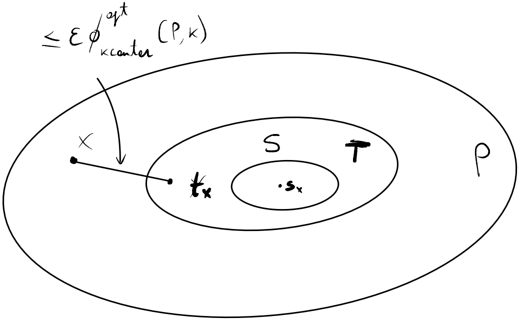
Coreset Technique

(Part 1 - Exercises)

Exercise

Let P be a set of N points in a metric space (M, d), and let $T \subseteq P$ be a coreset of $|T| > k$ points such that for each $x \in P$ we have $d(x, T) \leq \epsilon \Phi^{\mathrm{opt}}_{\mathrm{kcenter}}(P, k)$, for some $\epsilon \in (0, 1)$. Let S be the set of k centers obtained by running the Farthest-First Traversal algorithm on T. Prove an upper bound to $\Phi_{\mathrm{kcenter}}(P, S)$ as a function of ϵ and $\Phi^{\mathrm{opt}}_{\mathrm{kcenter}}(P, k)$.



$$\phi_{kcenter}(P,S) = f(\varepsilon) \cdot \phi^{opt}_{kcenter}(P,k)$$

$\forall x \in P: \quad d(x,S) \leq f(\varepsilon)\, \phi^{OPT}_{kcenter}(P,k)$

let t_x be the closest point to x in T

let s_x be the closest point to t_x in S

def of d triangular inequality

$$d(x,S) \leq d(x,s_x) \leq d(x,t_x) + d(t_x,s_x) \leq$$

$$\leq \varepsilon\, \phi^{OPT}_{kcenter}(P,k) + d(t_x, s_x)$$

$$\leq \varepsilon\, \phi^{OPT}_{kcenter}(P,k) + 2\,\phi^{OPT}_{kcenter}(P,k) = (2+\varepsilon)\,\phi^{OPT}_{kc}(P,k)$$

$2+\varepsilon$ - approximation

We now have to prove that $\forall t\ \exists s' \in S$ s.t.

$$d(x,s') \leq 2\,\phi^{OPT}_{kc}(P,k)$$

$S = \{s_1, \dots, s_k\} \rightarrow$ point s_i found at iteration i of the FFT

Run the FFT once more: $q = s_{k+1} \longrightarrow \hat{S} = S \cup \{q\} = \{s_1, s_2, \dots, s_k, s_{k+1}\}$

Then $\exists\, s_i, s_j$ that are in the same cluster C_ℓ^*

$$d(s_i, s_j) \leq d(s_i, c_\ell^*) + d(c_\ell^*, s_j)$$

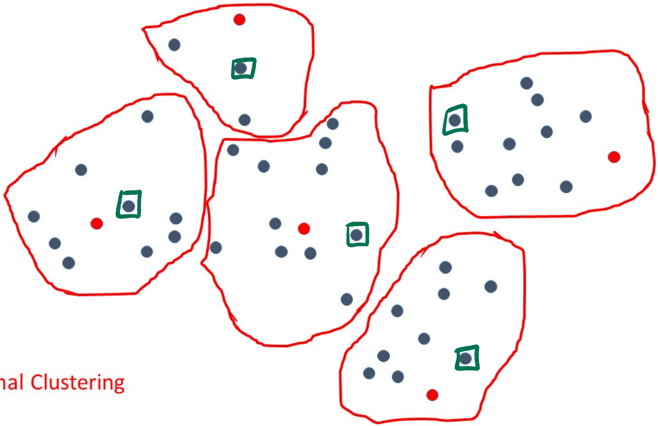
Optimality of the cluster $\quad \leq \phi^{OPT}(P,k) + \phi^{OPT}(P,k)$

$$\leq 2\,\phi^{OPT}(P,k)$$

$$\forall x \in T,\ d(x,S) = d(x, \{s_1, \dots, s_k\}) \leq d(x, \{s_1, \dots, s_{j-1}\}) \leq d(s_j, \{s_1, \dots, s_{j-1}\})$$
$$\leq d(s_j, \{s_i\}) \leq 2\,\phi^{OPT}(P,k)$$

Exercise

Let P be a set of points in a metric space (M, d), and let $T \subseteq P$. For any $k < |T|, |P|$, show that $\Phi^{\text{opt}}_{\text{kcenter}}(T, k) \leq 2\Phi^{\text{opt}}_{\text{kcenter}}(P, k)$. Is the bound tight?



Optimal Clustering

for each cluster take 1 point → T'

$$\phi^{OPT}(T,k) \le \phi(T,\hat{T})$$

set generated
by taking k points

1) $C^* = \{ C_1^* \dots C_k^* \}$ opt sol in P with centers $\hat{c}_1^*, \dots, \hat{c}_k^*$

2) Partition P using C^*

3) Remove all points not in T

4) $\forall C_i^*$ take one point in $T \rightarrow \hat{t}_i$

5) $\hat{T} = \{ \hat{t}_1, \dots, \hat{t}_k \}$

Exercise

Let P be a set of N points in a metric space (M, d), and let $\mathcal{C} = (C_1, C_2, \ldots, C_k; c_1, c_2, \ldots, c_k)$ be a k-clustering of P. Initially, each point $q \in P$ is represented by a pair $(\mathrm{ID}(q), (q, c(q)))$, where $\mathrm{ID}(q)$ is a distinct key in $[0, N-1]$ and $c(q) \in \{c_1, \ldots, c_k\}$ is the center of the cluster of q.

1. Design a 2-round MapReduce algorithm that for each cluster center c_i determines the most distant point among those belonging to the cluster C_i (ties can be broken arbitrarily).
2. Analyze the local and aggregate space required by your algorithm. Your algorithm must require $o(N)$ local space and $O(N)$ aggregate space.

Exercise

Let P be a set of N *bicolored points* from a metric space, partitioned into k clusters $C_1, C_2, \ldots, C_k$. Each point $x \in P$ is initially represented by the key-value pair $(\mathrm{ID}_x, (x, i_x, \gamma_x))$, where ID_x is a distinct key in $[0, N-1]$, i_x is the index of the cluster which x belongs to, and $\gamma_x \in \{0, 1\}$ is the color of x.

1. Design a 2-round MapReduce algorithm that for each cluster C_i checks whether all points of C_i have the same color. The output of the algorithm must be the k pairs (i, b_i), with $1 \le i \le k$, where $b_i = -1$ if C_i contains points of different colors, otherwise b_i is the color common to all points of C_i.
2. Analyze the local and aggregate space required by your algorithm. Your algorithm must require $o(N)$ local space and $O(N)$ aggregate space.