Spherical Sliced-Wasserstein

Clément Bonet\textsuperscript{1}, Paul Berg\textsuperscript{2}, Nicolas Courty\textsuperscript{2}, François Septier\textsuperscript{1}, Lucas Drumetz\textsuperscript{3}, Minh-Tan Pham\textsuperscript{2}

\textsuperscript{1}Université Bretagne Sud, LMBA
\textsuperscript{2}Université Bretagne Sud, IRISA
\textsuperscript{3}IMT Atlantique, Lab-STICC

CAp
03/07/2023
Motivation

Optimal Transport widely use nowadays in Machine Learning

- Domain Adaptation [Courty et al., 2016]
- Generative Models (e.g. WGAN [Arjovsky et al., 2017])
- Document Classification [Kusner et al., 2015]
- ...

Data generally lie on manifolds, e.g. on the sphere $S^{d-1} = \{x \in \mathbb{R}^d, \|x\|_2 = 1\}$:

- Directional data, meteorology, cosmology...
- Also used as embeddings for VAEs, Self-supervised learning...
Wasserstein Distance on the Sphere

- **Sphere:** $S^{d-1} = \{x \in \mathbb{R}^d, \|x\|_2 = 1\}$
- **Geodesic distance:** $\forall x, y \in S^{d-1}, d(x, y) = \arccos(\langle x, y \rangle)$

**Definition (Wasserstein distance)**

Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(S^{d-1})$, then

$$W^p_p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int d(x, y)^p \ d\gamma(x, y),$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(S^{d-1} \times S^{d-1}), \pi_1^\# \gamma = \mu, \pi_2^\# \gamma = \nu\}$ and $\pi^1(x, y) = x$, $\pi^2(x, y) = y$, $\pi_1^\# \gamma = \gamma \circ (\pi^1)^{-1}$. 
Let $\mu, \nu \in \mathcal{P}_p(S^{d-1})$, $x_1, \ldots, x_n \sim \mu$, $y_1, \ldots, y_n \sim \nu$, $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$.

Numerical computation with plug-in estimator: Linear program

$$W_p^p(\hat{\mu}_n, \hat{\nu}_n) = \min_{\gamma \in \Pi(\hat{\mu}_n, \hat{\nu}_n)} \langle C, \gamma \rangle,$$

with $C = (d(x_i, y_j))_{i,j}$.

Complexity: $O(n^3 \log n)$ [Peyré et al., 2019]
Wasserstein Distance on the Sphere

Let \( \mu, \nu \in \mathcal{P}_p(S^{d-1}) \), \( x_1, \ldots, x_n \sim \mu \), \( y_1, \ldots, y_n \sim \nu \), \( \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \) and \( \hat{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i} \).

Numerical computation with plug-in estimator: Linear program

\[
W_p^p(\hat{\mu}_n, \hat{\nu}_n) = \min_{\gamma \in \Pi(\hat{\mu}_n, \hat{\nu}_n)} \langle C, \gamma \rangle,
\]

with \( C = (d(x_i, y_j))_{i,j} \).

Complexity: \( O(n^3 \log n) \) [Peyré et al., 2019]

Proposed Solutions:

- Entropic regularization + Sinkhorn \( O(n^2) \) [Cuturi, 2013]
- Minibatch estimator [Fatras et al., 2020]
- Sliced-Wasserstein [Rabin et al., 2011b, Bonnotte, 2013] but only on Euclidean spaces
Sliced-Wasserstein on $\mathbb{R}^d$

(a) Samples and directions

(b) One dimensional densities

Figure: Illustration of the projection of distributions on different lines.

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \ W^p_\mu(\mu, \nu) = \int_0^1 |F^{-1}_\mu(u) - F^{-1}_\nu(u)|^p \, du$$  (3)
Sliced-Wasserstein on $\mathbb{R}^d$

**Definition (Sliced-Wasserstein [Rabin et al., 2011b])**

Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

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

where $P^\theta(x) = \langle x, \theta \rangle$, $\lambda$ uniform measure on $S^{d-1}$.

**Properties:**

- Distance
- Topologically equivalent to the Wasserstein distance [Nadjahi et al., 2019]
- Monte-Carlo approximation in $O(Ln(\log n + d))$
SW on the Sphere

Goal: defining SW discrepancy on the sphere taking care of geometry of the manifold

<table>
<thead>
<tr>
<th></th>
<th>SW</th>
<th>SSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed-form of $W$</td>
<td>Line</td>
<td>?</td>
</tr>
<tr>
<td>Projection</td>
<td>$P^\theta(x) = \langle x, \theta \rangle$</td>
<td>?</td>
</tr>
<tr>
<td>Integration</td>
<td>$S^{d-1}$</td>
<td>?</td>
</tr>
</tbody>
</table>

Table: SW to SSW
SW on the Sphere

Goal: defining SW discrepancy on the sphere taking care of geometry of the manifold

<table>
<thead>
<tr>
<th></th>
<th>SW</th>
<th>SSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed-form of $W$</td>
<td>Line</td>
<td>?</td>
</tr>
<tr>
<td>Projection</td>
<td>$P^\theta(x) = \langle x, \theta \rangle$</td>
<td>?</td>
</tr>
<tr>
<td>Integration</td>
<td>$S^{d-1}$</td>
<td>?</td>
</tr>
</tbody>
</table>

Table: SW to SSW

- Generalization of straight lines on manifolds: geodesics
- On $S^{d-1}$, geodesics $=$ great circles
Wasserstein on the Circle

Let \( \mu, \nu \in \mathcal{P}(S^1) \) where \( S^1 = \mathbb{R}/\mathbb{Z} \).

- Parametrize \( S^1 \) by \([0, 1]\)
- \( \forall x, y \in [0, 1], \ d_{S^1}(x, y) = \min(|x - y|, 1 - |x - y|) \)
- \( \forall \mu, \nu \in \mathcal{P}(S^1) \), [Rabin et al., 2011a]

\[
W_p^p(\mu, \nu) = \inf_{\alpha \in \mathbb{R}} \int_0^1 |F_{\mu}^{-1}(t) - (F_{\nu} - \alpha)^{-1}(t)|^p \, dt. \tag{5}
\]

- To find \( \alpha \): binary search [Delon et al., 2010]
Particular Cases

- For $p = 1$, [Hundrieser et al., 2021]

$$W_1(\mu, \nu) = \int_0^1 |F_\mu(t) - F_\nu(t) - \text{LevMed}(F_\mu - F_\nu)| \, dt,$$  \hspace{1cm} (6)

where

$$\text{LevMed}(f) = \inf \left\{ t \in \mathbb{R}, \text{Leb}\{x \in [0, 1[, f(x) \leq t\} \geq \frac{1}{2}\right\}.$$  \hspace{1cm} (7)

- For $p = 2$ and $\nu = \text{Unif}(S^1)$,

$$W_2^2(\mu, \nu) = \int_0^1 |F_\mu^{-1}(t) - t - \hat{\alpha}|^2 \, dt \quad \text{with} \quad \hat{\alpha} = \int x \, d\mu(x) - \frac{1}{2}.$$  \hspace{1cm} (8)

In particular, if $x_1 < \cdots < x_n$ and $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$, then

$$W_2^2(\mu_n, \nu) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2 + \frac{1}{n^2} \sum_{i=1}^n (n + 1 - 2i)x_i + \frac{1}{12}.$$  \hspace{1cm} (9)
Great circle: Intersection between 2-plane and \( S^{d-1} \)

Parametrize 2-plane by the Stiefel manifold

\[
\mathcal{V}_{d,2} = \{ U \in \mathbb{R}^{d \times 2}, \, U^T U = I_2 \}
\]

Projection on great circle \( C \): For a.e. \( x \in S^{d-1} \),

\[
P^C(x) = \arg\min_{y \in C} d_{S^{d-1}}(x, y),
\]

where \( d_{S^{d-1}}(x, y) = \arccos(\langle x, y \rangle) \).

For \( U \in \mathcal{V}_{d,2} \), \( C = \text{span}(UU^T) \cap S^{d-1} \),

\[
P^U(x) = U^T \arg\min_{y \in C} d_{S^{d-1}}(x, y)
= \frac{U^T x}{\|U^T x\|_2}.
\]

Figure: Illustration of the geodesic projections on a great circle (in black). In red, random points sampled on the sphere. In green the projections and in blue the trajectories.
Spherical Sliced-Wasserstein

**Definition (Spherical Sliced-Wasserstein)**

Let \( p \geq 1, \, \mu, \nu \in \mathcal{P}_p(S^{d-1}) \) absolutely continuous w.r.t. Lebesgue measure,

\[
SSW^p_\mu(\mu, \nu) = \int_{\mathbb{V}_{d,2}} W^p_{\mu}(P^U \# \mu, P^U \# \nu) \, d\sigma(U),
\]

(10)

with \( \sigma \) the uniform distribution over \( \mathbb{V}_{d,2} \).

<table>
<thead>
<tr>
<th>Table: Comparison SW-SSW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Closed-form of</strong></td>
</tr>
<tr>
<td><strong>Projection</strong></td>
</tr>
<tr>
<td><strong>Integration</strong></td>
</tr>
</tbody>
</table>
Is SSW a Distance?

Question: Is SSW a distance?

Proposition

Let $p \geq 1$, then $SSW_p$ is a pseudo-distance on $\mathcal{P}_{p,ac}(S^{d-1})$.

- Lacking property (for now): indiscernibility property, i.e.
  $$SSW_p(\mu, \nu) = 0 \implies \mu = \nu.$$  
- Need to show that $P_{\#}^U \mu = P_{\#}^U \nu$ for $\sigma$-ae $U \in \mathcal{V}_{d,2}$ implies $\mu = \nu$.
- Idea: relate $P^U$ to a well chosen (injective) Radon transform which integrates along $\{x \in S^{d-1}, P^U(x) = z\}$ for $U \in \mathcal{V}_{d,2}$ and $z \in S^1$. 
Proposition

Let $U \in \mathbb{R}^{d,2}$, $z \in S^1$. The projection set on $z \in S^1$ is

\[
\{ x \in S^{d-1}, \; P^U(x) = z \} = \{ x \in F \cap S^{d-1}, \; \langle x, Uz \rangle > 0 \},
\]

where $F = \text{span}(UU^T)_{\perp} \oplus \text{span}(Uz)$.

**Figure:** The set of projection on the blue point $Uz \in \text{span}(UU^T) \cap S^{d-1}$ is plotted in blue.
Definition (Spherical Radon Transform)

Let \( f \in L^1(S^{d-1}) \), then we define a Spherical Radon transform \( \tilde{R} : L^1(S^{d-1}) \rightarrow L^1(S^1 \times V_{d,2}) \) as

\[
\forall z \in S^1, \, \forall U \in V_{d,2}, \quad \tilde{R}f(z, U) = \int_{S^{d-1}} f(x) \, d\sigma^z_d(x), \quad (12)
\]

with \( \sigma^z_d \) a suitable measure on \( \{x \in S^{d-1}, \, P^U(x) = z\} \).

Results on the injectivity of \( \tilde{R} \) so far:

- In our work: linked it with the Hemispherical Radon transform studied in [Rubin, 1999]
- In [Quellmalz et al., 2023]: showed that it a distance on \( S^2 \)
Goal:

$$\argmin_{\mu} \ SSW^2_2(\mu, \nu),$$

where we have access to $\nu$ through samples, i.e. $\hat{\nu}_m = \frac{1}{m} \sum_{j=1}^{m} \delta_{y_j}$ with $(y_j)_j$ i.i.d samples of $\nu$.

**Figure:** Minimization of SSW with respect to a mixture of vMF.
Wasserstein Autoencoders

Autoencoder with spherical latent space [Davidson et al., 2018, Xu and Durrett, 2018]

SSWAE:

\[
\mathcal{L}(f, g) = \int c(x, g(f(x)))d\mu(x) + \lambda SSW^2_2(f_\#\mu, p_Z),
\]

(a) SWAE

(b) SSWAE

Figure: Latent space of SWAE and SSWAE for a uniform prior on \(S^2\) (on MNIST).

Table: FID on MNIST (Lower is better).

<table>
<thead>
<tr>
<th>Method / Prior</th>
<th>Unif((S^{10}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSWAE</td>
<td>14.91 ± 0.32</td>
</tr>
<tr>
<td>SWAE</td>
<td>15.18 ± 0.32</td>
</tr>
<tr>
<td>WAE-MMD IMQ</td>
<td>18.12 ± 0.62</td>
</tr>
<tr>
<td>WAE-MMD RBF</td>
<td>20.09 ± 1.42</td>
</tr>
<tr>
<td>SAE</td>
<td>19.39 ± 0.56</td>
</tr>
<tr>
<td>Circular GSWAE</td>
<td>15.01 ± 0.26</td>
</tr>
</tbody>
</table>
Density Estimation

Goal: learn a normalizing flow $T$ such that $T \# \mu = p_Z$ with $p_Z = \text{Unif}(S^{d-1})$:

$$\argmin_T SSW^2_2(T \# \mu, p_Z),$$  \hspace{1cm} (14)

where we have access to $\mu$ through samples.

Density:

$$\forall x \in S^{d-1}, \ f_\mu(x) = p_Z(T(x)) | \det J_T(x)|. \hspace{1cm} (15)$$

Table: Negative test log likelihood.

<table>
<thead>
<tr>
<th></th>
<th>Earthquake</th>
<th>Flood</th>
<th>Fire</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSW</td>
<td>$0.84 \pm 0.07$</td>
<td>$1.26 \pm 0.05$</td>
<td>$0.23 \pm 0.18$</td>
</tr>
<tr>
<td>SW</td>
<td>$0.94 \pm 0.02$</td>
<td>$1.36 \pm 0.04$</td>
<td>$0.54 \pm 0.37$</td>
</tr>
<tr>
<td>Stereo</td>
<td>$1.91 \pm 0.1$</td>
<td>$2.00 \pm 0.07$</td>
<td>$1.27 \pm 0.09$</td>
</tr>
</tbody>
</table>

Figure: Density estimation of models trained on earth data. We plot the density on the test data.
Conclusion

- First SW discrepancy on manifolds
- Good performance on ML tasks

Perspectives and follow-up works:
- Study statistical properties
- Try other Spherical Sliced-Wasserstein discrepancies via other Radon transforms
- Study other Riemannian manifolds: Hyperbolic spaces [Bonet et al., 2022], SPDs [Bonet et al., 2023]
- Implemented in POT [Flamary et al., 2021]
Conclusion

First SW discrepancy on manifolds
Good performance on ML tasks

Perspectives and follow-up works:

- Study statistical properties
- Try other Spherical Sliced-Wasserstein discrepancies via other Radon transforms
- Study other Riemannian manifolds: Hyperbolic spaces [Bonet et al., 2022], SPDs [Bonet et al., 2023]
- Implemented in POT [Flamary et al., 2021]

Thank you!


### Runtime Comparisons

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wasserstein + LP</td>
<td>$O(n^3 \log n)$</td>
</tr>
<tr>
<td>Sinkhorn</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>SSW$_2$ + BS</td>
<td>$O(L(n + m)(d + \log(\frac{1}{\epsilon}))) + Ln \log n + Lm \log m)$</td>
</tr>
<tr>
<td>SSW$_1$</td>
<td>$O(L(n + m)(d + \log(n + m)))$</td>
</tr>
<tr>
<td>SSW$_2$ + Unif</td>
<td>$O(Ln(d + \log n))$</td>
</tr>
</tbody>
</table>

**Table: Complexity**

![Runtime Comparisons Graph](image.png)
Wasserstein Autoencoders

SSWAES:

$$\mathcal{L}(f, g) = \int c(x, g(f(x))) d\mu(x) + \lambda SSW^2_{W_2}(f\#\mu, p_Z),$$

(16)

Much interest in using a spherical latent space [Davidson et al., 2018, Xu and Durrett, 2018], e.g. uniform.
Variational Inference

Goal:

$$\arg\min_{\mu} SSW_2^2(\mu, \nu),$$

where we know the density of $\nu$ up to a constant.

**Algorithm SWVI [Yi and Liu, 2021]**

**Input:** $V$ a potential, $K$ the number of iterations of SWVI, $N$ the batch size, $\ell$ the number of MCMC steps

**Initialization:** Choose $q_\theta$ a sampler

**for** $k = 1$ **to** $K$ **do**

Sample $(z_0^i)_{i=1}^N \sim q_\theta$

Run $\ell$ MCMC steps starting from $(z_0^i)_{i=1}^N$ to get $(z_\ell^j)_{j=1}^N$

// Denote $\hat{\mu}_0 = \frac{1}{N} \sum_{j=1}^N \delta_{z_0^j}$ and $\hat{\mu}_\ell = \frac{1}{N} \sum_{j=1}^N \delta_{z_\ell^j}$

Compute $J = SW_2^2(\hat{\mu}_0, \hat{\mu}_\ell)$

Backpropagate through $J$ w.r.t. $\theta$

Perform a gradient step

**end for**
Variational Inference

Goal:

\[ \arg\min_{\mu} SSW_2^2(\mu, \nu), \]

where we know the density of \( \nu \) up to a constant.

- Use SSW instead of SW
- Use Normalizing flows + MCMC on the sphere

Figure: Amortized SSWVI with a normalizing flow \( w.r.t. \) a mixture of vMF.

Figure: Comparison of the ESS between SWVI et SSWVI with the mixture target (mean over 10 runs).