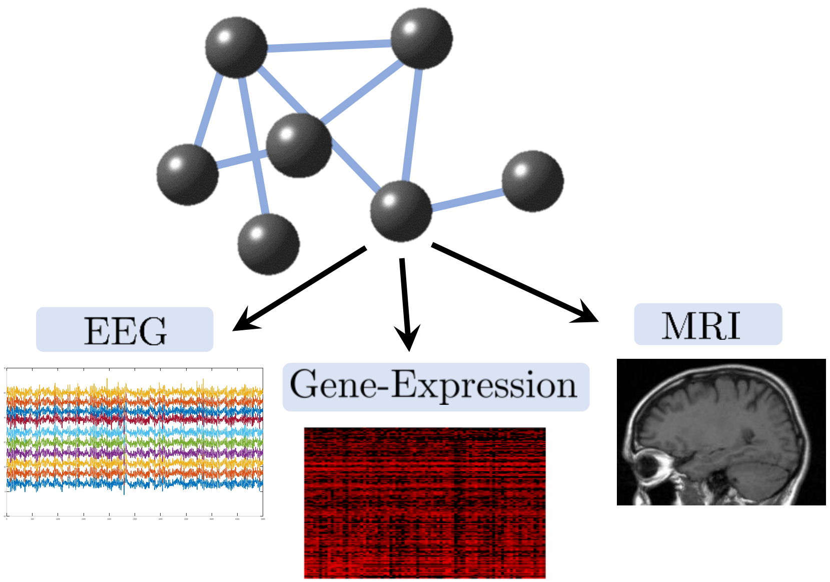
Optimal Transport on Manifolds for Domain Adaptation and Metric Learning

Almog Lahav and Ronen Talmon

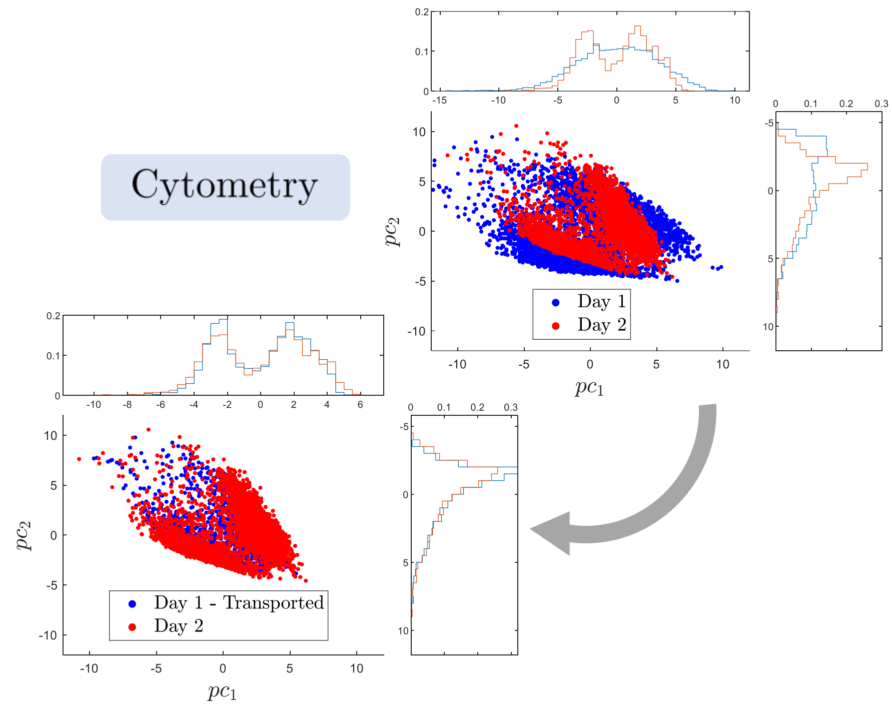
Technion - Israel Institute of Technology

High-dimensional Datasets

How to **compare** them?



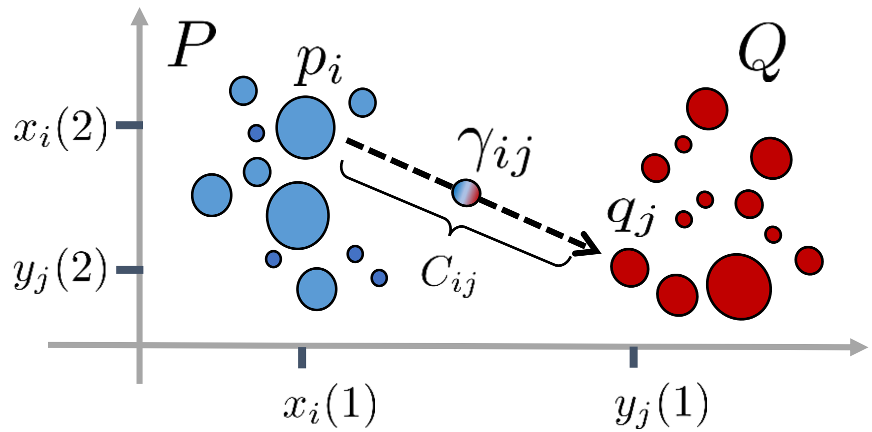
How to **adapt** one to the other?



OT - Kantorovich Problem

- We consider two distibutions: $P = \{(x_i, p_i)\}_{i=1}^m$ and $Q = \{(y_i, q_i)\}_{i=1}^n$
- Kantorovich's optimal **plan**:

$$\underset{\gamma}{\text{minimize}} \quad \langle C, \gamma \rangle_F$$
$$\text{subject to} \quad \gamma \mathbf{1} = p, \qquad \gamma^T \mathbf{1} = q$$
$$\gamma_{ij} \geq 0 \qquad \forall i, j$$



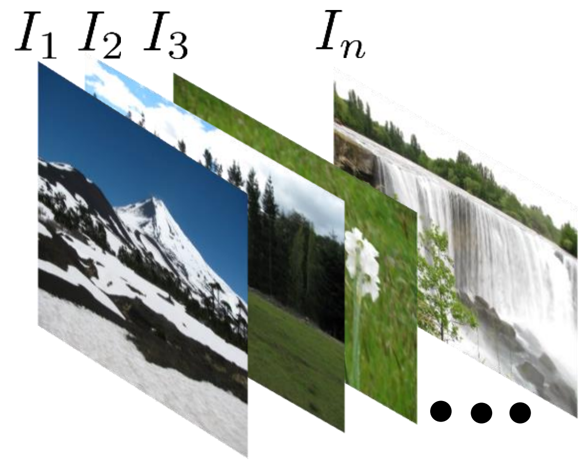
$C_{ij} = d(x_i, y_j)$ is the **ground distance**

γ_{ij} is the mass transported from x_i to y_j

Earth Mover's Distance (EMD)

- OT induces a **metric** [Y. Rubner et al., 2000]

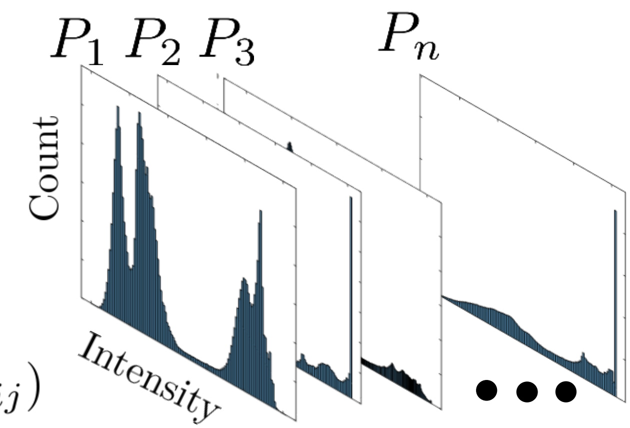
$$
\begin{aligned}
EMD(P_1, P_2) = \quad & \min_{\gamma} \langle C, \gamma \rangle_F \\
\text{subject to} \quad & \gamma \mathbf{1} = p_1, \qquad \gamma^T \mathbf{1} = p_2 \\
& \gamma_{ij} \geq 0 \; \forall i, j
\end{aligned}
$$



- Efficient implementation of approx. EMD:

Sinkhorn distance [M. Cuturi, 2013]

$$
EMD(P_1, P_2) = \quad \min_{\gamma} \langle C, \gamma \rangle_F + \lambda \sum_{ij} \gamma_{ij} log(\gamma_{ij})
$$

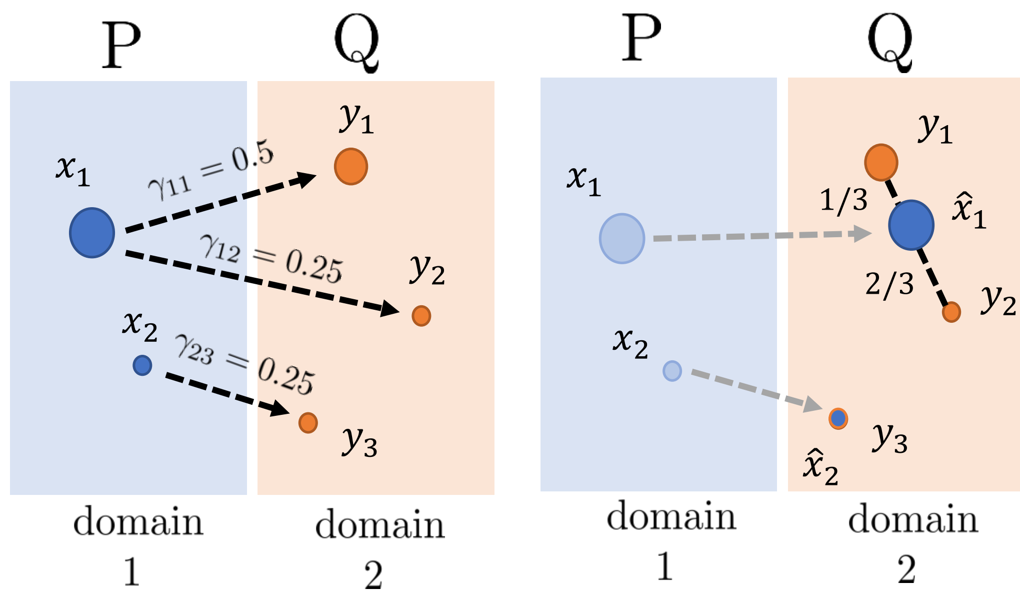


Optimal Transport for Domain Adaptation
[N. Courty, et al. 2017]

- **Test** set $P = \{(x_i, p_i)\}_{i=1}^{m}$ and **training** set $Q = \{(y_j, q_j)\}_{j=1}^{n}$
- To classify P with the classifier trained on Q: $P \xrightarrow{\text{OT}} Q$
- When $c(x_i, y_j) = \|x_i - y_j\|_2^2$:

$$\hat{x}_i = \frac{\sum_j \gamma_{ij} y_j}{\sum_j \gamma_{ij}}$$

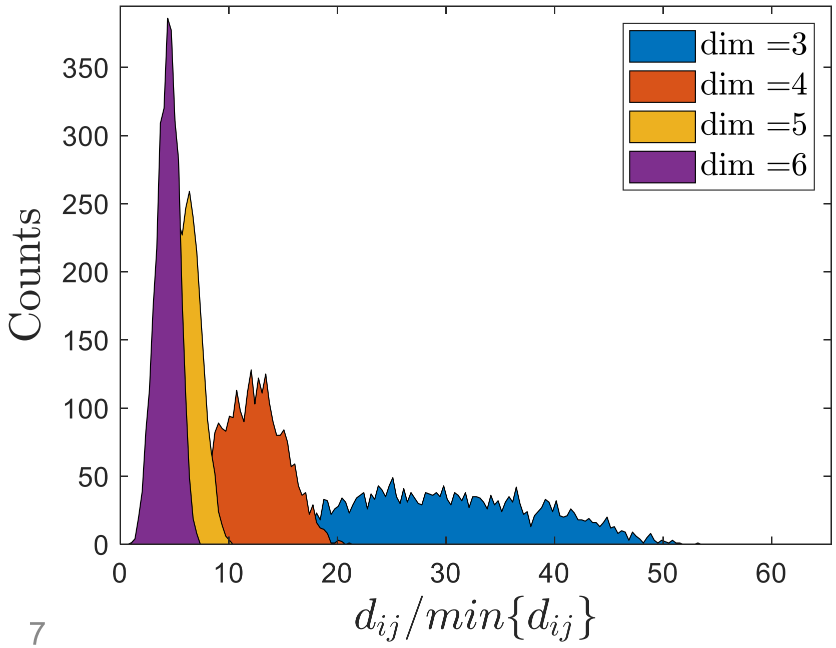
$$\hat{X} = diag(\gamma 1)^{-1} \gamma Y$$



OT in High-dimensional Space

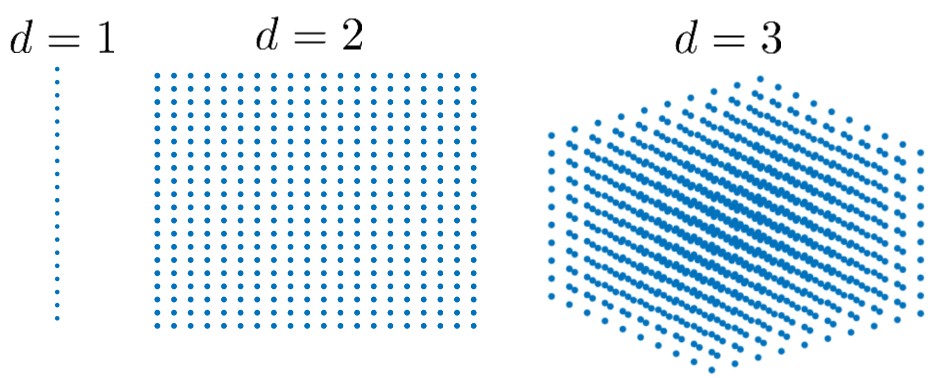
- A common choice: $c_{ij} = \|x_i - y_j\|_2^2$

Often fails to capture the essence of **high-dimensional data**



$$\lim_{d\to\infty} E\left(\frac{dist_{max}(d) - dist_{min}(d)}{dist_{min}(d)}\right) \to 0$$

[Beyer and Goldstein et al., 1999]

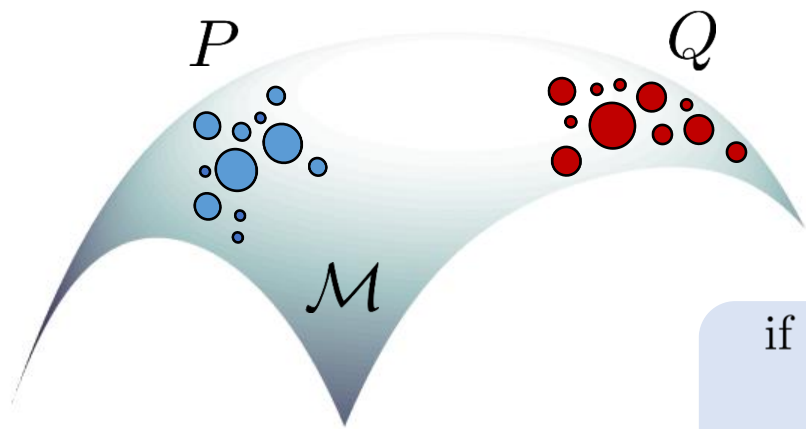


OT in High-dimensional Space

- Common practice: assuming an **intrinsic low-dimensional structure**

- For $P = \{(x_i, p_i)\}_{i=1}^m$ and $Q = \{(y_j, q_j)\}_{j=1}^n$ we assume:

  $\{x_i\}_{i=1}^m$ **and** $\{y_j\}_{j=1}^n$ **lie on a common low-dimensional manifold** $\mathcal{M}$



Examples for $\mathcal{M}$:

Sphere [Z. Su, et al., 2015]

Cone of SPD matrix [O. Yair et al., 2019]

if $\mathcal{M}$ is unknown

$\Downarrow$

Manifold Learning

Manifold Learning

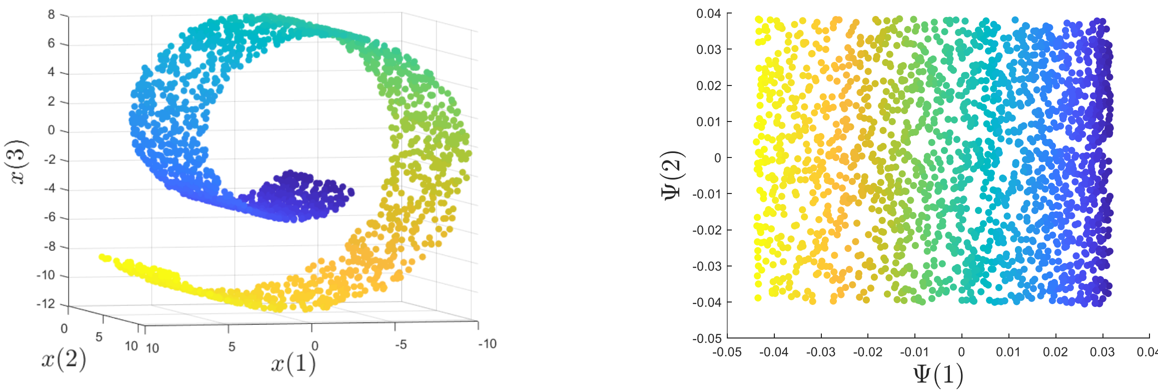
- Dataset: $\{x_i\}_{i=1}^{n} \in \mathcal{M} \subset \mathbb{R}^d$, where $\mathcal{M}$ is unknown
- Finding a map to a **low-dimensional embedded space**:

$$x_i \longrightarrow \Psi\{x_i\} \in \mathbb{R}^p \qquad p < d$$

$\Psi\{x_i\}$ should respect the structure of the manifold

- ISOMAP [J. B. Tenenbaum et al., 2000]
- LLE [S. T. Roweis et al., 2000]
- Laplacian eigenmaps [M. Belkin et al., 2003]

embbeding computed by LTSA [Z. Zhang et al., 2004]



Algorithm 1 Optimal Transport on Manifold

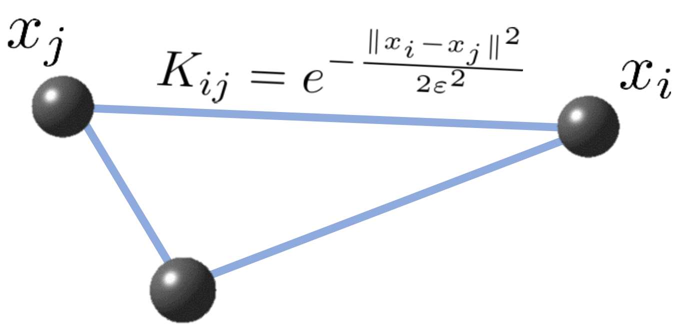
Input: Two distibutions: $P = \{(x_i, p_i)\}_{i=1}^m$ and $Q = \{(y_i, q_i)\}_{i=1}^n$

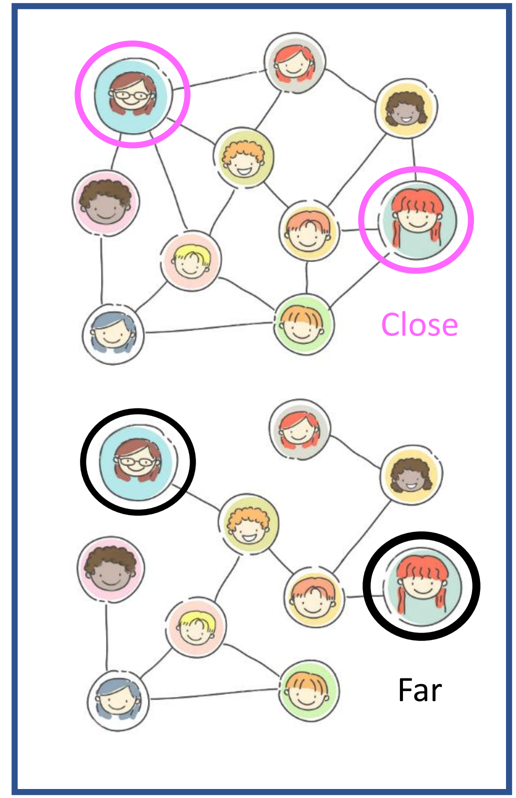
Output: Optimal plan $\gamma^\star$

1. Apply manifold learning algorithm to $S_{x \cup y} = \{x_i\}_{i=1}^m \cup \{y_j\}_{j=1}^n$
2. Use the obtained embedding $\Psi\{s_i\} \forall s_i \in S_{x \cup y}$ to compute:
$$c_{ij} = d\left(\Psi\left\{x_i\right\}, \Psi\left\{x_j\right\}\right)$$
3. Solve:
$$\begin{aligned} \gamma^\star = \quad & \underset{\gamma}{\operatorname{argmin}} \langle C, \gamma \rangle_F \\ \text{subject to} \quad & \gamma \mathbf{1} = p, \qquad \gamma^T \mathbf{1} = q \\ & \gamma_{ij} \geq 0 \ \forall i, j \end{aligned}$$

Diffusion Distance
[R.R. Coifman and S. Lafon, 2004]

- Given a set of samples $\{x_i\}_{i=1}^{n}$
- We define a markov chain on a graph:





transition matrix: $P = (diag\{K\mathbf{1}\})^{-1}K$

transition probability of t steps: $p_t(x_i, x_j) = P_{ij}^t$

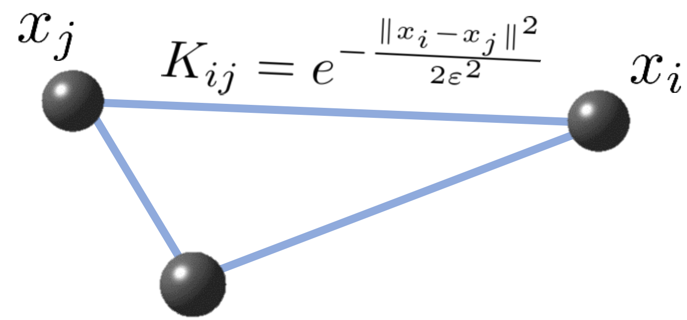
diffusion distance:

$$D_t(x_i, x_j) = \sum_{l=1}^{n} (p_t(x_i, x_l) - p_t(x_j, x_l))^2 / \varphi_0(l)$$

Diffusion Ground Distance

- Diffusion maps: $x_i \longrightarrow \Psi_t\{x_i\} = (\lambda_1^t \psi_1(x_i), \lambda_2^t \psi_2(x_i), ..., \lambda_d^t \psi_d(x_i))$

  where ψ_l and λ_l satisfy: $P\psi_l = \lambda_l \psi_l$



- $D_t(x_i, x_j) \approx \|\Psi_t(x_i) - \Psi_t(x_j)\|^2$

- Ground diffusion distance: $c_{ij} = \|\Psi_t(x_i) - \Psi_t(x_j)\|^2$

✓ low dimension

✓ local structures to global metric

✓ robustness to noise

✓ use of exisiting OT results, e.g.:

- **barycentric mapping**

  $\Psi_t\{\hat{X}\} = diag(\gamma \mathbf{1})^{-1} \gamma \Psi_t\{Y\}$

- **analitic OT solution**

Diffusion Ground Distance

- Two distibutions: $P = \{(x_i, p_i)\}_{i=1}^{m}$ and $Q = \{(y_i, q_i)\}_{i=1}^{n}$
- Theorem [A. Takatsu, 2011]:

If
1. $\mathbf{p} \sim \mathcal{N}(\mu_p, \sigma^2 I)$, $\mathbf{q} \sim \mathcal{N}(\mu_q, \sigma^2 I)$
2. $c_{ij} = \|x_i - y_j\|_2^2$

then the OT has a closed form and:

$$EMD(P, Q) = \|\mu_p - \mu_q\|_2$$

- If $c_{ij} = D_t(x_i, y_j)$:

$$EMD_{\mathcal{M}}(P, Q) \approx \|\Psi_t\{\mu_p\} - \Psi_t\{\mu_q\}\|_2$$

Diffusion Ground Distance

- Taylor series arround $x = y$:

$$D_t(x, y) = \sum_{n=0}^{\infty} \frac{D_t^{(n)}(x,y)|_{x=y}}{n!}(x-y)^n$$

- Assumimg a unfiorm density on $\mathcal{M}$:

$$D_t(x, y) = c\varepsilon(x-y)^2 + \mathcal{O}(\varepsilon^3(x-y)^4)$$

- For small ε:

$$D_t(x, y) \approx c\varepsilon(x-y)^2$$

$$\Downarrow$$

if $\sigma << 1/\varepsilon$, $\mathbf{p}$ and $\mathbf{q}$ are approx. Gaussians in the embedded space

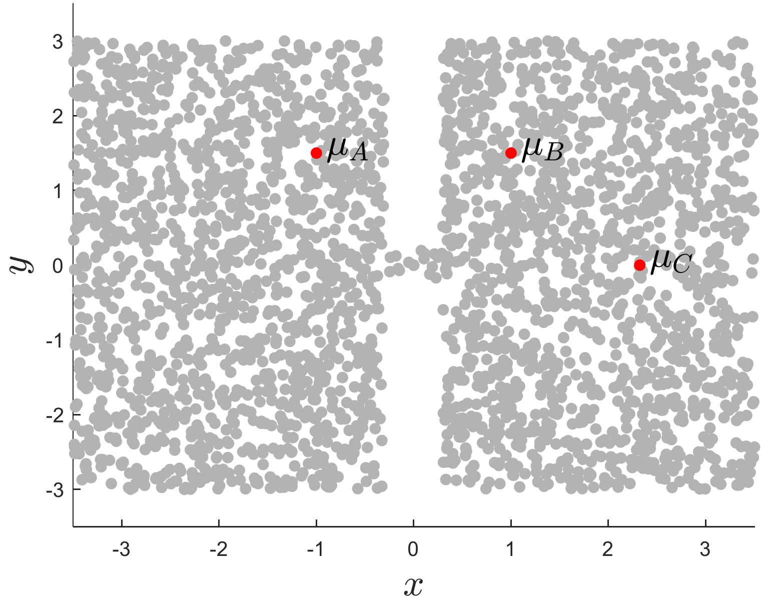
- Recall $D_t(x, y) = \|\Psi_t(x) - \Psi_t(y)\|_2^2 \quad \forall x, y$:

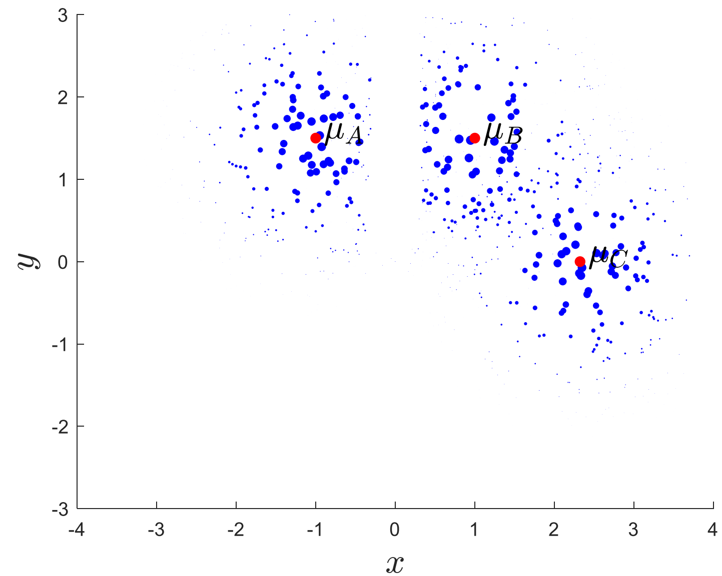
$$EMD_{\mathcal{M}}(P, Q) \approx \|\Psi_t\{\mu_p\} - \Psi_t\{\mu_q\}\|_2^2$$

Diffusion Ground Distance - Example

- Consider 3 distributions:

$$p_A \sim \mathcal{N}(\mu_A, \sigma^2 I), \quad p_B \sim \mathcal{N}(\mu_B, \sigma^2 I), \quad p_C \sim \mathcal{N}(\mu_C, \sigma^2 I)$$

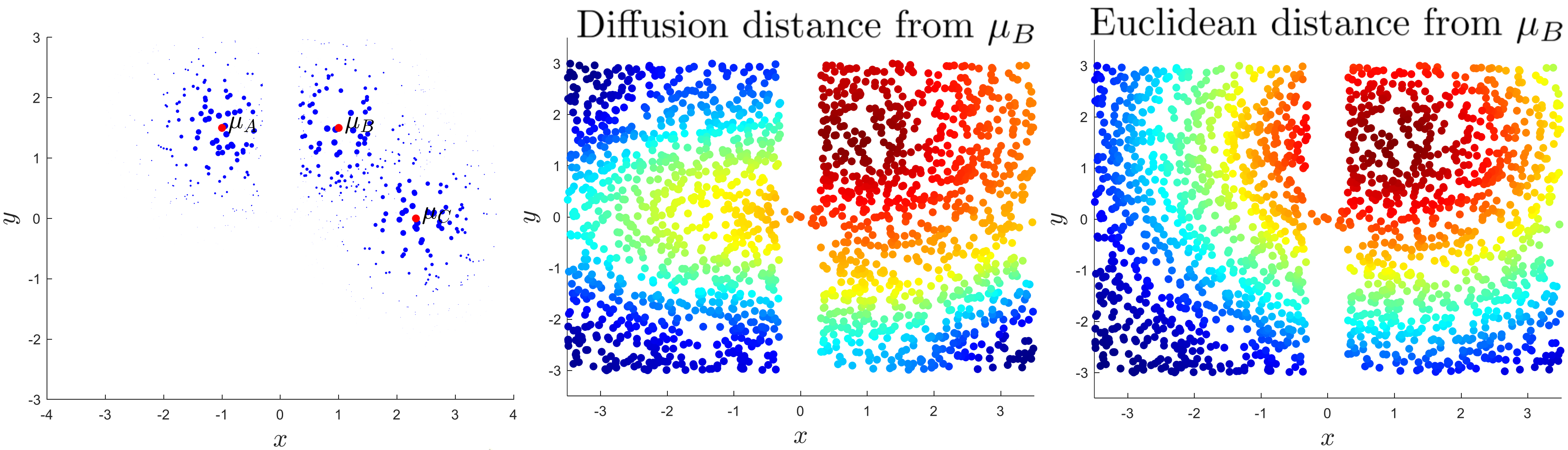




Diffusion Ground Distance - Example

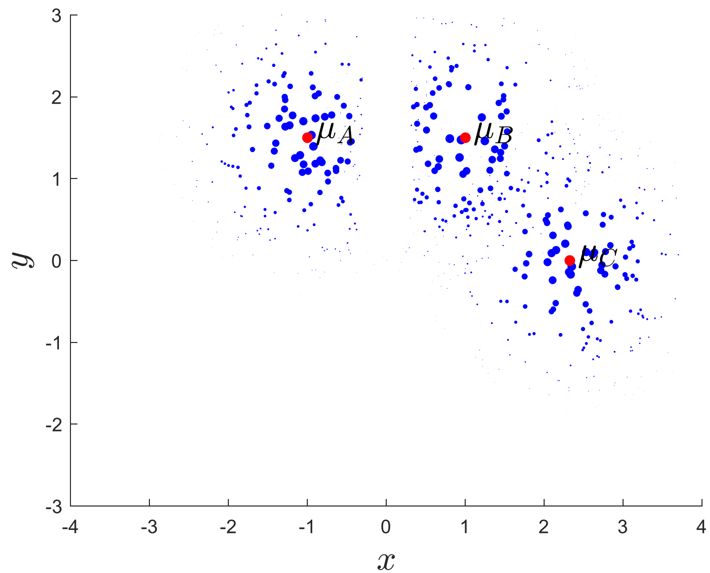
- Consider 3 distributions:

$$p_A \sim \mathcal{N}(\mu_A, \sigma^2 I), \quad p_B \sim \mathcal{N}(\mu_B, \sigma^2 I), \quad p_C \sim \mathcal{N}(\mu_C, \sigma^2 I)$$

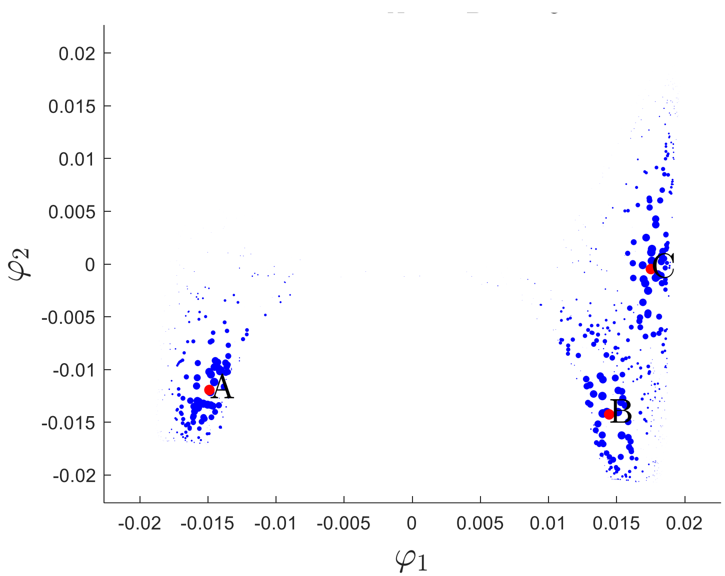


Diffusion Ground Distance - Example

Distribution in original space



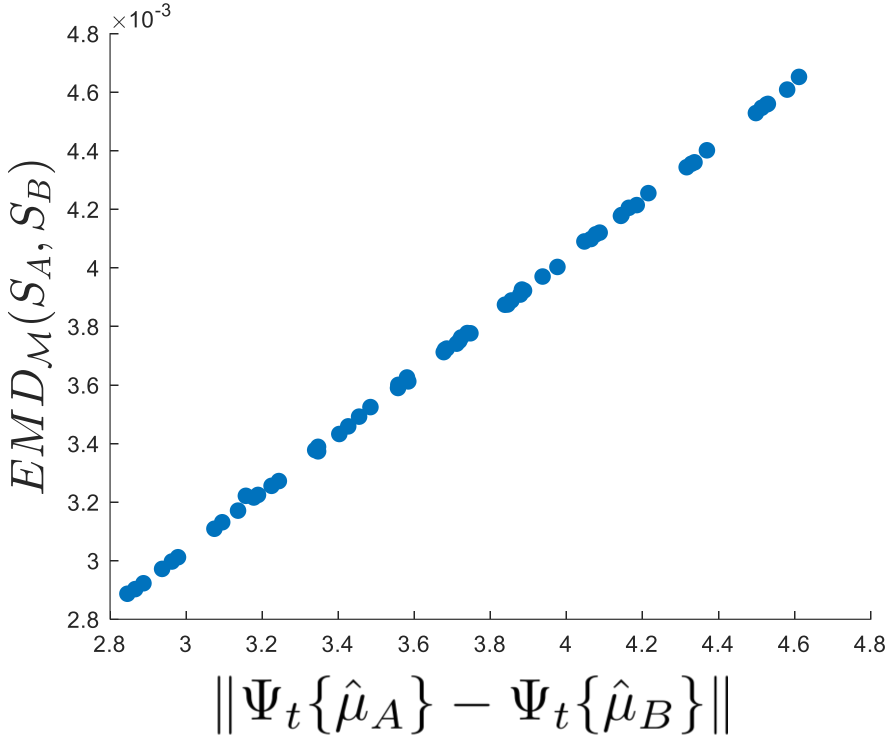
Distribution in diffusion space



	$EMD(S_A, S_B)$	$EMD(S_B, S_C)$
Euclidean GD	4.70	3.87
Diffusion GD	$17 \cdot 10^{-4}$	$5 \cdot 10^{-4}$

Diffusion Ground Distance - Example

Simulation

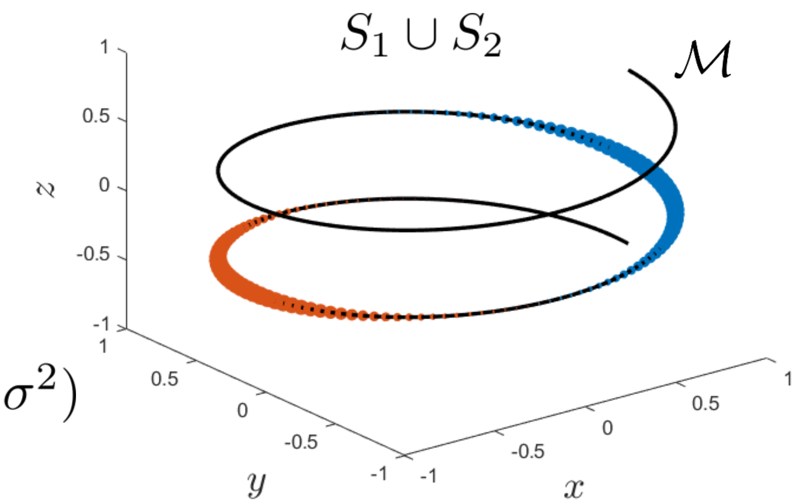


Theory

Diffusion Ground Distance - 1D Manifold

- Consider a dataset $S_k = \left\{ (x_i, y_i, z_i), p_i^{(k)} \right\}_{i=1}^{20}$:

$$x_i = r \cdot cos(t_i)$$
$$y_i = r \cdot sin(t_i)$$
$$z_i = t_i/10$$



- Weights $\mathbf{p^{(k)}}$ are samples of $\mathcal{N}(\mu_k, \sigma^2)$

- The heat diffusion generator Δ has eigenfunctions on the manifold $\mathcal{M}$:

$$\Delta \psi_l = -\lambda_l \psi_l$$

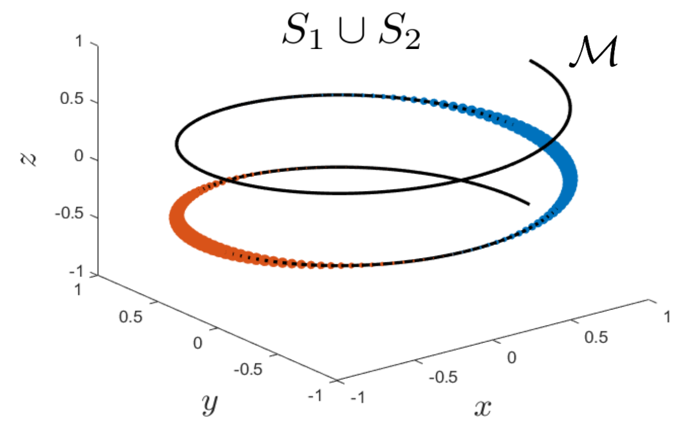
- For a curve of length L:

$$\psi_1 \left((x_i, y_i, z_i) \right) = cos \left(\frac{\pi}{L} t_i \right)$$

Diffusion Ground Distance - 1D Manifold

- The diffusion ground distance:

$$c_{ij} = \|\psi_1(t_i) - \psi_1(t_j)\|_2^2$$
$$= \left| cos\left(\frac{\pi}{L}t_i\right) - cos\left(\frac{\pi}{L}t_j\right)\right|^2$$

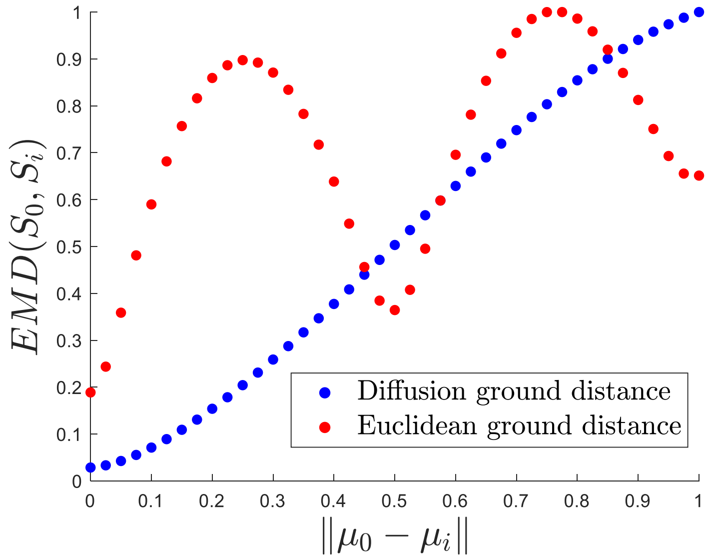


- If $|t_i - t_j|_2 > |t_m - t_n|_2$ then $c_{ij} > c_{mn}$

  $\Rightarrow c_{ij}$ respects $\mathcal{M}$

- For $\mathbf{p^{(k)}} \sim \mathcal{N}(\mu_k, \sigma^2)$:

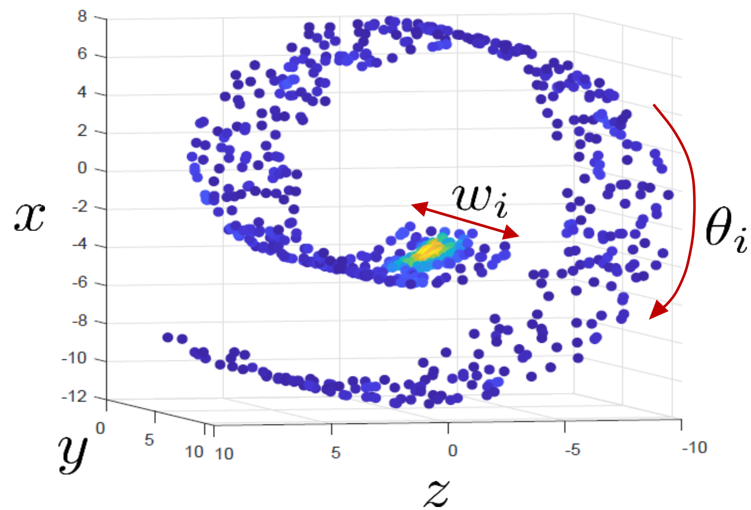
$$EMD_{\mathcal{M}}(S_k, S_l) \approx |cos(\mu_k) - cos(\mu_l)|$$



OT on 2D Manifold

- Consider a dataset $S_k = \{(x_i, y_i, z_i)\}_{i=1}^{700}$
- Parametrization:

$$x_i = t_i \cdot cos(t_i)$$
$$y_i = h_i$$
$$z_i = t_i \cdot sin(t_i)$$



- $N_u = 600$ realizations of uniform variables:

$$\mathbf{t_u} \sim U[1.25\pi, 3.75\pi] \qquad \mathbf{h_u} \sim U[0, 11]$$

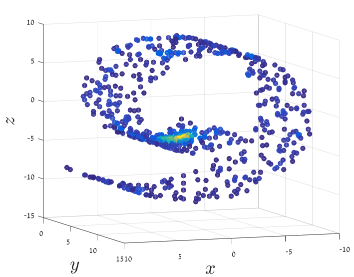
- $N_g = 100$ realizations of a Gaussian variable:

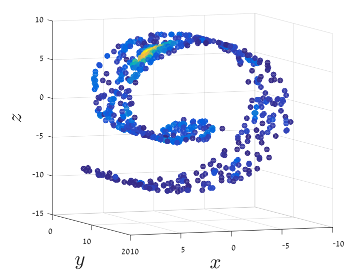
$$\mathbf{t_g} \sim \mathcal{N}(\theta_i, \sigma^2) \qquad \mathbf{h_g} \sim \mathcal{N}(w_i, 1)$$

OT on 2D Manifold

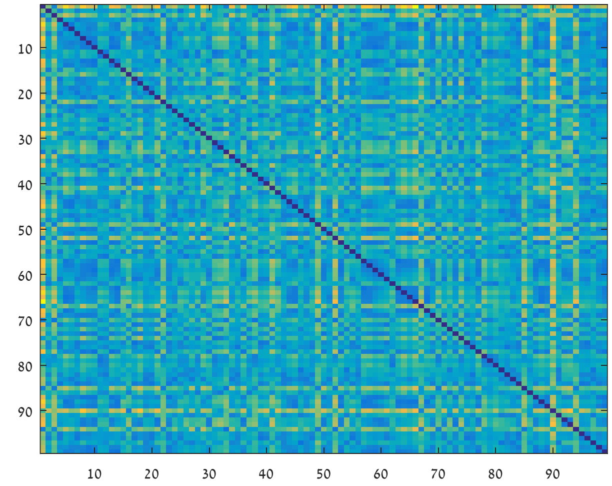
- Consider a set $\{S_k\}_{k=1}^{99}$
- We learn the manifold with LTSA [Z. Zhang et al., 2004]
- Compute the distance:

$$D_{k,l} = EMD_{\mathcal{M}}(S_k, S_l)\ \forall k, l$$



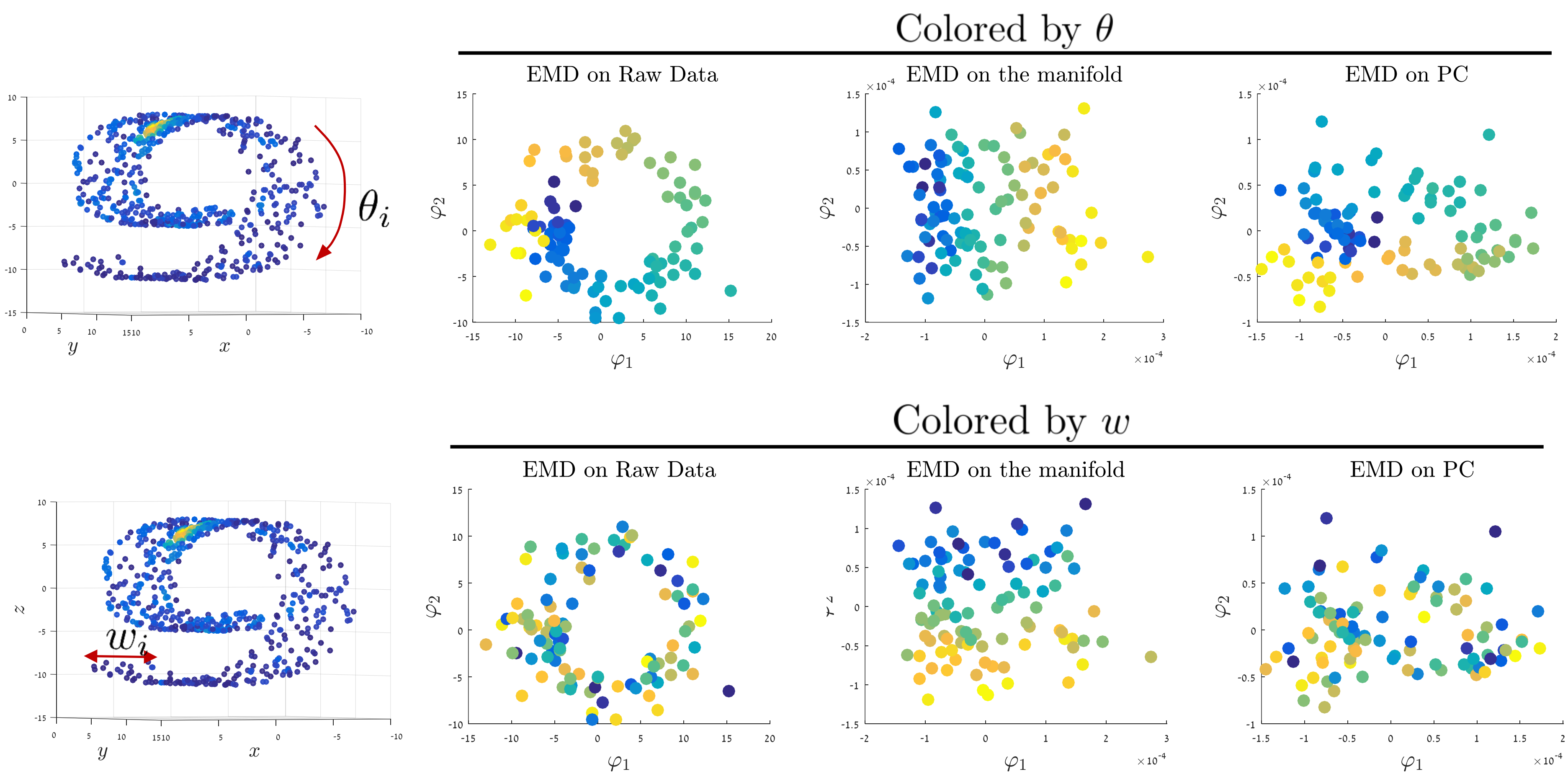


D



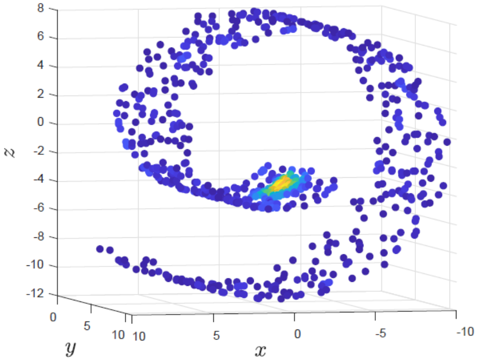
- Using MDS we visualize the obtained distances in $\mathbb{R}^2$

Low-dimensional representation obtained by MDS

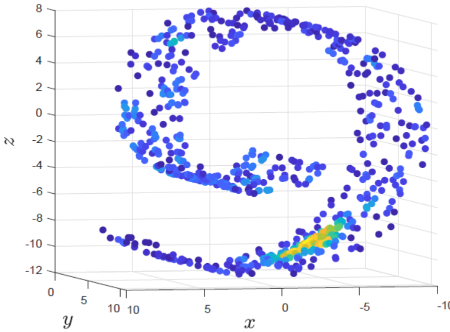


OT on 2D Manifold

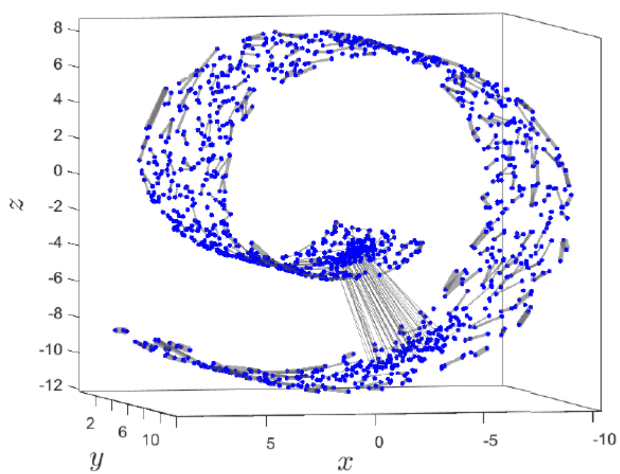
S_1



S_2



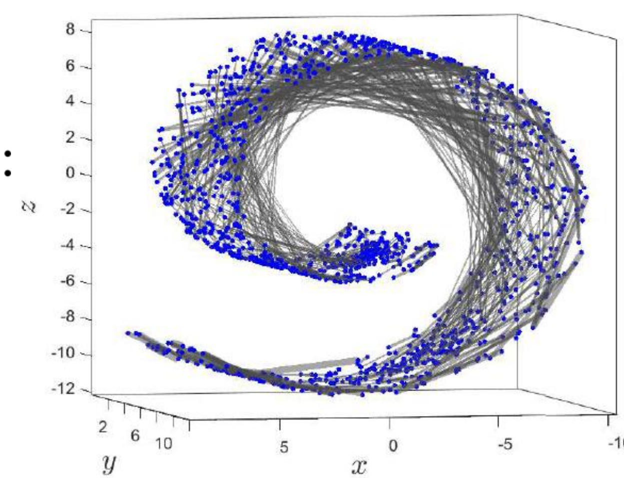
OT - Euclidean ground distance



j_{th} edge width:

$$\max_k \{\gamma^\star_{kj}\}$$

OT - on the manifold



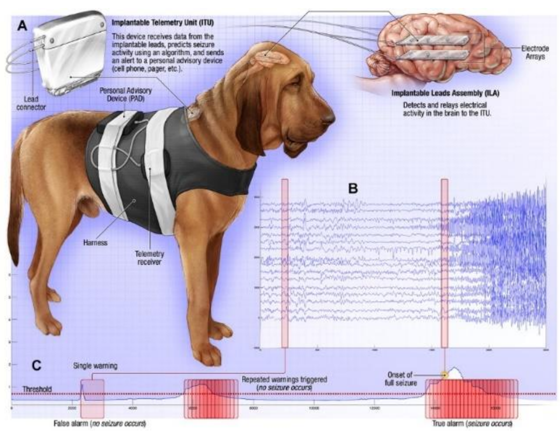
Application to Real Data

- EEG recordings for Epilepsy seizure detection acquired from dogs

The goal

Classify time series of 10 minutes

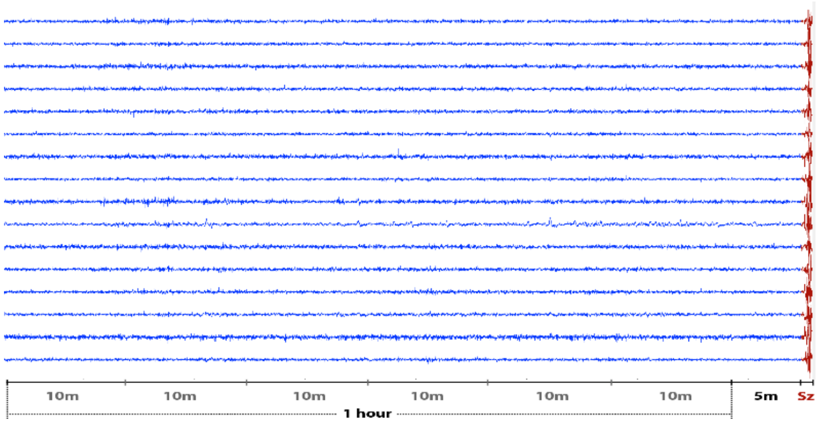
between seizures | prior to seizure



"Forecasting Seizures in Dogs with Naturally Occurring Epilepsy"
[Howbert JJ & Patterson EE et al., 2014]

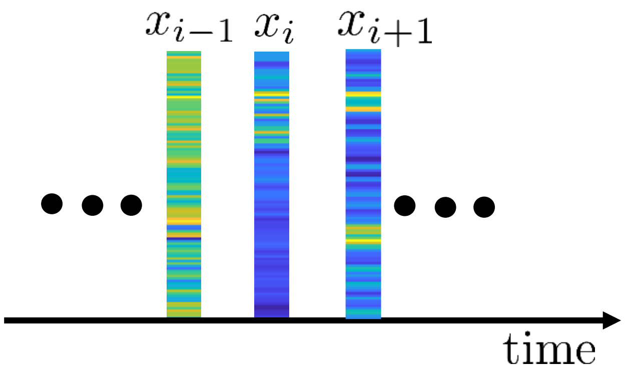
Application to Real Data

- "American Epilepsy Society Seizure Prediction Challenge" (Kaggle)
- Data: EEG recordings from a dog
  - 15-16 electrodes
  - 12 Segments of 10 minutes
  - Labels: interictal/preictal



- Features: scattering transform [Mallat S., 2012]
- Dataset (segment): $S = \{(x_i, p_i)\}_{i=1}^{168}$

$$x_i \in \mathbb{R}^{108}, \quad \mathbf{p} \text{ uniform}$$



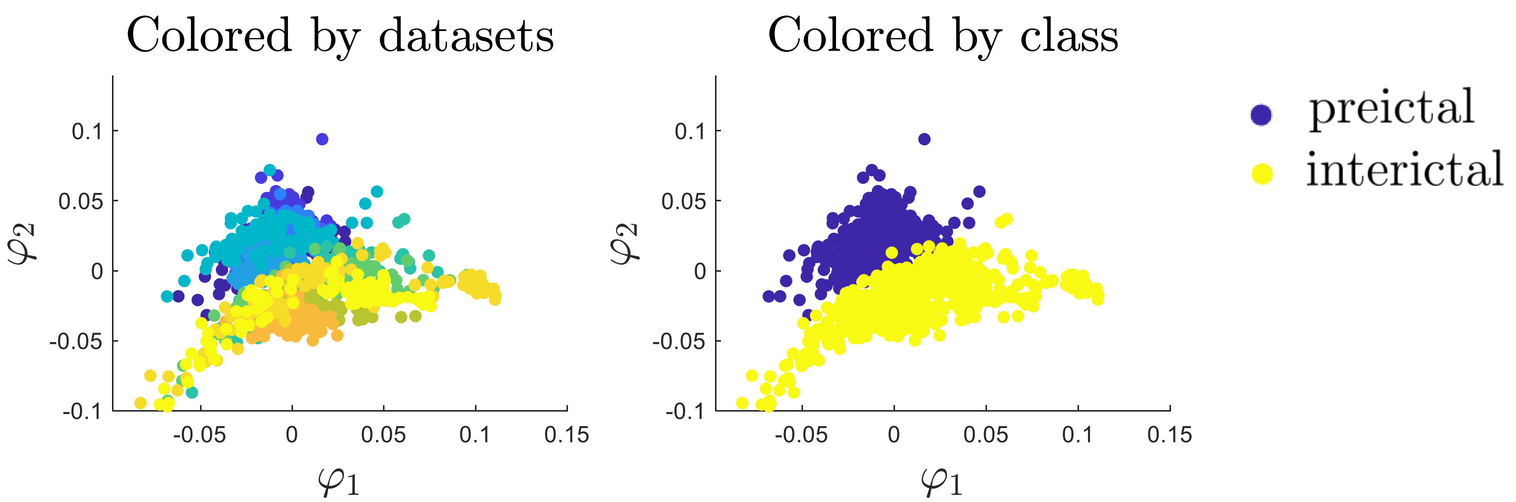
Application to Real Data

- We compute distances between segments:

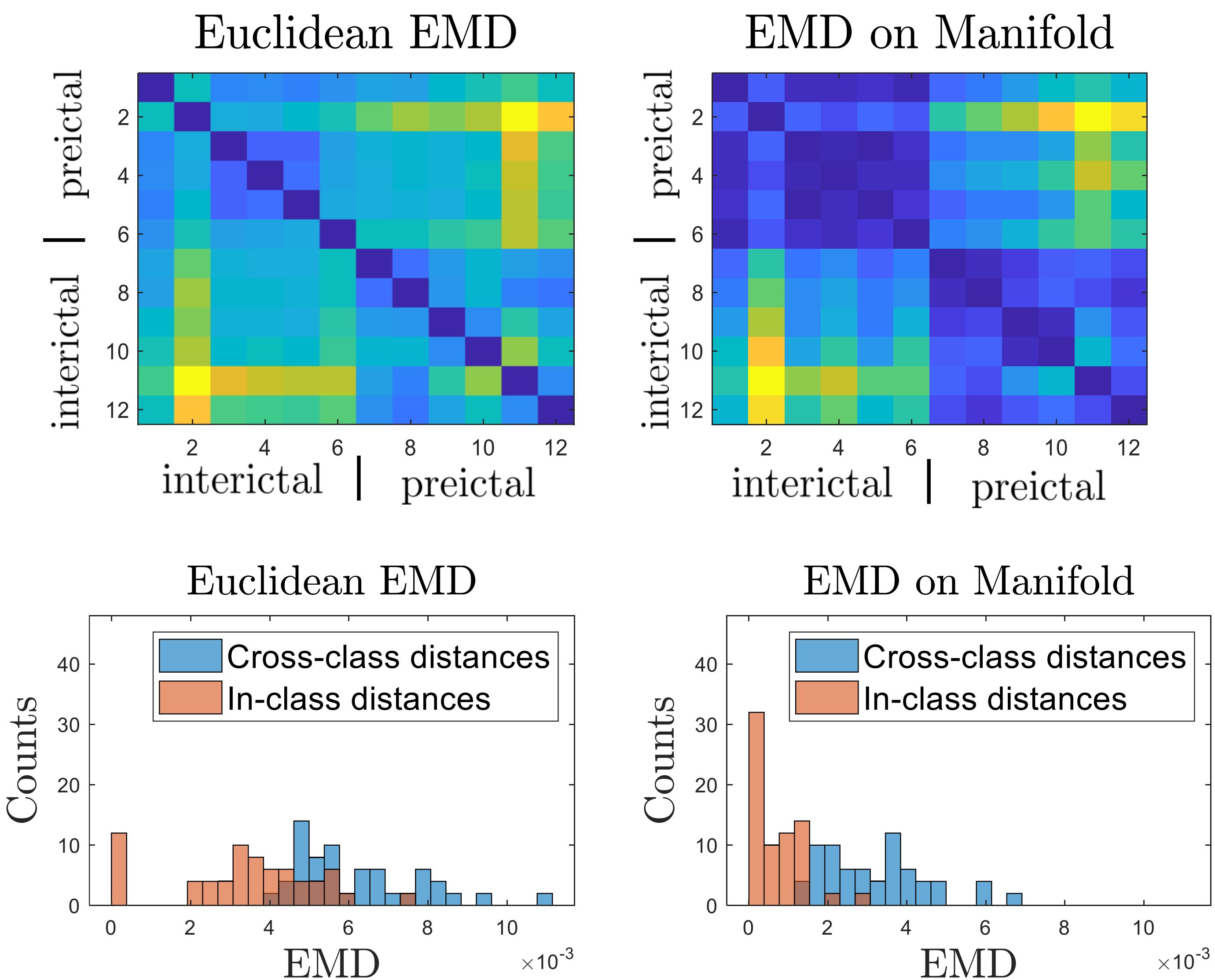
$$EMD(S_i, S_j) \quad 1 < i, j < 12$$

- We examine the **Euclidean** and the **diffusion** ground distance

Segments in the embedded diffusion space

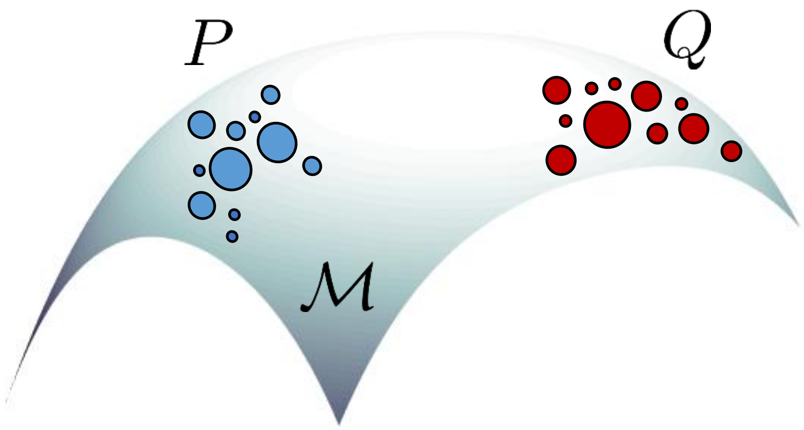


Results


Euclidean EMD
EMD on Manifold
preictal
interictal
preictal
interictal
2
4
6
8
10
12
interictal
preictal
interictal
preictal
Euclidean EMD
EMD on Manifold
Cross-class distances
In-class distances
Cross-class distances
In-class distances
Counts
Counts
0
10
20
30
40
0
2
4
6
8
10
EMD
EMD
×10^-3
×10^-3

Summary

- We presented a framework for computing meaningful distances and adapting high-dimensional datasets
- We assume that the data live on a low-dimensional **unknown** manifold
- We propose a solution that learns the manifold and solves OT on the learned manifold
- We showed both analytic and experimantal results



Thank You