Coreset for Ordered Weighted Clustering

Vladimir Braverman[1], Shaofeng H.-C. Jiang[2], Robert Krauthgamer[2], Xuan Wu[1]

[1]CS Department, Johns Hopkins University
[2]Weizmann Institute of Science
*All authors contribute equally to this work.

Key Word: Data-Reduction, OWA Framework, Ordered k-median, Simultaneous Core-set

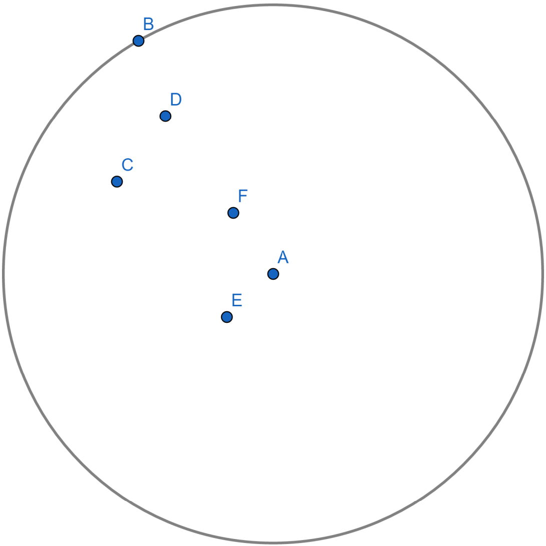
The Ordered k-Median Clustering

Let $X \subset \mathbb{R}^d$ be your data set.

k-center, k-median, and p-centrum

- k-center: $\min_{C \subset \mathbb{R}^d : |C| = k} \max_{x \in X} d(x, C)$.
- k-median: $\min_{C \subset \mathbb{R}^d : |C| = k} \sum_{x \in X} d(x, C)$.
- k-facility p-centrum: cost function is defined by the largest p connection cost.
- 1-centrum$=$ k-center
- n-centrum$=$ k-median.

k-center: $\{B\}$, k-median:$\{B, C, D, E, F\}$, 3-centrum: $\{B, C, D\}$.



The Ordered k-median Clustering

- Given a non-increasing weight vector $v \in \mathbb{R}_+^n$. Sort the data points by, $d(x_1, C) \geq ... \geq d(x_n, C)$
- $\min_{C \subset \mathbb{R}^d} \mathrm{cost}_v(X, C)$ where $\mathrm{cost}_v(X, C) := \sum_{i=1}^n v_i d(x_i, C)$.
- p-centrum Problem: $v = (1, ..., 1, 0, ..., 0)$.

Coreset and Simultaneous Coreset

Coreset

A weighted set D (with weight w) is called an (strong) ε-coreset of X for k-clustering problem (for a specific objective cost) if $\forall C \subset \mathbb{R}^d, |C| = k, \mathrm{cost}(D, C) \in (1 \pm \varepsilon)\mathrm{cost}(X, C)$.

Simultaneous Coreset

- Ordered k-median has multiple objectives, namely, cost_v for different v.
- Want to approximate them all.
- $\mathrm{cost}_v(D, C) \in (1 \pm \varepsilon)\mathrm{cost}_v(X, C)$ for every C and v.

Results

Upper Bounds

- Thm 1: We can construct Coreset for p-Centrum (for specific p) of size $O(\frac{k^2}{\varepsilon^{d+1}})$ efficiently.
- Thm 2: We can construct simultaneous Coreset for ordered k-median of size $O(\frac{k^2 \log^2 n}{\varepsilon^d})$ efficiently. This is the first simultaneous coreset for ordered weighted clustering.

Nearly Matching Lower Bound

- Thm 3:There is a constant c, s.t., c-Simultaneous coreset for ordered k-median problem has a size lower bound $\Omega(\log n)$.
- Previously Known Fact: $\Omega(\frac{1}{\varepsilon^d})$ is a lower bound of coreset size even for k-center problem.

Applications

- One coreset, multiple objectives.
- Can adjust the objective and optimize w.r.t it easily, via our coreset.

Thank you!

Future Work

- Closing the size bound gap for simultaneous coreset.
- Deriving lower bound when the objective is a specific v (depend on v).
- Study other objectives where similar coreset construction is useful.

Appendix

The Basic Case: p-Centrum Problem for $k = d = 1$

- Compute the optimal center c.
- Let $L \cup R$ be points contributed to $\mathrm{cost}_p(X, c)$, where L is left to c and R is right to c.
- Let $Q = X \setminus (L \cup R)$ denote the remaining points.
- Observation: $\max_{q \in Q} d(q, c) \leq \frac{1}{p}\mathrm{cost}_p(X, c)$.
- Partition L and R into buckets of small cumulative error $O(\varepsilon \mathrm{opt})$ (k-Median Part)
- Partition Q into buckets of small length $O(\varepsilon \mathrm{opt}/p)$.
- Pick D to be the mean of each bucket.

Moving to Simultaneous Coreset and High Dimension

Observation

- Although there are infinitely many possible weight, we only need to be simultaneous coreset for $O(\frac{\log n}{\varepsilon})$ many p-centrum problems in order to obtain simultaneous coreset.
- Buckets can be merged!

Dealing with high dimensional data

- Borrow Sariel's idea for k-median.
- Project into an ε-fan net (lines) shot from the approximate centers then apply the one dimensional construction.
- Need to take union of the approximate centers for all p_i-centrum problem.