Spherical Sliced-Wasserstein

Clément Bonet$^1$, Paul Berg$^1$, Nicolas Courty$^1$, François Septier$^1$, Lucas Drumetz$^2$, Minh-Tan Pham$^1$

$^1$Université Bretagne Sud
$^2$IMT Atlantique

Journée du laboratoire
07/06/2022
Motivation

Goal of Optimal Transport: Transport mass in an optimal way

\[ W_c(\mu, \nu) = \min_{T \# \mu = \nu} \int c(x, T(x)) \, d\mu(x) \]  

- Coupling
- Value of the min

Widely used nowadays in Machine Learning
- Generative Models (e.g. WGAN [Arjovsky et al., 2017])
- Domain Adaptation [Courty et al., 2016]
- ...

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

Definition (Wasserstein distance)

Let $M$ be a Riemannian manifold endowed with the Riemannian distance $d$, $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(M)$, then

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

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(M \times M), \ \pi^1_\# \gamma = \mu, \pi^2_\# \gamma = \nu\}$ and $\pi^1(x, y) = x$, $\pi^2(x, y) = y$, $\pi^1_\# \gamma = \gamma \circ (\pi^1)^{-1}$.

Numerical approximation: Linear program $O(n^3 \log n)$ [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
\begin{equation}
\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \quad W^p_\mu(\mu, \nu) = \int_0^1 |F^{-1}_\mu(u) - F^{-1}_\nu(u)|^p \, du
\end{equation}
Sliced-Wasserstein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \quad W_p^p(\mu, \nu) = \int_0^1 |F^{-1}_\mu(u) - F^{-1}_\nu(u)|^p \, du \quad (3)$$

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

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

$$SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(P^\theta_\# \mu, P^\theta_\# \nu) \, d\lambda(\theta), \quad (4)$$

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

Wasserstein on \( \mathbb{R} \):

\[
\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \ W_p^p(\mu, \nu) = \int_0^1 |F^{-1}_\mu(u) - F^{-1}_\nu(u)|^p \, du \tag{3}
\]

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

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

\[
SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(P^\theta_\#\mu, P^\theta_\#\nu) \, d\lambda(\theta), \tag{4}
\]

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

**Properties:**

- **Distance**
- Topologically equivalent to the Wasserstein distance
- Monte-Carlo approximation in \( O(Ln \log n) \)
Definition (Radon Transform)

Let \( f \in L^1(\mathbb{R}^d) \), then the Radon transform \( R : L^1(\mathbb{R}^d) \rightarrow L^1(\mathbb{R} \times S^{d-1}) \) is defined as

\[
\forall \theta \in S^{d-1}, \forall t \in \mathbb{R}, \quad Rf(t, \theta) = \int_{\mathbb{R}^d} f(x) \mathbb{1}_{\{\langle x, \theta \rangle = t\}} \, dx. \tag{5}
\]

Definition (Back-projection operator)

The back-projection operator \( R^* : C_0(\mathbb{R} \times S^{d-1}) \rightarrow C_0(\mathbb{R}^d) \) is defined as

\[
\forall g \in C_0(\mathbb{R} \times S^{d-1}), \forall x \in \mathbb{R}^d, \quad R^*g(x) = \int_{S^{d-1}} g(\langle x, \theta \rangle, \theta) \, d\theta. \tag{6}
\]

\[
\forall f, g, \quad \langle Rf, g \rangle_{\mathbb{R} \times S^{d-1}} = \langle f, R^*g \rangle_{\mathbb{R}^d}
\]
Radon Transform of Measures and Link with SW

- Radon transform of a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ defined as the measure $R\mu$ such that $\langle R\mu, g \rangle_{\mathbb{R} \times S^{d-1}} = \langle \mu, R^* g \rangle_{\mathbb{R}^d}$.

- Disintegration of $R\mu \in \mathcal{P}(\mathbb{R} \times S^{d-1})$ w.r.t. $\lambda$: $R\mu = \lambda \otimes K$

- [Bonneel et al., 2015, Proposition 6]: For $\lambda$-ae $\theta \in S^{d-1}$, $K(\theta, \cdot) = P^{\theta} # \mu$

$$\forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d), \quad SW^p_p(\mu, \nu) = \int_{S^{d-1}} W^p_p((R\mu)^{\theta}, (R\mu)^{\theta}) \, d\lambda(\theta). \quad (7)$$

Interest: If $R$ injective, $SW$ is a distance.
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
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
Let \( \mu, \nu \in \mathcal{P}(S^1) \) where \( S^1 = \mathbb{R}/\mathbb{Z} \).

- Parametrize \( S^1 \) by \([0, 1] \)

- \( \forall x, y \in [0, 1[, \; d_{S^1}(x, y) = \min(|x - y|, 1 - |x - y|) \)

- For a cost function \( c(x, y) = h(d_{S^1}(x, y)) \) with \( h : \mathbb{R} \rightarrow \mathbb{R}^+ \) increasing and convex

- \( \forall \mu, \nu \in \mathcal{P}(S^1) \), [Rabin et al., 2011a]

\[
W_c(\mu, \nu) = \inf_{\alpha \in \mathbb{R}} \int_0^1 h(|F_{\mu}^{-1}(t) - (F_{\nu} - \alpha)^{-1}(t)|) \, dt. \tag{8}
\]

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

- For $h = \text{Id}$, [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} (9)

where

$$\text{LevMed}(f) = \inf \left\{ t \in \mathbb{R}, \beta(\{x \in [0, 1[, f(x) \leq t\}) \geq \frac{1}{2} \right\}.$$ \hspace{1cm} (10)

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

$$W_2^2(\mu, \nu) = \int_0^1 |F^{-1}_\mu(t) - t - \hat{\alpha}|^2 \, dt \quad \text{with} \quad \hat{\alpha} = \int x \, d\mu(x) - \frac{1}{2}.$$ \hspace{1cm} (11)

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} (12)
Great circle: Intersection between 2-plane and $S^{d-1}$

Parametrize 2-plane by the Stiefel manifold

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

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

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

$$= \frac{U^T x}{\|U^T x\|_2}. $$

**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.
**Definition (Spherical Sliced-Wasserstein)**

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

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

(13)

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

<table>
<thead>
<tr>
<th>Closed-form of ( W )</th>
<th>SW</th>
<th>SSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projection</td>
<td>( P^\theta(x) = \langle x, \theta \rangle )</td>
<td>( P^U(x) = \frac{U^T x}{|U^T x|_2} )</td>
</tr>
<tr>
<td>Integration</td>
<td>( S^{d-1} )</td>
<td>( \mathbb{V}_{d,2} )</td>
</tr>
</tbody>
</table>

**Table:** Comparison SW-SSW
A New Spherical Radon Transform

Question: Is $SSW$ a distance?

**Definition (Spherical Radon Transform)**

Let $f \in L^1(S^{d-1})$, then we define a Spherical Radon transform
\[ \tilde{R} : L^1(S^{d-1}) \to L^1(S^1 \times \mathbb{V}_{d,2}) \]

as
\[ \forall z \in S^1, \forall U \in \mathbb{V}_{d,2}, \quad \tilde{R} f(z,U) = \int_{S^{d-1}} f(x) \mathbb{1}_{\{z=P_U(x)\}} \, dx. \quad (14) \]

**Definition (Back-projection operator)**

The back-projection operator $R^* : C_0(S^1 \times \mathbb{V}_{d,2}) \to C_b(S^{d-1})$ is defined as for a.e. $x \in S^{d-1},$
\[ \tilde{R}^* g(x) = \int_{\mathbb{V}_{d,2}} g(P_U(x),U) \, d\sigma(U). \quad (15) \]

For all $f \in L^1(S^{d-1}), \ g \in C_0(S^1 \times \mathbb{V}_{d,2}),$
\[ \langle \tilde{R} f, g \rangle_{S^1 \times \mathbb{V}_{d,2}} = \langle f, \tilde{R}^* g \rangle_{S^{d-1}}. \quad (16) \]
Define the Radon transform of $\mu \in \mathcal{M}(S^{d-1})$ as $\tilde{R}\mu$ such that

$$\forall g \in C_0(S^1 \times \mathbb{V}_{d,2}), \int_{S^1 \times \mathbb{V}_{d,2}} g(z, U) \, d(\tilde{R}\mu)(z, U) = \int_{S^{d-1}} \tilde{R}^* g(x) \, d\mu(x). \quad (17)$$

Disintegration: $\tilde{R}\mu = \sigma \otimes (\tilde{R}\mu)^U$

**Proposition**

Let $\mu \in \mathcal{M}(S^{d-1})$, then for $\sigma$-almost every $U \in \mathbb{V}_{d,2}$, $(\tilde{R}\mu)^U = P_{\#}^U \mu$.

$$\forall \mu, \nu \in \mathcal{P}_{p,ac}(S^{d-1}), \ SSW_p^p(\mu, \nu) = \int_{\mathbb{V}_{d,2}} W_p^p((\tilde{R}\mu)^U, (\tilde{R}\nu)^U) \, d\sigma(U). \quad (18)$$
SSW distance?

- \( \forall \mu, \nu \in \mathcal{P}_p(S^{d-1}), SSW_p(\mu, \nu) \geq 0, SSW_p(\mu, \nu) = SSW_p(\nu, \mu), \mu = \nu \implies SSW_p(\mu, \nu) = 0 \)

- Triangular inequality: \( \forall \mu, \nu, \alpha \in \mathcal{P}_p(S^{d-1}), \)

\[
SSW_p(\mu, \nu) = \left( \int_{V_{d,2}} W_p^p(P_{#\mu}, P_{#\nu}) \, d\sigma(U) \right)^{\frac{1}{p}} \\
\leq \left( \int_{V_{d,2}} (W_p(P_{#\mu}, P_{#\alpha}) + W_p(P_{#\alpha}, P_{#\nu}))^p \, d\sigma(U) \right)^{\frac{1}{p}} \\
\leq \left( \int_{V_{d,2}} W_p^p(P_{#\mu}, P_{#\alpha}) \, d\sigma(U) \right)^{\frac{1}{p}} \\
+ \left( \int_{V_{d,2}} W_p^p(P_{#\alpha}, P_{#\nu}) \, d\sigma(U) \right)^{\frac{1}{p}} \\
= SSW_p(\mu, \alpha) + SSW_p(\alpha, \nu). \tag{19}
\]

- Distance if \( SSW_p(\mu, \nu) = 0 \implies \mu = \nu, \)
  i.e. for \( \sigma \)-ae \( U, (\tilde{R}\mu)^U = (\tilde{R}\nu)^U \implies \mu = \nu, \) i.e. \( \tilde{R} \) injective
Properties of the Spherical Radon Transform

- \( \tilde{R} \) related to Hemispherical transform \( \mathcal{H} \) [Rubin, 2003] on \( S^{d-2} \), where for \( f \in L^1(S^{d-2}) \),

\[
\forall x \in S^{d-2}, \quad \mathcal{H} f(x) = \int_{S^{d-2}} f(y) 1_{\{\langle x, y \rangle > 0\}} \, dy.
\] (20)

**Proposition**

\[
\ker(\tilde{R}) = \{ \mu \in \mathcal{M}_{\text{even}}(S^{d-1}), \quad \forall H \in G_{d,d-1}, \quad \mu(H \cap S^{d-1}) = 0 \},
\]

where \( \mu \in \mathcal{M}_{\text{even}} \) if for all \( f \in C(S^{d-1}) \), \( \langle \mu, f \rangle = \langle \mu, f_+ \rangle \) with for all \( x \),
\[
f_+(x) = (f(x) + f(-x))/2.
\]

**Proposition**

Let \( p \geq 1 \), \( SSW_p \) is a pseudo-distance on \( \mathcal{P}_p(S^{d-1}) \).
## 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>$O(L(n + m)(d + \log(\frac{1}{\epsilon})) + L_n \log n + L_m \log m)$</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>$O(L_n(d + \log n))$</td>
</tr>
</tbody>
</table>

**Table: Complexity**

![Runtime Comparisons Graph](image-url)
Goal:

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

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.
Variational Inference

Goal:

$$\arg\min_{\mu} SW_2^2(\mu, \nu),$$

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 $(z_i^0)_{i=1}^N \sim q_\theta$

Run $\ell$ MCMC steps starting from $(z_i^0)_{i=1}^N$ to get $(z_j^\ell)_{j=1}^N$

// Denote $\hat{\mu}_0 = \frac{1}{N} \sum_{j=1}^N \delta_{z_j^0}$ and $\hat{\mu}_\ell = \frac{1}{N} \sum_{j=1}^N \delta_{z_j^\ell}$

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).
Wasserstein Autoencoders

Figure: Autoencoder with spherical latent space.

SSWAE:

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

(21)

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

$$\mathcal{L}(f, g) = \int c(x, g(f(x)))d\mu(x) + \lambda SSW_2^2(f\#\mu, p_Z), \quad (22)$$

Table: FID (Lower is better).

<table>
<thead>
<tr>
<th>Method / Prior</th>
<th>Unif($S^{10}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSWAE</td>
<td><strong>14.91 ± 0.32</strong></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>

Figure: Latent space of SWAE and SSWAE for a uniform prior on $S^2$.  

Wasserstein Autoencoders
Wasserstein Autoencoders

SSWAE:

\[ \mathcal{L}(f, g) = \int c(x, g(f(x)))d\mu(x) + \lambda SSW_{2}^{2}(f\#\mu, pz), \]  

(23)

(a) SSWAE  
(b) SWAE  
(c) SAE

Figure: Samples generated with Sliced-Wasserstein Autoencoders with a uniform prior on \( S^{10} \).
Conclusion

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

Future works
- SW on hyperbolic spaces
- Statistical analysis
Conclusion

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

Future works
- SW on hyperbolic spaces
- Statistical analysis

Thank you!
References I


