

# Clustering

Daniel Hsu

COMS 4772

1

## Finitely representing large sets

Let  $(\mathcal{X}, \rho)$  be a metric space.

- ▶ I.e.,  $\rho: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  is symmetric, non-negative (with  $\rho(x, y) = 0$  iff  $x = y$ ), and satisfies triangle inequality.

**Goal:** given a set  $S \subset \mathcal{X}$ , find a set  $C \subset \mathcal{X}$  (“centers”) that

- ▶ has small cardinality, and
- ▶ “represents” the set  $S$  well (as measured by a cost function).

2

## Covering / net formulations

3

### $k$ -center clustering

- ▶ Fix the cardinality  $k \in \mathbb{N}$  allowed for  $C$ .
- ▶ Cost function:

$$\text{cost}_\infty(S, C) := \max_{\mathbf{x} \in S} \rho(\mathbf{x}, C),$$

where  $\rho(\mathbf{x}, C) := \min_{\mathbf{y} \in C} \rho(\mathbf{x}, \mathbf{y})$ .

- ▶ Determines  $\varepsilon$  in  $\varepsilon$ -net criterion.
- ▶ NP-hard optimization problem.

4

## Farthest-first traversal (Gonzalez, 1985)

- ▶ **Input:** set  $S \subset \mathcal{X}$ .
- ▶ Let  $\mathbf{y}_1$  be any point in  $S$ .
- ▶ For  $t = 2, 3, \dots$ :
  - ▶ Let  $\mathbf{y}_t$  be a point in  $S$  farthest from all previous  $\mathbf{y}_i$ :

$$\mathbf{y}_t \in \arg \max_{\mathbf{x} \in S} \rho(\mathbf{x}, \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{t-1}\}).$$

- ▶ **Theorem.** For any  $k$ , cost of  $\hat{C} := \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$  is at most twice the cost of every  $C$  with  $|C| \leq k$ .

5

## Approximation analysis of farthest-first

- ▶ Let  $r_i := \rho(\mathbf{y}_{i+1}, \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_i\})$ , so

$$r_k = \rho(\mathbf{y}_{k+1}, \hat{C}) = \max_{\mathbf{x} \in S} \rho(\mathbf{x}, \hat{C}) = \text{cost}(S, \hat{C}).$$

- ▶ Pairwise distances among  $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{i+1}\}$  are at least  $r_i$ .
  - ▶ So  $r_1 \geq r_2 \geq \dots \geq r_k$ .
- ▶ Consider any set of at most  $k$  representatives  $C$ .
- ▶ At least two points in  $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{k+1}\}$  have same closest representative in  $C$ .
  - ▶ Say they are  $\mathbf{y}_i$  and  $\mathbf{y}_j$ , and they are represented by  $\mathbf{z} \in C$ .
  - ▶ By triangle inequality,

$$2 \cdot \text{cost}_\infty(S, C) \geq \rho(\mathbf{y}_i, \mathbf{z}) + \rho(\mathbf{y}_j, \mathbf{z}) \geq \rho(\mathbf{y}_i, \mathbf{y}_j) \geq r_k.$$

- ▶ So  $\text{cost}_\infty(S, \hat{C}) = r_k \leq 2 \cdot \text{cost}_\infty(S, C)$ . □

6

## $\varepsilon$ -nets

- ▶ Suppose we run farthest-first traversal to pick  $\mathbf{y}_1, \mathbf{y}_2, \dots$ , and stop as soon as

$$r_k = \text{cost}(S, \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}) \leq \varepsilon.$$

- ▶ Then  $\hat{C} := \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$  satisfies

$$\text{size of smallest } \varepsilon\text{-net} \leq |\hat{C}| \leq \text{size of smallest } \varepsilon/2\text{-net}.$$

- ▶ Size of smallest  $\varepsilon$ -net is called *covering number of  $S$*  (at scale  $\varepsilon$ , with respect to  $\rho$  metric).

7

## Set cover

- ▶ **Goal:** given set  $S$ , family of subsets  $\mathcal{F} := \{S_i : i \in \mathcal{I}\} \subseteq 2^S$ , pick  $S_{i_1}, S_{i_2}, \dots, S_{i_k}$ , with  $k$  as small as possible, that cover  $S$ :

$$\bigcup_{j=1}^k S_{i_j} = S.$$

- ▶ (Can assume  $\bigcup_{i \in \mathcal{I}} S_i = S$ .)
- ▶ **Example:**
  - ▶  $S \subseteq \mathcal{X}$  for some metric space  $(\mathcal{X}, \rho)$ .
  - ▶  $\mathcal{F} = \{B(c, \varepsilon) \cap S : c \in S\}$ , where  $B(c, r) := \{x \in \mathcal{X} : \rho(x, c) \leq r\}$  is ball of radius  $r$  around  $c$ .

8

## Greedy algorithm

- ▶ Assume  $S$  has cardinality  $n < \infty$ .
- ▶ Having already selected  $S_{i_1}, S_{i_2}, \dots, S_{i_t}$ , we next select

$$i_{t+1} \in \arg \max_{i \in \mathcal{I}} \left| S_i \cap \left( S \setminus \bigcup_{j=1}^t S_{i_j} \right) \right|.$$

(Halt when  $S$  is covered.)

- ▶ **Theorem.** If there is a cover of size  $k$ , then greedy finds a cover of size  $k(1 + \ln(n/k))$ .

9

## Analysis of greedy algorithm (Johnson, 1974)

- ▶ Suppose  $S_{i_1^*}, S_{i_2^*}, \dots, S_{i_k^*}$  covers  $S$ .
- ▶ After  $t$  steps of greedy, we have picked  $S_{i_1}, S_{i_2}, \dots, S_{i_t}$ .
  - ▶ Let  $n_t := |S \setminus \bigcup_{j=1}^t S_{i_j}|$  be the number of points in  $S$  not covered after  $t$  steps.
  - ▶ We know  $S_{i_1^*}, S_{i_2^*}, \dots, S_{i_k^*}$  would cover all  $n_t$  points.
  - ▶ So there is one of them covers at least  $n_t/k$  of the  $n_t$  points.
  - ▶ Greedy does at least well with its choice  $i_{t+1}$ .
- ▶ Starting with  $n_0 = n$ , we have

$$n_{t+1} \leq \left(1 - \frac{1}{k}\right) n_t.$$

- ▶ So  $n_t \leq k$  for  $t \geq k \ln(n/k)$ .
- ▶ After this, just need  $k$  more sets to cover remaining points.
- ▶ Total of  $k(1 + \ln(n/k))$  sets. □

10

## Average cost formulations

11

### $k$ -medians and $k$ -means cost functions

- ▶ Instead of requiring representatives close to every point in  $S$ , just require representatives close to random point in  $S$ .
- ▶ Some common cost functions:
  - ▶  **$k$ -medians**:  $\text{cost}(S, C) = \sum_{\mathbf{x} \in S} \rho(\mathbf{x}, C)$ .
  - ▶  **$k$ -means**:  $\text{cost}(S, C) = \sum_{\mathbf{x} \in S} \rho(\mathbf{x}, C)^2$ .

12

## $k$ -means

- ▶  $\mathcal{X} = \mathbb{R}^d$ ,  $\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ .
    - ▶  $\text{cost}(S, C) = \sum_{\mathbf{x} \in S} \min_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|_2^2$ .
  - ▶ NP-hard to approximate within some constant factor  $c > 1$  (Awasthi et al, 2015).
  - ▶ Easy cases:
    - ▶  $d = 1$ : dynamic programming in time  $O(n^2 k)$ .
    - ▶  $k = 1$ : bias-variance decomposition
- $$\sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{y}\|_2^2 = |S| \cdot \|\mathbf{y} - \text{mean}(S)\|_2^2 + \sum_{\mathbf{x} \in S} \|\mathbf{x} - \text{mean}(S)\|_2^2$$
- implies solution is  $\text{mean}(S)$ .
- ▶ Approximation schemes available when  $d = O(1)$  or  $k = O(1)$ .

13

## General case

- ▶ Notation: for  $C = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ ,
  - ▶  $C(\mathbf{x}) := \arg \min_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|_2^2$ , ties broken using some fixed rule.
  - ▶  $S_i^C = S_i := \{\mathbf{x} \in S : C(\mathbf{x}) = \mathbf{y}_i\}$  for each  $i = 1, 2, \dots, k$ .
- ▶ Improving  $C$ :

$$\begin{aligned} \text{cost}(S, C) &= \sum_{i=1}^k \text{cost}(S_i, C) \\ &= \sum_{i=1}^k \text{cost}(S_i, \mathbf{y}_i) \\ &\geq \sum_{i=1}^k \text{cost}(S_i, \text{mean}(S_i)) \\ &\geq \sum_{i=1}^k \text{cost}(S_i, \{\text{mean}(S_j) : j = 1, 2, \dots, k\}) \\ &= \text{cost}(S, \{\text{mean}(S_j) : j = 1, 2, \dots, k\}). \end{aligned}$$

14

## Local search algorithm (Lloyd, 1982)

- ▶ Start with  $C = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ ; repeat:
  - ▶ Partition  $S$  into  $S_1, S_2, \dots, S_k$  using  $C$ .
  - ▶ Set  $C := \{\text{mean}(S_i) : i = 1, 2, \dots, k\}$ .
- ▶ Alternative: start with partition of  $S$  into  $S_1, S_2, \dots, S_k$ .
- ▶ Cost is non-increasing.
- ▶ Eventually halts, because there are only  $O(n^{dk^2})$  ways to partition  $n$  points in  $\mathbb{R}^d$  with  $k$  Voronoi cells.
  - ▶ Could take  $2^{\Omega(n)}$  iterations (when  $k = \Theta(n)$ ), but atypical.
- ▶ How good is final solution?
  - ▶ **Depends on initialization.**
  - ▶ Could be arbitrarily worse than optimal.

15

## Bad case for Lloyd's algorithm

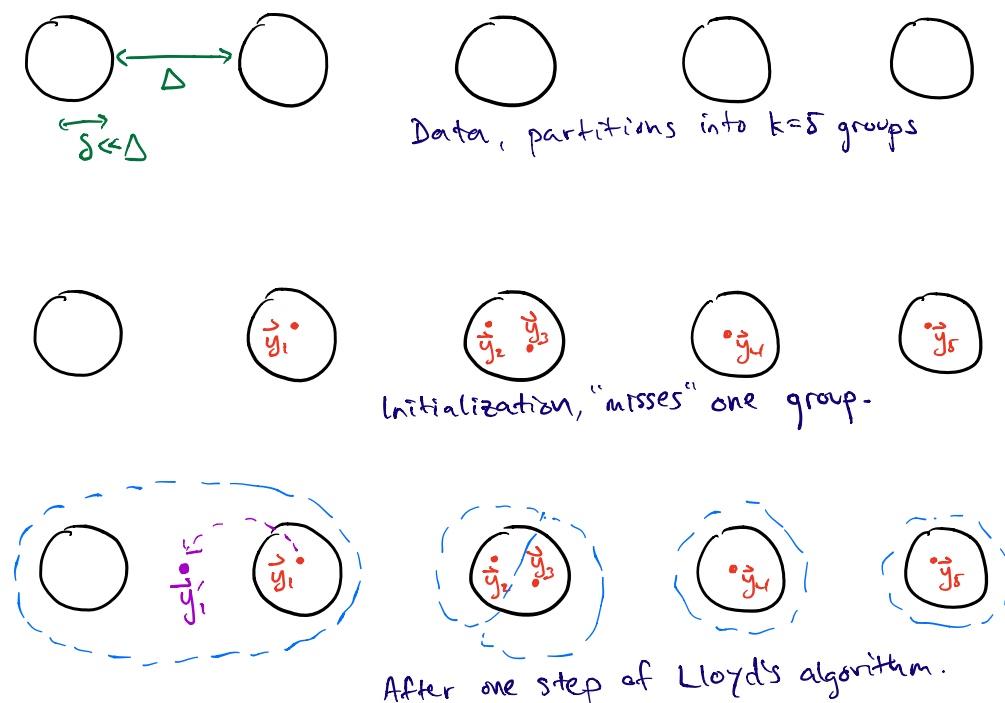


Figure 1: Bad case for Lloyd's algorithm

16



## Aside: dimension reduction

17

## Another look at bias-variance

$$\sum_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{y}\|_2^2 = |S| \cdot \|\mathbf{y} - \text{mean}(S)\|_2^2 + \sum_{\mathbf{x} \in S} \|\mathbf{x} - \text{mean}(S)\|_2^2.$$

Now averaging over  $\mathbf{y} \in S$ :

$$\begin{aligned} \frac{1}{|S|} \sum_{\mathbf{x}, \mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|_2^2 &= \sum_{\mathbf{y} \in S} \|\mathbf{y} - \text{mean}(S)\|_2^2 + \sum_{\mathbf{x} \in S} \|\mathbf{x} - \text{mean}(S)\|_2^2 \\ &= 2 \sum_{\mathbf{x} \in S} \|\mathbf{x} - \text{mean}(S)\|_2^2. \end{aligned}$$

18

## Dimension reduction for $k$ -means

Let  $S$  be partitioned into  $S_1, S_2, \dots, S_k$  by  $C = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ .

- ▶ Assume  $\mathbf{y}_i = \text{mean}(S_i)$  (i.e.,  $C$  is locally optimal).
- ▶ Bias-variance implies

$$\begin{aligned}\text{cost}(S, C) &= \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \text{mean}(S_i)\|_2^2 \\ &= \sum_{i=1}^k \frac{1}{2|S_i|} \sum_{\mathbf{x}, \mathbf{x}' \in S_i} \|\mathbf{x} - \mathbf{x}'\|_2^2,\end{aligned}$$

so cost only depends on pairwise distances between data.

- ▶ Can thus reduce dimension (using JL) to  $O(\log(n)/\varepsilon^2)$  and preserve cost of all locally-optimal solutions up to  $1 \pm \varepsilon$  factor.
- ▶ Also implies that we cannot expect  $\text{poly}(n, k, 2^{O(d)})$ -time exact algorithm for  $k$ -means.

19

$D^2$  sampling

20

## $D^2$ sampling

**Problem.** Lloyd's algorithm requires good initialization.

$D^2$  sampling /  $k$ -means++ (Arthur and Vassilvitskii, 2007)

- ▶ Pick  $\mathbf{Y}_1$  u.a.r. from  $S$ , and set  $C_1 := \{\mathbf{Y}_1\}$ .
- ▶ For  $t = 2, 3, \dots$ :
  - ▶ Pick  $\mathbf{Y}_t \sim p_t$ , where

$$p_t(\mathbf{y}) = \frac{\text{cost}(\{\mathbf{y}\}, C_{t-1})}{\text{cost}(S, C_{t-1})} \quad \text{for each } \mathbf{y} \in S.$$

- ▶ **Theorem.**

$$\mathbb{E} \text{cost}(S, C_k) \leq O(\log k) \cdot \min_{C \subseteq \mathbb{R}^d: |C| \leq k} \text{cost}(S, C).$$

21

## Analysis of the first center selection

- ▶ Let  $C^* := \{\mu_1, \mu_2, \dots, \mu_k\}$  be optimal solution, and let  $A_1, A_2, \dots, A_k$  be partitioning of  $S$  with respect to  $C^*$ .
- ▶ First analyze  $\mathbf{Y}_1$ , which is distributed uniformly at random in  $S$ .
- ▶ **Claim.**

$$\mathbb{E}[\text{cost}(A_i, C_1) \mid \{\mathbf{Y}_1 \in A_i\}] = 2 \text{cost}(A_i, C^*).$$

- ▶ **Proof.** By bias-variance,

$$\begin{aligned} & \mathbb{E} \left[ \sum_{\mathbf{x} \in A_i} \|\mathbf{x} - \mathbf{Y}_1\|_2^2 \mid \{\mathbf{Y}_1 \in A_i\} \right] \\ &= \mathbb{E} \left[ \sum_{\mathbf{x} \in A_i} \|\mathbf{x} - \mu_i\|_2^2 + |A_i| \cdot \|\mathbf{Y}_1 - \mu_i\|_2^2 \mid \{\mathbf{Y}_1 \in A_i\} \right] \\ &= 2 \sum_{\mathbf{x} \in A_i} \|\mathbf{x} - \mu_i\|_2^2. \quad \square \end{aligned}$$

- ▶ (Lose factor of two by restricting centers to data points.)

22

## Selection of subsequent centers

- ▶ Now consider  $\mathbf{Y}_t$  for  $t > 1$  (conditional on  $C_{t-1}$ ).
- ▶ Distribution of  $\mathbf{Y}_t$  not necessarily uniform in  $S$ .
  - ▶ Points farther from  $C_{t-1}$  get higher weight in  $p_t$ .
- ▶ Write, for  $\mathbf{y} \in A_i$ ,

$$p_t(\mathbf{y}) = \underbrace{\frac{\text{cost}(\{\mathbf{y}\}, C_{t-1})}{\text{cost}(A_i, C_{t-1})}}_{=: p_t(\mathbf{y}|A_i)} \cdot \underbrace{\frac{\text{cost}(A_i, C_{t-1})}{\text{cost}(S, C_{t-1})}}_{=: p_t(A_i)}.$$

- ▶ **Claim** (non-uniformity bound). For  $\mathbf{y} \in A_i$ ,

$$p_t(\mathbf{y} | A_i) \leq \frac{2}{|A_i|} \left( 1 + \frac{\text{cost}(A_i, \{\mathbf{y}\})}{\text{cost}(A_i, C_{t-1})} \right).$$

- ▶ **Claim** (cost bound).

$$\mathbb{E}[\text{cost}(A_i, C_{t-1} \cup \{\mathbf{Y}_t\}) | \{\mathbf{Y}_t \in A_i\}, C_{t-1}] \leq 8 \text{cost}(A_i, C^*).$$

23

## Non-uniformity bound

### Proof of non-uniformity bound.

- ▶ For any  $\mathbf{x} \in A_i$ ,
 
$$\text{cost}(\{\mathbf{y}\}, C_{t-1}) \leq \text{cost}(\{\mathbf{y}\}, \{C_{t-1}(\mathbf{x})\}) = \|\mathbf{y} - C_{t-1}(\mathbf{x})\|_2^2.$$

- ▶ By triangle inequality,

$$\text{cost}(\{\mathbf{y}\}, C_{t-1}) \leq 2 \left( \|\mathbf{x} - C_{t-1}(\mathbf{x})\|_2^2 + \|\mathbf{x} - \mathbf{y}\|_2^2 \right).$$

- ▶ Now average with respect to  $\mathbf{x} \in A_i$ :

$$\text{cost}(\{\mathbf{y}\}, C_{t-1}) \leq \frac{2}{|A_i|} \text{cost}(A_i, C_{t-1}) + \frac{2}{|A_i|} \text{cost}(A_i, \{\mathbf{y}\}).$$

- ▶ So

$$p_t(\mathbf{y} | A_i) = \frac{\text{cost}(\{\mathbf{y}\}, C_{t-1})}{\text{cost}(A_i, C_{t-1})} \leq \frac{2}{|A_i|} \left( 1 + \frac{\text{cost}(A_i, \{\mathbf{y}\})}{\text{cost}(A_i, C_{t-1})} \right). \quad \square$$

24

## Cost bound

### Proof of cost bound.

- ▶ Expected cost:

$$\sum_{\mathbf{y} \in A_i} p_t(\mathbf{y} \mid A_i) \cdot \text{cost}(A_i, C_{t-1} \cup \{v\mathbf{y}\})$$

- ▶ Using non-uniformity bound on  $p_t(\cdot \mid A_i)$ :

$$\leq \sum_{\mathbf{y} \in A_i} \frac{2}{|A_i|} \left( 1 + \frac{\text{cost}(A_i, \{\mathbf{y}\})}{\text{cost}(A_i, C_{t-1})} \right) \cdot \text{cost}(A_i, C_{t-1} \cup \{\mathbf{y}\})$$

- ▶ Using  $\text{cost}(A_i, C_{t-1} \cup \{\mathbf{y}\}) \leq \min\{\text{cost}(A_i, \{\mathbf{y}\}), \text{cost}(A_i, C_{t-1})\}$ :

$$\begin{aligned} &\leq \frac{4}{|A_i|} \sum_{\mathbf{y} \in A_i} \text{cost}(A_i, \{\mathbf{y}\}) = 8 \text{cost}(A_i, \text{mean}(A_i)) \\ &= 8 \text{cost}(A_i, C^*). \end{aligned}$$

□

25

## Cost of uncovered clusters

- ▶ So for any  $t$ ,

$$\mathbb{E}[\text{cost}(A_i, C_{t-1} \cup \{\mathbf{Y}_t\}) \mid \{\mathbf{Y}_t \in A_i\}, C_{t-1}] \leq 8 \text{cost}(A_i, C^*).$$

- ▶ **Problem:** some  $\mathbf{Y}_t$  land in already covered  $A_i$ .
- ▶ Define “good” and “bad” points:

$$\text{good (covered): } G_t := \bigcup_{i: A_i \cap C_t \neq \emptyset} A_i, \quad g_t := |\{i : A_i \cap C_t \neq \emptyset\}|,$$

$$\text{bad (uncovered): } B_t := \bigcup_{i: A_i \cap C_t = \emptyset} A_i, \quad b_t := |\{i : A_i \cap C_t = \emptyset\}|.$$

And define potential function

$$\Phi_t := \frac{t - g_t}{b_t} \text{cost}(B_t, C_t).$$

- ▶ Since  $g_k + b_k = k$ ,

$$\text{cost}(S, C_k) = \text{cost}(G_k, C_k) + \Phi_k.$$

26

## Change in uncovered clusters potential

- **Claim** (proof omitted).

$$\mathbb{E}[\Phi_{t+1} - \Phi_t \mid \{\mathbf{Y}_{t+1} \in B_t\}, C_t] \leq 0,$$

$$\mathbb{E}[\Phi_{t+1} - \Phi_t \mid \{\mathbf{Y}_{t+1} \in G_t\}, C_t] \leq \frac{\text{cost}(B_t, C_t)}{b_t}.$$

- Using this claim, it follows that

$$\begin{aligned} \mathbb{E}[\Phi_{t+1} - \Phi_t \mid C_t] &\leq \mathbb{P}(\mathbf{Y}_{t+1} \in G_t \mid C_t) \cdot \frac{\text{cost}(B_t, C_t)}{b_t} \\ &= \frac{\text{cost}(G_t, C_t)}{\text{cost}(S, C_t)} \cdot \frac{\text{cost}(B_t, C_t)}{b_t} \\ &\leq \frac{\text{cost}(G_t, C_t)}{k - t}. \end{aligned}$$

- Conclude that

$$\mathbb{E}[\Phi_k] \leq \mathbb{E}[\text{cost}(G_k, C_k)] \cdot (1 + 1/2 + 1/3 + \cdots + 1/k).$$

27

## Overall approximation bound

Use fact that  $\mathbb{E}[\text{cost}(G_k, C_k)] \leq 8 \text{cost}(S, C^*)$  to conclude:

$$\begin{aligned} \mathbb{E}[\text{cost}(S, C_k)] &= \mathbb{E}[\text{cost}(G_k, C_k) + \Phi_k] \\ &\leq 8 \text{cost}(S, C^*) \cdot (1 + H_k), \end{aligned}$$

where  $H_k = 1 + 1/2 + 1/3 + \cdots + 1/k$  is the  $k$ -th harmonic sum.  $\square$

28

## Bi-criteria approximation

29

### Bi-criteria guarantees for $D^2$ sampling

- ▶ Let  $C^*$  be optimal set of  $k$  centers for  $S$ .
- ▶ Algorithm provides  $(\alpha, \beta)$ -approximation if it returns  $\hat{C}$  with

$$|\hat{C}| \leq \alpha \cdot k, \quad \text{cost}(S, \hat{C}) \leq \beta \cdot \text{cost}(S, C^*).$$

- ▶ Akin to *proper* ( $\alpha = 1$ ) and *improper* ( $\alpha > 1$ ) learning.
- ▶  $D^2$  sampling provides (proper)  $(1, O(\log k))$ -approximation.
  - ▶ Also provides  $(O(1), O(1))$ -approximation!
  - ▶ Tight analysis:  $(O(1/\varepsilon^2), 2 + \varepsilon)$ -approximation (Wei, 2016).

30

## Simple bi-criteria analysis

- Define “good” and “bad” points:

$$\begin{aligned} \text{good: } G_t &:= \bigcup_{\substack{i \in \{1,2,\dots,k\}: \\ \text{cost}(A_i, C_t) \leq 16 \text{cost}(A_i, \{\mu_i\})}} A_i, \\ \text{bad: } B_t &:= \bigcup_{\substack{i \in \{1,2,\dots,k\}: \\ \text{cost}(A_i, C_t) > 16 \text{cost}(A_i, \{\mu_i\})}} A_i. \end{aligned}$$

- **Claim.** At least one of the following is true:

$$\begin{aligned} \text{cost}(S, C_t) &\leq 32 \text{cost}(S, C^*), \\ p_t(B_t) &\geq \frac{1}{2}. \end{aligned}$$

- **Proof.** If  $\text{cost}(S, C_t) > 32 \text{cost}(S, C^*)$ , then

$$p_t(B_t) = 1 - \frac{\text{cost}(G_t, C_t)}{\text{cost}(S, C_t)} \geq 1 - \frac{16 \text{cost}(G_t, C^*)}{32 \text{cost}(S, C^*)} \geq \frac{1}{2}. \quad \square$$

31

## Simple bi-criteria analysis (continued)

- Say round  $t$  is a “success” if
  - either  $\text{cost}(S, C_{t-1}) \leq 32 \text{cost}(S, C^*)$  already,
  - or  $\mathbf{Y}_t \in A_i \subseteq B_{t-1}$  for some cluster  $i$ , and

$$\text{cost}(A_i, C_t) \leq 16 \text{cost}(A_i, C^*) \quad (\text{i.e., } A_i \subseteq G_t).$$

- **Claim.** Round  $t$  succeeds with probability  $1/4$  (given  $C_{t-1}$ ).

- **Proof.**

- If first success criterion does not hold, then

$$p_{t-1}(B_{t-1}) \geq \frac{1}{2}.$$

- Furthermore, by Markov’s inequality and cost bound,

$$\mathbb{P}(\text{cost}(A_i, C_t) \leq 16 \text{cost}(A_i, C^*) \mid \{\mathbf{Y}_t \in A_i\}, C_{t-1}) \geq \frac{1}{2}. \quad \square$$

- $k$  success rounds guarantee  $\text{cost}(S, C_t) \leq 32 \text{cost}(S, C^*)$ ; this happens within  $t \leq 8k$  rounds with probability  $1 - e^{-\Omega(k)}$ .  $\square$

32



## Final remarks

- ▶ Can post-process the  $8k$  centers by solving LP to get proper  $O(1)$ -approximation (Aggarwal, Deshpande, Kannan, 2009).
- ▶ Different local search gets proper  $(9 + \epsilon)$ -approximation for any constant  $\epsilon > 0$  (Kanungo et al, 2003).
  - ▶ But seems to perform worse than  $D^2$  sampling in practice.
  - ▶ Can this be explained?
- ▶ Nearly all reasonable methods with theoretical analysis only pick centers from among data, thereby losing factor two in approximation. Can this be avoided?