Relax and Merge: A Simple Yet Effective Framework for Solving Fair k-means and k-sparse Wasserstein Barycenter Problems

Shihong Song Guanlin Mo Hu Ding

shihongsong@mail.ustc.edu.cn
University of Science and Technology of China

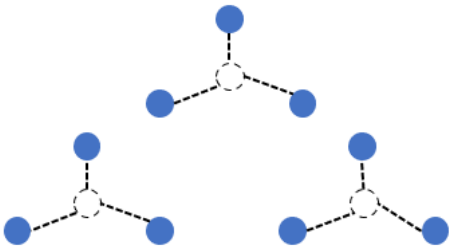
The Thirteenth International Conference on Learning Representations

k-means Objective

Problem (Vanilla Euclidean (Continues) k-means)

Input: a set of clients (data points) $P = \{p_1, p_2, \cdots, p_n\}$ in Euclidean space $\mathbb{R}^d$.
Objective: to find a (facility) set $S \subseteq \mathbb{R}^d$ with $|S| = k$ and an assignment map $\sigma : P \to S$ s.t.

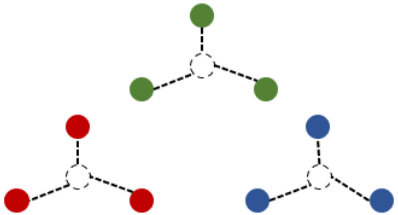
$$\min_{S,\sigma} \text{Cost}(S, \sigma) := \sum_{i=1}^{n} \|p_i - \sigma(p_i)\|^2 \tag{1}$$



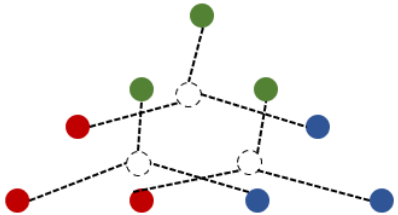
Fair k-means

- Suppose every client is **colored**, the fairness constraint requires the **balance** of colors in each cluster.
- Formally, if $P^{(i)}$ is the client set of i-th colored group, $C_j \leftarrow \{p|\sigma(p) = s_j\}$ and m is the number of colors, then the objective is

$$\min_{S,\sigma} \text{Cost}(S, \sigma) = \sum_{i=1}^{n} \|p_i - \sigma(p_i)\|^2 \tag{2}$$
$$s.t. \beta_i |C_j| \leq |C_j \cap P^{(i)}| \leq \alpha_i |C_j| \quad \forall i \in [m], \forall j \in [k],$$



Unfair



Fair

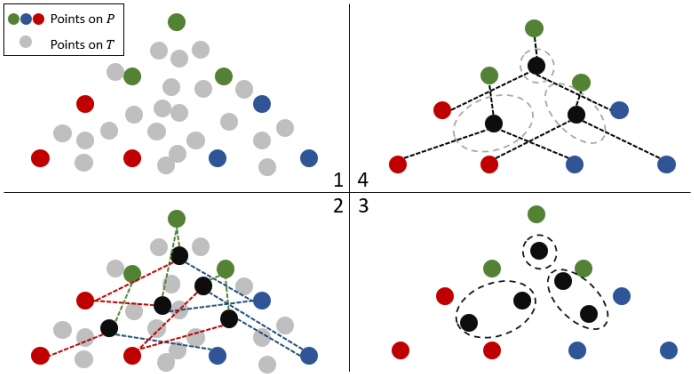
Outline of Our algorithm

We split the algorithms into three phases:

1. **finding the facility locations**;
2. calculating the fractional assignment (linear programming);
3. **rounding the fractional assignment to integral**.

Relax and Merge

- Firstly, we relax the constraint of k facilities, *i.e.*, we allow more than k facilities to be chosen. The chosen point set is denoted by T.
- Secondly, we use linear programming to calculate an assignment of T.
- Thirdly, we use vanilla k-means algorithm on T to merge the solution to k facilities.

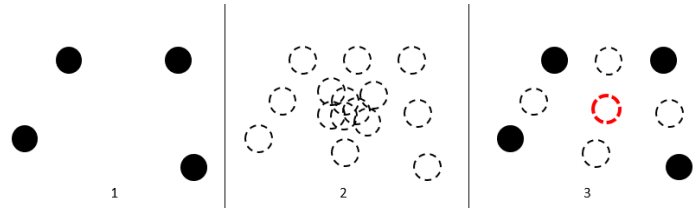


Construction of T

How to construct T?

- Ideally, we hope T could cover all the centroid of the subsets of clients.
- However, the number of subsets of clients could be large as 2^n.
- By introducing some small error factor $(1+\epsilon)$, the size of centroid set could be reduced to $O(n\epsilon^{-d}\log(1/\epsilon))$[1].

We use $(1+\epsilon)-$approximate centroid set as the set T.



[1] Jiří Matoušek. "On approximate geometric k-clustering". In: *Discrete & Computational Geometry* 24.1 (2000), pp. 61–84.

Theoretical Guarantee

Claim

"Relax and Merge" could return a feasible fractional solution for fair k-means.

Theorem (Theorem 1 (informal))

Given an instance of fair k-means and a ρ-approximate vanilla k-means clustering algorithm, there exists an algorithm that can return a fractional solution whose cost is at most $(1 + 4\rho + \epsilon)OPT$.

Theorem (2(informal))

"Relax and Merge" algorithm returns $(1 + 4\rho + \epsilon)$-approximate solution for k-sparse Wasserstein Barycenter problem.

Rounding Algorithm

Note

If we know the upper and lower bound in every cluster of each colored group, then we can use Minimum-Cost Circulation Flow algorithm to obtain the integral solution[a] to solve integral assignment perfectly!

[a] Jack Edmonds and Richard M Karp. "Theoretical improvements in algorithmic efficiency for network flow problems". In: *Journal of the ACM (JACM)* 19.2 (1972), pp. 248–264.

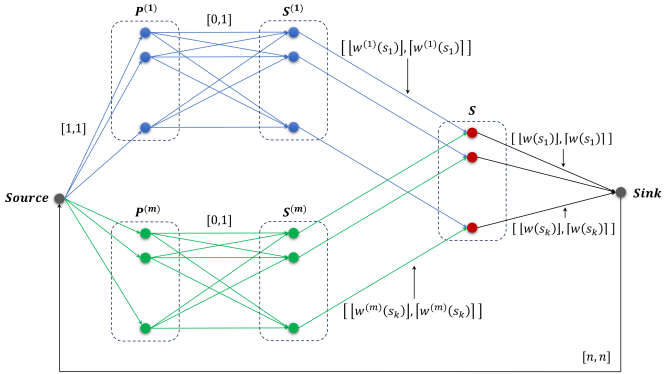
The main idea of rounding:

1. Use linear programming to estimate the size of every cluster, *e.g.*, $|\tilde{C}_j|$.
2. For each cluster C_j, the size bounds for q-th colored client group is set to $\lfloor \beta_q |\tilde{C}_j|, \alpha_q |\tilde{C}_j| \rceil$ approximately.
3. Use Minimum-Cost Circulation Flow algorithm to obtain the integral solution.

Rounding Algorithm

Lemma (4)

There exists an algorithm that can round a fractional solution of (α, β)-fair k-means to integral with at most 2-violation while the cost does not increase.



Strictly Fair k-means

- When $\alpha_i = \beta_i = \frac{|P^{(i)}|}{|P|}$, we call it **strictly** fair k-means.
- We assume each colored group has the same size, otherwise the instance is infeasible.

High Level idea:

1. Decompose the dataset to **fairlets** (fairlet is a collection composed of points, with each point being of a unique color, and the collection encompasses all colors).
2. Run vanilla k-means for fairlets.

Theorem (3)

Our algorithm returns a $(2 + 6\rho)$-approximate integral solution of strictly fair k-means.

Fairlet and Fairlet Decomposition

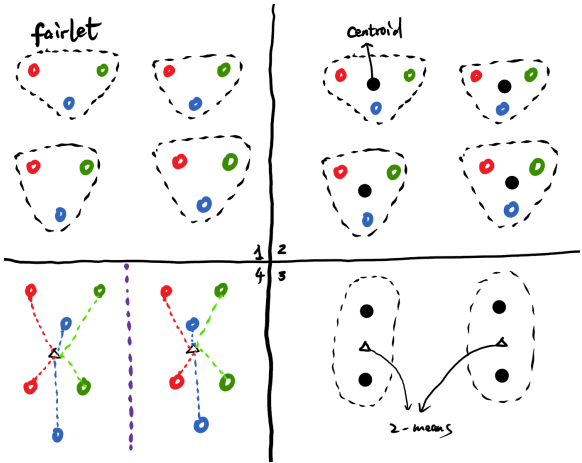


Figure: An example of strictly Fair 2-means

The End

Thanks.