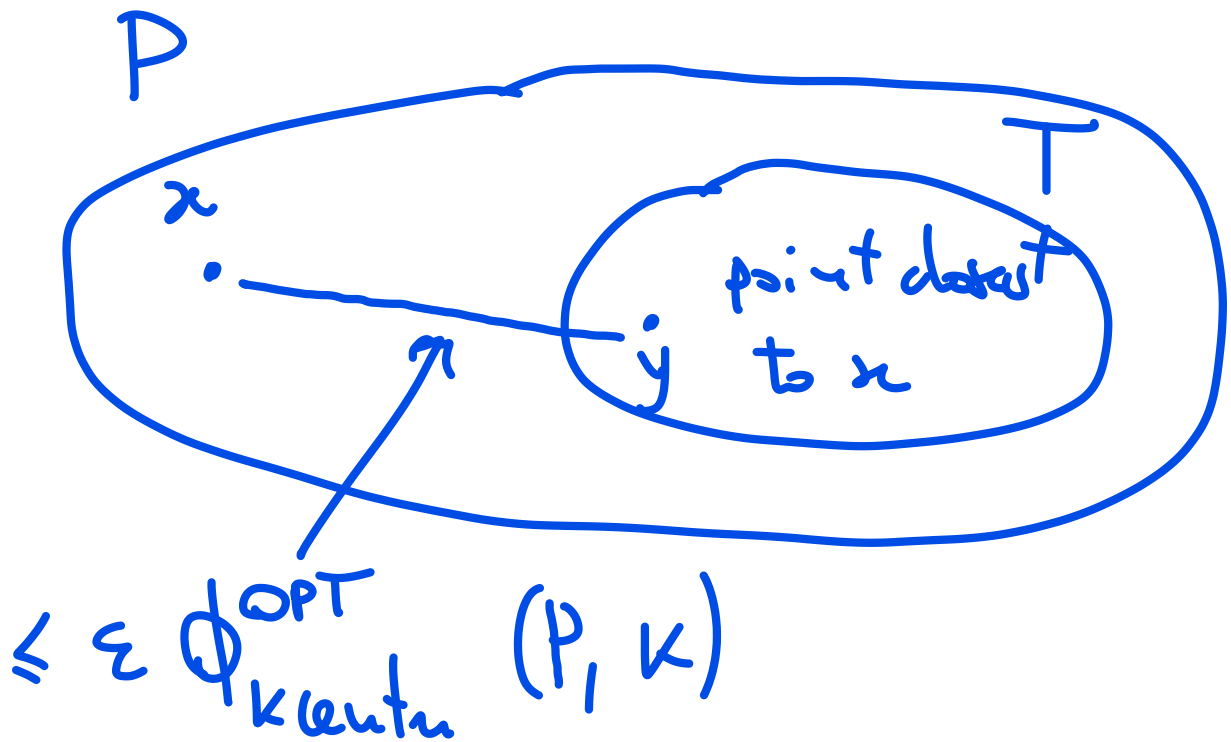
Clustering: Exercises

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 $d(x, T) \leq \epsilon \Phi^{\text{opt}}_{\text{kcenter}}(P, k)$, 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 $\Phi_{\text{kcenter}}(P, S)$ as a function of ϵ and $\Phi^{\text{opt}}_{\text{kcenter}}(P, k)$.



Hint: Repeat the same argument made in the analysis of MR-FFT. Specifically, show:

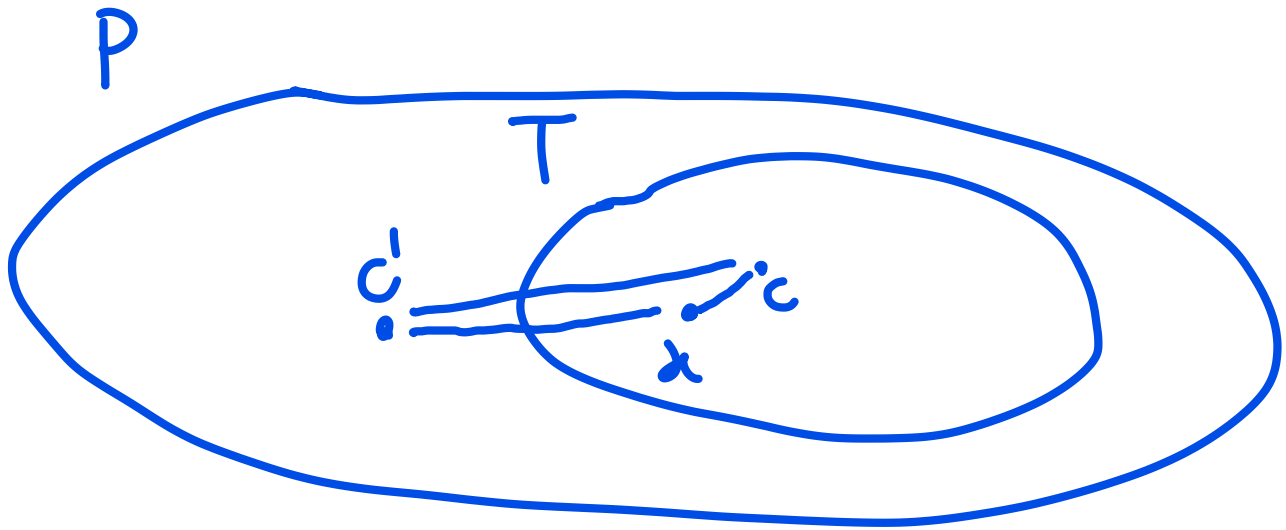
- $\forall y \in T \; \exists \, c \in S \subseteq T$ such that $d(y, c) \leq 2 \phi^{OPT}_{k\text{-center}}(P, k)$

* Putting things all together we conclude:
  - $\forall x \in P \; \exists \, y \in T : d(x, y) \leq \varepsilon \phi^{OPT}_{k\text{-cent}}(P, k)$ (by hypothesis)
  - $\forall y \in T \; \exists \, c \in S : d(y, c) \leq 2 \phi^{OPT}_{k\text{-center}}(P, k)$

$\Rightarrow d(x, c) \leq d(x, y) + d(y, c) \leq (2+\varepsilon) \phi^{OPT}_{k\text{-center}}(P, k)$

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?

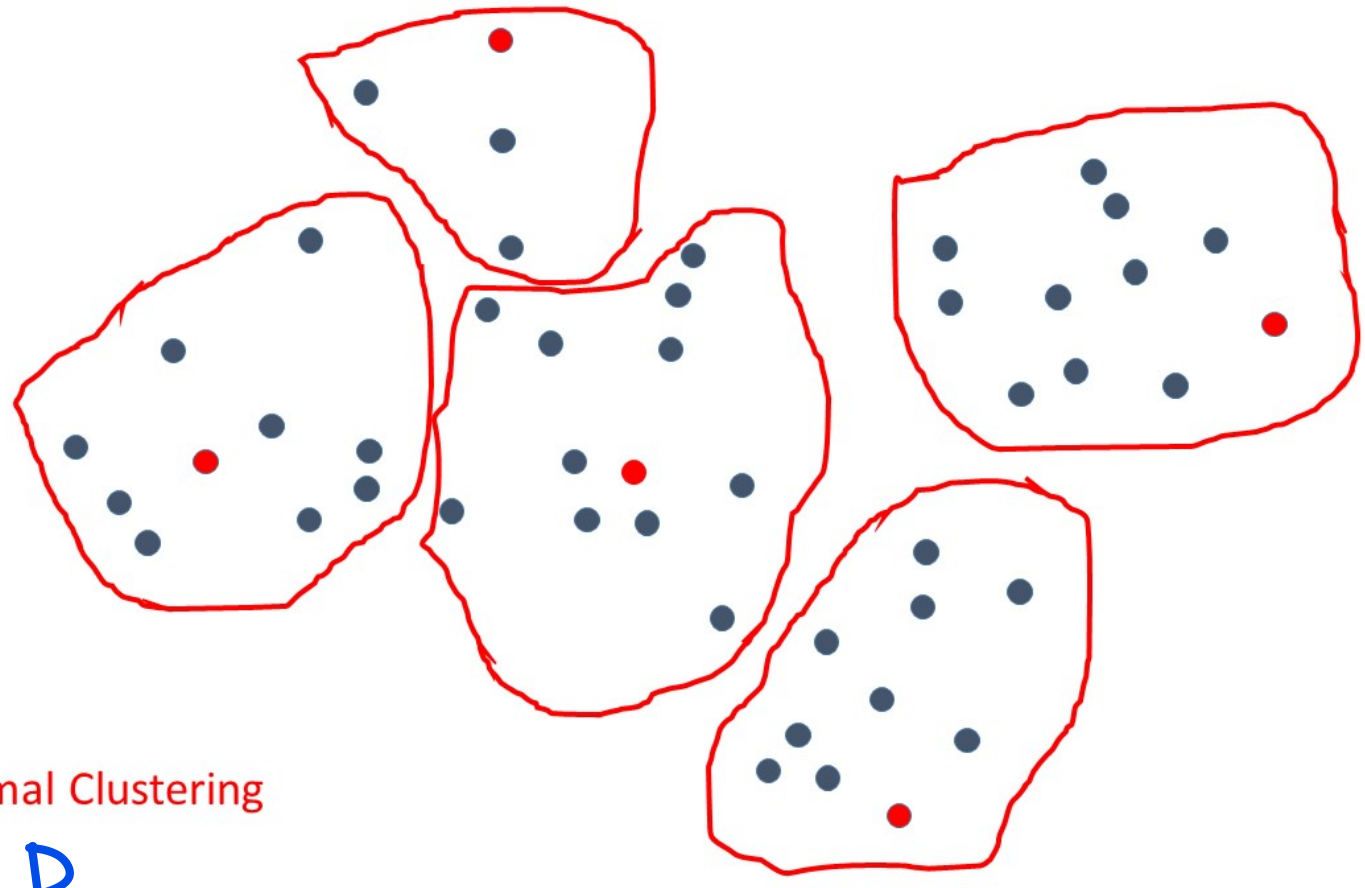


$d(x, c') \leq d(x, c) + d(c, c')$

if in the optimal solution for P x is "assigned" to a center c' not in T we must find a new center c for x in T not too far from c'

Let $(C_1, C_2, \ldots, C_k, c_1, c_2, \ldots, c_k)$ be the optimal clustering for P (k-center clustering)

* Remove from each C_i the points not in T
* Select an arbitrary point from $C_i \cap T$ (if nonempty) for every $1 \leq i \leq k$
* Let $S \subseteq T$ be the set of selected points



Optimal Clustering

Optimal Clustering after
removal of points not in T

$S \equiv$ set of points in $\square$

Consider a point $x \in T$ and suppose $x \in C_i$

let $c \in S$ be the point of S selected from C_i

Then $d(x, c) \leq d(x, c_i) + d(c_i, c)$

$$\leq 2 \phi^{opt}_{k\text{-center}}(P, k)$$

$$\Rightarrow \phi_{k\text{-center}}(T, S) \leq 2 \phi^{opt}_{k\text{-center}}(P, k)$$

$$\Rightarrow \phi^{opt}_{k\text{-cent}}(T, k) \leq \phi_{k\text{-center}}(T, S) \leq 2 \phi^{opt}_{k\text{-center}}(P, k)$$

Exercise

Let P be a set of N points from metric space (M, d), represented by key-value pairs (i, x_i), for $0 \leq i < N$. For an arbitrary $i \in [0, N)$ show that

$$d_{\max}(i) = \max\{d(x_i, x_j) \ : \ 0 \leq j < N\},$$

(which is a 2-approximation to the diameter) can be computed in MapReduce using 2 rounds, $M_L = O\left(\sqrt{N}\right)$ and $M_A = O(N)$.

APPROACH:

ROUND 1
* Subdivide points into $\sqrt{N}$ partitions (as usual) but include a copy of x_i in each partition
* In each partition compute the max distance between x_i and the other points of the partition

ROUND 2
* Compute the maximum of the local maxima

ROUND 1

Map phase:

$(j, x_j) \longrightarrow (j \bmod \sqrt{N}, (j, x_j)) \quad \forall j \neq i$

$(i, x_i) \longrightarrow \{(h, (i, x_i)) : 0 \leq h < \sqrt{N}\}$

Reduce phase: For each $h \in [0, \sqrt{N})$ separately, let L_h = list of values (j, x_j) from intermediate pairs with key h (note that L_h includes (i, x_i))

$(h, L_h) \rightarrow (0, d_h = \max_{(j,x_j) \in L_h} d(x_i, x_j))$

ROUND 2

Map phase : empty

Reduce phase : L_0 = list of values d_h with $0 \le h < \sqrt{N}$

$(0, L_0) \rightarrow (0, \max_{d_h \in L_0} d_h) \longrightarrow \equiv d_{max(i)}$

Analysis

$M_L = O(\sqrt{N})$ note that in Round 1 $O(\sqrt{N})$ local space is needed both in the Map phase, to create the replicas of x_i, and in the Reduce phase, to process each partition

In Round 2 (Reduce Phase) space $O(\sqrt{N})$ is needed to process L_0

$M_A = O(N)$: input, intermediate and output pairs in each round are $O(N)$

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 $1 \leq i \leq k$, 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.

1-ROUND Approach

Map Phase: $(ID_x, (x, i_x, \gamma_x)) \rightarrow (i_x, \gamma_x)$

Reduce Phase: For each $i \in [1, k]$ separately, let

L_i = list of colors from intermediate pairs with key i

(i.e., all color labels of points of C_i $\Rightarrow |L_i| = |C_i|$

$(i, L_i) \rightarrow (i, b_i)$ where

$$b_i = \begin{cases} -1 & \text{if } L_i \text{ contains } > 1 \text{ distinct colors} \\ \gamma & \text{if } L_i \text{ contains only color } \gamma \end{cases}$$

For this algorithm, M_L is proportional to $\max_{1 \leq i \leq k} |C_i|$ which can be close to N

2-ROUND APPROACH

ROUND 1

Map Phase: $(ID_x, (x, i_x, r_x)) \rightarrow (ID_x \bmod \sqrt{N}, (i_x, r_x))$

Reduce Phase: for every $j \in [0, \sqrt{N})$ separately, let

L_j = list of values (i, r) from intermediate pairs with key j

$(j, L_j) \to \{(i, b_i(j)) : 1 \le i \le k \text{ and } i \text{ occurs in } L_j\}$

$$b_i(j) = \begin{cases} -1 & \text{if } L_j \text{ contains } > 1 \text{ distinct colors for } C_i \\ \gamma & \text{if } L_j \text{ contains only color } \gamma \text{ for } C_i \end{cases}$$

ROUND 2

Map Phase: empty

Reduce Phase: for each $i \in [1, k]$ separately, let

L_i = list of all $b_i(j)$'s with $j \in [0, \sqrt{N})$

$(i, L_i) \to (i, b_i)$ → final

where

$$b_i = \begin{cases} -1 & \text{if some } b_i(j) = -1 \text{ or } L_i \text{ contains both 0's and 1's} \\ r & \text{if all } b_i(j)\text{'s are equal to } r \end{cases}$$

ANALYSIS

$M_L = O(\sqrt{N})$ since Map and Reduce phases of both rounds process at most $\sqrt{N}$ elements each

$M_A = O(N)$ since the number of input, intermediate and output pairs of each round is $O(N)$

Exercise

From before, we know that if S is the set of centers computed by k-means++ for P, then there is a value $\alpha > 1$ such that

$$\Pr\left(\Phi_{\text{kmeans}}(P,S) \leq \alpha \cdot \Phi^{\text{opt}}_{\text{kmeans}}(P,k)\right) \geq 1/2.$$

Show that the above probability can be made $\geq 1 - 1/N$, if we run several *independent instances* of k-means++, and return the set of centers which yields the minimum value of the objective function, among the ones computed in the various runs.

Suppose that we execute t independent runs of k-means++

S_i = set of centers computed in the i-th run of k-means++ $\quad 1 \le i \le t$

S_{min} = the t solution among the S_i's

$$S_{min} = \underset{1 \le i \le t}{\operatorname{argmin}} \, \phi_{kmeans}(P, S_i)$$

$$P_2\left(\phi_{kmeans}(P, S_{min}) \le \alpha \, \phi^{OPT}_{kmeans}(P, k)\right) \ge 1 - \frac{1}{N} \; ?$$

By the hypothesis $\forall \; 1 \le i \le k$

$$P_2\left(\phi_{kmeans}(P, S_i) \le \alpha \, OPT\right) \geq \frac{1}{2}$$

$$OPT \triangleq \phi^{OPT}_{kmeans}(P, k)$$

Define the following independent events

E_i = "$\phi_{kmeans}(P, S_i) > \alpha \, OPT$"

by the hypothesis we know that $P_2(E_i) \le \frac{1}{2}$

E = "$\phi_{kmeans}(P, S_{min}) > \alpha \, OPT$"

$$E = E_1 \cap E_2 \cap \dots \cap E_t$$

$$P_2\left(\Phi_{kmeans}(P, S_{min}) \leq \alpha\, OPT\right) = 1 - P_2(E)$$

$$P_2(E) = P_2(E_1 \cap E_2 \cap \dots \cap E_t)$$

$$= \prod_{i=1}^{t} P_2(E_i) \quad \text{by independence}$$

$$\leq \left(\frac{1}{2}\right)^t$$

By setting $t = \log_2 N$ we get

$$P_2(E) \leq \frac{1}{N} \Rightarrow 1 - P_2(E) \geq 1 - \frac{1}{N}$$