Hyperbolic Sliced-Wasserstein via Geodesic and Horospherical Projections

Clément Bonet\textsuperscript{1}, Laetitia Chapel\textsuperscript{1}, Nicolas Courty\textsuperscript{1}, François Septier\textsuperscript{1}, Lucas Drumetz\textsuperscript{2}

\textsuperscript{1}Université Bretagne Sud
\textsuperscript{2}IMT Atlantique

SAV
06/10/2022
Motivation

- Data with hierarchical structure: Hyperbolic spaces [Nickel and Kiela, 2017, 2018]
  - Trees
  - Graphs [Krioukov et al., 2010, Gupte et al., 2011]
  - Words [Tifrea et al., 2018]
  - Images [Khrulkov et al., 2020]
Motivation

- Data with hierarchical structure: Hyperbolic spaces [Nickel and Kiela, 2017, 2018]
  - Trees
  - Graphs [Krioukov et al., 2010, Gupte et al., 2011]
  - Words [Tifrea et al., 2018]
  - Images [Khrulkov et al., 2020]

Goal: develop new tools on hyperbolic spaces

- Distributions [Nagano et al., 2019]
- Neural networks [Ganea et al., 2018]
- Normalizing flows [Bose et al., 2020]
- Optimal transport (OT) [Alvarez-Melis et al., 2020, Hoyos-Idrobo, 2020]

Contribution: new OT discrepancy
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

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

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., 2011, Bonnotte, 2013] but only on Euclidean spaces
Sliced-Wasserstein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \ W^p_\mu(\mu, \nu) = \int_0^1 |F^{-1}_\mu(u) - F^{-1}_\nu(u)|^p \, du$$  \hspace{1cm} (2)
Sliced-Wasserstein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \quad W_p^p(\mu, \nu) = \int_0^1 |F_{\mu}^{-1}(u) - F_{\nu}^{-1}(u)|^p \, du \quad (2)$$

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

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

$$SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(P_{\# \mu}, P_{\# \nu}) \, d\lambda(\theta), \quad (3)$$

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

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}), \ W^p_p(\mu, \nu) = \int_0^1 |F_{\mu}^{-1}(u) - F_{\nu}^{-1}(u)|^p \, du \quad (2)$$

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

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

$$SW^p_p(\mu, \nu) = \int_{S^{d-1}} W^p_p(P^\theta \# \mu, P^\theta \# \nu) \, d\lambda(\theta), \quad (3)$$

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 $O(Ln(\log n + d))$
Hyperbolic space: Riemannian manifold of constant negative curvature

Different models:

Lorentz model

\[ L^d = \{ (x_0, \ldots, x_d) \in \mathbb{R}^{d+1}, \langle x, x \rangle_L = -1, x_0 > 0 \} \]

(4)

where

\[ \forall x, y \in L^d, \langle x, y \rangle_L = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i \]

(5)

Origin:

\[ x_0 = (1, 0, \ldots, 0) \]

Geodesic distance:

\[ d_L(x, y) = \arccosh(-\langle x, y \rangle_L) \]
Hyperbolic space

Hyperbolic space: Riemannian manifold of constant negative curvature
Different models:

- **Lorentz model** \( \mathbb{L}^d \subset \mathbb{R}^{d+1} \)

\[
\mathbb{L}^d = \{(x_0, \ldots, x_d) \in \mathbb{R}^{d+1}, \langle x, x \rangle_\mathbb{L} = -1, x_0 > 0 \}
\]

(4)

where

\[
\forall x, y \in \mathbb{L}^d, \langle x, y \rangle_\mathbb{L} = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i
\]

(5)

- **Origin**: \( x^0 = (1, 0, \ldots, 0) \)
- **Geodesic distance**: \( d_\mathbb{L}(x, y) = \arccosh(-\langle x, y \rangle_\mathbb{L}) \)
Hyperbolic space: Riemannian manifold of constant negative curvature
Different models:

- Lorentz model $\mathbb{L}^d \subset \mathbb{R}^{d+1}$
- Poincaré ball $\mathbb{B}^d = \{ x \in \mathbb{R}^d, \|x\|_2 < 1 \}$
  - Geodesic distance:
    \[
    d_{\mathbb{B}}(x, y) = \text{arccosh} \left( 1 + 2 \frac{\|x - y\|_2^2}{(1 - \|x\|_2^2)(1 - \|y\|_2^2)} \right)
    \]
  - Projection:
    \[
    \forall x \in \mathbb{L}^d, \quad P_{\mathbb{L} \rightarrow \mathbb{B}}(x) = \frac{1}{1 + x_0}(x_1, \ldots, x_d)
    \]
    \[
    \forall x \in \mathbb{B}^d, \quad P_{\mathbb{B} \rightarrow \mathbb{L}}(x) = \frac{1}{1 - \|x\|_2^2}(1 + \|x\|_2^2, 2x_1, \ldots, 2x_d).
    \]
SW on Hyperbolic space

Goal: defining SW discrepancy on Hyperbolic space

|                           | SW          | HSW |
|---------------------------|-------------|-----|
| Closed-form of $W$        | Line        | ?   |
| Projection                | $P^\theta(x) = \langle x, \theta \rangle$ | ?   |
| Integration               | $S^{d-1}$   | ?   |

Table: SW to HSW
Geodesics

- Generalization of straight lines on manifolds: geodesics
- On $\mathbb{L}^d$, geodesics = intersection between 2-plane and $\mathbb{L}^d$
- On $\mathbb{B}^d$, geodesics = circular arcs perpendicular to the boundary $S^{d-1}$

(a) Geodesics on Poincaré ball.  
(b) Geodesics in Lorentz model.
On hyperbolic spaces, geodesic lines, i.e. \( \gamma : \mathbb{R} \rightarrow \mathbb{L}^d \) such that
\[
\forall s, t \in \mathbb{R}, \quad d_{\mathbb{L}}(\gamma(s), \gamma(t)) = |t - s|.
\] (6)

Projection on \( \mathbb{R} \): Let \( v \in T_{x^0} \mathbb{L}^d = \text{span}(x^0)^\perp \),
\[
\forall x \in \gamma(\mathbb{R}) = \mathbb{L}^d \cap \text{span}(v, x^0), \quad t^v_{\mathbb{L}}(x) = \text{sign}(\langle x, v \rangle) d_{\mathbb{L}}(x, x^0)
\] (7)

**Proposition (Wasserstein distance on geodesics.)**

Let \( v \in T_{x^0} \mathbb{L}^d \cap S^d \) and \( \mathcal{G} = \text{span}(x^0, v) \cap \mathbb{L}^d \) a geodesic passing through \( x^0 \). Then, for \( \mu, \nu \) probability measures on \( \mathcal{G} \), we have
\[
\forall p \geq 1, \quad W^p_p(\mu, \nu) = W^p_p(t^v_{\mathbb{L}}\#\mu, t^v_{\mathbb{L}}\#\nu) = \int_0^1 |F^{-1}_{t^v_{\mathbb{L}}\#\mu}(u) - F^{-1}_{t^v_{\mathbb{L}}\#\nu}(u)|^p \, du.
\] (8)
Projection along geodesics

Let \( v \in T_{x^0} \mathbb{L}^d \cap S^d, \mathcal{G} = \text{span}(x^0, v) \cap \mathbb{L}^d \) a geodesic.

Geodesic projection:

\[
\forall x \in \mathbb{L}^d, \quad P^v(x) = \arg\min_{y \in \mathcal{G}} d_{\mathbb{L}}(x, y)
= \frac{1}{\sqrt{\langle x, x^0 \rangle_{\mathbb{L}}^2 - \langle x, v \rangle_{\mathbb{L}}^2}} \left( -\langle x, x^0 \rangle_{\mathbb{L}} x^0 + \langle x, v \rangle_{\mathbb{L}} v \right).
\]  

(c) Along geodesics.

Figure: Projection of (red) points on a geodesic (black line) in the Poincaré ball along geodesics. Projected points on the geodesic are in green.
Definition (Geodesic Hyperbolic Sliced-Wasserstein)

Let $p \geq 1$, $\mu, \nu \in \mathcal{P}_p(\mathbb{L}^d)$, 

$$GHSW^p_p(\mu, \nu) = \int_{T_{x^0} \mathbb{L}^d \cap S^d} W^p_p(t^v \# P^v \mu, t^v \# P^v \nu) \, d\lambda(v).$$  

(10)

| Closed-form of $W$ | SW | HSW |
|--------------------|----|-----|
| Projection         | $P^\theta(x) = \langle x, \theta \rangle$ | $P^v(x)$ |
| Integration        | $S^{d-1}$ | $T_{x^0} \mathbb{L}^d \cap S^d \cong S^{d-1}$ |

Table: Comparison SW-HSW
A second projection

- Geodesic projection:

\[ \langle x, \theta \rangle \theta = \arg\min_{y \in \text{span}(\theta)} \| x - y \|_2 \]  

(11)
A second projection

- Geodesic projection:

\[
\langle x, \theta \rangle \theta = \arg\min_{y \in \text{span}(\theta)} \| x - y \|_2
\]  

(11)

- Coordinate point of view

\[
\langle x, \theta \rangle = \lim_{t \to \infty} \left( t - \| x - t\theta \|_2 \right)
\]  

(12)

- Busemann function:

\[
B_{\gamma}(x) = \lim_{t \to \infty} \left( d(x, \gamma(t)) - t \right).
\]  

(13)

Proposition (Busemann function on hyperbolic space.)

- On \( \mathbb{L}^d \): \( \forall \nu \in T_{x^0} \mathbb{L}^d \cap S^d, \forall x \in \mathbb{L}^d, B_{\nu}(x) = \log(-\langle x, x^0 + \nu \rangle_{\mathbb{L}}) \)

- On \( \mathbb{B}^d \): \( \forall \tilde{\nu} \in S^{d-1}, \forall x \in \mathbb{B}^d, B_{\tilde{\nu}}(x) = \log \left( \frac{\| \tilde{\nu} - x \|^2}{1 - \| x \|^2} \right) \)
Projection along horospheres

- Projection along the level sets of $B_v$
- Level sets = horospheres
- Tend to better preserve the distances [Chami et al., 2021]

Figure: Projection of (red) points on a geodesic (black line) in the Poincaré ball along geodesics or horospheres (in blue). Projected points on the geodesic are in green.
Projection along horospheres

- Projection along the level sets of $B_v$
- Level sets = horospheres
- Tend to better preserve the distances [Chami et al., 2021]

**Proposition (Horospherical projection)**

1. Let $v \in T_{x^0} \mathbb{L}^d \cap S^d$ be a direction and $\mathcal{G} = \text{span}(x^0, v) \cap \mathbb{L}^d$ the corresponding geodesic passing through $x^0$. Then, for any $x \in \mathbb{L}^d$, the projection on $\mathcal{G}$ along the horosphere is given by

\[
\tilde{P}^v(x) = \frac{1 + u^2}{1 - u^2} x^0 + \frac{2u}{1 - u^2} v,
\]

where $u = \frac{1 + \langle x, x^0 + v \rangle_{\mathbb{L}}}{1 - \langle x, x^0 + v \rangle_{\mathbb{L}}}$.  

2. Let $\tilde{v} \in S^{d-1}$ be an ideal point. Then, for all $x \in \mathbb{B}^d$,

\[
\tilde{P}^{\tilde{v}}(x) = \left( \frac{1 - \|x\|_2^2 - \|\tilde{v} - x\|_2^2}{1 - \|x\|_2^2 + \|\tilde{v} - x\|_2^2} \right) \tilde{v}.
\]
Horospherical Hyperbolic Sliced-Wasserstein

**Definition (Horospherical Hyperbolic Sliced-Wasserstein)**

Let \( p \geq 1, \mu, \nu \in \mathcal{P}_p(\mathbb{L}^d), \)

\[
HHSW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(t^v \tilde{P}^v \# \mu, t^v \tilde{P}^v \# \nu) \, d\lambda(v).
\]  
(16)

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

\[
HHSW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(t^{\tilde{v}} \tilde{P}^{\tilde{v}} \# \mu, t^{\tilde{v}} \tilde{P}^{\tilde{v}} \# \nu) \, d\lambda(\tilde{v}).
\]  
(17)

**Proposition**

Let \( \mu, \nu \in \mathcal{P}(\mathbb{B}^d) \) and denote \( \tilde{\mu} = (P_{\mathbb{B} \to \mathbb{L}}) \# \mu, \tilde{\nu} = (P_{\mathbb{B} \to \mathbb{L}}) \# \nu. \) Then,

\[
\forall p \geq 1, \quad HHSW_p^p(\mu, \nu) = HHSW_p^p(\tilde{\mu}, \tilde{\nu}).
\]  
(18)
Summary

|                            | SW          | GHSW        | HHSW        |
|---------------------------|-------------|-------------|-------------|
| Closed-form of $\mathcal{W}$ | Line        | Geodesic    | Geodesic    |
| Projection along Straight line $S^{d-1}$ | Geodesic    | Geodesic Horosphere $S^{d-1}$ |
| Integration               | $S^{d-1}$   |             |             |
| Distance                   | Yes         | Pseudo      | Pseudo      |

Table: Comparison SW-HSW

HHSW/GHSW distances? Rely on related Radon transform injectivity.
Runtime Comparisons

| Method          | Complexity       |
|-----------------|------------------|
| Wasserstein + LP| $O(n^3 \log n)$  |
| Sinkhorn        | $O(n^2)$         |
| GHSW            | $O(Ln(d + \log n))$ |
| HHSW            | $O(Ln(d + \log n))$ |

**Table: Complexity**

![Graph showing runtime comparisons for Wasserstein, Sinkhorn, GHSW, and HHSW methods.](image-url)
Comparisons along Wrapped Normal distributions

Let $\mu = G(x^0, I_d)$, $\nu_t = G(x_t, I_d)$ where $x_t = \cosh(t)x^0 + \sinh(t)v$. 

Figure: Comparison of the Wasserstein distance (with the geodesic distance as cost), GHSW, HHSW and SW between Wrapped Normal distributions.
Gradient Flows

Goal:

$$\arg\min_{\mu} HSW_2^2(\mu, \nu),$$

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: Target distribution and evolution of the log 2-Wasserstein between the target and the gradient flow of GHSW, HHSW and SW. On the left, the target is a WND and on the right, a mixture of 4 WNDs.
Graph Clustering

- Embed a graph as $\nu \in \mathcal{P}(\mathbb{B}^d)$
- Fit a mixture:

$$\argmin_{(\mu_k)_k, (\Sigma_k)_k, (\alpha_k)_k} HSW(\nu, \sum_k \alpha_k \mathcal{G}(\mu_k, \Sigma_k))$$  \hspace{1cm} (19)

**Figure:** Fit of a mixture of WND on a SBM. Cross in black denote the centers learned.
Classification with Prototypes \cite{Ghadimi Atigh et al., 2021}

- \((x_i, y_i)_{i=1}^{n}\) training set, \(y_i \in \{1, \ldots, C\}\), \(\forall c \in \{1, \ldots, C\}\), \(p_c\) prototype.
- \(\forall i, z_i = \exp_0 (f_\theta(x_i))\)
- Loss:

\[
\ell(\theta) = \frac{1}{n} \sum_{i=1}^{n} B_p(z_i) + \lambda_{HSW} \left( \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}, \frac{1}{C} \sum_{c=1}^{C} G(\alpha_c p_c, \beta I_d) \right) \tag{20}
\]

Table: Test accuracy.

|         | CIFAR10  |          |          | CIFAR100 |          |          |          |          |
|---------|----------|----------|----------|----------|----------|----------|----------|----------|
|         | 2        | 3        | 4        | 3        | 5        | 10       | 50       |          |
| Dimensions |         |          |          |          |          |          |          |          |
| Busemann | 91.2     | 92.2     | 92.2     | 49.0     | 54.6     | 59.1     | 65.8     |          |
| GHSW     | **91.61**| **92.48**| **92.29**| **54.78**| **60.94**| **62.72**|          |          |
| HHSW     | 91.32    | 92.34    | 91.92    | 54.29    | 60.67    | 62.14    | 63.17    |          |
Conclusion

- SW discrepancies on hyperbolic spaces
- Application to different ML tasks

Future works
- Statistical analysis
- Distance?
- Applications: persistent diagrams...
Conclusion

- SW discrepancies on hyperbolic spaces
- Application to different ML tasks

Future works

- Statistical analysis
- Distance?
- Applications: persistent diagrams...

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

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

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

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
References II

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/fatras20a.html.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic busemann learning with ideal prototypes. *Advances in Neural Information Processing Systems*, 34:103–115, 2021.

Mangesh Gupte, Pravin Shankar, Jing Li, Shanmugaulayut Muthukrishnan, and Liviu Iftode. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World wide web*, pages 557–566, 2011.

Andrés Hoyos-Idrobo. Aligning hyperbolic representations: an optimal transport-based approach. *arXiv preprint arXiv:2012.01089*, 2020.
Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6418–6428, 2020.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. Physical Review E, 82(3):036106, 2010.

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

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems, 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In International Conference on Machine Learning, pages 3779–3788. PMLR, 2018.
Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.
Gradient Flows

Goal:

$$\text{argmin}_\mu HSW^2_2(\mu, \nu),$$

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 $\nu$.

![Image of particle evolution](image.png)

(a) With geodesic projection.

(b) With horospherical projection.

Figure: Evolution of the particles along the gradient flow of HSW (with geodesic or horospherical projection).
Each document = distribution of words
Embed words in $\mathbb{B}^{100}$
Compute the matrix of distances and use k-NN

Table: Document classification accuracy with $k$-NN ($k = 5$).

|          | W    | $W_e$ | SWp | SWI | GHSW | HHSW |
|----------|------|-------|-----|-----|------|------|
| Movie Reviews | 71.5 | 60.5  | 65  | 65.5| **69.3** | 58.8 |
| Twitter     | 69.7±0.7 | -    | 67.2±0.5 | **67.3±2.3** | 66.6±1.1 | 63.8±1 |
| BBCSport    | 94.7±1.1 | 89.8±0.5 | **89.8±1.4** | 89.8±0.8 | 89.4±1.5 | 75.6±1.7 |