■

We now derive the closed forms of the horospherical projections which we recall in Proposition 44.

Proposition 44 (Busemann function on Hyperbolic space)

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

$$\forall x \in \mathbb{L}_K^d,\ B^v(x) = \frac{1}{\sqrt{-K}} \log\left(-\sqrt{-K}\left\langle x, \sqrt{-K}x^0 + v\right\rangle_{\mathbb{L}}\right).$$

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

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

Proof of Proposition 44

1. *Lorentz model.* The geodesic in direction v can be characterized by

$$\forall t \in \mathbb{R},\ \gamma_v(t) = \cosh(\sqrt{-K}t)x^0 + \sinh(\sqrt{-K}t)\frac{v}{\sqrt{-K}}.$$

Hence, we have for all $x \in \mathbb{L}_K^d$,

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

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

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

Moreover,

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

Hence,

$$B^v(x) = \frac{1}{\sqrt{-K}}\log\left(K\left(\langle x, x^0\rangle_{\mathbb{L}} + \frac{\langle x, v\rangle_{\mathbb{L}}}{\sqrt{-K}}\right)\right).$$

2. *Poincaré ball.*

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

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

where

$$x(t) = -2K\frac{\|\tanh(\frac{\sqrt{-K}t}{2})\frac{p}{\sqrt{-K}} - x\|_2^2}{(1-\tanh^2(\frac{\sqrt{-K}t}{2}))(1+K\|x\|_2^2)}.$$

Now, on one hand, we have

$$B^p(x) = \lim_{t\to\infty}\ (d_{\mathbb{B}}(\gamma_p(t), x) - t)$$
$$= \lim_{t\to\infty}\ \frac{1}{\sqrt{-K}}\left(\log\left(1+x(t)+\sqrt{x(t)^2+2x(t)}\right) - \sqrt{-K}t\right)$$
$$= \lim_{t\to\infty}\ \frac{1}{\sqrt{-K}}\log\left(e^{-\sqrt{-K}t}(1+x(t)+\sqrt{x(t)^2+2x(t)})\right).$$

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

$$e^{-\sqrt{-K}t}x(t) = -2Ke^{-\sqrt{-K}t}\frac{\|\frac{e^{\sqrt{-K}t}-1}{e^{\sqrt{-K}t}+1}\frac{p}{\sqrt{-K}} - x\|_2^2}{(1-(\frac{e^{\sqrt{-K}t}-1}{e^{\sqrt{-K}t}+1})^2)(1+K\|x\|_2^2)}$$
$$= 2e^{-\sqrt{-K}t}\frac{\|e^{\sqrt{-K}t}p - p - \sqrt{-K}e^{\sqrt{-K}t}x - \sqrt{-K}x\|_2^2}{4e^{\sqrt{-K}t}(1+K\|x\|_2^2)}$$
$$= \frac{1}{2}\frac{\|p - e^{-\sqrt{-K}t}p - \sqrt{-K}x - \sqrt{-K}e^{-\sqrt{-K}t}x\|_2^2}{1+K\|x\|_2^2}$$
$$\underset{t\to\infty}{\to} \frac{1}{2}\frac{\|p-\sqrt{-K}x\|_2^2}{1+K\|x\|_2^2}.$$

Hence,

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

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

■

C.3 Proof of Proposition 11

First, we recall and show two lemmas.

Lemma 45 (Proposition 5.6.c in (Lee, 2006)) *Suppose $\phi : (\mathcal{M}, g) \to (\tilde{\mathcal{M}}, \tilde{g})$ is an isometry. Then, ϕ takes geodesics to geodesics,* i.e. *if γ is the geodesic in $\mathcal{M}$ with $\gamma(0) = p$ and $\gamma'(0) = v$, then $\phi\circ\gamma$ is the geodesic in $\tilde{\mathcal{M}}$ with $\phi(\gamma(0)) = \phi(p)$ and $(\phi\circ\gamma)'(0) = \phi_{*,p}(v)$.*

Lemma 46 *Let $\phi : (\mathcal{M}, g) \to (\tilde{\mathcal{M}}, \tilde{g})$ an isometry and $v \in T_o\mathcal{M}$ such that $\|v\|_o = 1$. Then for all $x \in \mathcal{M}$,*

$$B^v(x) = B^{\phi_{*,o}(v)}\big(\phi(x)\big), \tag{6}$$
$$P^v(x) = P^{\phi_{*,o}(v)}\big(\phi(x)\big). \tag{7}$$

Proof of Lemma 46 Let $v \in T_o\mathcal{M}$ such that $\|v\|_o = 1$, $x \in \mathcal{M}$. By Lemma 45, we have $\phi\big(\exp_o(tv)\big) = \exp_{\phi(o)}\big(t\phi_{*,o}(v)\big)$.
Proof of Equation (6). Let us show that $B^v(x) = B^{\phi_{*,o}(v)}\big(\phi(x)\big)$. By definition of the Busemann function, we have

$$\begin{aligned} B^v(x) &= \lim_{t\to\infty}\ d_{\mathcal{M}}\big(x, \exp_o(tv)\big) - t \\ &= \lim_{t\to\infty}\ d_{\tilde{\mathcal{M}}}\big(\phi(x), \phi(\exp_o(tv))\big) - t \quad \text{since } \phi \text{ is an isometry} \\ &= \lim_{t\to\infty}\ d_{\tilde{\mathcal{M}}}\big(\phi(x), \exp_{\phi(o)}(t\phi_{*,o}(v))\big) - t \\ &= B^{\phi_{*,o}(v)}\big(\phi(x)\big). \end{aligned}$$

Proof of Equation (7). Let us now show that $P^v(x) = P^{\phi_{*,o}(v)}\big(\phi(x)\big)$. Then,

$$\begin{aligned} P^v(x) &= \operatorname*{argmin}_{t\in\mathbb{R}}\ d_{\mathcal{M}}\big(x, \exp_o(tv)\big) \\ &= \operatorname*{argmin}_{t\in\mathbb{R}}\ d_{\tilde{\mathcal{M}}}\big(\phi(x), \phi(\exp_o(tv))\big) \quad \text{since } \phi \text{ is an isometry} \\ &= \operatorname*{argmin}_{t\in\mathbb{R}}\ d_{\tilde{\mathcal{M}}}\big(\phi(x), \exp_{\phi(o)}(t\phi_{*,o}(v))\big) \quad \text{by Lemma 45} \\ &= P^{\phi_{*,o}(v)}\big(\phi(x)\big). \end{aligned}$$

■

Proof of Proposition 11 First, let us show that for λ_o-almost all $v \in S_o$, $W_p^p(B^v_\# \mu, B^v_\# \nu) = W_p^p(B^{\phi_{*,o}(v)}_\# \tilde{\mu}, B^{\phi_{*,o}(v)}_\# \tilde{\nu})$ and $W_p^p(P^v_\# \mu, P^v_\# \nu) = W_p^p(P^{\phi_{*,o}(v)}_\# \mu, P^{\phi_{*,o}(v)}_\# \nu)$. Using Lemma 46, we have

$$W_p^p(B^v_\# \mu, B^v_\# \nu) = W_p^p(B^{\phi_{*,o}(v)}_\# \phi_\# \mu, B^{\phi_{*,o}(v)}_\# \phi_\# \nu) = W_p^p(B^{\phi_{*,o}(v)}_\# \tilde{\mu}, B^{\phi_{*,o}(v)}_\# \tilde{\nu}),$$
$$W_p^p(P^v_\# \mu, P^v_\# \nu) = W_p^p(P^{\phi_{*,o}(v)}_\# \phi_\# \mu, P^{\phi_{*,o}(v)}_\# \phi_\# \nu) = W_p^p(P^{\phi_{*,o}(v)}_\# \tilde{\mu}, P^{\phi_{*,o}(v)}_\# \tilde{\nu}).$$

These results are true for all $v \in S_o$, and therefore for λ_o-almost all $v \in S_o$. Thus, by integrating with respect to λ_o, and performing the change of $v \mapsto \phi_{*,o}(v)$ on the right side, we find

$$\mathrm{HCHSW}_p^p(\mu, \nu; \lambda_o) = \mathrm{HCHSW}_p^p(\tilde{\mu}, \tilde{\nu}; (\phi_{*,o})_\# \lambda_o),$$
$$\mathrm{GCHSW}_p^p(\mu, \nu; \lambda_o) = \mathrm{GCHSW}_p^p(\tilde{\mu}, \tilde{\nu}; (\phi_{*,o})_\# \lambda_o).$$

Finally, we can conclude by using that $\phi_{*,o}$ is an isometry between the tangent spaces and hence $(\phi_{*,o})_\# \lambda_o = \lambda_{\phi(o)}$. ■

C.4 Proof of Proposition 15

First, let us compute the differential of $\phi = \varphi \circ \mathcal{L}$. In that purpose, we first recall the differential of $\mathcal{L} : X = LL^T \mapsto L$ derived in (Lin, 2019, Proposition 4).

Lemma 47 (Proposition 4 in (Lin, 2019)) *Let $X \in S_d^{++}(\mathbb{R})$ and $V \in S_d(\mathbb{R})$. The differential operator $\mathcal{L}_{*,X} : T_X S_d^{++}(\mathbb{R}) \to T_{\mathcal{L}(X)} L_d^{++}(\mathbb{R})$ of $\mathcal{L}$ at X is given by*

$$\mathcal{L}_{*,X}(V) = \mathcal{L}(X)\Big(\lfloor \mathcal{L}(X)^{-1} V \mathcal{L}(X)^{-T} \rfloor + \frac{1}{2}\mathrm{diag}\big(\mathcal{L}(X)^{-1} V \mathcal{L}(X)^{-T}\big)\Big).$$

Lemma 48 *Let $\phi : X \mapsto \varphi\big(\mathcal{L}(X)\big)$ and $X = LL^T \in S_d^{++}(\mathbb{R})$ with $L \in L_d^{++}(\mathbb{R})$ obtained by the Cholesky decomposition. The differential operator of ϕ at X is given by*

$$\forall V \in T_X S_d^{++}(\mathbb{R}),\ \phi_{*,X}(V) = \lfloor \mathcal{L}_{*,X}(V) \rfloor + \mathrm{diag}\big(\mathcal{L}(X)\big)^{-1} \mathrm{diag}\big(\mathcal{L}_{*,X}(V)\big),$$

where

$$\mathcal{L}_{*,X}(V) = \mathcal{L}(X)\Big(\lfloor \mathcal{L}(X)^{-1} V \mathcal{L}(X)^{-T} \rfloor + \frac{1}{2}\mathrm{diag}\big(\mathcal{L}(X)^{-1} V \mathcal{L}(X)^{-T}\big)\Big).$$

Proof of Lemma 48 Using the chain rule, we have, for $X \in S_d^{++}(\mathbb{R})$ and $V \in S_d(\mathbb{R})$,

$$\begin{aligned}\phi_{*,X}(V) &= \varphi_{*,X}\big(\mathcal{L}_{*,X}(V)\big) = \lfloor \mathcal{L}_{*,X}(V) \rfloor + \log_{*,\mathrm{diag}(L)}\big(\mathrm{diag}(\mathcal{L}_{*,X}(V))\big) \\ &= \lfloor \mathcal{L}_{*,X}(V) \rfloor + \Sigma\big(\mathrm{diag}(\mathcal{L}_{*,X}(V))\big),\end{aligned}$$

using Lemma 12 for the differential of the log with

$$\begin{aligned}\Sigma\big(\mathrm{diag}(\mathcal{L}_{*,X}(V))\big) &= \mathrm{diag}(\mathcal{L}_{*,X}(V)) \odot \Gamma \\ &= \mathrm{diag}(\mathcal{L}_{*,X}(V)) \oslash \mathrm{diag}(\mathcal{L}(X)) \\ &= \mathrm{diag}(\mathcal{L}(X))^{-1} \mathrm{diag}(\mathcal{L}_{*,X}(V)).\end{aligned}$$

Thus, we conclude that $\phi_{*,X}(V) = \lfloor \mathcal{L}_{*,X}(V) \rfloor + \mathrm{diag}\big(\mathcal{L}(X)\big)^{-1}\mathrm{diag}\big(\mathcal{L}_{*,X}(V)\big)$. ■

Proof of Proposition 15 On one hand we have $\phi(I_d) = 0$, $\mathcal{L}_{*,I_d}(V) = \lfloor V \rfloor + \frac{1}{2}\mathrm{diag}(V)$ and thus $\phi_{*,I_d}(V) = \lfloor V \rfloor + \frac{1}{2}\mathrm{diag}(V)$ since $\mathcal{L}(I_d) = I_d$. Thus, using Proposition 7, the projection is given, for $A \in S_d(\mathbb{R})$ such that $\|A\|^2_{I_d} = \langle \phi_{*,I_d}(A), \phi_{*,I_d}(A)\rangle_F = 1$, by

$$\begin{aligned}\forall X = LL^T \in S_d^{++}(\mathbb{R}),\ P^A(X) &= \langle \phi(X), \phi_{*,I_d}(A)\rangle_F \\ &= \langle \lfloor L \rfloor + \log(\mathrm{diag}(L)), \lfloor A \rfloor + \frac{1}{2}\mathrm{diag}(A)\rangle_F \\ &= \langle \lfloor L \rfloor, \lfloor A \rfloor\rangle_F + \langle \log(\mathrm{diag}(L)), \frac{1}{2}\mathrm{diag}(A)\rangle_F.\end{aligned}$$

■

C.5 Proof of Proposition 16

Proof of Proposition 16 We use that $B^\gamma(x) = \lim_{t\to\infty} \frac{d(x,\gamma(t))^2 - t^2}{2t}$ (see *e.g.* (Bridson and Haefliger, 2013, II. 8.24)). Thus,

$$\begin{aligned}B^\gamma(x) &= \lim_{t\to\infty}\ d(x,\gamma(t)) - t \\ &= \lim_{t\to\infty}\ \frac{d(x,\gamma(t))^2 - t^2}{2t} \\ &= \lim_{t\to\infty}\ \sum_{i=1}^n \lambda_i \frac{d_i(x_i,\gamma_i(\lambda_i t))^2 - \lambda_i^2 t^2}{2\lambda_i t} \\ &= \sum_{i=1}^n \lambda_i B^{\gamma_i}(x_i).\end{aligned}$$

■

Appendix D. Proofs of Section 5

D.1 Proof of Proposition 17

Proof of Proposition 17 First, we will show that for any $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$, $\mathrm{CHSW}_p(\mu,\nu) < \infty$. Let $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$, and let $\gamma \in \Pi(\mu,\nu)$ be an arbitrary coupling between them. Then by using first Lemma 37 followed by the 1-Lipschitzness of the projections Lemma 38 and

Lemma 39, we obtain

$$\begin{aligned} W_p^p(P^v_\# \mu, P^v_\# \nu) &= \inf_{\gamma \in \Pi(\mu,\nu)} \int |P^v(x) - P^v(y)|^p \, \mathrm{d}\gamma(x,y) \\ &\leq \int |P^v(x) - P^v(y)|^p \, \mathrm{d}\gamma(x,y) \\ &\leq \int d(x,y)^p \, \mathrm{d}\gamma(x,y) \\ &\leq 2^{p-1} \left(\int d(x,o)^p \, \mathrm{d}\mu(x) + \int d(o,y)^p \, \mathrm{d}\nu(y) \right) \\ &< \infty. \end{aligned}$$

Hence, we can conclude that $\mathrm{CHSW}_p^p(\mu,\nu) < \infty$.

Now, let us show that it is a pseudo-distance. First, it is straightforward to see that $\mathrm{CHSW}_p(\mu,\nu) \geq 0$, that it is symmetric, *i.e.* $\mathrm{CHSW}_p(\mu,\nu) = \mathrm{CHSW}_p(\nu,\mu)$, and that $\mu = \nu$ implies that $\mathrm{CHSW}_p(\mu,\nu) = 0$ using that W_p is well a distance.

For the triangular inequality, we can derive it using the triangular inequality for W_p and the Minkowski inequality. Let $\mu, \nu, \alpha \in \mathcal{P}_p(\mathcal{M})$,

$$\begin{aligned} \mathrm{CHSW}_p(\mu,\nu) &= \left(\int_{S_o} W_p^p(P^v_\# \mu, P^v_\# \nu) \, \mathrm{d}\lambda_o(v) \right)^{\frac{1}{p}} \\ &\leq \left(\int_{S_o} \left(W_p(P^v_\# \mu, P^v_\# \alpha) + W_p(P^v_\# \alpha, P^v_\# \nu) \right)^p \, \mathrm{d}\lambda_o(v) \right)^{\frac{1}{p}} \\ &\leq \left(\int_{S_o} W_p^p(P^v_\# \mu, P^v_\# \alpha) \, \mathrm{d}\lambda_o(v) \right)^{\frac{1}{p}} + \left(\int_{S_o} W_p^p(P^v_\# \alpha, P^v_\# \nu) \, \mathrm{d}\lambda_o(v) \right)^{\frac{1}{p}} \\ &= \mathrm{CHSW}_p(\mu,\alpha) + \mathrm{CHSW}_p(\alpha,\nu). \end{aligned}$$

■

D.2 Proof of Proposition 18

Proof of Proposition 18 Let $f \in L^1(\mathcal{M})$, $g \in C_0(\mathbb{R} \times S_o)$, then by Fubini's theorem,

$$\begin{aligned} \langle \mathrm{CHR} f, g \rangle_{\mathbb{R} \times S_o} &= \int_{S_o} \int_{\mathbb{R}} \mathrm{CHR} f(t,v) g(t,v) \, \mathrm{d}t \mathrm{d}\lambda_o(v) \\ &= \int_{S_o} \int_{\mathbb{R}} \int_{\mathcal{M}} f(x) \mathbb{1}_{\{t = P^v(x)\}} g(t,v) \, \mathrm{dVol}(x) \mathrm{d}t \mathrm{d}\lambda_o(v) \\ &= \int_{\mathcal{M}} f(x) \int_{S_o} \int_{\mathbb{R}} g(t,v) \mathbb{1}_{\{t = P^v(x)\}} \, \mathrm{d}t \mathrm{d}\lambda_o(v) \mathrm{dVol}(x) \\ &= \int_{\mathcal{M}} f(x) \int_{S_o} g\big(P^v(x), v\big) \, \mathrm{d}\lambda_o(v) \mathrm{dVol}(x) \\ &= \int_{\mathcal{M}} f(x) \mathrm{CHR}^* g(x) \, \mathrm{dVol}(x) \\ &= \langle f, \mathrm{CHR}^* g \rangle_{\mathcal{M}}. \end{aligned}$$

■

D.3 Proof of Proposition 19

Proof of Proposition 19 We follow the proof of (Boman and Lindskog, 2009, Lemma 1). On one hand, $g \in C_0(\mathbb{R} \times S_o)$, thus for all $\epsilon > 0$, there exists $M > 0$ such that $|t| \geq M$ implies $|g(t, v)| \leq \epsilon$ for all $v \in S_o$.

Let $\epsilon > 0$ and $M > 0$ which satisfies the previous property. Denote $E(x, M) = \{v \in S_o,\ |P^v(x)| < M\}$. Then, as $d(x, o) > 0$, we have

$$E(x, M) = \{v \in S_o,\ |P^v(x)| < M\} = \left\{v \in S_o,\ \frac{P^v(x)}{d(x,o)} < \frac{M}{d(x,o)}\right\} \xrightarrow[d(x,o)\to\infty]{} \emptyset.$$

Thus, $\lambda_o\big(E(x, M)\big) \xrightarrow[d(x,o)\to\infty]{} 0$. Choose M' such that $d(x, o) > M'$ implies that

$$\lambda_o\big(E(x, M)\big) < \epsilon.$$

Then, for $x \in \mathcal{M}$ such that $|P^v(x)| \geq \max(M, M')$ (and thus $d(x, o) \geq M'$ since $|P^v(x)| \leq d(x, o)$ as P^v is Lipschitz,

$$\begin{aligned}
|\mathrm{CHR}^* g(x)| &\leq \left|\int_{E(x,M)} g(P^v(x), v)\ \mathrm{d}\lambda_o(v)\right| + \left|\int_{E(x,M)^c} g(P^v(x), v)\ \mathrm{d}\lambda_o(v)\right| \\
&\leq \|g\|_\infty\ \lambda_o\big(E(x, M)\big) + \epsilon\lambda_o\big(E(x, M)^c\big) \\
&\leq \|g\|_\infty \epsilon + \epsilon.
\end{aligned}$$

Thus, we showed that $\mathrm{CHR}^* g(x) \xrightarrow[d(x,o)\to\infty]{} 0$, and thus $\mathrm{CHR}^* g \in C_0(\mathcal{M})$. ■

D.4 Proof of Proposition 21

Proof of Proposition 21 Let $g \in C_0(\mathbb{R} \times S_o)$, as $\mathrm{CHR}\mu = \lambda_o \otimes K_\mu$, we have by definition

$$\int_{S_o}\int_{\mathbb{R}} g(t, v)\ K_\mu(v, \mathrm{d}t)\ \mathrm{d}\lambda_o(v) = \int_{\mathbb{R}\times S_o} g(t, v)\ \mathrm{d}(\mathrm{CHR}\mu)(t, v).$$

Hence, using the property of the dual, we have for all $g \in C_o(\mathbb{R} \times S_o)$,

$$\begin{aligned}
\int_{S_o}\int_{\mathbb{R}} g(t, v)\ K_\mu(v, \mathrm{d}t)\ \mathrm{d}\lambda_o(v) &= \int_{\mathbb{R}\times S_o} g(t, v)\ \mathrm{d}(\mathrm{CHR}\mu)(t, v) \\
&= \int_{\mathcal{M}} \mathrm{CHR}^* g(x)\ \mathrm{d}\mu(x) \\
&= \int_{\mathcal{M}}\int_{S_o} g(P^v(x), v)\ \mathrm{d}\lambda_o(v)\mathrm{d}\mu(x) \\
&= \int_{S_o}\int_{\mathcal{M}} g(P^v(x), v)\ \mathrm{d}\mu(x)\mathrm{d}\lambda_o(v) \\
&= \int_{S_o}\int_{\mathbb{R}} g(t, v)\ \mathrm{d}(P^v_\#\mu)(t)\mathrm{d}\lambda_o(v).
\end{aligned}$$

Hence, for λ_o-almost every $v \in S_o$, $K_\mu(v, \cdot) = P^v_\# \mu$. ■

D.5 Proof of Proposition 22

Proof of Proposition 22 Using Lemma 37 and that the projections are 1-Lipschitz (Lemma 38), we can show that, for any $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$,

$$\mathrm{CHSW}_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int |P^v(x) - P^v(y)|^p \, \mathrm{d}\gamma(x, y).$$

Let $\gamma^* \in \Pi(\mu, \nu)$ being an optimal coupling for the Wasserstein distance with ground cost d, then,

$$\begin{aligned}
\mathrm{CHSW}_p^p(\mu, \nu) &\leq \int |P^v(x) - P^v(y)|^p \, \mathrm{d}\gamma^*(x, y) \\
&\leq \int d(x, y)^p \, \mathrm{d}\gamma^*(x, y) \\
&= W_p^p(\mu, \nu).
\end{aligned}$$

■

D.6 Proof of Proposition 23

Proof of Proposition 23 Let $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$, then

$$\begin{aligned}
\mathrm{CHSW}_p^p(\mu, \nu) &= \int_{S_o} W_p^p(P^v_\# \mu, P^v_\# \nu) \, \mathrm{d}\lambda_o(v) \\
&= \int_{S_o} \|F^{-1}_{P^v_\# \mu} - F^{-1}_{P^v_\# \nu}\|^p_{L^p([0,1])} \, \mathrm{d}\lambda_o(v) \\
&= \int_{S_o} \int_0^1 \left(F^{-1}_{P^v_\# \mu}(q) - F^{-1}_{P^v_\# \nu}(q)\right)^p \, \mathrm{d}q \, \mathrm{d}\lambda_o(v) \\
&= \|\Phi(\mu) - \Phi(\nu)\|^p_{\mathcal{H}}.
\end{aligned}$$

Thus, CHSW_p is Hilbertian. ■

D.7 Proof of Lemma 25

Proof of Lemma 25

Since for any $v \in S_o$ and $x \in \mathcal{M}$, $P^v(x) = \langle \phi(x) - \phi(o), \phi_{*,o}(x) \rangle$, by using Lemma 37 we have

$$\begin{aligned}
W_p^p(P^v_\# \mu, P^v_\# \nu) &= \inf_{\gamma \in \Pi(P^v_\# \mu, P^v_\# \nu)} \int |x - y|^p \, \mathrm{d}\gamma(x, y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int |P^v(x) - P^v(y)|^p \, \mathrm{d}\gamma(x, y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} \int |\langle \phi(x) - \phi(y), \phi_{*,o}(v) \rangle|^p \, \mathrm{d}\gamma(x, y).
\end{aligned}$$

Let's note $Q^v(x) = \langle x, v \rangle$. Then, we obtain

$$\begin{aligned} W_p^p(P^v_{\#}\mu, P^v_{\#}\nu) &= \inf_{\gamma \in \Pi(\mu,\nu)} \int |Q^{\phi_{*,o}(v)}(\phi(x)) - Q^{\phi_{*,o}(v)}(\phi(y))|^p \, \mathrm{d}\gamma(x,y) \\ &= W_p^p(Q^{\phi_{*,o}(v)}_{\#}\phi_{\#}\mu, Q^{\phi_{*,o}(v)}_{\#}\phi_{\#}\nu). \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} \mathrm{CHSW}_p^p(\mu,\nu) &= \int_{S_o} W_p^p(Q^{\phi_{*,o}(v)}_{\#}\phi_{\#}\mu, Q^{\phi_{*,o}(v)}_{\#}\phi_{\#}\nu) \, \mathrm{d}\lambda_o(v) \\ &= \int_{S_{\phi(o)}} W_p^p(Q^v_{\#}\phi_{\#}\mu, Q^v_{\#}\phi_{\#}\nu) \, \mathrm{d}\big((\phi_{*,o})_{\#}\lambda_o\big)(v). \end{aligned}$$

Finally, since $\phi_{*,o}$ is an isometry between the tangent spaces by definition of the metric, we have $(\phi_{*,o})_{\#}\lambda_o = \lambda_{\phi(o)}$. ■

D.8 Proof of Proposition 26

Proof of Proposition 26

We know by Proposition 17 that CHSW_p is a finite pseudo-distance. For the indiscernible property, using Lemma 25 and the distance property of SW_p, we have that $\mathrm{CHSW}_p(\mu,\nu) = \mathrm{SW}_p^p(\phi_{\#}\mu, \phi_{\#}\nu) = 0$ implies that $\phi_{\#}\mu = \phi_{\#}\nu$ by applying the same proof of (Bonnotte, 2013, Proposition 5.1.2). Indeed, we have that $\mathrm{SW}_p^p(\phi_{\#}\mu, \phi_{\#}\nu) = 0$ implies $W_p^p(Q^v_{\#}\phi_{\#}\mu, Q^v_{\#}\phi_{\#}\nu) = 0$ for $\lambda_{\phi(o)}$-almost every $v \in S_{\phi(o)}$, and thus that $Q^v_{\#}\phi_{\#}\mu = Q^v_{\#}\phi_{\#}\nu$ since W_p is a distance. Hence, using the Fourier transform and that $\lambda_{\phi(o)}$ is absolutely continuous with respect to the Lebesgue measure, we obtain that $\phi_{\#}\mu = \phi_{\#}\nu$.

Then, as ϕ is a bijection from $\mathcal{M}$ to $\mathcal{N}$, we have for all Borelian $C \subset \mathcal{M}$,

$$\begin{aligned} \mu(C) &= \int_{\mathcal{M}} \mathbb{1}_C(x) \, \mathrm{d}\mu(x) \\ &= \int_{\mathcal{N}} \mathbb{1}_C\big(\phi^{-1}(y)\big) \, \mathrm{d}(\phi_{\#}\mu)(y) \\ &= \int_{\mathcal{N}} \mathbb{1}_C\big(\phi^{-1}(y)\big) \, \mathrm{d}(\phi_{\#}\nu)(y) \\ &= \int_{\mathcal{M}} \mathbb{1}_C(x) \, \mathrm{d}\nu(x) \\ &= \nu(C). \end{aligned}$$

■

D.9 Proof of Proposition 27

To prove Proposition 27, we will adapt the proof of Nadjahi et al. (2020) to our projection. First, we start to adapt Nadjahi et al. (2020, Lemma S1):

Lemma 49 (Lemma S1 in Nadjahi et al. (2020)) *Let $(\mu_k)_k \in \mathcal{P}_p(\mathcal{M})$ and $\mu \in \mathcal{P}_p(\mathcal{M})$ such that $\lim_{k\to\infty}$ $\mathrm{CHSW}_1(\mu_k, \mu) = 0$. Then, there exists $\varphi : \mathbb{N} \to \mathbb{N}$ non decreasing such that $\mu_{\varphi(k)} \xrightarrow[k\to\infty]{\mathcal{L}} \mu$.*

Proof of Theorem 49

Using Lemma 25, we know that $\mathrm{CHSW}_1(\mu, \nu) = \mathrm{SW}_1(\phi_\# \mu, \phi_\# \nu)$. Let's note $\alpha_k = \phi_\# \mu_k \in \mathcal{P}_p(\mathcal{N})$ and $\alpha = \phi_\# \mu \in \mathcal{P}_p(\mathcal{N})$ and $Q^v(x) = \langle v, x \rangle$.

Then, by Bogachev and Ruas (2007, Theorem 2.2.5),

$$\lim_{k\to\infty} \int_{S_{\phi(o)}} W_1(Q^v_\# \alpha_k, Q^v_\# \alpha) \, \mathrm{d}\lambda_{\phi(o)}(v) = 0$$

implies that there exists a subsequence $(\mu_{\varphi(k)})_k$ such that for $\lambda_{\phi(o)}$-almost every v,

$$W_1(Q^v_\# \alpha_{\varphi(k)}, Q^v_\# \alpha) \xrightarrow[k\to\infty]{} 0.$$

As the Wasserstein distance metrizes the weak convergence, this is equivalent to

$$Q^v_\# \mu_{\varphi(k)} \xrightarrow[k\to\infty]{\mathcal{L}} Q^v_\# \mu.$$

Then, by Levy's characterization theorem, this is equivalent with the pointwise convergence of the characterization function, *i.e.* for all $t \in \mathbb{R}$, $\Phi_{Q^v_\# \alpha_{\varphi(k)}}(t) \xrightarrow[k\to\infty]{} \Phi_{Q^v_\# \mu}(t)$. Then, working in $T_{\phi(o)}\mathcal{N}$ with the Euclidean norm, we can use the same proof of Nadjahi et al. (2020) by using a convolution with a gaussian kernel and show that it implies that $\alpha_{\varphi(k)} \xrightarrow[k\to\infty]{\mathcal{L}} \alpha$, *i.e.* $\phi_\# \mu_{\varphi(k)} \xrightarrow[k\to\infty]{\mathcal{L}} \phi_\# \mu$.

Finally, let's show that it implies the weak convergence of $(\mu_{\varphi(k)})_k$ towards μ. Let $f \in C_b(\mathcal{M})$, then

$$\int_{\mathcal{M}} f \, \mathrm{d}\mu_{\varphi(k)} = \int_{\mathcal{N}} f \circ \phi^{-1} \, \mathrm{d}(\phi_\# \mu_{\varphi(k)}) \xrightarrow[k\to\infty]{} \int_{\mathcal{N}} f \circ \phi^{-1} \, \mathrm{d}(\phi_\# \mu) = \int_{\mathcal{M}} f \, \mathrm{d}\mu.$$

Hence, we an conclude that $\mu_{\varphi(k)} \xrightarrow[k\to\infty]{\mathcal{L}} \mu$. ■

Proof of Proposition 27 First, we suppose that $\mu_k \xrightarrow[k\to\infty]{\mathcal{L}} \mu$ in $\mathcal{P}_p(\mathcal{M})$. Then, by continuity, we have that for λ_o almost every $v \in T_o\mathcal{M}$, $P^v_\# \mu_k \xrightarrow[k\to\infty]{} P^v_\# \mu$. Moreover, as the Wasserstein distance on $\mathbb{R}$ metrizes the weak convergence, $W_p(P^v_\# \mu_k, P^v_\# \mu) \xrightarrow[k\to\infty]{} 0$. Finally, as W_p is bounded and it converges for λ_o-almost every v, we have by the Lebesgue convergence dominated theorem that $\mathrm{CHSW}_p^p(\mu_k, \mu) \xrightarrow[k\to\infty]{} 0$.

For the opposite side, suppose that $\mathrm{CHSW}_p(\mu_k, \mu) \xrightarrow[k\to\infty]{} 0$. Then, since we generalized (Nadjahi et al., 2020, Lemma S1) to our setting in Theorem 49, we can use the same contradiction argument as Nadjahi et al. (2020) and we conclude that $(\mu_k)_k$ converges weakly to μ.

■

D.10 Proof of Proposition 28

Proof of Proposition 28 By using Lemma 37, let us first observe that

$$\begin{aligned}
W_1(\mu,\nu) &= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d_{\mathcal{M}}(x,y)\ \mathrm{d}\gamma(x,y) \\
&= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} \|\phi(x)-\phi(y)\|\ \mathrm{d}\gamma(x,y) \\
&= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{N}\times\mathcal{N}} \|x-y\|\ \mathrm{d}(\phi\otimes\phi)_{\#}\gamma(x,y) \\
&= \inf_{\gamma\in\Pi(\phi_{\#}\mu,\phi_{\#}\nu)} \int_{\mathcal{N}\times\mathcal{N}} \|x-y\|\ \mathrm{d}\gamma(x,y) \\
&= W_1(\phi_{\#}\mu,\phi_{\#}\nu).
\end{aligned}$$

Here, we note that W_1 must be understood with respect to the ground cost metric which makes sense given the space, *i.e.* $d_{\mathcal{M}}$ on $\mathcal{M}$ and $\|\cdot-\cdot\|$ on $\mathcal{N}$.

Then, using Lemma 25, we have

$$\mathrm{CHSW}_1(\mu,\nu) = \mathrm{SW}_1(\phi_{\#}\mu,\phi_{\#}\nu).$$

Since $\mathcal{N}$ is a Euclidean inner product space of dimension d, we can apply (Bonnotte, 2013, Lemma 5.14), and we obtain

$$W_1(\mu,\nu) = W_1(\phi_{\#}\mu,\phi_{\#}\nu) \leq C_{d,p,r}\mathrm{SW}_1(\phi_{\#}\mu,\phi_{\#}\nu)^{\frac{1}{d+1}} = C_{d,p,r}\mathrm{CHSW}_1(\mu,\nu)^{\frac{1}{d+1}}.$$

Then, using that $W_p^p(\mu,\nu) \leq (2r)^{p-1}W_1(\mu,\nu)$ and that by the Hölder inequality,

$$\mathrm{CHSW}_1(\mu,\nu) \leq \mathrm{CHSW}_p(\mu,\nu),$$

we obtain (with a different constant $C_{d,r,p}$)

$$W_p^p(\mu,\nu) \leq C_{d,r,p}\mathrm{CHSW}_p(\mu,\nu)^{\frac{1}{d+1}}.$$

■

D.11 Proof of Proposition 29

Proof of Proposition 29 First, using the triangular inequality, the reverse triangular inequality and the Jensen inequality for $x \mapsto x^{1/p}$ (which is concave since $p \geq 1$), we have the following inequality

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

Moreover, by Fubini-Tonelli,

$$\begin{aligned}\mathbb{E}[\mathrm{CHSW}_p^p(\hat{\mu}_n, \mu)] &= \mathbb{E}\left[\int_{S_o} W_p^p(P^v_{\#}\hat{\mu}_n, \mu)\ \mathrm{d}\lambda_o(v)\right] \\ &= \int_{S_o} \mathbb{E}[W_p^p(P^v_{\#}\hat{\mu}_n, P^v_{\#}\mu)]\ \mathrm{d}\lambda_o(v).\end{aligned}$$

Then, by applying Theorem 40, we get that for $q > p$, there exists a constant $C_{p,q}$ such that,

$$\begin{aligned}&\mathbb{E}[W_p^p(P^v_{\#}\hat{\mu}_n, P^v_{\#}\nu)] \\ &\qquad \leq C_{p,q}\tilde{M}_q(P^v_{\#}\mu)^{p/q}\left(n^{-1/2}\mathbb{1}_{\{q>2p\}} + n^{-1/2}\log(n)\mathbb{1}_{\{q=2p\}} + n^{-(q-p)/q}\mathbb{1}_{\{q\in(p,2p)\}}\right).\end{aligned}$$

Then, noting that necessarily, $P^v(o) = 0$ (for both the horospherical and geodesic projection, since the geodesic is of the form $\exp_o(tv)$), and using that P^v is 1-Lipschitz Lemma 38, we can bound the moments as

$$\begin{aligned}\tilde{M}_q(P^v_{\#}\mu) &= \int_{\mathbb{R}} |x|^q\ \mathrm{d}(P^v_{\#}\mu)(x) \\ &= \int_{\mathcal{M}} |P^v(x)|^q\ \mathrm{d}\mu(x) \\ &= \int_{\mathcal{M}} |P^v(x) - P^v(o)|^q\ \mathrm{d}\mu(x) \\ &\leq \int_{\mathcal{M}} d(x,o)^q\ \mathrm{d}\mu(x) \\ &= M_q(\mu).\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}[\mathrm{CHSW}_p^p(\hat{\mu}_n, \mu)] \\ &\qquad \leq C_{p,q}M_q(\mu)^{p/q}\left(n^{-1/2}\mathbb{1}_{\{q>2p\}} + n^{-1/2}\log(n)\mathbb{1}_{\{q=2p\}} + n^{-(q-p)/q}\mathbb{1}_{\{q\in(p,2p)\}}\right),\end{aligned}$$

and similarly,

$$\begin{aligned}&\mathbb{E}[\mathrm{CHSW}_p^p(\hat{\nu}_n, \nu)] \\ &\qquad \leq C_{p,q}M_q(\nu)^{p/q}\left(n^{-1/2}\mathbb{1}_{\{q>2p\}} + n^{-1/2}\log(n)\mathbb{1}_{\{q=2p\}} + n^{-(q-p)/q}\mathbb{1}_{\{q\in(p,2p)\}}\right).\end{aligned}$$

Hence, we conclude that

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

■

D.12 Proof of Proposition 30

Proof of Proposition 30 Let $(v_\ell)_{\ell=1}^L$ be iid samples of λ_o. Then, by first using Jensen inequality and then remembering that $\mathbb{E}_v[W_p^p(P^v_\#\mu, P^v_\#\nu)] = \mathrm{CHSW}_p^p(\mu,\nu)$, we have

$$\begin{aligned}
\mathbb{E}_v\left[\left|\widehat{\mathrm{CHSW}}^p_{p,L}(\mu,\nu) - \mathrm{CHSW}_p^p(\mu,\nu)\right|\right]^2 &\leq \mathbb{E}_v\left[\left|\widehat{\mathrm{CHSW}}^p_{p,L}(\mu,\nu) - \mathrm{CHSW}_p^p(\mu,\nu)\right|^2\right] \\
&= \mathbb{E}_v\left[\left|\frac{1}{L}\sum_{\ell=1}^L \left(W_p^p(P^{v_\ell}_\#\mu, P^{v_\ell}_\#\nu) - \mathrm{CHSW}_p^p(\mu,\nu)\right)\right|^2\right] \\
&= \frac{1}{L^2}\mathrm{Var}_v\left(\sum_{\ell=1}^L W_p^p(P^{v_\ell}_\#\mu, P^{v_\ell}_\#\nu)\right) \\
&= \frac{1}{L}\mathrm{Var}_v\left(W_p^p(P^{v}_\#\mu, P^{v}_\#\nu)\right) \\
&= \frac{1}{L}\int_{S_o}\left(W_p^p(P^{v}_\#\mu, P^{v}_\#\nu) - \mathrm{CHSW}_p^p(\mu,\nu)\right)^2 \,\mathrm{d}\lambda_o(v).
\end{aligned}$$

■

Appendix E. Proofs of Section 7

E.1 Proof of Proposition 31

Proof of Proposition 31 This proof follows the proof in the Euclidean case derived in (Bonnotte, 2013, Proposition 5.1.7) or in (Candau-Tilh, 2020, Proposition 1.33).

As μ is absolutely continuous, $P^v_\#\mu$ is also absolutely continuous and there is a Kantorovitch potential ψ_v between $P^v_\#\mu$ and $P^v_\#\nu$. Moreover, as the support is restricted to a compact, it is Lipschitz and thus differentiable almost everywhere.

First, using the duality formula, we obtain the following lower bound for all $\epsilon > 0$,

$$\frac{\mathrm{CHSW}_2^2\big((T_\epsilon)_\#\mu,\nu\big) - \mathrm{CHSW}_2^2(\mu,\nu)}{2\epsilon} \geq \int_{S_o}\int_{\mathcal{M}} \frac{\psi_v(P^v(T_\epsilon(x))) - \psi_v(P^v(x))}{\epsilon}\,\mathrm{d}\mu(x)\mathrm{d}\lambda_o(v).$$

Then, we know that the exponential map satisfies $\exp_x(0) = x$ and $\frac{\mathrm{d}}{\mathrm{d}t}\exp(tv)|_{t=0} = v$. Taking the limit $\epsilon \to 0$, the right term is equal to $\frac{\mathrm{d}}{\mathrm{d}t}g(t)|_{t=0}$ with $g(t) = \psi_v(P^v(T_t(x)))$ and is equal to

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t)|_{t=0} = \psi_v'(P^v(T_0(x)))\langle \nabla P^v(T_0(x)), \frac{\mathrm{d}}{\mathrm{d}t}T_t(x)|_{t=0}\rangle_x = \psi_v'(P^v(x))\langle \mathrm{grad}_{\mathcal{M}}P^v(x), \xi(x)\rangle_x.$$

Therefore, by the Lebesgue dominated convergence theorem (we have the convergence λ_o-almost surely and $|\psi_v(P^v(T_\epsilon(x))) - \psi_v(P^v(x))| \leq \epsilon$ using that ψ_v and P^v are Lipschitz and that $d\big(\exp_x(\epsilon\xi(x)), \exp_x(0)\big) \leq C\epsilon$),

$$\begin{aligned}
&\liminf_{\epsilon\to 0^+} \frac{\mathrm{CHSW}_2^2\big((T_\epsilon)_\#\mu,\nu\big) - \mathrm{CHSW}_2^2(\mu,\nu)}{2\epsilon} \\
&\geq \int_{S_o}\int_{\mathcal{M}} \psi_v'(P^v(x))\langle \mathrm{grad}_{\mathcal{M}}P^v(x), \xi(x)\rangle\,\mathrm{d}\mu(x)\mathrm{d}\lambda_o(v).
\end{aligned}$$

For the upper bound, first, let $\pi^v \in \Pi(\mu,\nu)$ a coupling such that $\tilde{\pi}^v = (P^v \otimes P^v)_\# \pi^v \in \Pi(P^v_\#\mu, P^v_\#\nu)$ is an optimal coupling for the regular quadratic cost. For $\tilde{\pi}^v$-almost every (x,y), $y = x - \psi_v'(x)$ and thus for π^v-almost every (x,y), $P^v(y) = P^v(x) - \psi_v'\big(P^v(x)\big)$. Therefore,

$$\begin{aligned}\mathrm{CHSW}_2^2(\mu,\nu) &= \int_{S_o} W_2^2(P^v_\#\mu, P^v_\#\nu)\ \mathrm{d}\lambda_o(v)\\ &= \int_{S_o}\int_{\mathbb{R}\times\mathbb{R}} |x-y|^2\ \mathrm{d}\tilde{\pi}^v(x,y)\ \mathrm{d}\lambda_o(v)\\ &= \int_{S_o}\int_{\mathcal{M}\times\mathcal{M}} |P^v(x)-P^v(y)|^2\ \mathrm{d}\pi^v(x,y)\ \mathrm{d}\lambda_o(v).\end{aligned}$$

On the other hand, $((P^v\circ T_\epsilon)\otimes P^v)_\#\pi^v \in \Pi(P^v_\#(T_\epsilon)_\#\mu, P^v_\#\nu)$ and hence

$$\begin{aligned}\mathrm{CHSW}_2^2\big((T_\epsilon)_\#\mu,\nu\big) &= \int_{S_o} W_2^2(P^v_\#(T_\epsilon)_\#\mu, P^v_\#\nu)\ \mathrm{d}\lambda_o(v)\\ &\le \int_{S_o}\int_{\mathcal{M}\times\mathcal{M}} |P^v(T_\epsilon(x))-P^v(y)|^2\ \mathrm{d}\pi^v(x,y)\ \mathrm{d}\lambda_o(v).\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{\mathrm{CHSW}_2^2\big((T_\epsilon)_\#\mu,\nu\big) - \mathrm{CHSW}_2^2(\mu,\nu)}{2\epsilon}\\ &\le \int_{S_o}\int_{\mathcal{M}\times\mathcal{M}} \frac{|P^v(T_\epsilon(x))-P^v(y)|^2 - |P^v(x)-P^v(y)|^2}{2\epsilon}\ \mathrm{d}\pi^v(x,y)\ \mathrm{d}\lambda_o(v).\end{aligned}$$

Note $g(\epsilon) = \big(P^v(T_\epsilon(x))-P^v(y)\big)^2$. Then, $\frac{\mathrm{d}}{\mathrm{d}\epsilon}g(\epsilon)|_{\epsilon=0} = 2\big(P^v(x)-P^v(y)\big)\langle \mathrm{grad}_{\mathcal{M}}P^v(x),\xi(x)\rangle_x$. But, as for π^v-almost every (x,y), $P^v(y) = P^v(x) - \psi_v'(P^v(x))$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}g(\epsilon)|_{\epsilon=0} = 2\psi_v'\big(P^v(x)\big)\langle \mathrm{grad}_{\mathcal{M}}P^v(x),\xi(x)\rangle_x.$$

Finally, by the Lebesgue dominated convergence theorem, we obtain

$$\begin{aligned}&\limsup_{\epsilon\to 0^+} \frac{\mathrm{CHSW}_2^2\big((T_\epsilon)_\#\mu,\nu\big) - \mathrm{CHSW}_2^2(\mu,\nu)}{2\epsilon}\\ &\le \int_{S_o}\int_{\mathcal{M}} \psi_v'(P^v(x))\langle \mathrm{grad}_{\mathcal{M}}P^v(x),\xi(x)\rangle_x\ \mathrm{d}\mu(x)\mathrm{d}\lambda_o(v).\end{aligned}$$

■

E.2 Proof of Proposition 34

Proof of Proposition 34 We apply Proposition 33. First, using that for $f : x \mapsto \langle x,y\rangle_{\mathbb{L}}$, $\nabla f(x) = -KJy$, for all $x \in \mathbb{L}^d_K$,

$$\nabla B^v(x) = \sqrt{-K}J\frac{\sqrt{-K}x^0 + v}{\langle x, \sqrt{-K}x^0 + v\rangle_{\mathbb{L}}}.$$

Thus, noticing that $J^2 = I_{d+1}$,

$$\begin{aligned}
\operatorname{grad}_{\mathbb{L}^d_K} B^v(x) &= \operatorname{Proj}_x^K\big(-KJ\nabla B^v(x)\big) \\
&= \operatorname{Proj}_x^K\left(-K\sqrt{-K}\frac{\sqrt{-K}x^0+v}{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}}\right) \\
&= -K\sqrt{-K}\frac{\sqrt{-K}x^0+v}{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}} - K\left\langle x, -K\sqrt{-K}\frac{\sqrt{-K}x^0+v}{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}}\right\rangle_{\mathbb{L}} x \\
&= K\sqrt{-K}\left(\frac{\sqrt{-K}x^0+v}{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}} + K\frac{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}}{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}}x\right) \\
&= K\sqrt{-K}\left(-\frac{\sqrt{-K}x^0+v}{\langle x,\sqrt{-K}x^0+v\rangle_{\mathbb{L}}} + Kx\right).
\end{aligned}$$

Similarly, we have

$$\nabla P^v(x) = \frac{-KJ\big(\langle x,x^0\rangle_{\mathbb{L}}v - \langle x,v\rangle_{\mathbb{L}}x^0\big)}{\langle x,v\rangle^2_{\mathbb{L}} + K\langle x,x^0\rangle^2_{\mathbb{L}}}.$$

Thus, observing that $\langle x, \nabla P^v(x)\rangle_{\mathbb{L}} = 0$, we have

$$\begin{aligned}
\operatorname{grad}_{\mathbb{L}^d_K} P^v(x) &= \operatorname{Proj}_x^K\big(-KJ\nabla P^v(x)\big) \\
&= -KJ\nabla P^v(x) - K\langle x, -KJ\nabla P^v(x)\rangle_{\mathbb{L}}x \\
&= -KJ\nabla P^v(x) \\
&= \frac{K^2\big(\langle x,x^0\rangle_{\mathbb{L}}v - \langle x,v\rangle_{\mathbb{L}}x^0\big)}{\langle x,v\rangle^2_{\mathbb{L}} + K\langle x,x^0\rangle^2_{\mathbb{L}}}.
\end{aligned}$$

■

E.3 Proof of Lemma 35

Proof of Lemma 35 By Lemma 12, we have $\phi_{*,X}(V) = U\Sigma(V)U^T$ with $\Sigma(V) = U^TVU \odot \Gamma$. Thus,

$$\begin{aligned}
U\Sigma(V)U^T = W &\iff \Sigma(V) = U^TWU \\
&\iff U^TVU \odot \Gamma = U^TWU \\
&\iff U^TVU = U^TWU \oslash \Gamma \\
&\iff V = U\big(U^TWU \oslash \Gamma\big)U^T.
\end{aligned}$$

■

E.4 Proof of Lemma 36

Proof of Lemma 36 By (Pennec, 2020, Equation 3.8), we know that $\langle \log_{*,X}(V), Y\rangle = \langle \log_{*,X}(Y), V\rangle$. Thus, by linearity, we have that

$$\forall V \in T_X S_d^{++}(\mathbb{R}),\ P^A_{*,X}(V) = \langle A, \log_{*,X}(V)\rangle_F = \langle \log_{*,X}(A), V\rangle_F.$$

Then, applying Lemma 12, we have the result. ■

Appendix F. Busemann Function on SPDs endowed with Affine-Invariant Metric

Let $A \in S_d(\mathbb{R})$, $M \in S_d^{++}(\mathbb{R})$, we recall from Section 4.3.1 that the Busemann function can be computed as

$$\begin{aligned} B^A(M) &= \lim_{t\to\infty} d_{AI}\big(\exp(tA), M\big) - t \\ &= -\langle A, \log(\pi_A(M)\rangle_F, \end{aligned}$$

where π_A is a projection on the spaces of matrices commuting with $\exp(A)$ which belongs to a group $G \subset GL_d(\mathbb{R})$ leaving the Busemann function invariant. In the next paragraph, we detail how we can proceed to obtain π^A.

When A is diagonal with sorted values such that $A_{11} > \cdots > A_{dd}$, then the group leaving the Busemann function invariant is the set of upper triangular matrices with ones on the diagonal (Bridson and Haefliger, 2013, II. Proposition 10.66), *i.e.* for any such matrix g in that group, $B^A(M) = B^A(gMg^T)$. If the points are sorted in increasing order, then the group is the set of lower triangular matrices. Let's note G_U the set of upper triangular matrices with ones on the diagonal. For a general $A \in S_d(\mathbb{R})$, we can first find an appropriate diagonalization $A = P\tilde{A}P^T$, where $\tilde{A}$ is diagonal sorted, and apply the change of basis $\tilde{M} = P^T M P$ (Fletcher et al., 2009). We suppose that all the eigenvalues of A have an order of multiplicity of one. By the affine-invariance property, the distances do not change, *i.e.* $d_{AI}(\exp(tA), M) = d_{AI}(\exp(t\tilde{A}), \tilde{M})$ and hence, using the definition of the Busemann function, we have that $B^A(M) = B^{\tilde{A}}(\tilde{M})$. Then, we need to project $\tilde{M}$ on the space of matrices commuting with $\exp(\tilde{A})$ which we denote $F(A)$. By Bridson and Haefliger (2013, II. Proposition 10.67), this space corresponds to the diagonal matrices. Moreover, by Bridson and Haefliger (2013, II. Proposition 10.69), there is a unique pair $(g, D) \in G_U \times F(A)$ such that $\tilde{M} = gDg^T$, and therefore, we can note $\pi_A(\tilde{M}) = D$. This decomposition actually corresponds to a UDU decomposition. If the eigenvalues of A are sorted in increasing order, this would correspond to a LDL decomposition.

Appendix G. Additional Details on Experiments

Table 3: Dataset characteristics.

	BBCSport	**Movies**	**Goodreads genre**	**Goodreads like**
Doc	737	2000	1003	1003
Train	517	1500	752	752
Test	220	500	251	251
Classes	5	2	8	2
Mean words by doc	116 ± 54	182 ± 65	1491 ± 538	1491 ± 538
Median words by doc	104	175	1518	1518
Max words by doc	469	577	3499	3499

We sum up the statistics of the different datasets in Table 3.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

David Alvarez-Melis, Youssef Mroueh, and Tommi Jaakkola. Unsupervised hierarchy matching with optimal transport over hyperbolic spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 1606–1617. PMLR, 2020.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32, 2019.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. *Fast and Simple Computations on Tensors with Log-Euclidean Metrics.* PhD thesis, INRIA, 2005.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(2): 411–421, 2006.

Iskander Azangulov, Andrei Smolensky, Alexander Terenin, and Viacheslav Borovitskiy. Stationary kernels and Gaussian processes on Lie groups and their homogeneous spaces I: the compact case. *Journal of Machine Learning Research*, 25(280):1–52, 2024a.

Iskander Azangulov, Andrei Smolensky, Alexander Terenin, and Viacheslav Borovitskiy. Stationary kernels and Gaussian processes on Lie groups and their homogeneous spaces II: non-compact symmetric spaces. *Journal of Machine Learning Research*, 25(281):1–51, 2024b.

Werner Ballmann, Mikhael Gromov, and Viktor Schroeder. Manifolds of non positive curvature. In *Arbeitstagung Bonn 1984: Proceedings of the meeting held by the Max-Planck-Institut für Mathematik, Bonn June 15–22, 1984*, pages 261–268. Springer, 2006.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019.

Jérôme Bertrand and Benoît Kloeckner. A geometric study of Wasserstein spaces: Hadamard spaces. *Journal of Topology and Analysis*, 4(04):515–542, 2012.

Rajendra Bhatia. Positive Definite Matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Kingshook Biswas. The Fourier transform on negatively curved harmonic manifolds. *arXiv preprint arXiv:1802.07236*, 2018.

Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2007.

Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure Theory*, volume 1. Springer, 2007.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

Jan Boman and Filip Lindskog. Support theorems for the Radon transform and Cramér-Wold theorems. *Journal of theoretical probability*, 22(3):683–710, 2009.

Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. Efficient Gradient Flows in Sliced-Wasserstein Space. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic Sliced-Wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 334–370. PMLR, 2023a.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical Sliced-Wasserstein. In *The Eleventh International Conference on Learning Representations*, 2023b.

Clément Bonet, Benoît Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-Wasserstein on symmetric positive definite matrices for M/EEG signals. In *International Conference on Machine Learning*, pages 2777–2805. PMLR, 2023c.

Clément Bonet, Kimia Nadjahi, Thibault Séjourné, Kilian Fatras, and Nicolas Courty. Slicing Unbalanced Optimal Transport. *Transactions on Machine Learning Research*, 2024.

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

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51: 22–45, 2015.

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation.* PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Joey Bose, Ariella Smofsky, Renjie Liao, Prakash Panangaden, and Will Hamilton. Latent variable modelling with hyperbolic normalizing flows. In *International Conference on Machine Learning*, pages 1045–1055. PMLR, 2020.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds.* Cambridge University Press, 2023.

Martin R Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319. Springer Science & Business Media, 2013.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Daniel Brooks, Olivier Schwander, Frederic Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. *Advances in Neural Information Processing Systems*, 32, 2019a.

Daniel A Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Exploring complex time-series representations for Riemannian machine learning of radar data. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3672–3676. IEEE, 2019b.

Yann Cabanes. *Apprentissage dans les disques de Poincaré et de Siegel de séries temporelles multidimensionnelles complexes suivant un modèle autorégressif gaussien stationnaire centré: application à la classification de données audio et de fouillis radar.* PhD thesis, Bordeaux, 2022.

Jules Candau-Tilh. Wasserstein and Sliced-Wasserstein Distances. Master's thesis, Université Pierre et Marie Curie, 2020.

James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115):2, 1997.

Mathieu Carriere, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *International conference on machine learning*, pages 664–673. PMLR, 2017.

Ines Chami, Albert Gu, Dat P Nguyen, and Christopher Ré. HoroPCA: Hyperbolic dimensionality reduction via horospherical projections. In *International Conference on Machine Learning*, pages 1419–1429. PMLR, 2021.

Ricky TQ Chen and Yaron Lipman. Riemannian flow matching on general geometries. In *International Conference on Machine Learning*, 2023.

Ziheng Chen, Yue Song, Gaowen Liu, Ramana Rao Kompella, Xiaojun Wu, and Nicu Sebe. Riemannian multinomial logistics regression for SPD neural networks. In *Conference on Computer Vision and Pattern Recognition 2024*, 2024a.

Ziheng Chen, Yue Song, Tianyang Xu, Zhiwu Huang, Xiao-Jun Wu, and Nicu Sebe. Adaptive Log-Euclidean metrics for SPD matrix learning. *IEEE Transactions on Image Processing*, 2024b.

Emmanuel Chevallier and Nicolas Guigui. Wrapped statistical models on manifolds: motivations, the case SE(n), and generalization to symmetric spaces. In *Workshop on Joint Structures and Common Foundations of Statistical Physics, Information Geometry and Inference for Learning*, pages 96–106. Springer, 2020.

Emmanuel Chevallier, Emmanuel Kalunga, and Jesús Angulo. Kernel density estimation on spaces of Gaussian distributions and symmetric positive definite matrices. *SIAM Journal on Imaging Sciences*, 10(1):191–215, 2017.

Emmanuel Chevallier, Didong Li, Yulong Lu, and David Dunson. Exponential-wrapped distributions on symmetric spaces. *SIAM Journal on Mathematics of Data Science*, 4(4): 1347–1368, 2022.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

Seunghyuk Cho, Juyong Lee, and Dongwoo Kim. GM-VAE: Representation Learning with VAE on Gaussian Manifold. *arXiv preprint arXiv:2209.15217*, 2022a.

Seunghyuk Cho, Juyong Lee, Jaesik Park, and Dongwoo Kim. A rotated hyperbolic wrapped normal distribution for hierarchical representation learning. *Advances in Neural Information Processing Systems*, 35:17831–17843, 2022b.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Andrej Cvetkovski and Mark Crovella. Multidimensional scaling in the Poincaré disk. *arXiv preprint arXiv:1105.5332*, 2011.

Biwei Dai and Uros Seljak. Sliced iterative normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

Haitz Sáez de Ocáriz Borde, Alvaro Arroyo, Ismael Morales López, Ingmar Posner, and Xiaowen Dong. Neural latent geometry search: Product manifold inference via Gromov-Hausdorff-Informed Bayesian optimization. In *Advances in Neural Information Processing Systems*, 2023a.

Haitz Sáez de Ocáriz Borde, Anees Kazi, Federico Barbero, and Pietro Lio. Latent graph inference using product manifolds. In *The Eleventh International Conference on Learning Representations*, 2023b.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

Hanze Dong, Xi Wang, LIN Yong, and Tong Zhang. Particle-based variational inference with preconditioned functional gradient flow. In *International Conference on Learning Representations*, 2023.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Kernel methods in hyperbolic spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10665–10674, 2021.

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra, editors, *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141. PMLR, 26–28 Aug 2020.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3032–3042, 2015.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1): 3571–3578, 2021.

P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

P Thomas Fletcher, John Moeller, Jeff M Phillips, and Suresh Venkatasubramanian. Computing hulls and centerpoints in positive definite space. *arXiv preprint arXiv:0912.1580*, 2009.

P Thomas Fletcher, John Moeller, Jeff M Phillips, and Suresh Venkatasubramanian. Horoball hulls and extents in positive definite space. In *Workshop on Algorithms and Data Structures*, pages 386–398. Springer, 2011.

Fernando Galaz-Garcia, Marios Papamichalis, Kathryn Turnbull, Simon Lunagomez, and Edoardo Airoldi. Wrapped Distributions on homogeneous Riemannian manifolds. *arXiv preprint arXiv:2204.09790*, 2022.

Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry*, volume 2. Springer, 1990.

Robert C. Garrett, Trevor Harris, Zhuo Wang, and Bo Li. Validating climate models with spherical convolutional Wasserstein distance. In *Advances in Neural Information Processing Systems*, 2024.

Baptiste Genest, Nicolas Courty, and David Coeurjolly. Non-Euclidean sliced optimal transport sampling. In *Computer Graphics Forum*, page e15020. Wiley Online Library, 2024.

Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, 17, 2004.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Statistical inference with regularized optimal transport. *arXiv preprint arXiv:2205.04283*, 2022.

Jumpei Goto and Hiroyuki Sato. Approximated logarithmic maps on Riemannian manifolds and their applications. *JSIAM Letters*, 13:17–20, 2021.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019.

Mehrtash T Harandi, Mathieu Salzmann, and Richard Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for SPD matrices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 17–32. Springer, 2014.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008.

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

Zihao Hu, Guanghui Wang, and Jacob Abernethy. Riemannian projection-free online learning. In *Advances in Neural Information Processing Systems*, 2023.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. *Advances in Neural Information Processing Systems*, 29, 2016.

Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Stephan Huckemann and Herbert Ziezold. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38(2):299–319, 2006.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2464–2477, 2015.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Ce Ju and Cuntai Guan. Deep optimal transport on SPD manifolds for domain adaptation. *arXiv preprint arXiv:2201.05745*, 2022.

Isay Katsman, Eric Chen, Sidhanth Holalkere, Anna Asch, Aaron Lou, Ser Nam Lim, and Christopher M De Sa. Riemannian residual neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Jun Kitagawa and Asuka Takatsu. Two new families of metrics via optimal transport and barycenter problems. *arXiv preprint arXiv:2311.15874*, 2023.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in Neural Information Processing Systems*, 32, 2019.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730, 2021.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Serge Lang. *Fundamentals of Differential Geometry*, volume 191. Springer Science & Business Media, 2012.

Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of Wasserstein distances. *Advances in Neural Information Processing Systems*, 32, 2019.

Alice Le Brigant. *Probability on the spaces of curves and the associated metric spaces via information geometry; radar applications*. PhD thesis, Université de Bordeaux, 2017.

Alice Le Brigant and Stéphane Puechmorel. Approximation of densities on Riemannian manifolds. *Entropy*, 21(1):43, 2019.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

John M Lee. *Riemannian Manifolds: an Introduction to Curvature*, volume 176. Springer Science & Business Media, 2006.

John M Lee. *Smooth Manifolds*. Springer, 2012.

Jere Lehtonen. The geodesic ray transform on two-dimensional Cartan-Hadamard manifolds. *arXiv preprint arXiv:1612.04800*, 2016.

Jere Lehtonen, Jesse Railo, and Mikko Salo. Tensor tomography on Cartan–Hadamard manifolds. *Inverse Problems*, 34(4):044004, 2018.

Wenjie Lei, Zhengming Ma, Shuyu Liu, and Yuanping Lin. EEG mental recognition based on RKHS learning and source dictionary regularized RKHS subspace learning. *IEEE Access*, 9:150545–150559, 2021.

Rémi Leluc, François Portier, Johan Segers, and Aigerim Zhuman. Speeding up monte carlo integration: Control neighbors for optimal convergence. *arXiv preprint arXiv:2305.06151*, 2023.

Rémi Leluc, Aymeric Dieuleveut, François Portier, Johan Segers, and Aigerim Zhuman. Sliced-Wasserstein estimation with spherical harmonics as control variates. In *International Conference on Machine Learning*, 2024.

Tao Li, Cheng Meng, Hongteng Xu, and Jun Yu. Hilbert curve projection distance for distribution comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ya-Wei Eileen Lin, Ronald R Coifman, Gal Mishne, and Ronen Talmon. Hyperbolic diffusion embedding and distance for hierarchical representation learning. In *International Conference on Machine Learning*, pages 21003–21025. PMLR, 2023.

Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.

Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. Wasserstein task embedding for measuring task similarities. *Neural Networks*, 181:106796, 2025.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR, 2019.

Federico López, Beatrice Pozzetti, Steve Trettel, Michael Strube, and Anna Wienhard. Symmetric spaces for graph embeddings: A Finsler-Riemannian approach. In *International Conference on Machine Learning*, pages 7090–7101. PMLR, 2021.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M De Sa. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems*, 33:17548–17558, 2020.

Brice Loustau. Hyperbolic geometry. *arXiv e-prints*, pages arXiv–2003, 2020.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A González, and Thamar Solorio. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, 2017.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

Tudor Manole, Sivaraman Balakrishnan, and Larry Wasserman. Minimax confidence intervals for the Sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252–2345, 2022.

Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional Statistics*, volume 2. Wiley Online Library, 2000.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in Neural Information Processing Systems*, 33:2503–2515, 2020.

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

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, pages 1–25, 2024.

Dimitri Meunier, Massimiliano Pontil, and Carlo Ciliberto. Distribution regression with Sliced Wasserstein kernels. In *International Conference on Machine Learning*, pages 15501–15523. PMLR, 2022.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning. In *International Conference on Machine Learning*, pages 24925–24949. PMLR, 2023.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Pytorch metric learning. *arXiv preprint arXiv:2008.09164*, 2020.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the Sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, pages 20802–20812, 2020.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pages 4693–4702. PMLR, 2019.

Khai Nguyen and Nhat Ho. Sliced wasserstein estimation with control variates. In *International Conference on Learning Representations*, 2024a.

Khai Nguyen and Nhat Ho. Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 36, 2024b.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional Sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021a.

Khai Nguyen, Son Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Improving relational regularized autoencoders with spherical sliced fused Gromov Wasserstein. In *International Conference on Learning Representations*, 2021b.

Khai Nguyen, Nicola Bariletto, and Nhat Ho. Quasi-monte carlo for 3d sliced Wasserstein. In *International Conference on Learning Representations*, 2024.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

Vanni Noferini. A formula for the Fréchet derivative of a generalized matrix function. *SIAM Journal on Matrix Analysis and Applications*, 38(2):434–457, 2017.

Ruben Ohana, Kimia Nadjahi, Alain Rakotomamonjy, and Liva Ralaivola. Shedding a PAC-Bayesian light on adaptive sliced-Wasserstein distances. In *International Conference on Machine Learning*, pages 26451–26473. PMLR, 2023.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

Jiwoong Park, Junho Cho, Hyung Jin Chang, and Jin Young Choi. Unsupervised hyperbolic representation learning via message passing auto-encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5516–5526, 2021.

Sangmin Park and Dejan Slepčev. Geometry and analytic properties of the sliced wasserstein space. *arXiv preprint arXiv:2311.05134*, 2023.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International Conference on Machine Learning*, pages 5072–5081. PMLR, 2019.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2021.

Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

Xavier Pennec. Manifold-valued image processing with SPD matrices. In *Riemannian geometric statistics in medical image analysis*, pages 75–134. Elsevier, 2020.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

Gabriel Peyré, Marco Cuturi, et al. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Aram-Alexandre Pooladian, Carles Domingo-Enrich, Ricky TQ Chen, and Brandon Amos. Neural optimal transport with Lagrangian costs. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Alison Pouplin, David Eklund, Carl Henrik Ek, and Søren Hauberg. Identifying latent distances with Finslerian geometry. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Michael Quellmalz, Robert Beinert, and Gabriele Steidl. Sliced optimal transport on the sphere. *Inverse Problems*, 39(10):105005, 2023.

Michael Quellmalz, Léo Buecher, and Gabriele Steidl. Parallelly sliced optimal transport on spheres and on the rotation group. *Journal of Mathematical Imaging and Vision*, pages 1–26, 2024.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.

Alain Rakotomamonjy, Mokhtar Z Alaya, Maxime Berar, and Gilles Gasso. Statistical and topological properties of Gaussian smoothed sliced probability divergences. *arXiv preprint arXiv:2110.10524*, 2021.

Danilo J Rezende and Sébastien Racanière. Implicit Riemannian concave potential maps. *arXiv preprint arXiv:2110.01288*, 2021.

Danilo Jimenez Rezende, George Papamakarios, Sébastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pages 8083–8092. PMLR, 2020.

Joel W Robbin and Dietmar A Salamon. Introduction to Differential Geometry. *ETH, Lecture Notes, preliminary version*, 18, 2011.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. *Advances in Neural Information Processing Systems*, 34:17669–17680, 2021.

Boris Rubin. Notes on radon transforms in integral geometry. *Fractional Calculus and Applied Analysis*, 6(1):25–72, 2003.

Raif M Rustamov and Subhabrata Majumdar. Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *International Conference on Machine Learning*, pages 29388–29415. PMLR, 2023.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Manifold-regression to predict from MEG/EEG brain signals without source modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Filippo Santambrogio. Optimal Transport for Applied Mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Christopher Scarvelis and Justin Solomon. Riemannian metric learning via optimal transport. In *International Conference on Learning Representations*, 2023.

Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. *Advances in Neural Information Processing Systems*, 35:6802–6814, 2022.

Zhongmin Shen. *Lectures on Finsler geometry*. World Scientific, 2001.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations*, 2020.

Stefan Sommer, Tom Fletcher, and Xavier Pennec. Introduction to Differential and Riemannian Geometry. In *Riemannian Geometric Statistics in Medical Image Analysis*, pages 3–37. Elsevier, 2020.

Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Fully-connected network on noncompact symmetric space and ridgelet transform based on Helgason-Fourier analysis. In *International Conference on Machine Learning*, pages 20405–20422. PMLR, 2022.

Yann Thanwerdas and Xavier Pennec. O(n)-invariant Riemannian metrics on SPD matrices. *Linear Algebra and its Applications*, 661:163–201, 2023.

James Thornton, Michael Hutchinson, Emile Mathieu, Valentin De Bortoli, Yee Whye Teh, and Arnaud Doucet. Riemannian diffusion Schrödinger bridge. *arXiv preprint arXiv:2207.03024*, 2022.

Huy Tran, Yikun Bai, Abihith Kothapalli, Ashkan Shahbazi, Xinran Liu, Rocio P Diaz Martin, and Soheil Kolouri. Stereographic spherical sliced Wasserstein distances. In *International Conference on Machine Learning*, 2024.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9*, pages 589–600. Springer, 2006.

Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Jörg A Walter. H-MDS: a new approach for interactive visualization with multidimensional scaling in the hyperbolic space. *Information systems*, 29(4):273–292, 2004.

Yifei Wang, Peng Chen, and Wuchen Li. Projected Wasserstein gradient descent for high-dimensional bayesian inference. *SIAM/ASA Journal on Uncertainty Quantification*, 10 (4):1513–1532, 2022.

Jiaqi Xi and Jonathan Niles-Weed. Distributional convergence of the sliced Wasserstein process. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xianliang Xu and Zhongyi Huang. Central limit theorem for the sliced 1-Wasserstein distance and the max-sliced 1-wasserstein distance. *arXiv preprint arXiv:2205.14624*, 2022.

Or Yair, Felix Dietrich, Ronen Talmon, and Ioannis G Kevrekidis. Domain adaptation with optimal transport on the manifold of SPD matrices. *arXiv preprint arXiv:1906.00616*, 2019.

Ryoma Yataka, Kazuki Hirashima, and Masashi Shiraishi. Grassmann manifold flows for stable shape generation. *Advances in Neural Information Processing Systems*, 36:72377–72411, 2023.

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

Rixin Zhuang, Zhengming Ma, Weijia Feng, and Yuanping Lin. SPD data dictionary learning based on kernel learning and Riemannian metric. *IEEE Access*, 8:61956–61972, 2020.