Clustering

Daniel Hsu

COMS 4772

Finitely representing large sets

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

- I.e., $\rho \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ 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).

Covering / net formulations

k-center clustering

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

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

  where $\rho(\boldsymbol{x}, S) := \min_{\boldsymbol{y} \in S} \rho(\boldsymbol{x}, \boldsymbol{y})$.
- Determines ε in ε-net criterion.
- NP-hard optimization problem.

Farthest-first traversal (Gonzalez, 1985)

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

$$\boldsymbol{y}_t \ \in \ \arg\max_{\boldsymbol{x} \in S} \rho(\boldsymbol{x}, \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_i\}) \,.$$

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

Approximation analysis of farthest-first

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

$$r_k \ = \ \rho(\boldsymbol{y}_{k+1}, \widehat{C}) \ = \ \max_{\boldsymbol{x} \in S} \rho(\boldsymbol{x}, \widehat{C}) \ = \ \mathrm{cost}(S, \widehat{C}) \,.$$

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

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

- So $\mathrm{cost}_\infty(S, \widehat{C}) = r_k \leq 2 \cdot \mathrm{cost}_\infty(S, C)$. $\square$

ε-nets

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

$$r_k \;=\; \operatorname{cost}(S, \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}) \;\leq\; \varepsilon\,.$$

- Then $\widehat{C} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$ satisfies

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

- Size of smallest ε-net is called *covering number of S* (at scale ε, with respect to ρ metric).

Set cover

- **Goal**: given set S, family of subsets $\mathcal{F} := \{S_i : i \in \mathcal{I}\} \subseteq 2^S$, pick $S_{i_1}, S_{i_2}, \ldots, S_{i_k}$, 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.

Greedy algorithm

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

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

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

Analysis of greedy algorithm (Johnson, 1974)

- Suppose $B_{i_1^\star}, B_{i_2^\star}, \ldots, B_{i_k^\star}$ covers S.
- After t steps of greedy, we have picked $B_{i_1}, B_{i_2}, \ldots, B_{i_t}$.
  - Let $n_t := |S \setminus \bigcup_{j=1}^{t} B_{i_j}|$ be the number of points in S not covered after t steps.
  - We know $B_{i_1^\star}, B_{i_2^\star}, \ldots, B_{i_k^\star}$ 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. $\square$

Average cost formulations

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**: $\text{cost}(S, C) = \sum_{x \in S} \rho(\boldsymbol{x}, S)$.
  - k-**means**: $\text{cost}(S, C) = \sum_{x \in S} \rho(\boldsymbol{x}, S)^2$.

k-means

- $\mathcal{X} = \mathbb{R}^d$, $\rho(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$.
  - $\text{cost}(S, C) = \sum_{x \in S} \min_{\boldsymbol{y} \in C} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
- 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

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

  implies solution is mean(S).
  - Approximation schemes available when $d = O(1)$ or $k = O(1)$.

General case

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

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

Local search algorithm (Lloyd, 1982)

- Start with $C = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$; repeat:
  - Partition S into $S_1, S_2, \ldots, S_k$ using C.
  - Set $C := \{\text{mean}(S_i) : i = 1, 2, \ldots, k\}$.
- Alternative: start with partition of S into $S_1, S_2, \ldots, S_k$.
- 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.

Bad case for Lloyd's algorithm

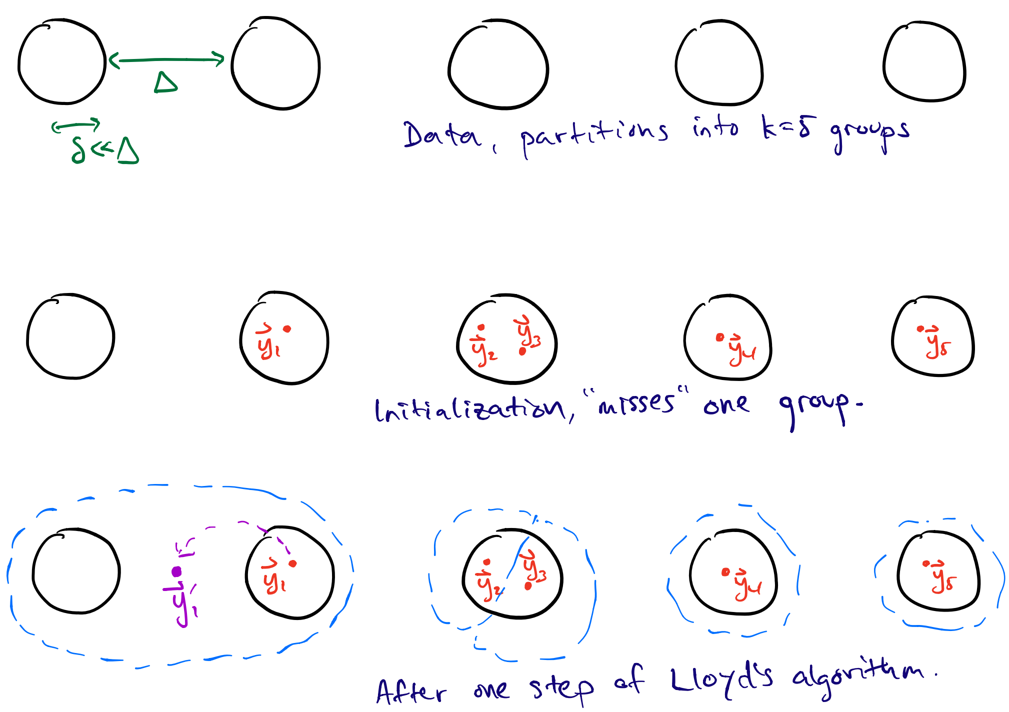

Figure 1: Bad case for Lloyd's algorithm

Aside: dimension reduction

Another look at bias-variance

$$\sum_{\boldsymbol{x} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \; = \; |S| \cdot \|\boldsymbol{y} - \mathrm{mean}(S)\|_2^2 + \sum_{\boldsymbol{x} \in S} \|\boldsymbol{x} - \mathrm{mean}(S)\|_2^2 \, .$$

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

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

Dimension reduction for k-means

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

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

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

  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.

D^2 sampling

D^2 sampling

Problem. Lloyd's algorithm requires good initialization.

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

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

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

- **Theorem**.

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

Analysis of the first center selection

- Let $C^\star := \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$ be optimal solution, and let $A_1, A_2, \ldots, A_k$ be partitioning of S with respect to $C^\star$.
- First analyze $\boldsymbol{Y}_1$, which is distributed uniformly at random in S.
- **Claim**.

$$\mathbb{E}\big[\operatorname{cost}(A_i, C_1) \mid \{\boldsymbol{Y}_1 \in A_i\}\big] \;=\; 2 \operatorname{cost}(A_i, C^\star)\,.$$

- **Proof**. By bias-variance,

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

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

Selection of subsequent centers

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

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

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

$$p_t(\boldsymbol{y} \mid A_i) \;\leq\; \frac{2}{|A_i|}\left(1 + \frac{\mathrm{cost}(A_i, \{\boldsymbol{y}\})}{\mathrm{cost}(A_i, C_{t-1})}\right) .$$

- **Claim** (cost bound).

$$\mathbb{E}\left[\mathrm{cost}(A_i, C_{t-1} \cup \{\boldsymbol{Y}_t\}) \mid \{\boldsymbol{Y}_t \in A_i\}, C_{t-1}\right] \;\leq\; 8\,\mathrm{cost}(A_i, C^\star) .$$

Non-uniformity bound

Proof of non-uniformity bound.

- For any $\boldsymbol{x} \in A_i$,

$$\mathrm{cost}(\{\boldsymbol{y}\}, C_{t-1}) \;\leq\; \mathrm{cost}(\{\boldsymbol{y}\}, \{C_{t-1}(\boldsymbol{x})\}) \;=\; \|\boldsymbol{y} - C_{t-1}(\boldsymbol{x})\|_2^2 .$$

- By triangle inequality,

$$\mathrm{cost}(\{\boldsymbol{y}\}, C_{t-1}) \;\leq\; 2\Big(\|\boldsymbol{x} - C_{t-1}(\boldsymbol{x})\|_2^2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\Big) .$$

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

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

- So

$$p_t(\boldsymbol{y} \mid A_i) \;=\; \frac{\mathrm{cost}(\{\boldsymbol{y}\}, C_{t-1})}{\mathrm{cost}(A_i, C_{t-1})} \;\leq\; \frac{2}{|A_i|}\left(1 + \frac{\mathrm{cost}(A_i, \{\boldsymbol{y}\})}{\mathrm{cost}(A_i, C_{t-1})}\right) . \qquad \square$$

Cost bound

Proof of cost bound.

- Expected cost:

$$\sum_{\boldsymbol{y}\in A_i} p_t(\boldsymbol{y} \mid A_i)\cdot \mathrm{cost}(A_i, C_{t-1}\cup\{vy\})$$

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

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

- Using $\mathrm{cost}(A_i, C_{t-1}\cup\{\boldsymbol{y}\}) \leq \min\{\mathrm{cost}(A_i,\{\boldsymbol{y}\}), \mathrm{cost}(A_i, C_{t-1})\}$:

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

$\square$

Cost of uncovered clusters

- So for any t,

$$\mathbb{E}\left[\mathrm{cost}(A_i, C_{t-1}\cup\{\boldsymbol{Y}_t\}) \mid \{\boldsymbol{Y}_t\in A_i\}, C_{t-1}\right] \leq 8\,\mathrm{cost}(A_i, C^\star)\,.$$

- **Problem**: some $\boldsymbol{Y}_t$ land in already covered A_i.
- Define "good" and "bad" points:

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

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

And define potential function

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

- Since $g_k + b_k = k$,

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

Change in uncovered clusters potential

- **Claim** (proof omitted).

$$\mathbb{E}\big[\Phi_{t+1} - \Phi_t \mid \{\boldsymbol{Y}_{t+1} \in B_t\},\, C_t\big] \;\le\; 0\,,$$
$$\mathbb{E}\big[\Phi_{t+1} - \Phi_t \mid \{\boldsymbol{Y}_{t+1} \in G_t\},\, C_t\big] \;\le\; \frac{\mathrm{cost}(B_t, C_t)}{b_t}\,.$$

- Using this claim, it follows that

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

- Conclude that

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

Overall approximation bound

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

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

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

Bi-criteria approximation

Bi-criteria guarantees for D^2 sampling

- Let $C^\star$ be optimal set of k centers for S.
- Algorithm provides (α, β)-*approximation* if it returns $\widehat{C}$ with

$$|\widehat{C}| \;\leq\; \alpha \cdot k\,, \qquad \mathrm{cost}(S, \widehat{C}) \;\leq\; \beta \cdot \mathrm{cost}(S, C^\star)\,.$$

- 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).

Simple bi-criteria analysis

- Define "good" and "bad" points:

$$\text{good:} \quad G_t := \bigcup_{\substack{i \in \{1,2,\dots,k\}: \\ \text{cost}(A_i, C_t) \le 16\,\text{cost}(A_i, \{\boldsymbol{\mu}_i\})}} A_i \,,$$

$$\text{bad:} \quad B_t := \bigcup_{\substack{i \in \{1,2,\dots,k\}: \\ \text{cost}(A_i, C_t) > 16\,\text{cost}(A_i, \{\boldsymbol{\mu}_i\})}} A_i \,.$$

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

$$\text{cost}(S, C_t) \le 32\,\text{cost}(S, C^\star)\,,$$

$$p_t(B_t) \ge \frac{1}{2}\,.$$

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

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

Simple bi-criteria analysis (continued)

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

$$\text{cost}(A_i, C_t) \le 16\,\text{cost}(A_i, C^\star) \qquad (\text{i.e., } A_i \subseteq G_t)\,.$$

- **Claim**. Round t succeeds with probability $1/4$ (given C_{t-1}).
- **Proof**.
  - If first success criterion does not hold, then

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

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

$$\mathbb{P}\big(\text{cost}(A_i, C_t) \le 16\,\text{cost}(A_i, C^\star) \mid \{\boldsymbol{Y}_t \in A_i\}, C_{t-1}\big) \ge \frac{1}{2}\,. \quad \square$$

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

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?