Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds

Clément Bonet[1]

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

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

SIAM UQ
28/02/2024

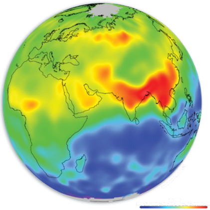
Motivations

Optimal Transport: Meaningful way to compare distributions

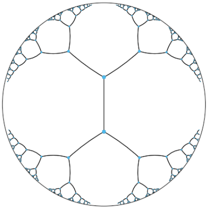
- Domain Adaptation (Courty et al., 2016)
- Generative Models (*e.g.* WGAN (Arjovsky et al., 2017))
- Document Classification (Kusner et al., 2015)

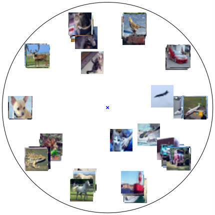
Data often lie on **manifolds**:

- Spherical data (geophysical data, directional data...)
- Hierarchical data (trees, graphs, words, images...) on Hyperbolic spaces
- M/EEG data on the space of Symmetric Positive Definite Matrices (SPDs)



Source: ESA





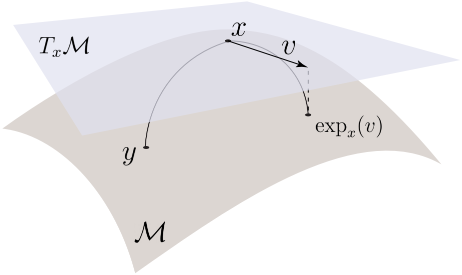
Riemannian Manifolds

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 curve $\gamma : [0, 1] \to \mathcal{M}$ 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

$\to$ Geodesics curves aperiodics of the form $\gamma(t) = \exp_x(tv)$ for $t \in \mathbb{R}$, $v \in T_x\mathcal{M}$

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

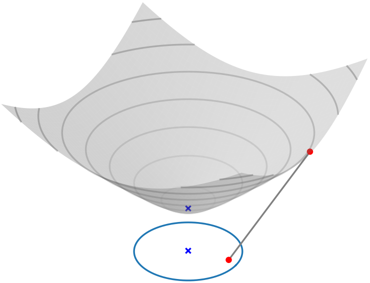
Different isometric models:

- **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 $\mu, \nu \in \mathcal{P}_2(\mathcal{M})$,

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

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

Properties:

- W_2 distance
- Metrizes the weak convergence
- Riemannian structure

Solving the OT Problem

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

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

Computational Complexity (Pele and Werman, 2009)

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

Solving the OT Problem

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

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

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 $\hat{\nu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$,

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

Solving the OT Problem

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

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

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 $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$,

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

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:

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

1D Wasserstein Distance

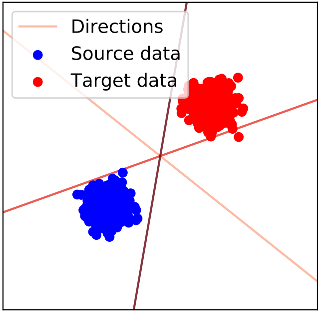
$$W_2^2(\mu,\nu) = \int_0^1 \big|F_\mu^{-1}(u) - F_\nu^{-1}(u)\big|^2\ \mathrm{d}u = \big\|F_\mu^{-1} - F_\nu^{-1}\big\|_{L^2([0,1])}^2$$

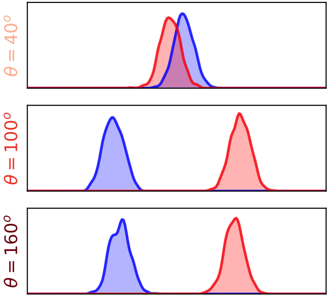
Let $x_1 < \dots < x_n,\ y_1 < \dots < y_n$, $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$,

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

$\rightarrow O(n \log n)$

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

SW on Cartan-Hadamard Manifolds

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

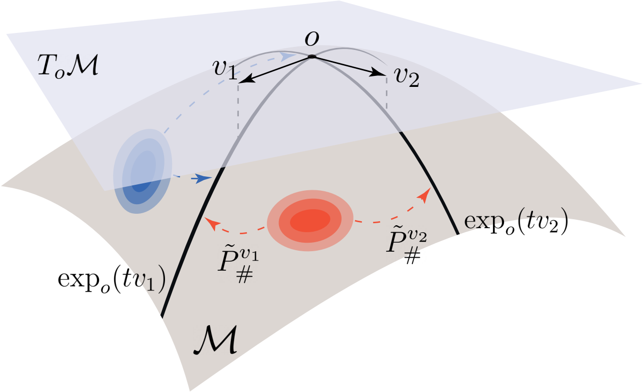
	SW	CHSW
Closed-form of W	Line	?
Projection	$P^\theta(x) = \langle x, \theta \rangle$	?
Integration	S^{d-1}	?

Projecting on Geodesics

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

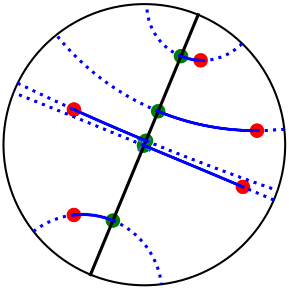
$$\forall v \in T_o\mathcal{M},\ \mathcal{G}_v = \{\exp_o(tv),\ t \in \mathbb{R}\}$$

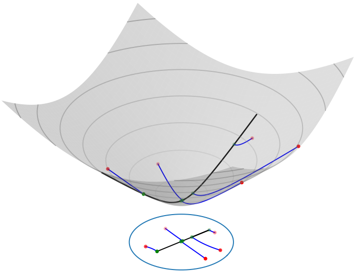
- Geodesics isometric to $\mathbb{R}$
- Integrate along all possible directions on $S_o = \{v \in T_o\mathcal{M},\ \|v\|_o = 1\}$



Projections

1. **Geodesic projections**:
   - On Euclidean space: For $\theta \in S^{d-1}$, $\mathcal{G}_\theta = \{t\theta,\ t \in \mathbb{R}\}$, $\exp_0(t\theta) = 0 + t\theta = t\theta$,
$$\forall x \in \mathbb{R}^d,\ P^\theta(x) = \langle x, \theta \rangle = \operatorname*{argmin}_{t\in\mathbb{R}}\ \|x - t\theta\|_2 = \operatorname*{argmin}_{t\in\mathbb{R}}\ d\big(x, \exp_0(t\theta)\big)$$
   - On Cartan-Hadamard manifold: For $v \in T_o\mathcal{M}$, $\mathcal{G}_v = \{\exp_o(tv),\ t \in \mathbb{R}\}$,
$$\forall x \in \mathcal{M},\ P^v(x) = \operatorname*{argmin}_{t\in\mathbb{R}}\ d\big(x, \exp_o(tv)\big)$$



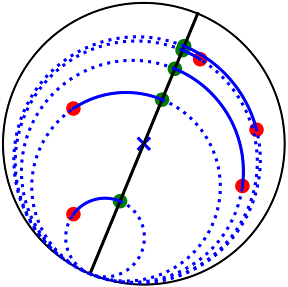


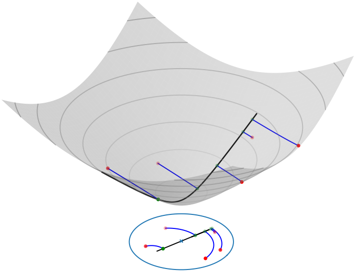
Projections

1. **Geodesic projections**: $\forall x \in \mathcal{M},\ P^v(x) = \operatorname*{argmin}_{t\in\mathbb{R}}\ d\big(x, \exp_o(tv)\big)$
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 $\mathcal{P}_2(\mathcal{M}) \to$ 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, $K(\mu,\nu) = \exp\big(-\gamma \mathrm{CHSW}_2^2(\mu,\nu)\big)$ 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 $\mathbb{L}^d$:

$$\forall v \in T_{x^0}\mathbb{L}^d \cap S^d,\ x \in \mathbb{L}^d,$$

$$P^v(x) = \operatorname{arctanh}\left(-\frac{\langle x, v\rangle_{\mathbb{L}}}{\langle x, x^0\rangle_{\mathbb{L}}}\right)$$

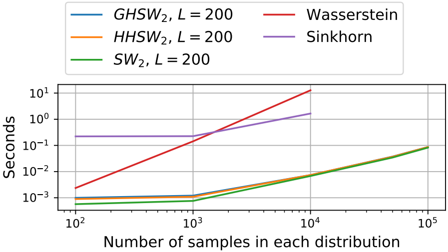
$$B^v(x) = \log\big(-\langle x, x^0 + v\rangle_{\mathbb{L}}\big)$$

$$\forall \tilde{v} \in S^{d-1},\ y \in \mathbb{B}^d,$$

$$P^{\tilde{v}}(y) = 2\operatorname{arctanh}\big(s(y)\big)$$

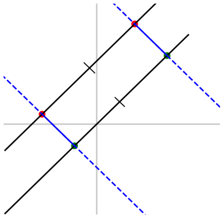
$$B^{\tilde{v}}(y) = \log\left(\frac{\|\tilde{v} - y\|_2^2}{1 - \|y\|_2^2}\right)$$

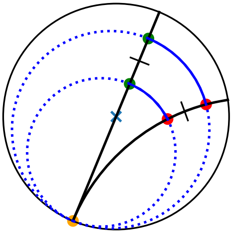
Method	Complexity
Wasserstein + LP	$O(n^3 \log n + n^2 d)$
Sinkhorn	$O(n^2 d)$
SW	$O\big(Ln(d + \log n)\big)$
GHSW	$O\big(Ln(d + \log n)\big)$
HHSW	$O\big(Ln(d + \log n)\big)$



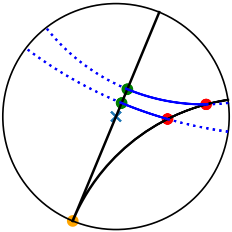
Comparison of the Projections

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





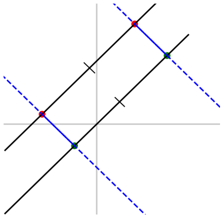
Horospherical projection

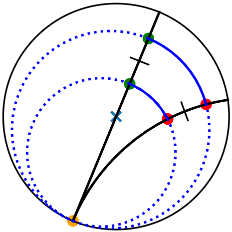


Geodesic projection

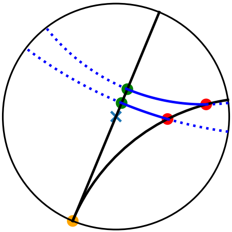
Comparison of the Projections

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



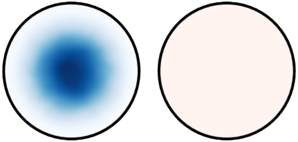


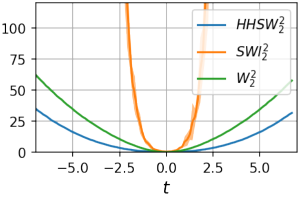
Horospherical projection

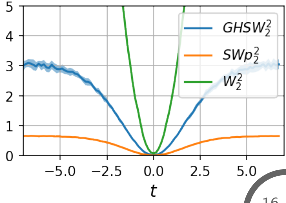


Geodesic projection

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

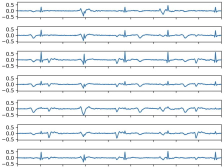
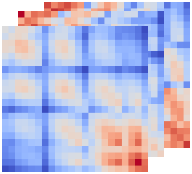
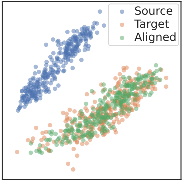






Applications to Different Hadamard Manifolds

- Hyperbolic spaces: Deep classification with prototypes (Bonet et al., 2023c)
- **Pullback Euclidean Manifolds**
  - Additional properties: distance, metrize the weak convergence
  - Mahalanobis manifolds: Application to document classification (Kusner et al., 2015)
- Space of **Symmetric Positive Definite matrices** (SPDs) (Bonet et al., 2023a)
  - Application to **M/EEG data**: Brain-age prediction
  - Application to **BCI**: Domain adaptation

- Product of Hadamard manifolds: Comparison of datasets

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)

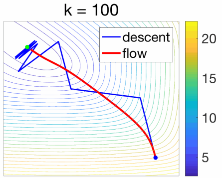
A gradient flow is a curve $\rho:[0,T]\to\mathcal{P}_2(\mathbb{R}^d)$ which decreases as much as possible along the functional $\mathcal{F}$.

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 $k \geq 0,\ i \in \{1,\dots,n\}$,

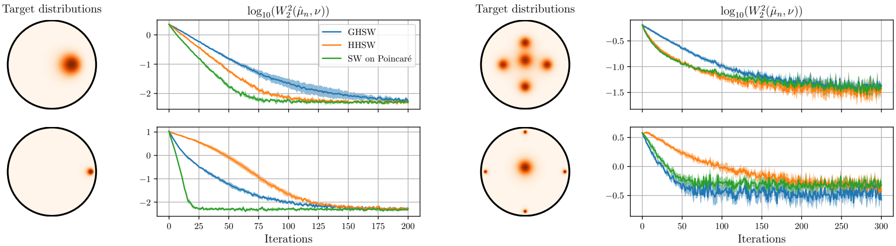
$$x_i^{k+1} = x_i^k - \frac{\tau}{L}\sum_{\ell=1}^L \psi'_{\theta_\ell,k}(\langle\theta_\ell, x_i^k\rangle)\theta_\ell$$

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

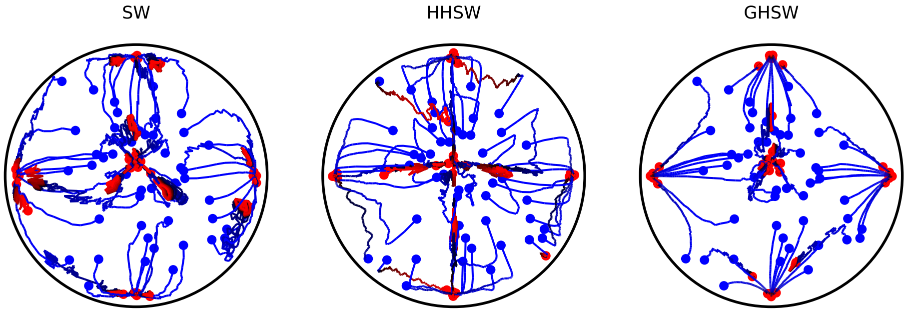


Application to Hyperbolic Space

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

Follow-up works and perspectives:

- Study other Riemannian manifolds: Sphere (Bonet et al., 2023b; Quellmalz et al., 2023, 2024; Tran et al., 2024; Garrett et al., 2024)
- 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

Follow-up works and perspectives:

- Study other Riemannian manifolds: Sphere (Bonet et al., 2023b; Quellmalz et al., 2023, 2024; Tran et al., 2024; Garrett et al., 2024)
- Extension to unbalanced setting (Séjourné et al., 2023)
- Study statistical properties
- Study convergence of the gradient flows

Thank you!

References I

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

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.

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.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

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.

Robert C Garrett, Trevor Harris, Bo Li, and Zhuo Wang. Validating climate models with spherical convolutional wasserstein distance. *arXiv preprint arXiv:2401.14657*, 2024.

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

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.

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.

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.

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

References V

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.

Michael Quellmalz, Léo Buecher, and Gabriele Steidl. Parallelly sliced optimal transport on spheres and on the rotation group. *arXiv preprint arXiv:2401.16896*, 2024.

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.

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.

Huy Tran, Yikun Bai, Abihith Kothapalli, Ashkan Shahbazi, Xinran Liu, Rocio Diaz Martin, and Soheil Kolouri. Stereographic spherical sliced wasserstein distances, 2024.