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

Paper review

Wasserstein K-means for clustering probability distributions (NeurIPS 2022) - 1

Presentation: **Jeiyoon Park**

6^{th} Generation, TAVE

TAVE Research

Outline

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

Outline

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

Background

1. Before we get started,

1) Summary:

- Authors observed and analyzed the peculiar behaviors of Wasserstein barycenters and their results in clustering probability

- This paper proposes distance-based K-means approach (D-WKW) and its semidefinite program relaxation (W-SDP) by showing the exact recovery results for Gaussians.

2) It will be a very long journey…

Background: Wasserstein Distance

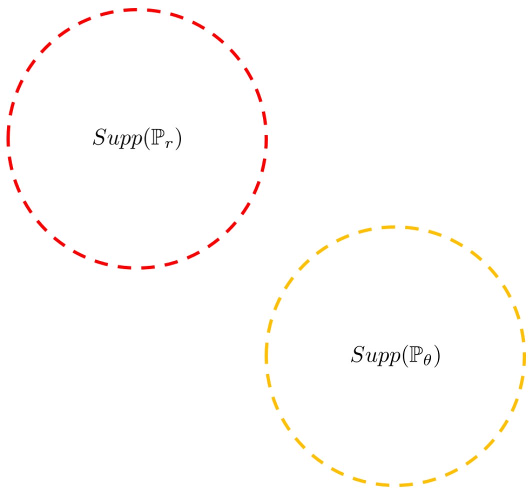
2. Why Wasserstein distance?

1) Maximum Likelihood Estimation (MLE)

- Given θ-parametrized distributions $(P_\theta)_{\theta \in R^d}$ and dataset $\{x^{(i)}\}_{i=1}^{m}$,
- Find the values of the model parameters that maximize the likelihood function over parameter space:

$$\max_{\theta \in R^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)}) \leftrightarrow Minimize\ KL - divergence$$

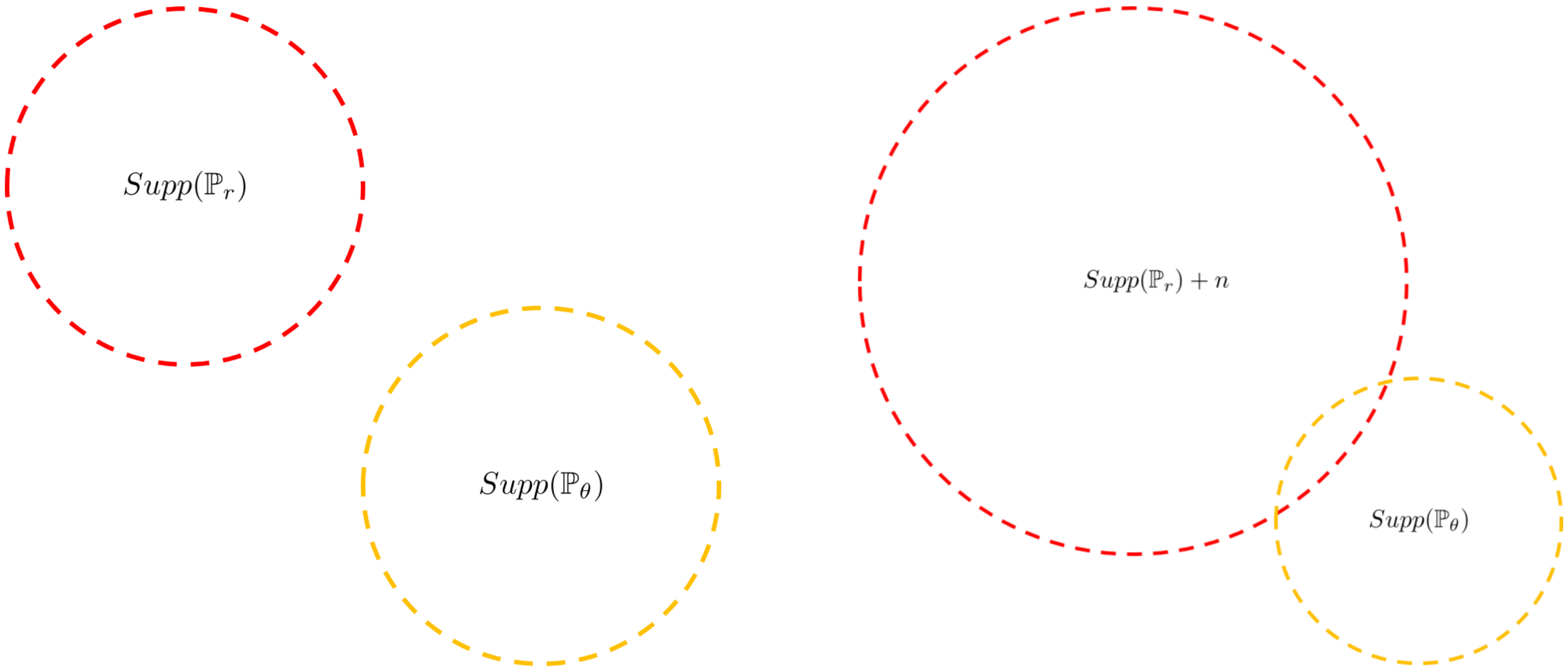
- However, If the supports of the two distributions don't overlap, the KL-divergence will diverge. (i.e., We can't compute KL-divergence)



Background: Wasserstein Distance

2. Why Wasserstein distance?

2) Wasserstein GAN (Arjovsky et al., ICML 2017)
- Adding Gaussian noise to images makes images very blurry
- Since GAN doesn't need to directly predict the distribution, we just input data into the model without prior.

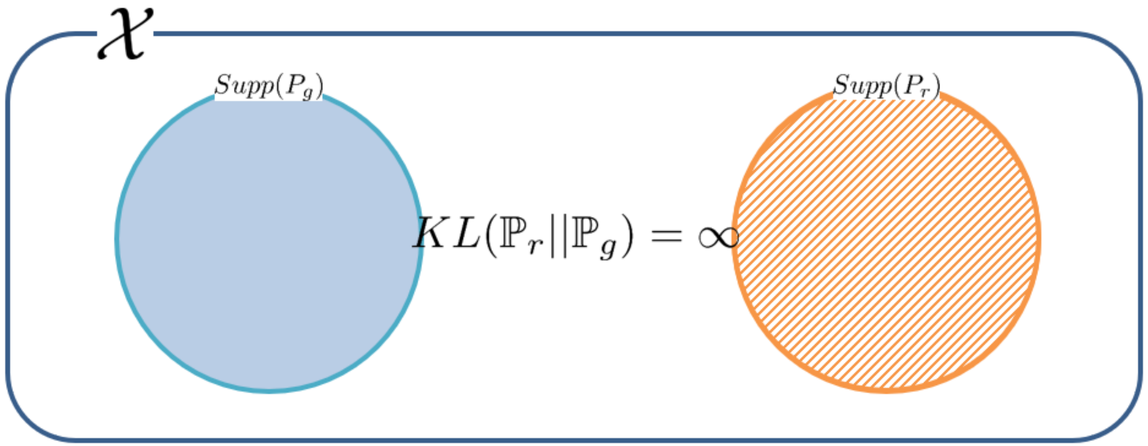


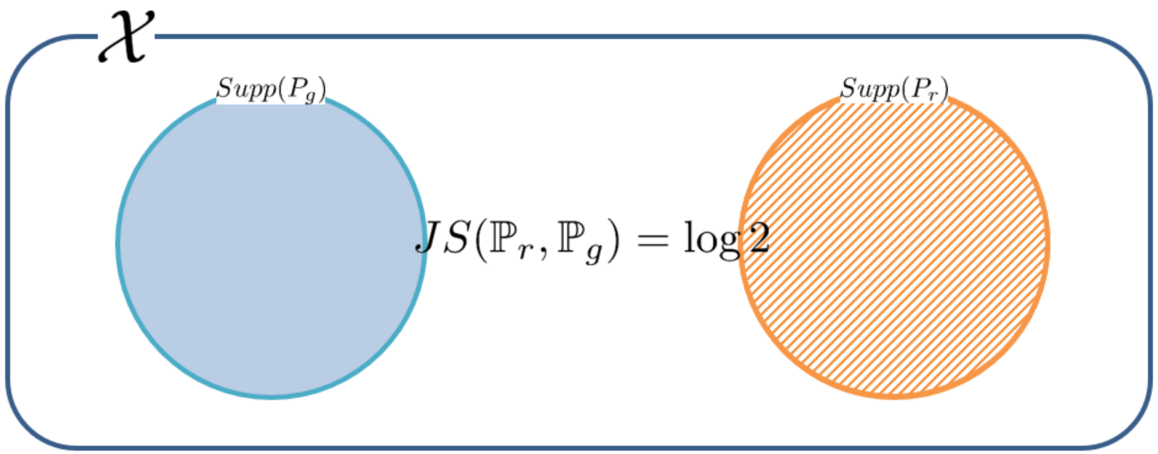
Background: Wasserstein Distance

2. Why Wasserstein distance?

3) Wasserstein distance (a.k.a., Earth Mover distance)

- Existing measure: KL divergence and Jensen-Shannon divergence (JS)
- If the supports of the two distributions don't overlap, the KLD will diverge
- If the supports of the two distributions don't overlap, the JSD will be $\log 2$ which can't provide information about how far away are they.





Background: Wasserstein Distance

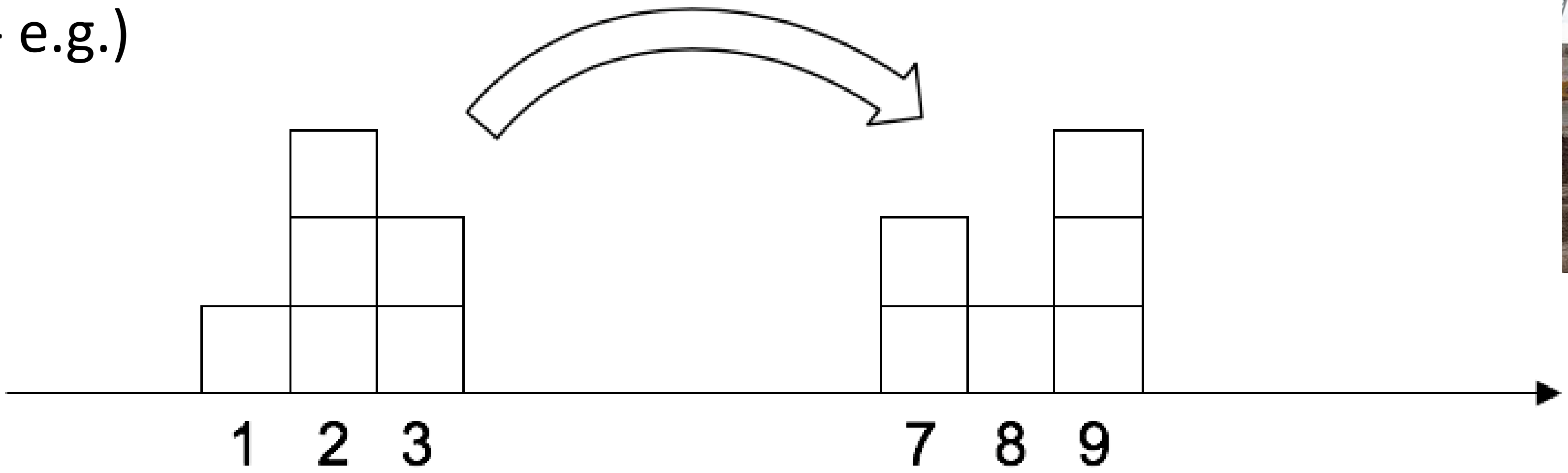
2. Why Wasserstein distance?

3) Wasserstein distance (a.k.a., Earth Mover distance)

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

, where $\prod(\mathbb{P}_r, \mathbb{P}_g)$ denotes a set of joint distribution.

- e.g.)



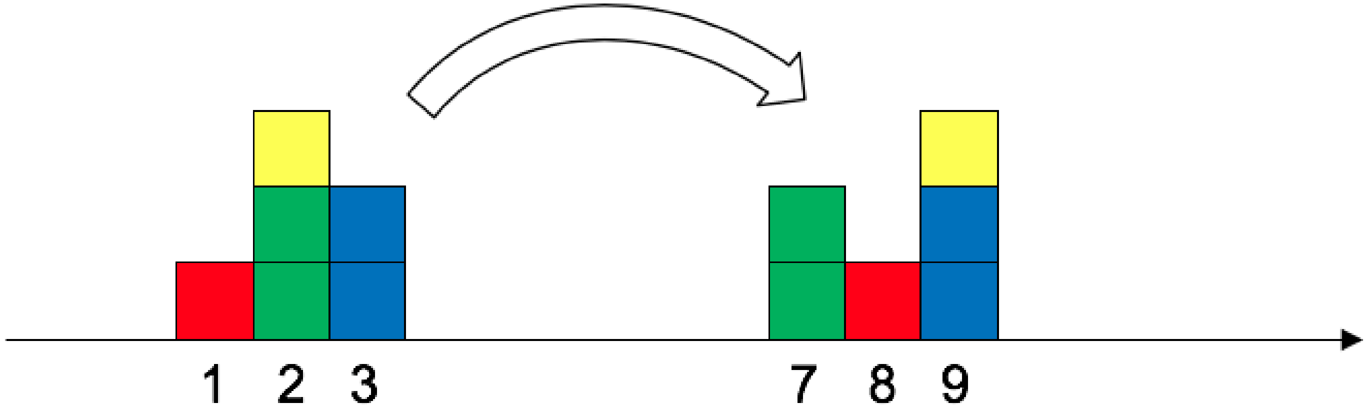
https://haawron.tistory.com/21

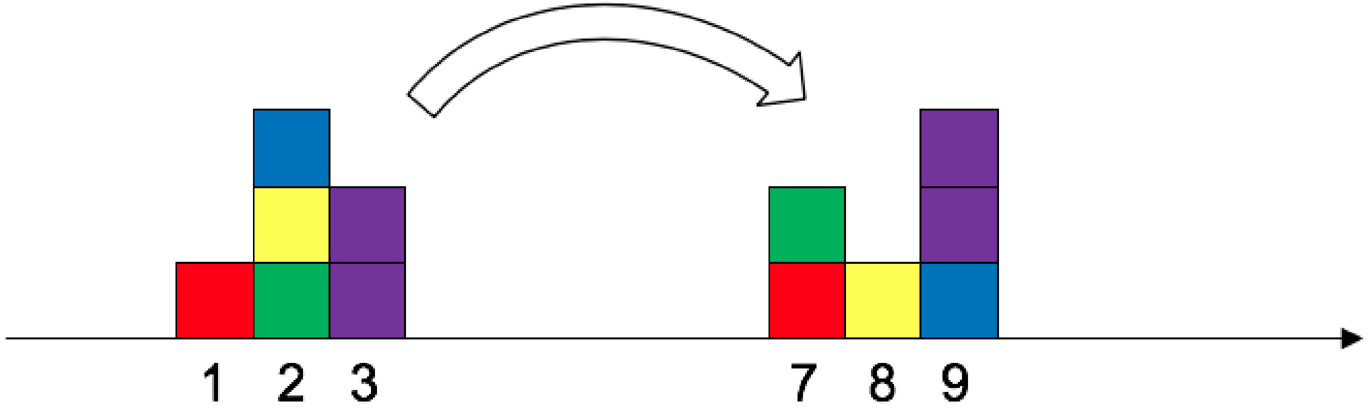
Background: Wasserstein Distance

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

2. Why Wasserstein distance?

- e.g.) Earth-Mover distance
- (left) $\mathbb{P}_r \rightarrow$ (right) $\mathbb{P}_g$



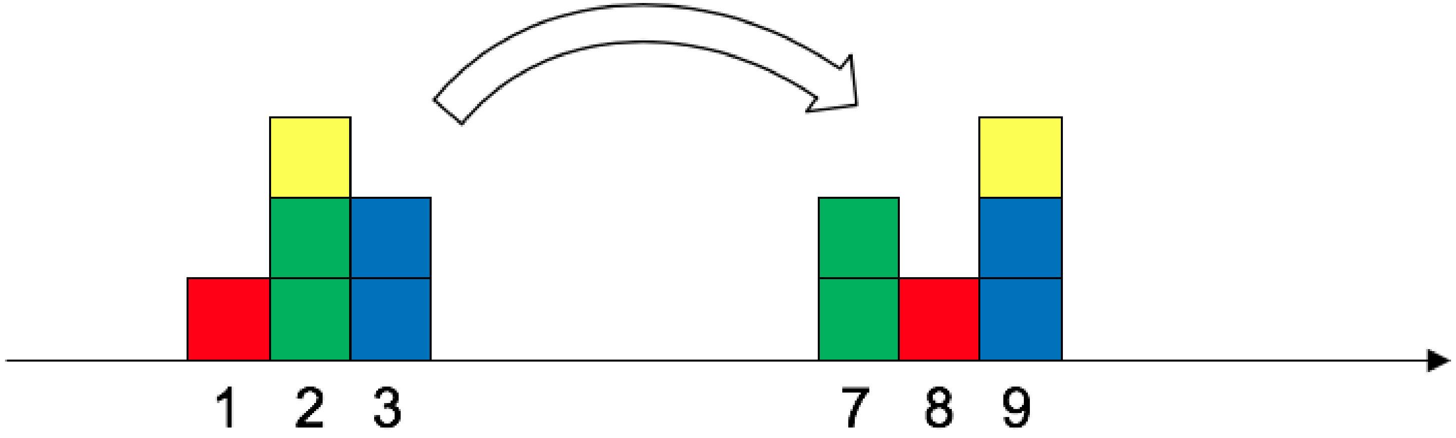


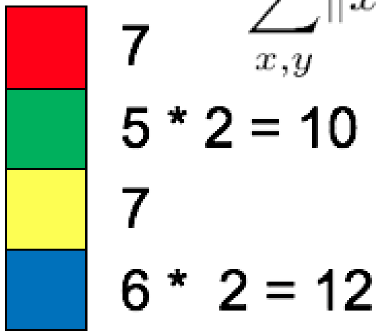
Background: Wasserstein Distance

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

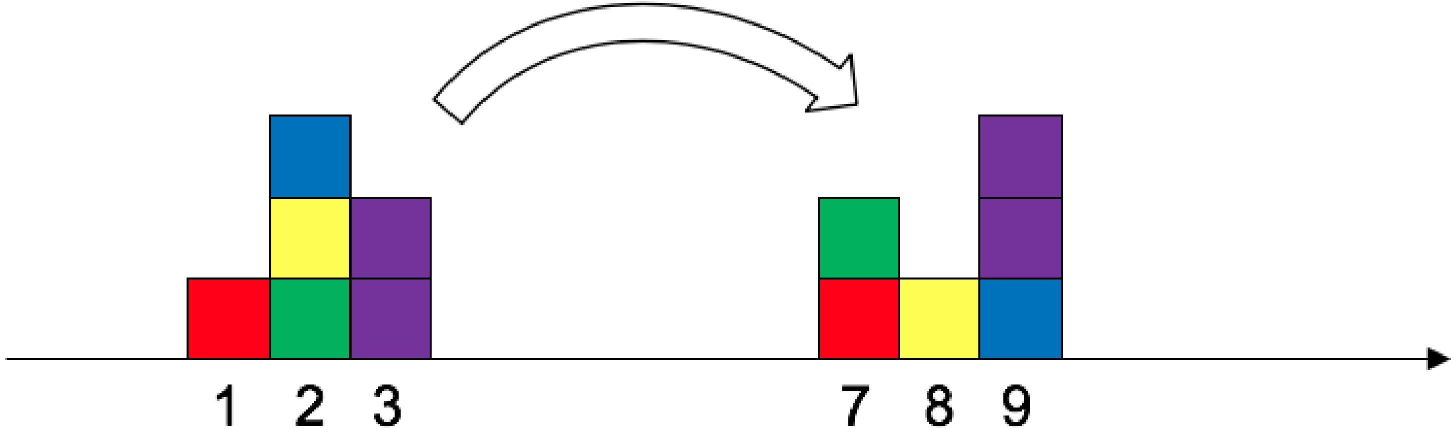
2. Why Wasserstein distance?

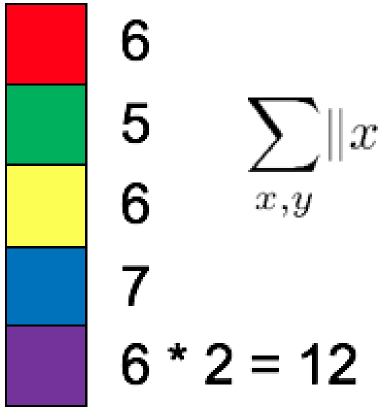
- e.g.) Earth-Mover distance





$$\sum_{x,y} ||x - y|| = 36$$





$$\sum_{x,y} ||x - y|| = 36$$

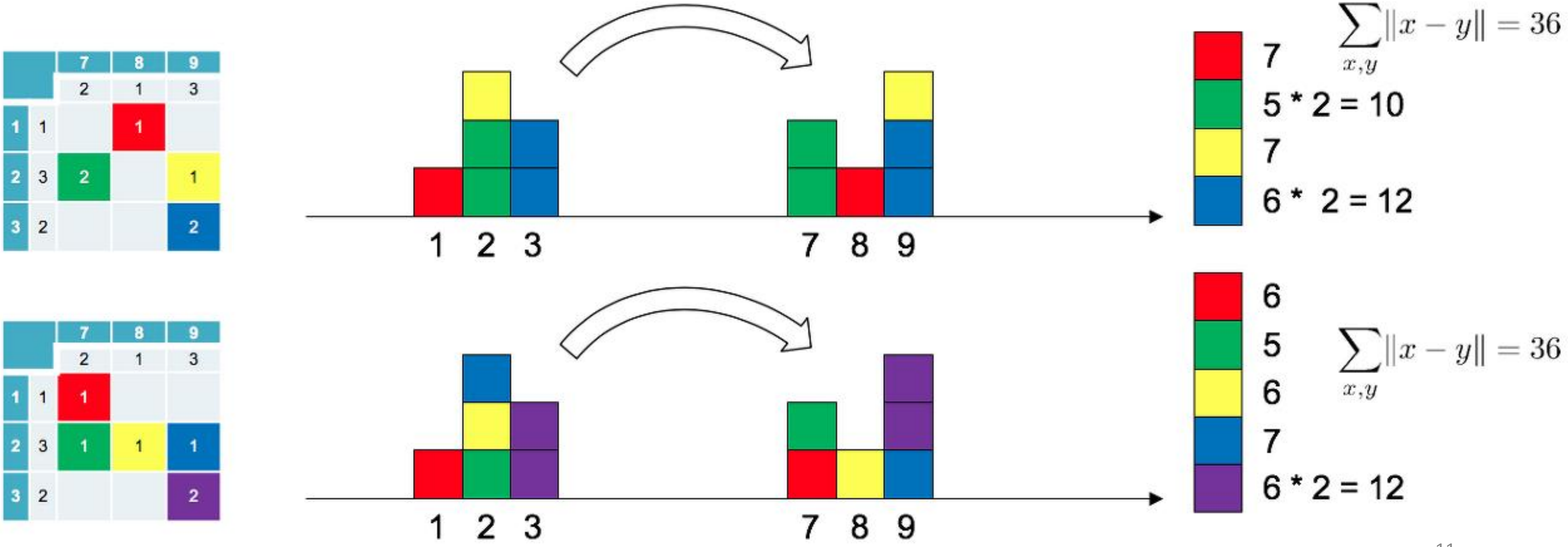
https://haawron.tistory.com/21

Background: Wasserstein Distance

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

2. Why Wasserstein distance?

- e.g.) Earth-Mover distance



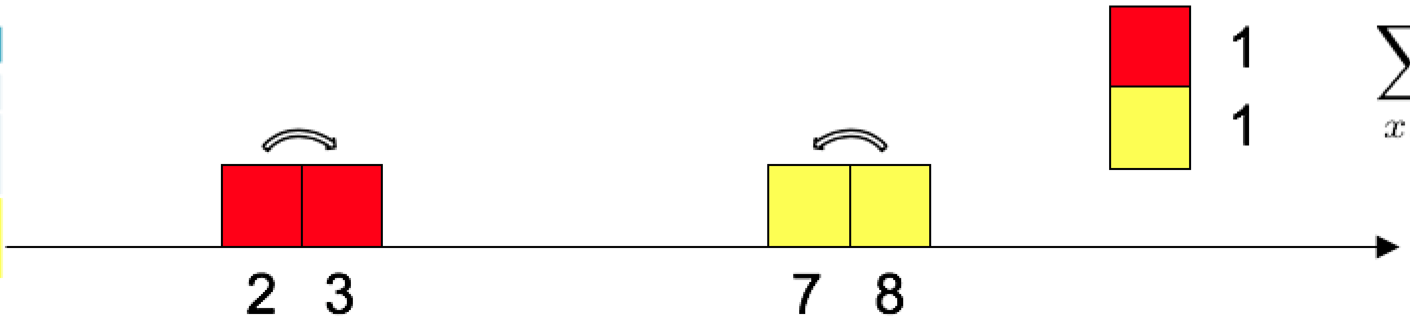
Background: Wasserstein Distance

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

2. Why Wasserstein distance?

- e.g.) Earth-Mover distance

		3	7
		1	1
2	1	1	
8	1		1



$$\sum_{x,y} \|x - y\| = 2$$

		3	7
		1	1
2	1	1	
8	1		1

		3	7
		0.5	0.5
2	0.5	0.5	
8	0.5		0.5

https://haawron.tistory.com/21

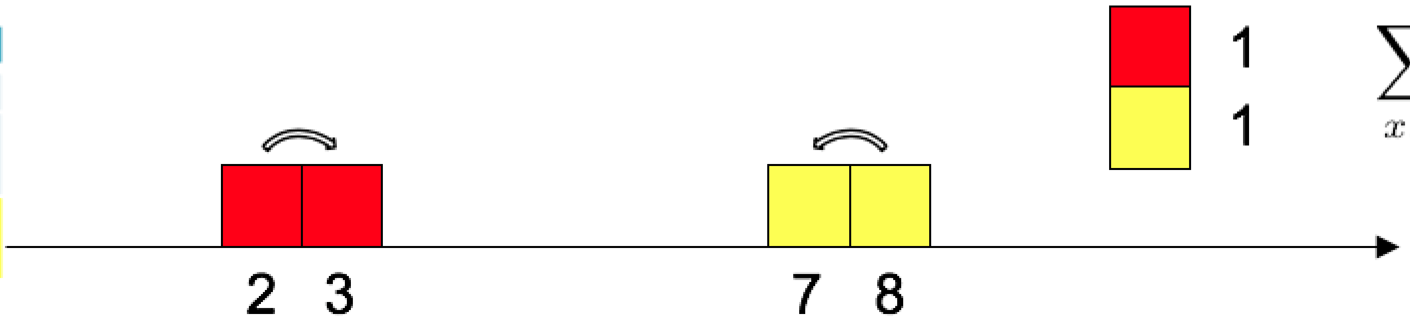
Background: Wasserstein Distance

2. Why Wasserstein distance?

- e.g.) Earth-Mover distance

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

	3	7
	1	1
2 1	1	
8 1		1



$$\sum_{x,y} \|x - y\| = 2$$

	3	7
	1	1
2 1	1	
8 1		1

	3	7
	0.5	0.5
2 0.5	0.5	
8 0.5		0.5

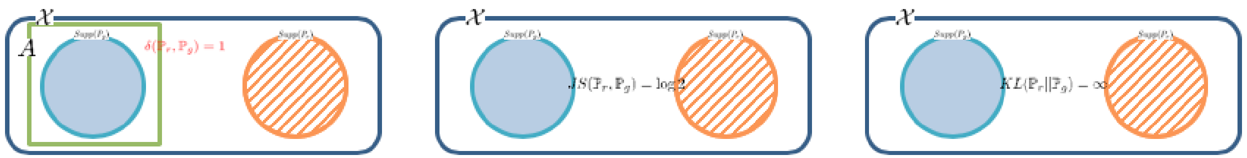
$$\begin{aligned} W(\mathbb{P}_r, \mathbb{P}_g) &= \gamma X, Y(2,3) \times |2 - 3| \\ &+ \gamma X, Y(8,7) \times |8 - 7| \\ &= 0.5 + 0.5 \end{aligned}$$

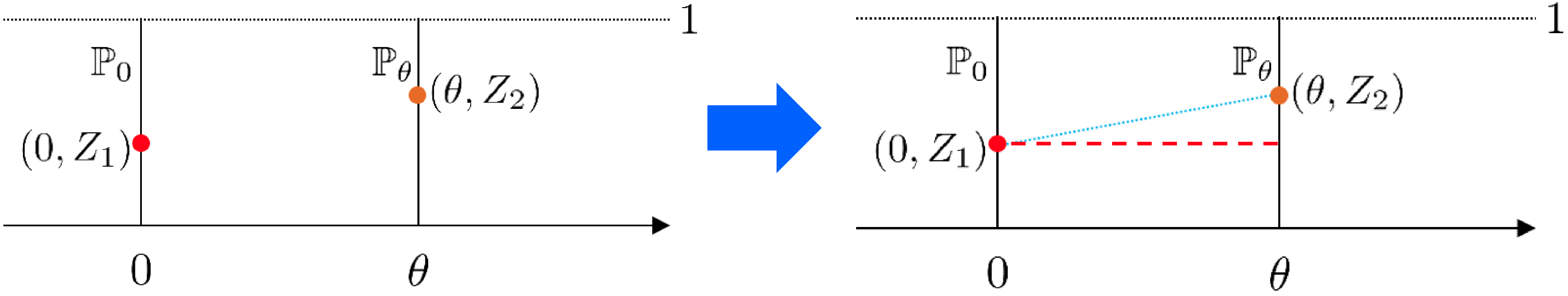
https://haawron.tistory.com/21

Background: Wasserstein Distance

2. Why Wasserstein distance?

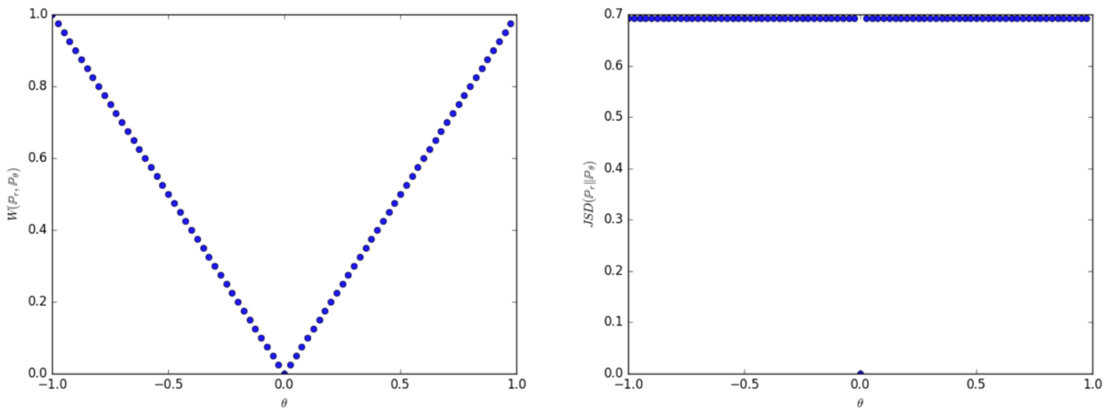
- e.g.) Learning parallel lines





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

$$\boxed{W(\mathbb{P}_r, \mathbb{P}_g) = |\theta|}$$



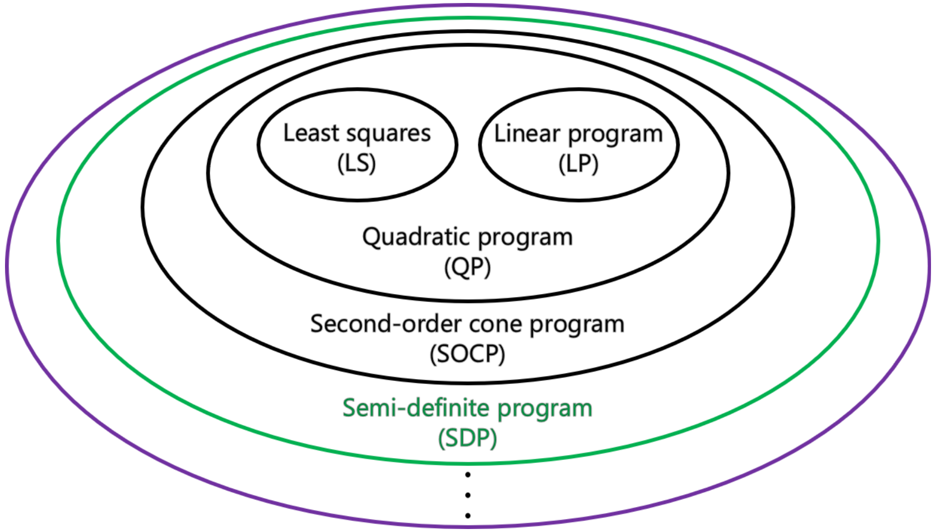
Background: SDP Relaxation

3. Semi-Definite Program (SDP)

1) Why SDP?

- A class of very difficult non-convex problems → approximation using SDP relaxation
- Finding maximum eigenvalue or minimizing nuclear norm

, where Nuclear norm $||A||_* \coloneqq \sum_i \sigma_i(A)$ and $\sigma_i(A)$ is i-th singular value of A

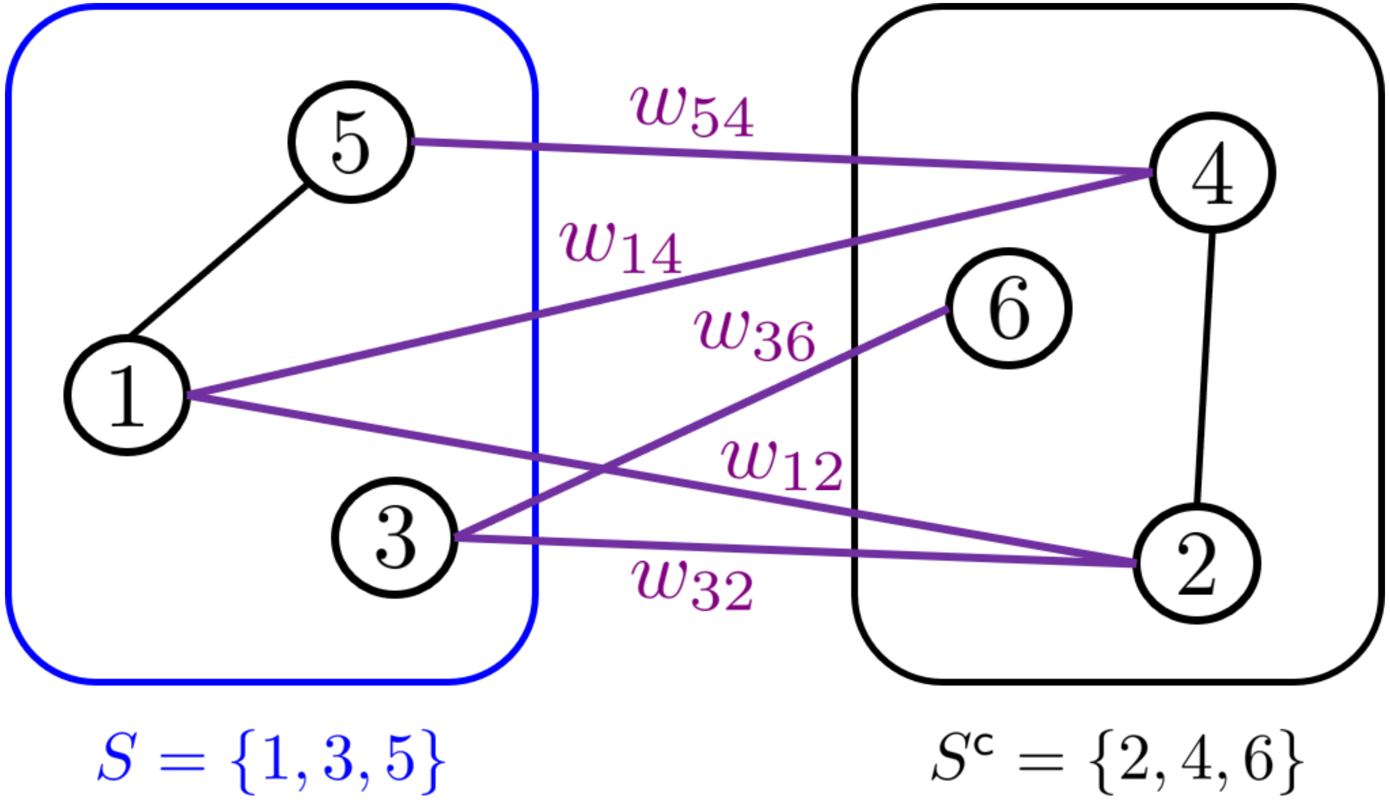


Background: SDP Relaxation

3. Semi-Definite Program (SDP)

2) A simple example: Maximum cut problem

- Maximum cut problem: Finding set that can maximize cut



- Set S: subset of set $\mathcal{V}$ (vertex)
- Cut: Sum of weights of edges

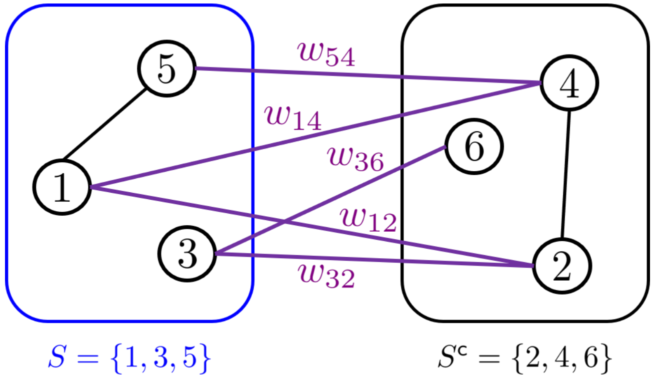
e.g.)

$S = \{1,3,5\} \subset \mathcal{V}$

$$Cut(S) = \omega_{54} + \omega_{14} + \omega_{36} + \omega_{12} + \omega_{32}$$

Background: SDP Relaxation

3. Semi-Definite Program (SDP)



2) A simple example: Maximum cut problem

- x_i denotes whether node i is in the set S:

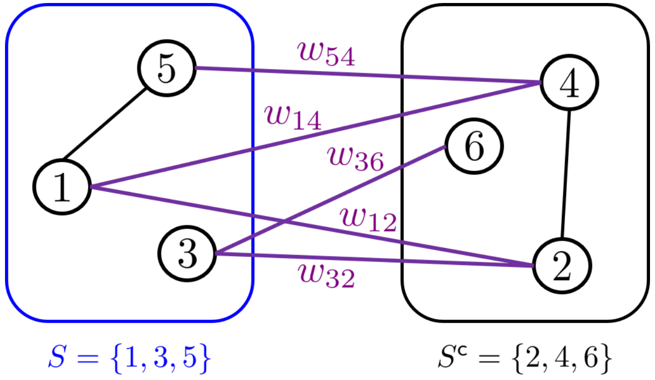
$$x_i = \begin{cases} +1 & if\ x \in S \\ -1 & Otherwise \end{cases}$$

- Maximum cut via optimization:

$$\max_{x_i} \sum_{i,j} \frac{1}{2} w_{ij}(1 - x_i x_j) : \ x_i^2 = 1 \ (i = 1, \dots, d)$$

Background: SDP Relaxation

3. Semi-Definite Program (SDP)



2) A simple example: Maximum cut problem

- x_i denotes whether node i is in the set S:

$$x_i = \begin{cases} +1 & if\ x \in S \\ -1 & Otherwise \end{cases}$$

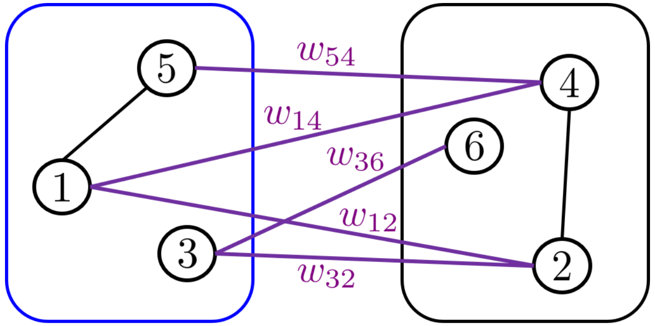
- Maximum cut via optimization:

$$\max_{x_i} \sum_{i,j} \boxed{\frac{1}{2}} w_{ij}(1 - x_i x_j) : \boxed{x_i^2 = 1} \ (i = 1, \ldots, d)$$

We want just ω_{ij}

x_i is +1 or -1

Background: SDP Relaxation



3. Semi-Definite Program (SDP)

2) A simple example: Maximum cut problem

- A simple and effective technique: Lifting
- It Lifts optimization variable space (i.e., vector to matrix):

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{d1} & \cdots & x_{dd} \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix} = xx^T$$

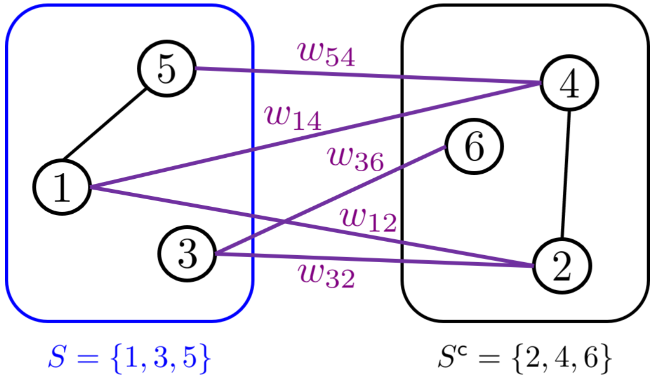
, where $\boxed{X_{ii} = 1, X \succcurlyeq 0, rank(X) = 1}$

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

Background: SDP Relaxation

3. Semi-Definite Program (SDP)

2) 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,

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

Background

4. Summary

1) Why This paper?
2) Wasserstein Distance
3) SDP Relaxation

Outline

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

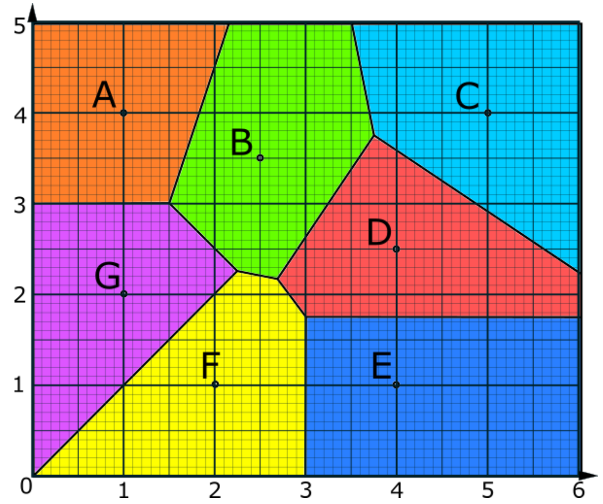
$$\min_{\beta_1,\ldots,\beta_K \in \mathbb{R}^d} \sum_{i=1}^{n} \min_{k\in[K]} \|X_i - \beta_k\|_2^2 = \min_{G_1,\ldots,G_K} \left\{ \sum_{k=1}^{K} \sum_{i \in G_k} \|X_i - \bar{X}_k\|_2^2 : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

- 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

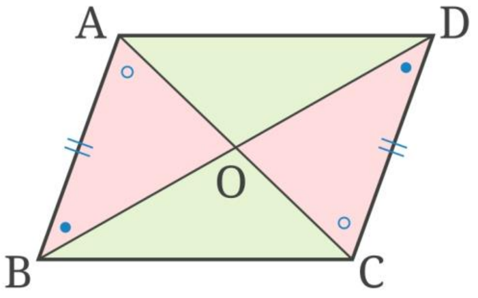
4) Actually, there are two kinds of K-means clustering: Centroid-based formulation and Distance-based formulation

- Distance-based formulation:

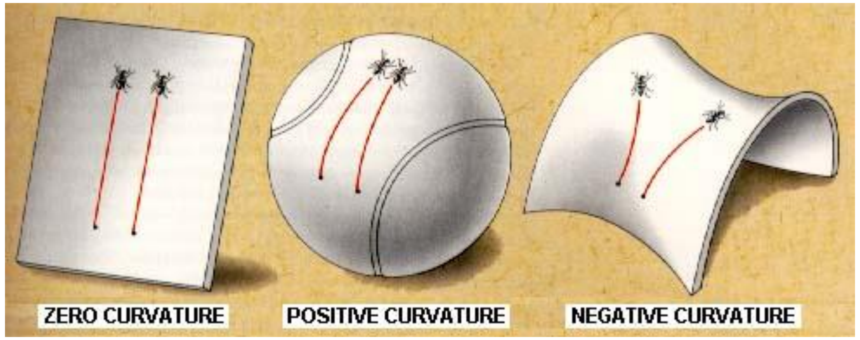
$$\min_{G_1,\ldots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j\in G_k} \|X_i - X_j\|_2^2 : \bigsqcup_{k=1}^{K} G_k = [n] \right\}$$

- 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

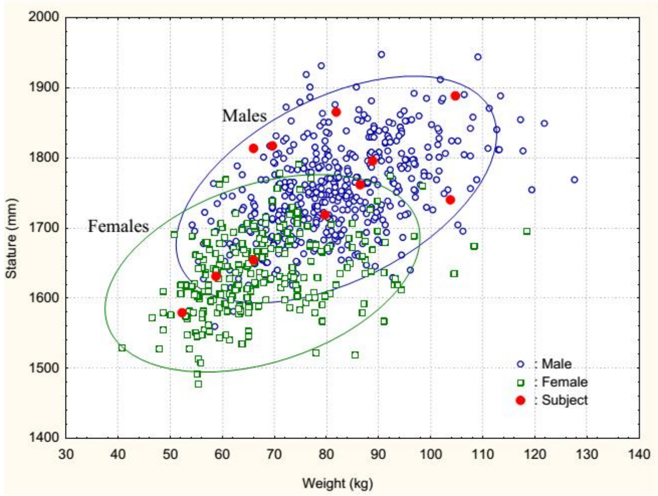


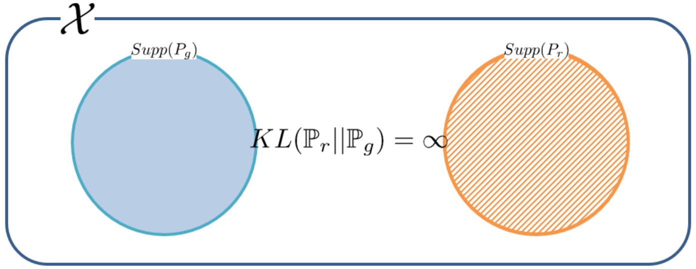
1. Detour: K-means clustering

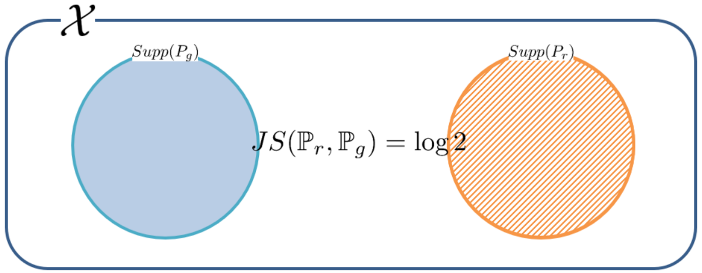
5) K-means clustering for Euclidean space

- 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:

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

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

Background

- Next Week

1) Method
   - Pitfalls of Barycenter-based clustering: Irregularity and Non-robustness
   - Failure of centroid-based Wasserstein K-means
   - Pairwise distance-based clustering (D-WKW)
   - SDP Relaxation (W-SDP)
2) Experiments
   - Counter-example
   - Exact recovery for clustering Gaussian
   - Real-data application (MNIST, Fashion-MNIST, USPS)
3) Discussion
   - Time Cost
   - Read-data application?

README.md

Others

- Singular Value Decomposition
- The Geometric Meaning of Covariance
- Operator norm calculation for simple matrix
- CUTOFF FOR EXACT RECOVERY OF GAUSSIAN MIXTURE MODELS
- Riemannian Manifold
- Is a sample covariance matrix always symmetric and positive definite?
- Matrix Decompositions
- Cholesky decomposition
- The hardness of k-means clustering in the plane
- Linear Programming
- Slack variable
- Concave and convex functions of a single variable
- Positive-Definite
- Rank
- SDP relaxation
- Quantile
- Quantile Function
- Pushfoward
- Diffeomorphism
- Density matrix
- Quantum state
- Bures metric
- ON THE BURES-WASSERSTEIN DISTANCE BETWEEN POSITIVE DEFINITE MATRICES
- Lloyd's algorithm
- Parallelogram law
- Optimal Transport
- Alexandrov space
- Alexandrov curvature
- A generalization of the parallelogram law to higher dimensions
- t-SNE
- Silhouette score
- Hyperopt
- t-SNE vs UMAP
- DBSCAN
- Wasserstein GAN
- Wasserstein distance
- Why Wasserstein distance?
- Why Wasserstein is indeed weak
- KL-Divergence
- Wasserstein Barycenter Applied to K-Means Clustering
- metric

To be continued…

https://jeiyoon.github.io/