Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds

Clément Bonet[1]

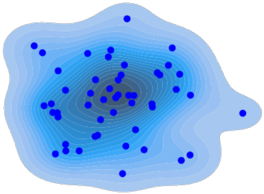
Joint work with Nicolas Courty[2], Lucas Drumetz[3], François Septier[4],
Laetitia Chapel[5], Benoît Malézieux[6], Alain Rakotomamonjy[7], Thomas Moreau[6], Matthieu Kowalski[8]

[1]ENSAE, CREST
[2]Université Bretagne Sud, IRISA
[3]IMT Atlantique, Lab-STICC
[4]Université Bretagne Sud, LMBA
[5]Institut Agro Rennes-Angers, IRISA
[6]Université Paris-Saclay, Inria, CEA
[7]Université de Rouen, Criteo AI Lab
[8]Universite Paris-Saclay, CNRS, LISN

GdT OT
22/01/2024

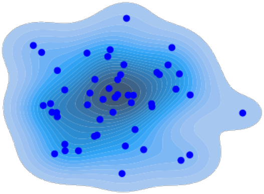
Probability Distributions

- Data: $x_1, \ldots, x_n \in \mathbb{R}^d \longleftrightarrow$ probability distribution $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

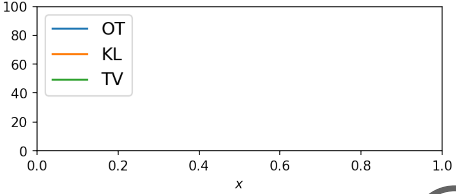


Probability Distributions

- Data: $x_1, \ldots, x_n \in \mathbb{R}^d \longleftrightarrow$ probability distribution $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$



- **Goals**:
  - Compare distributions using some discrepancy D
  - Learn distributions by minimizing D (*e.g.* for generative models)



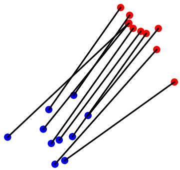
Optimal Transport

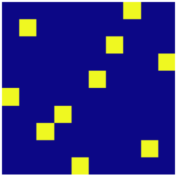
Kantorovich Problem

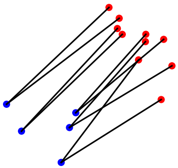
Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

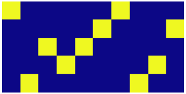
$$\mathrm{OT}_c(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int c(x,y)\ \mathrm{d}\gamma(x,y),$$

$$\Pi(\mu,\nu) = \left\{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d),\ \forall A \in \mathcal{B}(\mathbb{R}^d),\ \gamma(A \times \mathbb{R}^d) = \mu(A),\ \gamma(\mathbb{R}^d \times A) = \nu(A)\right\}$$









Optimal Transport

Wasserstein Distance

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y)$$

Properties:

- W_2 distance
- Metrizes the weak convergence
- Riemannian structure
- Geodesics between μ, ν: $\forall t \in [0,1],\ \mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\# \gamma$ for $\gamma \in \Pi_o(\mu,\nu)$

Condition to have a deterministic coupling, *i.e.* $\gamma = (\mathrm{Id}, T)_\# \mu$ with $T_\# \mu = \nu$ where $\forall A \in \mathcal{B}(\mathbb{R}^d),\ T_\# \mu(A) = \mu\big(T^{-1}(A)\big)$: **Brenier's theorem** (Brenier, 1991)

$\mu \ll \mathrm{Leb} \implies$ Optimal coupling γ^* unique and $\gamma^* = (\mathrm{Id}, \nabla\varphi)_\# \mu$ with φ convex

Solving the OT Problem

Let $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$,

$$W_2^2(\mu, \nu) = \min_{P \in \mathbb{R}_+^{n \times n},\ P\mathbb{1}_n = \alpha,\ P^T\mathbb{1}_n = \beta} \langle C, P \rangle_F \quad \text{with} \quad C = \left(\|x_i - y_j\|_2^2\right)_{i,j}$$

Computational Complexity (Pele and Werman, 2009)

Numerical computation: **Linear program** in $O(n^3 \log n)$

Solving the OT Problem

Let $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$,

$$W_2^2(\mu, \nu) = \min_{P \in \mathbb{R}_+^{n \times n},\ P\mathbb{1}_n = \alpha,\ P^T \mathbb{1}_n = \beta} \langle C, P \rangle_F \quad \text{with} \quad C = \left(\|x_i - y_j\|_2^2\right)_{i,j}$$

Computational Complexity (Pele and Werman, 2009)

Numerical computation: **Linear program** in $O(n^3 \log n)$

Sample Complexity (Boissard and Le Gouic, 2014)

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $x_1, \ldots, x_n \sim \mu$, $y_1, \ldots, y_n \sim \nu$, $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$,

$$\mathbb{E}\Big[\big|W_2(\hat{\mu}_n, \hat{\nu}_n) - W_2(\mu, \nu)\big|\Big] = O(n^{-1/d})$$

Solving the OT Problem

Let $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$,

$$W_2^2(\mu, \nu) = \min_{P \in \mathbb{R}_+^{n\times n},\ P\mathbb{1}_n = \alpha,\ P^T\mathbb{1}_n = \beta} \langle C, P \rangle_F \quad \text{with} \quad C = \left(\|x_i - y_j\|_2^2\right)_{i,j}$$

Computational Complexity (Pele and Werman, 2009)

Numerical computation: **Linear program** in $O(n^3 \log n)$

Sample Complexity (Boissard and Le Gouic, 2014)

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $x_1, \dots, x_n \sim \mu$, $y_1, \dots, y_n \sim \nu$, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$,

$$\mathbb{E}\big[|W_2(\hat{\mu}_n, \hat{\nu}_n) - W_2(\mu, \nu)|\big] = O(n^{-1/d})$$

Proposed solutions:

- Entropic regularization + Sinkhorn (Cuturi, 2013)
- Minibatch estimator (Fatras et al., 2020)
- Sliced-Wasserstein (Rabin et al., 2011; Bonnotte, 2013)

1D OT Problem

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$,

- Cumulative distribution function:

$$\forall t \in \mathbb{R},\ F_\mu(t) = \mu\big(]-\infty, t]\big) = \int \mathbb{1}_{]-\infty,t]}(x)\ \mathrm{d}\mu(x)$$

- Quantile function:

$$\forall u \in [0,1],\ F_\mu^{-1}(u) = \inf\left\{x \in \mathbb{R},\ F_\mu(x) \geq u\right\}$$

1D Wasserstein Distance

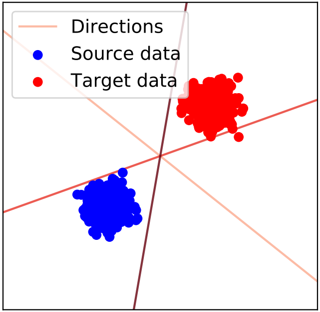
$$W_2^2(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^2\ \mathrm{d}u = \left\|F_\mu^{-1} - F_\nu^{-1}\right\|^2_{L^2([0,1])}$$

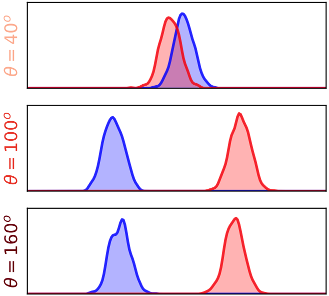
Let $x_1 < \cdots < x_n,\ y_1 < \cdots < y_n,\ \mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i},\ \nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i},$

$$W_2^2(\mu,\nu) = \frac{1}{n}\sum_{i=1}^n (x_i - y_i)^2$$

$\rightarrow O(n \log n)$

Sliced-Wasserstein Distance





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

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{SW}_2^2(\mu, \nu) = \int_{S^{d-1}} W_2^2(P^\theta_\# \mu, P^\theta_\# \nu)\ \mathrm{d}\lambda(\theta),$$

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

Properties of the Sliced-Wasserstein Distance

Let $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^d$, $\alpha, \beta \in \Sigma_n$, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \beta_i \delta_{y_i}$.

Approximation via Monte-Carlo:

$$\widehat{\mathrm{SW}}_{2,L}^2(\mu, \nu) = \frac{1}{L} \sum_{\ell=1}^L W_2^2(P_\#^{\theta_\ell}\mu, P_\#^{\theta_\ell}\nu),$$

$\theta_1, \dots, \theta_L \sim \lambda$.

Properties:

- Computational complexity: $O(Ln \log n + Lnd)$
- Sample complexity: independent of the dimension (Nadjahi et al., 2020)
- SW_2 distance (Bonnotte, 2013)
- Topologically equivalent to the Wasserstein distance (Nadjahi et al., 2019), *i.e.* $\lim_{n\to\infty} \mathrm{SW}_2^2(\mu_n, \mu) = 0 \iff \lim_{n\to\infty} W_2^2(\mu_n, \mu) = 0$.
- Differentiable, Hilbertian

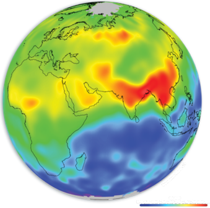
Table of Contents

Riemannian Manifolds in Machine Learning

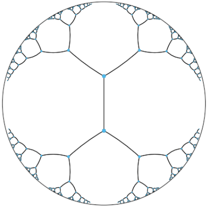
Data often lie on manifolds or have an underlying structure which can be captured on manifolds.

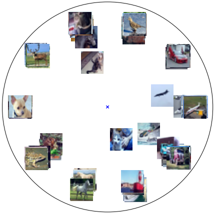
Example

- Directional data, Earth data, cyclic data on the sphere S^{d-1}
- Hierarchical data (trees, graphs, words, images) on Hyperbolic spaces
- M/EEG data on the space of Symmetric Positive Definite Matrices (SPDs)



Source: ESA





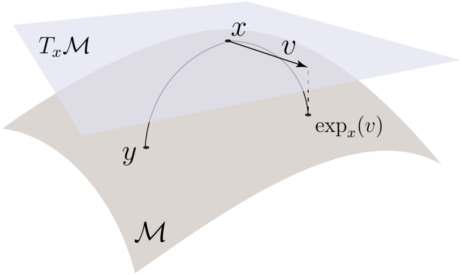
Riemannian Manifolds

Definition

A Riemannian manifold $(\mathcal{M}, g)$ of dimension d is a space that behaves locally as a linear space diffeomorphic to $\mathbb{R}^d$.

Properties:

- To any $x \in \mathcal{M}$, associate a tangent space $T_x\mathcal{M}$ with a smooth inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$.
- Geodesic between x and y: shortest path minimizing the length $\mathcal{L}$
- Geodesic distance: $d(x, y) = \inf_{\gamma} \mathcal{L}(\gamma)$
- Exponential map: $\forall x \in \mathcal{M}, \; \exp_x : T_x\mathcal{M} \to \mathcal{M}$



Cartan-Hadamard Manifolds

Particular case of Riemannian manifold: **Cartan-Hadamard** manifolds $(\mathcal{M}, g)$

Definition: Non-positive curvature, complete and connected

Properties:

- Geodesically complete: Any geodesic $\gamma : [0,1] \to \mathcal{M}$ between $x \in \mathcal{M}$ and $y \in \mathcal{M}$ can be extended to $\mathbb{R}$
- For any $x \in \mathcal{M}$, $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ diffeomorphism

Example

- Euclidean spaces
- Hyperbolic spaces (Nickel and Kiela, 2017, 2018; Khrulkov et al., 2020)
- SPDs endowed with specific metrics (Sabbagh et al., 2019, 2020; Pennec, 2020)
- Product of Cartan-Hadamard manifolds (Gu et al., 2019; Skopek et al., 2019)

Hyperbolic Space

Hyperbolic space: Riemannian manifold of constant negative curvature

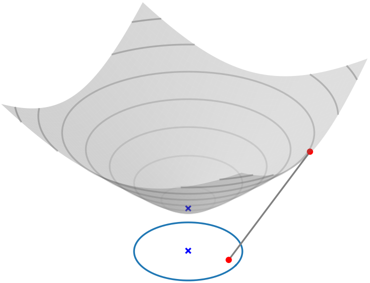
Different isometric models:

- **Lorentz model** $\mathbb{L}^d = \big\{(x_0, \dots, x_d) \in \mathbb{R}^{d+1},\ \langle x, x\rangle_{\mathbb{L}} = -1,\ x_0 > 0\big\}$,

$$d_{\mathbb{L}}(x,y) = \operatorname{arccosh}\big(-\langle x, y\rangle_{\mathbb{L}}\big), \quad \langle x, y\rangle_{\mathbb{L}} = -x_0 y_0 + \sum_{i=1}^d x_i y_i$$

- **Poincaré ball** $\mathbb{B}^d = \{x \in \mathbb{R}^d,\ \|x\|_2 < 1\}$,

$$d_{\mathbb{B}}(x,y) = \operatorname{arccosh}\left(1 + 2\frac{\|x-y\|_2^2}{(1-\|x\|_2^2)(1-\|y\|_2^2)}\right)$$



Optimal Transport on Riemannian Manifolds

Let $(\mathcal{M}, g)$ be a Riemannian manifold, d its geodesic distance.

Definition (Wasserstein distance)

Let $\mu, \nu \in \mathcal{P}_2(\mathcal{M})$, then

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int d(x,y)^2 \; \mathrm{d}\gamma(x,y)$$

In practice: same drawbacks of the Euclidean case.

SW on Cartan-Hadamard Manifolds

Goal: defining SW discrepancy on Cartan-Hadamard manifolds taking care of geometry of the manifold

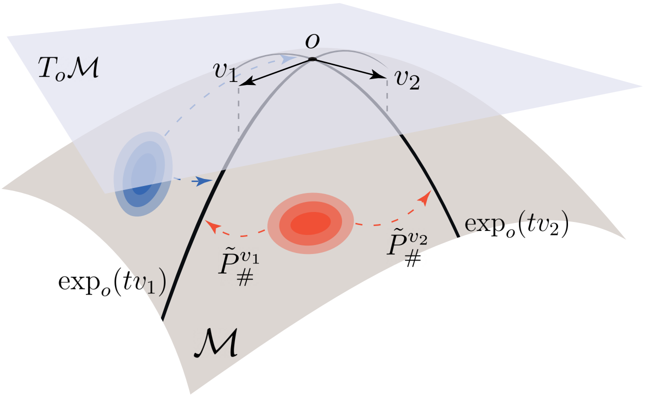
	SW	CHSW
Closed-form of W	Line	?
Projection	$P^\theta(x) = \langle x, \theta \rangle$	?
Integration	S^{d-1}	?

Projecting on Geodesics

- Generalization of straight lines on manifolds: **geodesics**

$$\forall v \in T_o\mathcal{M},\ \mathcal{G}_v = \{\exp_o(tv),\ t \in \mathbb{R}\}$$

- Geodesics isometric to $\mathbb{R}$
- Integrate along all possible directions on $S_o = \{v \in T_o\mathcal{M},\ \|v\|_o = 1\}$



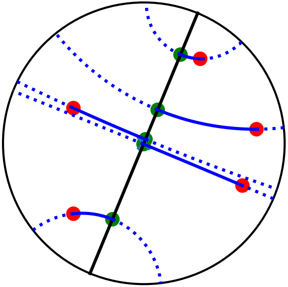
Projections

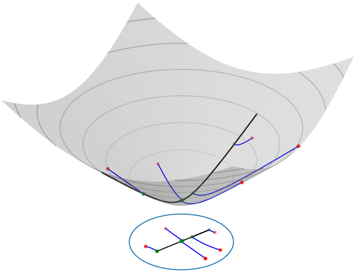
1. **Geodesic projections**:
   - On Euclidean space: For $\theta \in S^{d-1}$, $\mathcal{G}_\theta = \{t\theta,\ t \in \mathbb{R}\}$,

$$\forall x \in \mathbb{R}^d,\ P^\theta(x) = \langle x, \theta \rangle = \underset{t \in \mathbb{R}}{\mathrm{argmin}}\ \|x - t\theta\|_2$$

   - On Cartan-Hadamard manifold: For $v \in T_o\mathcal{M}$, $\mathcal{G}_v = \{\exp_o(tv),\ t \in \mathbb{R}\}$,

$$\forall x \in \mathcal{M},\ P^v(x) = \underset{t \in \mathbb{R}}{\mathrm{argmin}}\ d\big(x, \exp_o(tv)\big)$$





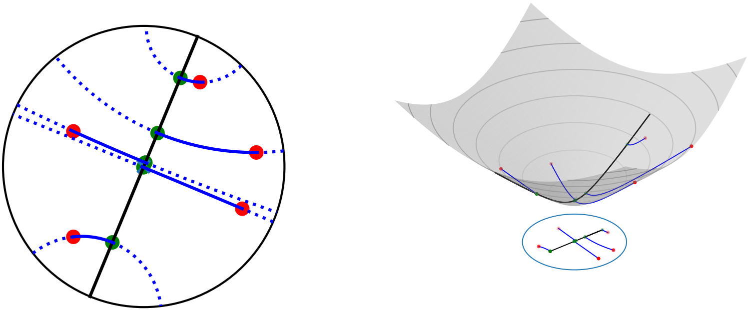
Projections

1. **Geodesic projections**:
   - On Euclidean space: For $\theta \in S^{d-1}$, $\mathcal{G}_\theta = \{t\theta,\ t \in \mathbb{R}\}$, $\exp_0(t\theta) = 0 + t\theta = t\theta$,

$$\forall x \in \mathbb{R}^d,\ P^\theta(x) = \langle x, \theta \rangle = \operatorname*{argmin}_{t\in\mathbb{R}} \ \|x - t\theta\|_2 = \operatorname*{argmin}_{t\in\mathbb{R}} \ d\big(x, \exp_0(t\theta)\big)$$

   - On Cartan-Hadamard manifold: For $v \in T_o\mathcal{M}$, $\mathcal{G}_v = \{\exp_o(tv),\ t \in \mathbb{R}\}$,

$$\forall x \in \mathcal{M},\ P^v(x) = \operatorname*{argmin}_{t\in\mathbb{R}} \ d\big(x, \exp_o(tv)\big)$$

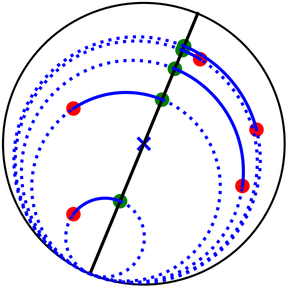


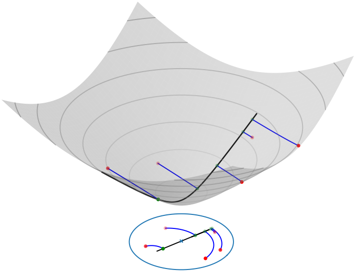
Projections

1. **Geodesic projections**: $\forall x \in \mathcal{M},\ P^v(x) = \underset{t\in\mathbb{R}}{\mathrm{argmin}}\ d\big(x, \exp_o(tv)\big)$
2. **Horospherical projections**: following level sets of the Busemann function

$$B^\gamma(x) = \lim_{t\to\infty}\ d\big(x,\gamma(t)\big) - t$$

- On Euclidean space: $B^\theta(x) = -\langle x, \theta\rangle$
- On Cartan-Hadamard manifold: $B^v(x) = \lim_{t\to\infty}\ d\big(x, \exp_o(tv)\big) - t$





Cartan-Hadamard Sliced-Wassertein

Let $(\mathcal{M}, g)$ a Hadamard manifold with o its origin. Denote λ the uniform distribution on $S_o = \{v \in T_o\mathcal{M},\ \|v\|_o = 1\}$.

Geodesic-Cartan Hadamard Sliced-Wasserstein

$$\forall \mu, \nu \in \mathcal{P}_2(\mathcal{M}),\ \mathrm{GCHSW}_2^2(\mu, \nu) = \int_{S_o} W_2^2(P^v_\# \mu, P^v_\# \nu)\ \mathrm{d}\lambda(v)$$

Horospherical-Cartan Hadamard Sliced-Wasserstein

$$\forall \mu, \nu \in \mathcal{P}_2(\mathcal{M}),\ \mathrm{HCHSW}_2^2(\mu, \nu) = \int_{S_o} W_2^2(B^v_\# \mu, B^v_\# \nu)\ \mathrm{d}\lambda(v)$$

$\mathrm{CHSW} = \mathrm{GCHSW}$ or HCHSW

General Properties

Some properties:

- Pseudo distance on $\mathcal{P}_2(\mathcal{M})$ $\rightarrow$ open question: distance?
- $\forall \mu, \nu \in \mathcal{P}_2(\mathcal{M}),\ \mathrm{CHSW}_2^2(\mu,\nu) \leq W_2^2(\mu,\nu)$
- Sample complexity independent of the dimension
- Computational complexity: $L \cdot O\big(\mathrm{sort}(n)\big) + Ln \cdot O\big(\mathrm{projection}(d)\big)$
- CHSW_2 is Hilbertian

Proposition

Define $K : \mathcal{P}_2(\mathcal{M}) \times \mathcal{P}_2(\mathcal{M}) \rightarrow \mathbb{R}$ as $K(\mu,\nu) = \exp\big(-\gamma \mathrm{CHSW}_2^2(\mu,\nu)\big)$ for $\gamma > 0$. Then K is a positive definite kernel.

Proposition

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{B}^d)$ and denote $\tilde{\mu} = (P_{\mathbb{B}\to\mathbb{L}})_\#\mu$, $\tilde{\nu} = (P_{\mathbb{B}\to\mathbb{L}})_\#\nu$. Then,

$$\mathrm{HHSW}_2^2(\mu,\nu) = \mathrm{HHSW}_2^2(\tilde{\mu},\tilde{\nu}),$$
$$\mathrm{GHSW}_2^2(\mu,\nu) = \mathrm{GHSW}_2^2(\tilde{\mu},\tilde{\nu}).$$

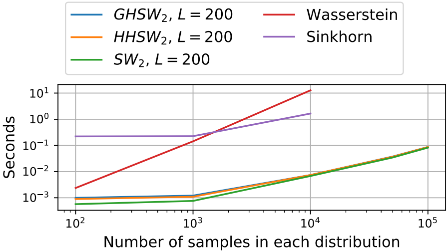
Runtime and Complexity (Bonet et al., 2023c)

Closed-forms for P^v and B^v on $\mathbb{B}^d$ and $\mathbb{L}^d$:

$$\forall v \in T_{x^0}\mathbb{L}^d \cap S^d,\ x \in \mathbb{L}^d,$$
$$P^v(x) = \operatorname{arctanh}\left(-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right)$$
$$B^v(x) = \log\big(-\langle x, x^0 + v\rangle_{\mathbb{L}}\big)$$

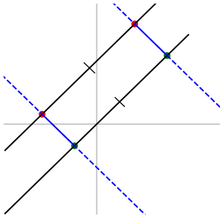
$$\forall \tilde{v} \in S^{d-1},\ y \in \mathbb{B}^d,$$
$$P^{\tilde{v}}(y) = 2\operatorname{arctanh}\big(s(y)\big)$$
$$B^{\tilde{v}}(y) = \log\left(\frac{\|\tilde{v} - y\|_2^2}{1 - \|y\|_2^2}\right)$$

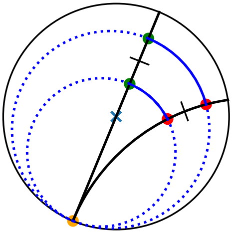
Method	Complexity
Wasserstein + LP	$O(n^3 \log n + n^2 d)$
Sinkhorn	$O(n^2 d)$
SW	$O\big(Ln(d + \log n)\big)$
GHSW	$O\big(Ln(d + \log n)\big)$
HHSW	$O\big(Ln(d + \log n)\big)$



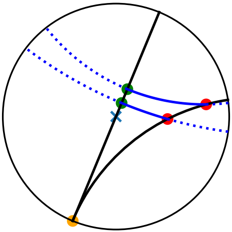
Comparison of the Projections

- Property of the Horospherical projection: conserves the distance between points on a parallel geodesic (Chami et al., 2021)





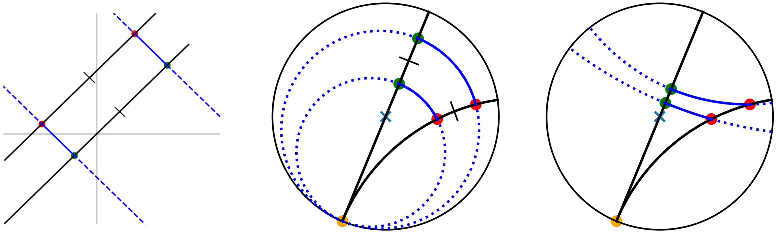
Horospherical projection



Geodesic projection

Comparison of the Projections

- Property of the Horospherical projection: conserves the distance between points on a parallel geodesic (Chami et al., 2021)



Horospherical projection

Geodesic projection

- Let $\mu = \text{WND}(0, I_d)$, $\nu_t = \text{WND}(x_t, I_d)$,

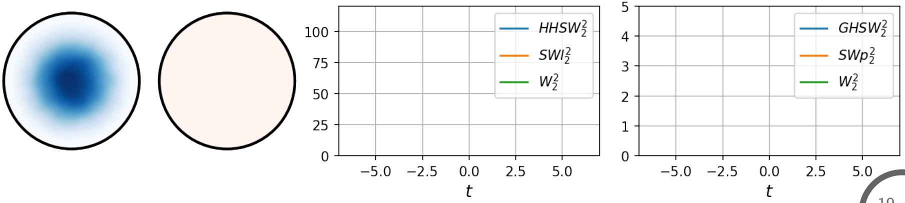


Table of Contents

Pullback Euclidean Manifold

Let $(\mathcal{N}, \langle \cdot, \cdot \rangle)$ an Euclidean space, $\phi : \mathcal{M} \to \mathcal{N}$ a diffeomorphism.

- $(\mathcal{M}, g^\phi)$ Riemannian manifold with $g_x^\phi(u,v) = \langle \phi_{*,x}(u), \phi_{*,x}(v) \rangle$ for $x \in \mathcal{M}$, $u, v \in T_x\mathcal{M}$
- Geodesic distance: $d_{\mathcal{M}}(x,y) = \|\phi(x) - \phi(y)\|$
- Geodesic through $o \in \mathcal{M}$ with direction $v \in T_o\mathcal{M}$:

$$\forall t \in \mathbb{R},\ \gamma_v(t) = \phi^{-1}\big(\phi(o) + t\phi_{*,o}(v)\big)$$

Proposition

Let $v \in S_o = \{v \in T_o\mathcal{M},\ \|v\|_o = \|\phi_{*,o}(v)\| = 1\}$, *then the projection coordinate on* $\mathcal{G}_v = \{\gamma_v(t),\ t \in \mathbb{R}\}$ *is*

$$\forall x \in \mathcal{M},\ P^v(x) = -B^v(x) = \langle \phi(x) - \phi(o), \phi_{*,o}(v) \rangle.$$

Pullback SW

Let $(\mathcal{M}, g^\phi)$ a Pullback Euclidean Manifold.

Proposition

Let $\mu, \nu \in \mathcal{P}_2(\mathcal{M})$. *Then,*

$$\begin{aligned} \mathrm{CHSW}_2^2(\mu,\nu) &= \int_{S_{\phi(o)}} W_2^2(Q^v_\# \phi_\# \mu, Q^v_\# \phi_\# \nu)\ \mathrm{d}\big((\phi_{*,o})_\# \lambda\big)(v) \\ &= \mathrm{SW}_2^2\big(\phi_\#\mu, \phi_\#\nu\big) \end{aligned}$$

with $Q^v(x) = \langle x, v\rangle$ *and* SW_2 *the Euclidean Sliced-Wasserstein distance.*

Additional Properties:

- CHSW_2 is a finite distance on $\mathcal{P}_2(\mathcal{M})$
- CHSW_2 metrizes the weak convergence
- For $\mu, \nu \in \mathcal{P}\big(B(o,r)\big)$,

$$\mathrm{CHSW}_2^2(\mu,\nu) \le W_2^2(\mu,\nu) \le C_{d,r}\mathrm{CHSW}_2(\mu,\nu)^{\frac{1}{d+1}}$$

Examples

Example

- Mahalanobis distance: $\langle u, v\rangle_x = u^T A v$ for $A \in S_d^{++}(\mathbb{R})$
- Squared geodesic distance where $\langle u, v\rangle_x = u^T A(x) v$ for $A(x) \in S_d^{++}(\mathbb{R})$
- SPD with ($O(n)$-Invariant) Log-Euclidean metric, Log-Cholesky metric

Mahalanobis distance: Let $A \in S_d^{++}(\mathbb{R})$,

$$\forall x, y \in \mathbb{R}^d,\ d(x,y)^2 = (x-y)^T A (x-y) = \|A^{\frac{1}{2}} x - A^{\frac{1}{2}} y\|_2^2$$

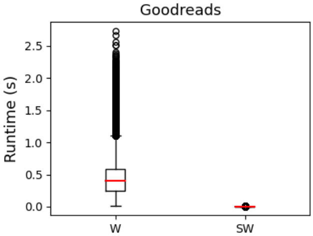
- $\phi(x) = A^{\frac{1}{2}} x$, $\phi_{*,0}(v) = A^{\frac{1}{2}} v$
- For $v \in S_0 = \{v \in \mathbb{R}^d,\ \|v\|_0^2 = v^T A v = 1\},\ P^v(x) = \langle A^{\frac{1}{2}} x, A^{\frac{1}{2}} v\rangle = x^T A v$

$$\mathrm{SW}_{2,A}^2(\mu,\nu) = \int_{S_0} W_2^2(P^v_\# \mu, P^v_\# \nu)\ \mathrm{d}\lambda(v)$$

Document Classification (Kusner et al., 2015)

Goal: Classify documents

- Words $x_1, \dots, x_n \in \mathbb{R}^d$
- Document $D_k = \sum_{i=1}^n w_i^k \delta_{x_i}$ with $\sum_{i=1}^n w_i^k = 1$
- Learn A (Huang et al., 2016)
- Compute $\big(d_A(D_k, D_\ell)\big)_{k,\ell}$
- Use a k-nearest neighbor classifier



Accuracy

	BBCSport	Movies	Goodreads genre	Goodreads like
W_2	94.55	74.44	56.18	71.00
W_A	98.36	76.04	56.81	68.37
SW_2	$89.42_{\pm 0.89}$	$67.27_{\pm 0.69}$	$50.01_{\pm 1.21}$	$65.90_{\pm 0.17}$
$\mathrm{SW}_{2,A}$	$97.58_{\pm 0.04}$	$76.55_{\pm 0.11}$	$57.03_{\pm 0.68}$	$67.54_{\pm 0.14}$

Manifold of SPD Matrices with Affine-Invariant Metric

Symmetric Positive Definite (SPD) Matrices:

$$S_d^{++}(\mathbb{R}) = \left\{ M \in S_d(\mathbb{R}),\ \forall x \in \mathbb{R}^d \setminus \{0\},\ x^T M x > 0 \right\}$$

- Affine-Invariant distance: $\forall X, Y \in S_d^{++}(\mathbb{R})$, $d_{AI}(X, Y) = \sqrt{\mathrm{Tr}\big(\log(X^{-1}Y)^2\big)}$
- Tangent space: $T_{I_d} S_d^{++}(\mathbb{R}) \cong S_d(\mathbb{R})$
- Geodesics through I_d: $\mathcal{G}_A = \{\exp(tA),\ t \in \mathbb{R}\}$ for $A \in S_d(\mathbb{R})$

Manifold of SPD Matrices with Affine-Invariant Metric

Symmetric Positive Definite (SPD) Matrices:

$$S_d^{++}(\mathbb{R}) = \big\{M \in S_d(\mathbb{R}),\ \forall x \in \mathbb{R}^d \setminus \{0\},\ x^T M x > 0\big\}$$

- Affine-Invariant distance: $\forall X, Y \in S_d^{++}(\mathbb{R})$, $d_{AI}(X,Y) = \sqrt{\mathrm{Tr}\big(\log(X^{-1}Y)^2\big)}$
- Tangent space: $T_{I_d} S_d^{++}(\mathbb{R}) \cong S_d(\mathbb{R})$
- Geodesics through I_d: $\mathcal{G}_A = \{\exp(tA),\ t \in \mathbb{R}\}$ for $A \in S_d(\mathbb{R})$

Projections:

- Closed-form for the geodesic projection?
- Busemann function:

$$\forall M \in S_d^{++}(\mathbb{R}),\ B^A(M) = -\langle A, \log\big(\pi_A(M)\big)\rangle_F,$$

with π_A projection on the space of matrices commuting with A.

$\rightarrow$ Very costly in practice

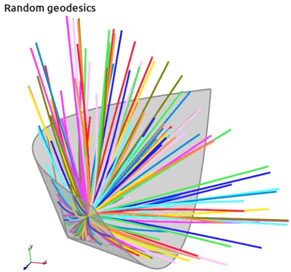
Manifold of SPD Matrices with Log-Euclidean Metric

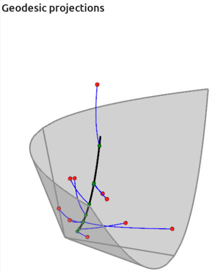
Symmetric Positive Definite (SPD) Matrices:

$$S_d^{++}(\mathbb{R}) = \left\{M \in S_d(\mathbb{R}),\ \forall x \in \mathbb{R}^d \setminus \{0\},\ x^T M x > 0\right\}$$

- Log-Euclidean distance: $\forall X, Y \in S_d^{++}(\mathbb{R}),\ d_{LE}(X,Y) = \| \log X - \log Y \|_F$
- Tangent space: $T_{I_d} S_d^{++}(\mathbb{R}) \cong S_d(\mathbb{R})$
- Projection on geodesics $\mathcal{G}_A = \{\exp(tA),\ t \in \mathbb{R}\}$ for $A \in S_{I_d}$:

$$\forall M \in S_d^{++}(\mathbb{R}),\ P^A(M) = -B^A(M) = \langle A, \log M \rangle_F$$



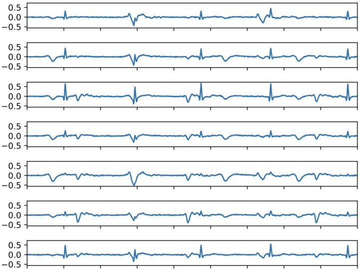


M/EEG data

M/EEG data:

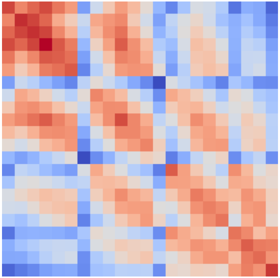
- Recorded from the brain
- Multivariate time series $X \in \mathbb{R}^{N \times T}$
- Transform X into SPDs

$\rightarrow$ Brain-Age prediction



Data X with T time samples

$\longrightarrow$



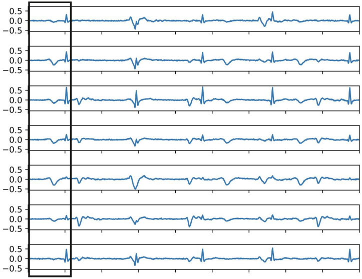
SPD matrix

M/EEG data

M/EEG data:

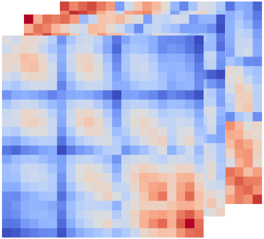
- Recorded from the brain
- Multivariate time series $X \in \mathbb{R}^{N \times T}$
- Transform X into distribution of SPDs

$\rightarrow$ Brain-Age prediction



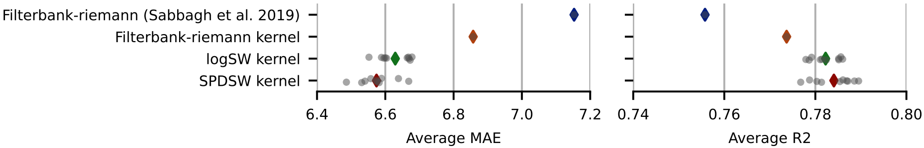
Data X with T time samples

$\longrightarrow$



Distribution of SPD matrices

Brain-Age Prediction



Positive definite Gaussian Kernel with SPDSW

$$K(\mu,\nu) = e^{-\gamma \mathrm{SPDSW}_2^2(\mu,\nu)} = e^{-\gamma \|\Phi(\mu)-\Phi(\nu)\|_{\mathcal{H}}^2}$$

Known feature map Φ, no need for expensive quadratic computations

$\rightarrow$ **Kernel Ridge** regression

SPD Matrices with Other Metrics

Other Pullback-Euclidean metrics over SPDs:

- $O(n)$-Invariant Log-Euclidean metric (Thanwerdas and Pennec, 2023):
  - $\forall X \in S_d^{++}(\mathbb{R}),\ \phi^{p,q}(X) = F^{p,q}\big(\log(X)\big)$ with, for $A \in S_d(\mathbb{R})$,

$$F^{p,q}(A) = qA + \frac{p-q}{d}\mathrm{Tr}(A)I_d$$

  - $\forall X \in S_d^{++}(\mathbb{R}),\ P^A(X) = \langle F^{p,q}\big(\log(X)\big), F^{p,q}(A)\rangle_F.$
- Log-Cholesky metric (Lin, 2019):
  - $\forall X = LL^T \in S_d^{++}(\mathbb{R}),\ \phi(X) = \lfloor L \rfloor + \log\big(\mathrm{diag}(L)\big)$
  - $\forall X = LL^T \in S_d^{++}(\mathbb{R}),\ P^A(X) = \langle \lfloor L \rfloor, \lfloor A \rfloor\rangle + \langle \log\big(\mathrm{diag}(L)\big), \frac{1}{2}\mathrm{diag}(A)\rangle_F.$
- Adaptative Log-Euclidean metric (Chen et al., 2023):
  - $\forall X \in S_d^{++}(\mathbb{R}),\ \phi(X) = \log_\alpha(X)$ with $\alpha = (a_1, \dots, a_d) \in \mathbb{R}_+^d \setminus \{(1, \dots, 1)\}$

Product of Manifolds

Let $\big((\mathcal{M}_i, g_i)\big)_{i=1}^n$ n Hadamard manifolds.

Product Manifold:

- $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_n$
- For $x = (x_1, \dots, x_n) \in \mathcal{M}$, $g(x) = \sum_{i=1}^n g_i(x_i)$
- $T_x\mathcal{M} = T_{x_1}\mathcal{M}_1 \times \cdots \times T_{x_n}\mathcal{M}$
- Geodesic distance: $\forall x, y \in \mathcal{M},\ d(x,y)^2 = \sum_{i=1}^n d(x_i, y_i)^2$
- Geodesic passing through $o = (o_1, \dots, o_n)$ in direction $v = (v_1, \dots, v_n) \in T_o\mathcal{M}$:

$$\forall t \in \mathbb{R},\ \gamma_o(t) = \big(\exp_{o_1}(tv_1), \dots, \exp_{o_n}(tv_n)\big)$$

Projections:

- Closed-form for the geodesic projection?
- Busemann function: For $(\lambda_i)_{i=1}^n$ such that $\sum_{i=1}^n \lambda_i^2 = 1$ and $\gamma : t \mapsto \big(\gamma_1(\lambda_1 t), \dots \gamma_n(\lambda_n t)\big)$,

$$B^\gamma(x) = \sum_{i=1}^n \lambda_i B^{\gamma_i}(x_i).$$

Table of Contents

Gradient Flows

Goal: $\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)$ for $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

Example

- $\mathcal{F}(\mu) = \mathrm{KL}(\mu||\nu)$ for sampling from $\nu \propto e^{-V(x)}$
- $\mathcal{F}(\mu) = D(\mu, \nu)$ for sampling from ν

Definition (Gradient Flow)

A gradient flow is a curve $\rho : [0, T] \to \mathcal{P}_2(\mathbb{R}^d)$ which decreases as much as possible along the functional $\mathcal{F}$.

Gradient Flows

Goal: $\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)$ for $\mathcal{F}:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$.

Example

- $\mathcal{F}(\mu) = \mathrm{KL}(\mu||\nu)$ for sampling from $\nu \propto e^{-V(x)}$
- $\mathcal{F}(\mu) = D(\mu,\nu)$ for sampling from ν

Definition (Gradient Flow)

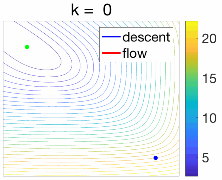
A gradient flow is a curve $\rho:[0,T]\to\mathcal{P}_2(\mathbb{R}^d)$ which decreases as much as possible along the functional $\mathcal{F}$.

For $F:\mathbb{R}^d\to\mathbb{R}$ differentiable:

- Need to solve

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F\big(x(t)\big) \\ x(0) = x_0 \end{cases}$$

- Or approximate it by a time discretization
- Gradient descent/Proximal point algorithm



From (Bach, 2020)

Wasserstein Gradient Flows

Goal: $\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)$ for $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

Wasserstein Gradient Flows

Wasserstein gradient flows of $\mathcal{F}$: curve $t \mapsto \rho_t$ satisfying (weakly)

$$\partial_t \rho_t - \mathrm{div}\big(\rho_t \nabla_{W_2} \mathcal{F}(\rho_t)\big) = 0,$$

where for all $\xi \in L^2(\mu)$,

$$\mathcal{F}\big((\mathrm{Id} + \epsilon \xi)_\# \mu\big) = \mathcal{F}(\mu) + \epsilon \int \langle \nabla_{W_2} \mathcal{F}(\mu)(x), \xi(x) \rangle \, \mathrm{d}\mu(x) + o(\epsilon).$$

- Approximated with the forward Euler scheme as:

$$\forall k \geq 0,\ \mu_{k+1} = \big(\mathrm{Id} - \tau \nabla_{W_2} \mathcal{F}(\mu_k)\big)_\# \mu_k = \exp_{\mathrm{Id}}\big(-\tau \nabla_{W_2} \mathcal{F}(\mu_k)\big)_\# \mu_k$$

- Particle approximation: $\hat{\mu}_k^n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i^k}$

$$\forall k \geq 0, i \in \{1, \dots, n\},\ x_i^{k+1} = \exp_{x_i^k}\big(-\tau \nabla_{W_2} \mathcal{F}(\hat{\mu}_k^n)(x_i^k)\big)$$

Wasserstein Gradient of CHSW

Let $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{CHSW}_2^2(\mu,\nu)$ for $\mu,\nu \in \mathcal{P}_2(\mathcal{M})$.

Wasserstein gradient of $\mathcal{F}$

For all $x \in \mathcal{M}$,

$$\nabla_{W_2}\mathcal{F}(\mu)(x) = \int_{S_o} \psi_v'\big(P^v(x)\big)\mathrm{grad}_{\mathcal{M}} P^v(x)\ \mathrm{d}\lambda(v),$$

with ψ_v the Kantorovich potential between $P^v_\#\mu$ and $P^v_\#\nu$:

$$\forall s \in \mathbb{R},\ \psi_v'(s) = s - F^{-1}_{P^v_\#\nu}\big(F_{P^v_\#\mu}(s)\big).$$

- Continuity equation:

$$\partial_t \mu_t + \mathrm{div}(\mu_t v_t) = 0 \quad \text{with} \quad v_t = -\int_{S_o} \psi_v'\big(P^v(x)\big)\mathrm{grad}_{\mathcal{M}} P^v(x)\ \mathrm{d}\lambda(v)$$

- Algorithm: For all $k \geq 0$, $i \in \{1,\dots,n\}$,

$$x_i^{k+1} = \exp_{x_i^k}\big(\tau \hat{v}_k(x_i^k)\big) \quad \text{with} \quad \hat{v}_k(x) = -\frac{1}{L}\sum_{\ell=1}^{L} \psi'_{v_\ell,k}\big(P^{v_\ell}(x)\big)\mathrm{grad}_{\mathcal{M}} P^{v_\ell}(x).$$

Wasserstein Gradient of SW

Let $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{SW}_2^2(\mu,\nu)$ for $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

Wasserstein gradient of $\mathcal{F}$ (Bonnotte, 2013; Liutkus et al., 2019)

For $\theta \in S^{d-1}$, $P^\theta(x) = \langle x,\theta\rangle$, $\mathrm{grad} P^\theta(x) = \nabla P^\theta(x) = \theta$. For all $x \in \mathbb{R}^d$,

$$\nabla_{W_2}\mathcal{F}(\mu)(x) = \int_{S^{d-1}} \psi_\theta'\big(P^\theta(x)\big)\theta \ \mathrm{d}\lambda(\theta),$$

with ψ_θ the Kantorovich potential between $P^\theta_\#\mu$ and $P^\theta_\#\nu$:

$$\forall s \in \mathbb{R},\ \psi_\theta'(s) = s - F^{-1}_{P^\theta_\#\nu}\big(F_{P^\theta_\#\mu}(s)\big).$$

- Continuity equation:

$$\partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0 \quad \text{with} \quad v_t = -\int_{S^{d-1}} \psi_\theta'\big(\langle\theta,x\rangle\big)\theta \ \mathrm{d}\lambda(\theta)$$

- Algorithm (SWF): For all $k \geq 0,\ i \in \{1,\dots,n\}$,

$$x_i^{k+1} = x_i^k - \frac{\tau}{L}\sum_{\ell=1}^L \psi'_{\theta_\ell,k}(\langle\theta_\ell, x_i^k\rangle)\theta_\ell$$

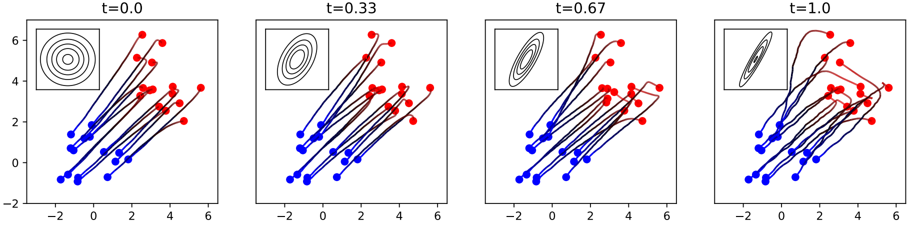
Application to Mahalanobis Space

On Mahalanobis manifold:

- $\exp_x(v) = x + v$
- $P^v(x) = x^T A v$
- $\mathrm{grad}_{\mathcal{M}} P^v(x) = v$

Algorithm: $\forall k \geq 0,\ i \in \{1, \dots, n\},\ x_i^{k+1} = x_i^k - \frac{\tau}{L} \sum_{\ell=1}^{L} \psi'_{v_\ell,k}(v_\ell^T A x_i^k) v_\ell$

SWF in the space $(\mathbb{R}^d, d_{A_t})$ with A_t interpolating between I_2 and $A \in S_d^{++}(\mathbb{R})$

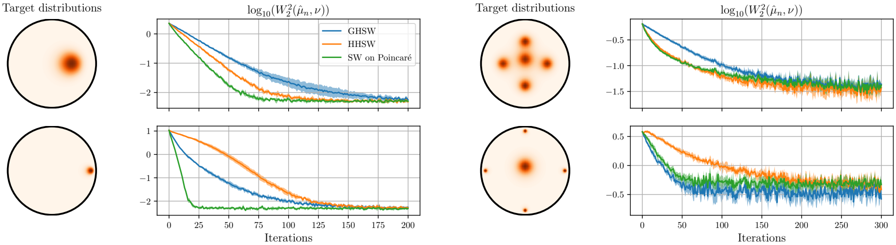


Application to Hyperbolic Space

On Lorentz model:

- $\forall x \in \mathbb{L}^d,\ v \in T_x\mathbb{L}^d,\ \exp_x(v) = \cosh(t\|v\|_{\mathbb{L}})x + \sinh(t\|v\|_{\mathbb{L}})\frac{v}{\|v\|_{\mathbb{L}}}$
- $P^v(x) = \operatorname{arctanh}\left(-\frac{\langle x,v\rangle_{\mathbb{L}}}{\langle x,x^0\rangle_{\mathbb{L}}}\right)$, $\operatorname{grad}_{\mathbb{L}} P^v(x) = -\frac{\langle x,x^0\rangle_{\mathbb{L}} v - \langle x,v\rangle_{\mathbb{L}} x^0}{\langle x,x^0\rangle_{\mathbb{L}}^2 - \langle x,v\rangle_{\mathbb{L}}^2}$
- $B^v(x) = \log\left(-\langle x, x^0+v\rangle_{\mathbb{L}}\right)$, $\operatorname{grad}_{\mathbb{L}} B^v(x) = \frac{x^0+v}{\langle x,x^0+v\rangle_{\mathbb{L}}} + x$

Algorithm: $\forall k \geq 0,\ x_i^{k+1} = \exp_{x_i^k}\left(-\frac{\tau}{L}\sum_{\ell=1}^L \psi'_{v_\ell,k}\left(P^{v_\ell}(x_i^k)\right)\operatorname{grad}_{\mathbb{L}} P^{v_\ell}(x_i^k)\right)$

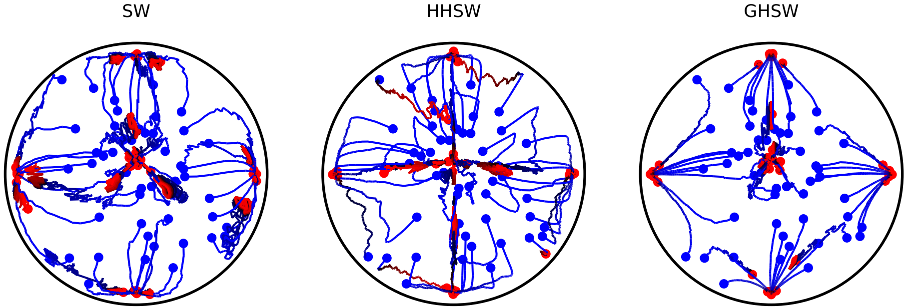


Application to Hyperbolic Space

On Lorentz model:

- $\forall x \in \mathbb{L}^d,\ v \in T_x\mathbb{L}^d,\ \exp_x(v) = \cosh(t\|v\|_{\mathbb{L}})x + \sinh(t\|v\|_{\mathbb{L}})\frac{v}{\|v\|_{\mathbb{L}}}$
- $P^v(x) = \operatorname{arctanh}\left(-\frac{\langle x,v\rangle_{\mathbb{L}}}{\langle x,x^0\rangle_{\mathbb{L}}}\right)$, $\operatorname{grad}_{\mathbb{L}} P^v(x) = -\frac{\langle x,x^0\rangle_{\mathbb{L}} v - \langle x,v\rangle_{\mathbb{L}} x^0}{\langle x,x^0\rangle_{\mathbb{L}}^2 - \langle x,v\rangle_{\mathbb{L}}^2}$
- $B^v(x) = \log\left(-\langle x, x^0+v\rangle_{\mathbb{L}}\right)$, $\operatorname{grad}_{\mathbb{L}} B^v(x) = \frac{x^0+v}{\langle x,x^0+v\rangle_{\mathbb{L}}} + x$

Algorithm: $\forall k \geq 0,\ x_i^{k+1} = \exp_{x_i^k}\left(-\frac{\tau}{L}\sum_{\ell=1}^L \psi'_{v_\ell,k}\big(P^{v_\ell}(x_i^k)\big)\operatorname{grad}_{\mathbb{L}} P^{v_\ell}(x_i^k)\right)$



Conclusion

Conclusion

- SW discrepancies on Cartan-Hadamard manifolds
- Can be applied to ML tasks on different manifolds
- Wasserstein gradient flows to minimize CHSW

Perspectives and follow-up works:

- Study other Riemannian manifolds: Sphere (Bonet et al., 2023b; Quellmalz et al., 2023)
- Extension to unbalanced setting (Séjourné et al., 2023)
- Study statistical properties
- Study convergence of the gradient flows

Conclusion

Conclusion

- SW discrepancies on Cartan-Hadamard manifolds
- Can be applied to ML tasks on different manifolds
- Wasserstein gradient flows to minimize CHSW

Perspectives and follow-up works:

- Study other Riemannian manifolds: Sphere (Bonet et al., 2023b; Quellmalz et al., 2023)
- Extension to unbalanced setting (Séjourné et al., 2023)
- Study statistical properties
- Study convergence of the gradient flows

Thank you!

References I

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33: 21428–21439, 2020.

Francis Bach. Effortless optimization through gradient flows, 2020. URL https://francisbach.com/gradient-flows/.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

Clément Bonet, Benoît Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-Wasserstein on Symmetric Positive Definite Matrices for M/EEG Signals. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2777–2805. PMLR, 23–29 Jul 2023a.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical sliced-wasserstein. In *The Eleventh International Conference on Learning Representations*, 2023b.

References II

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, pages 334–370. PMLR, 2023c.

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

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

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.

Ziheng Chen, Tianyang Xu, Zhiwu Huang, Yue Song, Xiao-Jun Wu, and Nicu Sebe. Adaptive riemannian metrics on spd manifolds. *arXiv preprint arXiv:2303.15477*, 2023.

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

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.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. *Advances in neural information processing systems*, 29, 2016.

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.

References IV

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40 (4):1353–1370, 2019.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR, 2019.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33: 20802–20812, 2020.

References V

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.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

Xavier Pennec. Manifold-valued image processing with spd matrices. In *Riemannian geometric statistics in medical image analysis*, pages 75–134. Elsevier, 2020.

Michael Quellmalz, Robert Beinert, and Gabriele Steidl. Sliced optimal transport on the sphere. *arXiv preprint arXiv:2304.09092*, 2023.

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.

References VI

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Manifold-regression to predict from meg/eeg brain signals without source modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A Engemann. Predictive regression modeling with meg/eeg: from source power to signals and cognitive states. *NeuroImage*, 222:116893, 2020.

Thibault Séjourné, Clément Bonet, Kilian Fatras, Kimia Nadjahi, and Nicolas Courty. Unbalanced optimal transport meets sliced-wasserstein. *arXiv preprint arXiv:2306.07176*, 2023.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. *arXiv preprint arXiv:1911.08411*, 2019.

Yann Thanwerdas and Xavier Pennec. O (n)-invariant riemannian metrics on spd matrices. *Linear Algebra and its Applications*, 661:163–201, 2023.

Or Yair, Felix Dietrich, Ronen Talmon, and Ioannis G Kevrekidis. Domain adaptation with optimal transport on the manifold of spd matrices. *arXiv preprint arXiv:1906.00616*, 2019.

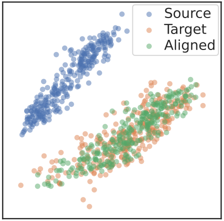
Domain Adaptation in BCI

Learning a map f_θ between a source μ and a target ν

$$\min_\theta \ \mathrm{SPDSW}_2^2\big((f_\theta)_\#\mu, \nu\big)$$

Minimizing over the particles

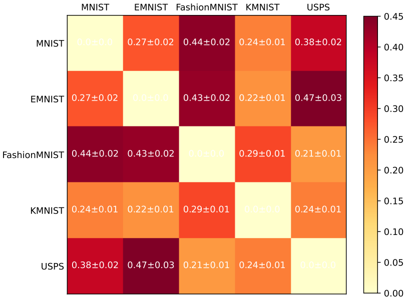
$$\min_{(x_i)_{i=1}^n} \ \mathrm{SPDSW}_2^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \nu\right)$$



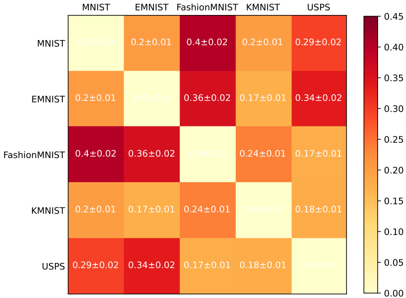
Subjects	Source	AISOTDA (Yair et al., 2019)	SPDSW	LogSW	LEW	LES	SPDSW	LogSW	LEW	LES
			Transformations in $S_d^{++}(\mathbb{R})$				Descent over particles			
1	82.21	80.90	84.70	84.48	84.34	84.70	85.20	85.20	77.94	82.92
3	79.85	87.86	85.57	84.10	85.71	86.08	87.11	86.37	82.42	81.47
7	72.20	82.29	81.01	76.32	81.23	81.23	81.81	81.73	79.06	73.29
8	79.34	83.25	83.54	81.03	82.29	83.03	84.13	83.32	80.07	85.02
9	75.76	80.25	77.35	77.88	77.65	77.65	80.30	79.02	76.14	70.45
Avg. acc.	77.87	82.93	82.43	80.76	82.24	82.54	83.71	83.12	79.13	78.63
Avg. time (s)	-	-	**4.34**	**4.32**	11.41	12.04	**3.68**	**3.67**	8.50	11.43

Dataset Comparisons (Alvarez-Melis and Fusi, 2020)

- Consider datasets as feature-label pairs
- Embed labels in $\mathbb{H}^{d_y}$
- Dataset: Distribution in $\mathbb{R}^{d_x} \times \mathbb{H}^{d_y}$



SW



Product HCHSW

For 10^4 samples, 0.05s vs 120s.