# Stereographic Spherical Sliced Wasserstein Distances

**Huy Tran** [*1] **Yikun Bai** [*1] **Abihith Kothapalli** [*1] **Ashkan Shahbazi** [1] **Xinran Liu** [1] **Rocio Diaz Martin** [2]
**Soheil Kolouri** [1]

## Abstract

Comparing spherical probability distributions is of great interest in various fields, including geology, medical domains, computer vision, and deep representation learning. The utility of optimal transport-based distances, such as the Wasserstein distance, for comparing probability measures has spurred active research in developing computationally efficient variations of these distances for spherical probability measures. This paper introduces a high-speed and highly parallelizable distance for comparing spherical measures using the stereographic projection and the generalized Radon transform, which we refer to as the Stereographic Spherical Sliced Wasserstein (S3W) distance. We carefully address the distance distortion caused by the stereographic projection and provide an extensive theoretical analysis of our proposed metric and its rotationally invariant variation. Finally, we evaluate the performance of the proposed metrics and compare them with recent baselines in terms of both speed and accuracy through a wide range of numerical studies, including gradient flows and self-supervised learning. Our code is available at https://github.com/mint-vu/s3wd.

## 1. Introduction

Applications involving distributions defined on a hypersphere are remarkably diverse, highlighting the importance of spherical geometries across various disciplines. These applications include: 1) mapping the distribution of geographic or geological features on celestial bodies, such as stars and planets (Jupp, 2020; Cabella & Marinucci, 2009; Perraudin et al., 2019), 2) magnetoencephalography (MEG) imaging (Vrba & Robinson, 2001) in medical domains, 3) spherical image representations and 360° images (Coors et al., 2018; Jiang et al., 2023), such as omnidirectional images in computer vision (Khasanova & Frossard, 2017), 4) texture mapping in computer graphics (Elad et al., 2005; Dominitz & Tannenbaum, 2010), and more recently, 5) deep representation learning, where the latent representation is often mapped to a bounded space, commonly a sphere, where cosine similarity is utilized for effective representation learning (Chen et al., 2020; Wang & Isola, 2020).

The analysis of distributions on hyperspheres is traditionally approached through directional statistics, also referred to as circular/spherical statistics (Jammalamadaka et al., 2001; Mardia et al., 2000; Ley & Verdebout, 2017; Pewsey & García-Portugués, 2021). This specialized field is dedicated to the statistical analysis of directions, orientations, and rotations. More recently, with the growing application of optimal transport theory (Villani, 2008; Peyré & Cuturi, 2018) in machine learning, due in part to its favorable statistical, geometrical, and topological properties, there has been an increasing interest in using optimal transport to compare spherical probability measures (Cui et al., 2019; Hamfeldt & Turnquist, 2022).

One of the main bottlenecks in optimal transport theory is its high computational cost, generally of cubic complexity. This has sparked extensive research to develop faster solvers (Cuturi, 2013; Scetbon & Cuturi, 2022; Charikar et al., 2023) or computationally superior equivalent distances (Rabin et al., 2012; Bonneel et al., 2015; Kolouri et al., 2019a). Notably, sliced variations of optimal transport distances, such as sliced Wasserstein distances (Rabin et al., 2012) and their various extensions (Kolouri et al., 2019a; Le et al., 2019; Nguyen et al., 2022), have emerged as effective solutions. These methods use integral geometry and the Radon transform (Helgason et al., 2011) to represent high-dimensional distributions via a set of their one-dimensional marginals. By doing so, they take advantage of more efficient optimal transport solvers designed for one-dimensional probability measures, offering a pragmatic approach to comparing probability measures. Owing to their computational efficiency and implementation simplicity, these methods have been recently adapted for spherical measures, leading to the development of spherical sliced optimal transport

[*]Equal contribution [1]Department of Computer Science, Vanderbilt University, Nashville, TN [2]Department of Mathematics, Vanderbilt University, Nashville, TN. Correspondence to: Soheil Kolouri <soheil.kolouri@vanderbilt.edu>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

methods (Bonet et al., 2023a; Quellmalz et al., 2023). This comes as part of a broader effort to extend sliced Wasserstein distances to arbitrary manifolds; in this regard, we also highlight extensions of these methods to measures supported on compact manifolds (Rustamov & Majumdar, 2023), hyperbolic spaces (Bonet et al., 2023b), and symmetric positive definite matrices (Bonet et al., 2023c).

The main challenge in developing spherical sliced optimal transport methods lies in extending the classical Radon transform to its spherical counterparts. In the context of sliced optimal transport, such an extension must: 1) map probability measures on the hypersphere to a family of probability measures on one-dimensional domains (e.g., $\mathbb{R}$ or $\mathbb{S}^1$), and 2) be injective so that a sliced distance can be defined (otherwise, one will obtain only a pseudo-metric). These requirements rule out many of the existing extensions of the Radon transform to the sphere, e.g., the classic Funk-Radon transform, which takes integrals along all great circles (Helgason et al., 2011; Quellmalz, 2020). Recently, (Quellmalz et al., 2023) proposed using two such spherical extensions for the Radon transform, namely, the vertical slice transform (Gindikin et al., 1994b) and a normalized version of the semicircle transform (Groemer, 1998), to define sliced optimal transport on the sphere. Notably, the semicircle transform was also used in (Bonet et al., 2023a) to define a spherical sliced Wasserstein discrepancy for empirical probability measures.

The recent works on sliced spherical optimal transport (Bonet et al., 2023a; Quellmalz et al., 2023) map a distribution defined on a hypersphere into its marginal distributions on a unit circle, thus requiring circular optimal transport to compare these marginals. Importantly, the calculation of optimal transport between two one-dimensional measures defined on a circle is more expensive (requiring an additional binary search) than when the measures are defined on the real line (Martin et al., 2023; Hundrieser et al., 2022). Motivated by this observation, in this paper, we explore a spherical Radon transform aimed at converting a spherical distribution into one-dimensional marginals along the real line. We utilize the stereographic projection, which is a smooth bijection, in composition with an injective (nonlinear) map to transform the hypersphere into a hyperplane, where we then apply the classic Radon transform. The key advantage of the injective map is its ability to manage the distortion of distances introduced by the stereographic projection. We support our proposed method with detailed theoretical analysis and comprehensive numerical studies.

**Contributions.** Our specific contributions are as follows:

- Introducing a computationally efficient transport distance, Stereographic Spherical Sliced Wasserstein distance (S3W), for spherical probability measures.
- Providing a rotationally invariant variation of the proposed distance, Rotationally Invariant Stereographic Spherical Sliced Wasserstein distance (RI-S3W).
- Offering theoretical analysis of the proposed distances.
- Demonstrating the performance, both in terms of speed and accuracy, of the proposed distances in diverse applications, including gradient flows on the sphere, representation learning with Sliced-Wasserstein Auto-Encoders (SWAEs), spherical density estimation via normalizing flows, sliced-Wasserstein variational inference on the sphere, and self-supervised learning.

## 2. Background

### 2.1. Stereographic Projection

Stereographic projection is a mathematical technique to map points on a sphere onto a plane. Originating in Greek astronomy, it projects points from a sphere, typically from one of its poles, onto a plane tangential to the opposite pole or onto the "equator plane." This projection is conformal, preserving angles and shapes locally, making it significant in fields like cartography, complex analysis, and computer graphics. It elegantly translates spherical geometry into planar terms, maintaining the intricate relationships between points and angles found on the sphere's surface.

Let $\mathbb{S}^d$ denote the $d$-dimensional sphere in $\mathbb{R}^{d+1}$ defined as $\mathbb{S}^d := \{s \in \mathbb{R}^{d+1} : \|s\|_2 = 1\}$. The stereographic projection $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ maps a point $s \in \mathbb{S}^d$ (excluding the "North Pole," $s_n = [0, \dots, 0, 1]$) to a point $x \in \mathbb{R}^d$ by the formula:

$$[x]_i = \frac{2[s]_i}{1 - [s]_{d+1}}, \quad \text{for } i = 1, 2, \dots, d. \tag{1}$$

This projection is a bijective and smooth mapping between $\mathbb{S}^d \setminus \{s_n\}$ and $\mathbb{R}^d$, i.e., the hyperplane tangent to the sphere at the south pole, $s_{d+1} = -1$, which is commonly used as a way to visualize spherical geometries in an Euclidean space.

### 2.2. Radon Transform

The Radon transform is a fundamental tool in integral geometry, widely used for image reconstruction, particularly in computed tomography (CT). It transforms a function defined in the $d$-dimensional Euclidean space, $\mathbb{R}^d$, into an infinite set of its one-dimensional slices, i.e., its integrals over hyperplanes parametrized with unit vectors $\theta \in \mathbb{S}^{d-1}$. Radon transform is instrumental in medical imaging and has recently attracted ample attention from the machine learning community for its favorable characteristics for measuring distances between probability measures (Bonneel et al., 2015; Kolouri et al., 2016a; 2019a) and, more generally, positive measures (Bai et al., 2023; Séjourné et al., 2023), and more recently on theoretical analysis of deep neural networks (Parhi & Nowak, 2023; Unser, 2023).

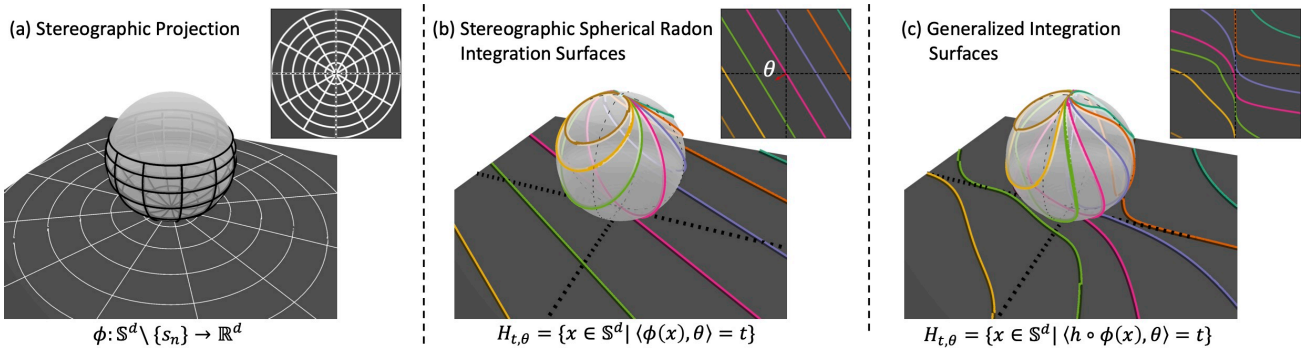



*Figure 1.* Depiction of stereographic projection from $\mathbb{S}^2 \setminus \{s_n\}$ to $\mathbb{R}^2$ (a), the stereographic Radon transform integration surfaces on the sphere, i.e., the level sets of $\langle \phi(x), \theta \rangle$ for a fixed $\theta \in \mathbb{R}^d$ (b), and the generalized stereographic Radon transform integration surfaces on the sphere, i.e. the level sets of $\langle h \circ \phi(x), \theta \rangle$ for a fixed $\theta \in \mathbb{R}^{d'}$.

More formally, the classic Radon transform, denoted by $\mathcal{R}$, maps a function $f$ in the Lebesgue space $L^1(\mathbb{R}^d)$ — the space of integrable functions on $\mathbb{R}^d$ — to its integrals over the hyperplanes of $\mathbb{R}^d$. Formally, it is defined as:

$$\mathcal{R}f(t, \theta) := \int_{\mathbb{R}^d} f(x)\delta(t - \langle x, \theta \rangle)dx, \tag{2}$$

for $(t, \theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$, where $\delta(\cdot)$ is the one-dimensional Dirac delta function, and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Here, $\mathcal{R}$ maps from $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{R} \times \mathbb{S}^{d-1})$. Each hyperplane, $H(t, \theta) = \{x \in \mathbb{R}^d \mid \langle x, \theta \rangle = t\}$, corresponds to a level set of a function $g : \mathbb{R}^d \times \mathbb{S}^{d-1} \to \mathbb{R}$, defined as $g(x, \theta) = \langle x, \theta \rangle$. For a fixed $\theta$, the set of all integrals over hyperplanes orthogonal to $\theta$ yields the continuous function $\mathcal{R}f(\cdot, \theta) : \mathbb{R} \to \mathbb{R}$, representing a projection or slice of $f$.

Notably, the Radon transform $\mathcal{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d-1})$ is a linear bijection with a closed-form inversion formula. See the Appendix for details.

### 2.3. Generalized Radon Transform

The Generalized Radon Transform (GRT) extends the foundational concept of the classic Radon transform, as introduced by Radon in (Radon, 1917), from integration over hyperplanes in $\mathbb{R}^d$ to integration over more complex structures, namely hypersurfaces or $(d-1)$-dimensional manifolds. This broader scope of the GRT has been developed and explored in various works (Beylkin, 1984; Denisyuk, 1994; Ehrenpreis, 2003; Gel'fand et al., 1969; Kuchment, 2006; Homan & Zhou, 2017), in many applications from impedance and acoustic tomography (Kuchment, 2006) to machine learning (Kolouri et al., 2019a).

The GRT of a function $f \in L^1(\mathbb{R}^d)$ involves the integration of $f$ over hypersurfaces in $\mathbb{R}^d$. These hypersurfaces are defined as the level sets of a 'defining function' (Kolouri et al., 2019a), $g : \mathbb{R}^d \times (\mathbb{R}^{d'} \setminus \{0\}) \to \mathbb{R}$, characterized by $H_{t,\theta} = \{x \in \mathbb{R}^d | \, g(x, \theta) = t\}$. The GRT of $f$, denoted as $\mathcal{G}f$, is formally given by:

$$\mathcal{G}f(t, \theta) := \int_{\mathbb{R}^d} f(x)\delta(t - g(x, \theta))\, dx, \tag{3}$$

where $\delta$ represents the Dirac delta function, enabling the integration over the specific level sets defined by $g$. Note that $g(x, \theta) = \langle x, \theta \rangle$ recovers the classic Radon transform.

The injectivity of the GRT is essential for defining distances between measures through their generalized slices. The choice of $g$ identifies whether GRT is injective or not. Kolouri et al. (2019a) enumerate a set of necessary conditions on $g$ to construct a bijective GRT. However, these conditions are not sufficient and are limited in practical utility, as they do not provide guidance on crafting specific defining functions that would ensure injectivity.

To achieve a practical injective GRT, Chen et al. (2022) recently introduced a variation of the GRT. They posited that by setting $g(x, \theta) = \langle h(x), \theta \rangle$ for an injective function $h : \mathbb{R}^d \to \mathbb{R}^{d'}$, one effectively applies the classic Radon transform, which is bijective, to the image of $h$. This approach leads to an injective GRT, thus resolving the issue of invertibility. For consistency of notation with Chen et al. (2022), we denote this variation of the GRT as:

$$\mathcal{H}f(t, \theta) := \int_{\mathbb{R}^d} f(x)\delta(t - \langle h(x), \theta \rangle)\, dx, \tag{4}$$

where $\mathcal{H} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d'-1})$. Importantly, a neural network can effectively parametrize the injective function $h$. For instance, $h$ could be a normalizing flow (Kobyzev et al., 2020). Alternatively, to avoid the complexities associated with normalizing flows, one can define $h(x)$ as $[x^T, \rho(x)^T]^T$, where $x^T$ and $\rho(x)^T$ are transposed vectors that are concatenated. Here, $\rho : \mathbb{R}^d \to \mathbb{R}^{d'-1}$ is any nonlinear function parametrized as a neural network of choice (Chen et al., 2022). In our experiments, we will use $\mathcal{H}$ as the default GRT due to its injectivity.

### 2.4. GRT of Radon Measures

Skipping much of the theoretical details (refer to the appendix), for a Radon measure $\mu \in \mathcal{M}(\mathbb{R}^d)$, its Generalized Radon Transform $\mathcal{G}(\mu) = \nu$ with respect to the defining function $g$, is defined as the measure $\nu \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1})$ such that for each $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$,

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\nu(t,\theta) = \int_{\mathbb{R}^d} (\mathcal{G}^*(\psi))(x)\, d\mu(x), \quad (5)$$

where $\mathcal{G}^*$ is the dual operator (aka adjoint operator), which for any $\psi \in L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1})$, is defined as

$$\mathcal{G}^*(\psi)(x) = \int_{\mathbb{S}^{d'-1}} \psi(g(x,\theta),\theta)\, d\sigma_{d'}(\theta) \quad \forall x \in \mathbb{R}^d, \quad (6)$$

where $\sigma_{d'}$ is the uniform probability measure defined in $\mathbb{S}^{d'-1}$. Importantly, the dual GRT operator satisfies,

$$\mathcal{G}(\mu)(\psi) = \mu(\mathcal{G}^*(\psi)).$$

Lastly, with a slight abuse of notation, we denote the corresponding slice for $\theta \in \mathbb{S}^{d'-1}$ as $\mathcal{G}(\mu)_\theta = g(\cdot,\theta)_\#\mu \in \mathcal{M}(\mathbb{R})$. Note that $f_\#\mu$ denotes the pushforward of measure $\mu$ with respect to $f$, defined as $f_\#\mu(A) = \mu(f^{-1}(A))$. When $\mu$ is a positive or a probability measure in $\mathbb{R}^d$, then $\mathcal{G}(\mu)_\theta$ is a positive/probability measure in $\mathbb{R}$.

### 2.5. Wasserstein and Sliced Wasserstein Distances

Let $M$ denote a Riemannian manifold endowed with the distance $d(\cdot,\cdot) : M \times M \to \mathbb{R}_+$. For $1 \le p < \infty$, let $\mu, \nu \in \mathcal{P}_p(M) := \{\mu \in \mathcal{P}(M) | \int_M d^p(x, x_0)\, d\mu(x) < \infty$ for some $x_0 \in M\}$ be two probability measures defined on manifold $M$ with a finite $p$-th moment. Then, the optimal transport (OT) problem (Villani, 2008) seeks to transport the mass in $\mu$ into $\nu$ such that the expected traversed distance is minimized. This leads to the $p$-Wasserstein distance:

$$W_p^p(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{M\times M} d^p(x,y) d\gamma(x,y), \quad (7)$$

where $\Gamma(\mu,\nu)$ denotes the joint probability measures $\gamma \in \mathcal{P}(M \times M)$ with marginals $\mu$ and $\nu$. Unfortunately, for discrete probability measures with $N$ particles, solving (7) generally has a $\mathcal{O}(N^3 \log N)$ complexity. However, for $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ the problem can be solved in $\mathcal{O}(N \log N)$:

$$W_p^p(\mu,\nu) = \int_0^1 \|F_\mu^{-1}(t) - F_\nu^{-1}(t)\|^p dt \quad (8)$$

where $F_\mu^{-1}$ and $F_\nu^{-1}$ are the quantile functions of $\mu$ and $\nu$. Notably, similar efficient solvers are developed for when $\mu, \nu \in \mathcal{P}_p(\mathbb{S}^1)$ (Delon et al., 2010; Hundrieser et al., 2022; Bonet et al., 2023a). For an injective generalized Radon transform $\mathcal{G}$, such that for $\mu \in \mathcal{P}_p(M)$ we have $\mathcal{G}(\mu)_\theta \in \mathcal{P}_p(\mathbb{R})$, the generalized Sliced-Wasserstein distance (Rabin et al., 2012; Kolouri et al., 2019a) can be defined as:

$$SW_{\mathcal{G},p}^p(\mu,\nu) := \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{G}(\mu)_\theta, \mathcal{G}(\nu)_\theta) d\sigma_{d'}(\theta), \quad (9)$$

where $\sigma_{d'} \in \mathcal{P}(\mathbb{S}^{d'-1})$ is a probability measure possessing a non-zero density on the sphere $\mathbb{S}^{d'-1}$, often simply the uniform measure. Recently, Bonet et al. (2023a) introduced the concept of Spherical Sliced Wasserstein distance, which involves projecting spherical measures onto great circles, resulting in $\mathcal{G}(\mu)_\theta \in \mathcal{P}_p(\mathbb{S}^1)$, and utilizing circular OT to measure distances between these slices. Notably, circular OT still necessitates solving an optimization problem to register cumulative distribution functions on a circle (i.e., finding an optimal cut). This results in a slower solver compared to OT on $\mathbb{R}$.

## 3. Method

Based on the extended definition of GRT for probability measures in 2.4, here we formally introduce the "Stereographic Spherical Radon Transform" and the corresponding sliced Wasserstein distance, "Stereographic Spherical Sliced Wasserstein Distance," for spherical probability measures.

### 3.1. Stereographic Spherical Radon Transform

Let $\mu \in \mathcal{M}(\mathbb{S}^d)$ denote a Radon measure defined on $\mathbb{S}^d$ that does not assign mass to the North Pole, i.e., $\mu(\{s_n\}) = 0$. We denote the stereographic projection as $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$, which is a bijection, and we have that $\phi_\#\mu$ is a Radon measure defined in $\mathbb{R}^d$. The **Stereographic Spherical Radon transform**, of $\mu$ is defined as

$$\mathcal{S}_\mathcal{R}(\mu) := \mathcal{R}(\phi_\#\mu) \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d-1}), \quad (10)$$

and we also define its generalized version as

$$\mathcal{S}_\mathcal{G}(\mu) := \mathcal{G}(\phi_\#\mu) \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1}). \quad (11)$$

Similarly we define $\mathcal{S}_\mathcal{H}(\mu) := \mathcal{H}(\phi_\#\mu)$, which provides an invertible transformation from $\mathcal{M}(\mathbb{S}^d)$ to $\mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1})$.

**Proposition 3.1** *For $\mu \in \mathcal{M}(\mathbb{S}^d)$ that does not give mass to the North Pole $\{s_n\}$ the Stereographic Spherical Radon transforms $\mathcal{S}_\mathcal{G}$ and $\mathcal{S}_\mathcal{H}$ satisfy the following properties:*

*(1) $\mathcal{S}_\mathcal{G}(\mu), \mathcal{S}_\mathcal{H}(\mu) \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1})$. In addition $\mathcal{S}_\mathcal{G}$ and $\mathcal{S}_\mathcal{H}$ preserves mass, and if $\mu$ is a positive measure, then $\mathcal{S}_\mathcal{G}(\mu), \mathcal{S}_\mathcal{H}(\mu)$ are positive measures too. Finally, if $\mu \in \mathcal{P}(\mathbb{S}^d \setminus \{s_n\})$, then $\mathcal{S}_\mathcal{G}(\mu), \mathcal{S}_\mathcal{H}(\mu)$ are probability measures defined on $\mathbb{R} \times \mathbb{S}^{d'-1}$.*

*(2) The disintegration theorem gives a unique $\mathcal{S}_\mathcal{G}(\mu)-$a.s. set of measures $(\mathcal{S}_\mathcal{G}(\mu)_\theta)_{\theta\in\mathbb{S}^{d'-1}} \subset \mathcal{M}(\mathbb{R})$,*

*and a unique $\mathcal{S}_{\mathcal{H}}(\mu)-$a.s. set of measures $(\mathcal{S}_{\mathcal{H}}(\mu)_\theta)_{\theta\in\mathbb{S}^{d'-1}} \subset \mathcal{M}(\mathbb{R})$ such that for any $\psi \in C_0(\mathbb{R}\times\mathbb{S}^{d'-1})$,*

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{G}}(\mu)(t,\theta) = \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{G}}(\mu)_\theta(t)\, d\sigma_{d'}(\theta), \tag{12}$$

*and similarly for $\mathcal{S}_{\mathcal{H}}$. Then, it holds that*

$$\mathcal{S}_{\mathcal{G}}(\mu)_\theta = (g(\cdot,\theta)\circ\phi)_{\#}\mu, \tag{13}$$

*and the same holds for $\mathcal{S}_{\mathcal{H}}$.*

*(3) $\mathcal{S}_{\mathcal{H}}$ is invertible.*

The proof of this proposition is included in the appendix Section D. Being equipped with the proposed stereographic spherical Radon transform, we are now ready to define our proposed distance.

### 3.2. Stereographic Spherical Sliced Wasserstein

Let $\mu,\nu \in \mathcal{P}_p(\mathbb{S}^d)$ denote two probability measures on the unit sphere in $\mathbb{R}^{d+1}$. We introduce the formal definition of the Stereographic Spherical Sliced Wasserstein ($S3W$) distances as follows:

$$S3W^p_{\mathcal{G},p}(\mu,\nu) := \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu)_\theta, \mathcal{S}_{\mathcal{G}}(\nu)_\theta)\, d\sigma_{d'}(\theta).$$

In this context, $\sigma_{d'} \in \mathcal{P}(\mathbb{S}^{d'-1})$ typically represents a probability measure possessing a non-zero density on the sphere $\mathbb{S}^{d'-1} \subset \mathbb{R}^{d'}$. However, for the sake of simplicity and in line with common practice, we opt to consider the uniform measure on $\mathbb{S}^{d'-1}$. Similarly, we can define $S3W^p_{\mathcal{H},p}$ by integrating over $W_p^p(\mathcal{S}_{\mathcal{H}}(\mu)_\theta, \mathcal{S}_{\mathcal{H}}(\nu)_\theta)$.

**Theorem 3.2** *The proposed $S3W_{\mathcal{G},p}(\cdot,\cdot)$ and $S3W_{\mathcal{H},p}(\cdot,\cdot)$ are well-defined. Furthermore, $S3W_{\mathcal{G},p}(\cdot,\cdot)$ is generally a pseudo-metric in $\mathcal{P}_p(\mathbb{S}^d\setminus\{s_n\})$, i.e., it is non-negative, symmetric and satisfies triangular inequality. In addition, $S3W_{\mathcal{H},p}(\cdot,\cdot)$ defines a metric in $\mathcal{P}_p(\mathbb{S}^d\setminus\{s_n\})$.*

The proof is in the appendix Section D.

### 3.3. Distance Distortion

The Stereographic Projection, while being conformal (i.e., preserving angles), severely distorts distances. To demonstrate the extent of this distortion, consider points $s = (\epsilon, 0, \ldots, 0, \sqrt{1-\epsilon^2})$ and $s' = (-\epsilon, 0, \ldots, 0, \sqrt{1-\epsilon^2})$. Then, as $\epsilon \to 0$, we have $\arccos(\langle s, s'\rangle) \to 0$, while $\|\phi(s) - \phi(s')\| \to \infty$! This distortion implies that the transportation cost after the stereographic projection would be significantly different from the transportation cost on the sphere, making the naive application of this projection with optimal transport unsuitable.

In this section, we aim to construct an injective function $h$ that closely approximates the arclength on the sphere $\mathbb{S}^d$ with the Euclidean distance in the embedded space. Specifically, we seek to satisfy:

$$\|h(\phi(s)) - h(\phi(s'))\| \approx \arccos(\langle s, s'\rangle),\ \forall s, s' \in \mathbb{S}^d.$$

We introduce two variants of $h$: an analytical function $h_1(\cdot)$ and a neural network-based learnable function $h_{NN}(\cdot)$, both mapping to a nearly-isometric embedding from $\phi(s)$.

We start by defining the analytic function:

$$h_1(x) := \arccos\left(\frac{\|x\|^2-1}{\|x\|^2+1}\right)\frac{x}{\|x\|}, \quad \forall x \in \mathbb{R}^d \tag{14}$$

and provide the following proposition.

**Proposition 3.3** *For $s_0, s, s' \in \mathbb{S}^d$, where $s_0$ denotes the South Pole, the stereographic projection $\phi$, and $h_1$ as defined in Eq. (14) we have:*

- $h_1(\phi(s)) = \angle(s, s_0)\frac{s[1:d]}{\|s[1:d]\|}$. *Thus, we have*
$$\|h_1(\phi(s)) - h_1(\phi(s'))\| \le 2\pi$$
- *The following inequality holds,*
$$\arccos(\langle s,s'\rangle) \le \|h_1(\phi(s)) - h_1(\phi(s'))\| + \epsilon(s,s')$$
*where the equality holds when $s$, $s'$, and $s_0$ are in the same great circle, and $\epsilon(s,s') \to 0$ as $d_{\mathbb{S}^d}(s,s') \to 0$.*

The proof is in the appendix Section H.

Next, we consider training a neural network to obtain a nearly-isometric Euclidean embedding. To do so, we define

$$h_{NN}(x) := [h_1^T(x)/C, \rho^T(x)]^T \tag{15}$$

where $\rho : \mathbb{R}^d \to \mathbb{R}^{d'-d}$ is a neural network, and $C \ge 2\pi$ a constant. We first note that such $h_{NN} : \mathbb{R}^d \to \mathbb{R}^{d'}$ is injective, then train $\rho$ by minimizing:

$$\mathcal{L}(\rho) = \mathbb{E}_{s,s'}\Big[(\arccos(\langle s, s'\rangle) - \|h_{NN}(\phi(s)) - h_{NN}(\phi(s'))\|)^2\Big], \tag{16}$$

where $s$ and $s'$ are sampled according to the uniform distribution in the sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$, i.e., $(s,s') \sim \sigma_{d+1}\times\sigma_{d+1}$. Figure 2 illustrates the arclength versus the distance in the embedding for random pairs of samples $s, s' \in \mathbb{S}^d$ with respect to the various scenarios proposed in this section. It is

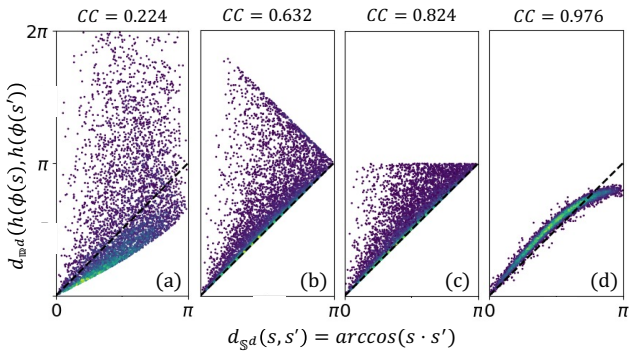



*Figure 2.* Spherical distance (i.e., the arclength) versus the distance after stereographic projection, where CC denotes Pearson's correlation coefficient. From left to right, when the injective function $h = id$, and the distance is $\|\phi(s) - \phi(s')\|$ (a), when $h(x) = h_1(x)$ (see Eq. (14)) and the distance is $\|h(\phi(s)) - h(\phi(s'))\|$ (b), when $h(x) = h_1(x)$ and the distance is $\min(\|h(\phi(s)) - h(\phi(s'))\|, 2\pi - \|h(\phi(s)) - h(\phi(s'))\|)$ (c), and finally when $h(x) = h_{NN}(x)$ (see Eq. (15)) where $\rho(x)$ is a trained neural network minimizing Eq. (16) and $C \geq 2\pi$ (d).

evident that an injective function $h_{NN}$ parameterized with a neural network can yield a nearly-isometric embedding. We discuss the use of this neural-network based $h_{NN}$ further in the appendix Section H.1.

To avoid any potential confusion, we emphasize that irrespective of the specific choice of $h$, as long as it maintains injectivity, $S3W_{\mathcal{H},p}$ remains a valid metric in $\mathcal{P}_p(\mathbb{S}^d)$. The discussion in this section is, however, significant, particularly when we aim to ensure that the transportation cost in the embedding closely resembles the spherical distance.

### 3.4. Rotationally Invariant Extension of $S3W$

Rotational symmetry is a significant property when dealing with probability metrics on the sphere. This symmetry implies that the probability distribution remains unchanged under rotations, making it a key consideration in spherical statistics and related applications. It simplifies calculations and enhances our understanding of the underlying processes on a spherical surface. Notably, the spherical OT leads to a rotationally symmetric metric on $\mathcal{P}_p(\mathbb{S}^d)$. However, our proposed Stereographic Spherical Sliced Wasserstein ($S3W$) metric is not rotationally invariant. Here, we propose a rotationally invariant variation of $S3W$, which leads to a robust and easy-to-implement rotationally invariant metric.

Let $\mathrm{SO}(d+1)$ denote the special orthogonal group in $\mathbb{R}^{d+1}$, and let $R \in \mathrm{SO}(d+1)$ denote a rotation matrix. For $\mu \in \mathcal{P}_p(\mathbb{S}^d)$ we denote the rotated measure as $R_\# \mu$. Given probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{S}^d)$, we define the rotationally invariant extension of $S3W$ as:

$$RI\text{-}S3W_{\mathcal{G},p}(\mu, \nu) := \mathbb{E}_{R\sim\omega}[S3W_{\mathcal{G},p}(R_\#\mu, R_\#\nu)], \quad (17)$$

where $\omega$ denotes the normalized Haar measure on $\mathrm{SO}(d+1)$ (i.e., the uniform probability measure which is left-rotation-invariant[1]). We similarly define RI-S3W$_{\mathcal{H},p}$ using S3W$_{\mathcal{H},p}$.

**Theorem 3.4** *$RI\text{-}S3W_{\mathcal{G},p}(\cdot,\cdot)$ and $RI\text{-}S3W_{\mathcal{H},p}(\cdot,\cdot)$ are well-defined. Furthermore, $RI\text{-}S3W_{\mathcal{G},p}(\cdot,\cdot)$ is generally a pseudo-metric in $\mathcal{P}_p(\mathbb{S}^d)$, i.e., it is non-negative, symmetric and satisfies triangular inequality. In addition, $RI\text{-}S3W_{\mathcal{H},p}(\cdot,\cdot)$ defines a metric in $\mathcal{P}_p(\mathbb{S}^d)$.*

The proof is included in the appendix Section E.

## 4. Numerical Implementation Details

**Stereographic Projection (SP)**

A key issue in implementing the SP for $S3W$ concerns the numerical handling of points near the North Pole. To ensure numerical stability, we introduce an $\epsilon$-cap around the North Pole, which serves to effectively establish an upper bound for the norm of the projected points. Specifically, any point where $x_{d+1} > 1 - \epsilon$ is initially mapped to the circle $x_{d+1} = 1 - \epsilon$, and then projected using SP. We refer to this modified SP as $\phi_\epsilon$. We discuss the stability of this $\epsilon$-cap in the appendix Section I.3.

**The $S3W$ Distances**

Let $\phi_\epsilon$ denote the Stereographic Projection operator that excludes the $\epsilon$-cap around $s_n \in \mathbb{S}^d$, and $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ an injective defining function, $L$ the number of slices, and $\theta_l \in \mathbb{S}^{d'-1}$ a slicing direction. For simplicity, we initially consider two data distributions with an equal number of samples and uniform mass distribution (the general case involving different numbers of samples and non-uniform mass distribution is discussed in the appendix). Let $\hat{\mu} = \frac{1}{M}\Sigma_{m=1}^M \delta_{x_m}$ and $\hat{\nu} = \frac{1}{M}\Sigma_{m=1}^M \delta_{y_m}$ denote the empirical distributions, where $\delta_{x_m}$ denotes a Dirac measure centered at $x_m$. Then, $S3W^p_{\mathcal{H},p}(\hat{\mu}, \hat{\nu})$ can be approximated via the following Monte Carlo estimator:

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{m=1}^{M}\left|\langle h(\phi_\epsilon(x_{\pi_l[m]})), \theta_l\rangle - \langle h(\phi_\epsilon(y_{\pi'_l[m]})), \theta_l\rangle\right|^p$$

where $\pi_l[m]$ and $\pi'_l[m]$ are the sorted indices of the projected samples on $\theta_l$ (see Algorithm 1 for the procedure).

[1]The Haar measure on a locally compact group $G$ is the unique, up to positive constants, left-translation-invariant regular Borel measure on $G$. If $G = \mathrm{SO}(d+1)$, the group translations are rotations. Here, we normalize it to obtain a probability measure.

**Algorithm 1** $S3W$

**Input:** $\{x_i\}_{i=1}^M \sim \mu$, $\{y_j\}_{j=1}^M \sim \nu$, $L$ projections, $p$-th order, $\epsilon$ for excluding the $\epsilon$-cap around $s_n$.
Initialize: $h$ (injective map), $\{\theta_l\}_{l=1}^L$ (projections)
Compute $\{u_i = h(\phi_\epsilon(x_i))\}$ and $\{v_j = h(\phi_\epsilon(y_j))\}$
Initialize distance $d = 0$
**for** $l = 1$ **to** $L$ **do**
  Compute $u_i^l = \langle u_i, \theta_l \rangle$, $v_j^l = \langle v_j, \theta_l \rangle$
  Sort $\{u_i^l\}$, $\{v_j^l\}$, s.t $u^l_{\pi_l[i]} \leq u^l_{\pi_l[i+1]}$, $v^l_{\pi'_l[j]} \leq v^l_{\pi'_l[j+1]}$
  $d = d + \frac{1}{L}\sum_{i=1}^M |u^l_{\pi_l[i]} - v^l_{\pi'_l[i]}|^p$
**end for**
**Return** $d^{\frac{1}{p}}$

### The $RI$-$S3W$ Distances

The Monte Carlo approximation of $RI$-$S3W$ distances can be written as:

$$RI\text{-}S3W_p(\hat{\mu}, \hat{\nu}) \approx \frac{1}{N_R}\sum_{n=1}^{N_R} S3W_p((R_n)_\#\hat{\mu}, (R_n)_\#\hat{\nu}),$$

where $\{R_n\}_{n=1}^{N_R} \subset \mathrm{SO}(d+1)$ are random rotation matrices. To generate $R_n$, we adopt the GeoTorch library (Lezcano-Casado, 2019), which provides a direct method to sample from $\mathrm{SO}(d+1)$. Our approach is efficient with vectorization and parallel processing on GPU. We note that generating $R_n$ is generally $\mathcal{O}(N_R \cdot d^3)$, which could become expensive for high-dimensional data or a large number of rotations. Instead, we could amortize this cost by presampling a rotation pool which could then be subsampled for every distance calculation. We denote this implementation as $ARI$-$S3W$, which involves a trade-off in memory and potential increased bias. In practice, we observe highly favorable performance in a variety of settings. We discuss this further in the appendix Section I.

### Computational Complexity

Stereographic projection requires $\mathcal{O}(Nd)$ where $N = n+m$ is the total number of data points from the source and target, and $(d+1)$ is the data dimensionality. If $n \gg m$, then we let $N = n$, and the same analysis holds. Applying $h(\cdot) = h_1(\cdot)$ to the projected data also requires $\mathcal{O}(Nd)$. Thereafter, slicing the data is done in $\mathcal{O}(LNd)$, where $L$ is the number of projections; sorting is done in $\mathcal{O}(LN\log N)$; and finally, the distance calculation requires $\mathcal{O}(LN)$. The overall time complexity for $S3W$ is therefore $\mathcal{O}(LN(d+\log N))$. For $RI$-$S3W$, the cost of calculating $S3W$ for all rotations is $\mathcal{O}(N_R LN(d+\log N))$. Generating the random rotations takes an overhead of $\mathcal{O}(N_R d^3)$ (which can be amortized), and applying these rotations to the data takes $\mathcal{O}(N_R N d^2)$. The total complexity is $\mathcal{O}(N_R(d^3 + Nd^2 + LN(d+\log N))$. If we amortize generating the rotation matrices, then the per-operation cost becomes $\mathcal{O}(N_R N(d^2 + Ld + L\log N))$.

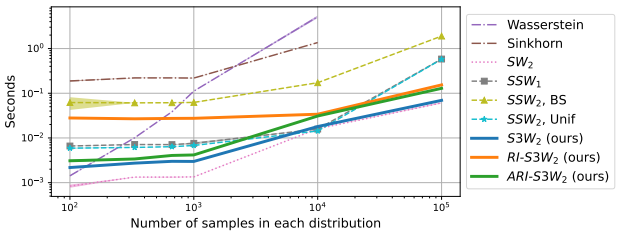



*Figure 3.* Runtime comparison for Wasserstein distance, Sinkhorn distance (Cuturi, 2013) with geodesic distance as cost function, $SW_2$ (sliced Wasserstein) distance, $SSW_1$ distance (using level median formula) (Bonet et al., 2023a), $SSW_2$ distance with binary search (BS) and antipodal closed form (for uniform distribution) (Bonet et al., 2023a), $S3W_2$ distance (ours), $RI$-$S3W_2$ distance (ours), and $ARI$-$S3W_2$ distance (ours).

### Runtime Comparison

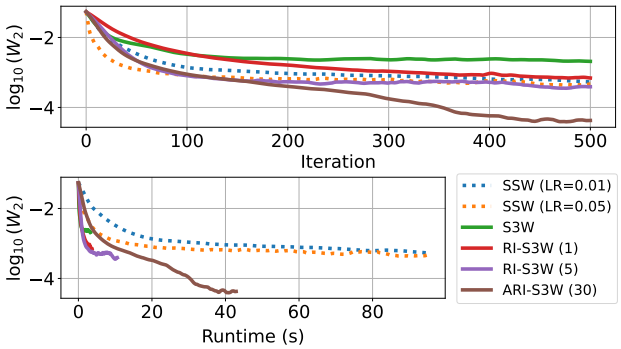
We compare the runtime to compute different distances between the uniform distribution and a von Mises-Fisher distribution on $\mathbb{S}^{100}$. The results in Figure 3 are averaged over 50 iterations for varying sample sizes of each distribution. We use $L = 200$ projections for all sliced methods, $N_R = 10$ rotations for $ARI$-$S3W_2$ and $RI$-$S3W_2$, and a pool size of 100 for $ARI$-$S3W_2$. We do not include the time required to generate the rotation pool in our $ARI$-$S3W_2$ measurements. Results for Wasserstein and Sinkhorn are based on the Python OT library (Flamary et al., 2021).

## 5. Experiments

Here, we present key results from our numerical experiments. We defer much of the visualization, discussions, and further results to the appendix Section I. All our experiments were executed on a Linux server with an AMD EPYC 7713 64-Core Processor, 8 × 32GB DIMM DDR4, 3200 MHz, and a NVIDIA RTX A6000 GPU.

### 5.1. Gradient Flow On The Sphere

Similar to the work of Bonet et al. (2023a), we apply our proposed distances as a loss function for the gradient flow problem. We consider a challenging target probability measure $\nu$ with 12 von Mises-Fisher distributions (vMFs), and aim to solve $\mathrm{argmin}_\mu S3W(\mu, \nu)$. Suppose we have access to the target measure only via i.i.d. samples $\{y_j\}_{j=1}^M$, i.e., $\hat{\nu} = \frac{1}{M}\sum_{j=1}^M \delta_{y_j}$ where $M = 2400$. We initialize 2400 particles uniformly sampled on $\mathbb{S}^2$, and directly optimize these particles with full-batch projected gradient descent on the surface of the sphere. Figure 4 shows the converged loss curves after 500 iterations and reports the runtime, negative log-likelihood (NLL), and the logarithm of the




| Method | Runtime(s) | NLL ↓ | $\log W_2$ ↓ |
|---|---|---|---|
| $SSW$ (LR=0.01) | 94.97 ± 0.96 | -4974.50 ± 2.27 | -3.27 ± 0.17 |
| $SSW$ (LR=0.05) | 94.98 ± 0.27 | -4974.46 ± 3.50 | -3.32 ± 0.13 |
| $S3W$ | **3.32 ± 0.62** | -4753.81 ± 83.70 | -2.57 ± 0.19 |
| $RI\text{-}S3W$ (1) | 4.00 ± 0.83 | -4957.22 ± 34.70 | -3.15 ± 0.26 |
| $RI\text{-}S3W$ (5) | 10.58 ± 0.27 | -4983.56 ± 6.62 | -3.50 ± 0.17 |
| $ARI\text{-}S3W$ (30) | 42.88 ± 0.05 | **-5025.37** ± 5.57 | **-4.37 ± 0.21** |

*Figure 4.* Learning a mixture of 12 vMFs. $ARI\text{-}S3W$ (30) has 30 rotations, pool size of 1000. $S3W$ variants use LR $= 0.01$. $SSW$ has an additional LR $= 0.05$ for better comparison. The plots show convergence of different distances w.r.t. iterations and runtime. The table summarizes numerical results for 10 independent runs. We provide more details of the plots in the appendix Section I.4.1.

2-Wasserstein distance between the distributions for $SSW$, $S3W$, $RI\text{-}S3W$ with $N_R \in \{1, 5\}$, and $ARI\text{-}S3W$ with $N_R = 30$ and pool size of 1000. We observe that the proposed distances provide on-par or better performance while being significantly faster than $SSW$ (up to 20X). Additionally, we provide mini-batch results in the appendix Section I.2. When the target is only known up to a constant, we use sliced-Wasserstein variational inference (Yi & Liu, 2022) (see the appendix, Section I.8).

### 5.2. Self-Supervised Learning (SSL)

We now show that our method can be an effective loss for contrastive SSL on the sphere. We adopt the contrastive objective proposed in (Wang & Isola, 2020), composed of alignment and uniformity loss terms, and replace the Gaussian kernel uniformity loss with an $S3W$-based loss:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\|z_i^A - z_i^B\|_2^2 + \frac{\lambda}{2}\left(\mathrm{S3W}_2(z^A,\nu) + \mathrm{S3W}_2(z^B,\nu)\right),$$

where $z^A, z^B \in \mathbb{R}^{n\times(d+1)}$ are two encoded views of the same images, $\nu = \mathrm{Unif}(\mathbb{S}^d)$ is the uniform distribution on $\mathbb{S}^d$ and $\lambda > 0$ is the regularization coefficient. We run our experiments on CIFAR-10 using a ResNet18 encoder. Figure 5 visualizes the learned embeddings when $d = 2$, and Table 1 assesses the quality of the learned embeddings for $d = 9$ using the standard linear classifier evaluation. The details of these experiments are included in the appendix Section I.7. We can see that the proposed metrics consistently lead to embeddings that are competitive both in terms of performance and runtime.

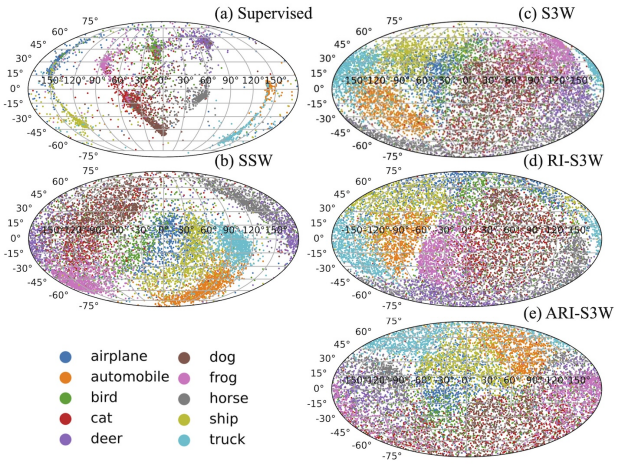



*Figure 5.* Projected features on $\mathbb{S}^2$ for CIFAR-10.

*Table 1.* Standard linear evaluation on CIFAR-10. Accuracy is for encoded (E) features and projected (P) features on $\mathbb{S}^9$ (encoder feature dimension is 10). Time reported is per epoch of pretraining.

| Method | Acc.(%) E/P ↑ | Time(s/ep.) |
|---|---|---|
| Supervised | 92.38 / 91.77 | —— |
| (Wang & Isola, 2020) | 79.76 / 74.57 | 24.28 |
| SimCLR (Chen et al., 2020) | 79.69 / 72.78 | **20.94** |
| $SSW$ (Bonet et al., 2023a) | 70.46 / 64.52 | 33.14 |
| $SW$ | 74.45 / 68.35 | 21.09 |
| $S3W$ | 78.54 / 73.84 | 21.36 |
| $RI\text{-}S3W$ (5) | **79.97** / 74.27 | 21.59 |
| $ARI\text{-}S3W$ (5) | 79.92 / **75.07** | 21.51 |

### 5.3. Sliced-Wasserstein Autoencoder (SWAE)

We now adopt the SWAE framework proposed by (Kolouri et al., 2019b) and demonstrate the application of $S3W$ distances in generative modeling. Let $\varphi : \mathcal{X} \to \mathbb{S}^d$ denote an encoder network and $\psi : \mathbb{S}^d \to \mathcal{X}$ denote a decoder network. The SWAE framework aims to enforce that the encoded data follow a prior distribution in the latent space. In our experiments, we use a mixture of vMF distributions with 10 components on $\mathbb{S}^2$ as our prior distribution, which we denote as $q$. Then, the training objective for the modified SWAE is:

$$\min_{\varphi,\psi} \mathbb{E}_{x\sim p}[c(x, \psi(\varphi(x))] + \lambda \cdot S3W(\varphi_{\#}p, q)$$

where $\lambda$ is the regularization coefficient, $c(\cdot,\cdot)$ the reconstruction loss, for which we use the standard Binary Cross Entropy (BCE) loss, and $p$ denotes the data distribution. Details about the network architectures and results on the MNIST benchmark can be found in the appendix Section I.9.

*Table 2.* CIFAR-10 results for SWAE. We evaluate the latent regularization loss ($\log(W_2)$ and NLL), along with the BCE reconstruction loss on the test data.

| Method | $\log W_2 \downarrow$ | NLL $\downarrow$ | BCE $\downarrow$ | Time (s/ep.) |
|---|---|---|---|---|
| Supervised | -0.5132 | 0.0060 | 0.6319 | **5.3243** |
| $SSW$ | -2.1949 | 0.0052 | 0.6323 | 15.4651 |
| $SW$ | -3.3229 | -0.0007 | 0.6348 | 5.4661 |
| $S3W$ | -3.3381 | 0.0025 | **0.6318** | 5.7511 |
| $RI\text{-}S3W$ (5) | -3.1424 | **-0.0043** | 0.6376 | 7.5443 |
| $ARI\text{-}S3W$ (5) | **-3.3853** | 0.0028 | 0.6332 | 5.8316 |

### 5.4. Earth Density Estimation

We extend our $S3W$ distances to the task of density estimation with normalizing flows on $\mathbb{S}^2$. Our focus is on the three datasets introduced by (Mathieu & Nickel, 2020), representing the Earth's surface as a perfect spherical manifold: Earthquake (NOAA, 2022), Flood (Brakenridge, 2017), and Fire (EOSDIS, 2020). Similar to (Bonet et al., 2023a), we use an exponential map normalizing flow model (Rezende et al., 2020) (see I.6) optimizing $\min_T S3W(T_\# \nu, q)$, where $T$ is the transformation induced by the model, $\nu$ is the data distribution known via samples $\{y_j\}_{j=1}^M$, and $q$ is the prior distribution on $\mathbb{S}^2$. The learned density at $y \in \mathbb{S}^2$ can then be approximated as $f_\nu(y) = q(T(y))|\det J_T(y)|$ where $J_T(y)$ denotes the Jacobian of $T$ at $y$.

*Table 3.* Earth datasets results. We use a pool size of 100 for $ARI\text{-}S3W$. Stereo denotes the approach introduced by (Gemici et al., 2016) which involves stereographically projecting the sphere $\mathbb{S}^d$ onto $\mathbb{R}^d$ and then performing RealNVP (Dinh et al., 2016).

| Method | Quake $\downarrow$ | Flood $\downarrow$ | Fire $\downarrow$ |
|---|---|---|---|
| Stereo | 2.04 ± 0.19 | 1.85 ± 0.03 | 1.34 ± 0.11 |
| $SW$ | 1.12 ± 0.07 | 1.58 ± 0.02 | 0.55 ± 0.18 |
| $SSW$ | 0.84 ± 0.05 | 1.26 ± 0.03 | 0.24 ± 0.18 |
| $S3W$ | 0.88 ± 0.09 | 1.33 ± 0.05 | 0.36 ± 0.04 |
| $RI\text{-}S3W(1)$ | 0.79 ± 0.07 | 1.25 ± 0.02 | 0.15 ± 0.06 |
| $ARI\text{-}S3W(50)$ | **0.78 ± 0.06** | **1.24 ± 0.04** | **0.10 ± 0.04** |

## 6. Conclusion

We introduced a new class of sliced Wasserstein ($SW$) distances for spherical data using the stereographic projection (SP). We rigorously addressed the distortion issue caused by SP and presented several high-speed, high-performing variants of our approach. $S3W$ maps data to a generalized hypersurface with minimal distortion (via composing SP with a novel injective function) and efficiently computes $SW$ over an order of magnitude faster than existing baselines in many settings. $RI\text{-}S3W$ encodes rotation invariance into $S3W$, further boosting its performance. Although $RI\text{-}S3W$ is highly parallelizable, we achieved additional efficiency with our implementation of amortization (denoted as $ARI\text{-}S3W$). Given that our approach uses stereographic projection to map the hypersphere into Euclidean space, there may exist novel extensions of our slicing framework to unbalanced settings by leveraging recent advancements in unbalanced and partial OT on $\mathbb{R}$ (Bai et al., 2023; Séjourné et al., 2023). Moreover, our approach opens up interesting directions to bridging the spherical manifold and classical $SW$ literature via pushforwards to appropriate hypersurfaces. Lastly, we highlight recent advancements in improving the projection complexity of sliced Wasserstein distances (Deshpande et al., 2019; Nguyen et al., 2020; Nguyen & Ho, 2022; Nguyen et al., 2022; Nguyen & Ho, 2024; Nguyen et al., 2024). These strategies are compatible with our proposed $S3W$ distances and can be integrated with our method to further enhance the projection efficiency of our metric in comparing high-dimensional spherical measures.

## Acknowledgements

This research was supported by the NSF CAREER Award No. 2339898. We are grateful to Clément Bonet for providing insights into the SSW implementation.

## Impact Statement

The work presented in this paper aims to advance the field of machine learning, particularly the supplementary theoretical developments and explorations of computational optimal transport. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## References

Abouelaz, A. and Daher, R. Sur la transformation de radon de la sphère sd. *Bulletin de la Société Mathématique de France*, 121(3):353–382, 1993. URL http://eudml.org/doc/87670.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Bai, Y., Schmitzer, B., Thorpe, M., and Kolouri, S. Sliced optimal partial transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13681–13690, 2023.

Beylkin, G. The inversion problem and applications of the generalized Radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984.

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

Bonet, C., Chapel, L., Drumetz, L., and Courty, N. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 334–370. PMLR, 2023b.

Bonet, C., Malézieux, B., Rakotomamonjy, A., Drumetz, L., Moreau, T., Kowalski, M., and Courty, N. Sliced-wasserstein on symmetric positive definite matrices for m/eeg signals. In *International Conference on Machine Learning*, pp. 2777–2805. PMLR, 2023c.

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

Bradawl, G. Statistics of earth science data, 2003.

Brakenridge, G. Global active archive of large flood events. Dartmouth Flood Observatory, University of Colorado, 2017. URL http://floodobservatory.colorado.edu/Archives/index.html.

Cabella, P. and Marinucci, D. Statistical challenges in the analysis of cosmic microwave background radiation. *Ann. Appl. Stat.*, 3(1):61–95, 2009.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Charikar, M., Chen, B., Ré, C., and Waingarten, E. Fast algorithms for a new relaxation of optimal transport. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4831–4862. PMLR, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X., Yang, Y., and Li, Y. Augmented sliced wasserstein distances. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iMqTLyfwnOO.

Coors, B., Condurache, A. P., and Geiger, A. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 518–533, 2018.

Cui, L., Qi, X., Wen, C., Lei, N., Li, X., Zhang, M., and Gu, X. Spherical optimal transportation. *Computer-Aided Design*, 115:181–193, 2019.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

De Cao, N. and Aziz, W. The power spherical distribution. *arXiv preprint arXiv:2006.04437*, 2020.

Delon, J., Salomon, J., and Sobolevski, A. Fast transport optimization for monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7):2239–2258, 2010.

Denisyuk, A. Inversion of the generalized Radon transform. *Translations of the American Mathematical Society-Series 2*, 162:19–32, 1994.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10648–10656, 2019.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Dominitz, A. and Tannenbaum, A. Texture mapping via optimal mass transport. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):419–433, 2010. doi: 10.1109/TVCG.2009.64.

Doucet, A., De Freitas, N., Gordon, N. J., et al. *Sequential Monte Carlo methods in practice*. Springer, 2001.

Ehrenpreis, L. *The universality of the Radon transform*. Oxford University Press on Demand, 2003.

Elad, A., Keller, Y., and Kimmel, R. Texture mapping via spherical multi-dimensional scaling. In *Scale Space and PDE Methods in Computer Vision: 5th International Conference, Scale-Space 2005, Hofgeismar, Germany, April 7-9, 2005. Proceedings 5*, pp. 443–455. Springer, 2005.

EOSDIS. Land, atmosphere near real-time capability for eos (lance) system operated by nasa's earth science data and information system (esdis). https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data, 2020.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Gel'fand, I. M., Graev, M. I., and Shapiro, Z. Y. Differential forms and integral geometry. *Functional Analysis and its Applications*, 3(2):101–114, 1969.

Gemici, M. C., Rezende, D., and Mohamed, S. Normalizing flows on riemannian manifolds. *arXiv preprint arXiv:1611.02304*, 2016.

Gindikin, S., Reeds, J., and Shepp, L. Spherical tomography and spherical integral geometry. *Tomography, Impedance Imaging, and Integral Geometry*, 30:83–92, 1994a.

Gindikin, S., Reeds, J., and Shepp, L. Spherical tomography and spherical integral geometry. In *Tomography, Impedance Imaging, and Integral Geometry: 1993 AMS-SIAM Summer Seminar on the Mathematics of Tomography, Impedance Imaging, and Integral Geometry, June 7-18, 1993, Mount Holyoke College, Massachusetts*, volume 30, pp. 83. American Mathematical Soc., 1994b.

Groemer, H. On a spherical integral transformation and sections of star bodies. *Monatshefte für Mathematik*, 126 (2):117–124, 1998.

Hamfeldt, B. F. and Turnquist, A. G. A convergence framework for optimal transport on the sphere. *Numerische Mathematik*, 151(3):627–657, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Helgason, S. et al. *Integral geometry and Radon transforms*. Springer, 2011.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017.

Hielscher, R., Potts, D., and Quellmalz, M. An svd in spherical surface wave tomography. *New Trends in Parameter Identification for Mathematical Models*, pp. 121–144, 2018.

Homan, A. and Zhou, H. Injectivity and stability for a generic class of generalized Radon transforms. *The Journal of Geometric Analysis*, 27(2):1515–1529, 2017.

Hundrieser, S., Klatt, M., and Munk, A. The statistics of circular optimal transport. In *Directional Statistics for Innovative Applications: A Bicentennial Tribute to Florence Nightingale*, pp. 57–82. Springer, 2022.

Jammalamadaka, S. R., Sengupta, A., and Sengupta, A. *Topics in circular statistics*, volume 5. world scientific, 2001.

Jiang, S., Li, Y., Weng, D., You, K., and Chen, W. 3d reconstruction of spherical images: A review of techniques, applications, and prospects. *arXiv preprint arXiv:2302.04495*, 2023.

Jupp, P. Some applications of directional statistics to astronomy. In *Multivariate Statistics and Matrices in Statistics: Proceedings of the 5th Tartu Conference, Tartu-Pühajärve, Estonia, 23–28 May, 1994*, pp. 123. Walter de Gruyter GmbH & Co KG, 2020.

Khasanova, R. and Frossard, P. Graph-based classification of omnidirectional images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 869–878, 2017.

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

Kobyzev, I., Prince, S. J., and Brubaker, M. A. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

Kolouri, S., Park, S. R., and Rohde, G. K. The Radon cumulative distribution transform and its application to image classification. *Image Processing, IEEE Transactions on*, 25(2):920–934, 2016a.

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced-Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4876–4884, 2016b.

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

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=H1xaJn05FQ.

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

Kuchment, P. Generalized transforms of Radon type and their applications. In *Proceedings of Symposia in Applied Mathematics*, volume 63, pp. 67, 2006.

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. Tree-sliced variants of wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Ley, C. and Verdebout, T. *Modern directional statistics*. CRC Press, 2017.

Lezcano-Casado, M. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems, NeurIPS*, pp. 9154–9164, 2019.

Mardia, K. V., Jupp, P. E., and Mardia, K. *Directional statistics*, volume 2. Wiley Online Library, 2000.

Martin, R. D., Medri, I., Bai, Y., Liu, X., Yan, K., Rohde, G. K., and Kolouri, S. Lcot: Linear circular optimal transport. *arXiv preprint arXiv:2310.06002*, 2023.

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

Nguyen, K. and Ho, N. Amortized projection optimization for sliced wasserstein generative models. *Advances in Neural Information Processing Systems*, 35:36985–36998, 2022.

Nguyen, K. and Ho, N. Energy-based sliced wasserstein distance. *Advances in Neural Information Processing Systems*, 36, 2024.

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2020.

Nguyen, K., Ren, T., Nguyen, H., Rout, L., Nguyen, T. M., and Ho, N. Hierarchical sliced wasserstein distance. In *The Eleventh International Conference on Learning Representations*, 2022.

Nguyen, K., Ren, T., and Ho, N. Markovian sliced wasserstein distances: Beyond independent projections. *Advances in Neural Information Processing Systems*, 36, 2024.

NOAA. Global Significant Earthquake Database, 2020. URL https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153. National Geophysical Data Center / World Data Service (NGDC/WDS): NCEI/WDS Global Significant Earthquake Database, NOAA National Centers for Environmental Information.

NOAA. Ncei/wds global significant earthquake database. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Parhi, R. and Nowak, R. D. Deep learning meets sparse regularization: A signal processing perspective. *arXiv preprint arXiv:2301.09554*, 2023.

Perraudin, N., Defferrard, M., Kacprzak, T., and Sgier, R. Deepsphere: Efficient spherical convolutional neural network with healpix sampling for cosmological applications. *Astronomy and Computing*, 27:130–146, 2019.

Pewsey, A. and García-Portugués, E. Recent advances in directional statistics. *Test*, 30(1):1–58, 2021.

Peyré, G. and Cuturi, M. Computational optimal transport. *arXiv preprint arXiv:1803.00567*, 2018.

Quellmalz, M. A generalization of the funk–radon transform. *Inverse Problems*, 33(3):035016, 2017.

Quellmalz, M. The funk–radon transform for hyperplane sections through a common point. *Analysis and Mathematical Physics*, 10(3):38, 2020.

Quellmalz, M., Beinert, R., and Steidl, G. Sliced optimal transport on the sphere. *arXiv preprint arXiv:2304.09092*, 2023.

Quinto, E. T. Null spaces and ranges for the classical and spherical radon transforms. *Journal of Mathematical Analysis and Applications*, 90(2):408–420, 1982.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2012.

Radon, J. Uber die bestimmug von funktionen durch ihre integralwerte laengs geweisser mannigfaltigkeiten. *Berichte Saechsishe Acad. Wissenschaft. Math. Phys., Klass*, 69: 262, 1917.

Rezende, D. J., Papamakarios, G., Racaniere, S., Albergo, M., Kanwar, G., Shanahan, P., and Cranmer, K. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pp. 8083–8092. PMLR, 2020.

Rubin, B. *Introduction to Radon transforms*, volume 160. Cambridge University Press, 2015.

Rubin, B. On the spherical slice transform. *Analysis and Applications*, 20(03):483–497, 2022.

Rustamov, R. M. and Majumdar, S. Intrinsic sliced wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *International Conference on Machine Learning*, 2023.

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

Séjourné, T., Bonet, C., Fatras, K., Nadjahi, K., and Courty, N. Unbalanced optimal transport meets sliced-wasserstein. *arXiv preprint arXiv:2306.07176*, 2023.

Snyder, J. P. *Flattening the earth: two thousand years of map projections*. University of Chicago Press, 1997.

Unser, M. Ridges, neural networks, and the radon transform. *Journal of Machine Learning Research*, 24(37): 1–33, 2023.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Vrba, J. and Robinson, S. E. Signal processing in magnetoencephalography. *Methods*, 25(2):249–271, 2001.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wang, X., Lei, Q., and Panageas, I. Fast convergence of langevin dynamics on manifold: Geodesics meet log-sobolev. *Advances in Neural Information Processing Systems*, 33:18894–18904, 2020.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. doi: 10.1109/CVPR.2018.00393.

Xu, J. and Durrett, G. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805*, 2018.

Yi, M. and Liu, S. Sliced wasserstein variational inference. In *Asian Conference on Machine Learning*, 2022.

# A. Notation

- $(\mathbb{R}^d, \|\cdot\|)$: d-dimensional Euclidean space, where $\|\cdot\|$ is the Euclidean norm (or $2-$norm): given $x = [x_1, \dots, x_d] \in \mathbb{R}^d$, $\|x\| = \sqrt{x_1^2 + \dots + x_d^2}$. Sometimes we will write $\|\cdot\|_2$ to stress that we are considering this "$2-$norm".
- $\langle \cdot, \cdot \rangle$: canonical inner product in Euclidean spaces.
- $L^p(\Omega)$, where $p \geq 1$ and $\Omega \subseteq \mathbb{R}^d$: functional space defined by

$$L^p(\Omega) := \{f : \Omega \to \mathbb{R} | \int_\Omega |f|^p < \infty\}$$

  endowed with the norm

$$\|f\|_p = \left(\int_\Omega |f|^p\right)^{\frac{1}{p}}.$$

  In the extreme case $p = \infty$, we have

$$L^\infty(\Omega) := \{f : \Omega \to \mathbb{R} | \sup |f| < \infty\}. \tag{18}$$

  In our case, $\Omega$ will be $\mathbb{R}^d$, $\mathbb{R} \times \mathbb{S}^{d-1}$, $\mathbb{R} \times \mathbb{R}^{d'} \setminus \{0\}$ or $\mathbb{S}^d$.
- $C(X)$: space of real-valued continuous functions defined on the space $X$.
- $C_0(\mathbb{R}^d)$: set of continuous functions "that vanish at infinity".
- $C^\infty$: class of infinite differentiable functions.
- $\mathbb{S}^{d-1}$: the unit sphere in $\mathbb{R}^d$, defined as $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|^2 = 1\}$.
- $\sigma_d$: the uniform probability measure defined in the sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$.
- $s_n = [0, \dots, 0, 1] \in \mathbb{R}^{d+1}$ is the North Pole and $s_0 = [0, \dots, 0, -1] \in \mathbb{R}^{d+1}$ is the South Pole in $\mathbb{S}^d \subset \mathbb{R}^{d+1}$.
- $\phi$: stereographic projection (SP).
- $\phi_\epsilon$: stereographic projection operator whose domain excludes the $\epsilon$-cap around $s_n$ in the sphere $\mathbb{S}^d$.
- $v^T$: transpose vector.
- $H(t, \theta) = \{x \in \mathbb{R}^d \mid \langle x, \theta \rangle = t\}$ hyperplane.
- $H_{t,\theta} = \{x \in \mathbb{R}^d | \, g(x, \theta) = t\}$ "level set" of $g : \mathbb{R}^d \times \mathbb{S}^{d-1} \to \mathbb{R}$ at level $t$ with fixed spherical variable $\theta$.
- $\mathcal{R}(\cdot)$: Radon transform with $\mathcal{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d-1})$.
- $\mathcal{G}(\cdot)$: Generalized Radon Transform (GRT) with

$$\mathcal{G}(f)(t, \theta) = \int_{\mathbb{R}^d} f(x)\delta(t - g(x, \theta))\, dx.$$

  In this formulation, $d' \geq d$ in general, and $g : \mathbb{R}^d \times (\mathbb{R}^{d'} \setminus \{0\}) \to \mathbb{R}$ is a function which satisfies the following:

(H.1) $g(x, \theta)$ is $C^\infty$ function on $\mathbb{R}^d \times (\mathbb{R}^{d'} \setminus \{0\})$.
(H.2) $g(x, \theta)$ is homogeneous of degree one in $\theta$, i.e., $g(x, \lambda\theta) = \lambda g(x, \theta)$ for all $\lambda \in \mathbb{R}$.
(H.3) $g(x, \theta)$ is non-degenerate with respect to $x$ in the sense that $\forall (x, \theta) \in \mathbb{R}^d \times (\mathbb{R}^{d'} \setminus \{0\})$, $\frac{\partial g}{\partial x}(x, \theta) \neq 0$.
(H.4) The mixed Hessian of $g$ is strictly positive, i.e., $\det\left(\frac{\partial^2 g}{\partial x^i \partial \theta^j}\right) > 0$.

  By property (H.2), $\mathcal{G} : L^1(\mathbb{R}^d) \to L(\mathbb{R} \times \mathbb{S}^{d'-1})$. We refer (Beylkin, 1984; Denisyuk, 1994; Gel'fand et al., 1969; Ehrenpreis, 2003; Kuchment, 2006; Homan & Zhou, 2017; Kolouri et al., 2019a) for more details.

- $\mathcal{H}(\cdot)$: a variant and simplified version of the Generalized Radon Transform defined by $\mathcal{H} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R} \times \mathbb{S}^{d'-1})$,

$$\mathcal{H}(f)(t,\theta) = \int_{\mathbb{R}^d} f(x)\delta(t - \langle h(x), \theta\rangle)dx,$$

  where $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ is injective.

- $h_1$: function defined by (14).

- $(t, \theta)$: the first and second inputs for (generalized) Radon transform, where $t \in \mathbb{R}$ and variable $\theta$ lies in a pre-defined sphere: $\theta \in \mathbb{S}^{d-1}$ for the classic Radon transform ($\mathcal{R}$), and $\theta \in \mathbb{S}^{d'-1}$ for the generalized cases $\mathcal{G}$ and $\mathcal{H}$.

- $\mathcal{R}^*, \mathcal{G}^*, \mathcal{H}^*$: the Hermitian adjoint operators for $\mathcal{R}, \mathcal{G}, \mathcal{H}$. In particular, $\mathcal{R}^* : L^\infty(\mathbb{R} \times \mathbb{S}^{d-1}) \to L^\infty(\mathbb{R}^d)$, and $\mathcal{G}^*, \mathcal{H}^* : L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1}) \to L^\infty(\mathbb{R}^d)$ with

$$\begin{aligned}
\mathcal{R}^*(\psi)(x) &= \int_{\mathbb{S}^{d-1}} \psi(\langle x, \theta\rangle, \theta)d\sigma_d(\theta) \\
\mathcal{G}^*(\psi)(x) &= \int_{\mathbb{S}^{d'-1}} \psi(g(x,\theta), \theta)d\sigma_{d'}(\theta) \\
\mathcal{H}^*(\psi)(x) &= \int_{\mathbb{S}^{d'-1}} \psi(\langle h(x), \theta\rangle, \theta)d\sigma_{d'}(\theta)
\end{aligned}$$

- $\mathcal{M}(\Omega)$: Set of all real Radon measures (finite regular Borel measures not necessarily positive, that is, it includes signed measures) defined on $\Omega$. In this article $\Omega$ can be $\mathbb{R}^d$, $\mathbb{R} \times \mathbb{S}^{d-1}$, or $\mathbb{R} \times \mathbb{S}^{d'-1}$.

  $\mathcal{M}(\Omega)$ is endowed with the total variation norm

$$\|\mu\|_{TV} = \mu^+(\Omega) + \mu^-(\Omega)$$

  where $\mu^\pm$ are the positive and negative parts of $\mu$.

- $\mathcal{M}_+(\Omega)$: Set of all positive Radon measures defined on $\Omega$. It is endowed with the total variation norm $\|\mu\|_{TV} = \mu(\Omega)$.

- $\mathcal{P}(\Omega)$: Set of all probability measures defined on $\Omega$ (i.e., positive measures with $\|\mu\|_{TV} = \mu(\Omega) = 1$).

- $\mathcal{P}_p(\Omega)$: Set of all probability measures defined on $\Omega$ with $p-$th finite moment, that is,

$$\mathcal{P}_p(\Omega) = \{\mu \in \mathcal{P}(\Omega) : \int_\Omega \|x\|^p d\mu < \infty\}$$

  We have the following relation:

$$\mathcal{P}_p(\Omega) \subseteq \mathcal{P}(\Omega) \subseteq \mathcal{M}_+(\Omega) \subseteq \mathcal{M}(\Omega).$$

- $T_\#\mu$: push-forward of the measure $\mu \in \mathcal{M}(X)$ by the measurable map $T : X \to Y$, which defines a measure on $\mathcal{M}(Y)$ such that $T_\#\mu(B) = \mu(\{x \in X : T(x) \in B\})$ for every measurable set $B \subseteq Y$.

- $W_p$: $p$-Wasserstein distance.

- $\hat{\mu} = \frac{1}{M}\Sigma_{m=1}^M \delta_{x_m}$: empirical distribution.

- Unif$(X), \sigma(X)$: Uniform distribution on a measure space $X$.

- $F_\mu$, $F_\mu^{-1}$: cumulative distribution function (CDF) and quantile function of the measure $\mu$, respectively.

- $\mathcal{S}_\mathcal{R}$, $\mathcal{S}_\mathcal{G}$, and $\mathcal{S}_\mathcal{H}$: Stereographic Spherical Radon Transforms (SSRT),

$$\mathcal{S}_\mathcal{R}(\mu) = \mathcal{R}(\phi_\#\mu), \qquad \mathcal{S}_\mathcal{G}(\mu) = \mathcal{G}(\phi_\#\mu), \qquad \mathcal{S}_\mathcal{H}(\mu) = \mathcal{H}(\phi_\#\mu).$$

- $S3W_{\mathcal{R},p}$, $S3W_{\mathcal{G},p}$, and $S3W_{\mathcal{H},p}$: Stereographic Spherical Sliced Wasserstein (S3W) distances

$$S3W^p_{\mathcal{R},p}(\mu,\nu) = \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{S}_{\mathcal{R}}(\mu)_\theta, \mathcal{S}_{\mathcal{R}}(\nu)_\theta)\, d\sigma_d(\theta),$$

$$S3W^p_{\mathcal{G},p}(\mu,\nu) = \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu)_\theta, \mathcal{S}_{\mathcal{G}}(\nu)_\theta)\, d\sigma_{d'}(\theta),$$

$$S3W^p_{\mathcal{H},p}(\mu,\nu) = \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{H}}(\mu)_\theta, \mathcal{S}_{\mathcal{H}}(\nu)_\theta)\, d\sigma_{d'}(\theta).$$

- $\mathbb{E}$: expected value.
- $\mathrm{O}(n)$, $\mathrm{SO}(n)$: orthogonal group (rotations and reflections) and special orthogonal group (rotations) of $n \times n$ real matrices, respectively.
- $\omega$: Haar probability measure on the compact group $\mathrm{SO}(d+1)$.
- $\approx$: approximately.
- $x \sim \mu$: element $x$ sampled from the distribution $\mu$.
- $\nabla_x$: gradient with respect to the variable $x$.
- $RI - S3W_{\mathcal{G},p}$: Rotationally Invariant Extension of S3W defined by (17).
- $ARI - S3W$: Amortized Rotationally Invariant Extension of S3W used in our experiments.
- $d_{\mathbb{S}^d}(\cdot,\cdot)$: The great circle distance on the sphere $\mathbb{S}^d$ defined by $d_{\mathbb{S}^d}(s_1,s_2) = \arccos(\langle s_1, s_2\rangle)$ for all $s_1, s_2 \in \mathbb{S}^d$.
- $\angle(s_1,s_2)$: the angle between the points $s_1$ and $s_2$ defined by $\angle(s_1,s_2) = \arccos(\langle s_1,s_2\rangle) \in [0,\pi]$.

## B. Radon Transform for Radon Measures

This section will review the classic Radon transform in the general Radon measure setting. In the next section, we introduce the Generalized Radon Transform (GRT) for general Radon measures.

As we discussed in the main section, given a function $f \in L^1(\mathbb{R}^d)$, the Radon transform, $\mathcal{R}(f)$, is a function in $L^1(\mathbb{R}\times\mathbb{S}^{d-1})$. In particular, for each $(t,\theta) \in \mathbb{R}\times\mathbb{S}^{d-1}$, we have:

$$\begin{aligned}\mathcal{R}(f)(t,\theta) &= \int_{\mathbb{R}^d} f(x)\delta(t - \langle x,\theta\rangle)\, dx \\ &= \int_{\mathbb{R}^{d-1}} f(t\theta + U_\theta \xi)\, d\xi\end{aligned}$$

where $U_\theta \in \mathbb{R}^{d\times(d-1)}$ is any matrix such that its columns are formed by an orthonormal basis of $\theta^\perp$ (the subspace which is perpendicular to the vector $\theta$), see equation (22) in (Bonneel et al., 2015). It is straightforward to verify the above definition is well-defined (i.e. $\mathcal{R}(f)$ is independent of the choice of $U_\theta$).

Notably, the Radon transform $\mathcal{R} : L^1(\mathbb{R}^d) \to L^1(\mathbb{R}\times\mathbb{S}^{d-1})$ is a linear bijection with a closed-form inversion formula. For a function $f \in L^1(\mathbb{R}^d)$, the Radon transform $\mathcal{R}f$ can be inverted using the inverse Radon transform, denoted by $\mathcal{R}^{-1}$. The inversion formula is given by:

$$f(x) = \mathcal{R}^{-1}(\mathcal{R}f)(x) = \int_{\mathbb{S}^{d-1}} \big(\mathcal{R}f(\cdot,\theta) * \eta(\cdot)\big)(\langle x,\theta\rangle)\, d\sigma_d(\theta), \tag{19}$$

where $\eta(\cdot)$ is a one-dimensional high-pass filter with corresponding Fourier transform $\mathcal{F}\eta(\omega) = c|\omega|^{d-1}$, which appears due to the Fourier slice theorem (Helgason et al., 2011), and '$*$' is the convolution operator. This integral formula effectively reconstructs $f$ from its Radon transform and, in the medical imaging community, is referred to as 'filtered back projection.'

Besides, the fact that the Radon transform $\mathcal{R}$ is a bounded linear operator from $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{R}\times\mathbb{S}^{d-1})$, yields to the definition of the dual operator [2], denoted as $\mathcal{R}^*$, which is generally called "back projection" (Helgason et al., 2011; Rubin, 2015). By definition of the dual operator and Riesz representation theorem that identifies the dual space of $L^1$ with $L^\infty$, for each $\psi \in L^\infty(\mathbb{R}\times\mathbb{S}^{d-1})$, we have

$$\int_{\mathbb{R}\times\mathbb{S}^{d-1}} \psi(t,\theta)\,\mathcal{R}(f)(t,\theta)\,dt\,d\sigma_d(\theta) = \int_{\mathbb{R}^d} f(x)\,\mathcal{R}^*(\psi)(x)\,dx. \tag{20}$$

Moreover, $\mathcal{R}^* : L^\infty(\mathbb{R}\times\mathbb{S}^{d-1}) \to L^\infty(\mathbb{R}^d)$ has closed form: For each $x \in \mathbb{R}^d$ and $\psi \in L^\infty(\mathbb{R}\times\mathbb{S}^{d-1})$,

$$\mathcal{R}^*(\psi)(x) = \int_{\mathbb{S}^{d-1}} \psi(\langle x,\theta\rangle,\theta)\,d\sigma_d(\theta), \tag{21}$$

where $\sigma_d$ is the uniform probability measure defined in the sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ .

As discussed in the main section, the classic Radon transform can describe the distribution in projected 1D space. Indeed, suppose $f$ is the density of a probability measure $\mu$, and $\theta \in \mathbb{S}^{d-1}$, then function $\mathcal{R}(f)(\cdot,\theta) : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is exactly the density of the projected probability measure of $\mu$ into the space spanned by $\theta$. We denote this projection $\theta_{\#}\mu$ by using pushforward notation where we identify $\theta \in \mathbb{S}^{d-1}$ with the function

$$\begin{aligned} \theta : \mathbb{R}^d &\to \mathbb{R}^d \\ x &\mapsto \langle x,\theta\rangle\,\theta, \end{aligned} \tag{22}$$

which range lies of the line generated by the direction $\theta$. Thus, we write

$$\mathcal{R}(f)(\cdot,\theta) = \theta_{\#}\mu \in \mathcal{M}(\mathbb{R}). \tag{23}$$

However, when $\mu$ is not a continuous measure, continuous (e.g. empirical distribution), then $\theta_{\#}\mu$ does not admit a density function. To address this limitation and extend $\mathcal{R}$ to any king of finite measures, we follow a measure-theoretic approach of the Radon transform as in (Bonneel et al., 2015) by using equation (20). We refer the reader also to (Helgason et al., 2011)[Ch.1, Sec.5]).

For this extension, we first recall the classical Riesz-Markov representation theorem and the lemma for the identity between Radon measures.

**Theorem B.1 (Riesz-Markov Representation theorem)** *Suppose $\Omega$ is a separable and locally compact space, then the dual space of the Banach space $(C_0(\Omega), \|\cdot\|_\infty)$, denoted as $(C_0(\Omega)^*, \|\cdot\|_*)$, is isomorphic to $(\mathcal{M}(\Omega), \|\cdot\|_{TV})$. In particular, for each bounded linear functional $\xi \in C_0(\Omega)^*$, there exists a unique $\nu \in \mathcal{M}(\Omega)$ such that*

$$\xi(\psi) = \int_\Omega \psi(x)\,d\nu(x), \qquad \textit{and} \qquad \|\xi\|_* = \|\nu\|_{TV}.$$

**Lemma B.2 (Identity of Radon measures)** *Given $\mu_1, \mu_2 \in \mathcal{M}(\Omega)$ where $\Omega \subset \mathbb{R}^d$, the following are equivalent:*

*(1)* $\nu_1 = \nu_2$

*(2)* $\int_\Omega \psi(x)\,d\nu_1(x) = \int_\Omega \psi(x)\,d\nu_2(x), \qquad \forall \psi \in C_0(\Omega)$

*(3)* $\int_\Omega \psi(x)\,d\nu_1(x) = \int_\Omega \psi(x)\,d\nu_2(x), \qquad \forall \psi \in L^\infty(\Omega)$

We will refer to the spaces $C_0(\Omega)$ and $L^\infty(\Omega)$ as the spaces of *test functions* for $\mathcal{M}(\Omega)$.

*Proof:* If $\nu_1 = \nu_2$, we directly have $(2), (3)$. Since $C_0(\Omega) \subset L^\infty(\Omega)$, then $(3)$ implies $(2)$. All we need to prove is $(2) \Rightarrow (1)$. By the Riesz-Markov Representation theorem B.1, $\mathcal{M}(\Omega)$ is isometric to $(C_0(\Omega)^*, \|\cdot\|_*)$. By the identity in the dual space, we have $\nu_1 = \nu_2$ given (2). □

[2] Some references call the back projection as adjoint (aka Hermitian conjugate) operator e.g. (Bonneel et al., 2015; Unser, 2023).

Based on the above theorem, for $\mu \in \mathcal{M}(\mathbb{R}^d)$, its Radon transform $\mathcal{R}(\mu) = \nu$ is defined as the measure $\nu \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d-1})$ such that for each $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d-1})$,

$$\int_{\mathbb{R}\times\mathbb{S}^{d-1}} \psi(t,\theta)\, d\nu(t,\theta) = \int_{\mathbb{R}^d} (\mathcal{R}^*(\psi))(x)\, d\mu(x), \tag{24}$$

where the dual operator $\mathcal{R}^*$ is defined in (21). Synthetically, one usually writes

$$\mathcal{R}(\mu)(\psi) = \mu(\mathcal{R}^*(\psi)).$$

Note that, by the above lemma, we can choose the $L^\infty(\mathbb{R} \times \mathbb{S}^{d-1})$ space of test functions. In particular, by setting $\psi = 1_A$ for any Borel set $A \subset \mathbb{R} \times \mathbb{S}^{d-1}$, we can verify that $\nu$ is Radon measure if $\mu$ is Radon measure. Similarly, $\nu$ is a positive Radon measure if $\mu$ is positive. In addition, one can obtain $\|\nu\|_{TV} = \|\mu\|_{TV}$. Thus, if $\mu$ is a probability measure on $\mathbb{R}^d$, then $\nu = \mathcal{R}(\mu)$ is a probability measure defined on $\mathbb{R} \times \mathbb{S}^{d-1}$.

By the disintegration theorem in classic measure theory, there exists a $\nu-$a.s. unique set of measures $(\nu_\theta)_{\theta\in\mathbb{S}^{d-1}} \subset \mathcal{M}(\mathbb{R})$ such that for any $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d-1})$, we have

$$\int_{\mathbb{R}\times\mathbb{S}^{d-1}} \psi(t,\theta)\, d\nu(t,\theta) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d\nu_\theta(t)\, d\sigma_d(\theta). \tag{25}$$

Similar to (23) for the classic Radon measure, given a fixed $\theta \in \mathbb{S}^{d-1}$, $\mathcal{R}(\mu)_\theta$ describes the projected measure of $\mu$ into the 1D space spanned by $\theta$ (Bonneel et al., 2015)[Proposition 6], that is:

$$\mathcal{R}(\mu)_\theta = \nu_\theta = \theta_\# \mu \in \mathcal{M}(\mathbb{R}).$$

## C. Generalized Radon Transform for Probability Measures

Extending the Radon transform as in (3) and (4) is natural. In this section, we review those definitions.

Given $g : \mathbb{R}^d \times (\mathbb{R}^{d'} \setminus \{0\}) \to \mathbb{R}$, such that $g$ satisfies the technical assumptions [H.1–H.4] listed in the Notation section and given in (Beylkin, 1984) and (Kolouri et al., 2019a), the Generalized Radon Transform (GRT) is defined as

$$\mathcal{G}(f)(t,\theta) = \int_{\mathbb{R}^d} f(x)\delta(t - g(x,\theta))\, dx, \qquad \forall (t,\theta) \in \mathbb{R} \times (\mathbb{R}^{d'} \setminus \{0\}).$$

Note, for each $(t,\theta) \in \mathbb{R} \times (\mathbb{R}^{d'} \setminus \{0\})$, we set the corresponded hyper-surface as $H_{t,\theta} = \{x \in \mathbb{R}^d \mid g(x,\theta) = t\}$. By the homogeneity property (H.2), $H_{\lambda t,\lambda\theta} = H_{t,\theta}$, $\forall \lambda \in \mathbb{R}$. Then,

$$\mathcal{G}f(\lambda t, \lambda\theta) = \mathcal{G}f(t,\theta), \qquad \forall (t,\theta,\lambda) \in \mathbb{R} \times (\mathbb{R}^{d'} \setminus \{0\}) \times \mathbb{R}.$$

Thus, the hypersurface can be parameterized by $(t,\theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$, and we can redefine $\mathcal{G}$ as a mapping from space $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{R} \times \mathbb{S}^{d'-1})$.

Furthermore, as $\mathcal{G}$ is an linear operator from $L^1(\mathbb{R}^d)$ to $L^1(\mathbb{R} \times \mathbb{S}^{d'-1})$, we can define the dual operator $\mathcal{G}^*$ of $\mathcal{G}$ as in (Homan & Zhou, 2017) (also called the *adjoint* operator). In analogy with the Radon transform, this is done by the duality between $L^1$ and $L^\infty$: For each $\psi \in L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1})$,

$$\mathcal{G}^*(\psi)(x) = \int_{\mathbb{S}^{d'-1}} \psi(g(x,\theta),\theta)\, d\sigma_{d'}(\theta), \qquad \forall x \in \mathbb{R}^d, \tag{26}$$

where $\sigma_{d'}$ is the uniform probability measure defined in $\mathbb{S}^{d'-1}$.

In the specific case, $g(x,\theta) = \langle h(x), \theta\rangle$, where $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a injective function, the induced GRT becomes

$$\mathcal{H}(f)(t,\theta) = \int_{\mathbb{S}^{d'-1}} f(x)\delta(t - \langle h(x),\theta\rangle)\, dx, \qquad \forall (t,\theta) \in \mathbb{R} \times \mathbb{S}^{d'-1},$$

with corresponding dual operator

$$\mathcal{H}^*(\psi)(x) = \int_{\mathbb{S}^{d'-1}} \psi(\langle h(x), \theta\rangle, \theta)\, d\sigma_{d'}(\theta), \qquad \forall \psi \in L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1}),\, x \in \mathbb{R}^d. \tag{27}$$

In this case, that is $g(x, \theta) = \langle h(x), \theta\rangle$ for $h = (h_1, \ldots, h_{d'}) : \mathbb{R}^d \to \mathbb{R}^{d'}$, properties [H.1-H.4] read as follows: First, notice that the real inner product in $\mathbb{R}^d$ defines, in each variable, a $C^\infty$ function which is homogeneous of degree one. In particular, (H.2) is trivially satisfied. Then, $h$ must satisfy:

(a) In place of (H.1): $h \in C^\infty$.

(b) In place of (H.3): For all $(x, \theta) \in \mathbb{R}^d \times \mathbb{S}^{d'-1}$, $\left(\langle \frac{\partial h}{\partial x^1}(x), \theta\rangle, \ldots, \langle \frac{\partial h}{\partial x^d}(x), \theta\rangle\right) \neq 0$.

(c) In place of (H.4): $\det\left(\left[\frac{\partial h_i}{\partial x^j}\right]_{1\leq i\leq 1, 1\leq j\leq d'}\right) > 0$.

In particular, if $d = d'$, let $\mathbf{J}_h$ denote the Jacobian of $h = (h_1, \ldots, h_d) : \mathbb{R}^d \to \mathbb{R}^d$, that is, the $d' \times d$ matrix $\mathbf{J}_h = \left[\frac{\partial h_i}{\partial x^j}\right]_{ij}$. Then, condition (b) can be written as the matrix-vector multiplication $\mathbf{J}_h(x) \cdot \theta \neq 0$, and condition (c) is equivalent $\det(\mathbf{J}_h(x)) > 0$ for all $x \in \mathbb{R}^d$, and so in this case (b) is redundant. Therefore, when $d = d'$, $h : \mathbb{R}^d \to \mathbb{R}^d$ must satisfy:

$$h \in C^\infty \qquad \text{and} \qquad \det(\mathbf{J}_h) > 0.$$

Inspired by the framework of classic Radon transform for Radon measure, we can extend the idea into the GRT setting. In particular, given $\mu \in \mathcal{M}(\mathbb{R}^d)$, we define the measures $\mathcal{G}(\mu)$ and $\mathcal{H}(\mu)$ as follows: For each test function $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$,

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t, \theta)\, d\mathcal{G}(\mu)(t, \theta) = \int_{\mathbb{R}^d} \mathcal{G}^*(\psi)(x)\, d\mu(x) \tag{28}$$

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t, \theta)\, d\mathcal{H}(\mu)(t, \theta) = \int_{\mathbb{R}^d} \mathcal{H}^*(\psi)(x)\, d\mu(x) \tag{29}$$

**Proposition C.1** *Given $\mu \in \mathcal{M}(\mathbb{R}^d)$, the measures $\mathcal{G}(\mu), \mathcal{H}(\mu)$ is defined in (28) and (29) have the following properties:*

*(1) $\mathcal{G}(\mu)$ is well-defined, i.e., $\mathcal{G}(\mu) \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1})$.*

*(2) If $\mu \in \mathcal{M}_+(\mathbb{R}^d)$, then $\mathcal{G}(\mu) \in \mathcal{M}_+(\mathbb{R} \times \mathbb{S}^{d'-1})$. Similarly, if $\mu \in \mathcal{P}(\mathbb{R}^d)$, then $\mathcal{G}(\mu) \in \mathcal{P}(\mathbb{R} \times \mathbb{S}^{d'-1})$.*

*(3) $\mathcal{G}$ preserves the the total variation norm, i.e., $\|\mu\|_{TV} = \|\mathcal{G}(\mu)\|_{TV}$. In particular, $\mathcal{G}$ preserves mass.*

*(4) If $\mu \in \mathcal{P}(\mathbb{R}^d)$, then $\mathcal{G}(\mu) \in \mathcal{P}(\mathbb{R} \times \mathbb{S}^{d'-1})$.*

*(5) Given $\mu \in \mathcal{M}(\mathbb{R}^d)$, the disintegration theorem gives a unique $\mathcal{G}(\mu)-$a.s. set of measures $(\mathcal{G}(\mu)_\theta)_{\theta\in\mathbb{S}^{d'-1}} \subset \mathcal{M}(\mathbb{R})$ such that for any $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$,*

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t, \theta)\, d\mathcal{G}(\mu)(t, \theta) = \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t, \theta)\, d\mathcal{G}(\mu)_\theta(t)\, d\sigma_{d'}(\theta)$$

*Then, it holds that*

$$\mathcal{G}(\mu)_\theta = g(\cdot, \theta)_\# \mu.$$

*(6) Properties $(1) - (5)$ hold true for $\mathcal{H}$ in place of $\mathcal{G}$.*

*(7) Given $\mu \in \mathcal{M}(\mathbb{R}^d)$, the disintegration theorem gives a unique $\mathcal{H}(\mu)-$a.s. set of measures $(\mathcal{H}(\mu)_\theta)_{\theta\in\mathbb{S}^{d'-1}} \subset \mathcal{M}(\mathbb{R})$ such that for any $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$,*

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t, \theta)\, d\mathcal{H}(\mu)(t, \theta) = \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t, \theta)\, d\mathcal{H}(\mu)_\theta(t)\, d\sigma_{d'}(\theta)$$

*Then, it holds that*

$$\mathcal{H}(\mu)_\theta = \theta_\# h_\# \mu,$$

*or equivalently, $\mathcal{H}(\mu)_\theta = (\theta \circ h)_\# \mu$, where the function $\theta$ is defined as in (22).*

*(8)* $\mathcal{H}(\mu) = \mathcal{R}(h_{\#}\mu)$ *for every* $\mu \in \mathcal{M}(\mathbb{R}^d)$.

*(9) When* $h$ *is injective,* $\mathcal{H}$ *is invertible.*

*Proof:*

*(1) Let* $\mu \in \mathcal{M}(\mathbb{R}^d)$. *By (Beylkin, 1984; Homan & Zhou, 2017), we have that (6) is the dual operator of* $\mathcal{G}$. *In fact, by functional analysis theory,* $L^\infty$ *is the dual space of Banach space* $L^1$, *and thus,* $\mathcal{G}^* : L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1}) \to L^\infty(\mathbb{R}^d)$ *is well-defined. In particular,* $\int_{\mathbb{R}^d} \mathcal{G}^*(\psi)(x)\, d\mu(x)$ *is finite for all* $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1}) \subset L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1})$. *Then, expression (28) is finite for all* $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$, *and moreover defines* $\mathcal{G}(\mu)$ *as an element in the dual space of* $C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$. *By Theorem B.1, we have* $\mathcal{G}(\mu) \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1})$.

*(2) Let* $\mu \in \mathcal{M}_+(\mathbb{R}^d)$ *and* $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$ *with* $\psi(t,\theta) \geq 0, \forall t, \theta$. *Thus, by definition (6),* $\mathcal{G}^*\psi(x) \geq 0, \forall x \in \mathbb{R}^d$. *Therefore,*

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{G}\mu(t,\theta) = \int_{\mathbb{R}^d} \mathcal{G}^*(\psi)(x)\, d\mu(x) \geq 0$$

*which follows from the facts* $\mathcal{G}^*(\psi) \geq 0$ *and* $\mu \in \mathcal{M}_+(\mathbb{R}^d)$. *Since the above property holds for all non-negative test function* $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$, *we obtain that* $\mathcal{G}(\mu) \in \mathcal{M}_+(\mathbb{R} \times \mathbb{S}^{d'-1})$.

*(3) Set* $\psi = 1_{\mathbb{R}\times\mathbb{S}^{d'-1}} \in L^\infty(\mathbb{R} \times \mathbb{S}^{d'-1})$. *By definition of* $\mathcal{G}^*$, *we have*

$$\mathcal{G}^*(\psi)(x) = \int_{\mathbb{S}^{d'-1}} 1\, d\sigma_{d'}(\theta) = 1, \qquad \forall x \in \mathbb{R}^d.$$

*That is,* $\mathcal{G}^*(1_{\mathbb{R}\times\mathbb{S}^{d'-1}}) = 1_{\mathbb{R}^d}$. *Now, let* $\mu \in \mathcal{M}(R^d)$. *We have that*

$$\mu(\mathbb{R}^d) = \int_{\mathbb{R}^d} 1_{\mathbb{R}^d}\, d\mu = \int_{\mathbb{R}^d} \mathcal{G}^*(1_{\mathbb{R}\times\mathbb{S}^{d'-1}})\, d\mu = \int_{\mathbb{R}\times\mathbb{S}^{d'-1}} 1_{\mathbb{R}\times\mathbb{S}^{d'-1}}\, d\mathcal{G}(\mu) = \mathcal{G}(\mu)(\mathbb{R} \times \mathbb{S}^{d'-1}). \tag{30}$$

*Let* $\mu^+$ *and* $\mu^-$ *are the positive and negative parts of* $\mu$, *that is,* $\mu = \mu^+ - \mu^-$, *where* $\mu^\pm \in \mathcal{M}_+(\mathbb{R}^d)$. *By property (2) we have that* $\mathcal{G}(\mu^\pm) \in \mathcal{M}_+(\mathbb{R} \times \mathbb{S}^{d'-1})$. *By linearity of* $\mathcal{G}$, *we have*

$$\mathcal{G}(\mu) = \mathcal{G}(\mu^+ - \mu^-) = \underbrace{\mathcal{G}(\mu^+)}_{\mathcal{M}_+} - \underbrace{\mathcal{G}(\mu^-)}_{\mathcal{M}_+},$$

*which by the uniqueness of the Hahn-Jordan decomposition theorem implies that the positive and negative parts of the measure* $\mathcal{G}(\mu)$ *are* $\mathcal{G}(\mu^+)$ *and* $\mathcal{G}(\mu^-)$, *respectively. That is,*

$$\mathcal{G}(\mu)^\pm = \mathcal{G}(\mu^\pm).$$

*Then,*

$$\begin{aligned}
\|\mathcal{G}(\mu)\|_{TV} &= \mathcal{G}(\mu)^+(\mathbb{R} \times \mathbb{S}^{d'-1}) + \mathcal{G}(\mu)^-(\mathbb{R} \times \mathbb{S}^{d'-1}) \\
&= \mathcal{G}(\mu^+)(\mathbb{R} \times \mathbb{S}^{d'-1}) + \mathcal{G}(\mu^-)(\mathbb{R} \times \mathbb{S}^{d'-1}) \\
&= \mu^+(\mathbb{R} \times \mathbb{S}^{d'-1}) + \mu^-(\mathbb{R} \times \mathbb{S}^{d'-1}) = \|\mu\|_{TV},
\end{aligned}$$

*that is,* $\|\mathcal{G}(\mu)\|_{TV} = \|\mu\|_{TV}$.

*(4) If* $\mu \in \mathcal{P}(\mathbb{R}^d)$, *combining the property* $\mathcal{G}(\mu) \in \mathcal{M}_+(\mathbb{R} \times \mathbb{S}^{d'-1})$ *with the property* $\|\mathcal{G}(\mu)\|_{TV} = \|\mu\|_{TV} = 1$, *we obtain that* $\mathcal{G}(\mu) \in \mathcal{P}(\mathbb{R} \times \mathbb{S}^{d'-1})$.

*(5) We will follow the ideas in (Bonneel et al., 2015)[Proposition 6]. For any test function $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$ we have*

$$\begin{aligned}
\int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d\mathcal{G}(\mu)_\theta(t)\, d\sigma_{d'}(\theta) &= \int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{G}(\mu)(t,\theta) \qquad (\textit{disintegration theorem}) \\
&= \int_{\mathbb{R}^d} \mathcal{G}^*(\psi)(x)\, d\mu(x) \qquad (\textit{definition of } \mathcal{G}(\mu)) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d'-1}} \psi(g(x,\theta),\theta)\, d\sigma_{d'}(\theta)\, d\mu(x) \qquad (\textit{definition of } \mathcal{G}^*) \\
&= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}^d} \psi(g(x,\theta),\theta)\, d\mu(x)\, d\sigma_{d'}(\theta) \\
&= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d(g(\cdot,\theta)_{\#}\mu)(t)\, d\sigma_{d'}(\theta),
\end{aligned}$$

*where the fourth equality follows from Fubini theorem and the fact $\psi$ is bounded and $\mu, \sigma_{d'}$ are finite measures; in the fifth equality, $t = g(x,\theta)$ and we use the definition of pushforward. By uniqueness of the disintegration of measures we have $\mathcal{G}(\mu)_\theta = g(\cdot,\theta)_{\#}\mu$. This also can be shown by choosing an arbitrary $\psi_0 \in C_0(\mathbb{R})$, defining $\psi(t,\theta) := \psi_0(t)$, $\forall (t,\theta) \in \mathbb{R} \times \mathbb{S}^{d'-1}$, so $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$, and therefore we have*

$$\begin{aligned}
\int_{\mathbb{R}} \psi_0(t)\, d\mathcal{G}(\mu)_\theta(t) &= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d\mathcal{G}(\mu)_\theta(t)\, d\sigma_{d'}(\theta) \\
&= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d(g(\cdot,\theta)_{\#}\mu)(t)\, d\sigma_{d'}(\theta) \\
&= \int_{\mathbb{R}} \psi_0(t)\, d(g(\cdot,\theta)_{\#}\mu)(t).
\end{aligned}$$

*(6) For $\mathcal{H}$, properties $(1) - (5)$ can be derived similarly.*

*(7) As in $(5)$, we will follow the ideas in (Bonneel et al., 2015)[Proposition 6]. For any test function $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$ we have*

$$\begin{aligned}
\int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d\mathcal{H}(\mu)_\theta(t)\, d\sigma_{d'}(\theta) &= \int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{H}(\mu)(t,\theta) \qquad (\textit{disintegration theorem}) \\
&= \int_{\mathbb{R}^d} \mathcal{H}^*(\psi)(x)\, d\mu(x) \qquad (\textit{definition of } \mathcal{H}(\mu)) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{S}^{d'-1}} \psi(g(x,\theta),\theta)\, d\sigma_{d'}(\theta)\, d\mu(x) \qquad (\textit{definition of } \mathcal{H}^*) \\
&= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}^d} \psi(\langle h(x),\theta\rangle,\theta)\, d\mu(x)\, d\sigma_{d'}(\theta) \qquad (\textit{Fubini's Theorem}) \\
&= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(\langle y,\theta\rangle,\theta)\, d(h_{\#}\mu)(y)\, d\sigma_{d'}(\theta) \qquad (\textit{definition of pushforward}) \\
&= \int_{\mathbb{S}^{d'-1}} \int_{\mathbb{R}} \psi(t,\theta)\, d(\theta_{\#}h_{\#}\mu)(t)\, d\sigma_{d'}(\theta) \qquad (\textit{definition of pushforward}).
\end{aligned}$$

*Thus, by the uniqueness of the disintegration of measures, we have $\mathcal{H}(\mu)_\theta = \theta_{\#}h_{\#}\mu = (\theta \circ h)_{\#}\mu$.*

*(8) Let $\mu \in \mathcal{M}(\mathbb{R}^d)$, and let $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$ be any test function. Following the ideas from (7), we have*

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{H}(\mu)(t,\theta) &= \int_{\mathbb{R}^d} \mathcal{H}^*(\psi)(x)\, d\mu(x) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d'-1}} \psi(\langle h(x),\theta\rangle,\theta)\, d\sigma_{d'}(\theta)\, d\mu(x) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{S}^{d'-1}} \psi(\langle y,\theta\rangle,\theta)\, d\sigma_{d'}(\theta)\, dh_{\#}\mu(y) \\ &= \int_{\mathbb{R}^d} (\mathcal{R}^*(\psi))(y)\, dh_{\#}\mu(y) \\ &= \int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{R}(h_{\#}\mu)(t,\theta),\end{aligned}$$

*where in the third equality $y = h(x)$ and we use the definition of pushforward; the first and fifth equation follows from the equations (28) and (5), respectively. Therefore, $\mathcal{H}(\mu) = \mathcal{R}(h_{\#}\mu)$.*

*(9) By Proposition 7 in (Bonneel et al., 2015), we have the mapping*

$$h_{\#}(\mu) \mapsto \mathcal{R}(h_{\#}(\mu)) = \mathcal{H}(\mu) \tag{31}$$

*is invertible. If $h$ is injective, we have $\mu \mapsto h_{\#}(\mu)$ is invertible. Thus, $\mu \mapsto \mathcal{H}(\mu)$ is invertible.*

□

## D. Stereographic Spherical Radon Transform and Stereographic Spherical Sliced Wasserstein Distance for Probability Measures

Based on the extended definition of GRT for probability measures, in this section, we formally introduce the "Stereographic Spherical Radon Transform" and the corresponding sliced Wasserstein distance, "Stereographic Spherical Sliced Wasserstein distance" for general probability measures.

Given $\mu \in \mathcal{M}(\mathbb{S}^d)$ and the stereographic projection $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$, which is a bijection, we have that $\phi_{\#}\mu$ is a measure defined in $\mathbb{R}^d$. The **generalized Stereopgraphic Spherical Radon transform**, of $\mu$ has two versions $\mathcal{S}_{\mathcal{G}}(\mu)$ and $\mathcal{S}_{\mathcal{H}}(\mu)$ defined as follows:

$$\mathcal{S}_{\mathcal{G}}(\mu) := \mathcal{G}(\phi_{\#}\mu) \qquad \text{and} \qquad \mathcal{S}_{\mathcal{H}}(\mu) := \mathcal{H}(\phi_{\#}\mu). \tag{32}$$

That is, from (28) and (29), we have that for each test function $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{G}}(\mu)(t,\theta) = \int_{\mathbb{R}^d} \mathcal{G}^*(\psi)(x)\, d\phi_{\#}\mu(x) = \int_{\mathbb{S}^d\setminus\{s_n\}} \mathcal{G}^*(\psi)(\phi(s)) d\mu(s), \tag{33}$$

$$\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{H}}(\mu)(t,\theta) = \int_{\mathbb{R}^d} \mathcal{H}^*(\psi)(x)\, d\phi_{\#}\mu(x) = \int_{\mathbb{S}^d\setminus\{s_n\}} \mathcal{H}^*(\psi)(\phi(s))\, d\mu(s). \tag{34}$$

**Proposition D.1** *Let $\mu \in \mathcal{M}(\mathbb{S}^d)$ that does not give mass to the North Pole $\{s_n\}$. The Stereographic Spherical Radon transforms defined in (32) have the following properties:*

*(1) $\mathcal{S}_{\mathcal{G}}(\mu), \mathcal{S}_{\mathcal{H}}(\mu) \in \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d'-1})$. In addition $\mathcal{S}_{\mathcal{G}}$ and $\mathcal{S}_{\mathcal{H}}$ preserves mass, and if $\mu$ is a positive measure, then $\mathcal{S}_{\mathcal{G}}(\mu), \mathcal{S}_{\mathcal{H}}(\mu)$ are positive measures too. Finally, if $\mu \in \mathcal{P}(\mathbb{S}^d \setminus \{s_n\})$, then $\mathcal{S}_{\mathcal{G}}(\mu), \mathcal{S}_{\mathcal{H}}(\mu)$ are probability measures defined on $\mathbb{R} \times \mathbb{S}^{d'-1}$.*

*(2) The disintegration theorem gives a unique $\mathcal{S}_{\mathcal{G}}(\mu)-$a.s. set of measures $(\mathcal{S}_{\mathcal{G}}(\mu)_\theta)_{\theta\in\mathbb{S}^{d'-1}} \subset \mathcal{M}(\mathbb{R})$, and a unique $\mathcal{S}_{\mathcal{H}}(\mu)-$a.s. set of measures $(\mathcal{S}_{\mathcal{H}}(\mu)_\theta)_{\theta\in\mathbb{S}^{d'-1}} \subset \mathcal{M}(\mathbb{R})$ such that for any $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$,*

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{G}}(\mu)(t,\theta) &= \int_{\mathbb{S}^{d'-1}}\int_{\mathbb{R}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{G}}(\mu)_\theta(t)\, d\sigma_{d'}(\theta), \qquad \textit{and} \\ \int_{\mathbb{R}\times\mathbb{S}^{d'-1}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{H}}(\mu)(t,\theta) &= \int_{\mathbb{S}^{d'-1}}\int_{\mathbb{R}} \psi(t,\theta)\, d\mathcal{S}_{\mathcal{H}}(\mu)_\theta(t)\, d\sigma_{d'}(\theta).\end{aligned} \tag{35}$$

*Then, it holds that*

$$\mathcal{S}_{\mathcal{G}}(\mu)_\theta = (g(\cdot,\theta)\circ\phi)_{\#}\mu, \qquad \text{and} \qquad \mathcal{S}_{\mathcal{H}}(\mu)_\theta = (\theta\circ h\circ\phi)_{\#}\mu, \tag{36}$$

*where we recall that the function $\theta$ is defined as in* (22)

*(3) $\mathcal{S}_{\mathcal{H}}$ is invertible.*

*Proof:*

(1) If $\mu \in \mathcal{M}(\mathbb{S}^d)$ is such that does not give mass to $\{s_n\}$, then $\phi_{\#}\mu$ is a measure defined in $\mathbb{R}^d$ with the same total mass. Thus, since $\mathcal{S}_{\mathcal{G}}(\mu) = \mathcal{G}(\phi_{\#}\mu)$, the conclusion follows from by (1)–(4) in Proposition (C.1). Similar to $\mathcal{S}_{\mathcal{H}}(\mu)$.

(2) By definition of $\mathcal{S}_{\mathcal{G}}$ and (5) in Proposition (C.1) we have

$$\mathcal{S}_{\mathcal{G}}(\mu)_\theta = g(\cdot,\theta)_{\#}(\phi_{\#}\mu) = g(\phi(\cdot),\theta)_{\#}\mu.$$

Similarly, by definition of $\mathcal{S}_{\mathcal{H}}$ and (7) in Proposition (C.1) we have

$$\mathcal{S}_{\mathcal{H}}(\mu)_\theta = (\theta\circ h)_{\#}\phi_{\#}\mu = (\theta\circ h\circ\phi)_{\#}\mu.$$

(3) Finally, since the stenographic projection $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ is invertible, then following composition

$$\begin{aligned}\mathcal{M}(\mathbb{S}^d\setminus\{s_n\}) &\to \mathcal{M}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}\times\mathbb{S}^{d'-1})\\ \mu &\longmapsto \phi_{\#}\mu \;\; \longmapsto \mathcal{H}(\phi_{\#}\mu) = \mathcal{S}_{\mathcal{H}}(\mu)\end{aligned}$$

is invertible as each mapping is invertible. Similarly,

$$\begin{aligned}\mathcal{P}(\mathbb{S}^d\setminus\{s_n\}) &\to \mathcal{P}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}\times\mathbb{S}^{d'-1})\\ \mu &\longmapsto \phi_{\#}\mu \;\; \longmapsto \mathcal{H}(\phi_{\#}\mu) = \mathcal{S}_{\mathcal{H}}(\mu)\end{aligned}$$

is invertible. Thus $\mu \mapsto \mathcal{S}_{\mathcal{H}}(\mu)$ is an invertible mapping either general measures or for probability measures.

□

**Remark D.2** *If $g(x,\theta) = \langle x,\theta\rangle$, then the generalized Radon transform $\mathcal{G}$ coincides with the classical Radon transform $\mathcal{R}$, and the corresponding Stereographic Spherical Radon transform coincides with the one defined in* (10)*:*

$$\mathcal{S}_{\mathcal{R}}(\mu) = \mathcal{R}(\phi_{\#}\mu)$$

*for any Radon measure $\mu \in \mathcal{M}(\mathbb{S}^d)$ that does not assign mass to $\{s_n\}$*

We introduce the notation $\phi_g^\theta, \phi_h^\theta : \mathbb{S}^d\backslash\{s_n\} \to \mathbb{R}^{d'}$

$$\phi_g^\theta(z) := g(\phi(z),\theta)\theta, \qquad \phi_h^\theta(z) := \langle h(\phi(z),\theta)\rangle\theta,$$

having range in the line generated by $\theta \in \mathbb{S}^{d'-1}$. The expressions (13) in (2) in the above Proposition can be written as

$$\mathcal{S}_{\mathcal{G}}(\mu)_\theta = \phi_{g\,\#}^\theta\mu, \qquad \mathcal{S}_{\mathcal{H}}(\mu)_\theta = \phi_{h\,\#}^\theta\mu$$

and we call them *slices* of the SSRT of $\mu$.

Given two measures $\mu_1,\mu_2 \in \mathcal{P}(\mathbb{S}^d)$, we introduce the formal definition of Stereographic Spherical Sliced Wasserstein (S3W) distance as follows:

$$\begin{aligned}S3W^p_{\mathcal{G},p}(\mu_1,\mu_2) &:= \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta, \mathcal{S}_{\mathcal{G}}(\mu_2)_\theta)\, d\sigma_{d'}(\theta)\\ &= \int_{\mathbb{S}^{d'-1}} W_p^p(\phi_{g\,\#}^\theta\mu_1, \phi_{g\,\#}^\theta\mu_2)\, d\sigma_{d'}(\theta) \qquad (37)\\ S3W^p_{\mathcal{H},p}(\mu_1,\mu_2) &:= \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{H}}(\mu_1)_\theta, \mathcal{S}_{\mathcal{H}}(\mu_2)_\theta)\, d\sigma_{d'}(\theta)\\ &= \int_{\mathbb{S}^{d'}} W_p^p(\phi_{h\,\#}^\theta\mu_1, \phi_{h\,\#}^\theta\mu_2)\, d\sigma_{d'}(\theta), \qquad (38)\end{aligned}$$

where $\sigma_{d'} \in \mathcal{P}(\mathbb{S}^{d'-1})$ is the uniform measure on $\mathbb{S}^{d'-1}$.

**Theorem D.3** *The above definitions $S3W_{\mathcal{G},p}(\cdot,\cdot)$ and $S3W_{\mathcal{H},p}(\cdot,\cdot)$ are well-defined. Furthermore, $S3W_{\mathcal{G},p}(\cdot,\cdot)$ is a pseudo-metric in $\mathcal{P}(\mathbb{S}^d\backslash\{s_n\})$, i.e., it is non-negative, symmetric and satisfies triangular inequality. In addition, $S3W_{\mathcal{H},p}(\cdot,\cdot)$ defines a metric in $\mathcal{P}(\mathbb{S}^d\setminus\{s_n\})$.*

*Proof:* Given probability measures $\mu_1,\mu_2\in\mathcal{P}(\mathbb{S}^d\setminus\{s_n\})$, by (1) in Proposition D.1 we have that $\mathcal{S}_{\mathcal{G}}(\mu_1),\mathcal{S}_{\mathcal{G}}(\mu_2)$ are probability measures defined on $\mathbb{R}\times\mathbb{S}^{d'-1}$. Thus, by (2) in Proposition D.1, $\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta$ are probability measure on $\mathbb{R}$. Then, $W_p^p(\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta)$ is well-defined and thus $S3W^p_{\mathcal{G},p}(\cdot,\cdot)$ is well defined. Analogously, $S3W^p_{\mathcal{H},p}(\cdot,\cdot)$ is well-defined.

Since the Wasserstein distance $W_p(\cdot,\cdot)$ in $\mathcal{P}(\mathbb{R})$ is non-negative and symmetric and $\sigma_{d'}$ is a positive measure we have that $S3W_{\mathcal{G},p}(\cdot,\cdot)$ and $S3W_{\mathcal{H},p}(\cdot,\cdot)$ are non-negative and symmetric.

For triangular inequality, let $\mu_1,\mu_2,\mu_3\in\mathcal{P}(\mathbb{S}^d\setminus\{s_n\})$,

$$\begin{aligned}
S3W_{\mathcal{G},p}(\mu_1,\mu_3) &= \left(\int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_3)_\theta)d\sigma_{d'}(\theta)\right)^{1/p}\\
&\le \left(\int_{\mathbb{S}^{d'-1}} \left(W_p(\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta)+W_p(\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_3)_\theta)\right)^p d\sigma_{d'}(\theta)\right)^{1/p}\\
&\le \left(\int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta)\right)^{1/p}+\left(\int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_3)_\theta)\right)^{1/p}\\
&= S3W_{\mathcal{G},p}(\mu_1,\mu_2)+S3W_{\mathcal{G},p}(\mu_2,\mu_3)
\end{aligned}$$

where the first inequality holds from the fact $W_p(\cdot,\cdot)$ in $\mathcal{P}(\mathbb{R})$ is a metric; the second inequality follows from Minkowski inequality in $L^p(\mathbb{S}^{d'-1})$. Then, $S3W_{\mathcal{G},p}(\cdot,\cdot)$ is a Pseudo-metric. Analogously for $S3W_{\mathcal{H},p}(\cdot,\cdot)$.

In addition, suppose $S3W_{\mathcal{H},p}(\mu_1,\mu_2)=0$, we have $\mathcal{S}_{\mathcal{H}}(\mu_1)_\theta=\mathcal{S}_{\mathcal{H}}(\mu_2)_\theta$ where the equality holds $\sigma_{d'}(\theta)-$a.s. Then, from (35) we can deduce that $\mathcal{S}_{\mathcal{H}}(\mu_1)=\mathcal{S}_{\mathcal{H}}(\mu_2)$. Since $\mathcal{S}_{\mathcal{H}}$ is invertible by Proposition (D.1), we have $\mu_1=\mu_2$. Thus, $S3W_{\mathcal{H},p}(\cdot,\cdot)$ is a metric.

□

Lastly, and similar to the max-sliced Wasserstein distance, here we define the max-S3W distance as:

$$\text{Max-}S3W^p_{\mathcal{G},p}(\mu_1,\mu_2):=\sup_{\theta\in\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_{\mathcal{G}}(\mu_1)_\theta,\mathcal{S}_{\mathcal{G}}(\mu_2)_\theta), \tag{39}$$

and similarly for $\mathcal{H}$ we define Max-$S3W^p_{\mathcal{H},p}$.

## E. Rotationally Invariant Stereographic Spherical Sliced Wasserstein Distances (RI-S3W)

In this section, we discuss the Rotationally Invariant $S3W$ distance (17):

$$RI\text{-}S3W_{\mathcal{G},p}(\mu,\nu):=\mathbb{E}_{R\sim\omega}[S3W_{\mathcal{G},p}(R_\#\mu,R_\#\nu)].$$

First, we describe the *uniform (Haar) distribution* in $\mathrm{O}(d+1)$ denoted as $\omega$. Precisely, we will describe how a $(d+1)\times(d+1)$ orthonormal matrix $R$ with rows $v_1,\ldots,v_{d+1}$ can be randomly generated as a sample of $\omega$. The procedure relies on the Gram-Schmidt algorithm:

First, we randomly select $v_1\in\mathbb{S}^d$, in particular, we denote it as $v_1\sim\sigma_{d+1}$, that is, we sample $v_1$ according to $\sigma_{d+1}$ which is the uniform distribution in $\mathbb{S}^d$.

Next, $v_2$ should selected from $\text{span}(v_1)^\perp=H(v_1,0)=\{v:\langle v_1,v\rangle=0\}$ and satisfy $\|v_2\|=1$, where $\text{span}(v_1)$ is the 1D sub-space in $\mathbb{R}^d$ spanned by $v_1$. Denote $\mathbb{S}^{d-1}(v_1):=\mathbb{S}^d\cap H(v_1,0)$. Note, it is essentially a $d-1$ sphere. We randomly select $v_2\in\mathbb{S}^{d-1}(v_1)$, denoted as $v_2\sim\sigma_d$ given $v_1$, where $\sigma_d$ denotes the uniform distribution in sphere $\mathbb{S}^{d-1}(v_1)$.

In the third step, similarly, let $\mathbb{S}^{d-2}(v_1,v_2):=\mathbb{S}^d\cap H(v_1,0)\cap H(v_2,0)$, and sample $v_3\sim\sigma_{d-1}$ given $v_1,v_2$, where $\sigma_{d-1}$ is the uniform distribution in the $d-2$ sphere $\mathbb{S}^{d-2}(v_1,v_2)$.

Continue this process recursively until step $d+1$. Note that at the last iteration, we have an orthonormal set of vectors $\{v_1, \dots, v_d\}$, and $\mathbb{S}^0(v_1, \dots v_d)$ is a 0-dimensional sphere, that is, it is a set of two points $\{u_1, u_2\}$. We randomly select $v_{d+1}$ as one of those two points in $\mathbb{S}^0(v_1, \dots v_d)$, denoting $v_{d+1} \sim \sigma_1$, where $\sigma_1$ is the uniform distribution in $\mathbb{S}^0(v_1, \dots, v_d)$ (that is, $\sigma_1 = 0.5\delta_{u_1} + 0.5\delta_{u_2}$).

Thus, the uniform probability measure $\omega$ in $\mathrm{O}(d+1)$ can be described in the following way: Consider an arbitrary Borel set $A \subset \mathrm{O}(d+1)$, and let $A_i := \{v \in \mathbb{S}^d : \ v \text{ is the } i\text{−th row of } R, \ R \in A\}$ for $i = 1, 2, \dots d+1$, then

$$\omega(A) := \int_{v_1 \in A_1, \dots, v_{d+1} \in A_{d+1}} d\sigma_1(v_{d+1}) \, d\sigma_2(v_d) \dots d\sigma_d(v_2) \, d\sigma_{d+1}(v_1). \tag{40}$$

where we are using the above notation, that is, $\sigma_{d+1}$ is the uniform distribution on $\mathbb{S}^d$, and $\sigma_{d+1-k}$ is the uniform distribution on $\mathbb{S}^d(v_1, \dots, v_k)$ for $| \leq k \leq d$ (i.e., $\sigma_d$ is the uniform distribution on $\mathbb{S}^d(v_1)$, $\dots$ , $\sigma_2$ is the uniform distribution on $\mathbb{S}^d(v_1, \dots, v_{d-1})$, $\sigma_1$ is the uniform distribution on $\mathbb{S}^d(v_1, \dots, v_d)$). In particular, $\omega$ is constant, in the sense that its density function is constant: Indeed, let $f_{\sigma_{d+1}}, \dots, f_{\sigma_2}$ denote the density functions of $\sigma_{d+1}, \dots, \sigma_2$ (which are all constant equal to the reciprocal of the "surface area" of the corresponding sphere). Then, let $f_\omega$ denote the density function of probability measure $\omega$, and so, for each $R = [v_1^T, v_2^T, \dots, v_{d+1}^T] \in \mathrm{O}(d+1)$

$$f_\omega(R) = f_{\sigma_{d+1}}(v_1) \cdot \dots \cdot f_{\sigma_2}(v_d) \cdot \sigma_1(v_{d+1}) = \frac{1}{2} \prod_{i=1}^{d} \frac{1}{\|\mathbb{S}^i\|_i} \tag{41}$$

Thus we call $\omega$ uniform distribution. We mention that it is also called a Haar measure for the special orthogonal group $\mathrm{SO}(d+1)$. Finally, we can normalize it so that we have a probability distribution.

Next, we discuss the proof of Theorem 3.4. We first introduce the following lemma.

**Lemma E.1** *Consider $s_n = [0, \dots, 0, 1] \in \mathbb{R}^{d+1}$ the North Pole of $\mathbb{S}^d$. Let $s_1 \in \mathbb{S}^d$ and $A = \{R \in \mathrm{O}(d+1) : \ Rs_1 = s_n\}$. Then $A = \{R \in \mathrm{O}(d+1) : \ \text{the } (d+1)\text{−th row of } R \text{ is } s_1\}$. Furthermore, $\omega(A) = 0$.*

*Proof:* Let $B = \{R \in \mathrm{O}(d+1) : \ \text{the } (d+1)\text{−th row of } R \text{ is } s_1\}$. It is straightforward to verify $B \subset A$. For the other direction, for any $R = [v_1^T, \dots, v_{d+1}^T] \in A$ we have

$$v_i^T s_1 = 0 \text{ if } i \neq d+1$$
$$v_{d+1}^T s_1 = 1 \text{ if } i = d+1$$

Thus, $Rs_1 = s_n$, and we have $A \subset B$.

Next, we will show $\omega(A) = 0$. Let $A' = \{R \in \mathrm{O}(d+1) : \ \text{1st row of } R \text{ is } s_1\}$. We have $\omega(A') = \omega(A)$. Then $A'_1 = \{s_1\}$. and (40), we have

$$\omega(A') \leq \int_{v_1 = s_1} \left( \int_{v_2, \dots v_d \in \mathbb{S}^d} d\sigma_1(v_{d+1}) \dots d\sigma_d(v_2) \right) d\sigma_{d+1}(v_1) = 0,$$

which completes the proof. □

*Proof:* [Proof of Theorem 3.4].

We first claim that (17) is well-defined.

Let $S(\mu, \nu) = \{s \in \mathbb{S}^d : \min(\mu(\{s\}), \nu(\{s\})) > 0\}$. Since $\mu, \nu$ are probability measures, we have that $S(\mu, \nu)$ is a finite set. For each $s \in S$, let $A(s) = \{R \in \mathrm{O}(d+1) : \ Rs = s_n\}$. Note that $S3W_{\mathcal{G},p}(R_\# \mu, R_\# \nu)$ is well-defined for each $R \in \mathrm{O}(d+1) \setminus A(\mu, \nu)$, where $A(\mu, \nu) = \bigcup_{s \in S(\mu,\nu)} A(s)$. By lemma E.1,

$$\omega(A(\mu, \nu)) = \omega \left( \bigcup_{s \in S(\mu,\nu)} A(s) \right) \leq \sum_{s \in S(\mu,\nu)} \omega(A(s)) = 0.$$

Thus, we have that

$$RI - S3W_{\mathcal{G},p}(\mu, \nu) := \mathbb{E}_{R \sim \omega}[S3W_{\mathcal{G},p}(R_\# \mu, R_\# \nu)] = \int_{R \in \mathrm{O}(d+1) \setminus A(\mu,\nu)} S3W_{\mathcal{G},p}(R_\# \mu, R_\# \nu) \, d\omega(R)$$

is well-defined.

Now, we will verify its properties.

By Theorem D.3, $S3W_{\mathcal{G},p}$ is non-negative, and symmetric. Thus $RI\text{-}S3W_{\mathcal{G},p}(\cdot,\cdot)$ is also non-negative and symmetric.

For the triangle inequality, let $\mu_1,\mu_2,\mu_3 \in \mathcal{P}(\mathbb{S}^d)$. Similarly to before, let $A(\mu_1,\mu_2,\mu_3) = \bigcup_{s\in S(\mu_1,\mu_2,\mu_3)} A(s)$. Since $\omega(A(\mu_1,\mu_2,\mu_3)) = 0$, we have

$$
\begin{aligned}
&RI - S3W_{\mathcal{G},p}(\mu_1,\mu_3) = \int_{\mathrm{O}(d+1)} S3W_{\mathcal{G},p}(R_\#\mu, R_\#\nu)\, d\omega(R) = \int_{\mathrm{SO}(d+1)\setminus A(\mu_1,\mu_2,\mu_3)} S3W_{\mathcal{G},p}(R_\#\mu, R_\#\nu)\, d\omega(R)\\
&\leq \int_{\mathrm{SO}(d+1)\setminus A(\mu_1,\mu_2,\mu_3)} \left(S3W_{\mathcal{G},p}(R_\#\mu_1, R_\#\mu_2) + S3W_{\mathcal{G},p}(R_\#\mu_2, R_\#\mu_3)\right)\, d\omega(R) \qquad \text{(Triangular inequality)}\\
&= \int_{\mathrm{SO}(d+1)\setminus A(\mu_1,\mu_2,\mu_3)} S3W_{\mathcal{G},p}(R_\#\mu_1, R_\#\mu_2)\, d\omega(R) + \int_{\mathrm{SO}(d+1)\setminus A(\mu_1,\mu_2,\mu_3)} S3W_{\mathcal{G},p}(R_\#\mu_2, R_\#\mu_3)\, d\omega(R)\\
&= \int_{\mathrm{SO}(d+1)} S3W_{\mathcal{G},p}(R_\#\mu_1, R_\#\mu_2)\, d\omega(R) + \int_{\mathrm{SO}(d+1)} S3W_{\mathcal{G},p}(R_\#\mu_2, R_\#\mu_3)\, d\omega(R)\\
&= RI - S3W_{\mathcal{G},p}(\mu_1,\mu_2) + RI - S3W_{\mathcal{G},p}(\mu_2,\mu_3)
\end{aligned}
$$

where the second and previous to last equalities hold from the fact $\omega(A(\mu_1,\mu_2,\mu_3)) = 0$. Thus, $RI\text{-}S3W_{\mathcal{G},p}(\cdot,\cdot)$ is a well-defined pseudo-metric.

Similarly, we can prove $RI\text{-}S3W_{\mathcal{H},p}(\cdot,\cdot)$ is a well-defined pseudo-metric.

Finally, it remains to show that $\mu = \nu$ if $RI\text{-}S3W_{\mathcal{H},p}(\mu,\nu) = 0$. If that holds, then for all $R \in \mathrm{SO}(d+1)\setminus A(\mu,\nu)$, we have $S3W_{\mathcal{H},p}(R_\#\mu, R_\#\nu) = 0$. Pick one of such rotations $R$. From Theorem D.3, and the fact $R_\#\mu, R_\#\nu \in \mathcal{P}(\mathbb{S}^d\setminus\{s_n\})$, we have $R_\#\mu = R_\#\nu$. Since

$$\mathbb{R}^{d+1} \ni v \mapsto Rv \in \mathbb{R}^{d+1}$$

is a one-by-one mapping, we have $\mu = \nu$.

□

## F. Funk Radon Transform and Stereographic Spherical Radon Transform

The Funk-Radon Transform (Quinto, 1982; Helgason et al., 2011; Quellmalz, 2017; 2020), which is also known as the Minkowski-Funk transform or the spherical Radon transform is a classical and significant mathematical tool used in integral geometry, with profound applications in image reconstruction and analysis. It plays a crucial role in fields like tomography, allowing for the reconstruction of images from projection data. In this section, we discuss the relation between the Funk Transform and the Stereographic Spherical Radon Transform.

First, we review some basic concepts about the Funk transform. In the sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$ centered at the origin with radius 1, every $(d-1)-$dimensional sub-sphere ("circle") is the intersection of $\mathbb{S}^d$ with a $d-$dimensional hyperplane, that is,

$$S(s,t) = \{s' \in \mathbb{S}^d \mid \langle s, s'\rangle = t\},$$

where $s \in \mathbb{S}^d$ is normal to the hyperplane and $t \in [-1,1]$ is the signed distance of the hyperplane to origin. For example, if $t = \pm 1$, $S(s,t)$ consists of only the singleton $\{\pm s\}$. As another example, when $t = 0$, for any $s \in \mathbb{S}^d$, the sub-sphere $S(s,0)$ is a "great circle" (a sub-sphere of dimension $d-1$ in $\mathbb{R}^{d+1}$ centered at the origin and with radius 1).

The *Spherical Transform* or *Spherical Mean Operator* (Quellmalz, 2017) of a function defines on $\mathbb{S}^d$ is formally defined as a function on $\mathbb{S}^d \times [-1,1]$ by the surface integral

$$\mathcal{U}f(s,t) := \frac{1}{\|S(s,t)\|_{d-1}} \int_{S(s,t)} f\, dS(s,t) = \frac{1}{\|S(s,t)\|_{d-1}} \int_{\mathbb{R}^{d+1}} f(s')\delta(t - \langle s, s'\rangle)\, ds', \quad (s,t) \in \mathbb{S}^d \times [-1,1], \quad (42)$$

where $\|S(s,t)\|_{d-1}$ denotes the "surface area" of the sub-sphere $S(s,t)$, and $dS(s,t)$ is the induced volume form on the surface $S(s,t)$. In the special case $d = 2$, given $f$ defined on $\mathbb{S}^2 \subset \mathbb{R}^3$, the integral in (42) is carried out with respect to the

arclength $dS$ of the circle $S(s,t)$, that is,

$$\mathcal{U}f(s,t) = \frac{1}{2\pi \cdot \text{radius}(S(s,t))} \int_{S(s,t)} f\, dS(s,t) := \frac{1}{2\pi \cdot \text{radius}(S(s,t))} \int_a^b f(r(u))\|r'(u)\|\, du,$$

where radius$(S(s,t))$ denotes the radius of the circle $S(s,t)$, $r(u)$ for $a \leq u \leq b$ is a parametrization of $S(s,t)$, and $\|r'(u)\|$ is the magnitude of the tangent vector $r'(u)$ to $S(s,t)$. Besides, in the special case $d = 3$, given $f$ defined on $\mathbb{S}^3 \subset \mathbb{R}^4$, we have that $S(s,t)$ is 2-dimensional sphere and we get

$$\mathcal{U}f(s,t) = \frac{1}{\frac{4}{3}\pi \cdot \text{radius}(S(s,t))^3} \int_{S(s,t)} f\, dS(s,t) := \frac{1}{\frac{4}{3}\pi \cdot \text{radius}(S(s,t))^3} \int\int f(r(u,v))\|r_u \times r_v\|\, du\, dv,$$

where radius$(S(s,t))$ denotes the radius of the sphere $S(s,t)$, $r(u,v)$ is a parametrization of $S(s,t)$, and $\|r_u \times r_v\|$ is the magnitude of the cross-product between the partial derivatives $r_u$ and $r_v$, which is a normal vector to $S(s,t)$.

The inversion of this Spherical Mean Operator is an over-determined problem (for instance, $\mathcal{U}f(s,t) = f(s)$ for all $s \in \mathbb{S}^d$). In application, one has access restrictions of $\mathcal{U}f(s,t)$ to specific sub-spheres $S(s,t)$, and this yields different transforms.

One important example arises when we restrict $\mathcal{U}f(s,t)$ to "great circles" $S(s,0)$ getting the classical Funk-Radon transform $\mathcal{F}_1$, that is,

$$\mathcal{F}_1 f(s) := \mathcal{U}f(s,0), \qquad s \in \mathbb{S}^d. \tag{43}$$

Another example, studied in (Abouelaz & Daher, 1993; Gindikin et al., 1994a; Helgason et al., 2011) arises when considering the restriction to the family of sub-spheres containing the North Pole $s_n = [0,\ldots,0,1] \in \mathbb{R}^{d+1}$.

Similarly as in (Quellmalz, 2017; 2020), we will interested in a generalized Funck-Radon transform given by the following restriction of $\mathcal{U}$. Let $s = [s_1,\ldots,s_{d+1}] \in \mathbb{S}^d \subset \mathbb{R}^{d+1}$, and $\xi \in [-1,1]$. Consider the point $[0,\ldots,0,\xi s_{d+1}] \in \mathbb{R}^{d+1}$ located in the same positive axis as the North Pole $s_n = [0,\ldots,0,1]$ and inside the sphere $\mathbb{S}^d$[3]. Then, we restrict $\mathcal{U}f$ to the integrals over the intersections of the sphere $\mathbb{S}^d$ with hyperplanes containing the point $[0,\ldots,0,\xi s_{d+1}]$, that is,

$$\zeta_\xi^s := S(s,\xi s_{d+1}) = \{s' \in \mathbb{S}^d |\, \langle s,s'\rangle = \xi s_{d+1}\}, \qquad \xi \in [-1,1],\ s \in \mathbb{S}^d.$$

Then, the generalized Funk-Radon transform is defined by

$$\mathcal{U}_\xi f(s) := \mathcal{U}f(s,\xi s_{d+1}) = \frac{1}{\|\zeta_\xi^s\|_{d-1}} \int_{\zeta_\xi^s} f\, d\zeta_\xi^s = \frac{1}{\|\zeta_\xi^s\|_{d-1}} \int_{\mathbb{R}^{d+1}} f(s')\delta(\xi s_{d+1} - \langle s,s'\rangle)\, ds', \tag{44}$$

where $\|\zeta_\xi^s\|_{d-1}$ is the volume ("surface area") of sub-sphere (circle, if $d = 2$) $\zeta_\xi^s$ and $d\zeta_\xi^s$ is the integration over the surface $\zeta_\xi^s$ (path, if $d = 2$).

**Remark F.1** *Note, the generalized Funk-Radon transform has been further generalized by (Rubin, 2022) under the name "spherical slice transform." In particular, choose $\alpha \in \mathbb{R}^{d+1}$ and let $\mathcal{T}(\alpha,k)$ denote the set of all $k$-dimensional affine planes that pass through $\alpha$ and intersect $\mathbb{S}^n$:*

$$\mathcal{T}(\alpha,k) = \{H(a) = \alpha + H : H \text{ is hyperplane}, \dim(H) = k, H(a) \cap \mathbb{S}^d \neq \emptyset\}.$$

*Pick $H \in \mathcal{T}(\alpha,k)$, then $H \cap \mathbb{S}^d$ is a $k-1$ dimensional sub-circle (or a point). The spherical slice transform of function $f \in L_1(\mathbb{S}^d)$ is defined as:*

$$\mathcal{SS}_\alpha f(H) = \int_{H\cap\mathbb{S}^d} f(s) dH \cap \mathbb{S}^d, \qquad \forall H \in \mathcal{T}(\alpha,k). \tag{45}$$

*When $k = d$, pick $s \in \mathbb{S}^d$ and $\xi \in [-1,1]$, we define affine plane*

$$H(s,\xi s_n) = \xi s_n + s^\perp,$$

[3] In (Quellmalz, 2017; 2020), $\xi \in [0,1)$ and so $[0,\ldots,0,\xi s_{d+1}]$ lies strictly inside the sphere.

*we have* $H \in \mathcal{T}(\xi s_n, d)$ *and* $\zeta_x i^s = H(s) \cap \mathbb{S}^d$. *Thus*

$$\mathcal{SS}f(H(s, \xi s_n)) = U_\xi(f)(s).$$

*That is, the generalized Funk transform* (43) *is a special case of the spherical slice transform* (45).

As said before, the goal of this section is to discuss the relationship between a generalized Funk-Radon transform (44) and the Stereographic Radon transform. For doing that let us first consider the *stereographic projection* $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ given by the formula

$$\phi(s) = \frac{s[1:d]}{1 - s_{d+1}}, \qquad \forall s = [s_1, \ldots, s_d, s_{d+1}] \in \mathbb{S}^d, \tag{46}$$

where $s[1:d] := [s_1, \ldots, s_d] \in \mathbb{R}^d$, which corresponds to the intersection with the equator plane $x_{d+1} = 0$, rather than (1) which corresponds to projection onto the plane $x_{d+1} = -1$ (which has a factor of 2 in (46)).

It is straightforward to verify that the inverse of $\phi$, denoted as $\phi^{-1}$, is as follows:

$$\phi^{-1}(x) = [(1 - s_{d+1})x, s_{d+1}], \quad \text{where } s_{d+1} = \frac{\|x\|^2 - 1}{\|x\|^2 + 1}, \qquad \forall x \in \mathbb{R}^d. \tag{47}$$

See Figures 6, 7 and 8.

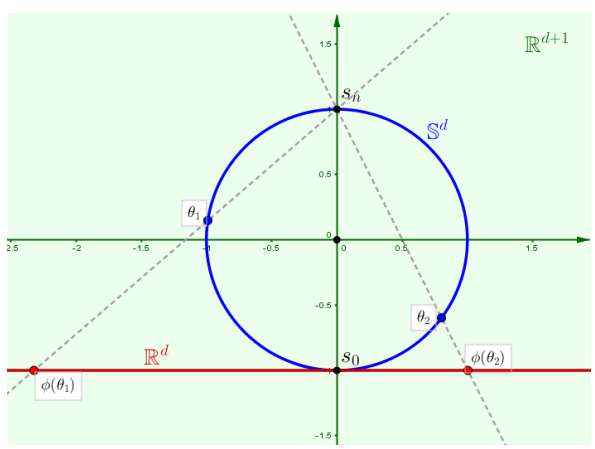



*Figure 6.* Visualization for $d = 1$ of the *stereographic projection* $\phi$ defined by the formula (1). The ambient space $\mathbb{R}^{d+1}$ is depicted in green, while the unit sphere (circle) $\mathbb{S}^d$ is in blue. The North and South Poles are labeled as $s_n$ and $s_0$, respectively. The projected space $\mathbb{R}^d = \phi(\mathbb{S}^d \setminus s_n)$ is highlighted in red, corresponding to the plane (line) in $\mathbb{R}^{d+1}$ defined by all points with the last coordinate equal to $-1$. Two points, $\theta_1$ and $\theta_2$ (blue dots), from $\mathbb{S}^d$ are projected through $\phi$ obtaining the points $\phi(\theta_1)$ and $\phi(\theta_2)$ (red dots).

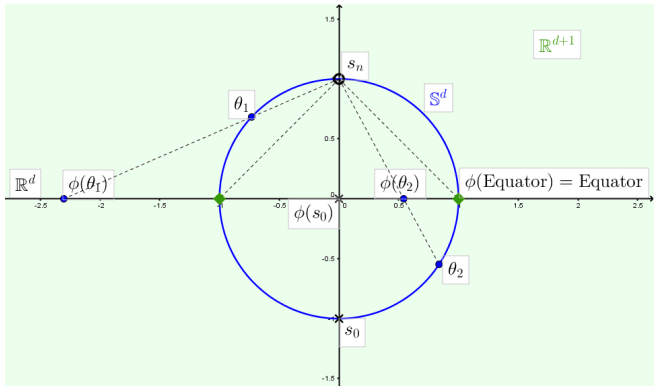



*Figure 7.* Visualization for $d = 1$ of the *stereographic projection* $\phi$ defined by the formula (46). The ambient space $\mathbb{R}^{d+1}$ is depicted in green, while the unit sphere (circle) $\mathbb{S}^d$ is in blue. The North and South Poles are labeled as $s_n$ and $s_0$, respectively. The projected space $\mathbb{R}^d = \phi(\mathbb{S}^d \setminus s_n)$ coincides with the hyperplane $x_{d+1} = 0$, which in the figure is nothing but the horizontal axis. Two points, $\theta_1$ and $\theta_2$ (blue dots), from $\mathbb{S}^d$ are projected through $\phi$ obtaining the points $\phi(\theta_1)$ and $\phi(\theta_2)$ (red dots). The South Pole is projected to the origin in $\mathbb{R}^d$ and the points in the Equator are fixed points for this stereographic projection.

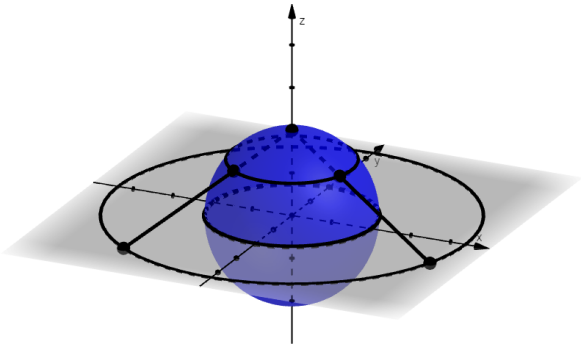



*Figure 8.* Visualization for $d = 2$ of the *stereographic projection* $\phi$ defined by the formula (46). The ambient space is the 3D space $\mathbb{R}^3$. The sphere $\mathbb{S}^2$ is depicted in blue. The projected space $\mathbb{R}^2 = \phi(\mathbb{S}^2 \setminus s_n)$ is depicted in grey. The points in the Equator are fixed points for this stereographic projection, and circles parallel to the Equator are mapped to circles in the plane as shown in the plot: If the height of the circle is positive, its projection is a circle with radius greater than 1 as the case depicted in the figure; if the height is negative, its projection is a circle with radius smaller than 1.

When we set $\xi = 1$ we have the following relations between the sub-spheres $\zeta_1^s$ and the hyperplanes $H(t, \theta)$ through the stereographic projection $\phi$.

**Proposition F.2** *Consider the sub-sphere (circle) $\zeta_1^s = \{s' \in \mathbb{S}^d | \langle s', s\rangle = s_{d+1}\}$, where $s = [s_1, \dots, s_d, s_{d+1}] \in \mathbb{S}^d$ is such that $s \neq s_n = [0, \dots, 0, 1]$. Then, $\phi(\zeta_1^s)$ is a hyperplane in $\mathbb{R}^d$, precisely,*

$$\phi(\zeta_1^s) = \left\{x \in \mathbb{R}^d \mid \langle x, \theta\rangle = t\right\},$$

*where $\theta = \frac{s[1:d]}{\|s[1:d]\|}$ and $t = \frac{s_{d+1}}{\sqrt{1-s_{d+1}^2}}$.*

*Conversely, given hyperplane $H(t, \theta) = \{x \in \mathbb{R}^d \mid \langle x, \theta\rangle = t\}$ for $\theta \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}$, we have*

$$\phi^{-1}(H(t, \theta)) = \zeta_1^s,$$

*where $s = [\sqrt{1 - s_{d+1}^2}\theta, s_{d+1}]$ with $s_{d+1} = sign(t)\frac{t^2}{t^2+1}$.*

*Proof:* Let $s \in \mathbb{S}^d$, and consider $s' \in \zeta_1^s$. Let $x' = \phi(s') = \frac{s'[1:d]}{1-s'_{d+1}}$. Then, we can write $s' = [(1 - s'_{d+1})x', s'_{d+1}]$. Besides, since

$$s_1^2 + \dots + s_d^2 + s_{d+1}^2 = 1,$$

by defining $\theta = \frac{s[1:d]}{\|s[1:d]\|}$, we can write $s = \left[\sqrt{1-s_{d+1}^2}\theta, s_{d+1}\right]$. Then, we have that

$$\begin{aligned} s_{d+1} &= \langle s', s\rangle \\ &= s'_{d+1}s_{d+1} + \langle (1-s'_{d+1})x', \sqrt{1-s_{d+1}^2}\theta\rangle \\ &= s'_{d+1}s_{d+1} + (1-s'_{d+1})\sqrt{1-s_{d+1}^2}\langle x', \theta\rangle. \end{aligned} \tag{48}$$

That is,

$$\langle x', \theta\rangle = \frac{s_{d+1}}{\sqrt{1-s_{d+1}^2}}.$$

Thus,

$$\phi(\zeta_1^s) = \left\{ x \in \mathbb{R}^d \mid \langle x, \theta\rangle = \frac{s_{d+1}}{\sqrt{1-s_{d+1}^2}} \right\}.$$

For the converse, given $\theta \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}$, consider $x' \in H(t, \theta)$. By using (47), let

$$s' = \phi^{-1}(x) = [(1-s'_{d+1})x, s'_{d+1}], \qquad \forall x \in \mathbb{R}^d$$

where $s'_{d+1} = \frac{\|x'\|^2-1}{\|x'\|^2+1}$.

We set $s = [\sqrt{1-s_{d+1}^2}\theta, s_{d+1}]$ with $s_{d+1} = \text{sign}(t)\sqrt{\frac{t^2}{t^2+1}}$. So, we have that $s \in \mathbb{S}^d$ because

$$(1-s_{d+1}^2)\underbrace{\theta^2}_{=1} + s_{d+1}^2 = 1,$$

and also we have that $t = \frac{s_{d+1}}{\sqrt{1-s_{d+1}^2}}$. Thus, we have that

$$\langle x', \theta\rangle = t = \frac{s_{d+1}}{\sqrt{1-s_{d+1}^2}}$$

and by reversing the above proof (see (48)), we have that it implies that

$$s_{d+1} = \langle s', s\rangle,$$

and so $s' \in \zeta_1^s$. Combining with the fact $\phi$ is invertible, we have

$$\phi^{-1}(H(t, \theta)) = \zeta_1^s.$$

□

**Lemma F.3** *Let $s \in \mathbb{S}^d$ different from the North Pole $s_n = [0, \dots, 0, 1]$, the South Pole $s_0 = [0, \dots, 0, -1]$, and consider $S_0$ the unique great circle (meridian) that passes through the North Pole $s_n$, the South Pole $s_0$, and the point $s$. For each $\hat{s} \in S_0$, we define the functions*

$$\theta(\hat{s}) := \frac{\hat{s}[1:d]}{\|\hat{s}[1:d]\|}, \qquad t(\hat{s}) := \frac{s_{d+1}}{\sqrt{1-s_{d+1}^2}}.$$

*For all $\hat{s} \in S_0$, $\theta(\hat{s}) = \theta(s)$. In addition $t : S_0 \setminus \{s_n\} \to \mathbb{R}$ is a bijection.*

*Proof:* Without loss of generality, we can suppose $s = [\cos\alpha, 0, \dots, 0, \sin\alpha]$ for some $\alpha \in (-\frac{3}{2}\pi, \frac{1}{2}\pi)$. Thus, $\theta(s) = [1, 0, \dots 0]$, and

$$S_0 = \left\{ \hat{s}_\alpha = [\cos\hat{\alpha}, 0, \dots, 0, \sin\hat{\alpha}] : \hat{\alpha} \in [-\frac{3}{2}\pi, \frac{1}{2}\pi] \right\}.$$

Pick $\hat{s}_\alpha \in S_0 \setminus \{s_n\}$, we have $\theta(\hat{s}_\alpha) = [1, 0, \ldots, 0]$. In addition, $t(\hat{s}_\alpha) = \frac{\sin\alpha}{|\cos\alpha|}$ is a one by one mapping. □

Notice that for $\mathbb{S}^2 \subset \mathbb{R}^3$, if we visualize $\mathbb{S}^2$ as the Earth and we pick the point $s \in \mathbb{S}^2$ to be the Royal Observatory in the town of Greenwich, London, England, then $S_0$ is the *Greenwich meridian* (or "prime meridian").

Finally, we discuss the relation between the variant Funk transform defined in (44) with $\xi = 1$ (i.e., $\mathcal{U}_1$) and the *Stereographic Spherical Radon transform* defined in (10) as

$$\mathcal{S}_\mathcal{R}(\mu) = \mathcal{R}(\phi_\# \mu)$$

for any Radon measure $\mu \in \mathcal{M}(\mathbb{S}^d)$ that does not assign mass to $\{s_n\}$. In particular, we can define the Stereographic Spherical Radon transform for $L^1$ functions $f : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}$, as they can be viewed as densities of continuous measures.

Indeed, let $\mu$ be the corresponding measure for $f$, that is,

$$\int_{\mathbb{S}^d \setminus \{s_n\}} \psi_0(s) f(s)\, d\sigma_{d+1}(s) = \int_{\mathbb{S}^d \setminus \{s_n\}} \psi_0(s)\, d\mu(s), \qquad \forall \psi_0 \in C(\mathbb{S}^d),$$

where $\sigma_{d+1}$ is the uniform measure on $\mathbb{S}^d \subset \mathbb{R}^{d+1}$.

Let $\hat{f}$ denote the density of $\phi_\# \mu$. By definition of push forward measure, we have:

$$\int_{\mathbb{S}^d \setminus \{s_n\}} \psi(\phi(s)) f(s) d\sigma_{d+1}(s) = \int_{\mathbb{S}^d \setminus \{s_n\}} \psi(\phi(s)) d\mu(s) = \int_{\mathbb{R}^d} \psi(x) d\phi_\# \mu(x) = \int_{\mathbb{R}^d} \psi(x) \hat{f}(x) dx, \tag{49}$$

for all test functions $\psi \in C_0(\mathbb{R}^d)$ that decay to zero. In short, we usually write $d\mu = f d\sigma_{d+1}$, and $d\phi_\# \mu = \hat{f} dx$, where $dx$ represents the Lebesgue measure in $\mathbb{R}^d$.

Thus, $\mathcal{S}_\mathcal{R}(f)$ is defined as

$$\mathcal{S}_\mathcal{R}(f)(t, \theta) = \mathcal{R}(\hat{f})(t, \theta) = \int_{\mathbb{R}^d} \hat{f}(x) \delta(t - \langle x, \theta \rangle)\, dx = \int_{H(t,\theta)} \hat{f}\, dH(t, \theta), \tag{50}$$

where $dH(t, \theta)$ is the surface/line integral over the hyperplane/plane/line $H(t, \theta)$.

**Proposition F.4** *Let $f \in L^1(\mathbb{S}^d \setminus \{s_n\})$, and let $s \in \mathbb{S}^d$. Then,*

$$\|\zeta_1^s\|_{d-1} \mathcal{U}_1 f(s) = \mathcal{S}_\mathcal{R} f(t, \theta),$$

*where*

$$\theta = \frac{s[1:d]}{\|s[1:d]\|}, \qquad t = \frac{s_{d+1}}{\sqrt{1 - s_{d+1}^2}}.$$

*Proof:* For convenience, we can extend $f$ to $\mathbb{S}^d$ by setting $f(s_n) = 0$.

Without loss of generality, we suppose $s = [\cos\theta, 0, \ldots, 0, \sin\theta]$ where $\theta \in (-\frac{3}{2}\pi, \frac{1}{2}\pi)$.

As before, let $\mu$ be the measure having density $f$ and let $\hat{f}$ denote the density of $\phi_\# \mu$. Since $\phi$ is invertible, it is straightforward to verify that $(\phi^{-1})_\# (\phi_\#)(\mu) = \mu$, that is,

$$\int_{\mathbb{R}^d} \psi_0(\phi^{-1}(x)) \hat{f}(x) dx = \int_{\mathbb{S}^d} \psi_0(s) f(s) d\sigma_{d+1}(s), \qquad \forall \psi_0 \in C(\mathbb{S}^d) \tag{51}$$

In short, $d\mu = f d\sigma_{d+1}$ and $d\phi_\# \mu = \hat{f} dx$.

Let $S_0$ denote the unique great circle (*meridian*) that passes through the North Pole $s_n = [0, \ldots, 0, 1]$, the South Pole $s_0 = [0, \ldots, 0, -1]$ and the point $s$. It is straightforward to verify

$$\mathbb{S}^d = \bigcup_{\hat{s} \in S_0} \zeta_1^{\hat{s}}$$

and that for any distinct $\hat{s}_1, \hat{s}_2 \in S_0$, $\zeta_1^{\hat{s}_1} \cap \zeta_1^{\hat{s}_2} = \{s_n\}$.

Similarly,

$$\mathbb{R}^d = \bigcup_{\hat{t} \in \mathbb{R}} H(\hat{t}, \theta)$$

and for any distinct $\hat{t}_1, \hat{t}_2 \in \mathbb{R}$, $H(\hat{t}_1, \theta) \cap H(\hat{t}_2, \theta) = \emptyset$.

Thus, we can "slice" and integrate in the following way:

$$\int_{\mathbb{S}^d} \psi_0 \, d\sigma_{d+1} = \int_{\hat{s} \in S_0} \int_{\zeta_1^{\hat{s}}} \psi_0 \, d\zeta_1^{\hat{s}} \, dS_0(\hat{s}), \tag{52}$$

where $d\zeta_1^{\hat{s}}$ denotes the surface/path integral over $\zeta_1^{\hat{s}}$, and $dS_0$ denotes the surface/path integral over $S_0$.

Also, by using Proposition F.2, we can "slice" and integrate in the following way:

$$\int_{\mathbb{R}^d} \psi(x) \, dx = \int_{\hat{s} \in \mathbb{R}} \int_{H(t(\hat{s}), \theta(\hat{s}))} \psi \, dH(t(\hat{s}), \theta(\hat{s})) \, d\hat{s}, \tag{53}$$

where for each $\hat{s} \in S_0$ we define $\theta(\hat{s}) := \frac{\hat{s}[1:d]}{\|\hat{s}[1:d]\|}$, $t(\hat{s}) := \frac{\hat{s}_{d+1}}{\sqrt{1-\hat{s}_{d+1}^2}}$, and where $dH(t(\hat{s}), \theta(\hat{s}))$ denotes the surface/line integral over $H(t(\hat{s}), \theta(\hat{s}))$, and $d\hat{s}$ denotes the integral over $\mathbb{R} = \phi(S_0 \setminus \{s_n\})$. (Notice that if $d = 2$, (53) is nothing but the Fubini-Tonelli expression $\int_{\mathbb{R}^2} \phi(z) \, dz = \int_{\mathbb{R}} \int_{\mathbb{R}} \psi(x, y) \, dx \, dy$, where $dz$ is the Lebesgue measure on $\mathbb{R}^2$ and $dx$, $dy$ the Lebesgue measure on $\mathbb{R}$.)

Refer to Figure 9 for an illustrative visualization.

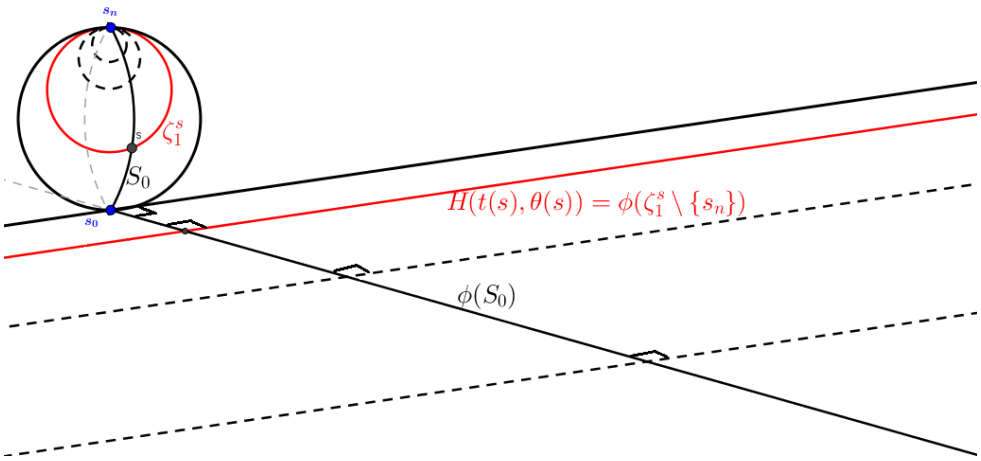



*Figure 9.* Representation of a unit sphere $\mathbb{S}^2$ in $\mathbb{R}^3$, featuring the North and South Poles ($s_n$ and $s_0$ in blue). The sketch includes five great circles passing through a point $s$: four intersecting $s_n$ and one meridian labeled as $S_0$. The stereographic projection of these five great circles is depicted as straight lines, which lie on the same plane. For the stereographic projection $\phi$ we used formula (1). The four circles passing through both $s_n$ and $s$ project onto parallel lines. One such circle is highlighted in red and labeled as $\zeta_1^s$, with its corresponding projection also in red, labeled as $H(t(s), \theta(s))$. These four parallel lines are perpendicular to the line labeled as $\phi(S_0)$, representing the stereographic projection of the meridian $S_0$.

By using Lemma F.3 combined with identity (51), together with with (52) and (53), we have

$$\begin{array}{ccc} \displaystyle\int_{\mathbb{R}^d} \psi_0(\phi^{-1}(x)) \hat{f}(x) dx & = & \displaystyle\int_{\mathbb{S}^d} \psi_0(s) f(s) d\sigma_{d+1}(s) \\ \| & & \| \\ \int_{\hat{s} \in \mathbb{R}} \int_{H(t(\hat{s}), \theta(\hat{s}))} \psi_0 \circ \phi^{-1} \, \hat{f} \, dH(t(\hat{s}), \theta(\hat{s})) \, d\hat{s} & & \int_{\hat{s} \in S_0} \int_{\zeta_1^{\hat{s}}} \psi_0 \, f \, d\zeta_1^{\hat{s}} \, dS_0(\hat{s}) \end{array}$$

Since the above equality holds for all $\psi_0 \in C(\mathbb{S}^d)$, by setting $\psi_0 \equiv 1$ we obtain

$$S_{\mathcal{R}} f(t,\theta) = \int_{H(t,\theta)} \hat{f}\, dH(t,\theta) = \int_{\zeta_1^s} f\, d\zeta_1^s = \|\zeta_1^s\|_{d-1}\, \mathcal{U}_1 f(s)$$

and we complete the proof. □

## G. Related Work: Vertical Sliced Transform, Semi-Circle Transform, and Related Wasserstein Distances

This section introduces the **vertical sliced transform** (Gindikin et al., 1994b), **semi-circle transform** (Groemer, 1998), and the related optimal transport distances (Quellmalz et al., 2023).

### G.1. Vertical Slice Transform and Vertical Sliced Wasserstein Distance

We first introduce the vertical slice transform technique introduced in (Quellmalz et al., 2023), which can only be defined in $\mathbb{S}^2$ space. Next, we extend it into $\mathbb{S}^d$ for $d \geq 2$. Finally, we discuss the vertical slice of Wasserstein distance and its relation to the classical slice of Wasserstein distance.

**Parametrization in $\mathbb{S}^2$.**

For each point $s \in \mathbb{S}^2$, we can parameterize it as

$$s = [\cos\alpha \sin\beta, \sin\alpha \sin\beta, \cos\beta], \tag{54}$$

where $\alpha \in [0, 2\pi), \beta \in [0, \pi]$.

When $s$ is the north (or south) pole, i.e., $s = [0,0,1]$ (or $s = [0,0,-1]$), the above parameterization in particular is not uniquely determined. In this case, we set $\alpha = 0, \beta = 0$ (or $\alpha = 0, \beta = \pi$). Otherwise, $\alpha, \beta$ are uniquely determined.

In particular, $\alpha$ is called the azimuth angle, denoted as $\alpha = \mathrm{azi}(s)$, and $\beta$ is called the zenith angle, denoted as $\beta = \mathrm{zen}(s)$.

In summary, the following mapping is bijective:

$$\begin{aligned} \Phi : ((0,2\pi) \times (0,\pi)) \cup (\{0\} \times \{0,\pi\}) &\to \mathbb{S}^2 \\ (\alpha,\beta) &\mapsto \Phi(\alpha,\beta) = s = [\cos\alpha\sin\beta, \sin\alpha\sin\beta, \cos\beta]. \end{aligned} \tag{55}$$

**Vertical Slice Transform and Vertical Sliced Wasserstein Distance.**

The equator in $\mathbb{S}^2$ can be defined by

$$\mathcal{E} := \{\Phi(\alpha, 0) = [\cos\alpha, \sin\alpha, 0] : \alpha \in [0, 2\pi)\}. \tag{56}$$

Each $\theta = \Phi(\alpha, 0) \in \mathcal{E}$ and $t \in [-1, 1]$ can determine a "vertical" circle

$$\mathcal{VS}(\alpha, t) := \mathcal{VS}(\theta, t) := \{s \in \mathbb{S}^2 : \langle \theta, s\rangle = t\}, \tag{57}$$

characterizing a circle centered at $t\theta$ with radius $\sqrt{1-t^2}$. Thus $\|\mathcal{VS}(\theta,t)\|_1 = 2\pi\sqrt{1-t^2}$.

Given $f \in \mathcal{L}_1(\mathbb{S}^2)$, the **vertical slice transform** $\mathcal{V}$ of $f$, denoted as $\mathcal{V}f$, is a $L_1$ function defined on $\mathcal{E} \times [-1,1]$:

$$\mathcal{V}f(\theta,t) := \begin{cases} \frac{1}{\|\mathcal{VS}(\theta,t)\|_1} \int_{\mathcal{VS}(\theta,t)} f(s) d\mathcal{VS}(\theta,t)(s) = \frac{1}{2\pi\sqrt{1-t^2}} \int_{\mathbb{S}^2} f(s)\delta(t - \langle\theta, s\rangle) ds & \text{if } t \in (-1,1) \\ f(\theta) & \text{if } t = 1 \\ f(-\theta) & \text{if } t = -1 \end{cases} \tag{58}$$

Its adjoint operator $\mathcal{V}^* : L^1(\mathcal{E} \times [-1,1]) \to L^1(\mathbb{S}^2)$ has closed form:

$$\mathcal{V}^* g(s) = \frac{1}{2\pi} \int_{\mathcal{E}} g(\theta, \langle\theta, s\rangle) d\mathcal{E}. \tag{59}$$

Thus, we can extend the vertical transform to any (positive Radon) measure as follows.

For each $\mu \in \mathcal{M}_+(\mathbb{S}^2)$, $\mathcal{V}(\mu)$ is defined by:

$$\int_{\mathcal{E}\times[-1,1]} \psi(\theta,t)d\mathcal{V}\mu(\theta,t) = \int_{\mathbb{S}^2} \mathcal{V}^*\psi(s)(s)d\mu(s), \qquad \forall \psi \in C_0(\mathbb{E}\times[-1,1]). \tag{60}$$

Note, $\mathcal{V}(\mu)$ can be equivalently defined as

$$\int \psi(\theta,t)d\mathcal{V}(\mu)(\theta,t) = \int_{\mathcal{E}}\int_{-1}^{1} \psi(\theta,t)d\mathcal{V}(\mu)_\theta(t)d\sigma_{\mathcal{E}}(\theta) \tag{61}$$

where $\sigma_{\mathcal{E}}$ is the uniformly probability measure defined on $\mathcal{E}$ and

$$\mathcal{V}(\mu)_\theta = \langle\theta,\cdot\rangle_{\#}\mu, \forall\theta\in\mathcal{E}.$$

is the integration of $\mathcal{V}\mu$ with respect to $\sigma_{\mathcal{E}}$.

In the discrete case, $\mu = \sum_{i=1}^n p_i\delta_{x_i}$, the above definition becomes $\mathcal{V}(\mu)_\theta = \sum_{i=1}^n p_i\delta_{\langle\theta,x_i\rangle}$.

Given $\mu,\nu\in\mathcal{P}(\mathbb{S}^2)$, the **vertical sliced Wasserstein distance** is defined by

$$VSW_p^p(\mu,\nu) := \int_{\mathcal{E}} W_p^p(\mathcal{V}(\mu)_\theta,\mathcal{V}(\nu)_\theta)d\sigma_{\mathcal{E}}(\theta) \tag{62}$$

$$\approx \frac{1}{N}\sum_{t=1}^N W_p^p(\langle\theta,\cdot\rangle_{\#}\mu,\langle\theta,\cdot\rangle_{\#}\nu) \tag{63}$$

where (63) is the Monte Carlo approximation and $\theta_1,\dots\theta_N$ are uniformly selected from $\mathcal{E}$.

By [Theorem 3.7, (Quellmalz et al., 2023)], the vertical transform (60) (or (61)) is injective. Thus, the above VSW problem defines a metric.

**Relation with Classical Radon Transform.**

We refer to (5) and (24) for the classical Radon transform for a measure $\mu\in\mathcal{P}(\mathbb{R}^3)$.

In addition, we have

$$\mathcal{R}(\mu)_\theta = \mathcal{V}(\mu)_\theta = \langle\theta,\cdot\rangle_{\#}\mu, \qquad \forall\theta\in\mathcal{E}\subset\mathbb{S}^2.$$

That is, the restricted transformed measure of $\mu$ under $\mathcal{R}$ and $\mathcal{V}$ coincide when $\theta\in\mathcal{E}$.

Recall that the Sliced Wasserstein (SW) distance is defined by

$$SW_p^p(\mu,\nu) := \int_{\mathbb{S}^2} W_p^p(\mathcal{R}(\mu)_\theta,\mathcal{R}(\nu)_\theta)d\sigma_{\mathbb{S}^2}(\theta) \tag{64}$$

$$\approx \frac{1}{N}\sum_{i=1}^N W_p^p(\langle\theta,\cdot\rangle_{\#}\mu,\langle\theta,\cdot\rangle_{\#}\nu), \tag{65}$$

where (65) is the Monte Carlo approximation of the SW problem and $\theta_1,\dots\theta_N$ are uniformly sampled from $\mathbb{S}^2$.

Comparing (63) and (65), we observe that when $\theta_1,\dots\theta_N$ are sampled from $\mathcal{E}$, (63) and (65) coincide.

**Generalization of Vertical Sliced Wasserstein in $\mathbb{S}^d$.**

It is natural to extend the vertical sliced transform given by (58), (60), (61) to $\mathbb{S}^d$ for $d\geq 2$.

Let

$$\mathcal{E} = \{s\in\mathbb{S}^d : s_{d+1} = 0\},$$

denote the equator.

Choose $(\theta, t) \in \mathcal{E} \times [-1, 1]$, the $d-1$ dimensional vertical sphere $\mathcal{VS}(\theta, t)$ is defined as

$$\mathcal{VS}(\theta, t) := \{s \in \mathbb{S}^d : \langle \theta, s \rangle = t\}.$$

Choose $f \in L_1(\mathbb{S}^d)$, $\mathcal{V}f$ is defined by

$$\mathcal{V}f(\theta, t) := \frac{1}{\|\mathcal{E}\|_{d-1}} \int_{VS(\theta,t)} f(s) dVS(\theta, t) = \frac{1}{\|\mathcal{E}\|_{d-1}} \int_{\mathbb{S}^d} f(s)\delta(\langle \theta, s \rangle - t) ds. \tag{66}$$

It is straightforward to verify $\mathcal{V}f \in L_1(\mathcal{E} \times [-1, 1])$.

In addition, its adjoint operator $\mathcal{V}^*$ becomes

$$\mathcal{V}^* g(s) = \frac{1}{\|\mathcal{E}\|_{d-1}} \int_{\mathcal{E}} g(\theta, \langle \theta, s \rangle) d\theta, \forall g \in L_1(\mathcal{E} \times [-1, 1]). \tag{67}$$

And similarly, for each $\mu \in \mathcal{M}(\mathbb{S}^d)$, vertical transformed measure $\mathcal{V}\mu$ can be defined by (60), or equivalently,

$$\int_{\mathcal{E}\times[-1,1]} g(\theta, t) d\mathcal{V}\mu(\theta, t) = \int_{\mathcal{E}} \int_{[-1,1]} g(\theta, t) d\mathcal{V}(\mu)_\theta(t) d\sigma_{\mathcal{E}}(\theta), \tag{68}$$

where $\mathcal{V}(\mu)_\theta = \langle \theta, \cdot \rangle_{\#}\mu$ is the integration of $d\mathcal{V}\mu$ with respect to $\sigma_{\mathcal{E}}$.

Thus, vertical sliced Wasserstein (62), (63) can be extended to the space $\mathbb{S}^d$.

We provide additional results in Appendix I.2.

### G.2. Semi-Circle Transform and Spherical Sliced Wasserstein Distance

We first introduce the semi-circle transform and then discuss its relation to spherical sliced Wasserstein (SSW).

The semi-circle can be defined by the following three equivalent formulations.

**Formulation 1 (Quellmalz et al., 2023)**: In $\mathbb{S}^2$, we define

$$\Psi(\alpha, \beta, \gamma) := \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{69}$$

Choose $s^* = \Phi(\alpha^*, \beta^*)$ (defined in (55)), $\xi \in [0, 2\pi)$, the semi-circle is defined by

$$\mathcal{SC}(s^*, \xi) = \{s \in \mathbb{S}^2 : \text{azi}(\Psi(\alpha^*, \beta^*, 0)^\top s) = \xi\} \cup \{\pm\Phi(\alpha^*, \beta^*)\}. \tag{70}$$

**Formulation 2 (Groemer, 1998)**:

In $\mathbb{S}^d$, where $d \geq 2$, given $u, v \in \mathbb{S}^d$ with $u \perp v$, the semi-circle is defined by

$$\mathcal{SC}(u, v) := \{s \in \mathbb{S}^d : s \perp u, \langle s, v \rangle \geq 0\}. \tag{71}$$

**Formulation 3 (Bonet et al., 2023c)**.

In $\mathbb{S}^d$, choose $U \in V_2(d) := \{U' \in \mathbb{R}^{d\times 2} : (U')^\top U' = I_2\}$, and let $U^\perp$ denote the perpendicular space of Range$(U)$. In addition, $U$ can determine a great circle:

$$\mathcal{S}_U := \{s \in \mathbb{R}^{d+1} : U^T s = 0\} \cap \mathbb{S}^d.$$

The projection from $\mathbb{S}^d$ to $\mathcal{S}_U$ admits the following closed form:

$$P_U(s) = \arg\min_{s' \in \mathcal{S}_U} d_{\mathbb{S}}(s, s') = \frac{UU^T s}{\|U^T s\|}, \forall s \in \mathbb{S}^d \setminus U^\perp. \tag{72}$$

The great circle $\mathcal{S}_U$ can be regarded as a $\mathbb{S}^1$ circle, and we can use $t \in [0, 2\pi)$ to represent each point in $\mathcal{S}_U$. Thus, the above projection mapping can be rewritten as $s \mapsto \frac{U^T s}{\|U^T s\|}$. The new formulation induces the semi-circle:

$$\mathcal{SC}(U,t) = \left\{ s \in \mathbb{S}^d : \frac{U^T s}{\|U^T s\|} = t \right\} \cup U^\perp \tag{73}$$

When $d = 2$, the above three formulations of semi-circle are equivalent. When $d \geq 2$, the second and third formulations are equivalent. For convenience, we select formulation 3.

**Semi-Circle Transform**. Given $f \in L_1(\mathbb{S}^d)$, we will introduce the semi-circle transform of $f$. Note, in (Quellmalz et al., 2023; Hielscher et al., 2018), this transform is called the *unnormalized semi-circle transform*; in (Bonet et al., 2023c), this transform is called the *spherical Radon transform*; and in (Groemer, 1998), it is called the *hemispherical transform*:

$$\mathcal{SC}(f)(U,t) := \int_{\mathcal{SC}(U,t)} f(s) d\mathcal{SC}(U,t). \tag{74}$$

Similar to the Radon transform (5) and the vertical slice transform (60), the above definition can be extended into $\mathcal{M}(\mathbb{S}^d)$. The corresponding Wasserstein problem is the spherical sliced Wasserstein problem (Bonet et al., 2023a):

$$SSW(\mu,\nu) := \int_{V_2(d)} W_p^p(\mathcal{SC}(\mu)_U, \mathcal{SC}(\nu)_U) d\sigma_{V_2(d)}(U), \tag{75}$$

where $\mathcal{SC}(\mu)_U$ is the integration of $\mathcal{SC}(\mu)$ with respect to the uniformly measure $\sigma_{V_2(d)}$. In addition, with $\sigma_{V_2(d)} - a.s.$, we have

$$\mathcal{SC}(\mu)_U = \left( \frac{U^T \cdot}{\|U^T \cdot\|} \right)_\# \mu.$$

## H. Distance Distortion

In $\mathbb{R}^d$ we consider the Euclidean distance and in $\mathbb{S}^d$ the *great circle distance* $d_{\mathbb{S}^d}(\cdot,\cdot)$. While in Euclidean spaces the distance between two points is the length of a straight segment between them, in the sphere, we measure the distance between two points as the length of the shortest path that lies in the sphere and connects them. In particular, for two angles in the unit circle $\theta_1, \theta_2 \in [0, 2\pi)$

$$d_{\mathbb{S}^1}(\theta_1, \theta_2) = \min\{|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|\}.$$

In general, the distance between $s_1, s_2 \in \mathbb{S}^d$ given by the arclength of the shortest path can be expressed as

$$d_{\mathbb{S}^d}(s_1, s_2) := \arccos(\langle s_1, s_2 \rangle).$$

We point out that in the stereographic projection, significant distortion can occur. For instance,

$$\text{as } d_{\mathbb{S}^d}(s_1, s_2) \longrightarrow 0, \text{ we might have } \|\phi(s_1) - \phi(s_2)\| \longrightarrow \infty.$$

See Figure 10 for a visualization.

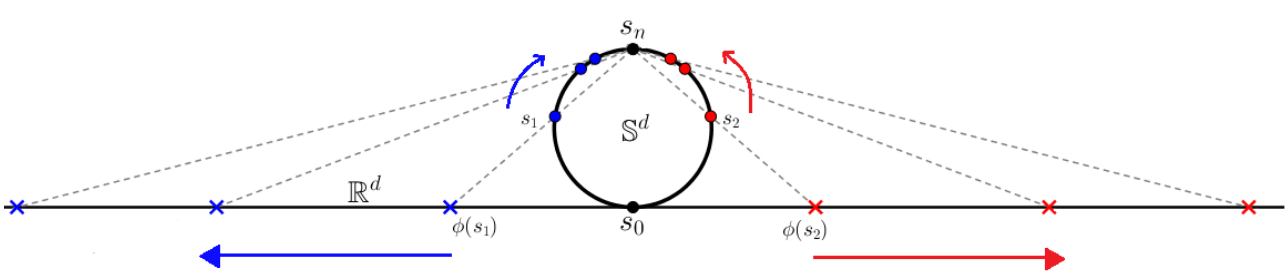



*Figure 10.* Illustration of the distortion phenomenon in one-dimensional space ($d = 1$) as a result of applying the stereographic projection with formula (1). As the points $s_1$ (blue) and $s_2$ (red) are brought closer within the sphere (circle), their respective stereographic projections, depicted as crosses, diverge from each other.

In this section, we aim to construct an injective function $h : \mathbb{R}^d \to \mathbb{R}^d$ such that the following holds:

$$\|h(\phi(s_1)) - h(\phi(s_2))\| \approx d_{\mathbb{S}^d}(s_1, s_2).$$

We define $h_1 : \mathbb{R}^d \to \mathbb{R}^d$ by

$$h_1(x) := \arccos(-s_{d+1})\frac{x}{\|x\|}, \quad \text{where } s_{d+1} = \frac{\|x\|^2 - 1}{\|x\|^2 + 1}. \tag{76}$$

**Proposition H.1** *Let $s, s_1, s_2 \in \mathbb{S}^d \setminus \{s_n\}$, and let $s_0 = [0, \dots, 0, -1]^T \in \mathbb{S}^d$ denote the South Pole. Consider $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ the stereographic projection as in (46) with inverse (47), and also consider the function $h_1 : \mathbb{R}^d \to \mathbb{R}^d$ defined by (76).*

*(1) $h_1(\phi(s)) = \angle(s, s_0)\frac{s[1:d]}{\|s[1:d]\|}$, and thus $\|h_1(\phi(s_1)) - h_1(\phi(s_2))\| \leq 2\pi$, where $\angle(s, s_0) = \arccos(\langle s, s_0\rangle) \in [0, \pi]$ denotes the angle between $s$ and $s_0$.*

*(2) If $s_1, s_2, s_0$ are in the same great circle, denoted as $S$, let $d_S$ denote the circle distance in $S$, then*

$$d_{\mathbb{S}^d}(s_1, s_2) = d_S(s_1, s_2) = \min\{\|h_1(\phi(s_1)) - h_1(\phi(s_2)\|, 2\pi - \|h_1(\phi(s_1)) - h_1(\phi(s_2)\|\} \tag{77}$$

*Proof:*

(1) Note that

$$h_1(\phi(s)) = \arccos(-s_{d+1})\frac{\phi(s)}{\|\phi(s)\|} = \arccos(\langle s, s_0\rangle)\frac{s[1:d]}{\|s[1:d]\|} = \angle(s, s_0)\underbrace{\frac{s[1:d]}{\|s[1:d]\|}}_{\leq 1}, \qquad \forall s \in \mathbb{S}^d \setminus \{s_n\} \subset \mathbb{R}^{d+1}.$$

Thus, we have that

$$\|h_1(\phi(s_1)) - h_1(\phi(s_2))\| \leq \angle(s_1, s_0) + \angle(s_2, s_0) \leq \pi + \pi = 2\pi.$$

(2) Without loss of generality, we suppose that the great circle $S$ (meridian) is of the form

$$S = \left\{ s = [\cos\theta, 0, \dots, 0, \sin\theta] : \theta \in [-\frac{3}{2}\pi, \frac{1}{2}\pi] \right\}$$

(where we go through the circle clockwise having $-\pi/2$ in the middle). Thus, we assume that $S$ passes through $s_0 = [0, \dots, 0, -1] = [\cos(-\frac{\pi}{2}), 0, \dots, 0, \sin(-\frac{\pi}{2})]$, $s_1 = [\cos\theta_1, 0, \dots 0, \sin\theta_1]$, and $s_2 = [\cos\theta_2, 0, \dots, 0, \sin\theta_2]$ in $S \setminus \{s_n\}$ with $\theta_1, \theta_2 \in (-\frac{3}{2}\pi, \frac{1}{2}\pi)$. It follows that

$$h_1(\phi(s_1)) = \angle(s_1, s_0)\frac{s_1[1:d]}{\|s_1[1:d]\|} = (\theta_1 + \frac{\pi}{2})[1, 0, \dots, 0]$$

and similarly, $h_1(\phi(s_2)) = (\theta_2 + \frac{\pi}{2})[1, 0, \dots, 0]$. Thus, the third term in (77) becomes

$$\min\{\|h_1(\phi(s_1)) - h_1(\phi(s_2))\|, 2\pi - \|h_1(\phi(s_1)) - h_1(\phi(s_2))\|\} = \min\{|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|\}.$$

For the first term in (77), we have

$$d_{\mathbb{S}^d}(s_1, s_2) = \arccos(\cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2) = \arccos(\cos(\theta_1 - \theta_2)) = \min\{|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|\}.$$

For the second term, we have $d_S(s_1, s_2) = \min(|\theta_1 - \theta_2|, 2\pi - |\theta_1 - \theta_2|)$.

And we are done.

□

**Remark H.2** *As a corollary from part 1 in Proposition H.1, we notice that $h_1$ preserves angles with the vertex at the origin in the projected space. (See angle $\beta$ in Figure 13.)*

*In addition, when $d = 2$, we observe that $h_1(\phi)$ is similar to the* ***Lambert azimuthal equal-area projection*** *(Bradawl, 2003):*

$$LP(s) = \sqrt{\frac{2}{1 - s[3]}} s[1:2]. \tag{78}$$

*The function $LP$ projects points in $\mathbb{S}^2$ into $\mathbb{R}^2$, while preserving the area in $\mathbb{S}^2$, i.e.*

$$\|A\|_2 = \|LP(A)\|_2.$$

*Meanwhile, $h_1 \circ \phi$ aims to preserve distance.*

*In addition, if we write $s = [\cos\alpha\cos\beta, \sin\alpha\cos\beta, \sin\beta]$, then*

$$\begin{aligned} h_1(\phi(s)) &= \angle(s, s_0)\frac{s[1:2]}{\|s[1:2]\|} \\ &= (\beta + \frac{\pi}{2})[\cos\alpha, \sin\alpha] \\ &= -\underbrace{(\beta + \frac{\pi}{2})[\sin(\alpha - \pi/2), -\cos(\alpha - \pi/2)]}_{A} \end{aligned} \tag{79}$$

*where $A$ is exactly the* ***azimuthal equidistant projection*** *(Snyder, 1997) of $s$ centered at south pole $[0, 0, -1]$.*

*Thus, $h_1 \circ \phi$ and the azimuthal equidistance projection are equivalent.*

**Conjecture H.3** *Let $s, s_1, s_2 \in \mathbb{S}^d \setminus \{s_n\}$, and let $s_0 = [0, 0, \ldots, -1]^T \in \mathbb{S}^d$ denote the South Pole. Consider $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ the stereographic projection as in (46) with inverse (47), and also consider the function $h_1 : \mathbb{R}^d \to \mathbb{R}^d$ defined by (76). Let $\mathrm{O}(d+1)$ denote the set of all $(d+1) \times (d+1)$ orthonormal matrices, then*

$$\begin{aligned} d_{\mathbb{S}^d}(s_1, s_2) &= \min_{R \in \mathrm{O}(d+1)} \min\{\|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\|, 2\pi - \|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\|\} \\ &= \min_{R \in \mathrm{O}(d+1)} \|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\| \end{aligned} \tag{80}$$

Numerically, we have shown that statement holds (see Figure 12). We will leave the theoretical proof for a future study.

Here we provide the intermediate results: Proposition H.4 and Proposition H.10.

**Proposition H.4** *Let $s, s_1, s_2 \in \mathbb{S}^d \setminus \{s_n\}$, and let $s_0 = [0, 0, \ldots, -1]^T \in \mathbb{S}^d$ denote the South Pole. Consider $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ the stereographic projection as in (46) with inverse (47), and also consider the function $h_1 : \mathbb{R}^d \to \mathbb{R}^d$ defined by (76). There exists $R^* \in \mathrm{SO}(d+1)$ such that*

$$\begin{aligned} d_{\mathbb{S}^d}(s_1, s_2) &= \min\{\|h_1(\phi(R^*s_1)) - h_1(\phi(R^*s_2))\|, 2\pi - \|h_1(\phi(R^*s_1)) - h_1(\phi(R^*s_2))\|\} \\ &\geq \min_{R \in \mathrm{O}(d+1)} \min\{\|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\|, 2\pi - \|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\|\} \end{aligned}$$

*Proof:*

Choose $s_1, s_2 \in \mathbb{S}^d$ two different points. Let $\zeta$ denote the shortest path from $s_1$ to $s_2$.

Then $\zeta$ can be parametrized as a function. In particular, there exists a mapping $\zeta : [0, 1] \to \mathbb{S}^d$, $\zeta(0) = s_1, \zeta(1) = s_2$, $\zeta$ is differentiable, $\|\zeta'(t)\|$ is constant, and the range of $\zeta$ is exactly the shortest path. By definition of the great circle distance $d_{\mathbb{S}^d}$, it is exactly the length of $\zeta$. That is,

$$d_{\mathbb{S}^d}(s_1, s_2) = |\zeta| = \int_0^1 \sqrt{(\zeta'(t))^2} dt.$$

Furthermore, $\zeta$ lies in a great circle, denoted as $S$.

Let $s_0' = \zeta(t_0)$ denote the middle point between $s_1, s_2$ (for an appropriate $t_0 \in (0,1)$),

By the Gram-Schmidt algorithm, there exists $R_0 \in \mathrm{SO}(d+1)$ such that $R_0 s_0' = s_0$ (for example, first set $r_1 = -s_0'$, then construct orthonormal vectors $r_2, \dots, r_{d+1}$, and finally define the matrix $R_0$ having rows $r_1, r_2, \dots, r_{d+1}$).

It holds that

$$d_{\mathbb{S}^d}(s_1, s_2) = d_{\mathbb{S}^d}(R_0 s_1, R_0 s_2),$$

i.e., $d_{\mathbb{S}^d}$ is invariant under rotations and reflections.

Besides, note that $R_0 s_1, R_0 s_2, R_0 s_0'$ are in the same great circle. Thus, by using that $d_{\mathbb{S}^2}$ in invariant under orthogonal transformations and by using part (2) in Proposition H.1, we have

$$\begin{aligned} d_{\mathbb{S}^d}(s_1, s_2) = d_{\mathbb{S}^d}(R_0 s_1, R_0 s_2) &= \min\{\|h_1(\phi(R_0 s_1)) - h_1(\phi(R_0 s_2))\|, 2\pi - \|h_1(\phi(R_0 s_1)) - h_1(\phi(R_0 s_2))\|\} \\ &= \|h_1(\phi(R_0 s_1)) - h_1(\phi(R_0 s_2))\| \end{aligned}$$

where the last equality holds because $s_0 = R_0 s_0'$ is a point between $R_0 s_1, R_0 s_2$.

In particular,

$$\min_{R \in \mathrm{O}(d+1)} \min\{\|h_1(\phi(R s_1)) - h_1(\phi(R s_2)\|, 2\pi - \|h_1(\phi(R s_1)) - h_1(\phi(R s_2)\|\} \leq d_{\mathbb{S}^d}(s_1, s_2)$$

□

For the other direction, we first recall that $d_{\mathbb{S}^d}$ is invariant under orthogonal transformations:

$$d_{\mathbb{S}^d}(s_1, s_2) = d_{\mathbb{S}^d}(R s_1, R s_2) \qquad \forall R \in \mathrm{O}(d+1).$$

In particular, every $R' \in \mathrm{O}(d+1)$ is a minimizer of: $d_{\mathbb{S}^d}(s_1, s_1) = d_{\mathbb{S}^d}(R' s_1, R' s_1) = \inf_{R \in \mathrm{O}(d+1)} d_{\mathbb{S}^d}(R s_1, R s_1)$. By using part (2) in Proposition H.1 we have that for all $R^*$ such that $R^* s_1$, $R^* s_2$, and $R^* s_0$ are in the same great circle it holds

$$d_{\mathbb{S}^d}(s_1, s_2) = d_{\mathbb{S}^d}(R^* s_1, R^* s_2) = \min\{\|h_1(\phi(R^* s_1)) - h_1(\phi(R^* s_2))\|, 2\pi - \|h_!(\phi(R^* s_1)) - h_1(\phi(R^* s_2))\|\} \quad (81)$$

However, we want to minimize over all the possible orthogonal matrices. At this point, the proof is more involved, and to prove this part we will need a series of auxiliary lemmas:

**Lemma H.5** *Given $s_1, s_2 \in \mathbb{S}^d$, with $(s_1)_{d+1} = (s_2)_{d+1}$ (i.e., $s_1$ and $s_2$ are in the same latitude), then*

$$d_{\mathbb{S}^d}(s_1, s_2) \leq \min(2\pi - \|h(\phi(s_1)) - h_1(\phi(s_2))\|, \|h_1(\phi(s_1)) - h_1(\phi(s_2))\|).$$

*Proof:* Suppose $(s_1)_{d+1} = (s_2)_{d+1} = \sin\alpha$ for some $\alpha \in [-\frac{\pi}{2}, \frac{\pi}{2})$. Let $\beta = \angle(s_1[1:d], s_2[1:d]) \in [0, \pi]$. We have

$$\begin{aligned} d_{\mathbb{S}^d}(s_1, s_2) &= \arccos(\langle s_1[1:d], s_2[1:d]\rangle + (s_1)_{d+1}(s_2)_{d+1}) \\ &= \arccos\left(\cos^2(\alpha)\cos(\beta) + \sin^2(\alpha)\right) \\ \|h_1(\phi(s_1)) - h_1(\phi(s_2))\| &= \sqrt{\|h_1(\phi(s_1)\|^2 + \|h_1(\phi(s_2)\|^2 - 2\|h_1(\phi(s_1))\|\|h_1(\phi(s_2))\|\cos(\angle(s_1[1:d], s_2[1:d]))} \\ &= \left(\alpha + \frac{\pi}{2}\right)\sqrt{2 - 2\cos(\beta)} \\ &= \left(\alpha + \frac{\pi}{2}\right) 2\sin\left(\frac{\beta}{2}\right) \end{aligned}$$

Let

$$\begin{aligned} L(\alpha, \beta) &= \|h_1(\phi(s_1)) - h_1(\phi(s_2))\| - d_{\mathbb{S}^d}(s_1, s_2) \\ &= 2(\alpha + \frac{\pi}{2})\sin\left(\frac{\beta}{2}\right) - \arccos\left(\cos^2(\alpha)\cos(\beta) + \sin^2(\alpha)\right) \end{aligned} \quad (82)$$

In hat follows we will show that

$$L(\alpha,\beta) \geq 0 \qquad \text{if } \alpha \in [-\frac{\pi}{2}, \frac{\pi}{2}),\ \beta \in [0,\pi].$$

First, if $\alpha = -\frac{\pi}{2}$ or $\beta = 0$, we have $L = 0$.

It remains to consider the case $\alpha \in (-\frac{\pi}{2}, \frac{\pi}{2})$, $\beta \in (0,\pi]$.

The procedure will be the following: We fix $\alpha \in (-\frac{\pi}{2}, \frac{\pi}{2})$ and consider the function $L(\alpha,\cdot)$ on the variable $\beta$.

We have:

$$\frac{d}{d\beta}L = \left(\alpha + \frac{\pi}{2}\right)\cos\left(\frac{\beta}{2}\right) + \frac{-\cos^2(\alpha)\sin(\beta)}{\sqrt{1-(\cos^2(\alpha)\cos(\beta)+\sin^2(\alpha))^2}} \tag{83}$$

We set $\frac{d}{d\beta}L = 0$, and obtain

$$\left(\alpha + \frac{\pi}{2}\right)^2 \cos^2\left(\frac{\beta}{2}\right) = \frac{\cos^4(\alpha)\sin^2(\beta)}{1-(\cos^2(\alpha)\cos(\beta)+\sin^2(\alpha))^2} \tag{84}$$

Let $A = \cos(\beta) = 2\cos^2\left(\frac{\beta}{2}\right) - 1 = 1 - 2\sin^2(\frac{\beta}{2})$, $B = \cos^2(\alpha) = \sin^2(\alpha') = 1 - \sin^2(\alpha)$, $C = (\alpha + \frac{\pi}{2})^2 = (\alpha')^2$, where $\alpha' = \alpha + \frac{\pi}{2} \in (0,\pi)$. The above equation (84) becomes:

$$\begin{aligned}
\frac{1}{2}C(A+1) &= \frac{B^2(1-A^2)}{1-(AB+1-B)^2} \\
&= \frac{B^2(1-A)(1+A)}{1-(A^2B^2+2AB-2AB^2+1-2B+B^2)} \\
&= \frac{B^2(1-A)(1+A)}{-A^2B^2-2AB+2AB^2+2B-B^2} \\
&= \frac{B(1-A)(1+A)}{-A^2B-2A+2AB+2-B}
\end{aligned}$$

thus, $C(-A^2B - 2A + 2AB + 2 - B) = 2B(1-A)$, or equivalently,

$$(-CB)A^2 + 2(BC - C + B)A + (2C - 2B - BC) = 0 \tag{85}$$

Denote by $a = -CB, b = 2(BC - C + B), c = (2C - 2B - BC)$ the coefficients of the above quadratic expression (85) as function of $A$: We have

$$\Delta^2 = b^2 - 4ac = 4(B-C)^2 \geq 0$$

and thus, we obtained real roots. Consider the following cases:

- Case 1: $a = 0$. Thus, $B = 0$ or $C = 0$, and so $\alpha = -\pi/2$, which was already analyzed.
- Case 2: $a \neq 0$. Since $B, C \geq 0$, we have that $a < 0$.

Besides, the two roots $A_1, A_2$ of (85) are given by

$$A_1 = \frac{BC + 2(B-C)}{BC} \qquad \text{and} \qquad A_2 = 1$$

Since $A = \cos(\beta) \in (-1,1)$ as $\beta \in (0,\pi)$, once we fix $\alpha$, there exists at most one $\beta^*$ such that $\frac{d}{d\beta}(\alpha,\beta^*)L = 0$. In fact, such $\beta^*$ exists if and only if $-1 < A_1 < 1$.

In addition, we have that

$$\begin{aligned}
\frac{d}{d\beta}L(\alpha,0) &= \lim_{\beta\searrow 0}\frac{d}{d\beta}L = (\alpha + \frac{\pi}{2}) > 0 \\
L(\alpha,0) &= 0 \\
L(\alpha,\pi) &= \begin{cases} 0 & \text{if } \alpha \in (-\frac{\pi}{2}, 0] \\ 4\alpha & \text{if } \alpha \in [0, \frac{\pi}{2}) \end{cases} \geq 0
\end{aligned}$$

We have the following:

If $\cos(\beta^*) = A_1 \in (-1,1)$, for some $\beta^* \in (0,\pi)$, we have $\frac{d}{d\beta}L(\alpha,\beta) \geq 0, \forall \beta \in [0,\beta^*]$, thus $L(\alpha,\beta)$ is an increasing function on $[0,\beta^*]$. Since $L(\alpha,0) > 0$, we have $L(\alpha,\beta) \geq 0$ for all $\beta \in [0,\beta^*]$.

On $(\beta^*,\pi]$, $\frac{dL}{d\beta} \neq 0$. By the Intermediate Value Theorem, we have $\frac{dL}{d\beta} < 0$. Thus, $L$ is a decreasing function here. Since $L(\alpha,\pi) \geq 0$, we have $L(\alpha,\beta) \geq 0$ on $(\beta^*,\pi]$.

If $A_1 \notin (-1,1)$, there is no $\beta^* \in (0,\pi)$ such that $\frac{d}{d\beta}L(\alpha,\beta) = 0$. Combining this with the fact $\frac{d}{d\beta}L(\alpha,0) > 0$, we have $\frac{d}{d\beta}L(\alpha,\beta) > 0$ on $(0,\pi)$. Thus, $L$ is an increasing function here, and thus $L(\alpha,\beta) \geq 0$.

Therefore, $L \geq 0$.

Through a similar process, one can show that

$$2\pi - \|h_1(\phi(s_1)) - h_1(\phi(s_2))\| - d_{\mathbb{S}^d}(s_1,s_2) \geq 0$$

and thus we complete the proof. □

**Lemma H.6** *In 3D, the function $h_1 \circ \phi : \mathbb{S}^2 \setminus \{s_n\} \to \mathbb{R}^2$, where $\phi$ defined by (46) and $h_1$ by (14) with $d = 2$ has range contained in the circle of radius $\pi$ centered at the origin. Moreover, $h_1 \circ \phi$ maps circles parallel to the Equator to circles in the plane centered at the origin. See Figure 11 for a visualization.*

*This can be generalized to the dimensions $d > 2$, where we consider the function $h_1 \circ \phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$, and in the preceding statement we have to change the word "circle" by "sphere". The Equator in $\mathbb{S}^d$ is the sub-sphere intersecting the hyperplane in $\mathbb{R}^{d+1}$ with equation $x_{d+1} = 0$.*

*Proof:* This is a consequence of part 1 in Proposition H.1, that is, $h_1(\phi(s)) = \angle(s,s_0)\frac{s[1:d]}{\|s[1:d]\|}$ for every $s \in \mathbb{S}^d \setminus \{s_n\}$, and the fact that the stereographic projection $\phi$ with formula given by (46) maps circles (resp., sub-spheres of dimension $d-1$ in $\mathbb{R}^{d+1}$) parallel to the Equator to circles (resp., spheres of dimension $d-1$) in $\mathbb{R}^2$ (resp., $\mathbb{R}^d$) -see Figure 8-.

Given a circle $S_\alpha$ (sub-sphere) which is the intersection between the sphere $\mathbb{S}^d$ and the (hyper)plane $x_{d+1} = \sin(\alpha)$ for some $\alpha \in [-\pi/2,\pi/2)$ (and so, $-1 \leq \sin(\alpha) < 1$), i.e.,

$$S_\alpha = \mathbb{S}^d \cap \left\{x \in \mathbb{R}^{d+1} :\ x_{d+1} = \sin(\alpha)\right\}, \tag{86}$$

we have that

$$\angle(s,s_0) = \angle(s',s_0) = \alpha + \frac{\pi}{2} \qquad \forall s,s' \in S_\alpha.$$

Thus, if for example we consider a representative $s = [0,\ldots,0,\cos(\alpha),\sin(\alpha)]$ in $S_\alpha$, we have that $h_1(\phi(s)) = \angle(s,s_0) = \alpha + \frac{\pi}{2}$. In general,

$$h_1(\phi(s')) = (\alpha + \frac{\pi}{2})\frac{s'[1:d]}{\|s'[1:d]\|} \qquad \forall s' \in S_\alpha.$$

Since the points of the form $\frac{s'[1:d]}{\|s'[1:d]\|}$ belong to the unit "circle" in $\mathbb{R}^d$, the function $h_1 \circ \phi$ maps the "circle" $S_\alpha$ (which is parallel to the Equator at height $\sin(\alpha)$) to the "circle" centered at the origin with radius $\alpha + \frac{\pi}{2}$:

$$h_1(\phi(S_\alpha)) = \left\{x \in \mathbb{R}^d :\ \|x\| = \alpha + \frac{\pi}{2}\right\}.$$

In particular, we have the following:

- The Equator corresponds to $\alpha = 0$ and so

$$h_1(\phi(\text{Equator})) = \left\{x \in \mathbb{R}^d :\ \sqrt{x_1^2 + \cdots + x_d^2} = \frac{\pi}{2}\right\}.$$

- The South Pole corresponds to $\alpha = -\pi/2$, i.e., $s_0 = [0,\ldots,0,-1] = [0,\ldots,\cos(-\pi/2),\sin(-\pi/2)]$, and so $h_1(\phi(s_0)) = [0,\ldots,0] \in \mathbb{R}^d$.

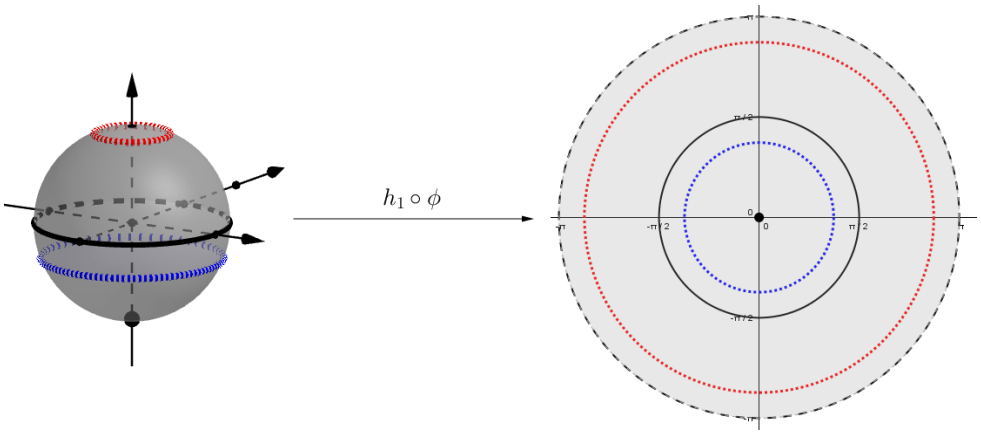



*Figure 11.* Depiction of the map $h_1 \circ \phi : \mathbb{S}^2 \setminus \{s_n\} \to \mathbb{R}^2$. The domain is depicted on the left. Its image, on the right, is inside the open disk of radius $\pi$ centered at the origin. The Equator (black circle on the left) goes to the circle of radius $\pi/2$ (black circle on the right). The South Pole (black big dot on the left) goes to the origin of coordinates of the plane on the right (black dot). Dotted blue "Parallel" on the left at negative height (in the South Hemisphere), goes to the small blue dotted circle on the right. Dotted red "Parallel" on the left at positive height (in the North Hemisphere), goes to the big red dotted circle on the right. We recall that "parallels" are circles on the sphere that are parallel to the Equator.

- For small numbers $\epsilon > 0$ we have that the "circle" $S^\epsilon$ parallel to the Equator at height $\pi/2 - \epsilon$ defined by

$$S^\epsilon = \mathbb{S}^d \cap \left\{ x \in \mathbb{R}^{d+1} : x_{d+1} = \frac{\pi}{2} - \epsilon \right\}$$

  is sent to

$$h_1(\phi(S^\epsilon)) = \left\{ x \in \mathbb{R}^d : \sqrt{x_1^2 + \cdots + x_d^2} = \pi - \epsilon \right\}.$$

  Thus, as $\epsilon \to 0$ we obtain that the range of $h_1 \circ \phi$ is contained on the open sphere/circle in $\mathbb{R}^d$ of radius $\pi$ centered at the origin.

- "Circles" at positive height, i.e., $S_\alpha$ as in (86) with $0 \leq \alpha < \pi/2$, are sent through $h_1 \circ \phi$ to spheres/circles in $\mathbb{R}^d$ with radius grater than $\pi/2$ and smaller that $\pi$ centered at the origin. "Circles" at at negative height, i.e., $S_\alpha$ as in (86) with $-\pi/2 \leq \alpha < 0$, are sent through $h_1 \circ \phi$ to spheres/circles in $\mathbb{R}^d$ with radius between 0 and $\pi/2$ centered at the origin.

□

**Corollary H.7** *Given two arbitrary points $s_1, s_2 \in \mathbb{S}^2 \setminus \{s_n\}$, we have the following bounds*

$$d_{\mathbb{S}^d}(s_1, s_2) \leq \pi \qquad \textit{and} \qquad \|h_1(\phi(s_1)) - h_1(\phi(s_2))\| \leq 2\pi.$$

*Proof:* In general $d_{\mathbb{S}^d}(s_1, s_2)$ is the minimum arc length between $s_1$ and $s_2$ taken among all possible paths on the sphere connecting those points, and so it is at most $\pi$. We recall that $d_{\mathbb{S}^d}(s_1, s_2) = \arccos(\langle s_1, s_2 \rangle) \in [0, \pi]$.

The other inequality is a consequence of Lemma H.6: The higher Ecudidean distance between points in a circle (sphere) with diameter $2\pi$ (radius $\pi$) is, in fact, $2\pi$. □

The above results (identity (81), Lemma H.5 and Corollary H.7) suggest that given two points $s_1, s_2$ on the sphere we would have that $d_{\mathbb{S}^d}(s_1, s_2) \leq \|h_1(\phi(s_1)) - h_1(\phi(s_2))\|$. This is shown experimentally in Figure 12. We will need some more lemmas to then sketch the proof of this fact theoretically.

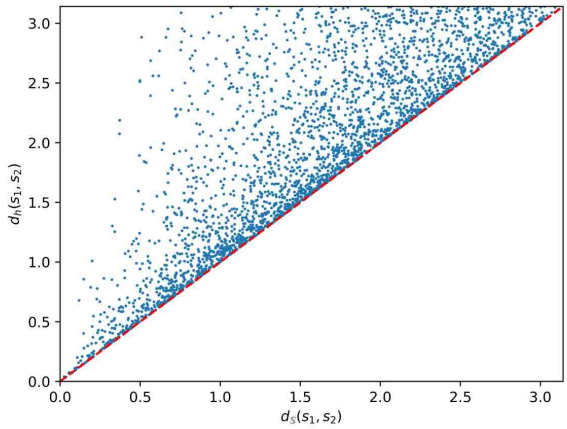



*Figure 12.* Experimental results for the relation $d_{\mathbb{S}^2}(s_1, s_2) \leq d_h(s_1, s_2) := \min\{\|h_1(\phi(s_1)) - h_1(\phi(s_2))\|, 2\pi - \|h_1(\phi(s_1)) - h_1(\phi(s_2))\|\}$. There were taken 1000 pairs $(s_1, s_2)$ uniformly at random on the unit sphere in $\mathbb{R}^3$. Additionally, we calculate the correlation between $d_{S^2}(s_1, s_2)$ and $d_h(s_1, s_2)$, resulting in an average correlation coefficient of 0.83.

**Lemma H.8** *Consider $s_1, s_2 \in \mathbb{S}^d$, with $(s_1)_{d+1} \leq (s_2)_{d+1}$, let $s_2'$ be the unique point such that*

$$(s_2')_{d+1} = (s_1)_{d+1} \qquad \textit{and} \qquad \angle(s_2'[1:d], s_2[1:d]) = 0. \tag{87}$$

*Let $\beta = \angle(s_1[1:d], s_2[1:d]) \in [0, \pi]$, and let $C_H$ denote the angle between line segments $[h_1(\phi(s_1)), h_1(\phi(s_2'))], [h_1(\phi(s_2')), h_1(\phi(s_2))]$. Then,*

$$C_H = \frac{\beta}{2} + \frac{\pi}{2} \tag{88}$$

*Proof:* We suggest the reader to look at Figure 13 for visualization purposes.

Consider $C_H$ the angle between line segments $[h_1(\phi(s_1)), h_1(\phi(s_2'))]$ and $[h_1(\phi(s_2)), h_1(\phi(s_2'))]$ in the projected space. The way we have chosen $s_2'$ guarantees that the segment $[0, h_1(\phi(s_2'))]$ is contained in the segment $[0, h_1(\phi(s_2))]$.

As an application of Lemma (H.6), we have that the segments $[0, h_1(\phi(s_2'))]$ and $[0, h_1(\phi(s_1))]$ have the same length as $s_2'$ and $s_1$ are at the same latitude. Then, the triangle in the projected space with vertices the origin or coordinates $O = [0, \ldots, 0] \in \mathbb{R}^d$, $s_1$ and $s_2'$ is isosceles. The angle of that triangle having vertex at $O$ is

$$\begin{aligned}\beta = \angle(s_1[1:d], s_2[1:d]) &= \angle(s_1[1:d], s_2[1:d]) = \angle(s_1[1:d], s_2'[1:d]) \\ &= \angle(\phi(s_1), \phi(s_2')) = \angle(h_1(\phi(s_1)), h_1(\phi(s_2')))\end{aligned} \tag{89}$$

and the other two angles are equal. Thus, since $C_H$ is adjacent to one of those two equal angles, and since the sum of the internal angles of a triangle is 180, we have $\pi = \beta + 2(\pi - C_H)$ or, equivalently, $C_H = \frac{\beta}{2} + \frac{\pi}{2}$. □

**Lemma H.9** *Consider $s_1, s_2 \in \mathbb{S}^d$, with $(s_1)_{d+1} \leq (s_2)_{d+1}$, let $s_2'$ be the unique point satisfying (87). Let $C \in [0, \pi]$ denote the spherical angle between $arc(s_1, s_2'), arc(s_2', s_2)$, and $C_H$ denote the angle between line segments $[h(\phi(s_1)), h(\phi(s_2')], [h_1(\phi(s_2')), h_1(\phi(s_2))]$. Then,*

$$C \leq C_H. \tag{90}$$

*Proof:* For better visualization and an easier way of parametrizing the points, we will work on the 3-dimensional space (see Figure 14). Thus, assuming $d = 2$, by using spherical coordinates let $\alpha \in [-\pi/2, \pi/2]$, $\beta = \angle(s_1[1:d], s_2[1:d]) \in [0, \pi]$, and $\alpha' \in [\alpha, \frac{\pi}{2})$ and, without loss of generality, let us parametrize

$$\begin{cases} s_1 = [\cos(\alpha), 0, \sin(\alpha)] \\ s_2 = [\cos(\alpha')\cos 2\beta, \cos(\alpha')\sin(\beta), \sin(\alpha)] \\ s_2' = [\cos\alpha\cos 2\beta, \cos(\alpha)\sin(\beta), \sin(\alpha)] \end{cases} \tag{91}$$

If $\beta = 0$, $s_2' = s_1$, $C = C_H = 0$ and there is nothing else to prove.

If $\beta > 0$, the angle $C$ is characterized by

$$\cos(C) = \frac{\langle n_{12'}, n_{22} \rangle}{\|n_{12'}\| \|n_{22'}\|},$$

where $n_{12'}$ is a vector normal (perpendicular) to the plane containing the vectors $\overrightarrow{Os_1}$ and $\overrightarrow{Os_2'}$, and $n_{22'}$ is a vector normal to the plane containing the vectors $\overrightarrow{Os_2}$ and $\overrightarrow{Os_2'}$, where $O$ denotes the origin of coordinates $O = [0, \dots, 0] \in \mathbb{R}^{d+1}$.

Since $\arccos(\cdot)$ is a decreasing function, want to show that

$$\cos(C) \geq \cos(C_H). \tag{92}$$

By using the parametrizations in (91), we can compute

$$\begin{cases} n_{22'} = [\sin(\beta), -\cos(\beta), 0] & \|n_{22'}\| = 1 \\ n_{12'} = [\sin(\alpha)\sin(\beta), \sin(\alpha)(1-\cos(\beta)), \cos(\alpha)\sin(\beta)] & \|n_{12'}\| = \sqrt{\sin^2(\beta) + \sin^2(\alpha)(1-\cos(\beta))^2} \\ \langle n_{12'}, n_{22'} \rangle = \sin(\alpha)(1-\cos(\beta)) \end{cases}$$

Thus,

$$\cos(C) = \frac{\sin(\alpha)(1-\cos(\beta))}{\sqrt{\sin^2(\beta) + \sin^2(\alpha)(1-\cos(\beta))^2}} \tag{93}$$

Notice that in the special case where $\alpha \in [0, \pi/2]$, that is, $(s_1)_{d+1} \geq 0$, and so both $s_1, s_2$ are in the same "positive hemisphere" (as in Figure 14), we have that $\sin(\alpha) \geq 0$, so, $\cos(C) \geq 0$, and therefore $0 \leq C \leq \pi/2$. Thus, in this particular case,

$$C \leq \pi/2 \leq C_H \qquad \text{if } 0 \leq \alpha < \frac{\pi}{2}. \tag{94}$$

(See Figure 14 for a visualization of this case.)

In general, by using (93) and (88), the inequality (92) reads as

$$\cos(C) = \frac{\sin(\alpha)(1-\cos(\beta))}{\sqrt{\sin^2(\beta) + \sin^2(\alpha)(1-\cos(\beta))^2}} \geq \cos\left(\frac{\beta}{2} + \frac{\pi}{2}\right) = -\sin\left(\frac{\beta}{2}\right)$$

By algebraic manipulations and by using trigonometric identities (92) holds if and only if

$$\sin\left(\frac{\beta}{2}\right)\left(1 + \frac{\sin(\alpha)\sin\left(\frac{\beta}{2}\right)}{\sqrt{\cos^2\left(\frac{\beta}{2}\right) + \sin^2(\alpha)\sin^2\left(\frac{\beta}{2}\right)}}\right) \geq 0 \qquad \text{if } 0 \leq \beta \leq \pi,\ \frac{-\pi}{2} \leq \alpha \leq \frac{\pi}{2} \tag{95}$$

Let us rewrite the inequality in (95) (by using that $\sin(\beta/2) \geq 0$ in our range) as

$$-\sin(\alpha) \leq \sqrt{\cos^2\left(\frac{\beta}{2}\right) + \sin^2(\alpha)\sin^2\left(\frac{\beta}{2}\right)} \tag{96}$$

The parameter $\alpha$ is constraint to the interval $[-\pi/2, \pi/2)$. We have already proved that for $\alpha \geq 0$ everything works (see (94)). Thus, let us suppose that $\alpha < 0$: In this case, $\sin(\alpha) < 0$ and so, both sides in (96) are positive. Therefore, it is equivalent to show that $\sin^2(\alpha) \leq \cos^2\left(\frac{\beta}{2}\right) + \sin^2(\alpha)\sin^2\left(\frac{\beta}{2}\right)$ or, equivalently, that $\sin^2(\alpha) \leq 1$, which is always true.

□

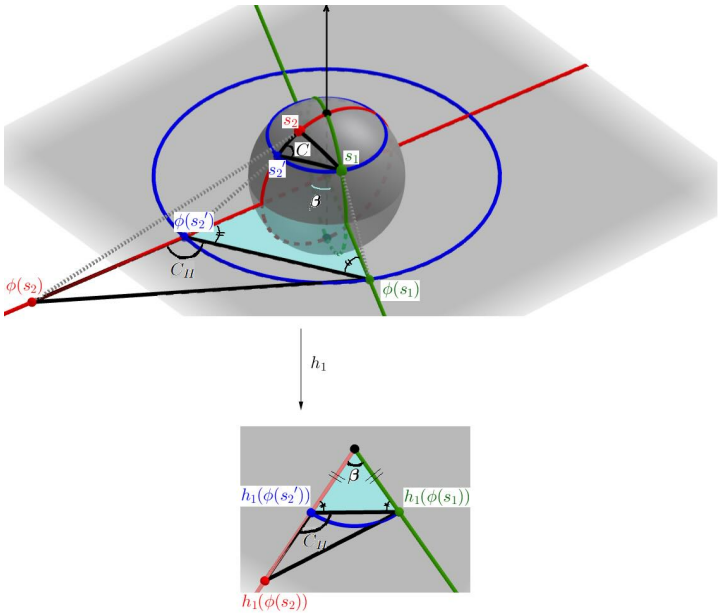



*Figure 13.* On top we have the sphere in the 3-dimensional space. Three points $s_1$, $s_2$, $s_2'$ are considered on the sphere. The point $s_2'$ satisfies (87), that is, $s_2', s_1$ are at the same latitude, and $s_2', s_2$ are in the same meridian (and among the points that satisfy these two conditions always choose the $s_2'$ which is "closer" to $s_2$). The corresponding projections $\phi(s_1)$, $\phi(s_2)$, $\phi(s_2')$ are plotted on the 2-dimensional plane. The circle on the sphere parallel to the Equator that passes through $s_1$ and $s_2'$ and its projected circle by $\phi$ are depicted in blue. The circle on the sphere that passes through $s_2$, $s_2'$, and the poles (meridian) and its projection by $\phi$ (which is a line) are depicted in red. Similarly, the circle on the sphere that passes through $s_1$, and the poles (meridian) and its projection by $\phi$ (which is a line) are depicted in green. On the bottom, we sketch the image by $h_1$ of the projected space. The angles $\beta$ given by (89), $C$ (spherical angle between the arcs $\text{arc}(s_1, s_2')$, $\text{arc}(s_2', s_2)$), and $C_H$ (defined by the line segments $[\phi(s_1), \phi(s_2')]$ and $[\phi(s_2), \phi(s_2')]$ or, equivalently, by the segments $[h_1(\phi(s_1)), h_1(\phi(s_2'))]$ and $[h_1(\phi(s_2)), h_1(\phi(s_2'))]$) are depicted. When applying the function $h_1$ after $\phi$, it preserves angles with vertex at the origin in the projected space, preserves the lines that pass through the origin (red and green lines), maps circles centered at the origin to circles centered at the origin (blue arc). Roughly speaking, the effect of $h_1$ is that it "compresses" the projected plane.

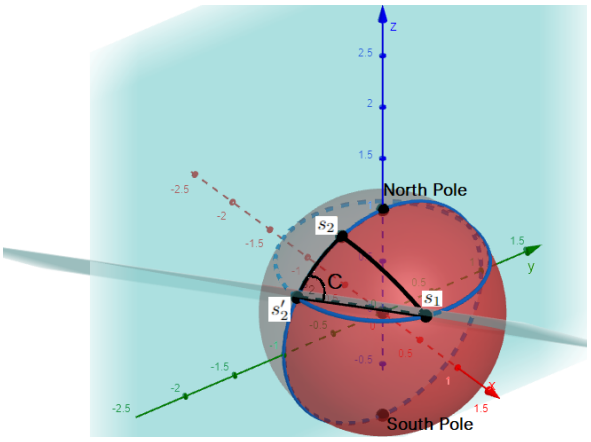



*Figure 14.* Given two arbitrary points $s_1, s_2 \in \mathbb{S}^2 \subset \mathbb{R}^3$ in the "positive" hemisphere. The point $s_1'$ is plotted such that it has the same latitude as $s_1$ and it is in the same meridian as $s_2$. The angle $C$ is depicted as the angle between the two planes containing the arcs $arc(s_1, s_2')$ and $arc(s_2, s_2')$. A spherical triangle connecting the points $s_1, s_2', s_2$ is depicted in black.

**Proposition H.10** *Let $s, s_1, s_2 \in \mathbb{S}^d \setminus \{s_n\}$, and let $s_0 = [0, 0, \ldots, -1]^T \in \mathbb{S}^d$ denote the South Pole. Consider $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ the stereographic projection as in (46) with inverse (47), and also consider the function $h_1 : \mathbb{R}^d \to \mathbb{R}^d$*

*defined by* (76)*. Then*

$$d_{\mathbb{S}^d}(s_1, s_2) \leq \min_{R \in \mathrm{O}(d+1)} \|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\| + \epsilon(s_1, s_2)$$

*where* $\epsilon(s_1, s_2) \to 0$ *as* $d_{\mathbb{S}^d}(s_1, s_2) \to 0$.

*Proof:*

Let $s_1$, $s_2$ in $\mathbb{S}^d$. Assume, without loss of generality, that with $(s_2)_{d+1} \geq (s_1)_{d+1}$.

Define $s_2'$ as in Lemma H.9: That is, such that $(s_2')_{d+1} = (s_1)_{d+1}$ (same latitude) and $s_2'$, $s_2$, $s_0$ are in the same great circle (same meridian). Since there are two possible choices for $s_2'$, choose it not only such that $(s_2')_{d+1} = (s_1)_{d+1}$ but also such that $\angle(s_2'[1:d], s_2[1:d]) = 0$. (This means we always choose the $s_2'$ which is "closer" to $s_2$.)

Let $C$ be the *spherical angle* angle between the arcs $arc(s_1, s_2')$ and $arc(s_2, s_2')$ given by the smallest portions of the great circles connecting the corresponding points $s_1$ with $s_2'$ and $s_2$ with $s_2'$, with vertex the origin of coordinates $[0, \ldots, 0] \in \mathbb{R}^{d+1}$. We recall that the angle $C$ is measured by the angle between the planes containing those two arcs.

See Figure 14 for a visualization in $\mathbb{R}^3$.

A *spherical triangle* on the surface of the unit sphere is defined by the great circles connecting three points on the sphere, in our case consider $s_1, s_2, s_2'$. The *spherical law of cosines* reads as follows:

$$\cos(d_{\mathbb{S}^d}(s_1, s_2)) = \cos(d_{\mathbb{S}^d}(s_1, s_2')) \cos(d_{\mathbb{S}^d}(s_2, s_2')) + \sin(d_{\mathbb{S}^d}(s_1, s_2')) \sin(d_{\mathbb{S}^d}(s_2, s_2')) \cos(C).$$

As a consequence, by using Taylor expansions,

$$\begin{aligned} d_{\mathbb{S}^d}(s_1, s_2)^2 = d_{\mathbb{S}^d}(s_1, s_2')^2 + d_{\mathbb{S}^d}(s_2, s_2')^2 - 2 d_{\mathbb{S}^d}(s_1, s_2') d_{\mathbb{S}^d}(s_2, s_2') \cos(C) \\ + \mathcal{O}(d^4_{\mathbb{S}^d}(s_1, s_2)) + \mathcal{O}(d^4_{\mathbb{S}^d}(s_1, s_2')) + \mathcal{O}(d^4_{\mathbb{S}^d}(s_2, s_2')) \end{aligned}$$

Thus, for "small" spherical triangles, i.e., when the lengths $d_{\mathbb{S}^d}(s_1, s_2)$, $d_{\mathbb{S}^d}(s_1, s_2')$, $d_{\mathbb{S}^d}(s_2, s_2')$ are small, we can use the "planar" approximation

$$d_{\mathbb{S}^d}(s_1, s_2)^2 \approx d_{\mathbb{S}^d}(s_1, s_2')^2 + d_{\mathbb{S}^d}(s_2, s_2')^2 - 2 d_{\mathbb{S}^d}(s_1, s_2') d_{\mathbb{S}^d}(s_2, s_2') \cos(C).$$

Now, as in the above lemmas, consider $C_H$ the angle between line segments $[h_1(\phi(s_1)), h_1(\phi(s_2'))]$ and $[h(\phi(s_2)), h(\phi(s_2'))]$ in the projected space.

By Lemma H.5 and by using $C \leq C_H$ together with fact that $\cos$ is a decreasing function on $[0, \pi]$, we have:

$$\begin{aligned} d^2_{\mathbb{S}^d}(s_1, s_2') + d^2_{\mathbb{S}^d}(s_2, s_2') - 2 d_{\mathbb{S}^d}(s_1, s_2') d_{\mathbb{S}^d}(s_2, s_2') \cos(C) &\leq \|h_1(\phi(s_1)) - h_1(\phi(s_2'))\|^2 + \|h_1(\phi(s_2)) - h_1(\phi(s_2'))\|^2 \\ &\quad - 2\|h_1(\phi(s_1)) - h_1(\phi(s_2'))\| \|h_1(\phi(s_2)) - h_1(\phi(s_2'))\| \cos(C_H) \\ &= \|h_1(\phi(s_1)) - h_1(\phi(s_2))\|^2 \end{aligned}$$

where in the last step we used the classical law of cosines.

Therefore, by using the approximation given by the spherical law of cosines (assuming $d_{\mathbb{S}^d}(s_1, s_2)$ small), we obtain

$$d_{\mathbb{S}^d}(s_1, s_2) \leq \|h_1(\phi(s_1)) - h_1(\phi(s_2))\|.$$

In particular,

$$d_{\mathbb{S}^d}(s_1, s_2) = d_{\mathbb{S}^d}(Rs_1, Rs_2) \leq \|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\| \qquad \forall R \in \mathrm{O}(d+1)$$

and so

$$d_{\mathbb{S}^d}(s_1, s_2) \leq \min_{R \in \mathrm{O}(d+1)} \|h_1(\phi(Rs_1)) - h_1(\phi(Rs_2))\|.$$

□

### H.1. Neural Network-Based Embedding

While any choice for $h$ which maintains injectivity will ensure $S3W_{\mathcal{H},p}$ remains a valid metric in $\mathcal{P}_p(\mathbb{S}^d)$, the particular choice for $h$ is significant in ensuring that the transportation cost in the embedding space resembles the spherical distance. As discussed in Section 3.3, as an alternative to the proposed analytic function $h_1(\cdot)$ (as given in Eq. (14)), we may consider training a neural network to obtain a nearly-isometric Euclidean embedding and correct for the distance distortion caused by stereographic projection.

Here, we consider the injective function defined by $h_{NN}(x) := [h_1^T(x)/C, \rho^T(x)]^T$ where $\rho : \mathbb{R}^2 \to \mathbb{R}^3$ is a neural network, and $C \geq 2\pi$ a constant. We parameterize $\rho$ using a multi-layer perceptron (MLP) with two hidden layers, each consisting of 128 neurons, and train $\rho$ by minimizing

$$\mathcal{L}(\rho) = \mathbb{E}_{s,s'}\left[(\arccos(\langle s, s'\rangle) - \|h(\phi(s)) - h(\phi(s'))\|)^2\right], \tag{97}$$

where $s$ and $s'$ are sampled according to the uniform distribution in the sphere $\mathbb{S}^d \subset \mathbb{R}^{d+1}$, i.e., $(s, s') \sim \sigma_{d+1} \times \sigma_{d+1}$. As shown in Figure 2, this choice for $h$ can yield a nearly-isometric embedding.

We then conduct two different gradient flow experiments to study the distinct behaviors of $h_1(\cdot)$ and $h_{NN}(\cdot)$. The visualizations demonstrating these gradient flows can be found at the following GitHub URL. We observe that the use of $h_{NN}(\cdot)$ provides more "geodesic-like" paths for the particles as compared to $h_1(\cdot)$, though the final outcomes are comparable.

We note that in our experiments, we primarily use only the analytic $h_1$ for its computational efficiency as opposed to the neural network-based learnable function $h_{NN}$. Specifically, in our experiments, the proposed distances are used in gradient-based optimization frameworks where the distance is computed in each iteration of (stochastic) gradient descent. Utilizing $h_{NN}$ in these optimization applications increases the computational cost of calculating the distance. More precisely, the computational cost would have the following components: 1) stereographic projection, 2) evaluation of $h_{NN}$, and 3) slicing, with a neural network becoming the computational bottleneck in the evaluation of $h_{NN}$.

## I. Numerical Experiments

In this section, we aim to demonstrate our proposed $S3W$ as an efficient metric suitable for various Machine Learning tasks. In I.2, we show that $S3W$ and its variants provide a high-speed alternative for $SSW$ as an effective loss for gradient-based optimization on the spherical manifold. In I.4, we study the evolution of our proposed distances w.r.t. varying parameters. In I.5, we discuss the computational efficiency of our method via runtime study. I.6, I.7, I.8, I.9 provide experiments in practical ML settings. Throughout our experiments, we use $h(x) = h_1(x) := \arccos\left(\frac{\|x\|^2-1}{\|x\|^2+1}\right)\frac{x}{\|x\|}$.

### I.1. Additional Algorithmic and Implementation Details

Let $\{\alpha_i, x_i\}_{i=1}^M \sim \hat{\mu}$ and $\{\beta_j, y_j\}_{j=1}^N \sim \hat{\nu}$ be iid samples from the empirical measures $\hat{\mu}$ and $\hat{\nu}$, respectively, where $\sum_{i=1}^M \alpha_i = 1$ and $\sum_{j=1}^N \beta_j = 1$. We provide below a general formulation for computing $S3W_p$ between $\hat{\mu}$ and $\hat{\nu}$. We then define the procedure to compute $RI\text{-}S3W_p$ and $ARI\text{-}S3W_p$. By abuse of notation, we denote these Algorithms as $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$, respectively.

#### I.1.1. ALTERNATIVE IMPLEMENTATION TO COMPUTE $S3W_p$

**Algorithm 2** Stereographic Spherical Sliced-Wasserstein ($S3W$)

**Input:** $\{\alpha_i, x_i\}_{i=1}^M \sim \hat{\mu}$, $\{\beta_j, y_j\}_{j=1}^N \sim \hat{\nu}$, $L$ projections, $p$-th order, $\epsilon$
Initialize: $h$ (injective map), $\{\theta_l\}_{l=1}^L$ (projection directions)
Initialize $d = 0$
Calculate $\{u_i = h(\phi_\epsilon(x_i))\}_{i=1}^M$ and $\{v_j = h(\phi_\epsilon(y_j))\}_{j=1}^N$
**for** $\ell = 1$ **to** $L$ **do**
  Project onto $\theta_l$: $u_i^l = \langle u_i, \theta_l \rangle$ and $v_j^l = \langle v_j, \theta_l \rangle$
  Sort $\{u_i^l\}$, $\{v_j^l\}$, s.t $u^l_{\pi_l[i]} \leq u^l_{\pi_l[i+1]}$, $v^l_{\pi'_l[j]} \leq v^l_{\pi'_l[j+1]}$
  Compute the (weighted) CDFs $\{\tilde{u}_i^l = \tilde{F}_{u^l}(u^l_{\pi_l[i]})\}_{i=1}^M$, $\{\tilde{v}_j^l = \tilde{F}_{v^l}(v^l_{\pi_l[j]})\}_{j=1}^N$
  Merge and sort the $\{\tilde{u}_i^l\}_i$ and $\{\tilde{v}_j^l\}_j$ into $\{z_k^l\}_{k=1}^{M+N}$ where $z_k^l \leq z_{k+1}^l$
  For each $z_k^l$, compute $\tilde{F}^{-1}_{u^l}(z_k^l)$ and $\tilde{F}^{-1}_{v^l}(z_k^l)$ by setting $\tilde{F}^{-1}_{u^l}(z_k^l)$ to the value of $u_i^l$ such that $\tilde{F}_{u^l}(u_i^l)$ is nearest to $z_k^l$.
  Proceed similarly for $\tilde{F}^{-1}_{v^l}(z_k^l)$.
  $d_\ell = \sum_{k=1}^{M+N} \left| \tilde{F}^{-1}_{u^l}(z_k^l) - \tilde{F}^{-1}_{v^l}(z_k^l) \right|^p (z_k^l - z_{k-1}^l)$
  Update total distance: $d = d + \frac{1}{L} d_\ell$
**end for**
Return S3W$_p(\hat{\mu}, \hat{\nu}) \approx d^{\frac{1}{p}}$

#### I.1.2. IMPLEMENTATION TO COMPUTE THE $RI\text{-}S3W_p$

**Algorithm 3** Rotationally Invariant $S3W$ ($RI\text{-}S3W$)

**Input:** $\{\alpha_i, x_i\}_{i=1}^M \sim \hat{\mu}$, $\{\beta_j, y_j\}_{j=1}^N \sim \hat{\nu}$, $L$ projections, $p$-th order, $\epsilon$, $N_R$ rotations
Sample $\{R_r\}_{r=1}^{N_R} \in \mathrm{SO}(d+1)$
Initialize $d = 0$
**for** $r = 1$ **to** $N_R$ **do**
  Apply rotation $R_r$ to obtain $X_r = \{\alpha_i R_r(x_i)\}_{i=1}^M$ and $Y_r = \{\beta_j R_r(y_j)\}_{j=1}^N$
  Use Algorithm 2 for inputs $X_r, Y_r$ to compute $d_r$
  Update total distance: $d = d + \frac{1}{N_R} d_r$
**end for**
Return $RI\text{-}S3W_p(\hat{\mu}, \hat{\nu}) \approx d$

#### I.1.3. IMPLEMENTATION TO COMPUTE THE $ARI\text{-}S3W_p$

Although $RI\text{-}S3W$ is highly parallelizable via batch matrix multiplication, generating random rotation matrices can still be computationally expensive. That is especially true for high-dimensional data and training setups with small batch size where $RI\text{-}S3W$ gets called repeatedly. One way to mitigate this issue is by pregenerating a batch of $N$ rotation matrices, $\{R_i\}_{i=1}^N \subset \mathrm{SO}(d+1)$, and subsample from this batch to amortize the generation cost. We introduce the Amortized Rotationally Invariant $S3W$ ($ARI\text{-}S3W$) as algorithm 4. This process can be made seamless by managing the rotation matrices in a stateful manner. That is, we maintain a Rotation manager class with static variables to ensure that once a set of matrices is generated, they can be efficiently reused. We describe the abstract procedure to compute $ARI\text{-}S3W$ in Algorithm 4.

**Algorithm 4** Amortized Rotationally Invariant $S3W$ ($ARI$-$S3W$)

**Input:** $\{\alpha_i, x_i\}_{i=1}^M \sim \hat{\mu}$, $\{\beta_j, y_j\}_{j=1}^N \sim \hat{\nu}$, $L$ projections, $p$-th order, pool size $N_{\text{total}}$, $N_R$ rotations, pregenerated rotations $\mathcal{R} = \{R_k\}_{k=1}^{N_{total}} \subset \text{SO}(d+1)$ where $N_{total} \geq N_R$
Initialize $d = 0$
Randomly subsample $N_R$ rotations $\tilde{\mathcal{R}} = \{\tilde{R}_r \sim \mathcal{R}\}_{r=1}^{N_R}$
**for** each r in $N_R$ **do**
  Apply $\tilde{R}_r$ to obtain $X_r = \{\alpha_i \tilde{R}_r(x_i)\}_{i=1}^M$ and $Y_r = \{\beta_j \tilde{R}_r(y_j)\}_{j=1}^N$
  Use Algorithm 2 where inputs are $\{\tilde{R}_r(x_i^r)\}_{i=1}^N$ and $\{\tilde{R}_r(y_j^r)\}_{j=1}^N$ to compute $d_r$
  Update total distance: $d = d + \frac{1}{N_R} d_r$
**end for**
Return $ARI\text{-}S3W_p(\hat{\mu}, \hat{\nu}) \approx d$

### I.2. Study: Gradient Flow on the Sphere

#### I.2.1. Background Overview

**The von Mises-Fisher Distribution** is a generalization of a Gaussian from $\mathbb{R}^d$ to $\mathbb{S}^d$. Its parametric form can be characterized by the mean direction $\mu \in \mathbb{S}^d$ and the concentration parameter $\kappa > 0$. That is

$$f(x; \mu, \kappa) = C_d(\kappa) \exp(\kappa \mu^T x) \tag{98}$$

where $C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$ is the normalization constant. In our experiment setting, we mostly work with $d = 2$. The pdf is

$$f(x; \mu, \kappa) = \frac{\kappa}{4\pi \sinh(\kappa)} \exp(\kappa \mu^T x). \tag{99}$$

**Gradient descent on the sphere** The sphere $\mathbb{S}^d$ is a Riemannian manifold where Euclidean geometry holds only locally. This local linearization is sufficient to adapt gradient descent to follow geodesic paths, which replace straight lines in Euclidean space. For a function $f : \mathbb{R}^m \to \mathbb{R}$, the classical Euclidean gradient update rule is given by

$$x_{t+1} = x_t - \gamma \nabla_{\text{Euc}} f(x_t), \tag{100}$$

where $\nabla_{\text{Euc}} f(x_t)$ denotes the Euclidean gradient at $x_t$, and from now on we simply write $\nabla f(x_t)$. On manifolds, where straight lines are generalized to geodesic curves, we replace $\nabla f(x_t)$ with the Riemannian gradient $\nabla_{\text{Riemannian}} f(x) = \nabla f(x) - \langle \nabla f(x), x \rangle x$ where $\langle \cdot, \cdot \rangle$ denotes the inner product. While this gives local linearized direction, it does not specify how to move along the manifold itself. To address this, one could use a retraction[4]. When that retraction is normalization, the update rule becomes

$$\begin{aligned} x_{t+1} &= \text{Retr}_{x_t}\left(-\gamma \nabla_{\text{Riemannian}} f(x_t)\right) \\ &= \frac{x_t - \gamma\left(\nabla f(x_t) - \langle \nabla f(x_t), x_t \rangle x_t\right)}{\|x_t - \gamma\left(\nabla f(x_t) - \langle \nabla f(x_t), x_t \rangle x_t\right)\|}. \end{aligned} \tag{101}$$

If we use $\nabla f(x_t)$ instead of $\nabla_{\text{Riemannian}} f(x_t)$, then we have the familiar Projected Gradient Descent. Another option for retraction is using the exponential map (see Eq. (111)) to project $\nabla_{\text{Riemannian}} f(x_t)$ from the tangent space $T_x \mathbb{R}^d$ back onto the manifold along the geodesic. The gradient update is

[4] a mapping that approximates geodesics and projects the updated point in the tangent space back onto the manifold.

$$
\begin{aligned}
x_{t+1} &= \exp_{x_t}\left(-\gamma \nabla_{\text{Riemannian}} f(x_t)\right) \\
&= \cos\left(|-\gamma(\nabla f(x_t) - \langle \nabla f(x_t), x_t\rangle x_t)|\right) x_t \\
&\quad + \sin\left(|-\gamma(\nabla f(x_t) - \langle \nabla f(x_t), x_t\rangle x_t)|\right) \times \frac{-\gamma(\nabla f(x_t) - \langle \nabla f(x_t), x_t\rangle x_t)}{|-\gamma(\nabla f(x_t) - \langle \nabla f(x_t), x_t\rangle x_t)|}.
\end{aligned} \tag{102}
$$

This variant is Riemannian Gradient Descent (Absil et al., 2008). For simplicity, we use Projected Gradient Descent in our gradient-based optimization, similar to (Bonet et al., 2023a).

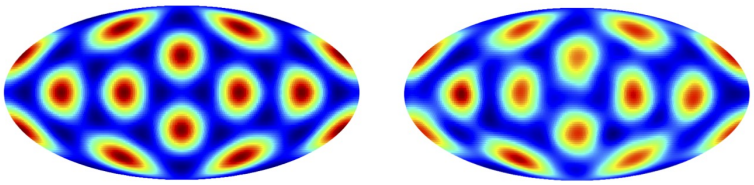
#### I.2.2. Experiment: Learning a Mixture of 12 von Mises-Fisher Distributions

Assuming we only have access to the target measure via its samples $\{y_j \in \hat{\nu_m}\}_{j=1}^{M}$ Our objective is to iteratively minimize $\text{argmin}_\mu d(\hat{\mu}_i, \hat{\nu}_{m_i})$. The Projected Gradient Descent algorithm for $S3W$ gives the following update rule

$$
\mathbf{x}'_{i,k+1} = \mathbf{x}_{i,k} - \gamma \nabla_{\mathbf{x}_{i,k}} \text{S3W}_p^p(\hat{\mu}_k, \hat{\nu}_{m_i}), \tag{103}
$$

$$
\mathbf{x}_{i,k+1} = \frac{\mathbf{x}'_{i,k+1}}{\|\mathbf{x}'_{i,k+1}\|_2}, \tag{104}
$$

where $\gamma$ denotes the learning rate, $i$ indexes the mini-batches, and $k$ the gradient step.



(a) Target Density (b) Target Empirical

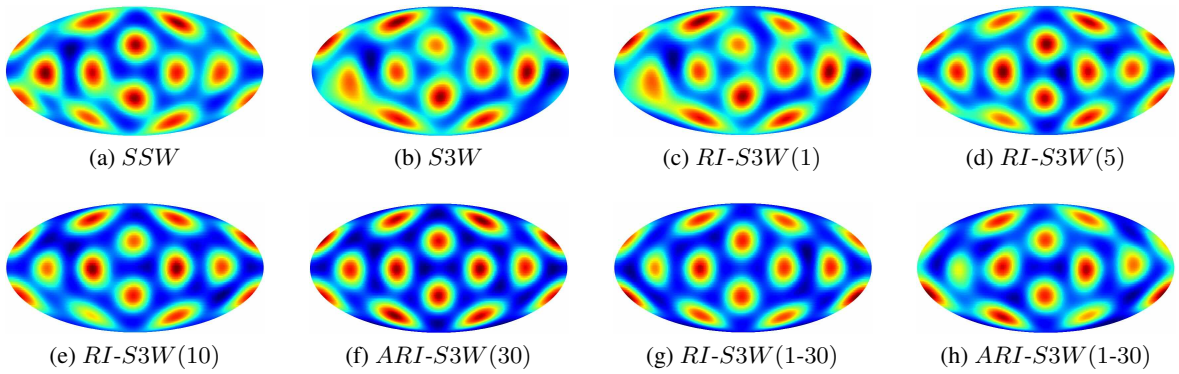
**Implementation** We consider a mixture of 12 vMFs as our target on $\mathbb{S}^2$, with each vMF centered on the vertices of an icosahedron determined by the golden ratio $\phi = (1 + \sqrt{5})/2$. We assume access to 2400 samples from the mixture (200 per vMF), each with $\kappa = 50$, and set batch sizes to $|m_i| = \{200, 2400\}$. We fix the number of projecions to 1000 for all distances. We optimize over 500 gradient steps using the Adam Optimizer (Kingma & Ba, 2014) (learning rate of $\gamma = 0.01$ for full-batch and 0.001 for mini-batch) for 10 in dependent runs. We report both qualitative and quantitative results (NLL, $W_2$); the latter comes with mean and standard deviation.

**Full-batch Results** In addition to the numerical results provided in the main paper (refer to Figure 4), we provide the Molleweide projections in Figure 16 to show that all distances work comparatively well to learn the target distribution. In this setting, we also include the evaluation of the vertical sliced Wasserstein (VSW) distance (Quellmalz et al., 2023) in Figure 16 and Table 4 (see Section G.1).

**Mini-batch results** We provide below our mini-batch results for all distances. In addition to the numerical results reported in Figure 4, we also show the Mollweide projections for the learned distributions as well as particle scatter-plots at epochs $\{0, 100, 300, 500\}$. In Table 4, $S3W$ shows the fastest runtime for both full-batch and mini-batch optimization. It is almost an order of magnitude faster than $SSW$. $RI\text{-}S3W(1)$ adds negligible computational overhead to $S3W$ with the single added rotation but performs almost on par with SSW. $RI\text{-}S3W$ (5) and $RI\text{-}S3W$ (10) are both high-performing candidates, exceeding $SSW$ albeit with some additional runtime compared to $S3W$. $ARI\text{-}S3W(30/1000)$ shows significant performance gain while being almost as fast as $S3W$. We remark that the smaller the batch size, the more frequently the distance function gets called, the bigger advantage $ARI\text{-}S3W$ will have over $RI\text{-}S3W$ since the rotation generation cost gets amortized over a large number of function calls and becomes negligible.

Both $RI\text{-}S3W$ and $ARI\text{-}S3W$ can achieve additional efficiency with rotation scheduler. In particular, one could start from few rotations and gradually (i.e. linearly) increase the number of rotations over time for fine-tuning. We provide additional

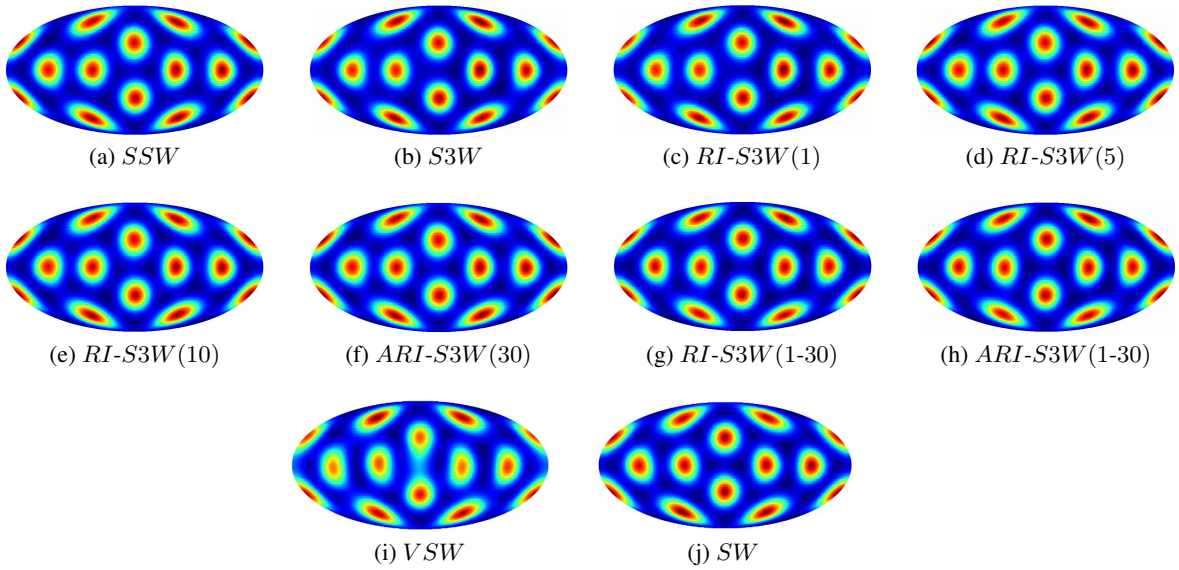
visualization and numerical results for $RI\text{-}S3W$ and $ARI\text{-}S3W$ with a linear schedule from 1 to 30 rotations over 500 epochs. We observe that $ARI\text{-}S3W(1\text{-}30)$ outperforms $ARI\text{-}S3W(30)$ while being twice as fast.



(a) $SSW$ (b) $S3W$ (c) $RI\text{-}S3W(1)$ (d) $RI\text{-}S3W(5)$

(e) $RI\text{-}S3W(10)$ (f) $ARI\text{-}S3W(30)$ (g) $RI\text{-}S3W(1\text{-}30)$ (h) $ARI\text{-}S3W(1\text{-}30)$

*Figure 15.* The Mollweide projections for mini-batch projected gradient descent. $ARI\text{-}S3W$ has pool size of 1000. $RI\text{-}S3W(1\text{-}30)$ and $ARI\text{-}S3W(1\text{-}30)$ denote $RI\text{-}S3W$ and $ARI\text{-}S3W$ with linear $N_R$ schedule from 1 to 30 over 500 epochs



(a) $SSW$ (b) $S3W$ (c) $RI\text{-}S3W(1)$ (d) $RI\text{-}S3W(5)$

(e) $RI\text{-}S3W(10)$ (f) $ARI\text{-}S3W(30)$ (g) $RI\text{-}S3W(1\text{-}30)$ (h) $ARI\text{-}S3W(1\text{-}30)$

(i) $VSW$ (j) $SW$

*Figure 16.* The Mollweide projections for full-batch projected gradient descent. $ARI\text{-}S3W$ has pool size of 1000. $RI\text{-}S3W(1\text{-}30)$ and $ARI\text{-}S3W(1\text{-}30)$ denote $RI\text{-}S3W$ and $ARI\text{-}S3W$ with linear $N_R$ scheduled from 1 to 30 over 500 epochs.

*Table 4.* Comparison between different distances as loss for gradient flows.

| | Distance | Runtime (s) ↓ | NLL ↓ | $\log(W_2)$ ↓ |
|---|---|---|---|---|
| Mini-batch | $SSW$ | 46.10 ± 0.28 | -284.61 ± 8.44 | -1.19 ± 0.05 |
| | $S3W$ | 5.16 ± 0.62 | -205.13 ± 17.56 | -1.17 ± 0.055 |
| | $RI\text{-}S3W$ (1) | 5.50 ± 0.96 | -254.54 ± 9.39 | -1.18 ± 0.06 |
| | $RI\text{-}S3W$ (5) | 8.62 ± 0.94 | -305.15 ± 11.79 | -1.22 ± 0.05 |
| | $RI\text{-}S3W$ (10) | 8.75 ± 0.61 | -320.74 ± 8.34 | -2.80 ± 0.12 |
| | $ARI\text{-}S3W$ (30/1000) | 6.75 ± 0.12 | **-343.71 ± 2.45** | **-3.02 ± 0.17** |
| | $RI\text{-}S3W$ (1-30) | 7.67 ± 0.10 | -322.35 ± 7.69 | -2.96 ± 0.20 |
| | $ARI\text{-}S3W$ (1-30/1000) | **4.38 ± 0.19** | -264.55 ± 13.31 | -2.84 ± 0.15 |
| Full-batch | $SW$ | **0.64 ± 0.1** | -4891.85 ± 3.59 | -4.70 ± 0.22 |
| | $VSW$ | 0.66 ± 0.01 | -4858.75 ± 6.37 | -2.32 ± 0.08 |
| | $RI\text{-}S3W$ (1-30) | 26.90 ± 0.13 | -5003.21 ± 1.88 | **-6.68 ± 0.06** |
| | $ARI\text{-}S3W$ (1-30/1000) | 23.78 ± 0.36 | **-5048.43 ± 6.16** | -5.40 ± 0.03 |

*Figure 17.* Particle Evolution for mini-batch projected gradient descent.

### I.3. Study: Stability of Metrics w.r.t. $\epsilon$-cap

While in theory, one only needs to exclude the "north pole" from the stereographic projection (SP) operation, the image of points near the north pole under SP can be made arbitrarily large in norm. Numerically, this may result in overflows and make the optimization in our ML applications unstable. Hence, we fix an $\epsilon$-cap for SP which ensures the norm of projected samples are bounded; in particular, any point such that $x_{d+1} > 1 - \epsilon$ is first mapped to the circle $x_{d+1} = 1 - \epsilon$ prior to the SP operation.

In this section, we provide additional analysis of the effect of this $\epsilon$-cap on the proposed metrics. We generate two vMFs (target at the North pole and source at the South pole) with $\kappa = 50$ and $N = 2048$ samples each. We then compute each of the proposed metrics between the source and target distributions as a function of $\epsilon$. We visualize the results over 100 runs for each metric in Figure 18. We note that it is expected for the behavior of $S3W$ to differ from $RI\text{-}S3W$ and $ARI\text{-}S3W$. Importantly, we observe that all of the proposed methods are stable for a wide range of values for $\epsilon$. We observe no notable difference when we vary the number of projections $L \in \{128, 512\}$.

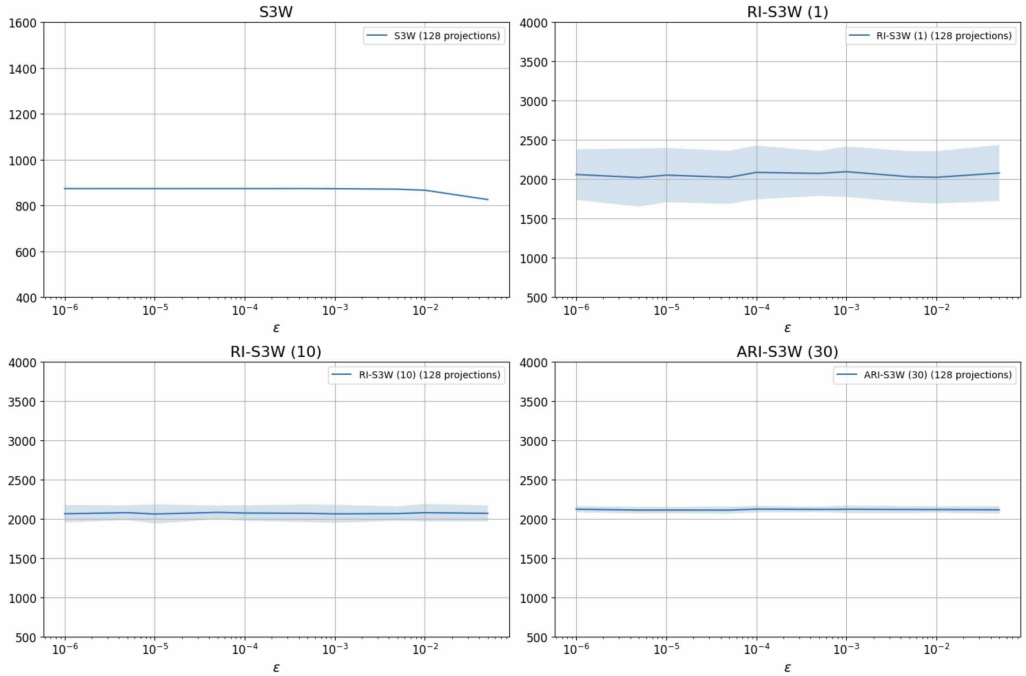



*Figure 18.* Stability of each metric w.r.t. $\epsilon$. Here, we use $N = 2048$ samples of each distribution and fix $L = 128$ projections.

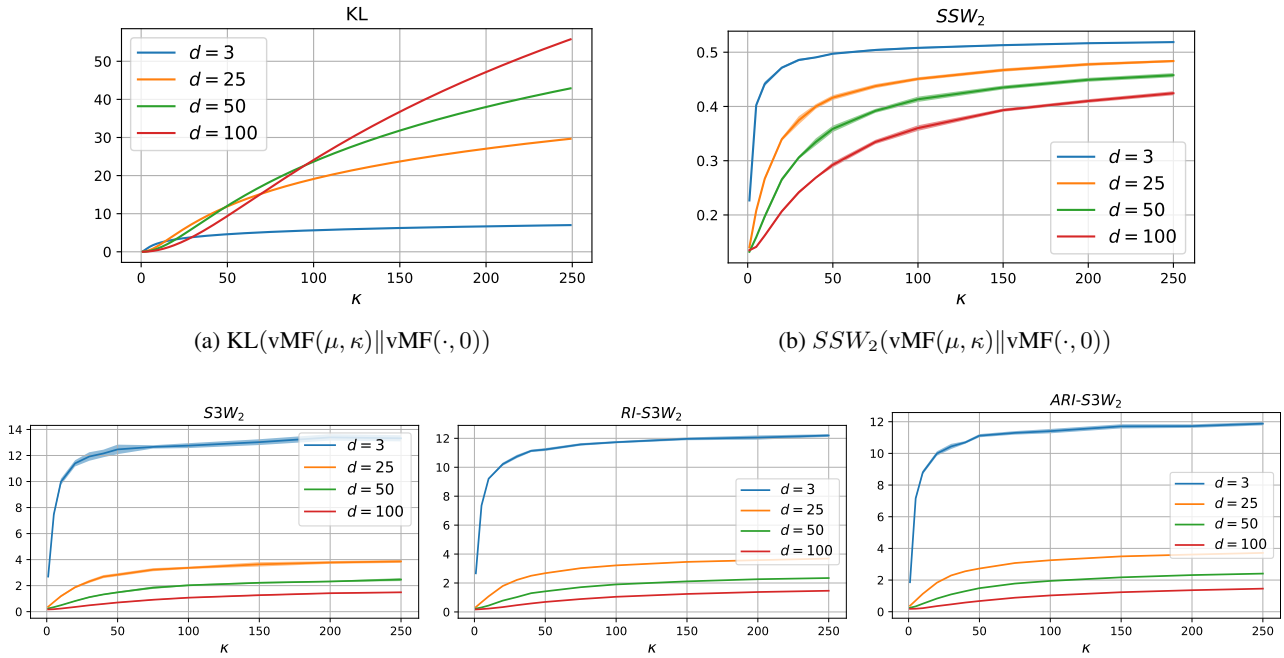
### I.4. Study: Evolution of the $S3W$ Distance

#### I.4.1. Experiment: Evolution of Metrics w.r.t. Iteration and Runtime

**Implementation** In this experiment, we aim to compare the evolution of different distances when used as a loss for gradient optimization. We randomly initialize a uniform source distribution $\hat{\mu} \sim \text{Unif}(\mathbb{S}^2)$ of 2400 samples. The objective is to learn a mixture $\hat{\nu}$ of 12 vMFs similar to Section 5.1, also consisting of 2400 samples. We use the Adam optimizer for projected gradient descent (see Section I.2) with a learning rate of $\gamma = 0.01$ for all distances. The $SSW$ baseline (Bonet et al., 2023a) will have an additional run with learning rate $\gamma' = 0.05$.

**Results** Figure 4 (in the main paper) shows comparative performance for: SSW, $S3W$, $RI\text{-}S3W$ (1 random rotation), $RI\text{-}S3W$ (5 random rotations), $ARI\text{-}S3W$ (30 random rotations, pool size of 1000). We show the evolution of the $\log(W_2)$ loss w.r.t. the cumulative runtime. We observe that at learning rate of 0.01, $SSW$ evolves nicely and starts converging around epoch 500, which is similar in performance with $RI\text{-}S3W$ (1) and $RI\text{-}S3W$ (5), albeit an order of magnitude slower. For a learning rate of 0.05, $SSW$ begins to display alternating behaviors but converges similarly to the case where learning rate is 0.01. $S3W$ is the fastest but slightly underperform other distances. $ARI\text{-}S3W$ (30) converges to the lowest loss and twice at fast as $SSW$ for both learning rates.

### I.4.2. Experiment: Evolution of Metrics w.r.t. $\kappa$ for Varying Dimensions




(a) $\mathrm{KL}(\mathrm{vMF}(\mu,\kappa)\|\mathrm{vMF}(\cdot,0))$ (b) $SSW_2(\mathrm{vMF}(\mu,\kappa)\|\mathrm{vMF}(\cdot,0))$

(c) $S3W_2(\mathrm{vMF}(\mu,\kappa)\|\mathrm{vMF}(\cdot,0))$ (d) $RI\text{-}S3W_2(\mathrm{vMF}(\mu,\kappa)\|\mathrm{vMF}(\cdot,0))$ (e) $ARI\text{-}S3W_2(\mathrm{vMF}(\mu,\kappa)\|\mathrm{vMF}(\cdot,0))$

*Figure 19.* Evolution between $\mathrm{vMF}(\mu,\kappa)$ and $\mathrm{vMF}(\cdot,0)$ w.r.t. $\kappa$ for varying dimension. We use $L = 200$ projections for all sliced metrics, $N_R = 100$ rotations for $RI\text{-}S3W$ and $ARI\text{-}S3W$, and pool size of 1000 for $ARI\text{-}S3W$. Each distribution has 500 samples. For each $\kappa \in \{1, 5, 10, 20, 30, 40, 50, 75, 100, 150, 200, 250\}$ we average each metric over 10 iterations.

We illustrate in Figure 19 the evolution of KL divergence, $SSW$, $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$ between $\mathrm{vMF}(\mu,\kappa)$ and $\mathrm{vMF}(\cdot,0)$ w.r.t. $\kappa$ for varying dimension. Just as (Bonet et al., 2023a) found that $SSW$ gets lower with the dimension contrary to KL divergence, we find that $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$ follow a similar trend.

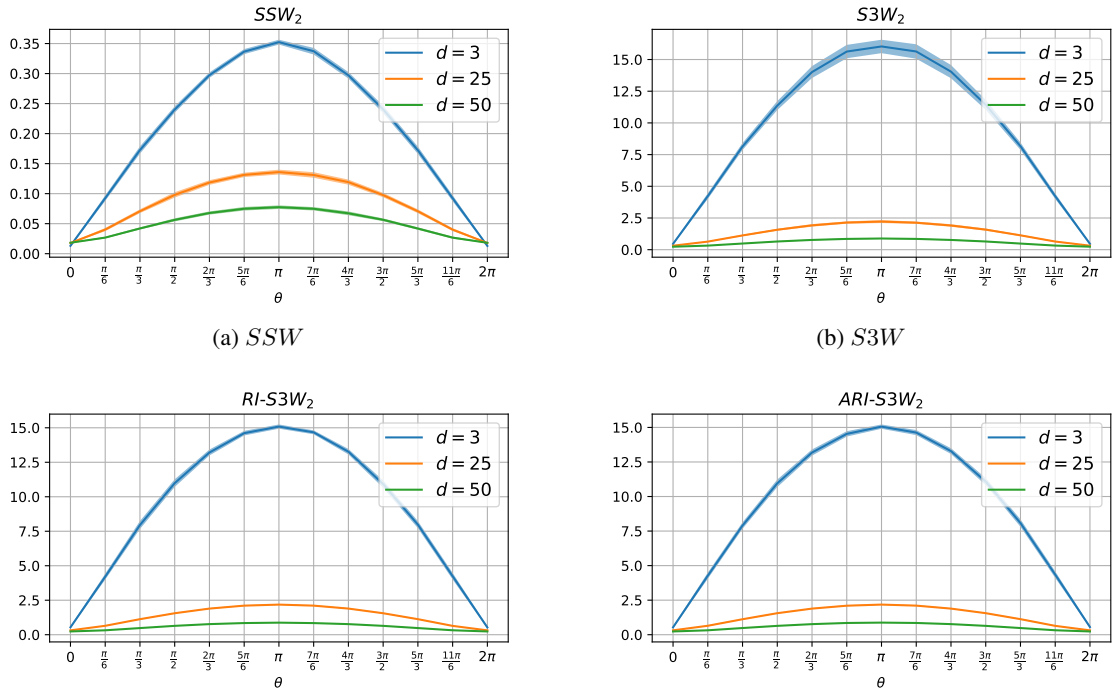
Here we use the analytic form for the KL divergence between the von Mises-Fisher distribution and the uniform distribution on $\mathbb{S}^d$ as derived in (Davidson et al., 2018; Xu & Durrett, 2018):

$$\begin{aligned}\mathrm{KL}\big(\mathrm{vMF}(\mu,\kappa)||\mathrm{vMF}(\cdot,0)\big) = \kappa\frac{I_{(d+1)/2}(\kappa)}{I_{(d+1)/2-1}(\kappa)} + \left(\frac{d+1}{2}-1\right)\log\kappa - \frac{d+1}{2}\log(2\pi) - \log I_{(d+1)/2-1}(\kappa)\\ + \frac{d+1}{2}\log\pi + \log 2 - \log\Gamma\left(\frac{d+1}{2}\right),\end{aligned} \tag{105}$$

where $I_v(\cdot)$ is the modified Bessel function of the first kind with order $v$, and $\Gamma(\cdot)$ is the gamma function.

### I.4.3. Experiment: Evolution of Metrics Between Rotated von Mises-Fisher Distributions

Similar to (Bonet et al., 2023a), we also compare the evolution of $SSW$, $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$ between a fixed vMF distribution and the rotation of this distribution along a great circle. That is, we compute each metric between $\mathrm{vMF}(\mu_0,\kappa_0)$ and $\mathrm{vMF}((\cos\theta,\sin\theta,0,\dots),\kappa_0)$ for $\theta \in \{\frac{k\pi}{6}\}_{k=0}^{12}$, where $\mu_0 = (1,0,0,\dots)$ and $\kappa_0 = 10$. We observe similar behavior between all metrics, with each distance being maximal between $\mathrm{vMF}(\mu_0,\kappa_0)$ and $\mathrm{vMF}(-\mu_0,\kappa_0)$ (corresponding to $\theta = \pi$).




(a) $SSW$ (b) $S3W$

(c) $RI\text{-}S3W$ (d) $ARI\text{-}S3W$

*Figure 20.* Evolution between vMF distributions on $\mathbb{S}^{d-1}$. Here we use $L = 200$ projections for all metrics, $N_R = 100$ rotations for $RI\text{-}S3W$ and $ARI\text{-}S3W$, and a pool size of 1000 for $ARI\text{-}S3W$. We average each metric over 100 iterations and use 500 samples for each distribution.

### I.4.4. Experiment: Evolution of $S3W$ w.r.t. Number of Slices and Number of Rotations

We perform several experiments to understand how $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$ evolve w.r.t. the number of slices used, the number of random rotations used, and the pool size used, respectively. In each of the following experiments, we measure the distance between a source uniform distribution $\text{vMF}(\cdot, 0)$ and a target von Mises-Fisher distribution $\text{vMF}(\mu, \kappa)$, where we sample 500 points from each distribution.

Figure 21 demonstrates that beyond $L = 100$ projections, variance becomes negligible across different dimensions and values of $\kappa$. Figure 22 demonstrates that $RI\text{-}S3W$ is generally stable for $R = 10$ rotations across varying dimensions and $\kappa$. Finally, Figure 23 demonstrates that variance is stable beyond a pool size of 20. In all plots, we average the metric over 20 iterations.

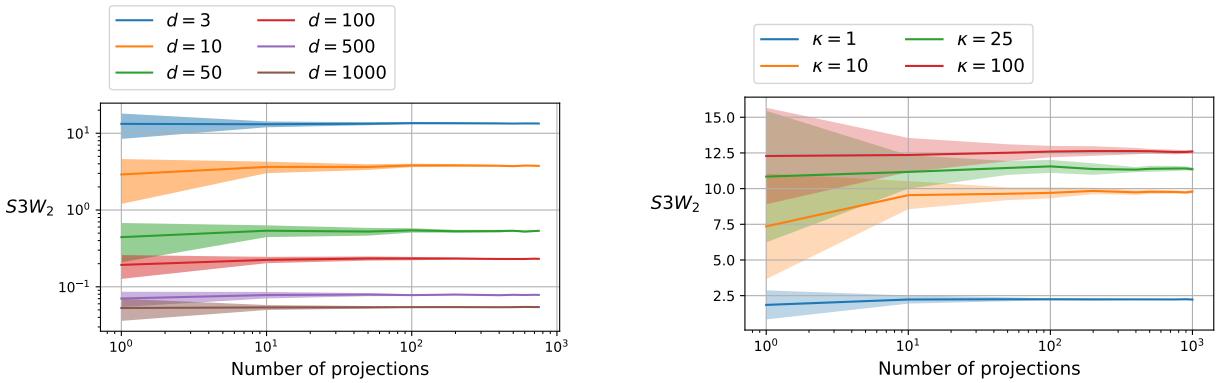



*Figure 21.* Evolution of $S3W$ between the source and target vMFs w.r.t. number of projections used. For the plot on the left, we fix $\kappa = 10$ for the target distribution, and for the plot on the right we use $d = 3$.

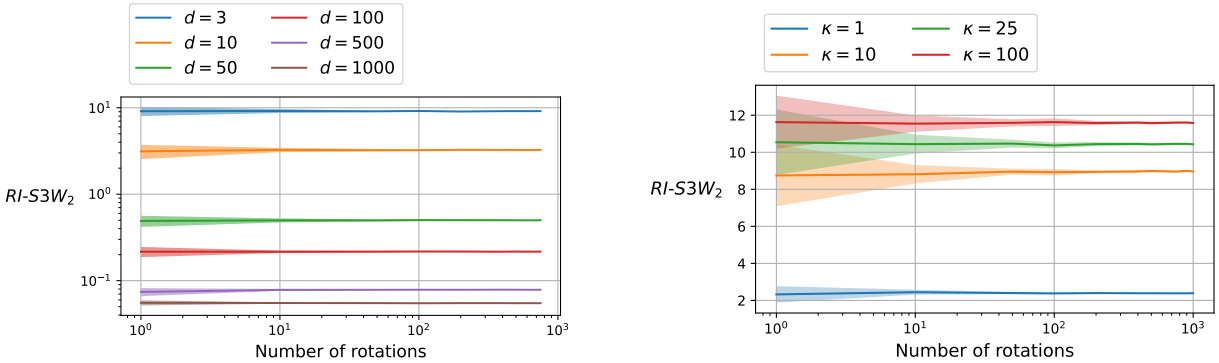



*Figure 22.* Evolution of $RI\text{-}S3W$ between the source and target vMFs w.r.t. number of rotations used. We fix $L = 10$ projections. For the plot on the left, we fix $\kappa = 10$ for the target distribution, and for the plot on the right we use $d = 3$.

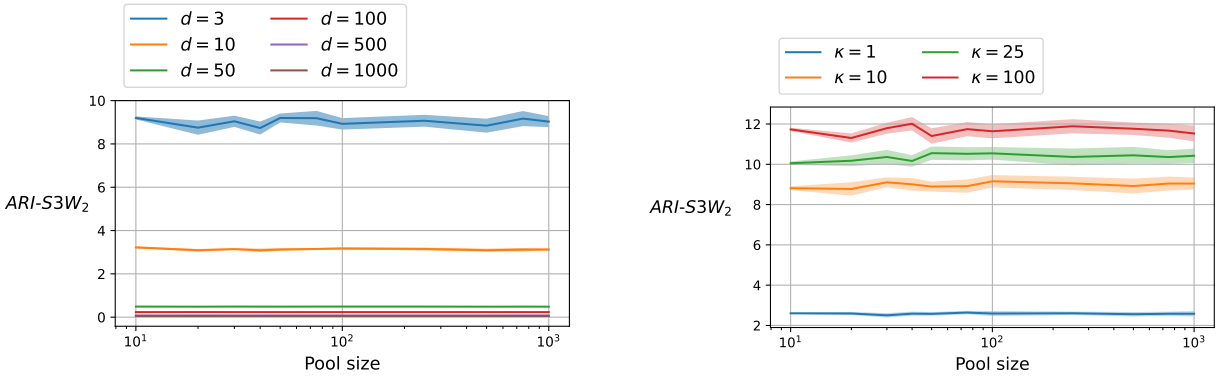
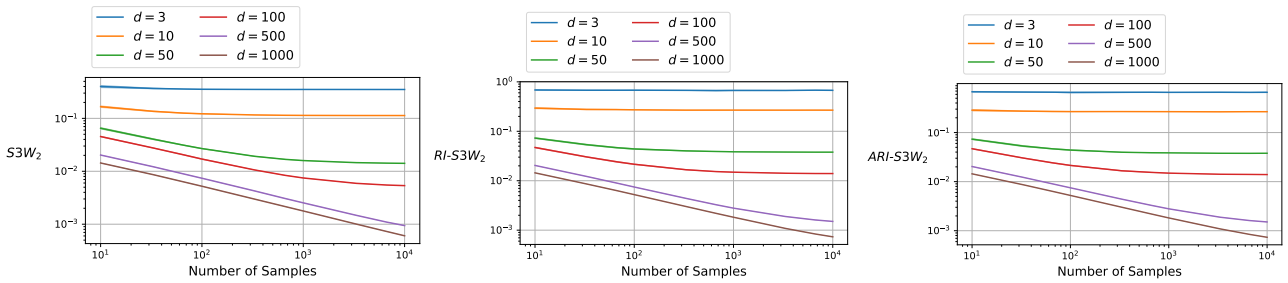



*Figure 23.* Evolution of $ARI\text{-}S3W$ between the source and target vMFs w.r.t. pool size used. We fix $L = 10$ projections and $N_R = 10$ rotations. For the plot on the left, we fix $\kappa = 10$ for the target distribution, and for the plot on the right we use $d = 3$.

### I.4.5. Experiment: Sample Complexity of Metrics

We perform a series of experiments to better understand the sample complexity of $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$. In application, since we employ empirical probability measures to approximate their continuous versions, we seek to understand how each of the proposed metrics evolves as we vary the number of supports in the empirical distributions.

In each of the following experiments, we measured the proposed distances between $N$ samples from two von Mises-Fisher distributions, centered at the north and south pole of $\mathbb{S}^d$ with $\kappa = 10$. We compute the distance as a function of the number of samples for $d \in \{3, 10, 50, 100, 500, 1000\}$, repeating each computation over 50 iterations. The resulting average and standard deviation are visualized in Figure 24.

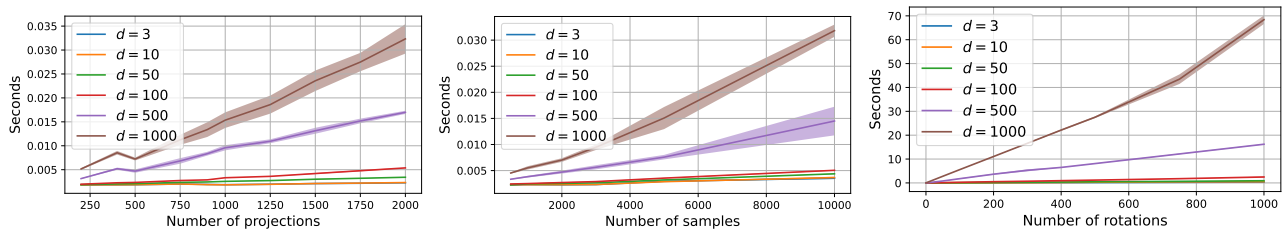



(a) $S3W$ (b) $RI\text{-}S3W$ (c) $ARI\text{-}S3W_2$

*Figure 24.* Sample complexity of $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$. Distances are computed between two vMF distributions on $\mathbb{S}^d$. We use $L = 1000$ projections for all metrics, $N_R = 10$ rotations for $RI\text{-}S3W$ and $ARI\text{-}S3W$, and a pool size of 100 for $ARI\text{-}S3W$. Each distance is measured over 50 iterations.

### I.5. Study: Runtime Analysis

In this section, we provide additional runtime analysis of the proposed method w.r.t. varying parameters. Figure 25 illustrates a comparative analysis of the runtime for calculating $S3W$ and $RI\text{-}S3W$ as we vary the number of projections, samples, and rotations, across different dimensions $d$ of the data. In each of these experiments, we measure the distance between a source uniform distribution $\text{vMF}(\cdot, 0)$ and a target von Mises-Fisher distribution $\text{vMF}(\mu, 10)$. The figure indicates a linear or quasi-linear relationship between the runtime and each parameter. Moreover, in Figure 26 we illustrate the time required to generate a pool of rotation matrices in $\text{SO}(d+1)$ as we vary the size of the pool and the dimension $d$. As expected, the runtime is linear in the size of rotation pool.




(a) (b) (c)

*Figure 25.* Runtime w.r.t. varying number of projections, samples, and rotations, respectively. In 25a, 25c we use $N = 500$ samples of each distribution; in 25b,25c we fix $L = 100$ projections; and in 25a,25b we do not use any rotations.

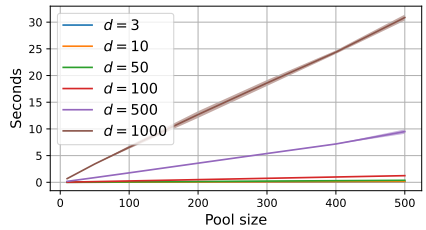



*Figure 26.* Runtime to generate pool of rotation matrices of varying sizes and dimension.

### I.6. Task: Density Estimation with Normalizing Flows

#### I.6.1. Background Overview

**Normalizing Flows** offers a framework for constructing complex probability distributions from simple ones through invertible transformations. It works by restricting the model class such that the likelihood function can be evaluated via simple sampling. Suppose we have access to the samples $\{x_i \sim \mu\}_{i=1}^n$ where $\mu$ is our target distribution. Let $p_Z$ be a base distribution over the latent $z$ (i.e. a Gaussian). We aim to find an invertible transformation $T$ such that the pushforward $T_{\#}p_\mu = p_Z$. Then, the density $p_\mu$ can be obtained through the change of variable formula

$$p_\mu(x) = p_Z(T(x))|\det J_T(x)|^{-1}, \tag{106}$$

where $|\det J_T(x)|$ denotes the determinant of the Jacobian of $T$ at $x$.

More concretely, consider the dataset $\mathcal{D}_{\text{train}} = \{x_1, ..., x_n\}$ sampled i.i.d from the empirical measure $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where $\delta_{x_i}$ is the Dirac delta at $x_i$. The log likelihood for training the neural network is given by

$$\log p(\mathcal{D}_{\text{train}}|w) = \sum_{i=1}^n \log p_X(x_i|w) = \sum_{i=1}^n \{\log p_Z(T(x_i)) - \log|\det J(x_i)|\} \tag{107}$$

Here, $T^{-1}$ represents the inverse transformation of $T$, and $|\det J(x_i)|$ is the determinant of the Jacobian of $T^{-1}$ at $x_i$.

**Real NVP and Stereographic Projection-based Normalizing Flows with Real NVP:**

Real NVP (Real-valued Non-Volume Preserving) (Dinh et al., 2016) uses a series of invertible layers to transform the base distribution into a more complex one. Each layer, known as a coupling layer, is designed such that its Jacobian is easy to compute. Specifically, given an input vector $x \in \mathbb{R}^m$, the coupling layer splits it into two parts $x_1 \in \mathbb{R}^{m_1}$ and $x_2 \in \mathbb{R}^{m_2}$ such that $m_1 + m_2 = m$. The transformation applied to $x_2$ is then conditioned on $x_1$ and defined by

$$x_2' = \exp(s(x_1)) \odot x_2 + t(x_1), \tag{108}$$

where $s : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$ and $t : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$ are scale and translation functions realized through neural networks. The operation $\odot$ denotes the Hadamard product. The key advantage of Real NVP is the traceability of both its inverse and the Jacobian determinant, that is

$$\log|\det J_T(x)| = \sum s(x_1) \tag{109}$$

which makes it a good candidate for high-dimensional density estimation.

To generalize to $\mathbb{S}^d$, (Gemici et al., 2016) proposes using a transformation $T$ composed of a stereographic projection $\phi$ (see section 2) followed by a real NVP flow $f$, denoted as $T = f \circ \phi$. The log density of the target distribution under this transformation is given by

$$\log p(x) = \log p_Z(z) + \log|\det J_f(z)| - \frac{1}{2}\log|\det(J_T\phi^{-1}J_{\phi^{-1}}(\phi(x)))|, \tag{110}$$

where $p_Z$ is the density of a prior on $\mathbb{R}^d$, typically a standard Gaussian, and $J_f(z)$ is the Jacobian of the Real NVP flow. The log det term accounts for the change in volume due to the stereographic projection $\phi$ and its inverse $\phi^{-1}$.

**Exponential Map and Exponential Map Normalizing Flows:**

Exponential maps serve as a tool for mapping data residing on curved spaces. In essence, it maps a vector in the tangent space of a point on a manifold $\mathcal{M}$ to the manifold itself. Specifically, for a point $x \in \mathcal{M}$ and a tangent vector $v_x$, the exponential map $\exp_x(v)$ traces a geodesic (i.e. shortest path) on $\mathcal{M}$. For the spherical manifold $\mathbb{S}^d$, it is given by

$$\exp_x(v) = x\cos(|v|) + \frac{v}{|v|}\sin(|v|). \tag{111}$$

Inspired by this, (Rezende et al., 2020) introduces the exponential map normalizing flows, which ensures that the construction of transformations happens directly on the manifold, avoiding the need for intermediate, non-diffeomorphic set mappings to Euclidean spaces (i.e. via the Stereographic projection). In this frame work, the transformation $T$ is defined by applying the exponential map to the projected gradient of a scalar field $\phi(x)$, that is

$$T(x) = \exp_x(\text{Proj}_x(\nabla\phi(x))) \tag{112}$$

where $\text{Proj}_x$ projects $\nabla\phi(x)$ onto the tangent space at $x$, $T_x\mathbb{S}^d$. The scalar field $\phi(x)$ is a convex combination of simple functions:

$$\phi(x) = \sum_{i=1}^{K} \alpha_i \beta_i e^{\beta_i(x^T\mu_i - 1)}, \tag{113}$$

with constraints $\alpha_i \geq 0$, $\sum_{i=1}^{K} \alpha_i \leq 1$, each $\beta_i > 0$, and $\mu_i \in \mathbb{S}^d$. The density update is then computed as

$$p(T(x)) = \frac{\pi(x)}{\sqrt{\det(E(x)^T J_T(x)^T J_T(x) E(x))}}, \tag{114}$$

where $J_T(x)$ is the Jacobian of $T$ at $x$, and $E(x)$ forms an orthonormal basis of $T_x\mathbb{S}^d$.

#### I.6.2. Experiment: Estimating Density on the Earth datasets

Our main focus is on $\mathbb{S}^2$ for the purpose of demonstration and visualization. Similar to (Bonet et al., 2023a), we use the three datasets introduced by (Mathieu & Nickel, 2020), representing the earth's surface as a perfect spherical manifold: volcano eruptions (NOAA, 2020), earthquakes (NOAA, 2020), and floods (Brakenridge, 2017). Our objective is to learn a normalizing flows model capable of reporting exact density at any point on the sphere, so that we can predict the likelihood of future events based on past data.

*Table 5.* Details of Earth datasets.

| | **Earthquake** | **Flood** | **Fire** |
|---|---|---|---|
| Train | 4284 | 3412 | 8966 |
| Test | 1836 | 1463 | 3843 |
| Total | 6120 | 4875 | 12809 |

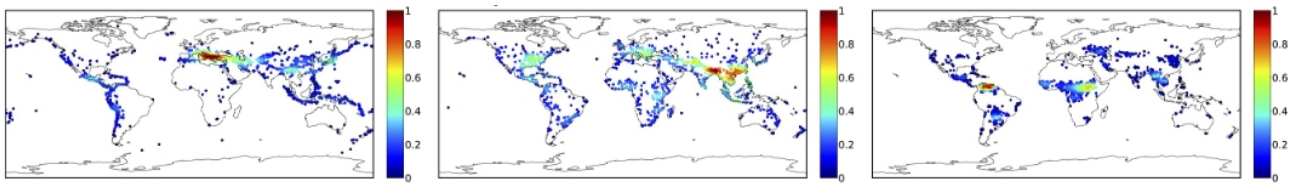



*Figure 27.* Ground truth as estimated with KDE (bandwidth 0.1) using test data (quakes, flood, fires, respectively). Densities are on a scale of 0 to 1.

Specifically, we demonstrate two approaches for training a normalizing flow model on spherical data. The first of which follows (Gemici et al., 2016), stereographically projecting the data from $\mathbb{S}^2$ onto $\mathbb{R}^2$ and then training a Real NVP model on $\mathbb{R}^2$. The second utilizes spherical exponential maps (Rezende et al., 2020) to train a normalizing flow natively on $\mathbb{S}^2$, using both $SW$ and $SSW$ (Bonet et al., 2023a) as baseline losses and comparing with our proposed S3W.

**Implementation** For exponential map normalizing flows, we construct the model using 48 radial blocks, each consisting of 100 components, totaling 24000 trainable parameters, similar to (Bonet et al., 2023a). Each block in the flow applies a transformation defined by a potential function $V$. The model projects the gradient of the potential function onto the sphere's tangent space and then maps it back to the sphere using the spherical exponential map.

For Stereographic Projection-based normalizing flows (Stereo-NF), we first project the manifold $\mathbb{S}^2$ onto $\mathbb{R}^2$ using the stereographic projection. Then, we train a Real NVP model, consisting of 10 Real NVP blocks with scaling and translation

parameterized by an MLP ($10$ layers of $25$ neurons each, $\text{LeakyReLU}_{0.2}(\cdot)$ activation). This setup follow (Bonet et al., 2023a) and has $27520$ parameters. We train the models for $20000$ epochs each using full batch gradient descent via the Adam optimizer. For exponential map NF, we use the learning rate of $0.1$ for $SSW$ and $SW$ as in the original setup and $0.05$ for our $S3W$and its variants. For Stereo-NF, we use a learning rate of $0.001$.

**Results** Figure 33 provides density maps of events (quakes, flood, fire from left to right), with the color scale representing the relative density on a scale from $0$ to $1$. We note that this normalization scheme does not change the underlying distribution but allows for an interpretable visualization, showing the relative differences in densities on a fixed scale. We observe that our proposed $S3W$ and its variants put relative density mass more accurately.

Table 3 provides the Negative Log-likelihood values for different distances. Despite having a higher parameter count, the increase in model complexity does not necessarily translate to better performance. This shows that estimating density natively on the sphere is superior for spherical data per this metric.

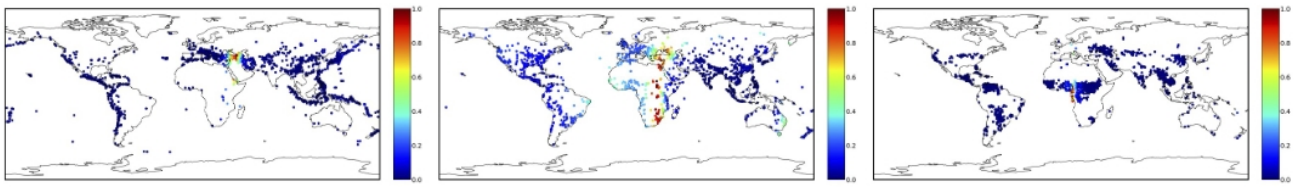


*Figure 28.* Stereo + RealNVP

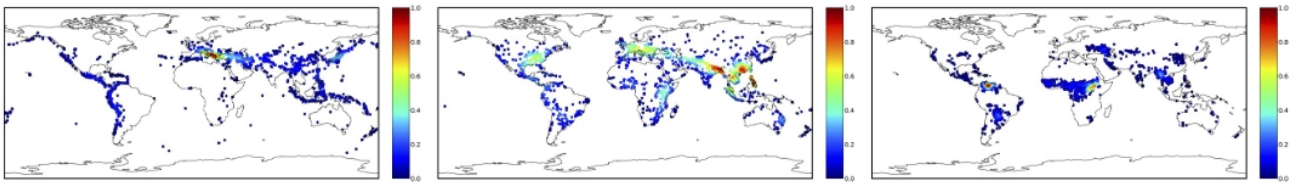


*Figure 29.* Sliced-Wasserstein ($SW$)

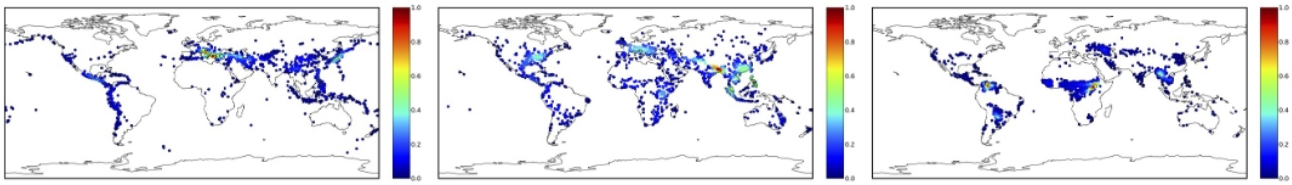


*Figure 30.* Spherical Sliced-Wasserstein ($SSW$) (Bonet et al., 2023a)

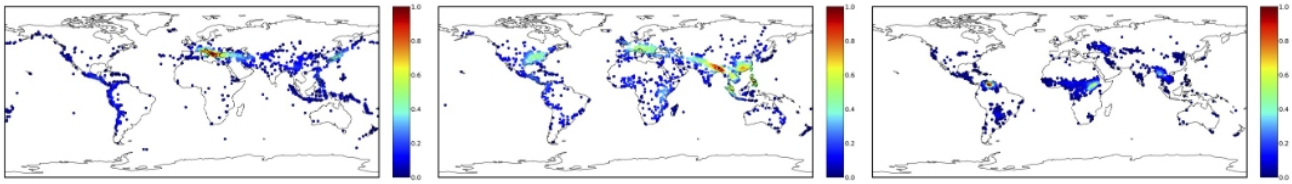


*Figure 31.* Stereographic Spherical Sliced-Wasserstein ($S3W$)

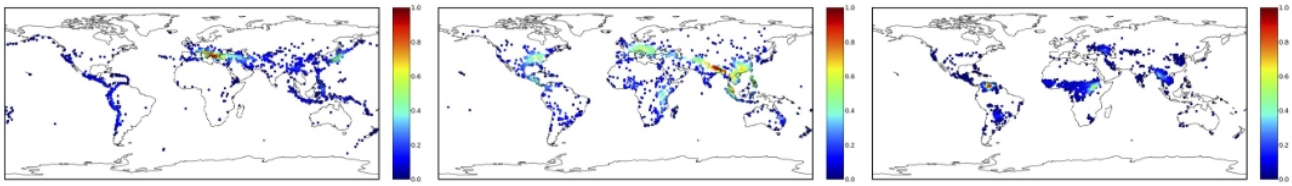


*Figure 32.* Rotation Invatiant $S3W$ ($RI$-$S3W$) with 1 random rotation

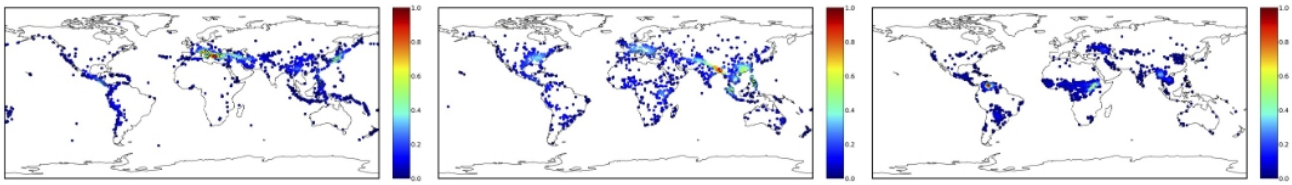


*Figure 33.* Amortized Rotation Invatiant $S3W$ ($ARI$-$S3W$) with 50 random rotations, pool size of 1000

### I.7. Task: Self-Supervised Representation Learning on the Sphere

The self-supervised learning (SSL) paradigm derives learning signals from its own training data without explicit labeling. Among various SSL approaches, contrastive learning-based methods have gained significant popularity due to their effectiveness. Previous works have found that constraining feature vectors to the hypersphere (i.e., having unit norm) offers additional benefits (Caron et al., 2020; Wu et al., 2018). In what follows, we provide a brief overview of (spherical) contrastive learning – currently one of the most potent approaches in SSL – and then outline the details of our experiment in Section I.7.2.

#### I.7.1. Background Overview

**Contrastive Learning with InfoNCE** in the SSL paradigm is fundamentally about distinguishing between similar (positive) and dissimilar (negative) pairs of data points. The objective is to minimize the distance between embeddings of similar data points while maximizing the distance between embeddings of dissimilar points. Let $p_{data}(\cdot)$ be the data distribution over $\mathbb{R}^m$ and $p_{pos}(\cdot,\cdot)$ the distribution of positive pairs over $\mathbb{R}^m \times \mathbb{R}^m$. Assuming symmetry, i.e., $\forall x, y : p_{pos}(x,y) = p_{pos}(y,x)$, and that the marginal distribution matches the data distribution, i.e., $\forall x : \int p_{pos}(x,y)dy = p_{data}(x)$, the popular InfoNCE (Oord et al., 2018) objective can be described as

$$\mathcal{L}_{\text{contrastive}} = -\mathbb{E}(x,y) \sim p_{\text{pos}} \left[ \log \frac{\exp(\text{sim}(f(x), f(y))/\tau)}{\sum_{k \in \text{Neg}(x)} \exp(\text{sim}(f(x), f(x_k))/\tau)} \right], \tag{115}$$

where $f(x)$ and $f(y)$ denote the embedding of the positive pairs, $\text{sim}(\cdot,\cdot)$ a similarity measure between two embeddings (e.g., the dot product), $\tau$ the temperature, and $\text{Neg}(x)$ the set of negative samples for $x$.

**Spherical Contrastive Loss as Alignment and Uniformity:**

(Wang & Isola, 2020) provides valuable insights into the mechanics of contrastive learning by decomposing the contrastive loss into two components

$$\mathcal{L}_{\text{align}}(f;\alpha) = -\mathbb{E}_{(x,y)\sim p_{\text{pos}}} \left[ \|f(x) - f(y)\|_\alpha^2 \right], \quad \alpha > 0. \tag{116}$$

and

$$\mathcal{L}_{\text{uniform}}(f;t) = \log \mathbb{E}_{x,y \text{ i.i.d.} \sim p_{\text{data}}} \left[ \exp(-t\|f(x) - f(y)\|_2^2) \right], t > 0. \tag{117}$$

The overall objective is therefore

$$\mathcal{L}_{\text{contrastive}}(f;\alpha,\beta) = \alpha \mathcal{L}_{\text{align}} + \beta \mathcal{L}_{\text{uniform}} \tag{118}$$

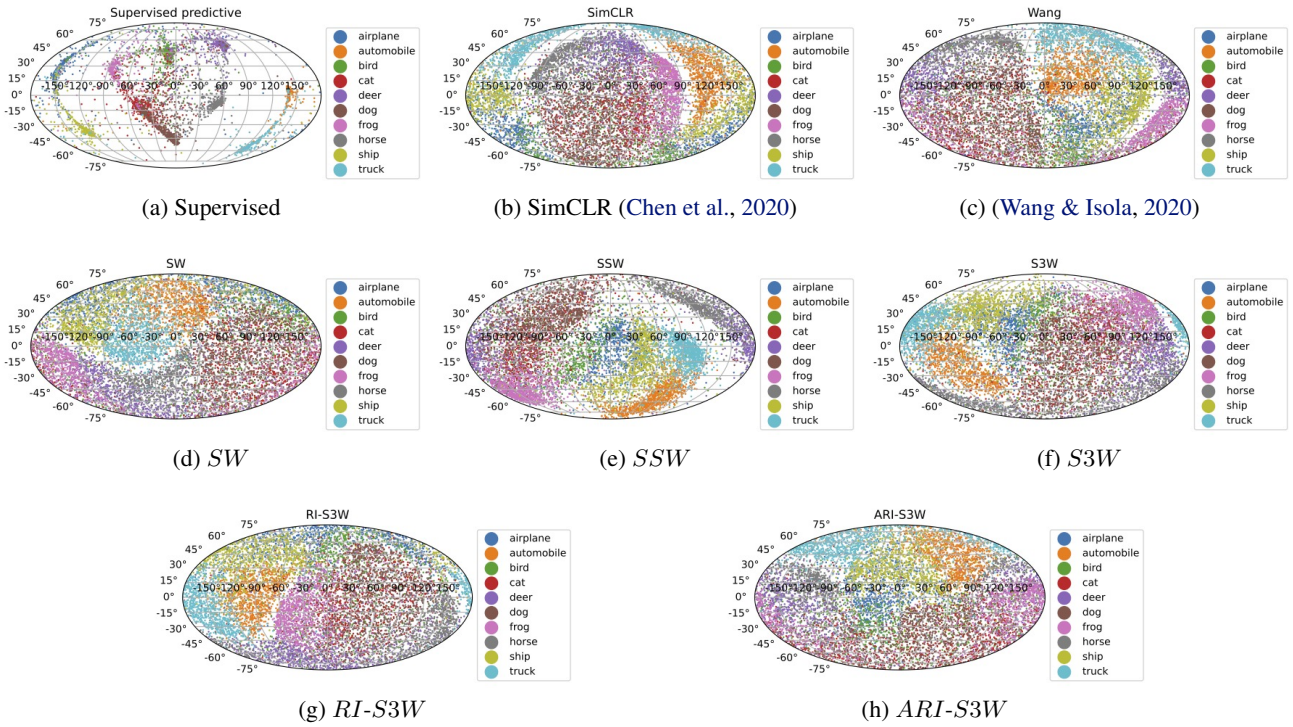
where $\alpha, \beta$ denotes the component weight for $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{uniform}}$, respectively. The alignment component $\mathcal{L}_{\text{align}}$ forces positive spherical views to be similar. The uniformity component $\mathcal{L}_{\text{uniformity}}$ encourages all spherical views to be spread out (i.e. uniformly distributed) on the hypersphere, which prevents representation collapse.

#### I.7.2. Experiment: $S3W$-Based Uniformity Loss

To demonstrate the efficacy of our proposed $S3W$ in SSL, we follow the same setup as described in (Bonet et al., 2023a). Specifically, they propose to replace the Gaussian kernel in $\mathcal{L}_{\text{uniform}}$, which operates on pairwise instances, with the distributional distance SSW, and optimize the following objective

$$\mathcal{L}_{\text{SSW-SSL}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \|z_i^A - z_i^B\|_2^2}_{\text{Alignment loss}} + \frac{\lambda}{2}\big(\underbrace{SSW_2^2(z^A,\nu) + SSW_2^2(z^B,\nu)}_{\text{Uniformity loss}}\big), \tag{119}$$

where $z^A, z^B \in \mathbb{R}^{n\times(d+1)}$ are the spherical representations of two views (i.e. augmentation) of the same images, $\nu = \text{Unif}(\mathbb{S}^d)$ is the uniform distribution on the hypersphere and $\lambda > 0$ is the regularization coefficient. They achieved comparable performance with (Wang & Isola, 2020; Chen et al., 2020) while benefiting from the subquadratic $O(Ln(d + \log n))$.




(a) Supervised (b) SimCLR (Chen et al., 2020) (c) (Wang & Isola, 2020)

(d) $SW$ (e) $SSW$ (f) $S3W$

(g) $RI\text{-}S3W$ (h) $ARI\text{-}S3W$

*Figure 34.* Evolution between the source and target vMFs, 500 samples each. For all sliced-Wasserstein variants, we use 200 projections, and for RI-S3W, we use 100 random rotations.

We will demonstrate the effectiveness of $S3W$ in terms of performance and runtime by modifying Eq. (119) and instead optimize

$$\mathcal{L}_{\text{S3W-SSL}} = \frac{1}{n}\sum_{i=1}^{n}\|z_i^A - z_i^B\|_2^2 + \frac{\lambda}{2}\left(\text{S3W}_2(z^A,\nu) + \text{S3W}_2(z^B,\nu)\right). \tag{120}$$

We use analogous objectives for $RI\text{-}S3W$ and $ARI\text{-}S3W$.

**Implementation** Similar to (Bonet et al., 2023a), we use a pretrained ResNet18 encoder (He et al., 2016) on CIFAR-10 (Krizhevsky, 2009) with $1024$ dimensional penultimate features, projected and $\ell_2-$normalized to be unit vectors on $\mathbb{S}^d$. The models are pretrained for $200$ epochs with minibatch SGD (momentum $0.9$, weight decay $0.001$, initial learning rate $0.05$). We select the batch size to be $512$ samples and use the standard random augmentation set consisting of random crop, horizontal flipping, color jittering, and gray scale transformation, as done in (Bonet et al., 2023a; Wang & Isola, 2020).

For evaluation, we fit a linear classifier on the pretrained representations and report the test accuracy. We report the accuracy both when this linear classifier is fit on the encoded feature representations and when it is fit on the projected features in $\mathbb{S}^d$. We train this linear layer for $100$ epochs with the Adam optimizer (learning rate $0.001$, weight decay of $0.2$ at epochs $60$ and $80$). For comparison, we also test the hypersphere method in (Wang & Isola, 2020) and SimCLR (Chen et al., 2020) to aid in evaluating the results. We also train a fully supervised model by training the encoder and linear classifier jointly with cross entropy loss, in order to serve as a baseline for performance comparison.

We use $L = 200$ projections for all sliced distances, $N_R = 5$ rotations for $RI\text{-}S3W$ and $ARI\text{-}S3W$, and a pool size of $100$ for $ARI\text{-}S3W$. We first perform a series of experiments with $d = 9$. In this setting, we let the regularization coefficient $\lambda = 0.5$ for $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$; $\lambda = 20.0$ for $SSW$; and $\lambda = 1.0$ for $SW$. These results are found in the main paper Table 1. We also perform a series of experiments with $d = 2$ in order to visualize the quality of the learned representations. In this setting, we instead use a regularization coefficient of $\lambda = 0.1$ for $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$. These visualizations for all the considered methods are provided in Figure 34.

### I.8. Task: Sliced-Wasserstein Variational Inference on the Sphere

#### I.8.1. Background Overview

**Variational Inference** turns the Bayesian inference problem into a variational one, benefiting from the rich optimization literature and reduced computational costs. The usual goal is to approximate an unnormalized distribution by minimizing the KL divergence between a tractable density and the true posterior. Formally, let $p(\cdot|x)$ denote the target posterior and and $q \in \mathcal{Q}$ the approximate posterior from a family of tractable distributions. The standard objective is to minimize

$$\min_{q\in\mathcal{Q}} \mathrm{KL}(q||p(\cdot|x)) = \int q(Z)\log\left(\frac{q(Z)}{p(Z|x)}\right)dZ = \mathbb{E}_q[\log\left(\frac{q(Z)}{p(Z|x)}\right)]. \tag{121}$$

**Sliced-Wasserstein Variational Inference (on the sphere):** The non-metric nature of KL, namely its asymmetry and failure to satisfy the triangle inequality, can result in undesirable behaviors. As an alternative, (Yi & Liu, 2022) introduces the sliced Wasserstein variational inference, optimized via MCMC without optimization or requiring a tractable $\mathcal{Q}$ family. Hence, this can be armortized with deep neural networks. In spherical data contexts, (Bonet et al., 2023a) recommends replacing $SW$ with $SSW$ and employs the Geodesic Langevin Algorithm (GLA) (see I.8.1).

To perform SWVI, we first select a sampler $q_\theta$ for the sphere. At each iteration $k$, we draw a batch of $N$ samples $\{z_0^i\}_{i=1}^N$from $q_\theta$. These samples are then propagated using $\ell$ MCMC steps on the sphere to obtain a new set of samples $\{z_\ell^j\}_{j=1}^N$. The updates are constrained to the sphere by employing the exponential map $\exp_{x_k}$ or normalization of the projection onto the tangent space $\mathrm{Proj}_{x_k}$ at each point $x_k \in \mathbb{S}^d$. We then compute sliced Wasserstein distance (i.e. SSW, S3W) between the empirical $\hat{\mu}_0$ and $\hat{\mu}_\ell$, which is then used to compute gradients w.r.t. $\theta$.

**Unadjusted Langevin Algorithm (ULA)** is used to sample from a probability distribution with density proportional to $e^{-V(x)}$ where $V(x)$ is a potential function (i.e. log density) and $x \in \mathbb{R}^d$. The update rule simply combines a deterministic gradient step with a stochastic noise term. That is

$$x_{t+1} = x_t - \gamma\nabla V(x_t) + \sqrt{2\gamma}Z, \quad Z \sim \mathcal{N}(0, I) \tag{122}$$

**Geodesic Langevin Algorithm (GLA)** (Wang et al., 2020) adapts ULA to a Riemannian manifold like $\mathbb{S}^d$ by modifying the update rule to respect the manifold structure. Similar to I.2.1, we first replace the Euclidean gradient with the Riemannian gradient, as follows

$$x_{t+1} = \mathrm{Retr}_{x_t}\left(-\gamma\nabla_{\text{Riemannian}}V(x_t) + \sqrt{2\gamma}Z\right), \quad Z \sim \mathcal{N}(0, I), \tag{123}$$

Here, $\mathrm{Retr}_{x_t}$ is a retraction. If we select $\mathrm{Retr}_{x_t} = \exp_{x_t}$ (see Eq. (111)) then we arrive at the Geodesic Langevin Algorithm (GLA) whose update rule is simply

$$x_{t+1} = \exp_{x_t}\left(-\gamma\nabla_{\text{Riemannian}}V(x_t) + \sqrt{2\gamma}Z\right), \quad Z \sim \mathcal{N}(0, I) \tag{124}$$

where $\nabla_{\text{Riemannian}}V(x_t) = \nabla V(x_t) - \langle\nabla V(x_t), x_t\rangle x_t$.

In our experiments, since $\mathbb{S}^d$ has constant curvature and the step size is small, we use normalization as the retraction for simplicity. The update rule is hence given by:

$$x_{t+1} = \frac{\tilde{x}_{t+1}}{\|\tilde{x}_{t+1}\|}, \quad \text{where} \quad \tilde{x}_{t+1} = x_t - \gamma\nabla_{\text{Riemannian}}V(x_t) + \sqrt{2\gamma}Z, \tag{125}$$

where $Z \sim \mathcal{N}(0, I)$ is the Gaussian noise term. Putting it all together, we arrive at

$$x_{t+1} = \frac{x_t - \gamma\left(\nabla V(x_t) - \langle\nabla V(x_t), x_t\rangle x_t\right) + \sqrt{2\gamma}Z}{\left\|x_t - \gamma\left(\nabla V(x_t) - \langle\nabla V(x_t), x_t\rangle x_t\right) + \sqrt{2\gamma}Z\right\|}. \tag{126}$$

**Power Spherical Distribution** (De Cao & Aziz, 2020) provides a parametric alternative to vMFs (refer to H2). It demonstrates greater stability and has a univariate marginal with a closed-form expression for both its CDF and its inverse. This enables fast and efficient sampling, in contrast to the vMF, which depends on rejection sampling. The PDF of the Power Spherical is defined as

$$p_X(x)(x;\mu,\kappa) \propto (1+\mu^\top x)^\kappa, \tag{127}$$

where $\mu \in \mathbb{S}^d$, $\kappa > 0$, for all $x \in \mathbb{S}^d$. We sample by drawing $Z \sim Beta(\frac{d-1}{2}+\kappa, \frac{d-1}{2})$ and $v \sim Unif(\mathbb{S}^{d-1})$. These samples are then used to construct $T = 2Z - 1$ and subsequently a vector $Y = [T, v^\top\sqrt{1-T^2}]^\top$. The final sample is obtained by applying a Householder[5] reflection about the mean direction $\mu$ to $Y$.

**Effective Sample Size (ESS)** estimates the equivalent number of independent samples that would provide the same amount of information as the correlated samples generated by an MCMC process (Doucet et al., 2001). For a chain of $n$ samples, the autocorrelation at lag[6] $t$ ($t \geq 0$) is given by

$$\rho_t = \frac{1}{\sigma^2}\int(\theta(n)-\mu)(\theta(n+t)-\mu)p(\theta)d\theta. \tag{128}$$

The ESS is then calculated from the total number of samples, $N$, adjusted for the sum of autocorrelation at all lags

$$ESS = \frac{N}{1+2\sum_{t=1}^{\infty}\rho_t} \tag{129}$$

Direct calculation for the ESS is often impractical. (Rezende et al., 2020) proposes estimating it using importance sampling weights $w_s$. That is

$$\text{ESS} = \frac{\text{Var}_{Unif}(e^{-\beta u(X)})}{\text{Var}_q\left(\frac{e^{-\beta u(X)}}{q_\eta(X)}\right)} \approx \frac{\left(\sum_{s=1}^{S} w_s\right)^2}{\sum_{s=1}^{S} w_s^2}, \tag{130}$$

where $w_s = e^{-\beta u(x_s)/q_\eta(x_s)}$. Here, in the context of normalizing flows (see experiments I.8.2, I.8.3), it is reported as a percentage of the sample size. If the ESS can be estimated reliably, then higher ESS means the flow matches the target better.

#### I.8.2. Experiment 1: Learning a Power Spherical Distribution

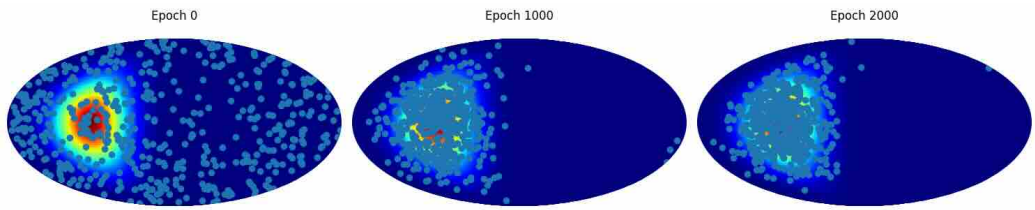
**Implementation** Following the setup described in (Bonet et al., 2023a), we set initial Power Spherical parameters $\mu = (1,1,1)$ and $\kappa = 0.1$ for source,$\mu = (0,1,0)$ and $\kappa = 10$ for target. We perform 2000 Riemannian Gradient Descent (Absil et al., 2008) steps and then 20 steps of GLA (step size 0.001) with 1000 projections (learning rate 2). We select $K = 2000$ steps with $N = 500$ particles.

We report runtimes and parameter convergence behaviors for all the distances. Although they all converge similarly, the runtimes differ drastically. For SSW, we report an average of 215.5 seconds over 5 independent runs. For $S3W$, $RI\text{-}S3W$ (1), $RI\text{-}S3W$ (5), $RI\text{-}S3W$ (10), $ARI\text{-}S3W$ (10 rotations, pool size 1000), they are 55.3, 57.3, 63.6, 73.1, 63.04 seconds , respectively.

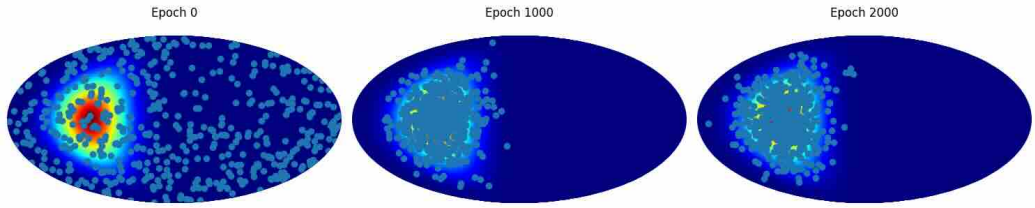
**Results**

[5]a linear operation $H(v) = v - 2u(u^\top v)$ reflecting $v$ across the hyperplane perpendicular to $u$.

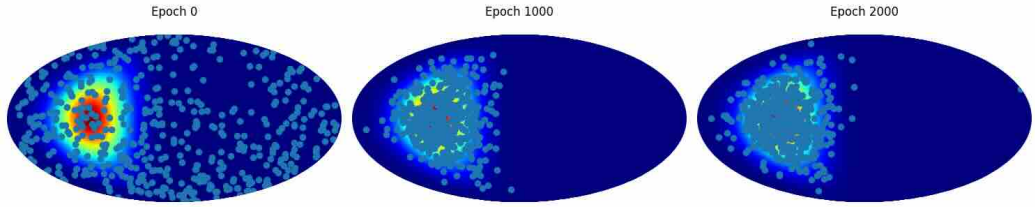
[6]the number of time steps separating sequential data points in a time series, used for measuring temporal correlation.



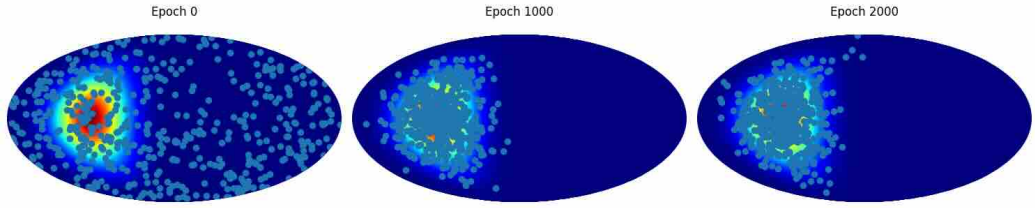

(a) $SSW$



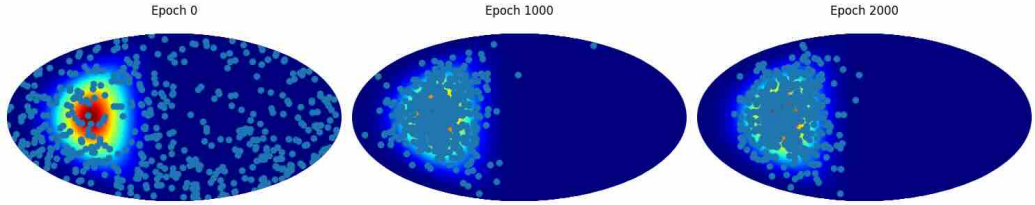

(b) $S3W$



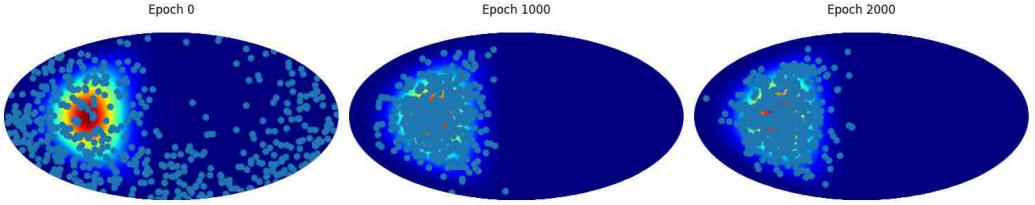

(c) $RI\text{-}S3W(1)$




(d) $RI\text{-}S3W(5)$




(e) $RI\text{-}S3W(5)$



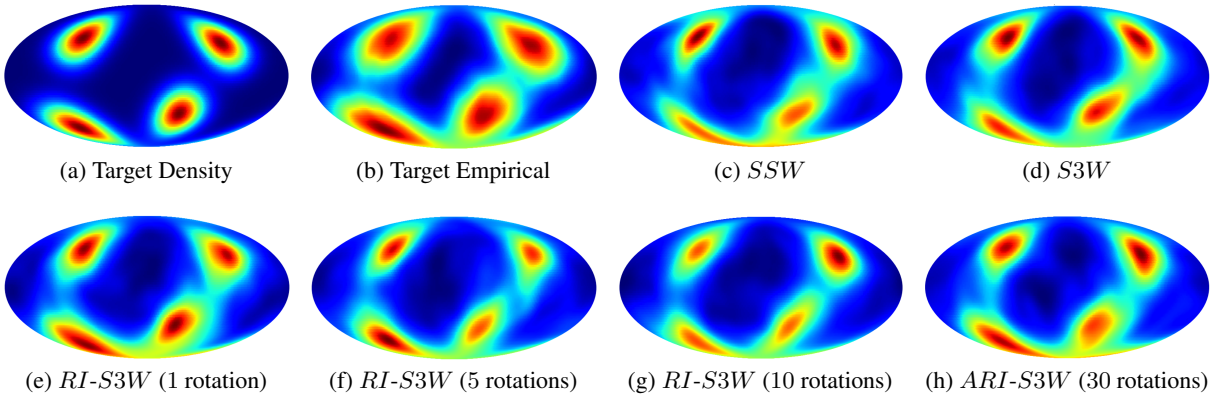

(f) $RI\text{-}S3W(10)$

*Figure 35.* Learning a mixture of 4 von Mises-Fisher distributions. To obtain the Mollweide projections, we perform Kernel Density Estimation with the Scott adaptive bandwidth.

### I.8.3. Experiment 2: Learning a Mixture of 4 von Mises-Fisher Distributions

**Implementation** We perform the Geodesic Langevin Algorithm (GLA) (see I.8.1) to generate samples that track the target distribution and used to guide our variational model, which is an exponential map normalizing flows $f_{\mu_T}(x; T_{exp})$ ($N = 6$ blocks, 5 components). The objective of $f_{\mu_T}$ is to transform uniform noise on the sphere into a close approximation of the target, which is a mixture of 4 von Mises-Fisher distributions. The parametric target has the mean directions $\mu_1 = (1.5, 0.7 + \frac{\pi}{2}), \mu_2 = (1.0, -1.0 + \frac{\pi}{2}), \mu_3 = (5.0, 0.6 + \frac{\pi}{2}), \mu_4 = (4.0, -0.7 + \frac{\pi}{2})$ and uniform concentration parameters $\kappa = 10$, equally weighted. The optimization is run for 10000 iterations with 20 GLA steps each, using the Adam optimizer (learning rate of 0.01, matching the baseline).

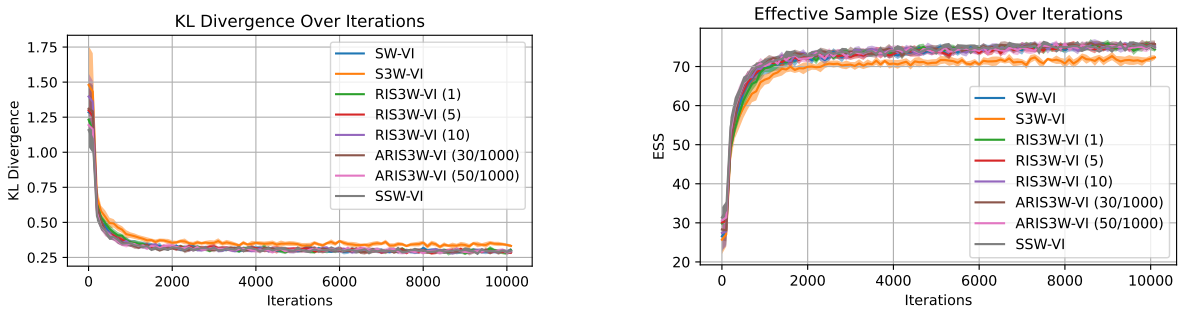
**Results** Figure 36 demonstrates the qualitative results for our experiments. While all models are capable of closely matching the target, it appears that $ARI$-$S3W$ (30), $RI$-$S3W$ (10), $RI$-$S3W$ (5) outperform $RI$-$S3W(1)$ , which fares better than $SSW$ and $S3W$.



(a) Target Density (b) Target Empirical (c) $SSW$ (d) $S3W$

(e) $RI$-$S3W$ (1 rotation) (f) $RI$-$S3W$ (5 rotations) (g) $RI$-$S3W$ (10 rotations) (h) $ARI$-$S3W$ (30 rotations)

*Figure 36.* Learning a mixture of 4 von Mises-Fisher distributions. To obtain the Mollweide projections, we perform KDE with the Scott adaptive bandwidth. For $ARI$-$S3W$, we use pool size of 1000.




(a) (b)

*Figure 37.* Evolution between the source and target vMFs, 500 samples each. For all sliced-Wasserstein variants, we use 200 projections, and for $ARI$-$S3W$, we use the pool size of 1000 random rotations.

To compare the performance of the different distances more rigorously, we use the KL Divergence and the Effective Sample Size (ESS, see I.8.1), similar to (Bonet et al., 2023a). We perform 10 runs per distance, and also add $ARI$-$S3W(50)$ for comparison. Figure 37 shows the convergence trend across different distances. The KL Divergence plot shows most models (with different distances) similarly and rapidly converge to approximating the target distribution. $S3W$ performs slightly worse than the rest while $SSW$ is on par with other variants of $S3W$, in contrast to our qualitative finding based on figure 36. The ESS plot demonstrates sampling efficacy. Again, similar trend is observed, with $S3W$ slightly underperforming other distances. We note that greater ESS values signal more independence among drawn samples.

Additionally, we tried the learning rate of 0.001 (not shown) for all distances and observed a slightly different trend. $SSW$ still outperforms $S3W$ and is on par with $RI$-$S3W(1)$, while underperforming the rest by a small margin.

### I.9. Task: Generative Modeling with Sliced-Wasserstein Autoencoder (SWAE)

Autoencoders (AEs) map an input $x \in \mathbb{R}^d$ back to itself through an intermediate latent space and play a central role in discovering low-dimensional representations useful for downstream tasks. Deterministic AEs focus on point estimates in the latent space, while their probabilistic counterparts extend this concept to probability distributions. A common requirement for both is regularization to avoid the trivial solutions. SWAEs (Kolouri et al., 2016b), a type of probabilistic AE, offer a simple yet effective approach to regularizing the latent space. In this section, we show the effectiveness of our proposed $S3W$ method when the SWAE latents are constrained to the unit sphere.

#### I.9.1. Experiment: $S3W$-Based Latent Regularization Loss

**Algorithm 5** S3W-AE

**Require:** Reg. coefficient $\lambda$, $L$ projections, encoder $f_\theta$, decoder $g_\eta$, dimension $d$ of the hypersphere.
**while** $f_\theta$ and $g_\eta$ have not converged **do**
Sample $x_1, ..., x_M$ from the training set (i.e., $p_X$).
Encode the samples: $z_1, ..., z_M \leftarrow f_\theta(x_1), ..., f_\theta(x_M)$.
Normalize encoded samples to $\mathbb{S}^d$: $\hat{z}_1, ..., \hat{z}_M \leftarrow \text{Normalize}(z_1, ..., z_M)$.
Sample $\tilde{z}_1, ..., \tilde{z}_M$ from $q_Z$ and normalize to $\mathbb{S}^d$.
Use algorithm 1 or its variants to calculate $\mathcal{L}^{(l)}_{\text{S3W}}(\{\tilde{z}_m\}, \{\hat{z}_m\})$.
Compute $\hat{\mathcal{L}}_{\text{S3W}} = \frac{1}{L}\sum_{l=1}^{L} \mathcal{L}^{(l)}_{\text{S3W}}$.
Update $\theta$ and $\eta$ by gradient descending

$$\hat{\mathcal{L}} = \sum_{m=1}^{M} c(x_m, g_\eta(\hat{z}_m)) + \lambda \cdot \hat{\mathcal{L}}_{\text{S3W}}$$

**end while**

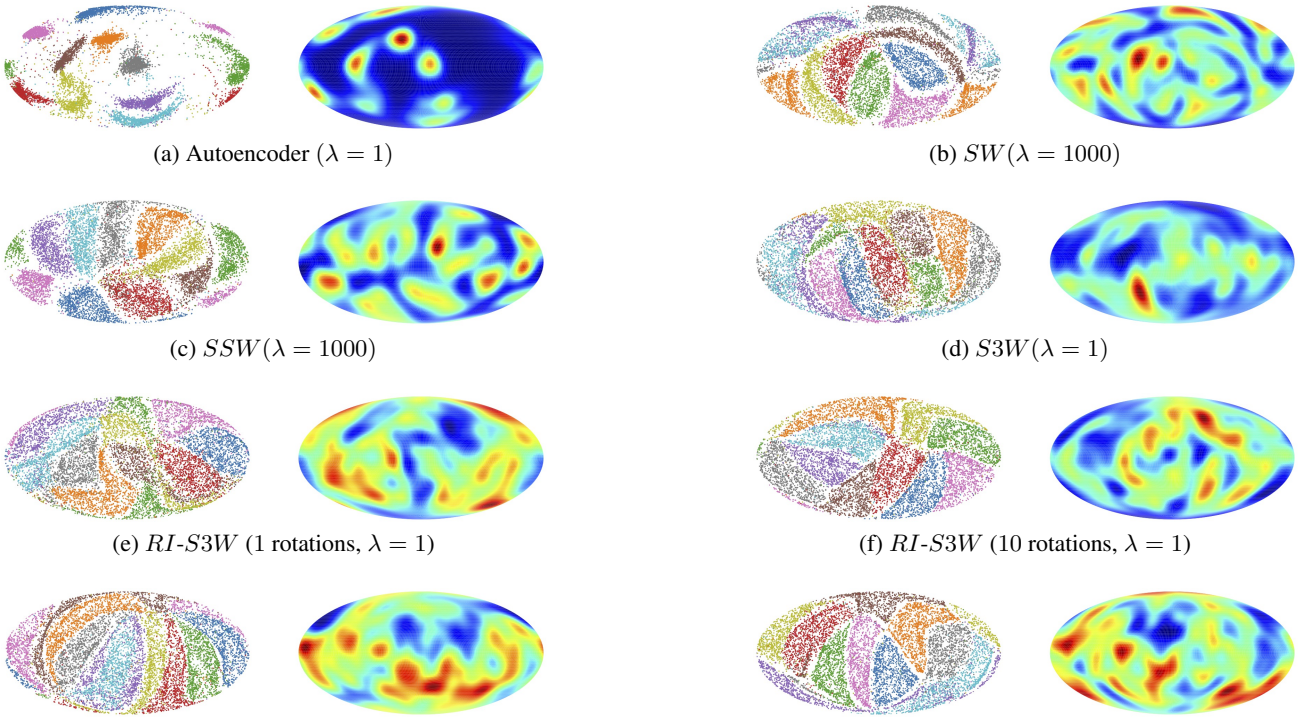
We perform experiments with both MNIST and CIFAR-10. The details of the architectures used in our experiments can be found in Table 7. In all experiments, we train the networks with the Adam optimizer (learning rate of $10^{-3}$) with a batch size of 500 over 100 epochs. The procedure is described in Algorithm 5.

For all experiments, we use the standard binary cross entropy (BCE) loss as our reconstruction loss, with a tradeoff parameter $\lambda$. For CIFAR-10, we use $\lambda = 10^{-3}$ for $SW$, $S3W$, $RI\text{-}S3W$, and $ARI\text{-}S3W$, and $\lambda = 10$ for $SSW$. We use $L = 100$ projections for all sliced methods, $N_R = 5$ rotations, and a pool size of 100 random rotations. Moreover, we use a vMF mixture with 10 components for the latent prior. These results are given in Table 2.

For MNIST, we provide a comprehensive evaluation of the considered methods as we vary the regularization parameter $\lambda$. These results can be found in Table 8. In these experiments, we once again use $L = 100$ projections for all sliced methods and a pool size of 100 for $ARI\text{-}S3W$. Here, we use the uniform distribution Unif($\mathbb{S}^2$) as our latent prior. Lastly, we repeat a set of experiments for each distance using the MNIST dataset and compare the Fréchet Inception Distance (FID) (Heusel et al., 2017) of each resulting model in Table 6. We compute the FID using 10000 samples and report the mean and standard deviation over 5 independent runs. Moreover, we visualize samples generated by each model in Figure 39. For these experiments, we again use $L = 100$ projections for all methods and a pool size of 100 for $ARI\text{-}S3W$, as well as a uniform prior on $\mathbb{S}^2$. We use $\lambda = 10^{-3}$ for all models for fair comparison.

*Table 6.* FID scores for SWAE models.

| Method | FID ↓ |
|---|---|
| $SW$ | 74.2301 ± 2.7727 |
| $SSW$ | 75.0354 ± 2.8609 |
| $S3W$ | 75.0717 ± 2.6026 |
| $RI\text{-}S3W$ (1) | 73.5115 ± 3.4937 |
| $RI\text{-}S3W$ (10) | 70.2262 ± 3.4730 |
| $ARI\text{-}S3W$ (30) | **69.5562 ± 1.5386** |



(a) Autoencoder ($\lambda = 1$)

(b) $SW(\lambda = 1000)$

(c) $SSW(\lambda = 1000)$

(d) $S3W(\lambda = 1)$

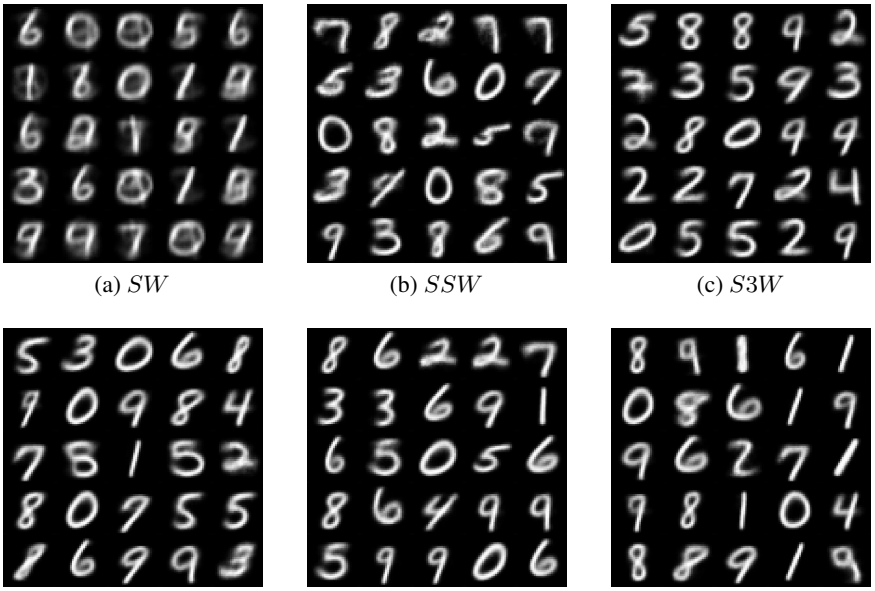
(e) $RI\text{-}S3W$ (1 rotations, $\lambda = 1$)

(f) $RI\text{-}S3W$ (10 rotations, $\lambda = 1$)

(g) $ARI\text{-}S3W$ (10 rotations, pool size of 100, $\lambda = 1$)

(h) $ARI\text{-}S3W$ (30 rotations, pool size of 100, $\lambda = 1$)

*Figure 38.* Latent space visualization (MNIST)



(a) $SW$

(b) $SSW$

(c) $S3W$

(d) $RI\text{-}S3W$ (1)

(e) $RI\text{-}S3W$ (10)

(f) $ARI\text{-}S3W$ (30)

*Figure 39.* Samples generated by SWAE models with a uniform prior on $\mathbb{S}^2$.

| Dataset | Encoder | Decoder |
|---|---|---|
| MNIST | $x \in \mathbb{R}^{28\times 28} \to \text{Conv2d}_{32} \to \text{ReLU}$ | $s \in \mathbb{S}^2 \to \text{FC}_{512} \to \text{FC}_{512} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{32} \to \text{ReLU}$ | $\to \text{Reshape}(128 \times 2 \times 2)$ |
| | $\to \text{Conv2d}_{64} \to \text{ReLU}$ | $\to \text{Conv2dT}_{128} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{64} \to \text{ReLU}$ | $\to \text{Conv2dT}_{64} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{128} \to \text{ReLU}$ | $\to \text{Conv2dT}_{64} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{128}$ | $\to \text{Conv2dT}_{32} \to \text{ReLU}$ |
| | $\to \text{Flatten} \to \text{FC}_{512} \to \text{ReLU}$ | $\to \text{Conv2dT}_{32} \to \text{ReLU}$ |
| | $\to \text{FC}_3 \to \ell_2$ normalization $\to s \in \mathbb{S}^2$ | $\to \text{Conv2dT}_1 \to \text{Sigmoid}$ |
| CIFAR-10 | $x \in \mathbb{R}^{3\times 32\times 32} \to \text{Conv2d}_{32} \to \text{ReLU}$ | $s \in \mathbb{S}^2 \to \text{FC}_{512} \to \text{FC}_{2048} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{32} \to \text{ReLU}$ | $\to \text{Reshape}(128 \times 4 \times 4)$ |
| | $\to \text{Conv2d}_{64} \to \text{ReLU}$ | $\to \text{Conv2dT}_{128} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{64} \to \text{ReLU}$ | $\to \text{Conv2dT}_{64} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{128} \to \text{ReLU}$ | $\to \text{Conv2dT}_{64} \to \text{ReLU}$ |
| | $\to \text{Conv2d}_{128}$ | $\to \text{Conv2dT}_{32} \to \text{ReLU}$ |
| | $\to \text{Flatten} \to \text{FC}_{512} \to \text{ReLU}$ | $\to \text{Conv2dT}_{32} \to \text{ReLU}$ |
| | $\to \text{FC}_3 \to \ell_2$ normalization $\to s \in \mathbb{S}^2$ | $\to \text{Conv2dT}_3 \to \text{Sigmoid}$ |

*Table 7.* Architecture of the Encoder and Decoder for MNIST and CIFAR datasets

*Table 8.* SWAE with different regularizations on the MNIST dataset with uniform prior for $d = 2$. Here we compare among the Sliced Wasserstein ($SW$), Spherical Sliced Wasserstein ($SSW$) (Bonet et al., 2023a), $S3W$, $RI\text{-}S3W$ (1 rotation), $RI\text{-}S3W$ (10 rotations), $ARI\text{-}S3W$ (10 rotations), $ARI\text{-}S3W$ (30 rotations, pool size of 100). The metrics are $log_{10}(W_2)$ and Cross Entropy (CE). For $SSW$, we include a wide range of regularization parameter $\lambda = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ for thorough comparison. For other methods, we include either $\lambda = \{0.001, 0.01, 0.1, 1\}$ or $\lambda = \{1, 10, 100, 1000\}$.

| Method | $\lambda$ | $\log(W_2)\downarrow$ | CE $\downarrow$ | Time(s/ep.) |
|---|---|---|---|---|
| Supervised AE | 1 | $-1.8438 \pm 0.1667$ | $0.1585 \pm 0.0018$ | $5.2500 \pm 0.0502$ |
| $SW$ | 1 | $-1.1294 \pm 0.0102$ | $0.1619 \pm 0.0608$ | $5.2113 \pm 0.1191$ |
| | 10 | $-1.1827 \pm 0.0096$ | $0.1652 \pm 0.0631$ | $5.4121 \pm 0.9013$ |
| | 100 | $-1.6033 \pm 0.0148$ | $0.1641 \pm 0.0631$ | $5.7512 \pm 0.3019$ |
| | 1000 | $-1.7856 \pm 0.0296$ | $0.1725 \pm 0.0668$ | $6.1400 \pm 0.4853$ |
| $SSW$ | 0.001 | $-0.7585 \pm 0.0020$ | $0.1620 \pm 0.0615$ | $16.1380 \pm 1.2041$ |
| | 0.01 | $-0.8523 \pm 0.0097$ | $0.1619 \pm 0.0623$ | $14.4800 \pm 1.4353$ |
| | 0.1 | $-0.8555 \pm 0.0119$ | $0.1609 \pm 0.0610$ | $12.4125 \pm 0.0844$ |
| | 1 | $-0.9393 \pm 0.0123$ | $0.1681 \pm 0.0647$ | $12.1478 \pm 1.462$ |
| | 10 | $-0.9168 \pm 0.0099$ | $0.1603 \pm 0.0604$ | $13.3211 \pm 0.2393$ |
| | 100 | $-1.1430 \pm 0.0119$ | $0.1598 \pm 0.0608$ | $12.3703 \pm 0.2108$ |
| | 1000 | $-1.3658 \pm 0.0198$ | $0.1624 \pm 0.0616$ | $12.9221 \pm 0.1441$ |
| $S3W$ | 0.001 | $-1.4883 \pm 0.0298$ | $0.1649 \pm 0.0634$ | $6.0980 \pm 0.1298$ |
| | 0.01 | $-1.6289 \pm 0.0230$ | $0.2013 \pm 0.0706$ | $6.1580 \pm 0.0264$ |
| | 0.1 | $-1.7331 \pm 0.0543$ | $0.2262 \pm 0.0643$ | $6.1620 \pm 0.0117$ |
| | 1 | $-1.7446 \pm 0.0403$ | $0.2188 \pm 0.0669$ | $6.1400 \pm 0.0303$ |
| $RI\text{-}S3W(1)$ | 0.001 | $-1.4061 \pm 0.0098$ | $0.1648 \pm 0.0624$ | $5.8920 \pm 0.0546$ |
| | 0.01 | $-1.6955 \pm 0.0422$ | $0.1655 \pm 0.0422$ | $5.9120 \pm 0.0286$ |
| | 0.1 | $-1.8754 \pm 0.0384$ | $0.2097 \pm 0.0721$ | $5.8780 \pm 0.0293$ |
| | 1 | $-1.8308 \pm 0.0618$ | $0.2315 \pm 0.0699$ | $5.9100 \pm 0.0460$ |
| $RI\text{-}S3W(10)$ | 0.001 | $-1.6410 \pm 0.0205$ | $0.1599 \pm 0.0613$ | $5.7860 \pm 0.4198$ |
| | 0.01 | $-1.6724 \pm 0.0256$ | $0.1642 \pm 0.0644$ | $5.9725 \pm 0.0286$ |
| | 0.1 | $-1.7934 \pm 0.0401$ | $0.1784 \pm 0.0679$ | $5.9900 \pm 0.0374$ |
| | 1 | $-1.7332 \pm 0.0437$ | $0.2178 \pm 0.0718$ | $6.0040 \pm 0.0162$ |
| $ARI\text{-}S3W(10)$ | 0.001 | $-1.5523 \pm 0.0347$ | $0.1649 \pm 0.0638$ | $6.2960 \pm 0.0372$ |
| | 0.01 | $-1.5835 \pm 0.0311$ | $0.1572 \pm 0.0018$ | $6.3120 \pm 0.0271$ |
| | 0.1 | $-1.6448 \pm 0.0381$ | $0.2141 \pm 0.0698$ | $6.3240 \pm 0.0242$ |
| | 1 | $-1.7876 \pm 0.0586$ | $0.2273 \pm 0.0688$ | $6.1980 \pm 0.0970$ |
| $ARI\text{-}S3W(30)$ | 0.001 | $-1.6453 \pm 0.0154$ | $0.1952 \pm 0.0176$ | $5.9540 \pm 0.5100$ |
| | 0.01 | $-1.6290 \pm 0.0126$ | $0.1756 \pm 0.0647$ | $6.1600 \pm 0.0253$ |
| | 0.1 | $-1.7889 \pm 0.0458$ | $0.2305 \pm 0.0631$ | $6.1625 \pm 0.0286$ |
| | 1 | $-1.8890 \pm 0.0506$ | $0.2263 \pm 0.0637$ | $6.1840 \pm 0.0185$ |