VANDERBILT UNIVERSITY

Stereographic Spherical Sliced Wasserstein Distances

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

[1]Department of Computer Science, Vanderbilt University, [2]Department of Mathematics, Vanderbilt University

Background and Introduction

- A **spherical probability distribution** is a probability distribution defined on the d-dimensional hypersphere, denoted $\mathbb{S}^d$.

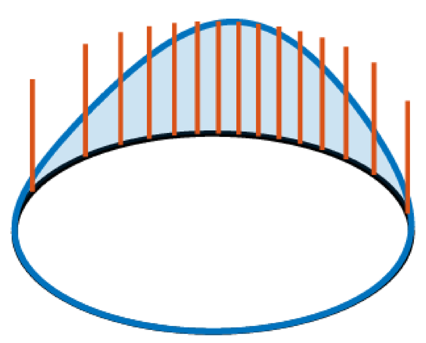


Figure 1. Visualization of a spherical probability distribution on $\mathbb{S}^1$ (the unit circle).

- The field of **optimal transport (OT)** allows us to compare two probability distributions and measure the distance between them. Existing distances that accomplish this task include the **Wasserstein** and **Sinkhorn** distances.
- There is a wide range of applications where we need to compare spherical probability distributions including astronomy, geophysics, meteorology, cosmology, medical imaging, computer vision, and deep learning [1].
- One of the main bottlenecks in OT theory is its high computational cost, with Wasserstein's $\boxed{O(n^3 \log n)}$ runtime and Sinkhorn's $\boxed{O(n^2 \log n)}$ runtime [2]. This high cost renders them impractical for use in large-scale settings.
- This work introduces a numerically efficient distance to compare spherical probability distributions, the **Stereographic Spherical Sliced Wasserstein (S3W) distance**. We demonstrate the superior performance, both in terms of speed and accuracy, of the proposed distance when used across a variety of deep learning problems.

Preliminaries

- The **stereographic projection** $\phi : \mathbb{S}^d \setminus \{s_n\} \to \mathbb{R}^d$ is a bijective, smooth, and conformal transformation from the hypersphere $\mathbb{S}^d$ (excluding the "north pole" $s_n = (0, \dots, 0, 1)$) into a hyperplane $\mathbb{R}^d$.
- The **generalized Radon transform (GRT)** of a probability distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ maps μ to its 1D marginals over hypersurfaces given by the level sets of a defining function $g : \mathbb{R}^d \times (\mathbb{R}^{d'} \setminus \{0\}) \to \mathbb{R}$. Formally, $\mathcal{G}(\mu) = \nu \in \mathcal{P}(\mathbb{R} \times \mathbb{S}^{d'-1})$ s.t.

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

for any test function $\psi \in C_0(\mathbb{R} \times \mathbb{S}^{d'-1})$.
Here, $\mathcal{G}^*$ is the dual operator of $\mathcal{G}$ satisfying $\mathcal{G}(\mu)(\psi) = \mu(\mathcal{G}^*(\psi))$. We denote a specific slice of the resulting measure as $\mathcal{G}(\mu)_\theta = g(\cdot,\theta)_\# \mu$, the pushforward measure of μ w.r.t. $g(\cdot,\theta)$ for a fixed θ. The level sets onto which we project μ can be characterized by $H_{t,\theta} = \{x \in \mathbb{R}^d \,|\, g(x,\theta) = t\}$.
- For probability distributions $\mu, \nu \in \mathcal{P}(M)$, the **p-Wasserstein distance** is

$$W_p^p(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{M\times M} d^p(x,y) d\gamma(x,y). \tag{2}$$

Here, $\gamma \in \mathcal{P}(M \times M)$ is any joint probability distribution with marginals μ and ν. When $\mu, \nu \in \mathcal{P}(\mathbb{R})$, with quantile functions F_μ^{-1} and F_ν^{-1}, Eq. (2) simplifies to

$$W_p^p(\mu,\nu) = \int_0^1 \|F_\mu^{-1}(t) - F_\nu^{-1}(t)\|^p dt. \tag{3}$$

Stereographic Spherical Radon Transform

Definition 1. We introduce the novel **stereographic spherical Radon transform** of a spherical probability distribution $\mu \in \mathcal{P}(\mathbb{S}^d \setminus \{s_n\})$ as

$$\mathcal{S}_\mathcal{G}(\mu) := \mathcal{G}(\phi_\# \mu) \in \mathcal{P}(\mathbb{R} \times \mathbb{S}^{d'-1}), \tag{4}$$

where $\phi_\# \mu$ is the pushforward measure of μ w.r.t. ϕ.

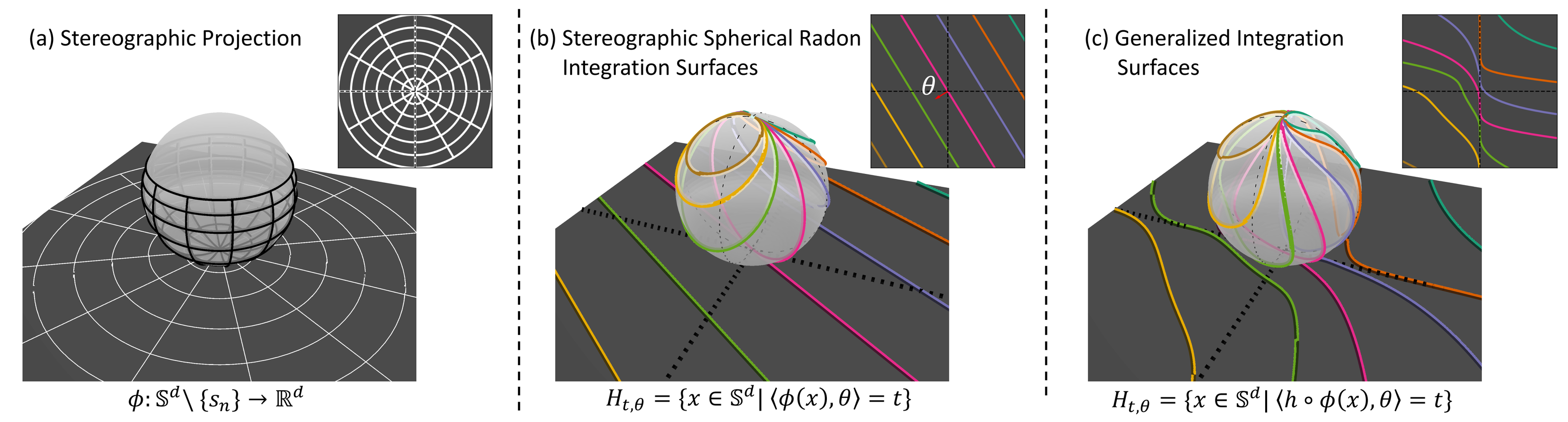


Figure 2. (a) Depiction of stereographic projection from $\mathbb{S}^2$ to $\mathbb{R}^2$. (b) The stereographic Radon transform integration surfaces on $\mathbb{S}^2$, i.e., the level sets of the defining function $g(x,\theta) = \langle \phi(x), \theta \rangle$ for a fixed $\theta \in \mathbb{R}^d$.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

S3W Distances

- For two spherical probability distributions $\mu, \nu \in \mathcal{P}(\mathbb{S}^d \setminus \{s_n\})$, we define their <u>S3W distance</u> as:

$$S3W_{\mathcal{G},p}^p(\mu,\nu) := \int_{\mathbb{S}^{d'-1}} W_p^p(\mathcal{S}_\mathcal{G}(\mu)_\theta, \mathcal{S}_\mathcal{G}(\nu)_\theta)\, d\sigma_{d'}(\theta) \tag{5}$$

where $\sigma_{d'} = \mathrm{Unif}(\mathbb{S}^{d'-1})$. Note that $\mathcal{S}_\mathcal{G}(\mu)_\theta, \mathcal{S}_\mathcal{G}(\nu)_\theta \in \mathcal{P}(\mathbb{R})$, and so the p-Wasserstein distance can be computed efficiently with Eq. (3).
- We introduce a rotationally invariant variation of S3W, the <u>RI-S3W distance</u>, given as:

$$RI\text{-}S3W_{\mathcal{G},p}(\mu,\nu) := \mathbb{E}_{R\sim\omega}[S3W_{\mathcal{G},p}(R_\# \mu, R_\# \nu)] \tag{6}$$

where ω is the Haar measure on the special orthogonal group $\mathrm{SO}(d+1)$ and $R \in \mathrm{SO}(d+1)$ is a rotation matrix.

Theorem 1. *$S3W_{\mathcal{G},p}(\cdot,\cdot)$ and $RI\text{-}S3W_{\mathcal{G},p}(\cdot,\cdot)$ are well-defined and are generally pseudo-metrics on $\mathcal{P}_p(\mathbb{S}^d \setminus \{s_n\})$. When the defining function $g(x,\theta) = \langle h \circ \phi(x), \theta \rangle$ for h injective, $S3W_{\mathcal{G},p}(\cdot,\cdot)$ and $RI\text{-}S3W_{\mathcal{G},p}(\cdot,\cdot)$ define metrics on $\mathcal{P}_p(\mathbb{S}^d \setminus \{s_n\})$.*

- Numerically, we amortize the cost of generating the rotation matrices in Eq. (6) by presampling a rotation pool which we then subsample for every distance calculation. We call this implementation the <u>ARI-S3W distance</u>.

Experiment: Runtime Comparison

- The theoretical runtime of computing S3W is $\boxed{O(LN(d + \log N))}$ and that of RI-S3W is $\boxed{O(N_R(d^3 + Nd^2 + LN(d + \log N))}$, where N is the number of samples, L is the number of level sets considered, d is the dimension, and N_R is the number of rotations used. The $N_R \cdot d^3$ term is avoided by amortizing the generation of rotation matrices as in ARI-S3W.
- We empirically benchmark the runtime of our distances against the **Sliced Wasserstein (SW)*** [3], **Spherical Sliced Wasserstein (SSW)** [1], Wasserstein, and Sinkhorn distances.

*SW is designed for Euclidean distributions, not spherical distributions. We provide it primarily for runtime comparison.

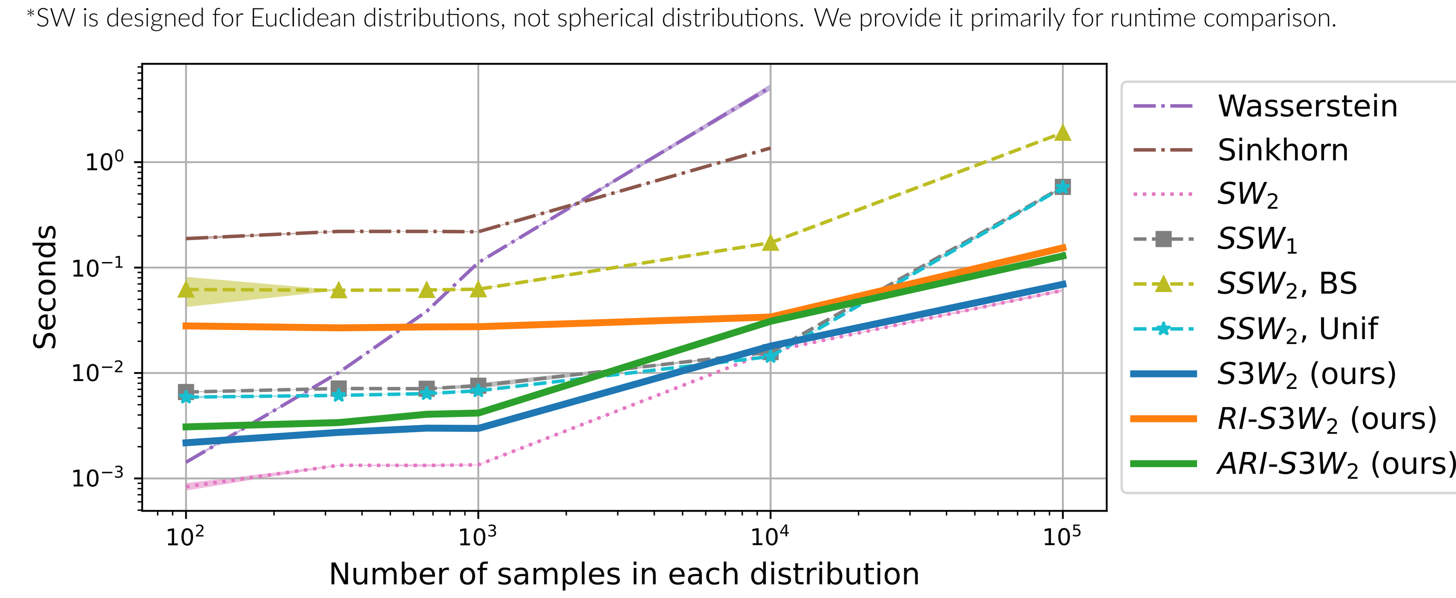


Figure 3. Empirical runtime comparison of $ARI\text{-}S3W$, $RI\text{-}S3W$, $S3W$, SW, SSW_1 (with level median formula), SSW_2 with binary search (BS), SSW_2 with antipodal closed form (only applicable for uniform distribution), Wasserstein, and Sinkhorn. The results demonstrate the improved runtime and scalability of our proposed distances over the benchmark distances.

Experiment: Gradient Flows on Sphere

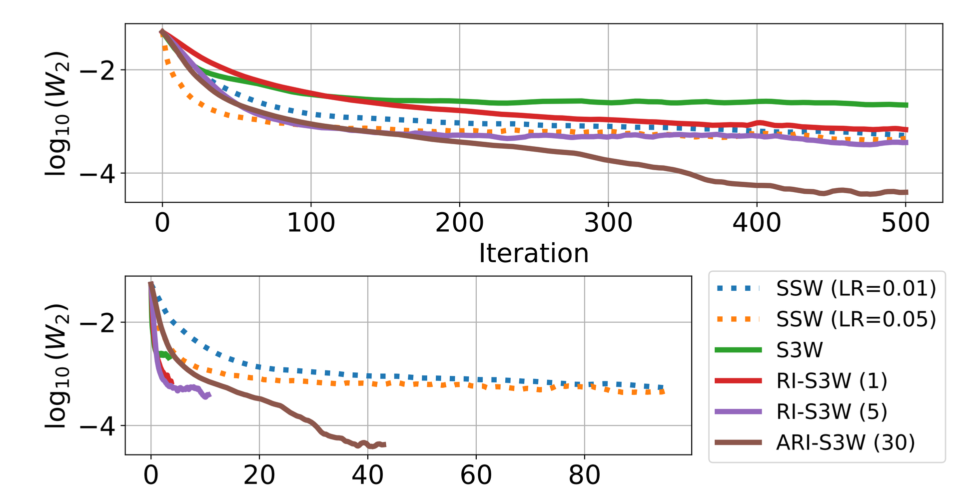


Figure 4. Performance of different distances when used as loss in gradient flow to learn target mixture of 12 von Mises-Fisher distributions. We test SSW with 2 learning rates and $RI\text{-}S3W$ with 1 and 5 rotations. We use 30 rotations subsampled from a pool of size 1000 for $ARI\text{-}S3W$. The top plot demonstrates that ARI-S3W obtains the best performance, and the bottom plot demonstrates that S3W converges the fastest.

Experiment: Self-Supervised Learning (SSL)

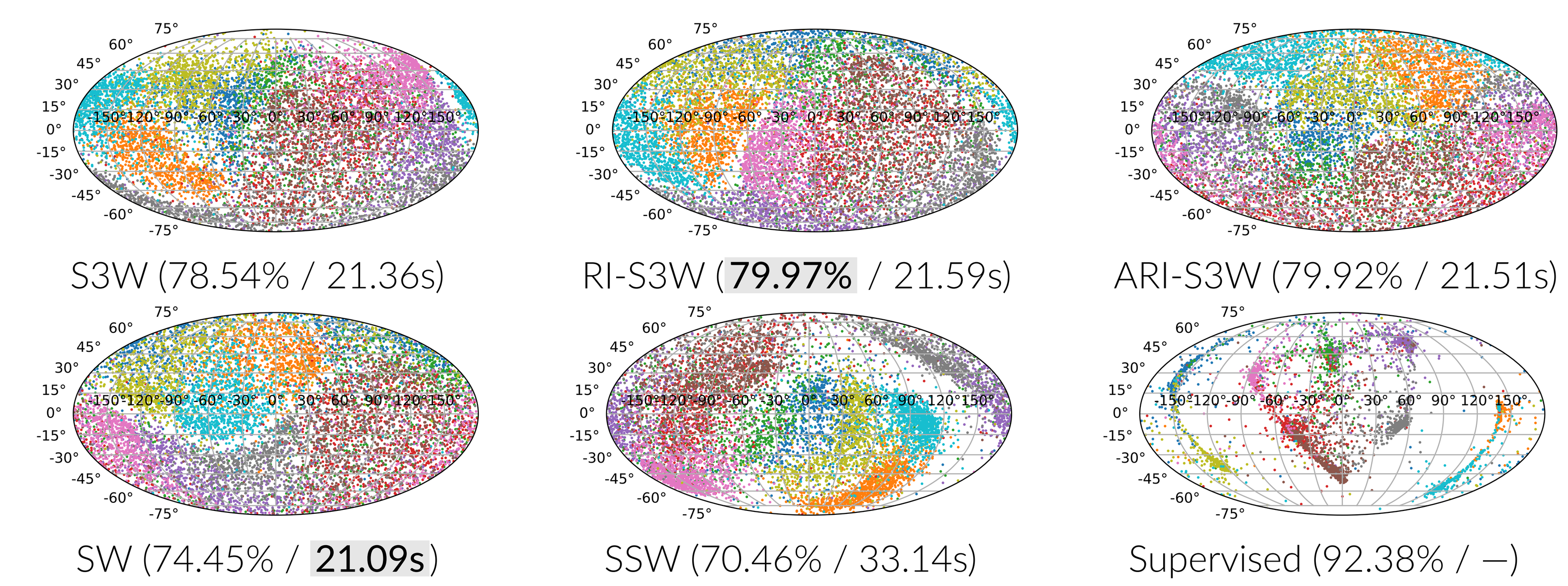
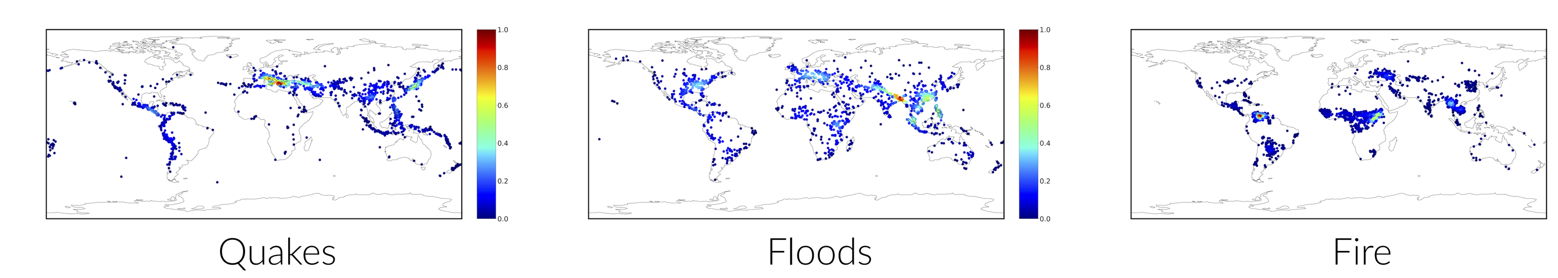


Figure 5. Visualization of the learned latent space embeddings (with $\mathbb{S}^2$ as latent space) of the 10 classes in the CIFAR-10 image dataset when each distance is used to train an image classification model with SSL. The resulting classification accuracy on test data (%) and the time per epoch of SSL pretraining (s) is given in parenthesis, with the highest accuracy and fastest runtime in bold. The result of training a fully supervised model is given as a baseline for comparison. The plots demonstrate the improved latent space utilization and cluster separation when our proposed distances are used, which is reflected in the improved accuracies.

Experiment: Earth Density Estimation



Method	Quake ↓	Flood ↓	Fire ↓
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
$RI\text{-}S3W$	0.79 ± 0.07	1.25 ± 0.02	0.15 ± 0.06
$ARI\text{-}S3W$	**0.78 ± 0.06**	**1.24 ± 0.04**	**0.10 ± 0.04**

Figure 6. Use of distances as loss in normalizing flow model for density estimation of natural disaster events (earthquakes, floods, and fires). The table reports the negative log-likelihood of the learned distribution evaluated on test data and the figures visualize the learned distribution on each dataset when ARI-S3W is used.

Conclusions

- We introduce a new set of distances for spherical probability distributions and prove that the proposed distances indeed comprise metrics on the space of spherical probability distributions.
- We then show that the distances yield superior performance, both in terms of speed and accuracy, over existing alternatives through runtime, gradient flow, SSL, and earth density estimation experiments.

References + Author Contributions

[1] C. Bonet, P. Berg, N. Courty, F. Septier, L. Drumetz, and M.-T. Pham, *Spherical sliced-wasserstein*, 2023.
[2] G. Peyré and M. Cuturi, *Computational optimal transport*, 2020.
[3] J. Rabin, G. Peyré, J. Delon, and M. Bernot, *Wasserstein barycenter and its application to texture mixing*, 2012.

<u>Author Contributions:</u> AK designed poster, performed runtime and SSL experiments. AK, HT implemented methodology/software, performed gradient flow experiment. HT performed earth density estimation experiment. YB, RDM did theoretical derivations. AS, XL aided in visualization. SK provided supervision, funding, and resources.

Contact: abi.kothapalli@vanderbilt.edu