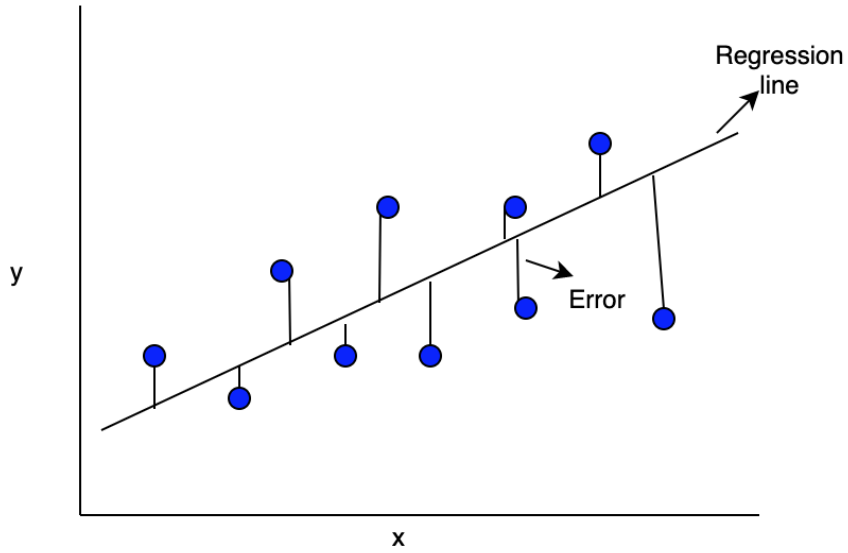
Fréchet Regression on the Bures-Wasserstein Manifold

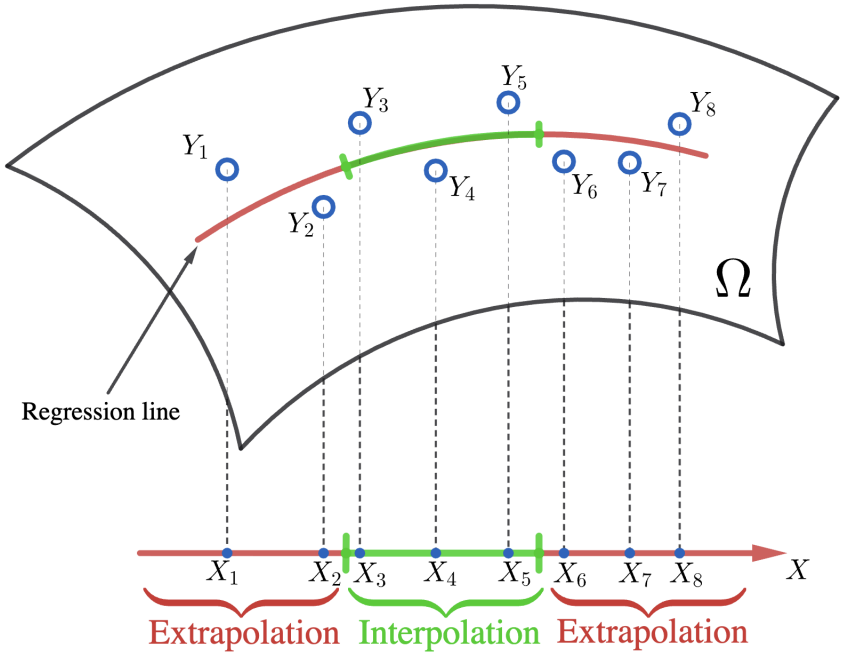
Duc Toan Nguyen[1], César A. Uribe [1]

[1]Department of Electrical and Computer Engineering, Rice University
duc.toan.nguyen@rice.edu, cauribe@rice.edu

Generalizing Linear Regression



(a) Euclidean space



(b) Non-Euclidean space

Fréchet Regression

Set-up: Assume the existence of a joint distribution $(X, Y) \sim \mathcal{F}$, where the sample spaces of X and Y are in $(\mathbb{R}^p, \|.\|_2)$ and (Ω, d), respectively.
Global Fréchet regression:

$$\begin{aligned} m(x) &= \arg\min_{\omega \in \Omega} \mathbb{E}_{(X,Y)\sim\mathcal{F}} \left[d^2(Y, \omega) \mid X = x \right] \\ &= \arg\min_{\omega \in \Omega} \mathbb{E}_{(X,Y)\sim\mathcal{F}} \left[s_G(x) d^2(Y, \omega) \right], \end{aligned} \tag{1}$$

where $s_G(x) = 1 + (X - \mu)^\top \Sigma^{-1} (x - \mu)$, with $\mu = \mathbb{E}[X]$ and $\Sigma = \mathrm{Var}(X)$.

Fréchet Regression

Set-up: Assume the existence of a joint distribution $(X, Y) \sim \mathcal{F}$, where the sample spaces of X and Y are in $(\mathbb{R}^p, \|.\|_2)$ and (Ω, d), respectively.
Global Fréchet regression:

$$\begin{aligned} m(x) &= \arg\min_{\omega \in \Omega} \mathbb{E}_{(X,Y)\sim\mathcal{F}} \left[d^2(Y, \omega) \mid X = x \right] \\ &= \arg\min_{\omega \in \Omega} \mathbb{E}_{(X,Y)\sim\mathcal{F}} \left[s_G(x) d^2(Y, \omega) \right], \end{aligned} \tag{1}$$

where $s_G(x) = 1 + (X - \mu)^\top \Sigma^{-1}(x - \mu)$, with $\mu = \mathbb{E}[X]$ and $\Sigma = \mathrm{Var}(X)$.
Empirical version: Given n independent samples $(X_k, Y_k) \sim \mathcal{F}$, $k \in \{1, ..., n\}$,

$$\hat{m}_G(x) = \arg\min_{\omega \in \Omega} \frac{1}{n} \sum_{k=1}^{n} s_{G,k}(x) d^2(Y_k, \omega), \tag{2}$$

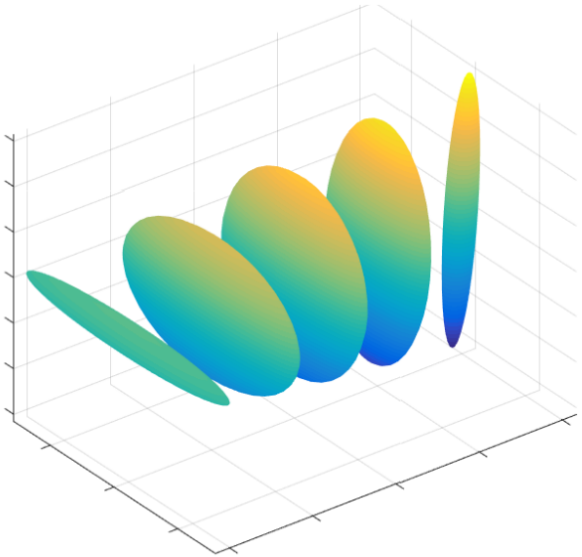
where $s_{G,k}(x) = 1 + (X_k - \bar{X})^\top \hat{\Sigma}^{-1}(x - \bar{X})$, for $k \in \{1, ..., n\}$, $\bar{X}$ is the sample mean, and $\hat{\Sigma}$ is the sample covariance matrix of $\{X_k\}_{k=1}^n$.

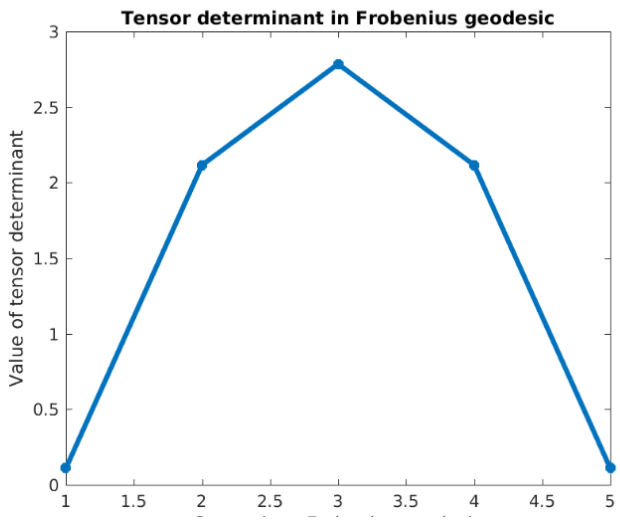
Bures-Wasserstein Manifold

- The Bures-Wasserstein (BW) manifold $\mathbb{S}^d_{++}$ is the space of $d \times d$ symmetric positive definite (SPD) matrices.
- Distance square $W^2(S_1, S_2)$ is the 2-Wasserstein distance between the zero-mean Gaussians $\mathcal{N}(0, S_1)$ and $\mathcal{N}(0, S_2)$.

Bures-Wasserstein Manifold

- The Bures-Wasserstein (BW) manifold $\mathbb{S}^d_{++}$ is the space of $d \times d$ symmetric positive definite (SPD) matrices.
- Distance square $W^2(S_1, S_2)$ is the 2-Wasserstein distance between the zero-mean Gaussians $\mathcal{N}(0, S_1)$ and $\mathcal{N}(0, S_2)$.
- **Why BW?**
  - BW avoids Frobenius "swelling effect" (Feragen and Fuster (2016)),
  - BW does **NOT** require matrix logarithms, unlike affine-invariant (Fisher-Rao) and log-Euclidean metrics.





Problem Formulation

Main Problem

Given n SPD matrices Σ_k and their weights $\lambda_k \in \mathbb{R}$, for $k \in \{1, ..., n\}$, such that $\sum_{k=1}^{n} \lambda_k = 1$, we want to solve

$$\begin{aligned} \min_{S \in \mathbb{S}^d_{++}} F(S) &:= \sum_{k=1}^{n} \lambda_k W_2^2(S, \Sigma_k) \\ &= \sum_{i \in \mathcal{I}} \lambda_i^+ W_2^2(S, \Sigma_i) - \sum_{j \in \mathcal{J}} \lambda_j^- W_2^2(S, \Sigma_j), \end{aligned} \tag{3}$$

for $\lambda_i^+, \lambda_j^- > 0, \mathcal{I} = \{k : \lambda_k > 0\}, \mathcal{J} = \{k : \lambda_k < 0\}$.

Research Questions

Positive weights (Interpolation): Agueh and Carlier (2011) showed that Problem (3) has a unique minimizer for any set of SPD matrices [1].

Example (Negative weights - Extrapolation): Let $\lambda_1 = 2$, $\lambda_2 = -1$, $\Sigma_1 = I$, $\Sigma_2 = 9I$, and $f(S) := \lambda_1 W_2^2(S, \Sigma_1) + \lambda_2 W_2^2(S, \Sigma_2)$. Then, the gradient $\nabla f(S) \succ 0$, for all $S \in \mathbb{S}_{++}^d$.

Research Questions

Positive weights (Interpolation): Agueh and Carlier (2011) showed that Problem (3) has a unique minimizer for any set of SPD matrices [1].

Example (Negative weights - Extrapolation): Let $\lambda_1 = 2$, $\lambda_2 = -1$, $\Sigma_1 = I$, $\Sigma_2 = 9I$, and $f(S) := \lambda_1 W_2^2(S, \Sigma_1) + \lambda_2 W_2^2(S, \Sigma_2)$. Then, the gradient $\nabla f(S) \succ 0$, for all $S \in \mathbb{S}_{++}^d$.

Questions:

1. What **conditions** guarantee the **existence** of the solution of Problem (3) in the case of (possibly) **negative weights**?

Research Questions

Positive weights (Interpolation): Agueh and Carlier (2011) showed that Problem (3) has a unique minimizer for any set of SPD matrices [1].

Example (Negative weights - Extrapolation): Let $\lambda_1 = 2$, $\lambda_2 = -1$, $\Sigma_1 = I$, $\Sigma_2 = 9I$, and $f(S) := \lambda_1 W_2^2(S, \Sigma_1) + \lambda_2 W_2^2(S, \Sigma_2)$. Then, the gradient $\nabla f(S) \succ 0$, for all $S \in \mathbb{S}^d_{++}$.

Questions:

1. What **conditions** guarantee the **existence** of the solution of Problem (3) in the case of (possibly) **negative weights**?

2. What **conditions** does the **unique existence** of the minimizer hold?

Research Questions

Positive weights (Interpolation): Agueh and Carlier (2011) showed that Problem (3) has a unique minimizer for any set of SPD matrices [1].

Example (Negative weights - Extrapolation): Let $\lambda_1 = 2$, $\lambda_2 = -1$, $\Sigma_1 = I$, $\Sigma_2 = 9I$, and $f(S) := \lambda_1 W_2^2(S, \Sigma_1) + \lambda_2 W_2^2(S, \Sigma_2)$. Then, the gradient $\nabla f(S) \succ 0$, for all $S \in \mathbb{S}^d_{++}$.

Questions:

1. What **conditions** guarantee the **existence** of the solution of Problem (3) in the case of (possibly) **negative weights**?
2. What **conditions** does the **unique existence** of the minimizer hold?
3. What **algorithms** can solve Problem (3) with **theoretical convergence guarantee**?

Conditions for the existence of minimizer

Theorem (Spectral Dominance of Positive Weights)

Let $\Sigma_1, \ldots, \Sigma_n \in \mathbb{S}^d_{++}$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ with $\sum_{k=1}^n \lambda_k = 1$. If the **Spectral Dominance of Positive Weights** *condition holds, i.e.,*

$$\sum_{i \in \mathcal{I}} \lambda_i^+ \sqrt{\lambda_{\min}(\Sigma_i)} > \sum_{j \in \mathcal{J}} \lambda_j^- \sqrt{\lambda_{\max}(\Sigma_j)}, \tag{4}$$

then Problem (3) admits a solution.

Properties of stationary points:

Since the function $F(S)$ is differentiable on $\mathbb{S}^d_{++}$, under condition (4), there is at least one stationary point S_* that satisfies

$$S_* = \sum_k \lambda_k \left(S_*^{1/2} \Sigma_k S_*^{1/2} \right)^{1/2}. \tag{5}$$

Bounded Region: If the condition in Theorem 1 holds, then any stationary point S_* has the following property:

$$\left(\sum_{i \in \mathcal{I}} \lambda_i^+ \sqrt{\lambda_{\min}(\Sigma_i)} - \sum_{j \in \mathcal{J}} \lambda_j^- \sqrt{\lambda_{\max}(\Sigma_j)} \right)^2 I \prec S_*,$$

$$\left(\sum_{i \in \mathcal{I}} \lambda_i^+ \sqrt{\lambda_{\max}(\Sigma_i)} - \sum_{j \in \mathcal{J}} \lambda_j^- \sqrt{\lambda_{\min}(\Sigma_j)} \right)^2 I \succ S_*.$$

No local maximum: S_* is a local minimum or a saddle.

Conditions for the unique existence

Following the approach in (Wintraecken, 2015, Theorem 3.4.9), we have

Theorem (Unique existence of the minimizer)

Let $\Sigma_k \in \mathbb{S}^d_{++}$, *for* $k \in \{1, ..., n\}$, $\lambda := \min \lambda_{\min}(\Sigma_k)$ *and* $\mathcal{S}_\lambda = \{\Sigma \in \mathbb{S}^d_{++} \ : \ \lambda_{\min}(\Sigma) \geq \lambda\}$. *Denote* $\mu_+ = \sum_i \lambda_i^+, \mu_- = \sum_j \lambda_j^-$. *Assume that there exist* $\rho > 0$, $r > 0$ *and* $\Sigma_0 \in \mathcal{S}_\lambda$ *such that*

- $\Sigma_k \in B_r(\Sigma_0)$ *for all* $k \in \{1, ..., n\}$, *where* $B_r(\Sigma_0)$ *is the geodesic ball centered at* Σ_0 *with radius* r,
- $r < \rho/(\mu_+ + \mu_-)$, $\rho < \sqrt{\lambda}/2$, $B_\rho(\Sigma_0) \subset \mathcal{S}_\lambda$, *and*
- $\mu_+/\mu_- > (2\rho\sqrt{\Lambda^+})/(\tanh(2\rho\sqrt{\Lambda^+}))$.

Then, Problem (3) *has a unique minimizer in* $B_\rho(\Sigma_0)$.

Remark: The conditions in Proposition 2 are **stronger** than the conditions in Theorem 1.

Riemannian Gradient Descent

Based on BWGD (Altschuler et al. (2021)),

Algorithm General BW Barycenter Gradient Descent

1: **Input**: SPD matrices Σ_k, weights λ_k, initial S_0, step-size η, epochs T.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad \tilde{S}_t = (1-\eta)I + \eta \sum_{k=1}^{n} \lambda_k \, \mathrm{GM}(S_{t-1}^{-1}, \Sigma_k)$ // $-\eta \nabla_{\mathrm{bw}} F(S_{t-1})$
4: $\quad S_t = \tilde{S}_t S_{t-1} \tilde{S}_t$ // $\exp_{S_{t-1}}(-\eta \nabla_{\mathrm{bw}} F(S_{t-1}))$
5: **end for**
6: **Return** $\bar{\Sigma} = S_T$

General BWGD - Convergence Analysis

Proposition (All iterations stay inside $\mathbb{S}^d_{++}$)

With **Spectral Dominance of Positive Weights**, All iterations S_t generated from Algorithm 1 stay inside $\mathbb{S}^d_{++}$.

Lemma (L-smoothness of F)

The function $F(S)$ is $L-$smooth with $L = \sum_k |\lambda_k| > 1$.

Theorem (Convergence rate of RGD)

Let the conditions of Theorem 1 hold, $\eta \leq 1/L$ where $L = \sum_k |\lambda_k|$, $T > 0$ and $S_0 \in \mathbb{S}^d_{++}$. Let F_* be the minimum value of Problem (3). Then, the sequence $\{S_t\}_{t=0}^{T-1}$ generated by Algorithm 1 has the following property:

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla_{\mathrm{bw}} F(S_t)\|^2_{S_t} \leq \frac{2L(F(S_0) - F_*)}{T}.$$

Pairwise Reformulation

$$
\begin{aligned}
F(S) &= \sum_{i\in\mathcal{I}} \lambda_i^+ W_2^2(S,\Sigma_i) - \sum_{j\in\mathcal{J}} \lambda_j^- W_2^2(S,\Sigma_j) \\
&= \sum_{i\in\mathcal{I},j\in\mathcal{J}} \frac{\lambda_i^+}{\mu_+}.\frac{\lambda_j^-}{\mu_-}\left(\mu_+ W_2^2(S,\Sigma_i) - \mu_- W_2^2(S,\Sigma_j)\right) \\
&= \sum_{i\in\mathcal{I},j\in\mathcal{J}} \frac{\lambda_i^+}{\mu_+}.\frac{\lambda_j^-}{\mu_-} f_{ij}(S), \qquad (6)
\end{aligned}
$$

for $\lambda_i^+, \lambda_j^- > 0, \mu_+ := \sum_i \lambda_i^+, \mu_- := \sum_j \lambda_j^-, S \in \mathbb{S}_{++}^d$.

Pairwise BWGD

Stochastic Gradient: $\nabla f_{ij}(S) = I - \left(\mu_+ \mathrm{GM}(S^{-1}, \Sigma_i) - \mu_- \mathrm{GM}(S^{-1}, \Sigma_j)\right)$.
We can use some off-the-shelf Stochastic Riemannian Optimization algorithms, such as R-SGD [2], R-SVRG [6], R-SRG [3], or R-SPIDER [7].

Pairwise BWGD

Stochastic Gradient: $\nabla f_{ij}(S) = I - \left(\mu_+ \mathrm{GM}(S^{-1}, \Sigma_i) - \mu_- \mathrm{GM}(S^{-1}, \Sigma_j)\right)$. We can use some off-the-shelf Stochastic Riemannian Optimization algorithms, such as R-SGD [2], R-SVRG [6], R-SRG [3], or R-SPIDER [7].

Conditions for Pairwise Formulation

$$(\mu_+) \min_{i \in \mathcal{I}} \sqrt{\lambda_{\min}(\Sigma_i)} > (\mu_-) \max_{j \in \mathcal{J}} \sqrt{\lambda_{\max}(\Sigma_j)} \tag{7}$$

Corollary

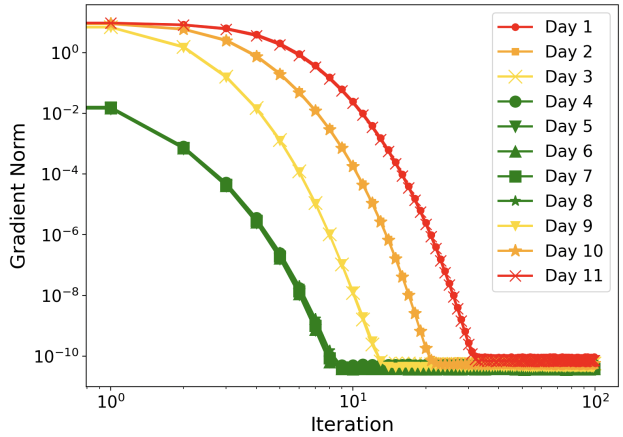
Let $\Sigma_k \in \mathbb{S}^d_{++}$ and corresponding weights $\lambda_k \in \mathbb{R}$, for $k \in \{1, ..., n\}$, and let (7) hold. Then, for any $S \in \mathbb{S}^d_{++}$, $i \in \mathcal{I}$, and $j \in \mathcal{J}$, it follows that

$$\mu_+ \mathrm{GM}(S^{-1}, \Sigma_i) - \mu_- \mathrm{GM}(S^{-1}, \Sigma_j) \in \mathbb{S}^d_{++}.$$

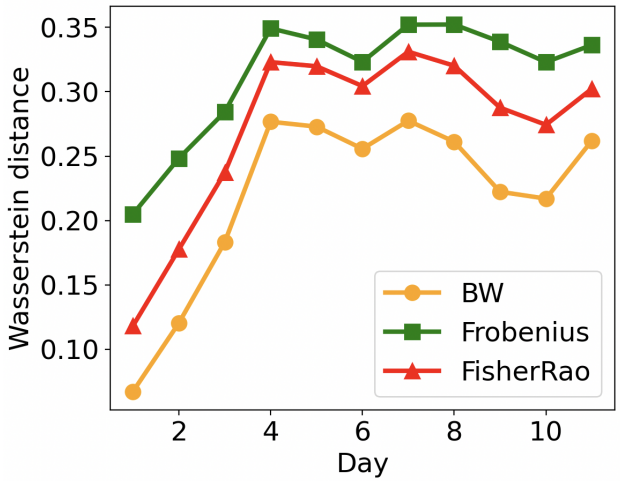
Ant Social Network: Experiment Set-up

- Network Regression on the Ant Social Organization dataset [4, 5].
- **Temporal network** consists of 11 **connected** networks (11 days), each with 113 nodes representing the ants in the first colony.
- Modify all the Laplacians as $\Sigma := L^{\dagger} + \frac{1}{d}\mathbf{1}_{d\times d}$ (Haasler et al. (2024)).
- We frame the problem as a regression task where the covariates $X = \tau$ (ranging over the 11 days) and the response is the corresponding modified Laplacian.

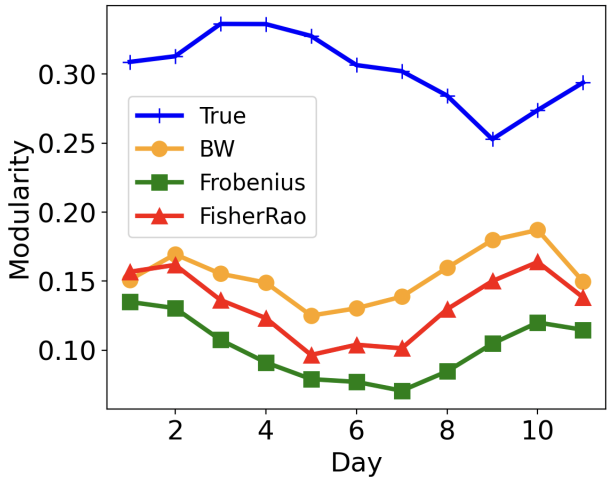
Ant Social Network: Results



(a) Gradient norms from Algorithm 1. The green lines represent the interpolation range (from day 4 to day 8 over total 11 days)



(b) Degree distributions

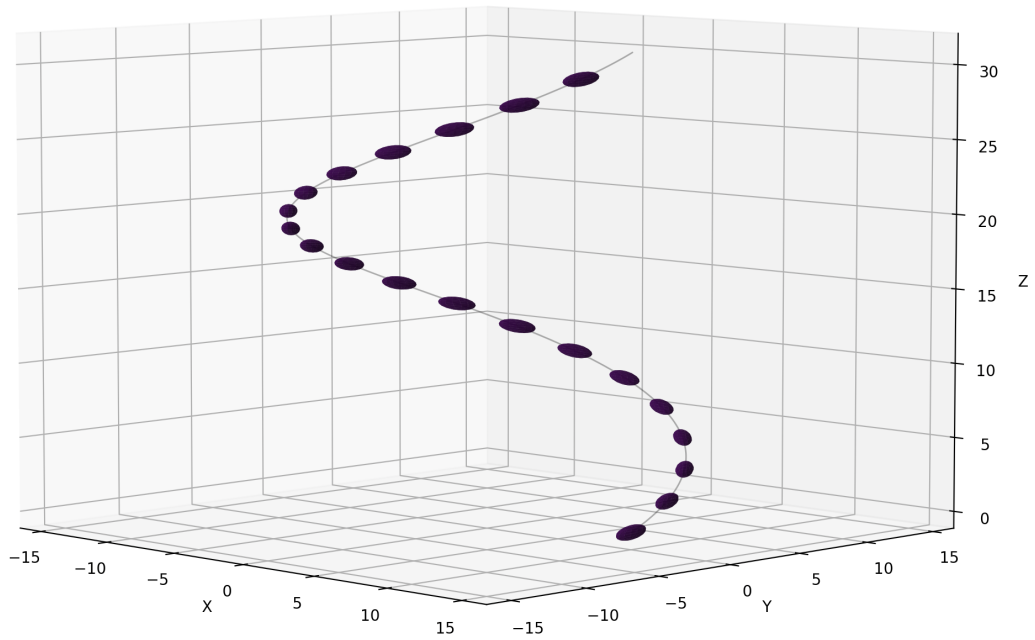


(c) Modularity

Diffusion Tensor Imaging: Experiment Set-up

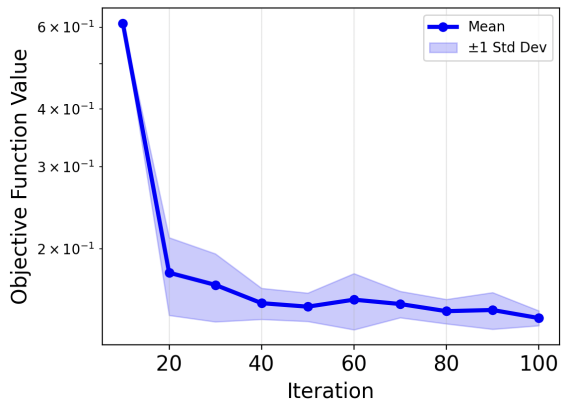
- **Dataset:** $n = 100,000$ diffusion tensors (3×3 SPD matrices).
- **Geometry:** Tensors generated along a helical backbone:

$$x(t) = 10\cos(t), \quad y(t) = 10\sin(t), \quad z(t) = 5t$$

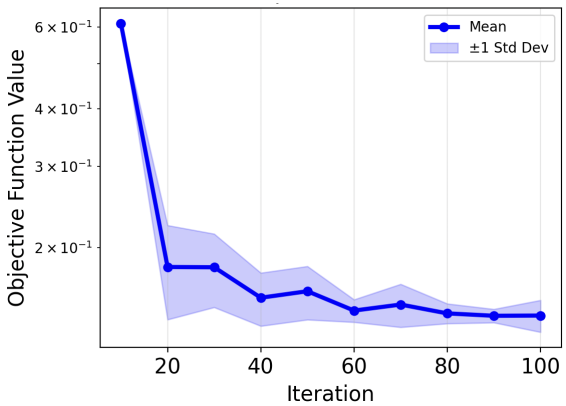


- **Task:** Fréchet Regression at covariates $X \in \{20k, 40k, 60k, 80k\}$.
- Protocol: 10 independent runs per target point.

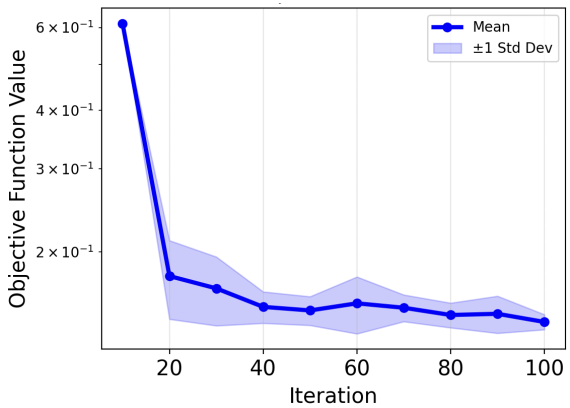
Diffusion Tensor Imaging: Results (1/2)



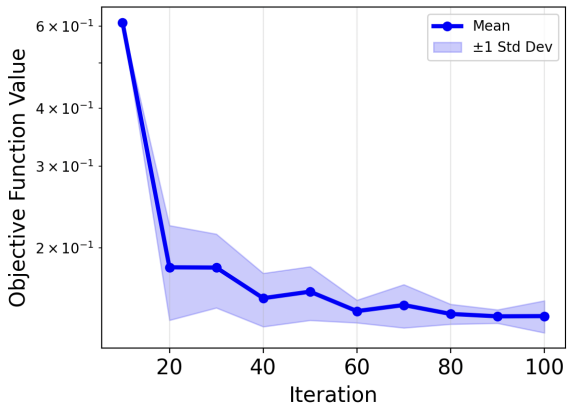
(a) Tensor at index 20000



(b) Tensor at index 40000



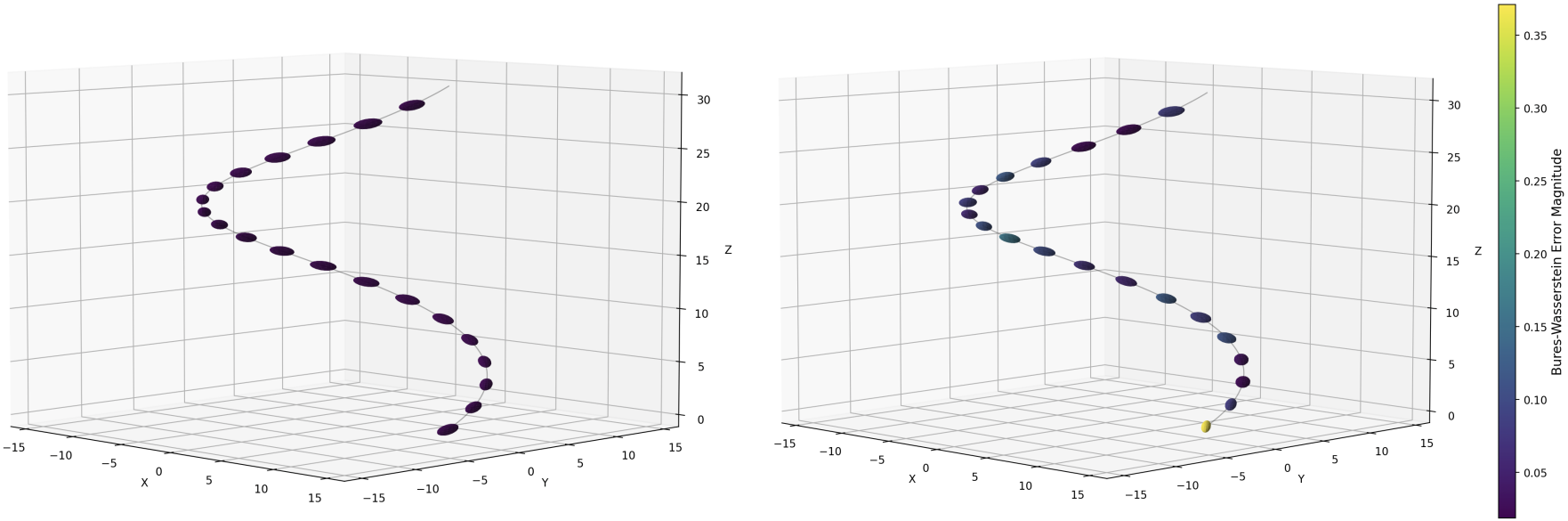
(c) Tensor at index 60000



(d) Tensor at index 80000

Figure: Objective Values over 100 iterations of RSGD from regression process of four tensors

Diffusion Tensor Imaging: Results (2/2)



(a) Ground-truth tensors

(b) Predicted tensors

Figure: Helix visualization for 20 ground-truth tensors (from 100,000 samples) and their prediction from the model

Conclusion and Future Work

Conclusion:

- We propose the **Spectral Dominance of Positive Weights** condition for the existence of minimizer of Problem 3

$$\sum_{i\in\mathcal{I}} \lambda_i^+ \sqrt{\lambda_{\min}(\Sigma_i)} > \sum_{j\in\mathcal{J}} \lambda_j^- \sqrt{\lambda_{\max}(\Sigma_j)},$$

- We further have stronger conditions for the **unique existence** of the minimizer.
- We propose Riemannian Gradient Descent algorithm and its pairwise version for large-scale set-up.

Future Work:

- Incorporate **accelerated, distributed, or adaptive Riemannian methods** to enhance convergence speed and computational efficiency.
- Generalize the framework to **positive semi-definite matrices** and **high-dimensional matrices** (large d).

Reference


Martial Agueh and Guillaume Carlier.
Barycenters in the Wasserstein space.
SIAM Journal on Mathematical Analysis, 43(2):904–924, 2011.


Silvere Bonnabel.
Stochastic gradient descent on Riemannian manifolds.
IEEE Transactions on Automatic Control, 58(9):2217–2229, 2013.


Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra.
Riemannian stochastic recursive gradient algorithm.
In *International conference on machine learning*, pages 2516–2524. PMLR, 2018.


Danielle P Mersch, Alessandro Crespi, and Laurent Keller.
Tracking individuals shows spatial fidelity is a key regulator of ant social organization.
Science, 340(6136):1090–1093, 2013.


Ryan A. Rossi and Nesreen K. Ahmed.