# Coresets

Representative smaller sample of a dataset

Coreset approximates whole dataset

Example: K-median problem

Given n points on plane $P \in [\Delta]^2$

Find K centers to serve the groups $c_1, \dots, c_K \in [\Delta]^T$

$$d(P, C) = \sum_{i=1}^{n} \min_{j \in [K]} \|P_i - C_j\|$$

"total distance to centers"

Why "median"? If $\|\cdot\| = l_1$, center of each group = median in each coordinate

(K-means: $\sum_{i=1}^{n} \min_{j \in [K]} \|P_i - C_j\|_2^2$)

$$OPT(P, K) = \min_{|C|=K} d(P, C)$$

K-median problem: find C minimizing cost

NP-hard, so approximate:

Find $C$, $|C| \le k$,

$$d(P, C) \le (1+\varepsilon) \cdot OPT(P, k)$$

Kolliopoulos-Rao '99: $O\left(\left(\frac{1}{\varepsilon}\right)^{O(1/\varepsilon)} \cdot n \log n \log k\right)$

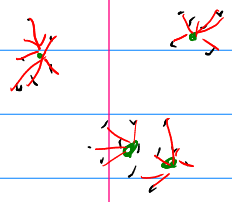
What about a streaming algorithm?

Today: generic technique for insertion streams using coresets

A coreset for $P$ is a weighted set of points:

$$s_1, \ldots, s_m \in [\Delta]^2$$

$$w_1, \ldots, w_m \in \mathbb{Z}$$



$\pi : [n] \to [m]$

$w_i = |\pi^{-1}(i)| = \#\, j$ s.t. $\pi(j) = i$

$$dist(P, (S, w)) := \sum_{j \in [n]} \|P_j - S_{\pi(j)}\|$$

$(k, \varepsilon)$ coreset if $dist(P, (S, w)) \le \varepsilon \cdot OPT(P, k)$

Lets us find appx. $k$-median solution:

$$d((S, w), C) := \sum_{j \in [m]} w_j \min_{i \in [k]} \|C_i - S_j\|$$

$$\tilde{C} := \operatorname*{argmin}_{C} d((S, w), C)$$

For any $C$,

$$d(P, C) = \sum_{i \in [n]} \min_j \|p_i - c_j\|$$

$$= \sum_{i \in [n]} \min_j \left( \|s_{\pi(i)} - c_j\| \pm \|p_i - s_{\pi(i)}\| \right)$$

$$= d((S,w), C) \pm \text{dist}(P, (S,w))$$

$$= d((S,w), C) \pm \varepsilon \cdot OPT(P,k)$$

Let $C^* = \operatorname{argmin}_C d(P, C)$. Then

$$d(P, \tilde{C}) = d((S,w), \tilde{C}) \pm \varepsilon \cdot OPT(P,k)$$
$$d(P, C^*) = d((S,w), C^*) \pm \varepsilon \cdot OPT(P,k)$$

$$d(P, \tilde{C}) \le d((S,w), \tilde{C}) + \varepsilon \cdot OPT$$
$$\le d((S,w), C^*) + \varepsilon \cdot OPT$$
$$\le d(P, C^*) + 2\varepsilon \cdot OPT$$
$$= (1 + 2\varepsilon) \cdot OPT$$

So it suffices to find a coreset, then run approximation alg. on result.

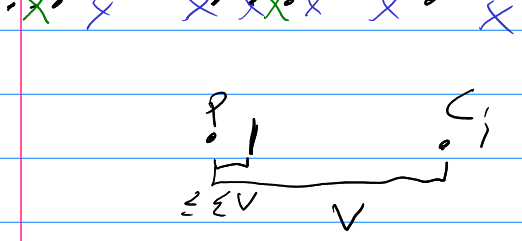
# Non-Streaming Coresets

First, suppose we know good set of centers $C$. (e.g. from KP99)

Want to get lots more centers so $d(P, S) \le \varepsilon \cdot d(P, C)$.
(then set weight = # mapping there)

1d:

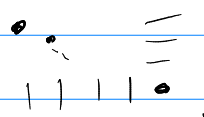
replace each $c_i$ with $\{c_i \pm x \mid x = (1+\varepsilon)^j\} \cup \{c_i\}$






Previously contributed $V$
Now $\le \varepsilon V$.

2d? $\{c_i + (ax, bx) \mid x = 2^j$
$a \in [-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}]$
$b \in [-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}]\}$



for $O(\varepsilon) \cdot OPT$,
$\Rightarrow O\left(\frac{k}{\varepsilon^2} \log \Delta\right)$ size.

(insertion)

# Streaming Coresets

Merge-and-reduce (useful on $+(w.l)$

Suppose $(S_1, w_1) =$ $(k, \varepsilon)$ coreset for $P_1$,
$(S_2, w_2) =$ $(k, \varepsilon)$ coreset for $P_2$.

Construct a coreset for $P_1 \cup P_2$.

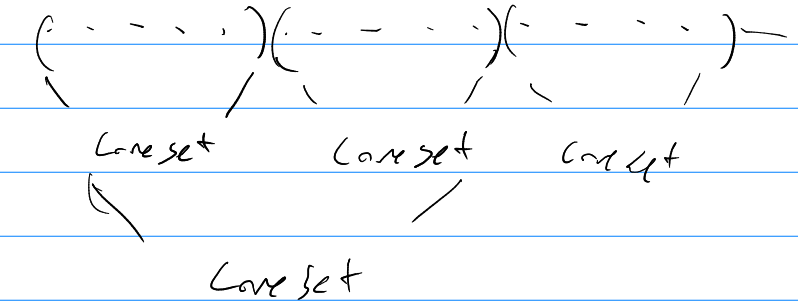
Ans just run the algorithm, get $(\bar{S}, \bar{w})$ that is $(k, \varepsilon')$ coreset.

$(S_1, w_1)$ moved $P_1$ by $\leq \varepsilon \cdot OPT(P_1, k)$
so $(S_1 \cup S_2, w_1 \cup w_2)$ moved $P_1 \cup P_2$ by $\leq \varepsilon \cdot opt(P_1 \cup P_2, k)$

Hence $(S_1, w_1) \cup (S_2, w_2)$ is $(k, \varepsilon)$ coreset for $P_1 \cup P_2$
But bigger. Run coreset alg. again to be smaller,
now

$$dist((\bar{S}, \bar{w}), P_1 \cup P_2) \leq \varepsilon \cdot OPT(P_1 \cup P_2, k) + dist((\bar{S}, \bar{w}), (S_1 \cup S_2, w_1 \cup w_2))$$
$$\leq \varepsilon \cdot OPT(P_1 \cup P_2, k) + \varepsilon' \cdot opt((S_1 \cup S_2, w_1 \cup w_2), k)$$
$$\leq \varepsilon \cdot OPT(P_1 \cup P_2, k) + \varepsilon'(1+\varepsilon) OPT(P_1 \cup P_2, k)$$
$$\leq (\varepsilon + 2\varepsilon') OPT(P_1 \cup P_2, k)$$

Stream of Points




Store tree. Merge any two at same level to coreset of size $O(\frac{k}{\varepsilon^2}\log\Delta)$.

the approximation error has

$$\varepsilon_i \le 2\varepsilon + \varepsilon_{i-1} \le 2i\varepsilon.$$

So Final result is a $(k, 2\varepsilon\cdot\log n)$ coreset.

replace $\varepsilon$ by $\varepsilon/2\log n$

$\Rightarrow$ $O(\frac{k}{\varepsilon^2}\cdot\log^2 n\log\Delta)$ space per level

$O(\frac{k}{\varepsilon^2}\log^3 n\log\Delta)$ total.