https://arxiv.org/pdf/2209.06975.pdf

Paper review

Wasserstein K-means for clustering probability distributions (NeurIPS 2022) - 2

Presentation: **Jeiyoon Park**
6th Generation, TAVE

TAVE Research

Outline

1. Background
2. Method
3. Experiments
4. Discussion

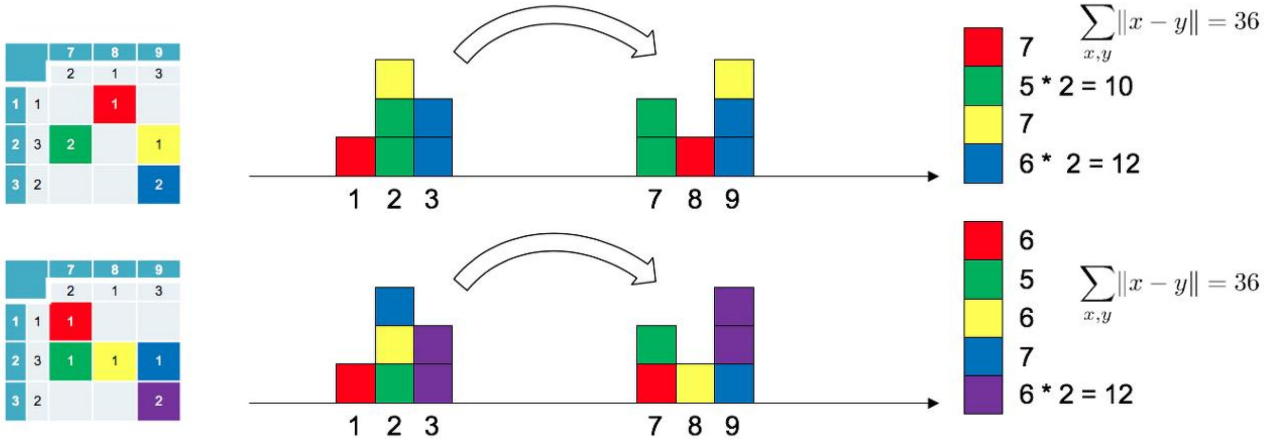
Outline

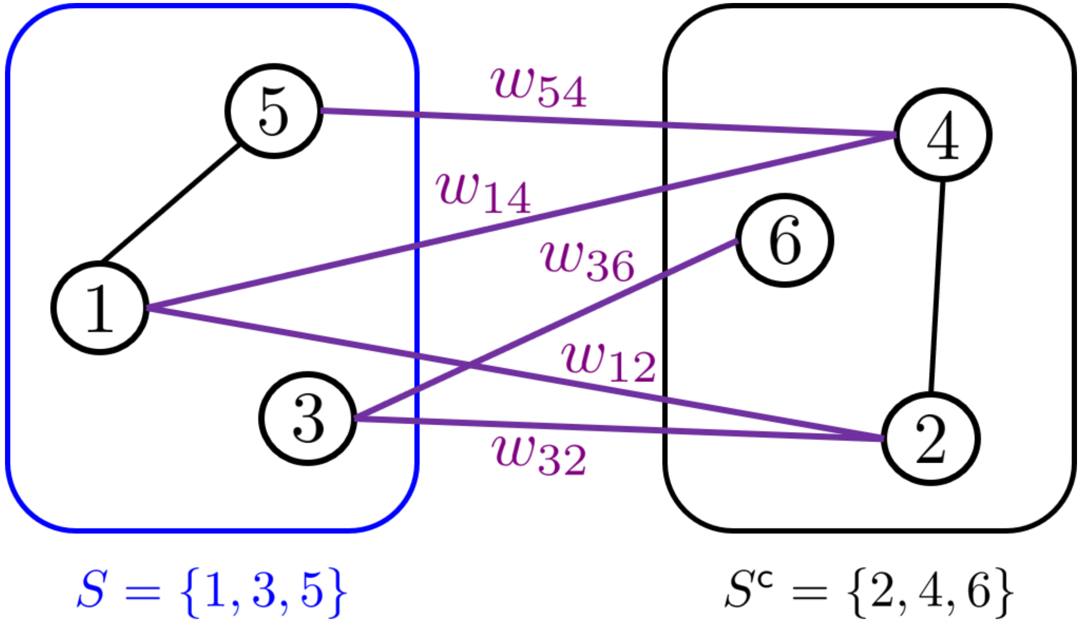
Recap

- Summary

  1) Detour: Wasserstein Distance

  2) Detour: SDP Relaxation





Plan for Today

- Summary

  1) Centroid-based Wasserstein K-means
  2) Three pitfalls of 1)
  3) Distance-based Wasserstein K-means
  4) Experiments: Real-data applications
  5) Discussion

Method

1. Detour: K-means clustering

1) K-means clustering: Setup K number of centroids and cluster data points by the distance from the points to the nearest centroid (or barycenter)

2) We are familiar to this notation:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

, where r_{nk} stands for the assignment of data points to clusters and μ_k is the location of centroids

3) Iterative optimization (a.k.a., Expectation and Maximization)

Method

1. Detour: K-means clustering

4) Actually, there are two kinds of K-means clustering: Centroid-based formulation and Distance-based formulation

- Centroid-based formulation:

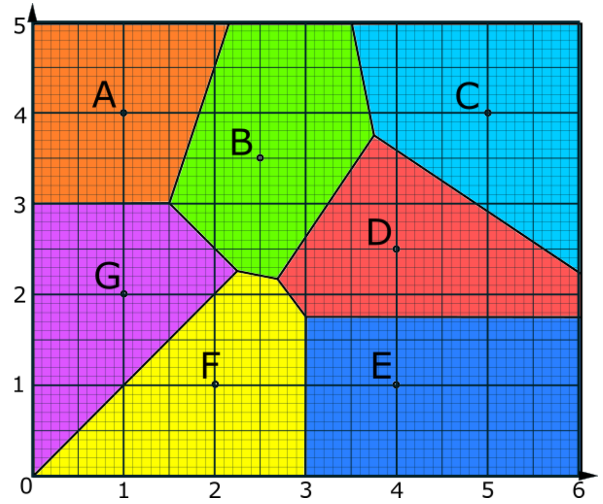
$$\min_{\beta_1,\ldots,\beta_K \in \mathbb{R}^d} \sum_{i=1}^{n} \min_{k\in[K]} \|X_i - \beta_k\|_2^2 = \min_{G_1,\ldots,G_K} \left\{ \sum_{k=1}^{K} \sum_{i\in G_k} \|X_i - \bar{X}_k\|_2^2 : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

- Assign each data point (Expectation)

$$G_k^{(t)} = \left\{ i \in [n] : \|X_i - \beta_k^{(t)}\|_2 \leqslant \|X_i - \beta_j^{(t)}\|_2,\ \forall j \in [K] \right\}$$

- Update the centroid for each cluster (Maximization)

$$\beta_k^{(t+1)} = \frac{1}{|G_k^{(t)}|} \sum_{i\in G_k^{(t)}} X_i$$



Method

1. Detour: K-means clustering

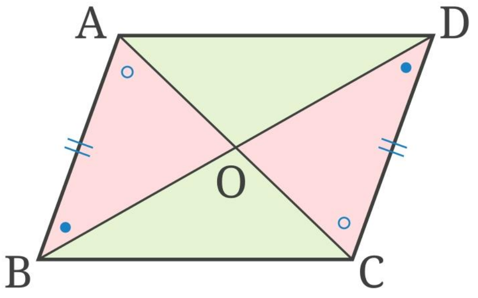
4) Actually, there are two kinds of K-means clustering: Centroid-based formulation and Distance-based formulation

- Distance-based formulation:

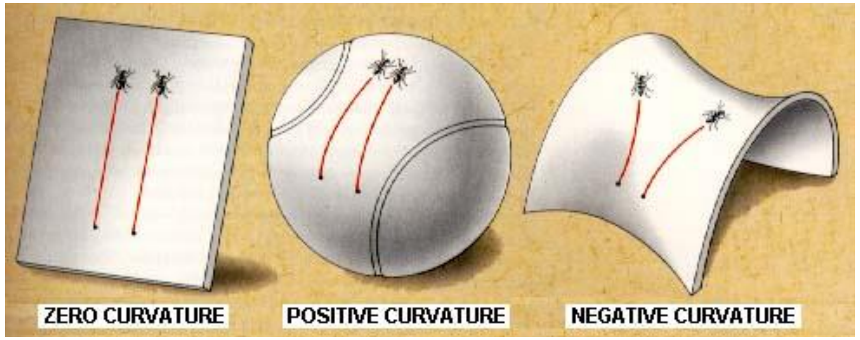
$$\min_{G_1,\dots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \|X_i - X_j\|_2^2 : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

- Note that both formulation yield the same partition (by parallelogram law):

$$\boxed{\sum_{i,j=1}^{n} \|X_i - X_j\|_2^2 = 2n \sum_{i=1}^{n} \|X_i - \bar{X}\|_2^2, \quad \text{with} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad X_i \in \mathbb{R}^p}$$



Method

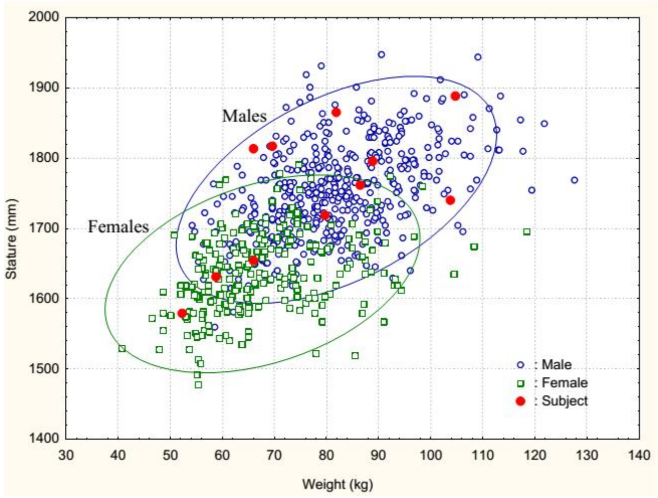


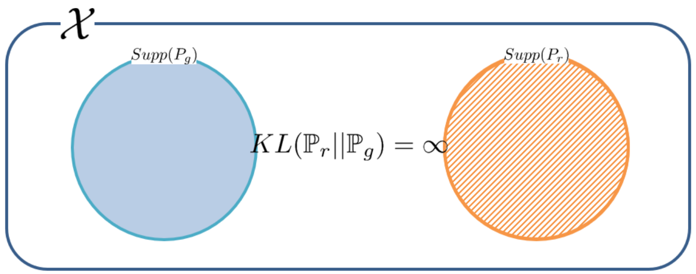
1. Detour: K-means clustering

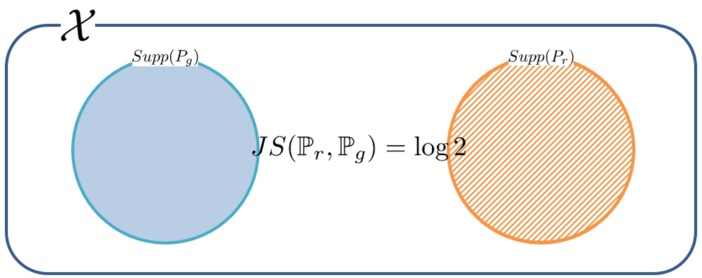
5) K-means clustering for Euclidean space

- It may not be well suited to analyze some data (e.g., ellipse-shaped dataset)
- This would lose important geometric information
- K-means clustering is an NP-hard optimization problem even in two dimensions

→ K-means clustering using different metric space







Method

2. Wasserstein K-means clustering

1) Why this paper?

- Authors provide evidence for pitfalls (irregularity and non-robustness) of barycenter-based Wasserstein K-means

- Authors generalize the distance-based formulation of K-means to the Wasserstein space

- Authors establish the exact recovery property of its SDP relaxation for clustering Gaussian measures

Method

2. Wasserstein K-means clustering

2) Clustering based on barycenters

- 2-Wasserstein distance between two distributions μ and ν:

$$W_2^2(\mu,\nu) := \min_{\gamma} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \right\}$$

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \prod(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||]$$

- Assign each probability measure μ to nearest centroid in the Wasserstein geometry:

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : \|X_i - \beta_k^{(t)}\|_2 \leqslant \|X_i - \beta_j^{(t)}\|_2, \ \forall j \in [K] \right\}$$

- Then update the centroid for each cluster:

$$\nu_k^{(t+1)} = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

$$\beta_k^{(t+1)} = \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} X_i$$

$$W_2^2(\mu,\nu) := \min_{\gamma} \Big\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \Big\}$$

$$G_k^{(t)} = \Big\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \Big\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

Method

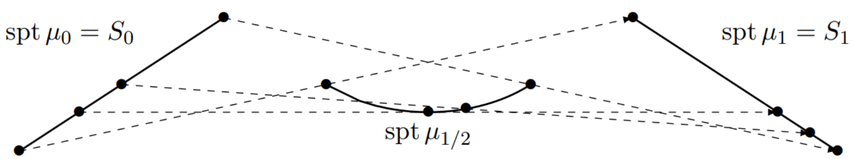


Figure 1: The support of $\mu_{1/2}$ when μ_0 and μ_1 have linear densities on the segments S_0 and S_1.

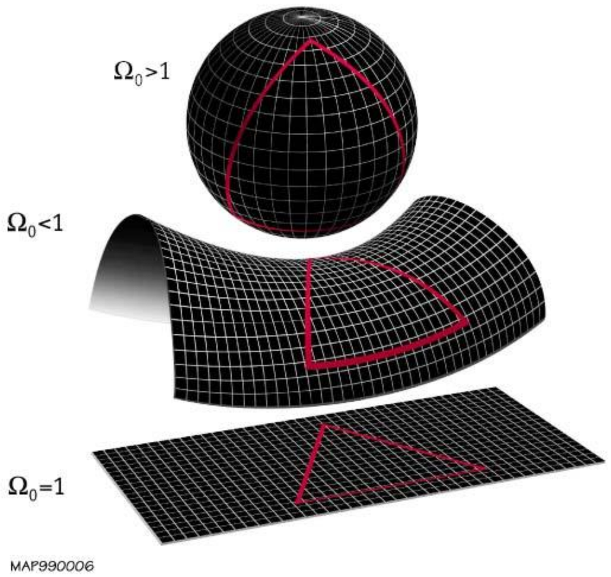
2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: Irregularity and Non-robustness

- ***Example 1***: Irregularity of Wasserstein barycenters

- Wasserstein barycenter has much less regularity than the sample mean in the Euclidean space *(Santambrogio and Wang., 2016)*

- **Lemma 1.** *Given two smooth and positive densities (i.e., $lim\rho_n > 0$) ρ_0, ρ_1 on two compact sets K_0, K_0, respectively, the support of the measure ρ_t obtained as the geodesic interpolant of ρ_0 and ρ_1 in the Wasserstein space $\mathbb{W}_2(\mathbb{R}^d)$ is not necessarily convex.*



$$W_2^2(\mu,\nu) := \min_\gamma \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$\nu_k^{(t+1)} = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- [Detour 1] Notations (Wasserstein distance and Transport plan γ)

$$W_2^2(\mu,\nu) = min\left\{\int_{\Omega\times\Omega} ||x-y||^2 \, d\gamma : \gamma \in \Pi(\mu,\nu)\right\}$$

$$\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(\Omega\times\Omega) : (\pi_x)_{\#\gamma} = \mu, \left(\pi_y\right)_{\#\gamma} = \nu, \}$$

,where $\Omega \subset R^d$ denotes a domain (compact and convex), $\pi_x(x,y) \coloneqq x$ and $\pi_y(x,y) \coloneqq y$ are the standard projections on the two factors of $\Omega \times \Omega$

$$W_2^2(\mu,\nu) := \min_{\gamma} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

Method

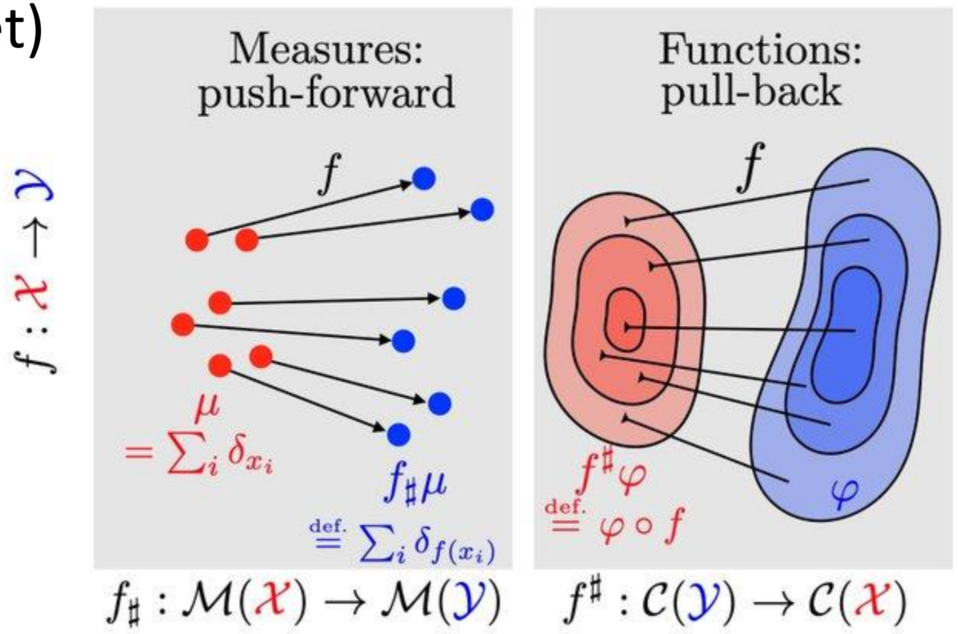
2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- [Detour 2] Pushforward: It is obtained by transferring a measure from one measurable space (i.e., Borel Set) to another using a measurable function

$$\boxed{\gamma \xrightarrow{(\pi_x)} (\pi_x)_{\#\gamma} = \mu}$$

$$\boxed{\gamma \xrightarrow{(\pi_y)} (\pi_y)_{\#\gamma} = \nu}$$



Remark: $f^\sharp$ and $f_\sharp$ are adjoints

$$\int_{\mathcal{Y}} \varphi \, \mathrm{d}(f_\sharp \mu) = \int_{\mathcal{X}} (f^\sharp \varphi) \, \mathrm{d}\mu$$

$$W_2^2(\mu,\nu) := \min_{\gamma} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

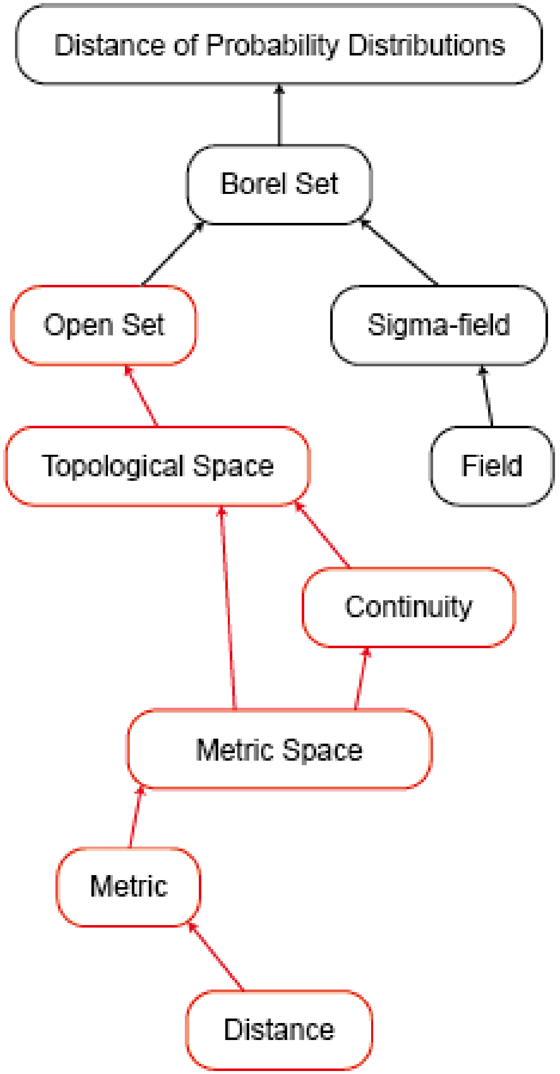
Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- [Detour 3] Metric space: Note that “Metric space” is distance space where it can be “measurable” using a specific distance (metric)

- *Distance → Metric → Metric space → Topological space*
- Distance: (w.r.t., elements) Points
- Metric: (w.r.t., elements) Distribution, Set
- Metric space: Measurable space using a specific metric
- Topological space: Metric space $\sqcup$ Metric space^c



$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i \in [n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}), \quad \forall j \in [K]\right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- Wasserstein barycenter has much less regularity than the sample mean in the Euclidean space
- So, we consider two measures ρ_0 and ρ_1 have smooth and positive densities on their supports are convex sets, but the support of the interpolant $\rho_{1/2}$ cannot convex

- **Lemma 1.** *Let Ω be a convex domain and μ_0 and μ_1 two measures on the segments S_0 and S_1 as in the **"geometric setting"**. Let ρ_0^n and ρ_1^n be smooth densities weakly converging to ρ_0 and ρ_1, respectively, and concentrated on Ω. Let $\rho_{1/2}^n$ be the middle point of the geodesic in $\mathbb{W}_2(\Omega)$ between them. Then, for n large enough, the support of $\rho_{1/2}^n$ is not convex.*

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i\in[n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}),\quad \forall j\in[K]\right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

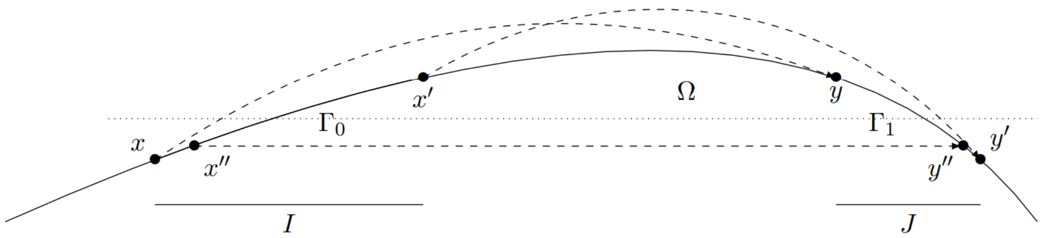


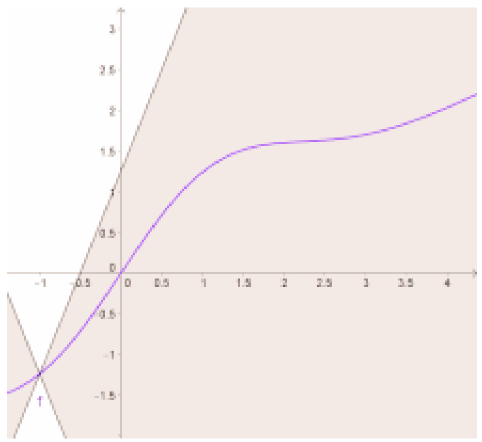
Figure 3: The configuration of Ω, Γ_0, Γ_1.

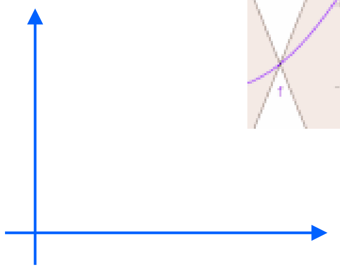
2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- Geometric setting.

- A convex set $\Omega \subset R^d$
- Two portions: $\Gamma_0 = \{(t, f(t) : t \in I)\}$ and $\Gamma_1 = \{(t, g(t) : t \in J)\}$
- $I = [a, b]$ and J $= [c, d]$ and $f: I \to R$, G: $J \to R$ and f, g are Lipschitz functions
- $x = (a, f(a))$ and $x' = (b, f(b))$ the endpoints of Γ_0
- y $= (c, g(c))$ and $y' = (d, g(b))$ the endpoints of Γ_1
- $0 < f' < \lambda$ and $-\lambda < g' < 0$; $\lambda \in (0,1)$
- $f(b) = g(c)$ and $f(a) = g(d)$
- Two points: $x'' \in \Gamma_0$ and $y'' \in \Gamma_1$, with same ordinate
- $x'' < (f(a) + f(b))/2$ and $y'' < (f(c) + f(d))/2$
- Two measures: $\mu_0 \in \mathcal{P}(\Gamma_0)$ and $\mu_1 \in \mathcal{P}(\Gamma_1)$
- A map T: $T(x) = y$, $T(x') = y'$, and $T(x'') = y''$





$$W_2^2(\mu,\nu) := \min_{\gamma} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$\nu_k^{(t+1)} = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

Method

*Hausdorff limit (briefly):
(1) Every bounded and closed subset of metric space is compact
(2) If some subsequence converges to A ($\lim_i A_i = A$)

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*
- ***proof.*** Consider the optimal transport plan $\gamma^n \in \Pi(\rho_0^n, \rho_1^n)$ for quadratic cost
- By uniqueness of optimal plan γ from μ_0 to μ_1 (See *Santambrogio and Wang., 2016*), it is clear that $\gamma^n \rightharpoonup \gamma$ ($\because$Monotonicity is preserved in 1D)
- From the fact that the support of γ is included in the Hausdorff limit of the supports of γ^n, we deduce the existence of points $(x_n, y_n), (x'_n, y'_n), (x''_n, y''_n) \in spt(\gamma^n)$ converging to $(x, y), (x', y')$, and (x'', y''), respectively
- Note that n is a power of cost in $\mathbb{W}_2$(e.g., quadratic cost $||x-y||^2$)

- Since $((x_n + y_n)/2, (x'_n + y'_n)/2)$ belong to the support of $\rho_{1/2}^n$, if this support were convex, it should also contain p_n:

$$p_n := \left(\frac{x_n+y_n}{2} + \frac{x'_n+y'_n}{2}\right)/2 = (x_n + y_n + x'_n + y'_n)/4$$

$$W_2^2(\mu,\nu) := \min_\gamma \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y) \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$\nu_k^{(t+1)} = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- For simplicity, and without loss of generality, suppose $(x_n + y_n + x_n' + y_n')/4 = 0$
- Suppose $p_n \in spt(\rho_{1/2}^n)$
- This means there exist $z_n, w_n \in \Omega$ such that $(z_n + \omega_n)/2 = p_n$ and $(z_n, w_n) \in spt(\gamma^n)$

- Note that the monotonicity of $spt(\gamma^n)$ implies the inequality:

$$(\omega_n - y_n'') \cdot (z_n - x_n'') \geq 0$$

$x'' < (f(a) + f(b))/2$ and $y'' < (f(c) + f(d))/2$

- Using $z + \omega = 0$,

$$\left| z - \frac{x'' - y''}{2} \right|^2 \leq \left| \frac{x'' + y''}{2} \right|^2$$

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i \in [n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}), \quad \forall j \in [K]\right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

- For simplicity, and without loss of generality, suppose $(x_n + y_n + x_n' + y_n')/4 = 0$
- Suppose $p_n \in spt(\rho^n_{1/2})$
- This means there exist $z_n, w_n \in \Omega$ such that $(z_n+\omega_n)/2 = p_n$ and $(z_n, w_n) \in spt(\gamma^n)$

- Note that the monotonicity of $spt(\gamma^n)$ implies the inequality:

$$(\omega_n - y_n'') \cdot (z_n - x_n'') \geq 0$$

$x'' < (f(a) + f(b))/2$ and
$y'' < (f(c) + f(d))/2$

- Using $z + \omega = 0$,

$$\left|z - \frac{x'' - y''}{2}\right|^2 \leq \left|\frac{x'' + y''}{2}\right|^2 - x''y'' \geq 0$$

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \|x-y\|_2^2\,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i\in[n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}),\quad \forall j\in[K]\right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

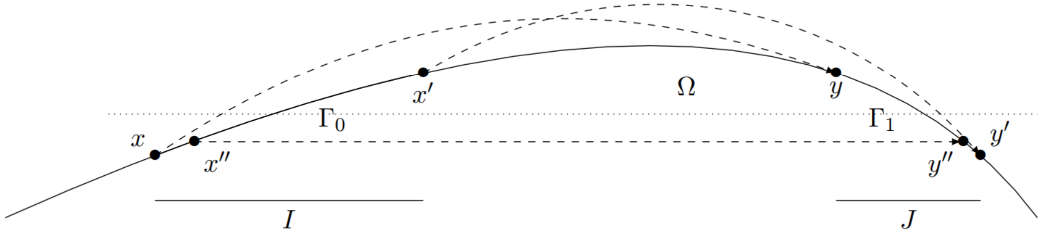


Figure 3: The configuration of Ω, Γ_0, Γ_1.

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: *Example 1 (Irregularity)*

$$\left|z-\frac{x''-y''}{2}\right|^2 \leq \left|\frac{x''+y''}{2}\right|^2 - x''y'' \geq 0 \;\&\; z\in\overline{\Omega}\cap-\overline{\Omega}$$

$0 < f' < \lambda$ and $-\lambda < g' < 0$; $\lambda \in (0,1)$

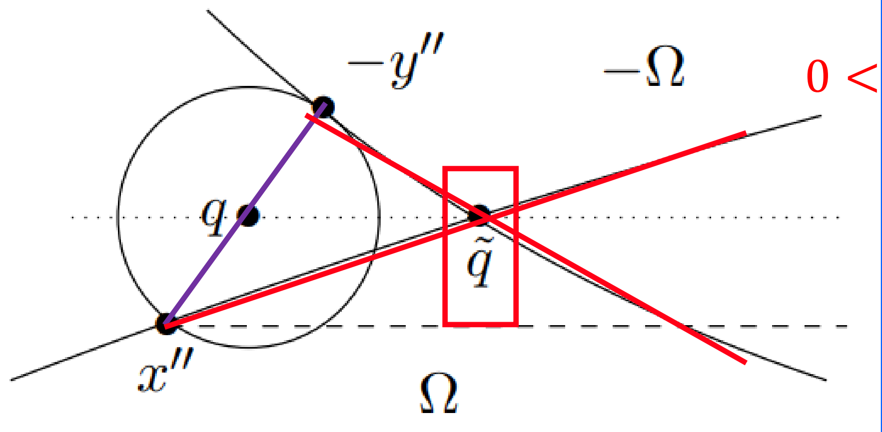


Figure 5: A zoom around q and $\tilde{q}$.

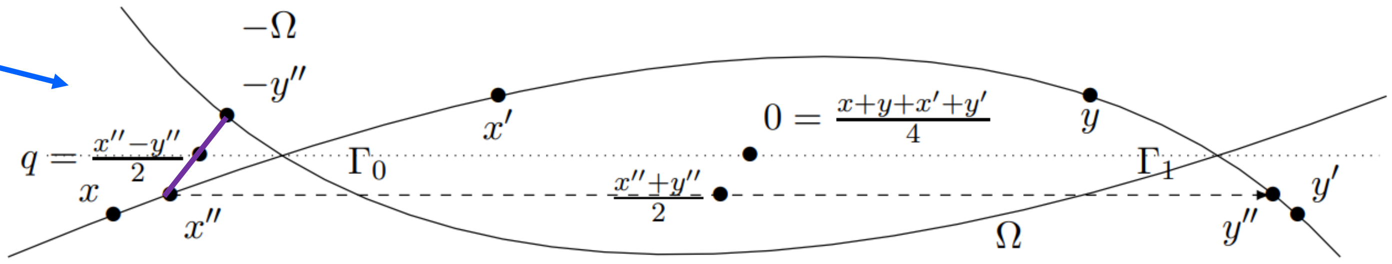


Figure 4: Ω, $-\Omega$, and the reflected points.

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i \in [n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}), \quad \forall j \in [K]\right\}$$

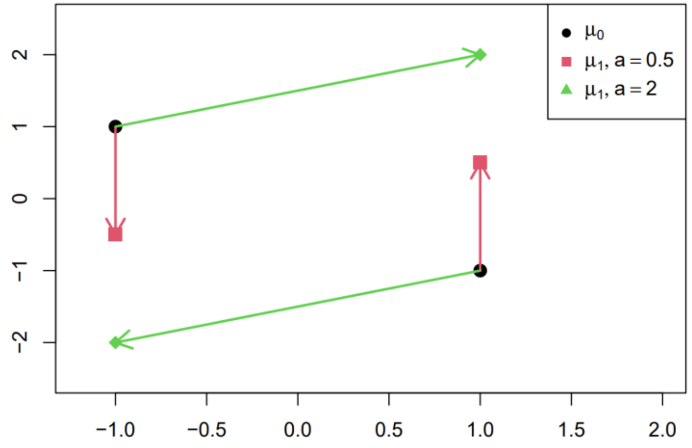
$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: Irregularity and Non-robustness

- ***Example 2***: Non-robustness of Wasserstein barycenters
- Wasserstein barycenter is its sensitivity to data perturbation: A small change mad lead to large (or global) changes in the resulting barycenter
- Source: $\mu_0 = 0.5\delta_{(-1,1)} + 0.5\delta_{(1,-1)}$
- Target: $\mu_1 = 0.5\delta_{(-1,-a)} + 0.5\delta_{(1,a)}$, where $a > 0$ and δ is point mass measure
- Optimal transport map $T := T_{\mu_0\to\mu_1}$



$$T(-1,1) = \begin{cases} (-1,-a) & \text{if } 0 < a < 1 \\ (1,a) & \text{if } a > 1 \end{cases}$$

$$T(1,-1) = \begin{cases} (1,a) & \text{if } 0 < a < 1 \\ (-1,-a) & \text{if } a > 1 \end{cases}$$

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p}\|x-y\|_2^2\,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i\in[n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}),\quad \forall j\in[K]\right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

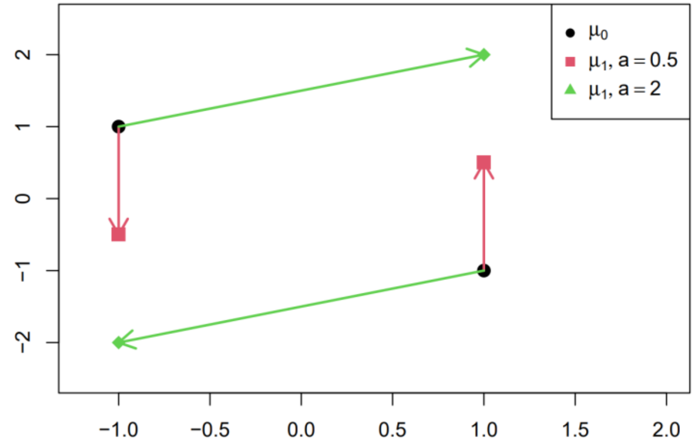
Method

$$\mu_0 \xrightarrow{[(1-t)id+tT]} \mu_t = [(1-t)id+tT]_{\#\mu_0}$$

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: Irregularity and Non-robustness

- ***Example 2***: Non-robustness of Wasserstein barycenters
- Wasserstein barycenter is its sensitivity to data perturbation: A small change mad lead to large (or global) changes in the resulting barycenter
- Source: $\mu_0 = 0.5\delta_{(-1,1)} + 0.5\delta_{(1,-1)}$
- Target: $\mu_1 = 0.5\delta_{(-1,-a)} + 0.5\delta_{(1,a)}$, where $a > 0$ and δ is point mass measure
- Optimal transport map $T := T_{\mu_0\to\mu_1}$



$$T(-1,1) = \begin{cases}(-1,-a) & \text{if } 0<a<1\\ (1,a) & \text{if } a>1\end{cases}$$

$$T(1,-1) = \begin{cases}(1,a) & \text{if } 0<a<1\\ (-1,-a) & \text{if } a>1\end{cases}$$

If not, it's not the optimal.

$\therefore\ \mu_t = [(1-t)id+tT]_{\#\mu_0}$ is a discontinuous function at $a = 1$

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p} \|x-y\|_2^2 \, \mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K]\right\}$$

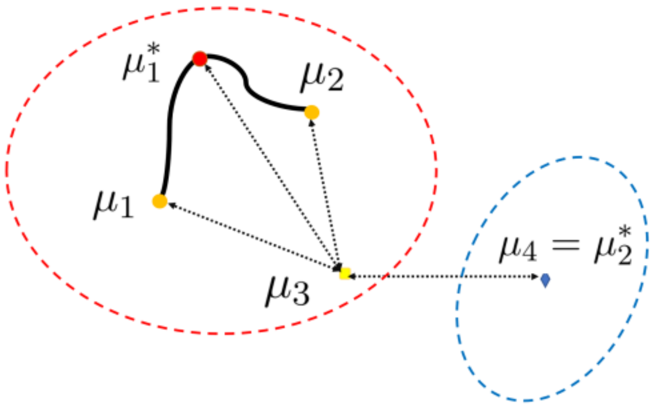
$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: Irregularity and Non-robustness

- ***Example 3***: Failure of centroid-based Wasserstein K-means
- Some distribution μ_3 in the Wasserstein space may have larger W_2 distance to Wasserstein barycenter μ_1^* than every distribution $\mu_i (i = 1,2)$



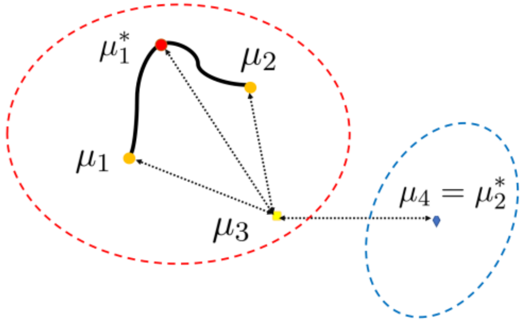
- i.e., $W_2(\mu_3, \mu_1^*) > W_2(\mu_3, \mu_2^*) > \max\{W_2(\mu_3,\mu_1), W_2(\mu_3,\mu_2)\}$

$$W_2^2(\mu, \nu) := \min_{\gamma} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y) \right\}$$

$$G_k^{(t)} = \left\{ i \in [n] : W_2(\mu_i, \nu_k^{(t)}) \leqslant W_2(\mu_i, \nu_j^{(t)}), \quad \forall j \in [K] \right\}$$

$$\nu_k^{(t+1)} = \arg \min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{|G_k^{(t)}|} \sum_{i \in G_k^{(t)}} W_2^2(\mu_i, \nu)$$

Method



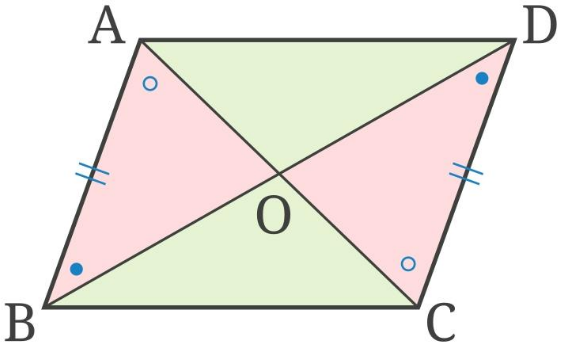
2. Wasserstein K-means clustering

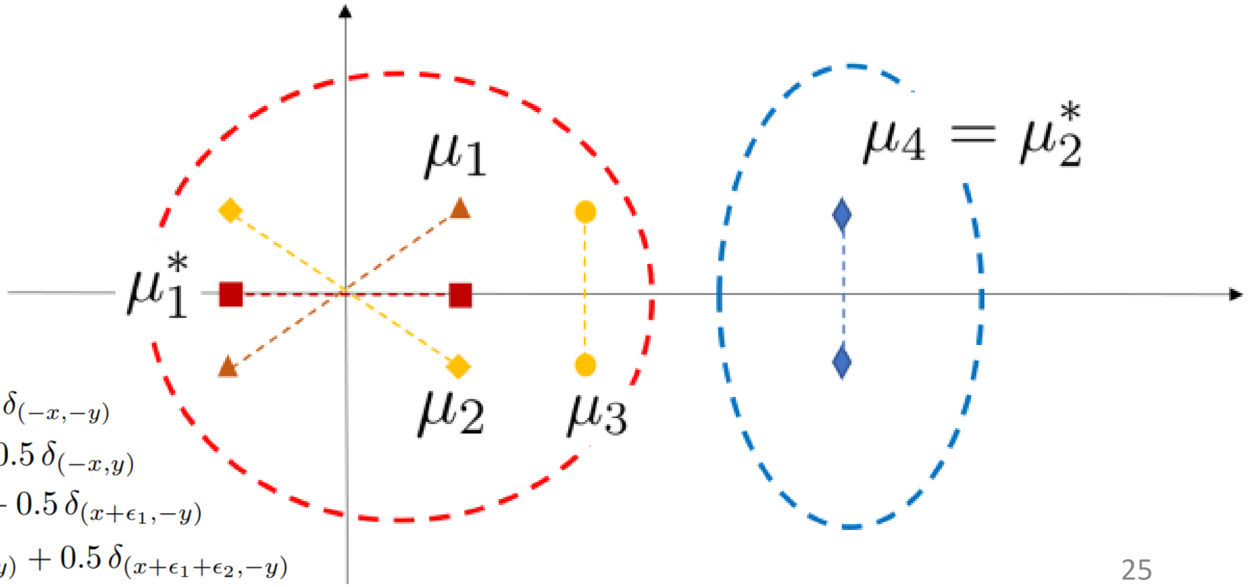
3) Pitfalls of Barycenters-based clustering: Irregularity and Non-robustness

- ***Example 3***: Failure of centroid-based Wasserstein K-means
- In contrast, for Euclidean spaces:

$$\sum_{i=1}^{n} \|X - X_i\|_2^2 = n\|X - \bar{X}\|_2^2 + \sum_{i=1}^{n} \|X_i - \bar{X}\|_2^2 \geqslant n\|X - \bar{X}\|_2^2, \quad \text{for any } X \in \mathbb{R}^p$$

- There is always some point $X_{i^\dagger}$ satisfying $||X - X_{i^\dagger}||_2 \geqslant ||X - \bar{X}||_2$





$\mu_1 = 0.5\,\delta_{(x,y)} + 0.5\,\delta_{(-x,-y)}$

$\mu_2 = 0.5\,\delta_{(x,-y)} + 0.5\,\delta_{(-x,y)}$

$\mu_3 = 0.5\,\delta_{(x+\epsilon_1,y)} + 0.5\,\delta_{(x+\epsilon_1,-y)}$

$\mu_4 = 0.5\,\delta_{(x+\epsilon_1+\epsilon_2,y)} + 0.5\,\delta_{(x+\epsilon_1+\epsilon_2,-y)}$

$$W_2^2(\mu,\nu) := \min_{\gamma}\left\{\int_{\mathbb{R}^p\times\mathbb{R}^p}\|x-y\|_2^2\,\mathrm{d}\gamma(x,y)\right\}$$

$$G_k^{(t)} = \left\{i\in[n] : W_2(\mu_i,\nu_k^{(t)}) \leqslant W_2(\mu_i,\nu_j^{(t)}),\quad \forall j\in[K]\right\}$$

$$\nu_k^{(t+1)} = \arg\min_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} W_2^2(\mu_i,\nu)$$

Method

2. Wasserstein K-means clustering

3) Pitfalls of Barycenters-based clustering: Irregularity and Non-robustness

- ***Example 3***: Failure of centroid-based Wasserstein K-means

Lemma 4 (**Configuration characterization**). If $(x,y,\epsilon_1,\epsilon_2)$ satisfies

$(\epsilon_1+x)^2+x^2$

$$y^2 < \min\{x^2, 0.25\,\Delta_{\epsilon_1,x}\} \quad\text{and}\quad \Delta_{\epsilon_1,x} < \epsilon_2^2 < \Delta_{\epsilon_1,x} + y^2,$$

where $\Delta_{\epsilon_1,x} := \epsilon_1^2 + 2x^2 + 2x\epsilon_1$, then for all sufficiently large m (number of copies of μ_1 and μ_2),

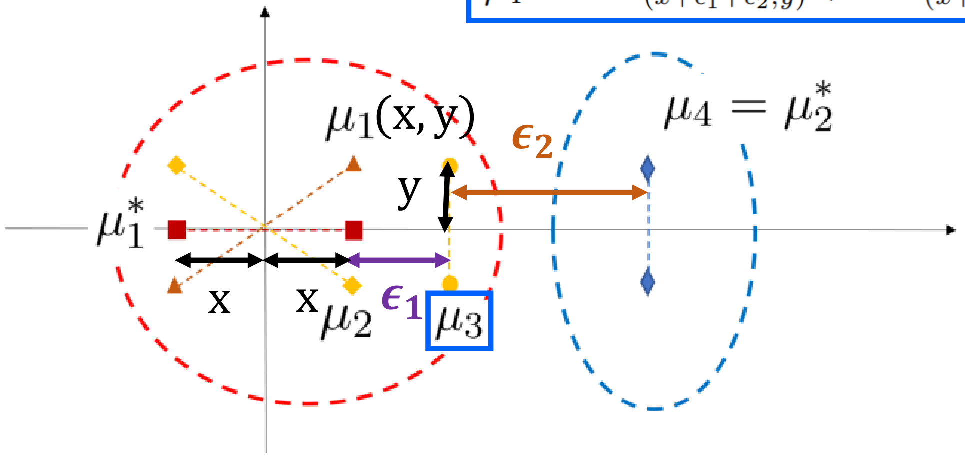
$$W_2(\mu_3,\mu_2^*) < W_2(\mu_3,\mu_1^*) \quad\text{and}\quad \underbrace{\max_{k=1,2}\max_{i,j\in G_k} W_2(\mu_i,\mu_j)}_{\text{largest within-cluster distance}} < \underbrace{\min_{i\in G_1, j\in G_2} W_2(\mu_i,\mu_j)}_{\text{least between-cluster distance}},$$

where μ_k^* denotes the Wasserstein barycenter of cluster G_k for $k=1,2$.

$$\boxed{\begin{aligned}\mu_1 &= 0.5\,\delta_{(x,y)} + 0.5\,\delta_{(-x,-y)}\\ \mu_2 &= 0.5\,\delta_{(x,-y)} + 0.5\,\delta_{(-x,y)}\\ \mu_3 &= 0.5\,\delta_{(x+\epsilon_1,y)} + 0.5\,\delta_{(x+\epsilon_1,-y)}\\ \mu_4 &= 0.5\,\delta_{(x+\epsilon_1+\epsilon_2,y)} + 0.5\,\delta_{(x+\epsilon_1+\epsilon_2,-y)}\end{aligned}}$$

- Where $(x,y,\epsilon_1,\epsilon_2)$ are positive constants

- Lemma 4 stands for $x > y$



Plan for Today

- Summary

1) Centroid-based Wasserstein K-means
2) Three pitfalls of 1)
3) Distance-based Wasserstein K-means
4) Experiments: Real-data applications
5) Discussion

Method

$$\min_{G_1,\dots,G_K}\left\{\sum_{k=1}^{K}\frac{1}{|G_k|}\sum_{i,j\in G_k}\|X_i-X_j\|_2^2 : \bigsqcup_{k=1}^{K} G_k=[n]\right\}$$

2. Wasserstein K-means clustering

4) Clustering based on pairwise distance

- Authors extends the Euclidean distance-based K-means formulation into the Wasserstein space:

$$\min_{G_1,\dots,G_K}\left\{\sum_{k=1}^{K}\frac{1}{|G_k|}\sum_{i,j\in G_k}W_2^2(\mu_i,\mu_j) : \bigsqcup_{k=1}^{K} G_k=[n]\right\}$$

- Given an initial cluster membership estimate $G_1^{(1)},\dots,G_K^{(1)}$, one assigns each probability measure $\mu_1,\dots,\mu_n$ based on minimizing the averaged squared W_2:

$$G_k^{(t+1)}=\left\{i\in[n]:\frac{1}{|G_k^{(t)}|}\sum_{s\in G_k^{(t)}}W_2^2(\mu_i,\mu_s)\leqslant\frac{1}{|G_j^{(t)}|}\sum_{s\in G_j^{(t)}}W_2^2(\mu_i,\mu_s),\quad\forall j\in[K]\right\}$$

Method

$$\min_{G_1,\dots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \|X_i - X_j\|_2^2 : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

2. Wasserstein K-means clustering

5) Connections to the standard K-means clustering in Euclidean space

Example 5 (**Degenerate probability measures**). If the probability measures are Dirac at point $X_i \in \mathbb{R}^p$, i.e., $\mu_i = \delta_{X_i}$, then the Wasserstein K-means is the same as the standard K-means since $W_2(\mu_i, \mu_j) = \|X_i - X_j\|_2$. ■

Example 6 (**Gaussian measures**). If $\mu_i = N(m_i, V_i)$ with positive-definite covariance matrices $\Sigma_i \succ 0$, then the squared 2-Wasserstein distance can be expressed as the sum of the squared Euclidean distance on the mean vector and

$$d^2(V_i, V_j) = \mathrm{Tr}\left[V_i + V_j - 2\left(V_i^{1/2} V_j V_i^{1/2} \right)^{1/2} \right], \tag{13}$$

the squared *Bures distance* on the covariance matrix [Bhatia et al., 2019]. Here, we use $V^{1/2}$ to denote the unique symmetric square root matrix of $V \succ 0$. That is,

$$W_2^2(\mu_i, \mu_j) = \|m_i - m_j\|_2^2 + d^2(V_i, V_j). \tag{14}$$

Then the Wasserstein K-means, formulated either in (7) or (11), can be viewed as a *covariance-adjusted* Euclidean K-means by taking account into the shape or orientation information in the (non-degenerate) Gaussian inputs. ■

Example 7 (**One-dimensional probability measures**). If μ_i are probability measures on $\mathbb{R}$ with cumulative distribution function (cdf) F_i, then the Wasserstein distance can be written in terms of the *quantile transform*

$$W_2^2(\mu_i, \mu_j) = \int_0^1 [F_i^-(u) - F_j^-(u)]^2 \, \mathrm{d}u, \tag{15}$$

where F^- is the generalized inverse of the cdf F on $[0, 1]$ defined as $F^-(u) = \inf\{x \in \mathbb{R} : F(x) > u\}$ (cf. Theorem 2.18 [Villani, 2003]). Thus the one-dimensional probability measures in Wasserstein space can be isometrically embedded in a flat L^2 space, and we can bring back the equivalence of the Wasserstein and Euclidean K-means clustering methods. ■

Examples [edit]

Point masses (degenerate distributions) [edit]

Let $\mu_1 = \delta_{a_1}$ and $\mu_2 = \delta_{a_2}$ be two degenerate distributions (i.e. Dirac delta distributions) located at points a_1 and a_2 in $\mathbb{R}$. There is only one possible coupling of these two measures, namely the point mass $\delta_{(a_1,a_2)}$ located at $(a_1, a_2) \in \mathbb{R}^2$. Thus, using the usual absolute value function as the distance function on $\mathbb{R}$, for any $p \geq 1$, the p-Wasserstein distance between μ_1 and μ_2 is

$$W_p(\mu_1, \mu_2) = |a_1 - a_2|.$$

By similar reasoning, if $\mu_1 = \delta_{a_1}$ and $\mu_2 = \delta_{a_2}$ are point masses located at points a_1 and a_2 in $\mathbb{R}^n$, and we use the usual Euclidean norm on $\mathbb{R}^n$ as the distance function, then

$$W_p(\mu_1, \mu_2) = \|a_1 - a_2\|_2.$$

Normal distributions [edit]

Let $\mu_1 = \mathcal{N}(m_1, C_1)$ and $\mu_2 = \mathcal{N}(m_2, C_2)$ be two non-degenerate Gaussian measures (i.e. normal distributions) on $\mathbb{R}^n$, with respective expected values m_1 and $m_2 \in \mathbb{R}^n$ and symmetric positive semi-definite covariance matrices C_1 and $C_2 \in \mathbb{R}^{n\times n}$. Then,[3] with respect to the usual Euclidean norm on $\mathbb{R}^n$, the 2-Wasserstein distance between μ_1 and μ_2 is

$$W_2(\mu_1, \mu_2)^2 = \|m_1 - m_2\|_2^2 + \mathrm{trace}\left(C_1 + C_2 - 2\left(C_2^{1/2} C_1 C_2^{1/2}\right)^{1/2}\right).$$

This result generalises the earlier example of the Wasserstein distance between two point masses (at least in the case $p = 2$), since a point mass can be regarded as a normal distribution with covariance matrix equal to zero, in which case the trace term disappears and only the term involving the Euclidean distance between the means remains.

One-dimensional distributions [edit]

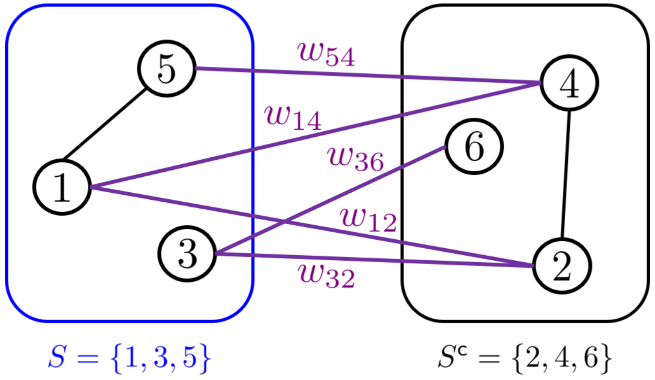
Let $\mu_1, \mu_2 \in P_p(\mathbb{R})$ be probability measures on $\mathbb{R}$, and denote their cumulative distribution functions by $F_1(x)$ and $F_2(x)$. Then the transport problem has an analytic solution: Optimal transport preserves the order of probability mass elements, so the mass at quantile q of μ_1 moves to quantile q of μ_2. Thus, the p-Wasserstein distance between μ_1 and μ_2 is

$$W_p(\mu_1, \mu_2) = \left(\int_0^1 \left|F_1^{-1}(q) - F_2^{-1}(q)\right|^p \mathrm{d}q \right)^{1/p}$$

where F_1^{-1} and F_2^{-1} are the quantile functions (inverse CDFs). In the case of $p = 1$, a change of variables leads to the formula

$$W_1(\mu_1, \mu_2) = \int_{\mathbb{R}} |F_1(x) - F_2(x)| \, \mathrm{d}x.$$

Method



3. SDP relaxation

1) [Detour] A simple example: Maximum cut problem

- SDP relaxation
- Relaxation means constraint is ignored
- This represents there is more space to explore for optimization:

$$p^*_{SDP} \coloneqq \max_X \sum_{i,j} \frac{1}{2}\omega_{ij}(1 - X_{ij}) : X_{ii} = 1, X \succcurlyeq 0, \cancel{rank(X) = 1}$$

Here, we employ a relaxation for maximization. So,

$$p^*_{SDP} \geq p^*$$

Method

3. SDP relaxation

2) Computational complexity

- Wasserstein Lloyd's algorithm requires to use and compute the barycenter at each iteration

- The distance-based K-means is worse-case NP-hard for Euclidean data

- Common way is to consider convex relaxations to approximate the solution below:

$$\min_{G_1,\ldots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} W_2^2(\mu_i, \mu_j) : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

$$\min_{G_1,\dots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j\in G_k} W_2^2(\mu_i,\mu_j) : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

Method

$$p^*_{SDP} := \max_X \sum_{i,j} \frac{1}{2}\omega_{ij}(1 - X_{ij}) : X_{ii} = 1, X \succcurlyeq 0, \cancel{rank(X) = 1}$$

3. SDP relaxation

3) Cutoff for exact recovery of gaussian mixture models

- We obtain the SDP relaxation of the equation by only preserving these convex constraints:

$$\min_{Z\in\mathbb{R}^{n\times n}} \left\{ \langle A, Z\rangle : Z^\top = Z,\ Z \succeq 0,\ \mathrm{Tr}(Z) = K,\ Z\mathbf{1}_n = \mathbf{1}_n,\ Z \geqslant 0 \right\}$$

- Without loss of generality, we focus on mean-zero Gaussian distributions since optimal separation conditions for exact recovery based on the Euclidean mean component:

$$V_i = (I + tX_i)V^{(k)}(I + tX_i) \quad \text{with } X_1,\dots,X_n \overset{i.i.d.}{\sim} SymN(0,1)$$

$$\min_{G_1,\dots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j\in G_k} W_2^2(\mu_i,\mu_j) : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

Method

$$p^*_{SDP} \coloneqq \max_X \sum_{i,j} \frac{1}{2}\omega_{ij}(1 - X_{ij}) : X_{ii} = 1, X \succcurlyeq 0, \cancel{rank(X) = 1}$$

3. SDP relaxation

Theorem 8 (**Exact recovery for clustering Gaussians**). Let $\Delta^2 := \min_{k\neq l} d^2(V^{(k)}, V^{(l)})$ denote the minimal pairwise separation among clusters, $\bar{n} := \max_{k\in[K]} n_k$ (and $\underline{n} := \min_{k\in[K]} n_k$) the maximum (minimum) cluster size, and $m := \min_{k\neq l} \frac{2n_k n_l}{n_k+n_l}$ the minimal pairwise harmonic mean of cluster sizes. Suppose the covariance matrix V_i of Gaussian distribution $\nu_i = N(0, V_i)$ is independently drawn from model (18) for $i = 1, 2, \dots, n$. Let $\beta \in (0, 1)$. If the separation Δ^2 satisfies

$$\Delta^2 > \bar{\Delta}^2 := \frac{C_1 t^2}{\min\{(1-\beta)^2, \beta^2\}} \mathcal{V} p^2 \log n, \tag{19}$$

then the SDP (17) achieves exact recovery with probability at least $1 - C_2 n^{-1}$, provided that

$$\underline{n} \geq C_3 \log^2 n, \quad t \leq C_4 \sqrt{\log n} / \left[(p + \log \bar{n}) \mathcal{V}^{1/2} T_v^{1/2}\right], \quad n/m \leq C_5 \log n,$$

where $\mathcal{V} = \max_k \left\|V^{(k)}\right\|_{\mathrm{op}}$, $T_v = \max_k \mathrm{Tr}\left[\left(V^{(k)}\right)^{-1}\right]$, and $C_i, i = 1, 2, 3, 4, 5$ are constants.

Outline

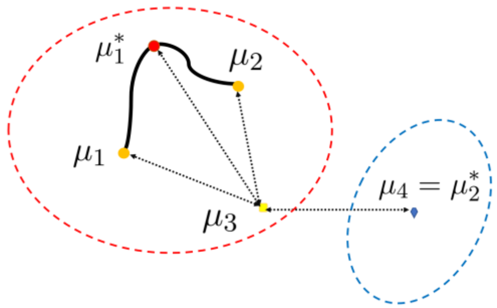
1. Background
2. Method
3. Experiments
4. Discussion

Experiments

- Summary

  1) Counter-example
  2) Gaussian distribution
  3) Real-data applications

Experiments



1. Counter-example in Example 3 revisited

- Instead of using point mass measures, we use Gaussian distributions with small variance as a smoothed version

- Authors consider $K = 2$, where cluster G_1^* consists of m_1 many copies of (μ_1, μ_2) pairs and m_2 many μ_3, and cluster G_2^* consists of m_3 many copies of μ_4.

- Authors choose μ_i as the following two-dimensional mixture of Gaussian distributions:

$$\mu_i = 0.5N\left(a_{i,1}, \Sigma_{i,1}\right) + 0.5N\left(a_{i,2}, \Sigma_{i,2}\right), for\ i = 1,2,3,4$$

Table 6: The time cost with standard deviation shown in parentheses for the counter example. TC: Time cost, W-SDP: Wasserstein SDP, D-WKM: Distance-based Wasserstein K-means, B-WKM: Barycenter-based Wasserstein K-means.

n	TC for W-SDP (SD)	TC for D-WKM	TC for B-WKM (SD)
101	14.50 (0.5873)	14.15 (0.5132)	181.1 (372.4)
202	56.94 (1.490)	54.98 (1.516)	341.0 (136.2)
303	128.4 (3.640)	123.9 (3.606)	549.2 (200.2)

Experiments

1. Counter-example in Example 3 revisited

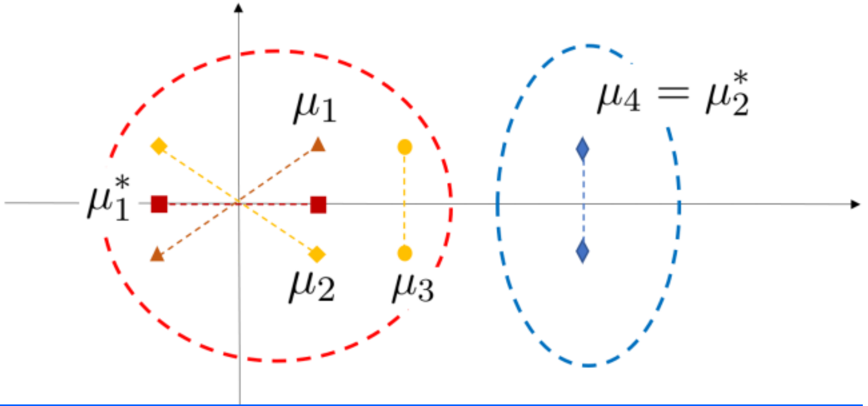


Table 1: Exact recovery rates and frequency of $\Delta_1 > \Delta_2$ for B-WKM among total 50 repetitions in the counter example. W-SDP: Wasserstein SDP, D-WKM: Distance-based Wasserstein K-means, B-WKM: Barycenter-based Wasserstein K-means. n: total number of distributions.

n	W-SDP	D-WKM	B-WKM	Frequency of $\Delta_1 > \Delta_2$
101	1.00	0.82	0.40	0.32
202	1.00	0.84	0.34	0.26
303	1.00	0.72	0.46	0.20

$\Delta_k := W^2(\mu_3, \mu_k^*)$

as the squared distance between μ_3 and μ_k^*

for $k = 1, 2$, where μ_k^* is the barycenter of G_k^*

Table 7: Estimated Wasserstein distances with standard deviation shown in parentheses and frequency of $\Delta^* > \Delta_*$ for the counter example.

n	Δ_*	Δ^*	Frequency of $\Delta_* < \Delta^*$
101	0.1978 (0.0055)	0.2046 (0.0050)	0.8200
202	0.1990 (0.0058)	0.2050 (0.0051)	0.8200
303	0.1996 (0.0067)	0.2052 (0.0050)	0.7600

$\Delta_* := \max_{k=1,2} \max_{i,j \in G_k} W_2(\mu_i, \mu_j)$

$\Delta^* := \min_{i \in G_1, j \in G_2} W^2(\mu_i, \mu_j)$

$$\underbrace{\max_{k=1,2} \max_{i,j \in G_k} W_2(\mu_i, \mu_j)}_{\text{largest within-cluster distance}} < \underbrace{\min_{i \in G_1, j \in G_2} W_2(\mu_i, \mu_j)}_{\text{least between-cluster distance}},$$

Experiments

2. Gaussian distribution

- K = 4 and all cluster size equal

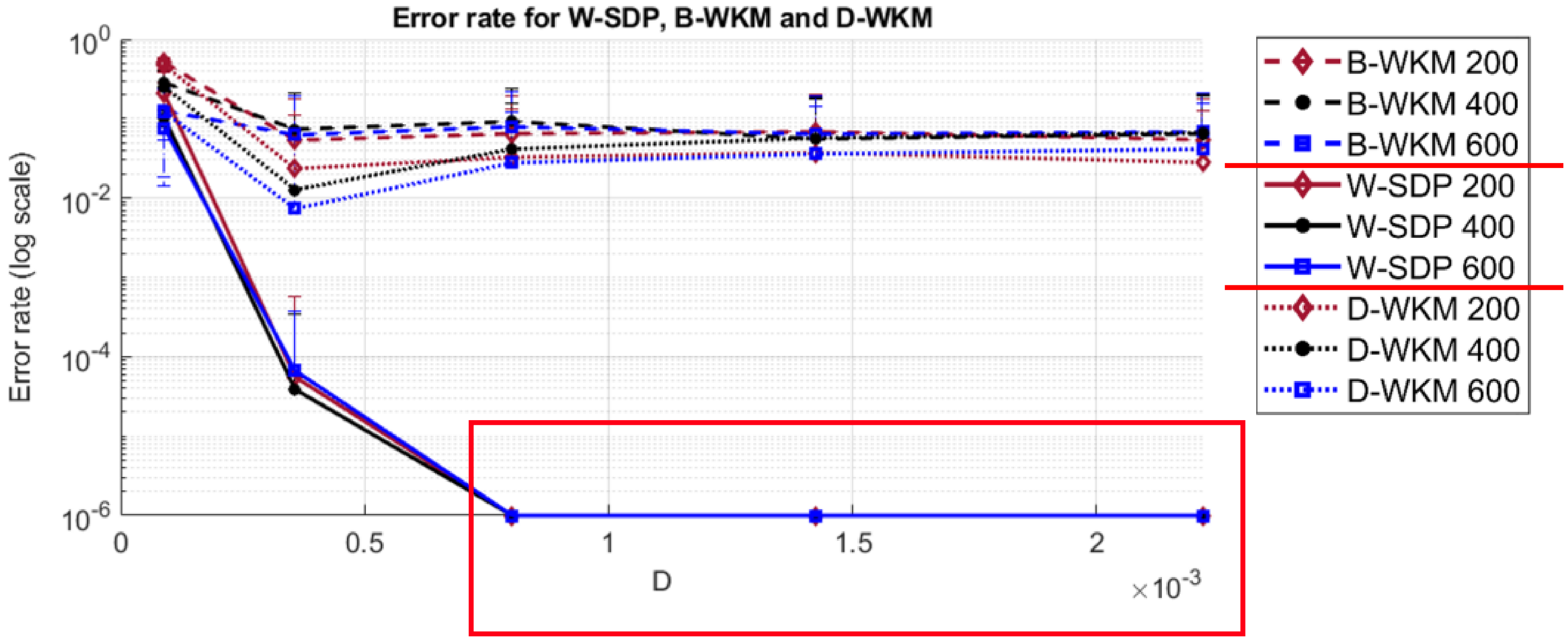


Figure 4: Mis-classification error versus squared distance D from Wasserstein SDP (W-SDP) and barycenter/distance-based Wasserstein K-means (B-WKM and D-WKM) for clustering Gaussians under $n \in \{200, 400, 600\}$. Due to the log-scale, 10^{-6} corresponds to exact recovery.

Experiments

3. Real-data applications

- 100 iterations, Sinkhorn divergence
- MNIST Case 1: 200 # “0”, 100 # “5”
- MNIST Case 2: 400 # “0”, 200 # “5”
- Fashion-MNIST (“T-shirt/top” and “Trouser”)
- USPS (handwriting digits; “5” and “7”)

Table 2: Error rate (SD) for clustering three benchmark datasets: MNIST, Fashion-MNIST and USPS handwriting digits. MNIST_1 (MNIST_2) refers to the results of Case 1 (Case 2) for MNIST dataset.

	W-SDP	D-WKM	B-WKM	KM
MNIST_1	0.235 (0.045)	0.156 (0.057)	0.310 (0.069)	0.295 (0.066)
MNIST_2	0.279 (0.050)	0.185 (0.097)	0.324 (0.032)	0.362 (0.033)
Fashion-MNIST	0.082 (0.020)	0.056 (0.014)	0.141 (0.059)	0.138 (0.099)
USPS handwriting	0.206 (0.020)	0.159 (0.061)	0.240 (0.045)	0.284 (0.025)

Outline

1. Background
2. Method
3. Experiments
4. Discussion

Discussion

- Future Work
  - Their approaches could be serious when sample size is large
  - Complexity issue

{"dialogue_id": "000", "turns": [{"turn_id": "insurance_000_000", "speaker_role": "Agent", "utterance": "Hello this is Jane at Rivertown Insurance. How can I help you today?", "d
{"dialogue_id": "001", "turns": [{"turn_id": "insurance_001_000", "speaker_role": "Agent", "utterance": "Hello this is Allision with Rivertown Insurance, returning a call to mister
{"dialogue_id": "002", "turns": [{"turn_id": "insurance_002_000", "speaker_role": "Agent", "utterance": "Thank you for calling Rivertown Insurance. How may I assist you today?", "di
{"dialogue_id": "003", "turns": [{"turn_id": "insurance_003_000", "speaker_role": "Customer", "utterance": "Hello.", "dialogue_acts": [], "intents": []}, {"turn_id": "insurance_003_
{"dialogue_id": "004", "turns": [{"turn_id": "insurance_004_000", "speaker_role": "Agent", "utterance": "Hello, thank you for calling Rivertown Insurance helpdesk. My name is Russ.
{"dialogue_id": "005", "turns": [{"turn_id": "insurance_005_000", "speaker_role": "Agent", "utterance": "Hi. Thank you for holding. My name is Rebecca.", "dialogue_acts": [], "inten
{"dialogue_id": "006", "turns": [{"turn_id": "insurance_006_000", "speaker_role": "Agent", "utterance": "Hello. Thank you for calling Rivertown Insurance helpline. My name is Michel
{"dialogue_id": "007", "turns": [{"turn_id": "insurance_007_000", "speaker_role": "Agent", "utterance": "Thank you for calling Rivertown Insurance helpline. My name is Meg.", "dialo
{"dialogue_id": "008", "turns": [{"turn_id": "insurance_008_000", "speaker_role": "Agent", "utterance": "Hello, thank you for calling Rivertown Insurance helpdesk. My name is Elizab
{"dialogue_id": "009", "turns": [{"turn_id": "insurance_009_000", "speaker_role": "Agent", "utterance": "Thank you for calling Rivertown Insurance helpline. My name is Ann. How may
{"dialogue_id": "010", "turns": [{"turn_id": "insurance_010_000", "speaker_role": "Agent", "utterance": "Hello my name is Betty, thank you for calling Rivertown Insurance. How may I
{"dialogue_id": "011", "turns": [{"turn_id": "insurance_011_000", "speaker_role": "Agent", "utterance": "Hello, thanks for calling Rivertown Insurance. How can I help you?", "dialog
{"dialogue_id": "012", "turns": [{"turn_id": "insurance_012_000", "speaker_role": "Agent", "utterance": "Hello. Thank you for calling Rivertown Insurance helpdesk. How may I help yo
{"dialogue_id": "013", "turns": [{"turn_id": "insurance_013_000", "speaker_role": "Agent", "utterance": "Thank you for calling Rivertown Insurance. How may I assist you today?", "di
{"dialogue_id": "014", "turns": [{"turn_id": "insurance_014_000", "speaker_role": "Agent", "utterance": "Hello. Thank you for calling Rivertown Insurance helpline. My name is Anthon
{"dialogue_id": "015", "turns": [{"turn_id": "insurance_015_000", "speaker_role": "Agent", "utterance": "Hello. Thank you for calling Rivertown Insurance, my name is Alice. How may
{"dialogue_id": "016", "turns": [{"turn_id": "insurance_016_000", "speaker_role": "Agent", "utterance": "Hello good morning. My name is Emily. Thank you for calling Rivertown Insura
{"dialogue_id": "017", "turns": [{"turn_id": "insurance_017_000", "speaker_role": "Agent", "utterance": "Hello. Thank you for calling Rivertown Insurance helpline. How may I help yo
{"dialogue_id": "018", "turns": [{"turn_id": "insurance_018_000", "speaker_role": "Agent", "utterance": "Hello, thanks for calling Rivertown Insurance helpline. How can I help you?"
{"dialogue_id": "019", "turns": [{"turn_id": "insurance_019_000", "speaker_role": "Customer", "utterance": "Hello.", "dialogue_acts": [], "intents": []}, {"turn_id": "insurance_019_
{"dialogue_id": "020", "turns": [{"turn_id": "insurance_020_000", "speaker_role": "Agent", "utterance": "Hello. Thank you for calling Rivertown Insurance helpline. My name is Anna.
{"dialogue_id": "021", "turns": [{"turn_id": "insurance_021_000", "speaker_role": "Agent", "utterance": "Hello, thank you for holding the line. My name is Curtis. How may I help you
{"dialogue_id": "022", "turns": [{"turn_id": "insurance_022_000", "speaker_role": "Agent", "utterance": "Hello, Thanks for calling Rivertown Insurance, my name is Jen, can I help yo
{"dialogue_id": "023", "turns": [{"turn_id": "insurance_023_000", "speaker_role": "Agent", "utterance": "Thank you for Calling Rivertown Insurance my name id Dale, how may I help yo
{"dialogue_id": "024", "turns": [{"turn_id": "insurance_024_000", "speaker_role": "Agent", "utterance": "Good afternoon, and thanks for calling Rivertown Insurance. How may I help y
{"dialogue_id": "025", "turns": [{"turn_id": "insurance_025_000", "speaker_role": "Agent", "utterance": "Hi there! Thank you for calling Rivertown Insurance, my name is Rachel, how

Thank you

https://jeiyoon.github.io/