This Lecture

- Core-sets for k-median
Coreset for 1-means

**Given:** a point set \( P \subset \mathbb{R}^d \)

**Find:** \( C \subseteq P \) such that for any query \( q \in \mathbb{R}^d \)

\[
\text{Cost}(P, q) \approx \text{Cost}(C, q)
\]

• A coreset from which we can estimate the 1-means cost, e.g.,

\[
\text{Cost}(P, q) = \sum_{p \in P} \text{dist}(p, q)^2
\]

• \( \sum_{p \in P} \|p - q\|^2 = \sum_{p \in P} \langle p - q, p - q \rangle = \sum_{p \in P} \|p\|^2 + n\|q\|^2 - 2q \sum_{p \in P} p \)

• So only keep the mean \( \sum_{p \in P} p \)

• Not a core-set exactly.
Coreset for k-center

**Given:** a point set \( P \subset \mathbb{R}^d \)

**Find:** \( C \subseteq P \) such that for any query \( q \in \mathbb{R}^d \)

\[
Cost(P, q) \approx Cost(C, q)
\]

- We showed a coreset from which we can estimate the 1-center cost, e.g., \( Cost(P, q) = far(P, q) = \max_{p \in P} \text{dist}(p, q) \)
- What about \( k \)-center cost?
Naïve Uniform Sampling

**Given:** a point set $P \subset \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$\text{Cost}(P, q) \approx \text{Cost}(C, q)$$
Importance Sampling

**Given:** a point set $P \subset \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

- $Pr \approx 1/n_i$ proportional to size of the cluster
- $weight \approx n_i$ proportional to the size
Importance Sampling

Given: a point set $P \subset \mathbb{R}^d$

Find: $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$Cost(P, q) \approx Cost(C, q)$$

• $Pr \approx 1/n_i$ proportional to size of the cluster
• weight $\approx n_i$ proportional to the size
• Do we need the clusters?
Importance Sampling

**Given:** a point set $P \subset \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

\[ Cost(P, q) \approx Cost(C, q) \]

- $Pr \approx 1/n_i$ proportional to size of the cluster
- *weight* $\approx n_i$ proportional to the size
- Do we need the clusters?
  - Answer: some approximation suffices
Importance Sampling

**Given:** a point set $P \subset \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$\text{Cost}(P, q) \approx \text{Cost}(C, q)$$

- $Pr \approx 1/n_i$ proportional to size of the cluster
- *weight* $\approx n_i$ proportional to the size
- Do we need the clusters?
  - Answer: some approximation suffices
  - Even bi-criteria approximation (pick more centers)

Low Probability
Large weight

High Probability
Low weight
General approach

1. Find an approximate clustering (usually much easier)
2. Sample points based on their cluster size
Even use previous approach

$k$-center

1. Find an approximate clustering (usually much easier)
2. Apply the grid on each cluster

• Exponential dependence on $d$
Coreset for 1-means

**Given:** a point set $P \subset \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $q \in \mathbb{R}^d$

$$\text{Cost}(P, q) \approx \text{Cost}(C, q)$$

- A coreset from which we can estimate the 1-means cost, e.g.,

$$\text{Cost}(P, q) = \sum_{p \in P} \text{dist}(p, q)^2$$
Coreset for k-median

**Given:** a point set $P \subset \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $Q \subseteq (\mathbb{R}^d)^k$

$\text{Cost}(P, Q) \approx \text{Cost}(C, Q)$

• A coreset from which we can estimate the k-median cost, e.g.,

$\text{Cost}(P, Q) = \sum_{p \in P} \text{dist}(p, Q) = \sum_{p \in P} \min_{q \in Q} \text{dist}(p, q)$

• Opening facilities, test several candidates.
Coreset for k-median

**Given:** a point set $P \subseteq \mathbb{R}^d$

**Find:** $C \subseteq P$ such that for any query $Q \subseteq \left(\mathbb{R}^d\right)^k$

- $\sum_{p \in P} \text{dist}(p, Q) \approx (1 + \epsilon) \sum_{c \in C} w(c) \text{dist}(c, Q)$

- A coreset from which we can estimate the k-median cost, e.g.,

$$Cost(P, Q) = \sum_{p \in P} \text{dist}(p, Q) = \sum_{p \in P} \min_{q \in Q} \text{dist}(p, q)$$
A general theorem

• Suppose the cost function satisfies

\[ \text{Cost}(P, Q) = \sum_{p \in P} w(p) \text{dist}(p, Q) \]

• Sample \( C \) proportional to sensitivity \( (p) = \max_{Q \in Q} \frac{\text{dist}(p, Q)}{\sum_{p' \in P} \text{dist}(p', Q)} \)

• Number of samples: \( |C| \geq O \left( \frac{VC(Q)}{\epsilon^2} \cdot \sum_p \text{sensitivity}(p) \right) \)

• Need to bound
  • \( VC(Q) \): (roughly how many parameters one need to describe the query, e.g., kd)
  • Total sensitivity \( \sum_p \text{sensitivity}(p) \) (for k-median can be bounded by k)
  • Gives coreset of size \( O\left(\frac{k^2 d}{\epsilon^2}\right) \)
Applications

- k-means, k-median, k-center
- j-subspace: query q is a j-dimensional subspace.
- Projective clustering (j,k): query Q is a set of k j-dimensional subspaces.
K-median

- sensitivity\( (p) = \frac{\text{dist}(p,Q^*)}{\sum_{p' \in P} \text{dist}(p',Q^*)} + \frac{1}{n_p} \)
  - \( Q^* \) is the optimal k-means clustering (again we can use approximation)
  - \( n_p \) is the number of points in \( p \)'s cluster

- Total sensitivity = 1 + \( k \)

- \( |C| = \left( \frac{k^2 d}{\epsilon^2} \right) \)

- Combining with PCA gives \( \left( \frac{k^2 \left( \frac{k}{\epsilon} \right)}{\epsilon^2} \right) \)
  - Independent of \( n \)
  - Independent of \( d \)
Proof of the Theorem
Setup

Given: \((P, w), P \subseteq X, w: P \to [0,1]\), sum of weights are 1

- Core-sets for core-sets

Query space: \((P, w, Q, f)\)

- \(f: P \times Q \to [0, \infty)\)

- \(\bar{f}(P, w, Q) = \sum_{p \in P} w(p) \cdot f(p, q)\)

Change multiplicative \(\epsilon\) to additive error \(\epsilon\)

- Goal: find \((C, u, Q, f)\) such that for any \(Q \in Q\):

- \(|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)| \leq \epsilon\)

- Why?

- Let \(f(p, Q) := \frac{\text{dist}(p, Q)}{\text{Cost}(p, Q)} = \frac{\text{dist}(p, Q)}{\sum_{p \in P} w(p) \cdot \text{dist}(p, Q)}\)
Why

We want $|\text{Cost}(P, Q) - \text{Cost}(C, q)| \leq \epsilon \cdot \text{Cost}(P, Q)$

- $|\text{Cost}(P, Q) - \text{Cost}(C, q)| = |\sum_{p \in P} w(p) \cdot \text{dist}(p, Q) - \sum_{p \in C} u(p) \cdot \text{dist}(p, Q)| =$
- $\text{Cost}(P, Q) \cdot |\sum_{p \in P} w(p) \cdot f(p, Q) - \sum_{p \in C} u(p) \cdot f(p, Q)| =$
- $\text{Cost}(P, Q) \cdot |\bar{f}(P, w, Q) - \bar{f}(C, u, Q)|$

Change multiplicative $\epsilon$ to additive error $\epsilon$

- Goal: find $(C, u, Q, f)$ such that for any $Q \in Q$:
- $|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)| \leq \epsilon$
- Why?

- Let $f(p, Q) := \frac{\text{dist}(p, Q)}{\text{Cost}(P, Q)} = \frac{\text{dist}(p, Q)}{\sum_{p \in P} w(p) \cdot \text{dist}(p, Q)}$
Intermediate goal

Given: \((P, w), P \subseteq X, w: P \rightarrow [0,1]\), sum of weights are 1

Query space: \((P, w, Q, f)\)

- \(f: P \times Q \rightarrow [0, \infty)\)
- \(\bar{f}(P, w, Q) = \sum_{p \in P} w(p) \cdot f(p, q)\)

Let \(f(p, Q) := \frac{\text{dist}(p, Q)}{\text{Cost}(P, Q)} = \frac{\text{dist}(p, Q)}{\sum_{p \in P} w(p) \cdot \text{dist}(p, Q)}\)

Intermediate Goal: find \((C, u)\) such that for any \(Q \in Q\):

- \(|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)| \leq \varepsilon \cdot \max_{p \in P} f(p, q)\)

- Why? (roughly, probability bounds work when parameters are between 0,1. otherwise a single input could have the maximum which is quite large. It is also hard to detect using uniform sampling. In other words the variance depends on the maximum).
Intermediate goal

Given: \((P, w), P \subseteq X, w: P \to [0,1]\), sum of weights are 1

Query space: \((P, w, Q, f)\)

- \(f: P \times Q \to [0, \infty)\)
- \(\bar{f}(P, w, Q) = \sum_{p \in P} w(p) \cdot f(p, q)\)

Let \(f(p, Q) := \frac{\text{dist}(p, Q)}{\text{Cost}(p, Q)} = \frac{\text{dist}(p, Q)}{\sum_{p \in P} w(p) \cdot \text{dist}(p, Q)}\)

Intermediate Goal: find \((C, u)\) such that for any \(Q \in Q\):

- \(|\bar{f}(P, w, Q) - \bar{f}(C, u, Q)| \leq \epsilon \cdot \max_{p \in P} f(p, q)\)

Define \(s(p) = \max_{Q \in Q} f(p, Q)\)

Let \(t = \sum_{p \in P} w(p) \cdot s(p)\)

Let \(t = \sum_{p \in P} w(p) \cdot s(p)\)

Now let: \(w'(p) := w(p) \cdot \frac{s(p)}{t}\) and let \(f'(p, Q) := \frac{f(p, Q)}{s(p)}\) Thus \(w(p)f(p, Q) = t \cdot w'(p)f'(p, Q)\)
Core-set for the new weights

• Define \( s(p) = \max_{Q \in Q} f(p, Q) \)
• Let \( t = \sum_{p \in P} w(p) \cdot s(p) \)
• Now let: \( w'(p) := w(p) \cdot \frac{s(p)}{t} \) and let \( f'(p, Q) := \frac{f(p, Q)}{s(p)} \)
  Thus \( w(p)f(p, Q) = t \cdot w'(p)f'(p, Q) \)

Suppose \((C, u)\) is \( \frac{\epsilon}{t} \) coreset for \((P, w', Q, f')\), i.e., for any \( Q \in Q\):
• \( |\bar{f}'(P, w', Q) - \bar{f}'(C, u, Q)| \leq \left(\frac{\epsilon}{t}\right) \cdot \max_{p \in P} f'(p, q) \)

Goal: for any \( Q \in Q\):
• \( |\bar{f}(P, w, Q) - t \cdot \bar{f}(C, u, Q)| \leq \epsilon \)

Proof:
• \( \bar{f}(P, w, Q) = t \cdot \bar{f}'(P, w', Q) \)
• \( |\bar{f}(P, w, Q) - t \cdot \bar{f}(C, u, Q)| = t \cdot |\bar{f}'(P, w', Q) - \bar{f}'(C, u, Q)| \leq t \cdot \left(\frac{\epsilon}{t}\right) \cdot \max_{p \in P} f'(p, q) \leq \epsilon \)
Goal: compute $\epsilon$-approximation for $f'$

- For every positive $r > 0$ define $\text{range}(q, r) = \{p \in P | w(p) \cdot f(p, q) \leq r\}$
- Then the dimension of $(P, w, Q, f)$ is the smallest $d$ s.t. for any $S \subseteq P$,
  - $|\{\text{range}(q, r) | q \in Q, r > 0\}| \leq 2^d$

- E.g. how many subsets can you cover with balls in $\mathbb{R}^d$? $n^{O(d)}$

- Coreset: Let $C$ be a random sample of size $O((\frac{1}{\epsilon^2})(d + \log \frac{1}{\delta}))$, then with probability $(1 - \delta)$, it is a $\epsilon$-core-set

- $dist \rightarrow Cost \rightarrow f \rightarrow s(p) \rightarrow f', w' \rightarrow \epsilon$-approx $\rightarrow$ random sampling
Bounding Sensitivity for k-median

- \( s(p) = \max_{Q \in Q} \frac{\text{dist}(p, Q)}{\sum_{p'} \text{dist}(p', Q)} \)

- For a specific \( Q \),

  \[
  \frac{\text{dist}(p, Q)}{\sum_{p'} \text{dist}(p', Q)} \leq \frac{\text{dist}(p, q_i^*)}{\sum_{p'} \text{dist}(p', q_i^*)} + \frac{\text{dist}(q_i^*, Q)}{\sum_{p'} \text{dist}(p', q_i^*)}
  \]

  \[
  \text{dist}(q_i^*, Q) \leq \text{dist}(q_i^*, p') + \text{dist}(p', Q)
  \]

  \[
  |P_i| \cdot \text{dist}(q_i^*, Q) \leq \sum_{p' \in P_i} \text{dist}(q_i^*, p') + \text{dist}(p', Q) \leq 2 \sum_{p' \in P} \text{dist}(p', Q)
  \]

- \( s(p) \leq \frac{\text{dist}(p, q_i^*)}{\sum_{p'} \text{dist}(p', q_i^*)} + \frac{2}{|P_i|} \)

- \( \sum_{p \in P} s(p) = 1 + 2k \)
K-median

- sensitivity\(p\) = \frac{\text{dist}(p, Q^*)}{\sum_{p' \in P} \text{dist}(p', Q^*)} + \frac{1}{n_p}
  
  - \(Q^*\) is the optimal k-means clustering (again we can use approximation)
  - \(n_p\) is the number of points in \(p\)’s cluster

- Total sensitivity = 1 + \(k\)

- \(|C| = \left(\frac{k^2d}{\epsilon^2}\right)\)

- Combining with PCA gives \(\left(\frac{k^2\left(\frac{k}{\epsilon}\right)}{\epsilon^2}\right)\)
  - Independent of \(n\)
  - Independent of \(d\)
Rough approximation

• Any bi-criteria approximation, e.g.,

Repeat for $\log n$ iterations:
1. Randomly sample $k$ centers.
2. Remove half of the points that are closest to the centers