Coresets

Coreset

(informally): a small subset of the input that "preserves" useful properties of the input

ideally very small compared to n

Example

- Minimum Enclosing Ball problem

Given: $P = \{p_1 \dots p_n\} \subseteq \mathbb{R}^d$

Find: $x \in \mathbb{R}^d$ minimizing

$$\max_{p \in P} \{d(p,x)\}$$

Coreset: $Q \subseteq P$ such that for all $y \in \mathbb{R}^d$, $\max_{q \in Q} \{d(q,y)\} \approx \max_{p \in P} \{d(p,y)\}$.

Coreset

(formally): given a cost function

S is an <u>ε-coreset</u> for P if

- $(1-\varepsilon)\,\mathrm{cost}(P) \leq \mathrm{cost}(S) \leq \mathrm{cost}(P)$
- $(1-\varepsilon)\,\mathrm{cost}(P) \leq \mathrm{cost}(S) \leq (1+\varepsilon)\,\mathrm{cost}(P)$
- $\mathrm{cost}(S) \leq \mathrm{cost}(P) \leq (1+\varepsilon)\,\mathrm{cost}(S)$

depends on the problem...

Example

- Minimum Enclosing Ball problem

Given: $P = \{p_1 \ldots p_n\} \subseteq \mathbb{R}^d$

Find: $x \in \mathbb{R}^d$ minimizing

$$\mathrm{cost}(P,x) = \max_{p \in P} \{ d(p,x) \}$$

Coreset: $Q \subseteq P$ such that for all $y \in \mathbb{R}^d$, $(1-\varepsilon)\,\mathrm{cost}(P,y) \leq \mathrm{cost}(Q,y) \leq \mathrm{cost}(P,y)$

ε-coreset for MEB

Alg:

$S_1 \leftarrow \{\text{an arbitrary } p \in P\}$

for $i = 1$ to T ⟵ determine later

$c_i \leftarrow$ MEB center of S_i

$p_i \leftarrow \arg\max_{p \in P} d(c_i, p)$

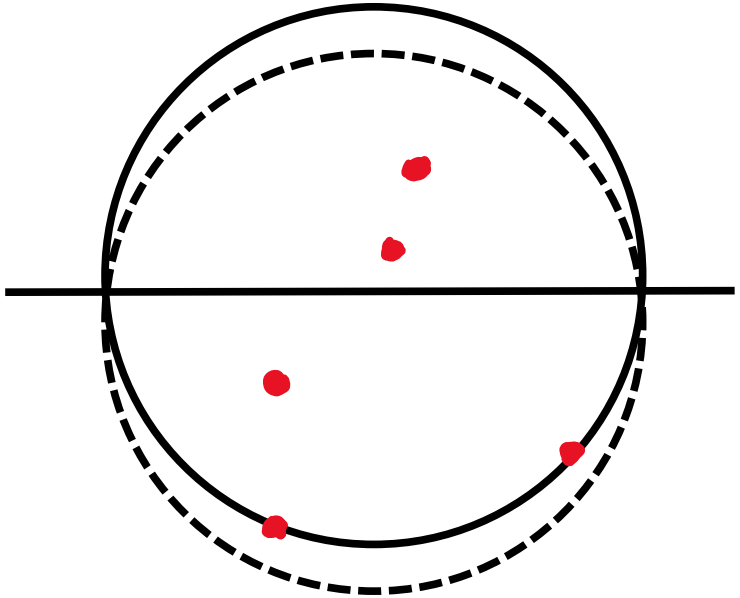
$S_{i+1} \leftarrow S_i \cup \{p_i\}$

return S_{T+1}

ε-coreset for MEB

Analysis:

Lemma: Let B be the MEB for P with center C & radius R. Any halfspace containing C contains $p \in P$ with $d(p, C) = R$.



Pf: Suppose not. Then can shift the ball perpendicular to halfplane $\Rightarrow$ no points on boundary.

We can then shrink ball $\Rightarrow$ not minimum.

ε-coreset for MEB

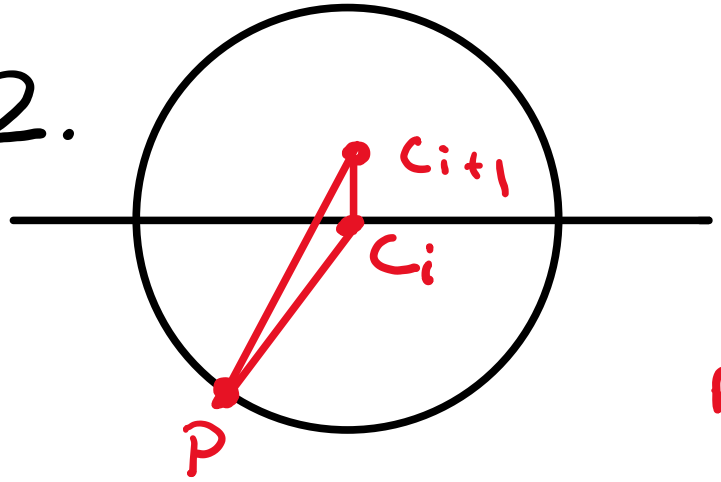
Analysis: R = radius of MEB(P) $\qquad \lambda_i = \frac{r_i}{R}$

r_i = radius of MEB(S_i)

Goal: identify T such that $\lambda_T \geq 1-\varepsilon$

1. $\exists\, q \in P$ s.t. $d(q, c_i) \geq R$

$$r_{i+1} \geq d(q, c_{i+1}) \geq d(q, c_i) - d(c_i, c_{i+1})$$
$$\geq R - d(c_i, c_{i+1})$$

2. 

Suppose $d(c_i, c_{i+1}) > 0$ (else done)

By Lemma $\exists\, p$, $d(p, c_i) = r_i$

$$r_{i+1} \geq d(c_{i+1}, p) \geq \sqrt{r_i^2 + d(c_i, c_{i+1})^2}$$

ε-coreset for MEB

Analysis: R = radius of MEB(P) $\qquad \lambda_i = \frac{r_i}{R}$

r_i = radius of MEB(S_i)

Goal: identify T such that $\lambda_T \geq 1-\varepsilon$

$$\Rightarrow \lambda_{i+1} \geq \frac{1}{R} \max\left\{R - d(c_i, c_{i+1}), \sqrt{\lambda_i^2 R^2 + d(c_i, c_{i+1})^2}\right\}$$

minimized when $R - d(c_i, c_{i+1}) = \sqrt{\lambda_i^2 R^2 + d(c_i, c_{i+1})^2}$

$$\Rightarrow d(c_i, c_{i+1}) = \frac{(1-\lambda_i^2)R}{2}$$

$$\lambda_{i+1} \geq \frac{R - d(c_i, c_{i+1})}{R} = \frac{1+\lambda_i^2}{2} \Rightarrow \lambda_i \geq 1 - \frac{1}{1+i/2}$$

ε-coreset for MEB

Analysis: R = radius of MEB(P) $\lambda_i = \dfrac{r_i}{R}$

r_i = radius of MEB(S_i)

Goal: identify T such that $\lambda_T \geq 1-\varepsilon$

$$\Rightarrow \lambda_T \geq 1 - \frac{1}{1+T/2} \geq 1-\varepsilon$$

suffices to set $T \geq \dfrac{2}{\varepsilon}$.

ε-coreset for MEB

Alg:

$S_1 \leftarrow \{$ an arbitrary $p \in P\}$

for $i = 1$ to $T = 2/\varepsilon$

$c_i \leftarrow$ MEB center of S_i

$p_i \leftarrow \arg\max_{p \in P} d(c_i, p)$

$S_{i+1} \leftarrow S_i \cup \{p_i\}$

return S_{T+1}

Size of ε-coreset is independent of n

Stre

Us

—

Sc

— (Strong Merge: It

Streaming Coresets

Reduce + Strong Merge Properties

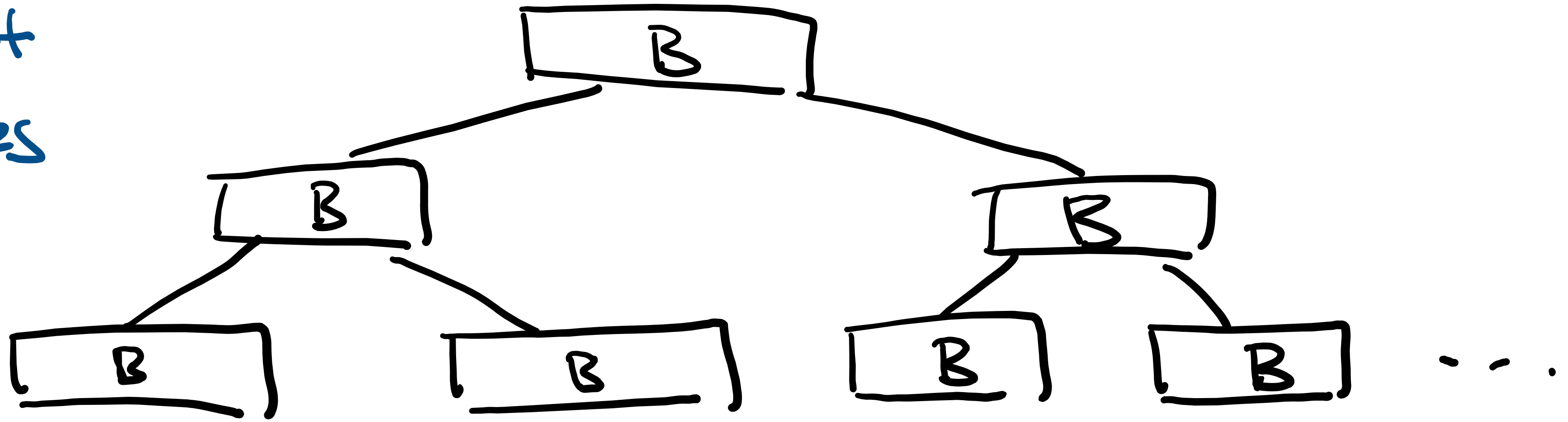
$\Rightarrow$ use Merge & Reduce Paradigm

Suppose $f(\varepsilon)$ space for ε-coreset.

Set $B = f(\varepsilon / \log n)$

at each level:

$\varepsilon / \log n$ -coreset of child nodes



Streaming Coresets

Reduce + Strong Merge Properties

$\Rightarrow$ use Merge & Reduce Paradigm

Suppose $f(\varepsilon)$ space for ε-coreset.

Set $B = f(\varepsilon/\log n)$

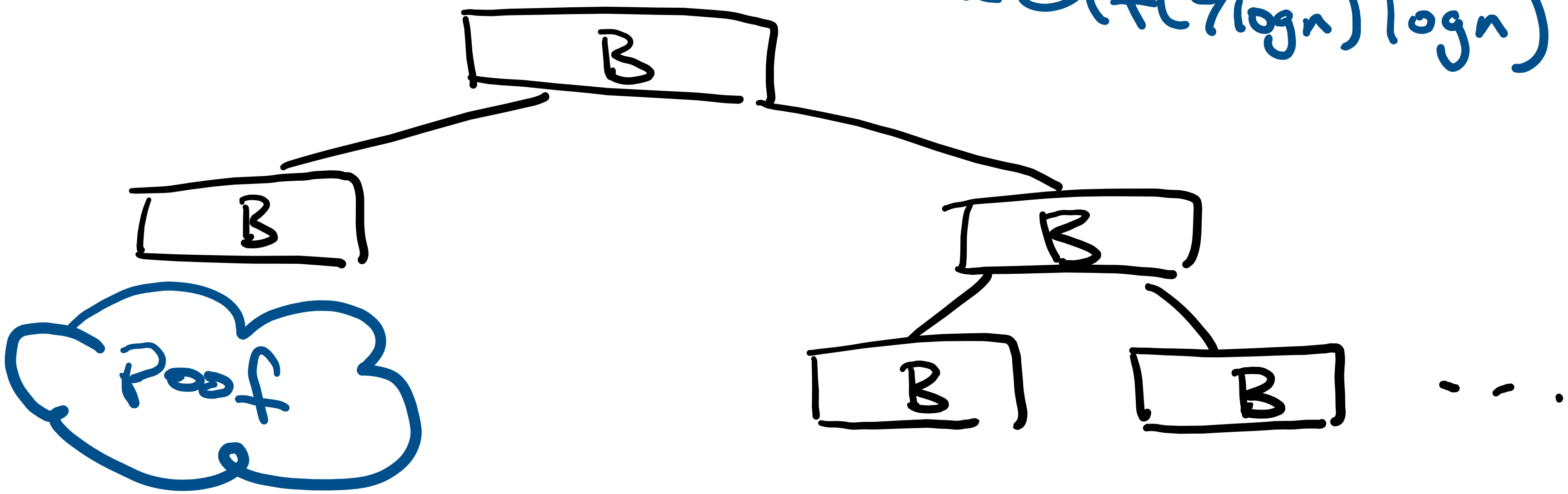
Space: $O(B \log n/B)$

$= O(f(\varepsilon/\log n) \log n)$

at each level:

$\varepsilon/\log n$-coreset of child nodes

keep at most $2B$ points at each level



Streaming Coresets

Reduce + Strong Merge Properties

$\Rightarrow$ use Merge & Reduce Paradigm

Suppose $f(\varepsilon)$ space for ε-coreset.

Correctness?

by induction.

j-th level is $j\dfrac{\varepsilon}{\log n}$-coreset.

height $= O(\log n) \Rightarrow$ ε-coreset at top.

Space: $O(B \log n/B)$

$= O(f(\varepsilon/\log n)\log n)$

$\Rightarrow O\left(\dfrac{(\log n)^2}{\varepsilon}\right)$ space for streaming MEB

Clustering

k-center: Given $P \subseteq \mathbb{R}^d$, $k \in \mathbb{Z}$

find $C = \{c_1 .. c_k\}$ minimizing

$$\mathrm{cost}(P, C) = \max_{p \in P} \min_{c \in C} \{d(p,c)\}.$$

MEB: k is always 1.

k-median

$$\mathrm{cost}(P,C) = \sum_{p \in P} \min_{c \in C} d(p,c)$$

k-means

$$\mathrm{cost}(P,C) = \sum_{p \in P} \min_{c \in C} d(p,c)^2$$

NP-Hard if k is part of input

Coreset for k-median

(k, ε)-coreset: $Q \subseteq P$ such that

for all $C \subseteq \mathbb{R}^d$, $|C| = k$,

$$(1-\varepsilon)\,\mathrm{cost}(P, C) \le \mathrm{cost}(Q, C) \le (1+\varepsilon)\,\mathrm{cost}(P, C)$$

will need idea of <u>weighted</u> k-median

$$\mathrm{cost}(P, C) = \sum_{p \in P} w(p) \min_{c \in C} d(p, c)$$

Coreset for k-median

(α, β) - bicriterion approx:

Find $A = \{a_1 \dots a_m\}$ $\quad m \le \alpha k$

s.t. $\mathrm{cost}(P, A) \le \beta \, \mathrm{opt}(P, k)$

Will convert into $(1+\varepsilon)$-approx

using k "centers"

via weighted ε-coreset.

(w/ prob $\ge 1-\delta$)

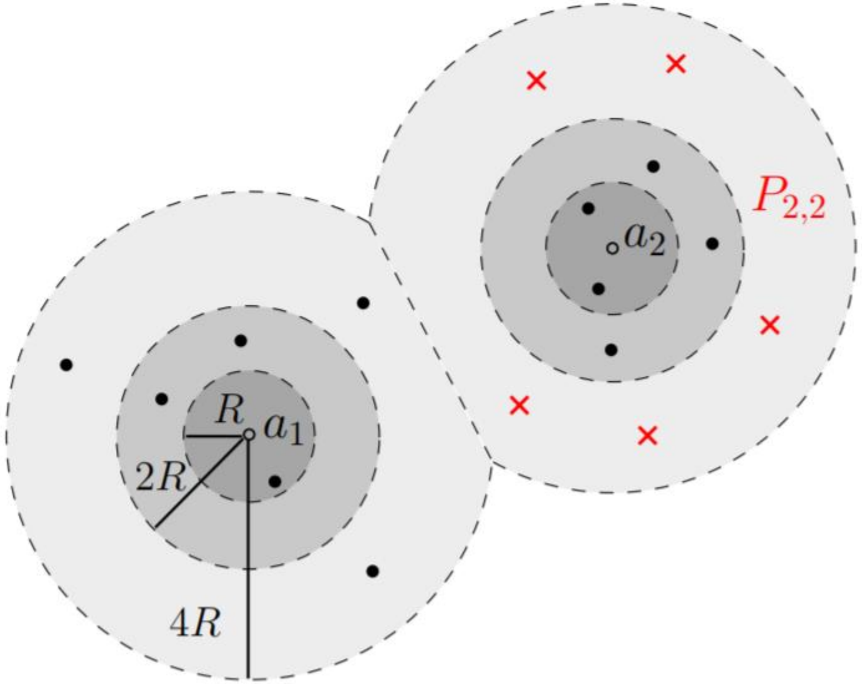
Coreset for k-median

1. Let A be (α, β) approx to k-median

   P_i : points in P "assigned" to a_i

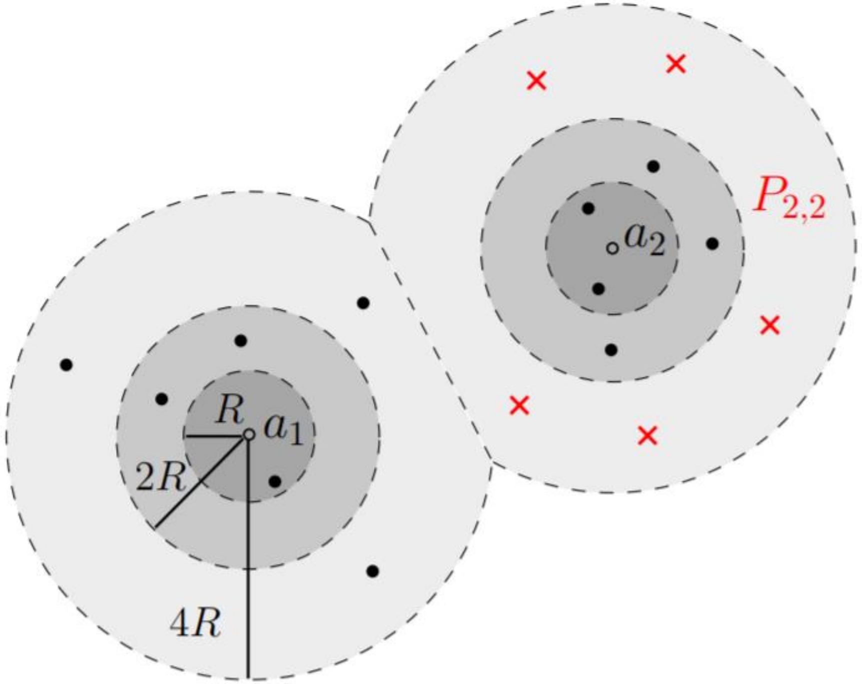
2. Split P_i into P_{ij} , $j \in \{0 .. \lg(\beta n)\}$

P_{ij} = points in P_i w/ distance to a_i between $2^{j-1}R$ & $2^j R$



$$R = \frac{\mathrm{cost}(P, A)}{\beta n}$$

Coreset for k-median



3. Sample s points from P_{ij}

$$s = O\left(\frac{\beta^2}{\varepsilon^2}\left(k \log n + \log \frac{1}{\delta}\right)\right)$$

& assign weight $\frac{|P_{ij}|}{s}$

$\longrightarrow S_{ij} \subseteq P_{ij}$.

Claim: $S = \bigcup_{i,j} S_{ij}$ is ε-coreset

w/ prob $\geq 1 - \delta$

Coreset for k-median

$$s = O\left(\frac{\beta^2}{\varepsilon^2}\left(k \log n + \log \frac{1}{\delta}\right)\right)$$

Key Lemma: U = sample $\frac{\ln(2/\Delta)}{\xi^2}$ points from V

Give each sample pt weight $\frac{|U|}{|V|}$

$|\mathrm{cost}(V,C) - \mathrm{cost}(U,C)| \leq \xi |V| \mathrm{diam}(V)$ w/ prob $\geq 1-\Delta$.

Sampling S_{ij} from P_{ij} $\Rightarrow$ need bound on $|P_{ij}|\,\mathrm{diam}(P_{ij})$

"easy" calculation: $\sum |P_{ij}|\,\mathrm{diam}(P_{ij}) \leq 6\beta\, \mathrm{opt}(P,k)$

Coreset for k-median

$$s = O\left(\frac{\beta^2}{\varepsilon^2}\left(k\log n + \log\frac{1}{\delta}\right)\right)$$

$|\text{cost}(P_{ij}, C) - \text{cost}(S_{ij}, C)| \le \xi |P_{ij}| \operatorname{diam}(S_{ij})$ w/ prob $\ge 1-\Delta$.

$\sum |P_{ij}| \operatorname{diam}(P_{ij}) \le 6\beta \operatorname{opt}(P,k)$

Set $\xi = \varepsilon/6\beta$, $\Delta = \frac{n^{-2k}\delta}{2} \Rightarrow s = \frac{36\beta^2}{\varepsilon^2} \ln \frac{4n^{2k}}{\delta}$

$$|\text{cost}(P,C) - \text{cost}(S,C)| \le \sum_{ij} |\text{cost}(P_{ij}, C) - \text{cost}(S_{ij}, C)|$$

$$\le \frac{\varepsilon}{6\beta} \sum_{ij} |P_{ij}| \operatorname{diam}(P_{ij}) \le \varepsilon \, \text{cost}(P,C)$$

Coreset for k-median

$$s = O\left(\frac{\beta^2}{\varepsilon^2}\left(k\log n + \log\frac{1}{\delta}\right)\right)$$

now solve k-median directly on ε-coreset

$\Rightarrow$ $(1+\varepsilon)$-approximation.

Similar idea works for k-means.

!! cost functions for k-median & k-means satisfy strong merge $\Rightarrow$ streaming!!