Today:

- Wrap up LSH (see previous notes)
- New theme: representative subsets / coresets

  Will cover:

  - quantiles / median ← Today
  - clustering (k-median)

Representative subsets / coresets

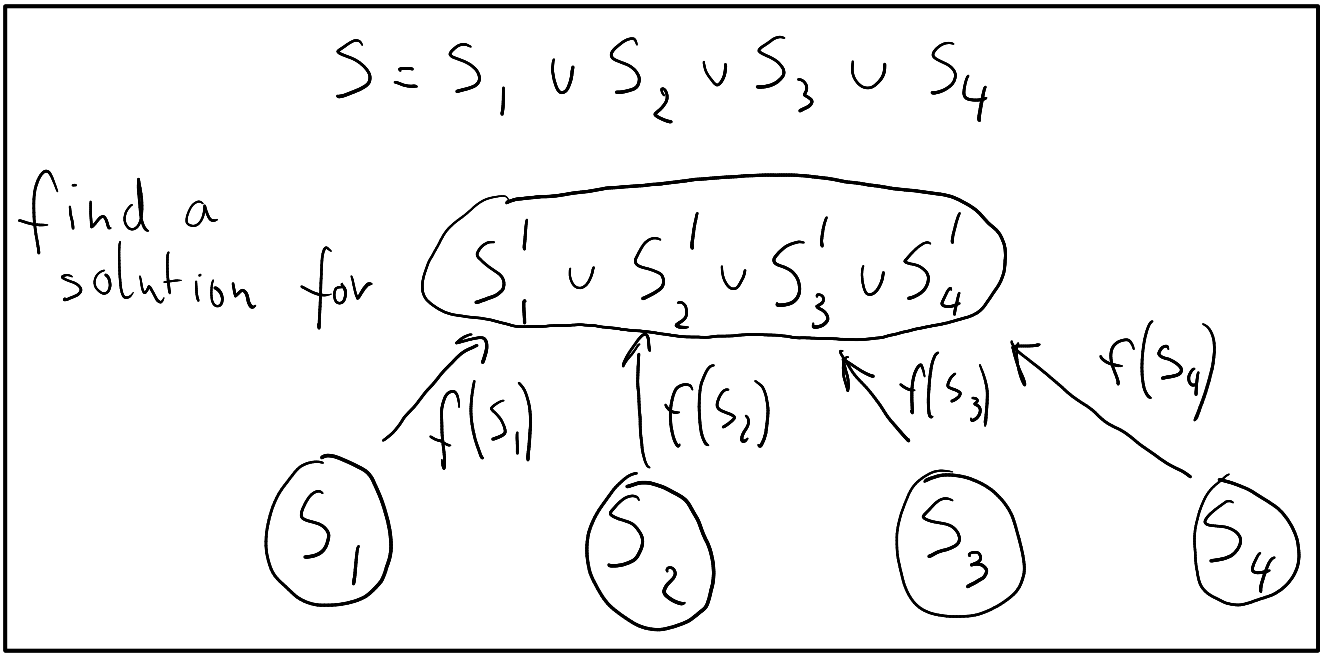
General goal:

- retain a small subset
- sufficient for solving the original problem(s)

Nice bonus properties (sometimes):

- replacing any subset with its coreset still leads to good approximation

- can combine coresets for separate subsets and still find a good solution for the original problem



Exact median : have to keep everything

$S = S_1 \cup S_2$ $\quad S_1 = \{s_1 < s_2 < s_3 < \dots < s_n\}$

S_1: size n $\quad S_2$: size $n-1$

$n-i$ elements $<$ coreset(S_1) $<$ $i-1$ elements

S_2

median of S = i-th element of S_1

S_2 can select any element of S_1

Instead: ε-approximate median

output an element of S that is:

- greater than or equal to at least $\frac{n}{2}(1-\varepsilon)$ elements

and

- less than or equal to at least $\frac{n}{2}(1+\varepsilon)$ elements

---

Notation

multiset S (finite) $\subseteq \mathbb{R}$

$$\text{rank}_S(x) = |\{y \in S : y \leq x\}|$$

= number of elements in S that are at most x

Fact: median(S) = smallest x s.t. $\text{rank}_S(x) \geq \frac{|S|}{2}$

Coresets for quantiles

(S, S' are multisets)

S' is an ε-coreset for S if

- same number of elements
- every element $x \in S'$ is present in S (possibly fewer copies of x in S)
- for all $x \in \mathbb{R}$

$$\text{rank}_S(x) \leq \text{rank}_{S'}(x) \leq \text{rank}_S(x) + \varepsilon|S|$$

Our construction

$$S = \{s_1 \leq s_2 \leq s_3 \leq \cdots \leq s_n\}$$

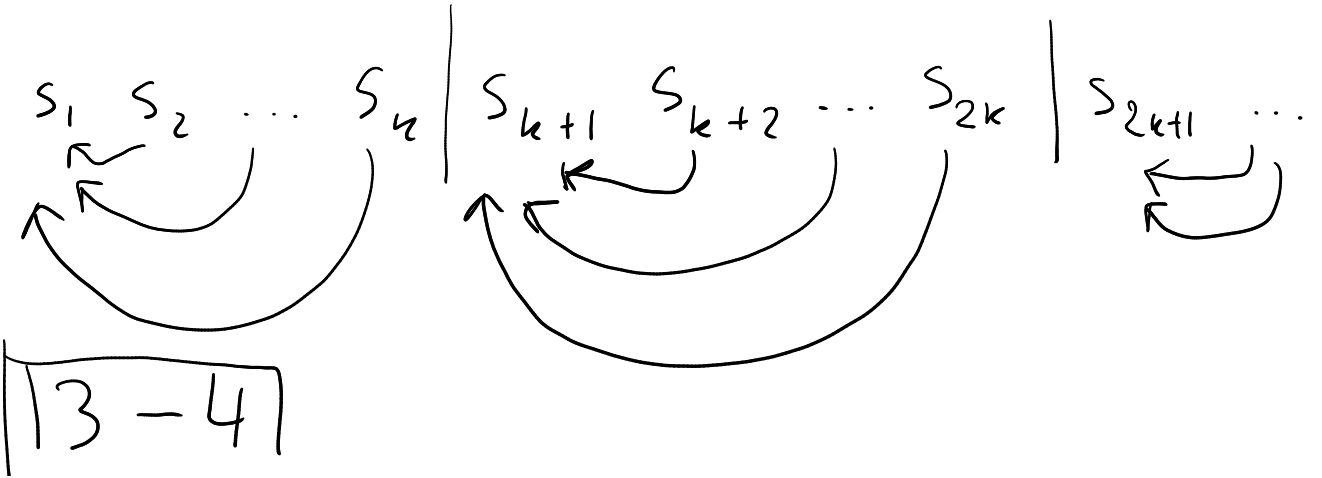
Set $k = \lceil \varepsilon n \rceil$

Replace $s_1, \ldots, s_k$ with k copies of s_1

$s_{k+1}, \ldots s_{2k}$ with k copies of s_{k+1}

⋮

(last block might be shorter)



Output: multiset with at most $1/\varepsilon$ different elements

Storage: $\leq 1/\varepsilon \times$ (element of S + count)

Why this works

S' = output of our construction for S

- first two properties of a coreset met trivially
- consider any $x \in \mathbb{R}$:
  - each block $\{s_1, \ldots, s_k\}, \{s_{k+1}, \ldots, s_{2k}\}, \ldots$ spans a range $[\underbrace{a_1}_{s_1}, \underbrace{b_1}_{s_k}], [\underbrace{a_2}_{s_{k+1}}, \underbrace{b_2}_{s_{2k}}], \ldots$
  - $a_1 \leq b_1 \leq a_2 \leq b_2 \leq a_3 \leq \ldots$
  - if x not in any range

    $\text{rank}_S(x) = \text{rank}_{S'}(x)$ ✓
  - otherwise let $[a_j, b_j]$ be the last range to which x belongs

- elements move only within ranges and only decrease
- ranges $[a_i, b_i]$ for $i < j$ are such that $b_i \leq x \Rightarrow$ no impact on rank of x
- ranges $[a_i, b_i]$ for $i > j$ are such that $x < a_i \Rightarrow$ no impact on rank of x
- range $[a_j, b_j] \ni x$:

  at most $k-1 \leq \varepsilon|S|$ elements can decrease, and hence rank of x can grow by at most $\varepsilon|S|$

$$\text{rank}_S(x) \leq \text{rank}_{S'}(x) \leq \text{rank}_S(x) + \varepsilon|S|$$

Approximate median from a coreset

How: output $\underbrace{\text{the median}}_{s_*}$ of an ε-coreset

Why: S - original set, S' - coreset, $s_* \in S$

① s_* less than or equal to $|S|/2$ elements of S

because $s_* \leq \text{median}(S)$

(2) $\text{rank}_{S'}(s_*) \leq \text{rank}_S(s_*) + \varepsilon|S|$

s_* median of $S' \Rightarrow \text{rank}_{S'}(s_*) \geq \frac{|S'|}{2} = \frac{|S|}{2}$

$$\boxed{\text{rank}_S(s_*) \geq \frac{|S|}{2}(1-2\varepsilon)}$$

s_* is an (2ε)-approximate median of S

Nice property (1):

$$S = S_1 \cup S_2 \cup \ldots \cup S_k$$

Union of ε-coresets for S_i's is an ε-coreset for S

Why: S_i' - coreset for S_i, $S' = \bigcup_{i=1}^{k} S_i'$

- first two properties of coresets met trivially
- consider any $x \in \mathbb{R}$

  For any $i \in [k]$: $\text{rank}_{S_i}(x) \leq \text{rank}_{S_i'}(x) \leq \text{rank}_{S_i}(x) + \varepsilon|S_i|$

$$\underbrace{\sum_{i=1}^{k} \text{rank}_{S_i}(x)}_{\text{rank}_S(x)} \leq \underbrace{\sum_{i=1}^{k} \text{rank}_{S_i'}(x)}_{\text{rank}_{S'}(x)} \leq \underbrace{\sum_{i=1}^{k} \left(\text{rank}_{S_i}(x) + \varepsilon|S_i|\right)}_{\text{rank}_S(x) + \varepsilon|S|}$$

Nice property ②

ε_1 - coreset of ε_2 - coreset is $(\varepsilon_1+\varepsilon_2)$ - coreset

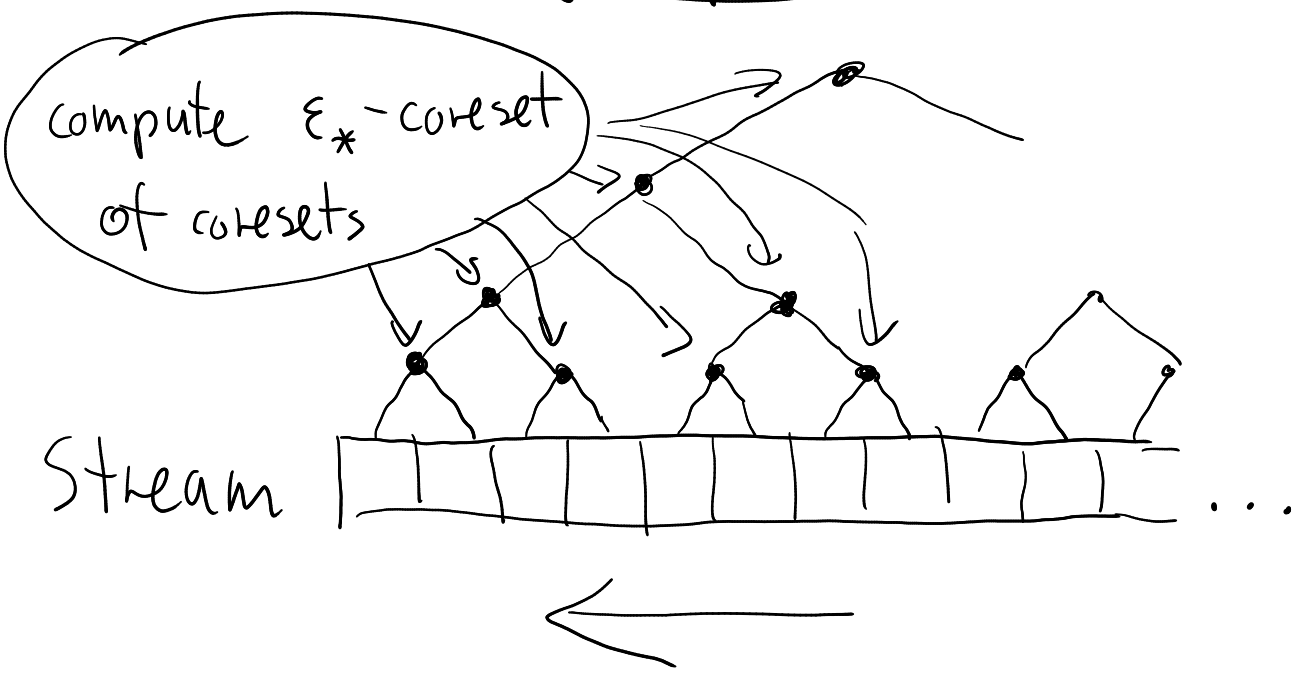
Why:

S_1 is ε_2 - coreset for S

S_2 is ε_1 - coreset for S_1

- two first properties are trivial
- need to show two inequalities for any $x \in \mathbb{R}$:
  - ① $\text{rank}_S(x) \leq \text{rank}_{S_1}(x) \leq \text{rank}_{S_2}(x)$ ✓
  - ② $\text{rank}_{S_2}(x) \leq \text{rank}_{S_1}(x) + \varepsilon_1 |S_1|$
    $\leq \text{rank}_S(x) + \varepsilon_2 |S| + \varepsilon_1 |S|$
    $= \text{rank}_S(x) + (\varepsilon_1 + \varepsilon_2)|S|$ ✓

Streaming application



- coresets computed on level i are $(\varepsilon_* \cdot i)$-coresets
- $\log n$ levels
- to get ε-approximate median, set $\varepsilon_* = \frac{\varepsilon}{2 \log n}$

Storage: # levels $\times$ (coreset size) =

$$\log n \cdot \frac{2 \log n}{\varepsilon} \times (\text{element} + \text{counter})$$

$$O\left(\frac{\log^2 n}{\varepsilon}\right) \text{ space}$$

Alternate solution:

- Sample $O(1/\varepsilon^2)$ elements
- Output their median, which is ε-approximate median with good probability
- Space: $O(1/\varepsilon^2)$

Combine the two approaches:

- approximate median of the sample is still likely to be an approximate median of the original set (see Homework 2)
- Feed the sample into our algorithm
- Output approximate median of the sample
- Space: $O\left(\frac{\log^2(1/\varepsilon)}{\varepsilon}\right)$