Wasserstein k-means with sparse simplex projection

Takumi Fukunaga[1] Hiroyuki Kasai [2,3]

[1]Dept. of Computer Science and Engineering,School of Fundamental Science and Engineering,Waseda University,Japan

[2]Dept. of Communications and Computer Engineering, School of Fundamental Science and Engineering, Waseda University,Japan

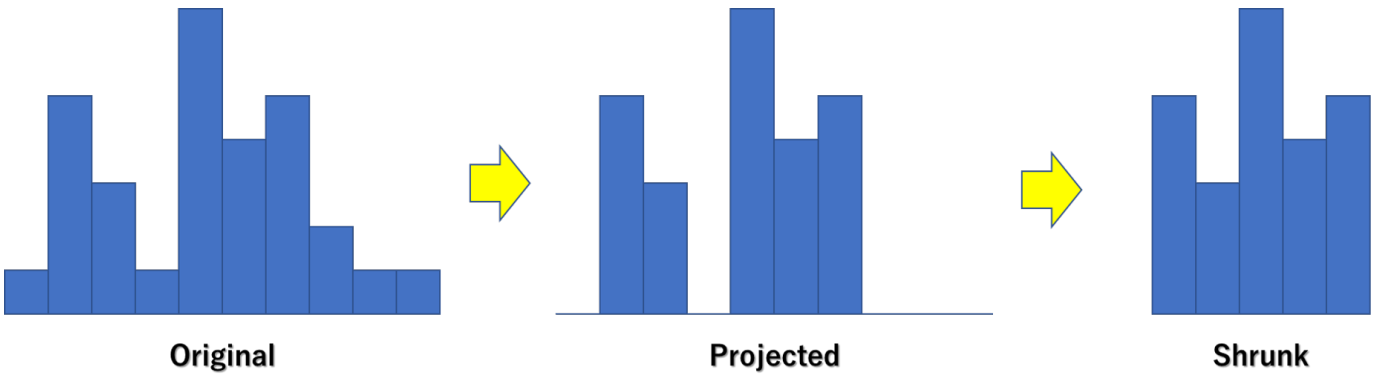
[3]Dept. of Computer Science and Communications Engineering, Graduate School of Fundamental Science and Engineering, Waseda University,Japan

December 9,2020

Introduction

Wasserstein k-means

- Propose
  - Improvement of computational cost of Wasserstein k-means
- Fast and efficient approaches
  - [Cuturi and Doucet, 2014, Bonneel et al., 2015, Anderes et al., 2016]
- Contribution of our methods
  - Sparsifying data on probability simplex and shrinking them by removing the zero elements



Introduction

Clustering Algorithm and its issue

- k-means [Lloyd, 1982]
  - High computational cost per iteration $\mathcal{O}(qk)$
    - Efficient approaches
      [Arthur and Vassilvitskii, 2007, Kanungo et al., 2002]
- Wasserstein k-means [Ye et al., 2017]
  - Adopting Wasserstein distance and Wasserstein barycenter
    - High computational cost to caluculate its distance
      $\mathcal{O}(n^3 \log n)$[Cuturi, 2013, Rubner et al., 2000]

assignment step

$$s_i = \underset{j=1,\ldots,k}{\arg\min}\, d(\boldsymbol{x}_i, \boldsymbol{c}_j), \forall i \in [q]$$

update step

$$\boldsymbol{c}_j = \mathrm{mean}(\{\boldsymbol{x} | s_i = j\}) \quad \text{or} \quad \mathrm{barycenter}(\{\boldsymbol{x} | s_i = j\}), \forall j \in [k]$$

Introduction

Optimal Transport

- Caluculate the minimum transport cost [Peyre and Cuturi, 2019]
- When $n = m$, its cost is called Wasserstein distance of order p
- Using this distance, calculate Wasserstein barycenter [Benamou et al., 2015]

$$\begin{aligned} \mathbf{T}^* &= \underset{\mathbf{T}\in\mathcal{U}_{mn}}{\arg\min}\langle \mathbf{T}, \mathbf{C}\rangle \\ W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) &= \min_{\mathbf{T}\in\mathcal{U}_{mn}} \langle \mathbf{T}, \mathbf{C}\rangle = \langle \mathbf{T}^*, \mathbf{C}\rangle \\ g(\boldsymbol{\mu}) &= \frac{1}{n}\sum_i W_p(\boldsymbol{\mu}, \boldsymbol{\nu}_i) \end{aligned}$$

- $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ of points
- $\boldsymbol{a}$ and $\boldsymbol{b}$ are in probability simplex
- $\mathbf{C}$ is ground matrix
- Row and Colunmn marginal constraint

$$\mathcal{U}_{mn} = \{\mathbf{T} \in \mathbb{R}_+^{m\times n} : \mathbf{T}\mathbf{1}_n = \boldsymbol{a}, \mathbf{T}^T\mathbf{1}_m = \boldsymbol{b}\}$$

Proposal

Motivation

- Reduce the size of data and centroid
- Adopt two following approaches

1. Sparsify datas
   - Make data sparser than the original ones
   - Maintain degradation of the clustering quality as small as possible
2. Shrink datas
   - No degradation
   - Key operator to reduce the computational complexities

Proposal

Basic idea - Sparse simplex projection

- Sparse simplex projection GSHP [Kyrillidis et al., 2013]

$$\hat{\boldsymbol{\beta}} \quad = \quad \mathrm{Proj}^{\gamma(t)}(\boldsymbol{\beta}) = \begin{cases} \hat{\boldsymbol{\beta}}_{|\mathcal{S}^\star} & = \mathcal{P}_{\Delta_\kappa}(\boldsymbol{\beta}_{|\mathcal{S}^\star}) \\ \hat{\boldsymbol{\beta}}_{|(\mathcal{S}^\star)^c} & = 0, \end{cases}$$

- $\mathcal{S}$ is the subset of $\mathcal{N} = \{1, \ldots, n\}$
- $\boldsymbol{a}_{|\mathcal{S}}$ extracts the elements of $\mathcal{S}$ in $\boldsymbol{a}$
- $(\mathcal{P}_{\Delta_\kappa}(\boldsymbol{\beta}_{|\mathcal{S}^*}))_v = [(\boldsymbol{\beta}_{|\mathcal{S}^*})_v + \tau]_+, \tau := \frac{1}{\kappa}(1 + \sum^{|\mathcal{S}^*|} \boldsymbol{\beta}_{|\mathcal{S}^*})$
- $\mathcal{S}^* = \mathrm{supp}(\mathcal{P}_{\Delta_\kappa})$
  - $\mathrm{supp}(\boldsymbol{a}) = \{i : \boldsymbol{a}_i \neq 0\}$

Proposal

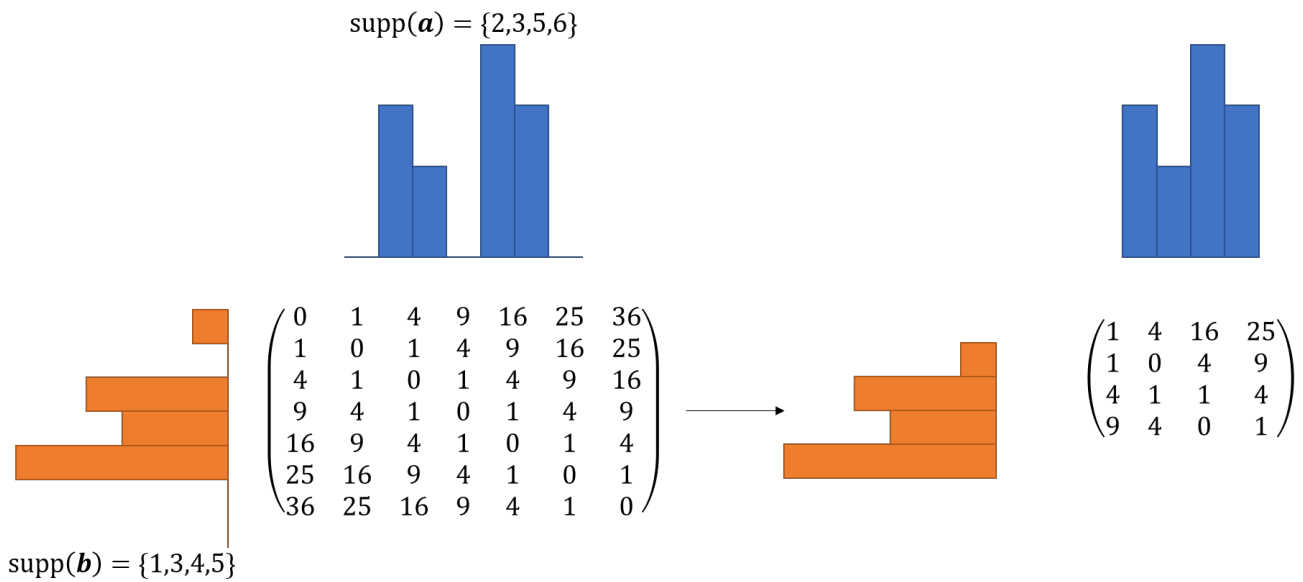
Basic idea - Shrinking datas

- The zero elements don't have effect on transport matrix
  - Removing the zero elements
- Define shrinking operator to vector and matrix

$$\tilde{\boldsymbol{\nu}}_i = \text{shrink}(\hat{\boldsymbol{\nu}}_i) = (\hat{\boldsymbol{\nu}}_i)_{|\mathcal{S}_{\text{samp}}} \in \mathbb{R}^{|\mathcal{S}_{\text{samp}}|}$$

$$\tilde{\boldsymbol{c}}_i = \text{shrink}(\hat{\boldsymbol{c}}_i) = (\hat{\boldsymbol{c}}_i)_{|\mathcal{S}_{\text{cent}}} \in \mathbb{R}^{|\mathcal{S}_{\text{cent}}|}$$

$$\tilde{\mathbf{C}} = \text{Shrink}(\mathbf{C}_{\boldsymbol{\nu}\boldsymbol{c}}) = \mathbf{C}_{\text{supp}(\hat{\boldsymbol{\nu}}_i),\text{supp}(\hat{\boldsymbol{c}}_i)} \in \mathbb{R}^{|\mathcal{S}_{\text{samp}}| \times |\mathcal{S}_{\text{cent}}|}$$



Proposal

Basic procedure

1. Update sparsity ratio $\gamma(t)$
2. Project $\boldsymbol{\nu}_i$ into $\hat{\boldsymbol{\nu}}_i$ and shrink $\hat{\boldsymbol{\nu}}_i$ into $\tilde{\boldsymbol{\nu}}_i$
3. Project $\boldsymbol{c}_j$ into $\hat{\boldsymbol{c}}_j$ and shrink $\hat{\boldsymbol{c}}_j$ into $\tilde{\boldsymbol{c}}_i$
4. Shrink ground cost matrix $\mathbf{C}$ into $\tilde{\mathbf{C}}$
5. Find closest centroids and update centroids
6. Unless cluster centroids stop changing, repeat step1

- Control parameter of sparse ratio $\gamma(t)$

$$\gamma(t) \quad := \quad \begin{cases} \gamma_{\min} & \text{(FIX)} \\ 1 - \dfrac{(1-\gamma_{\min})}{T_{\max}} t & \text{(DEC)} \\ \gamma_{\min} + \dfrac{(1-\gamma_{\min})}{T_{\max}} t & \text{(INC),} \end{cases}$$

- $T_{\max}$ is maximum iteration of k-means

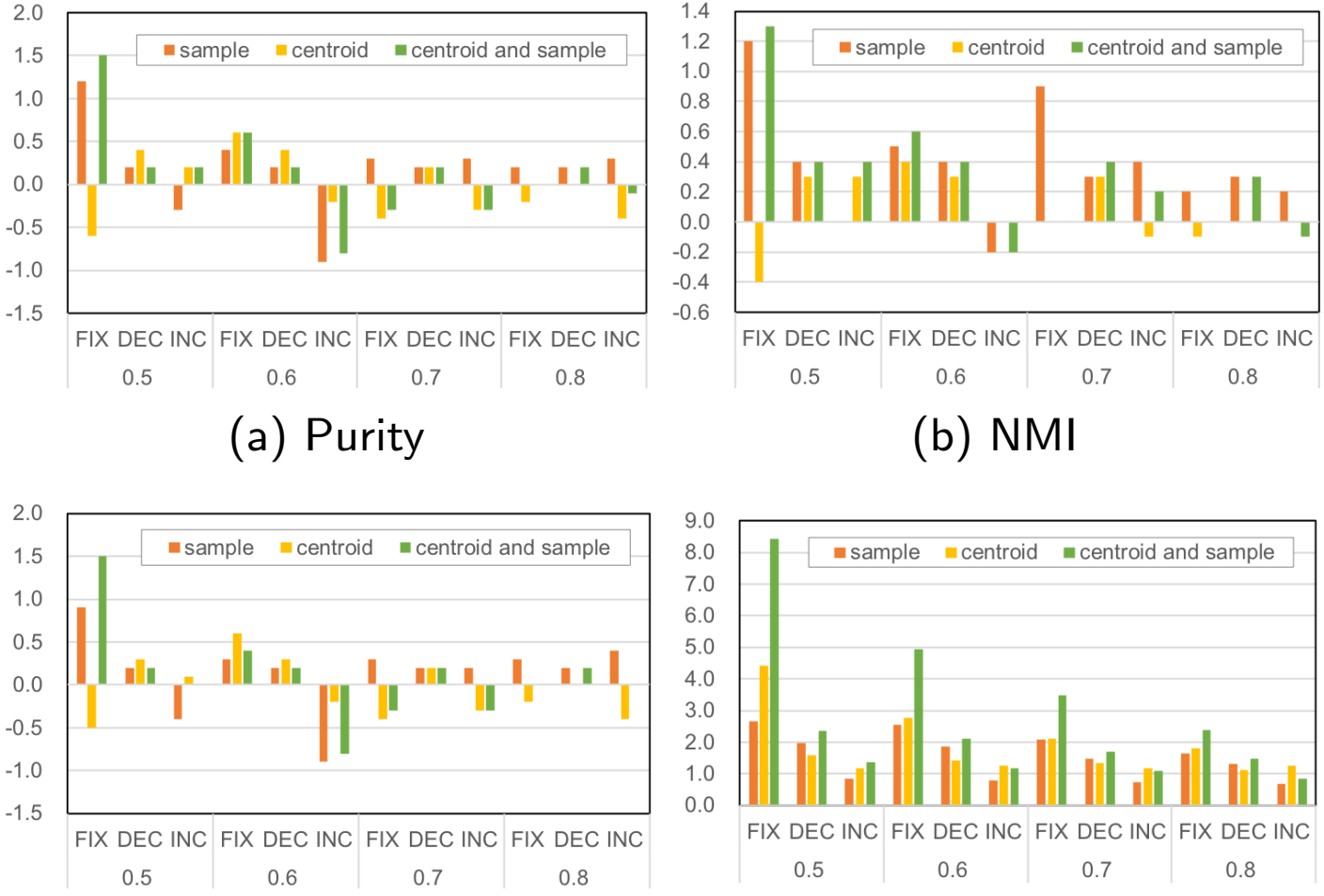
Experiment

Settings

- Use of Algorithm
  - Wasserstein barycenter [Cuturi and Doucet, 2014]
  - k-means with **litekmeans**
  - **linprog** of Mosek to solve LP [Andersen et al., 2000]
- Datasets
  - COIL-100 [Nene et al., 1996]
  - the USPS handwritten dataset

Experiment

2-D histogram evaluation



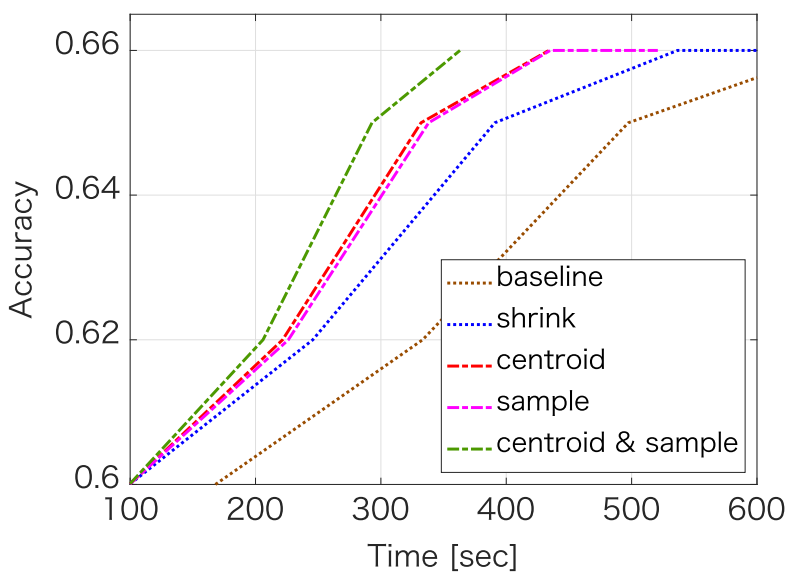
(a) Purity (b) NMI

(c) Accuracy (d) Computation time

Figure: Performance results of 2-D histogram data on the USPS dataset.

Experiment

Convergence Performance



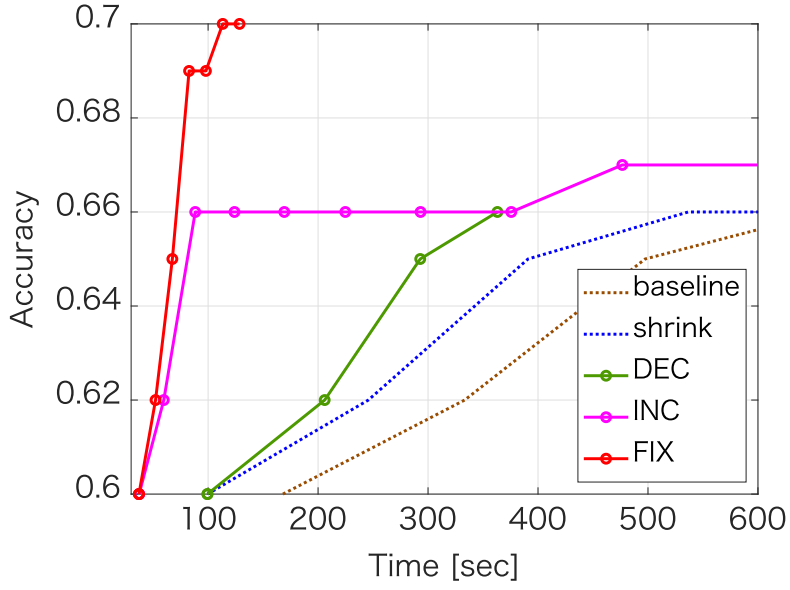


Figure: Left : Convergence performance with different projection data using DEC algorithm of $\gamma_{\min} = 0.5$ Right:Convergence performance comparison of different algorithm of $\gamma(t)$ of $\gamma_{\min} = 0.5$

Experiment

Comparoson on different sparsity

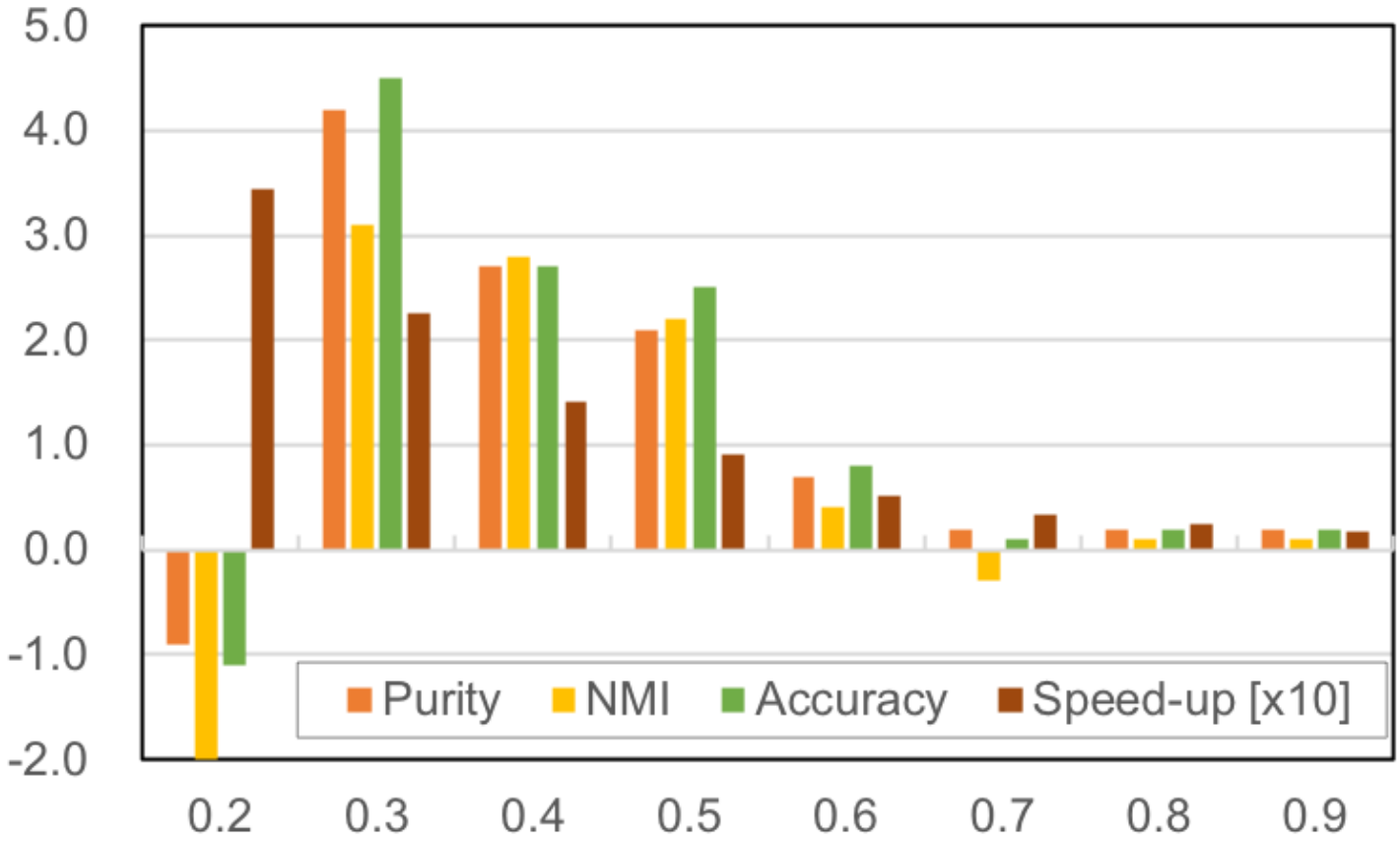


Figure: Performance comparison on different ratios on the USPS dataset.

Experiment

Conclusion

- We propose a faster Wasserstein k-means algorithm
- Our experiments demonstrate the effectiveness of our method
  - Reducing the computational complexity of Wasserstein distance
  - Keeping accuracy before sparsifying and shrinking

References I

- Anderes, E., Borgwardt, S., and Miller, J. (2016).
  Discrete wasserstein barycenters: optimal transport for discrete data.
  Mathematical Methods of Operations Research, 84:389–409.
- Andersen, E. D., Roos, C., and Terlaky, T. (2000).
  High Performance Optimization, volume 33, chapter The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm.
  Springer, Boston, MA.
- Arthur, D. and Vassilvitskii, S. (2007).
  K-means++: The advantages of careful seeding.
  In Proc. 18th Ann. ACM-SIAM Symp. Discr. Alg. (SODA).
- Benamou, J. D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015).
  Iterative bregman projections for regularized transportation problems.
  SIAM Journal on Scientific Computing,, 37(2):1111–A1138.

References II

- Bonneel, N., Rabin, J., and Peyré, G. (2015).
  Sliced and radon wasserstein barycenters of measures.
  Journal of Mathematical Imaging and Vision, 51:22–45.
- Cuturi, M. (2013).
  Sinkhorn distances: Lightspeed computation of optimal transport.
  In Annual Conference on Neural Information Processing Systems (NIPS).
- Cuturi, M. and Doucet, A. (2014).
  Fast computation of wasserstein barycenters.
  In ICML.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002).
  An efficient k-means clustering algorithm: analysis and implementation.
  IEEE Trans. Pattern Anal. Mach. Intell., 24(7):881–892.

References III

- Kyrillidis, A., Becker, S., Cevher, V., and Koch, C. (2013).
  Sparse projections onto the simplex.
  In ICML.
- Lloyd, S. (1982).
  Least squaares quantization in pcm.
  IEEE Trans. Inf. Theory, 28(2):129–137.
- Nene, S. A., Nayar, S. K., and Murase, H. (1996).
  Columbia university image library (COIL-20).
  Technical Report CUCS-005-96.
- Peyre, G. and Cuturi, M. (2019).
  Computational optimal transport.
  Foundations and Trends in Machine Learning, 11(5-6):355–607.
- Rubner, Y., Tomasi, C., and Guibas, L. J. (2000).
  The earth mover's distance as a metric for image retrieval.
  International Journal of Computer Vision, 40(2):99–121.

References IV

- Ye, Y., Wu, P., Wang, J. Z., and Li, J. (2017).
  Fast discrete distribution clustering using wasserstein barycenter with sparse support.
  IEEE Trans. Signal Process, 65(9):2317–2332.

Thank you for listening.

WASEDA University