CS 57800 – Statistical Machine Learning

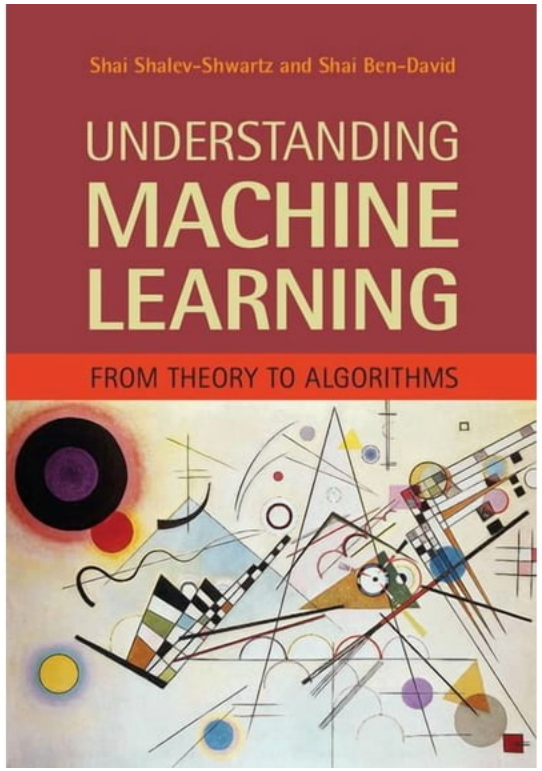
Lecture 2 Unsupervised Learning

https://ruizhezhang.com/course_fall_2026.html

Outline

Unsupervised Learning

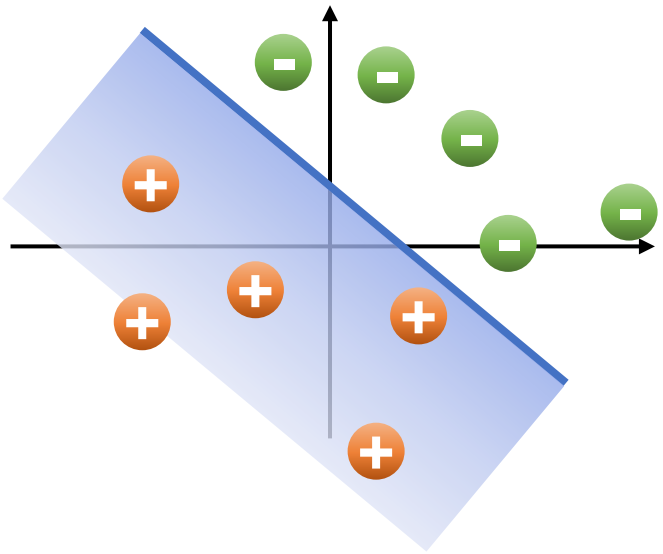
- Intro to Clustering
- K-Means
- Spectral Clustering
- …



Unsupervised Learning

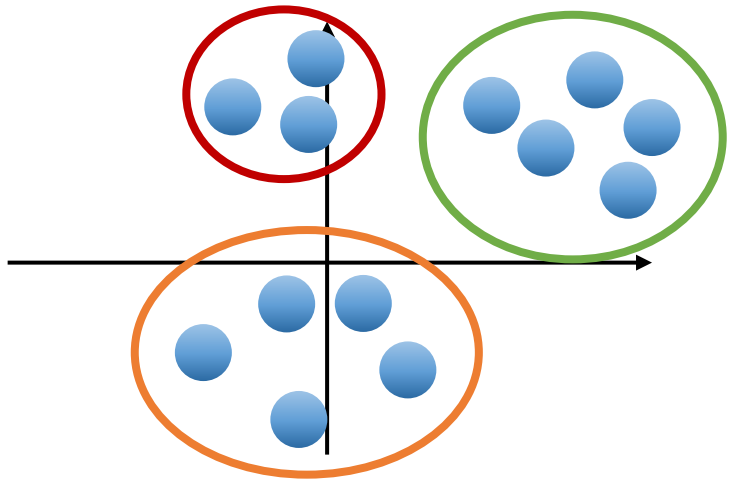
Supervised learning

- Training data has labels
- Make predictions for new data
- Measure success by population/empirical risk



Unsupervised learning

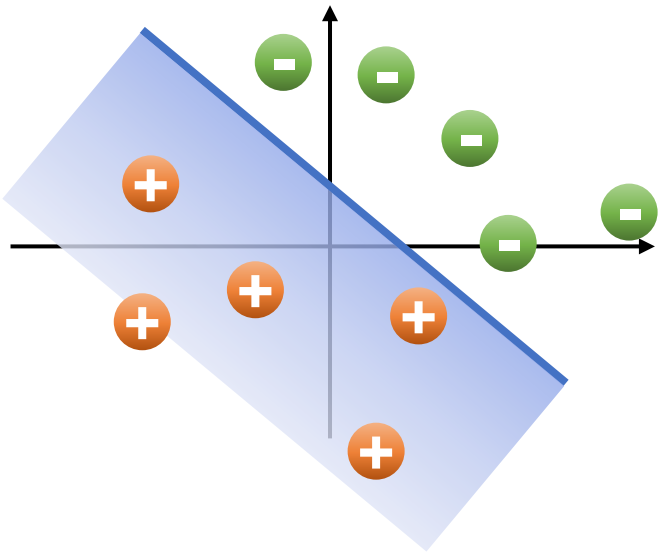
- Data without label
- Identify hidden structures in the data
- No “ground truth”



Unsupervised Learning

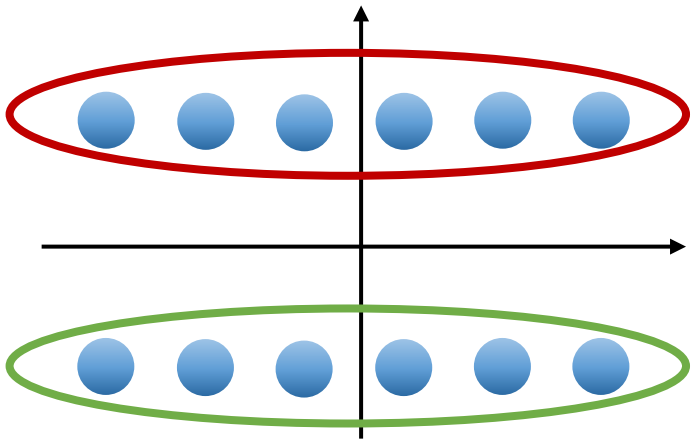
Supervised learning

- Training data has labels
- Make predictions for new data
- Measure success by population/empirical risk



Unsupervised learning

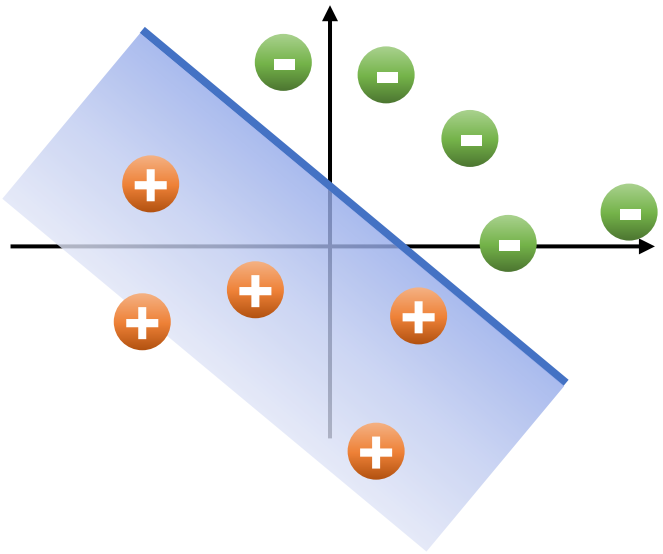
- Data without label
- Identify hidden structures in the data
- No “ground truth”
  - Not separate close-by points:



Unsupervised Learning

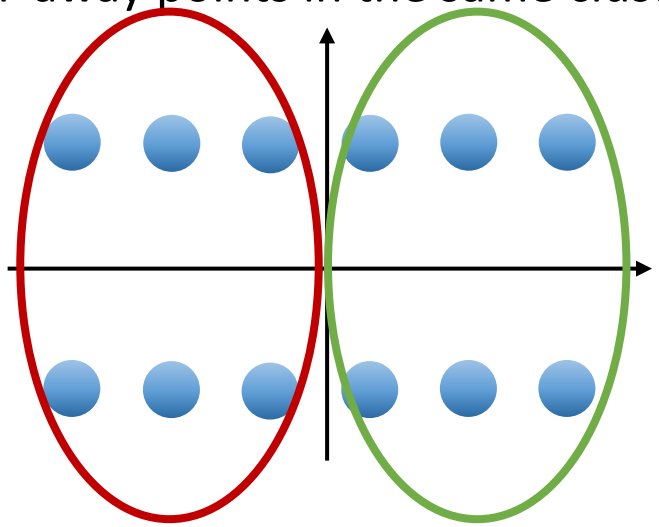
Supervised learning

- Training data has labels
- Make predictions for new data
- Measure success by population/empirical risk



Unsupervised learning

- Data without label
- Identify hidden structures in the data
- No “ground truth”
  - Not separate close-by points
  - No far-away points in the same cluster:



Clustering

- **Input:** a set of elements $\mathcal{X}$, a distance (or similarity) function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ that is symmetric, $d(x, x) = 0 \; \forall x \in \mathcal{X}$, and satisfies triangle inequality, (and a parameter k for the number of required clusters)
- **Output:** a partition of the domain set $C = (C_1, \dots, C_k)$ such that $\mathcal{X} = C_1 \cup \cdots \cup C_k$ and $C_i \cap C_j = \emptyset \; \forall i \neq j$
- **Goal:** an objective function $G\big((\mathcal{X}, d), C\big)$ is minimized

Clustering

Center-based objectives:

- Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotonic cost function
- Define the corresponding objective function:

μ_i is the "center" of C_i

$$G\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) \coloneqq \min_{\mu_1,\dots,\mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} f\big(d(x, \mu_i)\big)$$

where $\mathcal{X}'$ can be $\mathcal{X}$ or some superset of $\mathcal{X}$ (such as $\mathbb{R}^d$)

Example 1: k-means

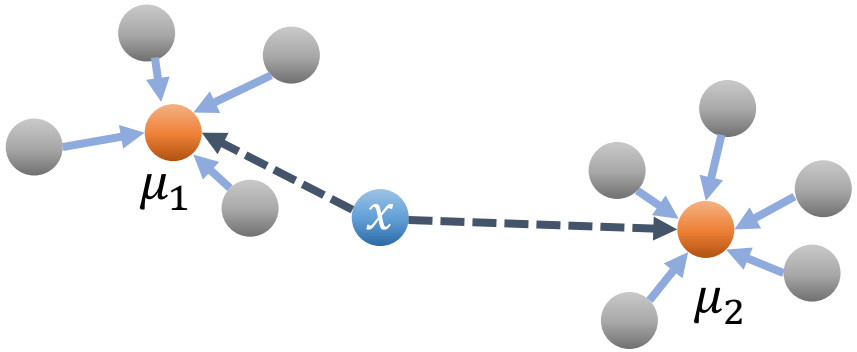
$$G_{k-\text{means}}\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) \coloneqq \min_{\mu_1,\dots,\mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

Clustering

Example 1: k-means

$$\mu_i := \operatorname*{argmin}_{\mu \in X'} \sum_{x \in C_i} d(x, \mu)^2 \,, \qquad G_{k-\text{means}} = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

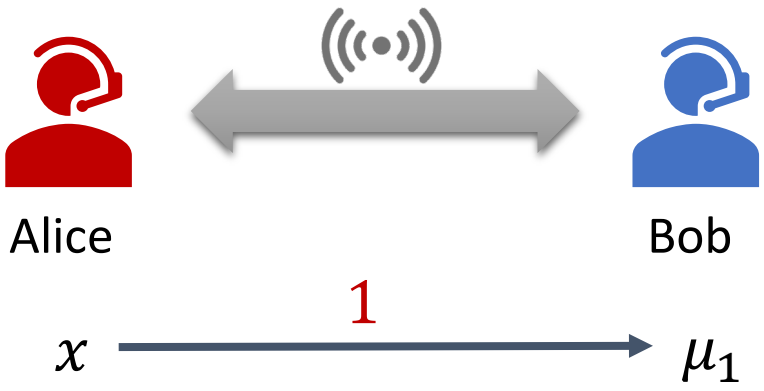
x

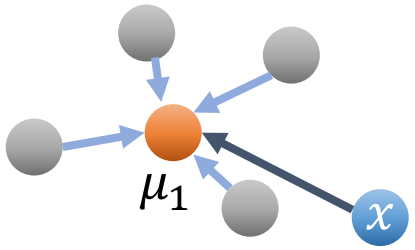


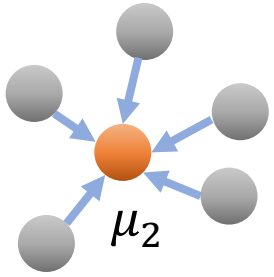
Clustering

Example 1: k-means

$$\mu_i := \operatorname*{argmin}_{\mu \in X'} \sum_{x \in C_i} d(x, \mu)^2 , \qquad G_{k-\text{means}} = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$







Distortion $= d(x, \mu_1)^2$

Squared distance is a convention in signal processing, roughly the "energy" of the signal

Clustering

Center-based objectives:

- Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotonic cost function
- Define the corresponding objective function:

μ_i is the "center" of C_i

$$G\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) \coloneqq \min_{\mu_1,\dots,\mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} f\big(d(x, \mu_i)\big)$$

where $\mathcal{X}'$ can be $\mathcal{X}$ or some superset of $\mathcal{X}$ (such as $\mathbb{R}^d$)

Example 2: k-medoids

$$G_{k-\text{medoids}}\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) \coloneqq \min_{\mu_1,\dots,\mu_k \in \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

Clustering

Center-based objectives:

- Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotonic cost function
- Define the corresponding objective function:

μ_i is the "center" of C_i

$$G\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) \coloneqq \min_{\mu_1,\dots,\mu_k \in \mathcal{X}'} \sum_{i=1}^{k} \sum_{x \in C_i} f\big(d(x, \mu_i)\big)$$

where $\mathcal{X}'$ can be $\mathcal{X}$ or some superset of $\mathcal{X}$ (such as $\mathbb{R}^d$)

Example 3: k-median

$$G_{k-\text{median}}\big((\mathcal{X}, d), (C_1, \dots, C_k)\big) \coloneqq \min_{\mu_1,\dots,\mu_k \in \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)$$

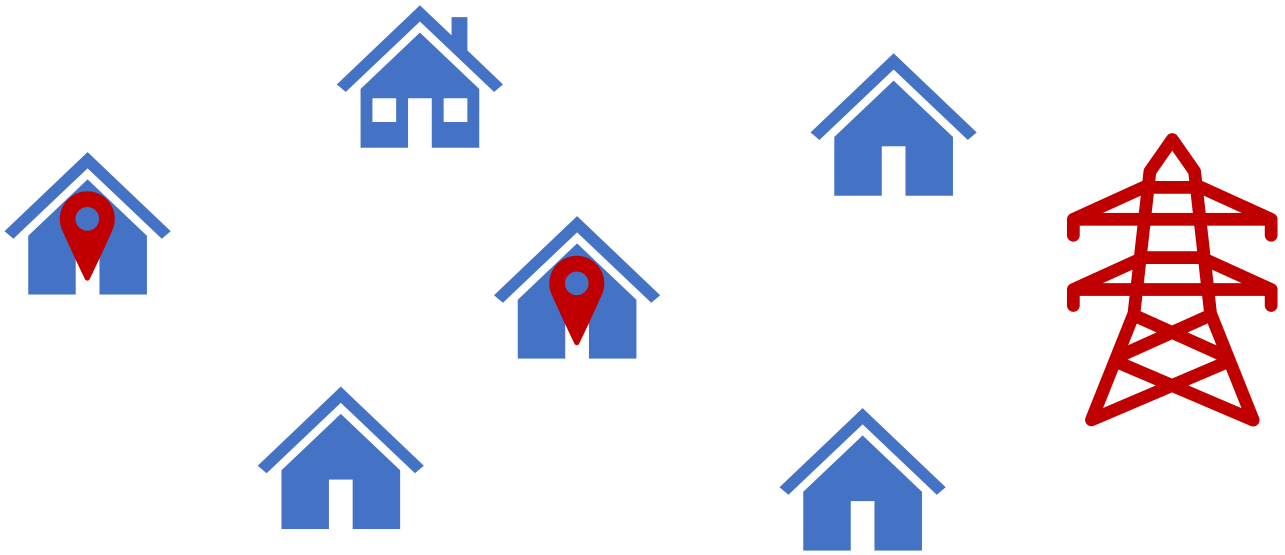
Clustering

Example 3: k-median

$$\mu_i \coloneqq \operatorname*{argmin}_{\mu \in C_i} \sum_{x \in C_i} d(x, \mu)\,, \qquad G_{k-\text{median}} = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)$$

$$G^{\star}_{k-\text{median}} = \min_{\mu_1, \ldots, \mu_k \in \mathcal{X}} \sum_{x} \min_{i \in [k]} d(x, \mu_i)$$

Facility location problem:



$G_{k-\text{median}}$ is the average distance

Clustering

Example 3': k-median

$$\mu_i \coloneqq \operatorname*{argmin}_{\mu \in \mathcal{X}'} \sum_{x \in C_i} d(x, \mu)\,, \qquad G_{k-\text{median}} = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)$$

$$G^{\star}_{k-\text{median}} = \min_{\mu_1, \ldots, \mu_k \in \mathcal{X}'} \sum_{x} \min_{i \in [k]} d(x, \mu_i)$$

Facility location problem:

$G_{k-\text{median}}$ is the average distance

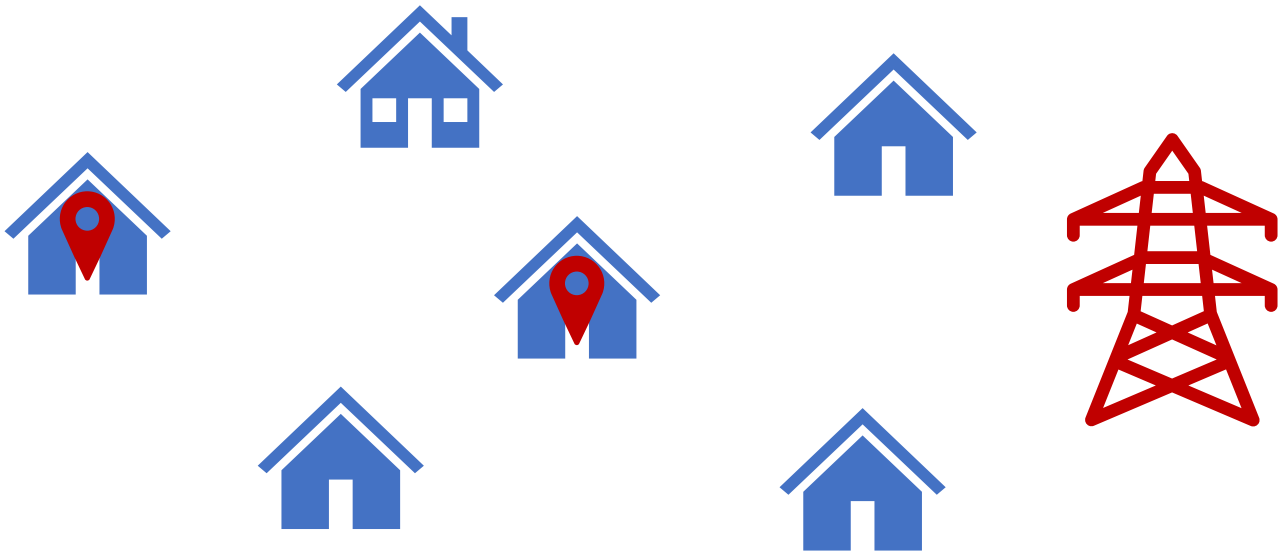
How about the worst distance?

Clustering

Example 4: k-center

$$G^{\star}_{k-\text{center}} = \min_{\mu_1,\ldots,\mu_k \in \mathcal{X}} \max_{x} \min_{i \in [k]} d(x, \mu_i)$$

Facility location problem:



$G_{k-\text{center}}$ is the worst distance

Clustering

- Different objective functions and different distance functions result in different algorithms and complexities
- Clustering has become a subfield in theoretical computer science and machine learning

Exact solutions:

- k-means, k-medoids, k-median, and k-center are **NP**-hard!

Approximate solutions:

- An α-approximation for a minimization problem returns a solution of cost at most
$$\alpha \cdot \mathrm{OPT}$$
- A $(c-\eta)$-hardness of approximation means, for any constant $\eta > 0$, there is no polynomial time $(c-\eta)$-approximation algorithm under certain complexity assumptions

Clustering

$\mu_i \in \mathcal{X}$, general metric space

Problems	Best poly-time approx.	Best hardness of approx.	References
k-centers	2	$2-\eta$	[Hsu-Nemhauser '79, Hochbaum-Shmoys '85,]
k-median	$2+\epsilon$	$1.416-\eta$	[Cohen-Addad-Grandoni-Lee-Schwiegelshohn-Svensson '25, Anand-Lee '24]
k-medoids	$4.9+\epsilon$		[Anand-Charikar-Cohen-Addad-Gao-Grandoni-Lee-Sharma-Wijland '26]

Euclidean space

Problems	Best poly-time approx.	Best hardness of approx.	References
k-means	$3.693+\epsilon$	$1.36-\eta$	[Anand-Charikar-Cohen-Addad-Gao-Grandoni-Lee-Sharma-Wijland '26, Cohen-Addad-C. S.-Lee '21]
k-median	$2+\epsilon$	$1.08-\eta$	

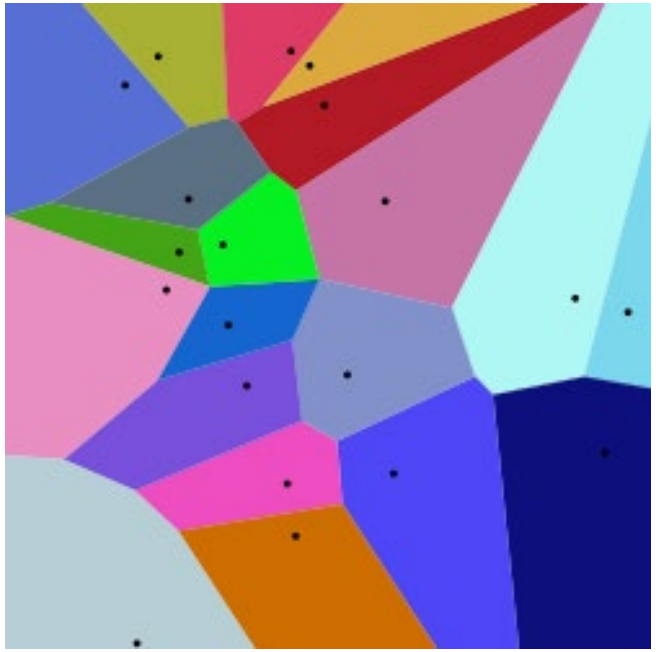
k-Means

Given m points $\mathcal{X} = \{x_1, \dots, x_m\} \subset \mathbb{R}^d$ and $k \in \mathbb{N}_+$, find $\{\mu_1, \dots, \mu_k\} \subset \mathbb{R}^d$ that minimizes the objective function

$$G(\mu_1, \dots, \mu_k) = \sum_{i=1}^{m} \min_{j \in [k]} \|x_i - \mu_j\|^2$$

- The partition $(C_1, \dots, C_k)$ of $\mathcal{X}$ is induced by the centers $(\mu_1, \dots, \mu_k)$
- Moreover, $(\mu_1, \dots, \mu_k)$ gives a Voronoi partition of $\mathbb{R}^d$:
  - $\mathbb{R}^d$ is decomposed into k convex cells $C_1, \dots C_k$
  - We can recover the previous form of the objective function:

$$G(C_{1:k}; \mu_{1:k}) = \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \mu_j)^2$$



k-Means

Lemma (One cluster). For any set $C \subset \mathbb{R}^d$ and any $\mu \in \mathbb{R}^d$,

$$G(C;\mu) = G\left(C;\overline{C}\right) + |C| \cdot \left\|\mu - \overline{C}\right\|^2, \qquad \overline{C} \coloneqq \frac{1}{|C|}\sum_{x\in C} x$$

Proof.

- Consider the following bias-variance decomposition of random vectors:

$$\mathbb{E}[\|X-\mu\|^2] = \mathbb{E}[\|X-\mathbb{E}[X]\|^2] + \|\mu-\mathbb{E}[X]\|^2, \qquad \forall \mu \in \mathbb{R}^d$$

- Define X to be a uniformly random point in C
- Then, $\mathbb{E}[X] = \overline{C}$, and

$$\mathbb{E}[\|X-\mu\|^2] = \frac{1}{|C|}\sum_{x\in C}\|x-\mu\|^2 = \frac{1}{|C|}G(C;\mu)$$

$$\mathbb{E}[\|X-\mathbb{E}[X]\|^2] = \frac{1}{|C|}G\left(C;\overline{C}\right)$$

■

k-Means

Lemma (One cluster). For any set $C \subset \mathbb{R}^d$ and any $\mu \in \mathbb{R}^d$,

$$G(C;\mu) = G\left(C;\overline{C}\right) + |C| \cdot \left\|\mu - \overline{C}\right\|^2, \qquad \overline{C} \coloneqq \frac{1}{|C|}\sum_{x \in C} x$$

- For a single cluster, the optimal choice for μ is the centroid of C

k-Means

k-means algorithm:

1. Initialize centers $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and clusters $C_1, \dots, C_k$
2. Repeat until there is no further change in the objective function:
   i. For each $j \in [k]$: $C_j \leftarrow \{x \in \mathcal{X} \text{ whose closest center is } \mu_j\}$
   ii. For each $j \in [k]$: $z_j \leftarrow \overline{C_j}$

Lemma (Convergence). During the k-means algorithm, the cost monotonically decreases.

Proof.

- Let $\mu_1^{(t)}, \dots, \mu_k^{(t)}, C_1^{(t)}, \dots, C_k^{(t)}$ denote the centers and clusters at the start of the t-th iteration

$$G\left(C_{1:k}^{(t+1)}; \mu_{1:k}^{(t+1)}\right) \leq G\left(C_{1:k}^{(t+1)}; \mu_{1:k}^{(t)}\right) \leq G\left(C_{1:k}^{(t)}; \mu_{1:k}^{(t)}\right)$$

(2.ii.) (2.i.)

■

k-Means

k-means algorithm:

1. Initialize centers $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and clusters $C_1, \dots, C_k$
2. Repeat until there is no further change in the objective function:
   i. For each $j \in [k]$: $C_j \leftarrow \{x \in \mathcal{X} \text{ whose closest center is } \mu_j\}$
   ii. For each $j \in [k]$: $z_j \leftarrow \overline{C_j}$

Lemma (Convergence). During the k-means algorithm, the cost monotonically decreases.

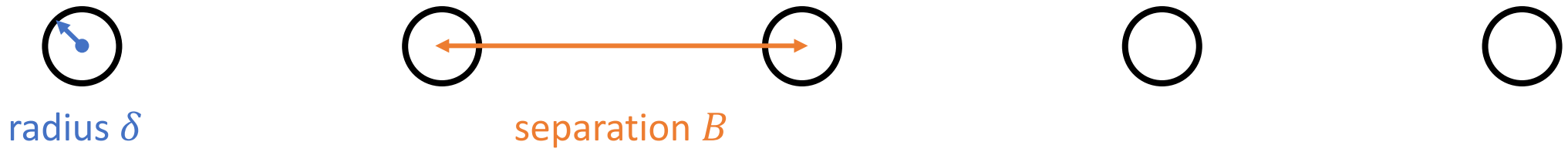
- The k-means algorithm will converge to a local optimum of the objective function

k-Means

k-means algorithm:

1. Initialize centers $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and clusters $C_1, \dots, C_k$
2. Repeat until there is no further change in the objective function:
   i. For each $j \in [k]$: $C_j \leftarrow \{x \in \mathcal{X} \text{ whose closest center is } \mu_j\}$
   ii. For each $j \in [k]$: $z_j \leftarrow \overline{C_j}$

5-cluster example:



$$G_{\text{OPT}} \approx \delta^2 n$$

k-Means

k-means algorithm:

1. Initialize centers $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and clusters $C_1, \dots, C_k$
2. Repeat until there is no further change in the objective function:
   i. For each $j \in [k]$: $C_j \leftarrow \{x \in \mathcal{X} \text{ whose closest center is } \mu_j\}$
   ii. For each $j \in [k]$: $z_j \leftarrow \overline{C_j}$

5-cluster example:

radius δ

separation B

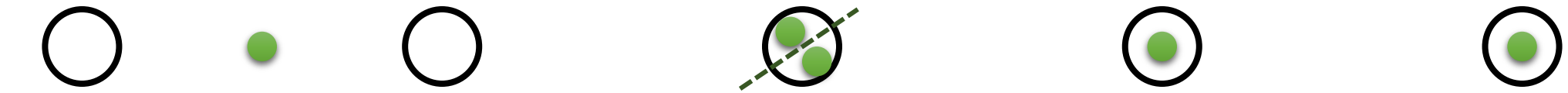
$$G_{\text{OPT}} \approx \delta^2 n$$

k-Means

k-means algorithm:

1. Initialize centers $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ and clusters $C_1, \dots, C_k$
2. Repeat until there is no further change in the objective function:
   i. For each $j \in [k]$: $C_j \leftarrow \{x \in \mathcal{X} \text{ whose closest center is } \mu_j\}$
   ii. For each $j \in [k]$: $z_j \leftarrow \overline{C_j}$

5-cluster example:



radius δ separation B

$$G_{\text{OPT}} \approx \delta^2 n \quad \ll \quad G_{\text{Local}} \approx \Omega(B^2 n)$$

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

Theorem (Arthur-Vassilvitskii '07**).** Let $\mu_1, \dots, \mu_k \in \mathcal{X}$ be the output of k-means++ and $\mu_1^\star, \dots, \mu_k^\star$ be the optimal centers. Then,

$$\mathbb{E}[G(\mu_1, \dots, \mu_k)] \leq O(\log k) \cdot G(\mu_1^\star, \dots, \mu_k^\star)$$

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

Lemma (One cluster). For any set $C \subset \mathbb{R}^d$ and any $\mu \in \mathbb{R}^d$,

$$G(C;\mu) = G(C;\overline{C}) + |C| \cdot \|\mu - \overline{C}\|^2, \qquad \overline{C} \coloneqq \frac{1}{|C|}\sum_{x \in C} x$$

By averaging over $\mu \in C$, we have

$$\mathbb{E}_{\mu \sim C}[G(C;\mu)] = \frac{1}{|C|}\sum_{\mu \in C} G(C;\mu) = G(C;\overline{C}) + \sum_{\mu \in C}\|\mu - \overline{C}\|^2 = 2G(C;\overline{C}) = 2\mathrm{OPT}(C)$$

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

$$\mathbb{E}_{\mu \sim C}[G(C; \mu)] = 2\mathrm{OPT}(C)$$

- Let $C_1^\star, \dots, C_k^\star$ be the optimal clusters induced by $\mu_1^\star, \dots, \mu_k^\star$
- Conditioned on $\mu_1 \in C_i^\star$, we have μ_1 is uniformly sampled from $C_i^\star$, and
$$\mathbb{E}_{\mu_1 \sim C_i}[G(C_i^\star; \mu_1)] = 2\mathrm{OPT}(C_i^\star)$$
- So, the first center is good

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

- Later centers are not uniform within their clusters
- The sampling distribution is biased towards points far away from the current centers
- Nevertheless, we can prove that this biased choice will not select a terrible center

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

Lemma. In the j-th iteration, let μ_j be the sampled center. Then, for any $i \in [k]$,

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] \leq 8\mathrm{OPT}(C_i^\star)$$

Proof.

- Conditioned on $\mu_j \in C_i^\star$, for any $x \in C_i^\star$,

$$\Pr\left[\mu_j = x \middle| \mu_j \in C_i^\star\right] = \frac{G\left(x; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)}$$

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

Lemma. In the j-th iteration, let μ_j be the sampled center. Then, for any $i \in [k]$,

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] \leq 8\mathrm{OPT}(C_i^\star)$$

Proof.

- Conditioned on $\mu_j \in C_i^\star$, for any $\mu \in C_i^\star$,

$$\Pr\left[\mu_j = \mu \middle| \mu_j \in C_i^\star\right] = \frac{G(\mu; \mu_{1:j-1})}{G(C_i^\star; \mu_{1:j-1})}$$

k-Means++

Lemma. In the j-th iteration, let μ_j be the sampled center. Then, for any $i \in [k]$,

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] \leq 8\mathrm{OPT}(C_i^\star)$$

$G\left(C_i^\star; \mu_{1:j}\right)$

Proof.

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] = \sum_{\mu \in C_i^\star} \frac{G\left(\mu; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} \min\left\{G\left(x; \mu_{1:j-1}\right), \|x-\mu\|^2\right\}$$

- For $G\left(\mu; \mu_{1:j-1}\right)$, pick any $x \in C_i^\star$ and let $\mu_x \in \left\{\mu_1, \ldots \mu_{j-1}\right\}$ be the closest center to x

$$\begin{aligned} G\left(\mu; \mu_{1:j-1}\right) \leq \|\mu - \mu_x\|^2 &\leq (\|\mu - x\| + \|x - \mu_x\|)^2 \\ &\leq 2\|\mu - x\|^2 + 2\|x - \mu_x\|^2 \\ &= 2\|\mu - x\|^2 + 2G\left(x; \mu_{1:j-1}\right) \end{aligned}$$

$(a+b)^2 \leq 2a^2 + 2b^2$

- Averaging over $x \in C_i^\star$:

$$G\left(\mu; \mu_{1:j-1}\right) \leq \frac{2}{\left|C_i^\star\right|}\left(G(C_i^\star, \mu) + G\left(C_i^\star; \mu_{1:j-1}\right)\right)$$

k-Means++

Lemma. In the j-th iteration, let μ_j be the sampled center. Then, for any $i \in [k]$,

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] \leq 8\mathrm{OPT}(C_i^\star)$$

Proof.

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] = \sum_{\mu \in C_i^\star} \frac{G\left(\mu; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} \min\left\{G\left(x; \mu_{1:j-1}\right), \|x-\mu\|^2\right\}$$

$$\leq \frac{2}{\left|C_i^\star\right|}\left(\sum_{\mu \in C_i^\star} \frac{G(C_i^\star, \mu)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} G\left(x; \mu_{1:j-1}\right) + \sum_{\mu \in C_i^\star} \frac{G\left(C_i^\star; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} \|x-\mu\|^2\right)$$

$$G\left(\mu; \mu_{1:j-1}\right) \leq \frac{2}{\left|C_i^\star\right|}\left(G(C_i^\star, \mu) + G\left(C_i^\star; \mu_{1:j-1}\right)\right)$$

k-Means++

Lemma. In the j-th iteration, let μ_j be the sampled center. Then, for any $i \in [k]$,

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] \le 8\mathrm{OPT}(C_i^\star)$$

Proof.

$$\mathbb{E}\left[G\left(C_i^\star; \mu_{1:j}\right) \middle| \mu_j \in C_i^\star\right] = \sum_{\mu \in C_i^\star} \frac{G\left(\mu; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} \min\left\{G\left(x; \mu_{1:j-1}\right), \|x-\mu\|^2\right\}$$

$$\le \frac{2}{|C_i^\star|}\left(\sum_{\mu \in C_i^\star} \frac{G(C_i^\star, \mu)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} G\left(x; \mu_{1:j-1}\right) + \sum_{\mu \in C_i^\star} \frac{G\left(C_i^\star; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot \sum_{x \in C_i^\star} \|x-\mu\|^2\right)$$

$$= \frac{2}{|C_i^\star|}\left(\sum_{\mu \in C_i^\star} \frac{G(C_i^\star, \mu)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot G\left(C_i^\star; \mu_{1:j-1}\right) + \sum_{\mu \in C_i^\star} \frac{G\left(C_i^\star; \mu_{1:j-1}\right)}{G\left(C_i^\star; \mu_{1:j-1}\right)} \cdot G(C_i^\star; \mu)\right)$$

$$= \frac{4}{|C_i^\star|} \sum_{\mu \in C_i^\star} G(C_i^\star; \mu) = 4\mathbb{E}_{\mu \sim C_i^\star}[G(C_i^\star; \mu)] = 8\mathrm{OPT}(C_i^\star)$$

■

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

Lemma. In the j-th iteration, let μ_j be the sampled center. Then, for any $i \in [k]$,

$$\mathbb{E}\left[G(C_i^\star; \mu_{1:j}) \middle| \mu_j \in C_i^\star\right] \leq 8\mathrm{OPT}(C_i^\star)$$

- Why can't it achieve 8-approximation?

k-Means++

k-means++ initialization:

1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

- For any $t \leq k$, define the set of clusters that have been "hit":

$$H_t := \{i \in [k] : C_i^\star \cap \mu_{1:t} \neq \emptyset\}, \qquad C_{H_t}^\star := \bigcup_{i \in H_t} C_i^\star$$

- Let $U_t := [k] \backslash H_t$ be the remaining "uncovered" clusters, and $W_t := t - |H_t|$ be the number of "wasted" iterations
- We claim that $\mathbb{E}\left[G\left(C_{H_t}^\star; \mu_t\right)\right] \leq 8\mathrm{OPT}$ for any $t \leq k$

k-Means++

We claim that $\mathbb{E}\left[G\left(C^\star_{H_t};\mu_t\right)\right] \le 8\mathrm{OPT}$ for any $t \le k$

- Note that

$$G\left(C^\star_{H_t};\mu_t\right) = \sum_{i\in H_t} G(C^\star_i;\mu_{1:t})$$

- Let $s_i \in [t]$ be the first time that $C^\star_i$ was "hit". Then $G(C^\star_i;\mu_{1:t}) \le G\left(C^\star_i;\mu_{1:s_i}\right) = G\left(C^\star_i;\mu_{s_i}\right)$

$$\mathbb{E}[G(C^\star_i;\mu_{1:t})] \le \sum_{s=1}^{k} \mathbb{E}\left[\mathbf{1}_{s_i=s} G(C^\star_i;\mu_s)\right] = \sum_{s=1}^{k} \mathbb{E}\left[\mathbb{E}_{\mu_s}\left[\mathbf{1}_{s_i=s} G(C^\star_i;\mu_s)\middle|\mu_{1:s-1}\right]\right]$$

- $\mathbf{1}_{s_i=s}$ is equivalent to $\mathbf{1}\{i \notin H_{s-1}\} \cdot \mathbf{1}\{\mu_s \in C^\star_i\}$; thus, we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{s_i=s} G(C^\star_i;\mu_s)\middle|\mu_{1:s_i-1}\right] &= \mathbf{1}\{i \notin H_{s-1}\} \cdot \mathbb{E}[G(C^\star_i;\mu_s)|\mu_{1:s-1}, \mu_s \in C^\star_i] \cdot \Pr[\mu_s \in C^\star_i|\mu_{1:s-1}] \\ &\le 8\mathrm{OPT}(C^\star_i) \cdot \mathbf{1}\{i \notin H_{s-1}\} \cdot \Pr[\mu_s \in C^\star_i|\mu_{1:s-1}] \\ &= 8\mathrm{OPT}(C^\star_i) \cdot \Pr[s_i = s|\mu_{1:s-1}]\end{aligned}$$

k-Means++

We claim that $\mathbb{E}\left[G\left(C^{\star}_{H_t};\mu_t\right)\right] \leq 8\mathrm{OPT}$ for any $t \leq k$

- Note that

$$G\left(C^{\star}_{H_t};\mu_t\right) = \sum_{i\in H_t} G(C_i^{\star};\mu_{1:t})$$

- Let $s_i \in [t]$ be the first time that $C_i^{\star}$ was "hit". Then $G(C_i^{\star};\mu_{1:t}) \leq G\left(C_i^{\star};\mu_{1:s_i}\right) = G\left(C_i^{\star};\mu_{s_i}\right)$

$$\mathbb{E}[G(C_i^{\star};\mu_{1:t})] \leq \sum_{s=1}^{k} \mathbb{E}_{\mu_{1:s-1}}[8\mathrm{OPT}(C_i^{\star}) \cdot \Pr[s_i = s|\mu_{1:s-1}]]$$

$$= 8\mathrm{OPT}(C_i^{\star}) \cdot \sum_{s=1}^{k} \Pr[s_i = s] \leq 8\mathrm{OPT}(C_i^{\star})$$

- Thus, $\mathbb{E}\left[G\left(C^{\star}_{H_t};\mu_t\right)\right] \leq \sum_{i\in H_t} 8\mathrm{OPT}(C_i^{\star}) \leq 8\mathrm{OPT}$

k-Means++

To analyze those uncovered clusters, we define a potential function

$$\Phi_t := W_t \frac{G\left(C_{U_t}; \mu_{1:t}\right)}{|U_t|}$$

Average cost of an uncovered cluster

- Initial potential $t = 0$:

$$W_0 = 0, \qquad \Phi_0 = 0$$

- Final potential $t = k$:

$$W_k = |U_k|, \qquad \Phi_k = G\left(C_{U_k}; \mu_{1:k}\right)$$

- Thus,

$$G(\mu_1, \dots, \mu_k) = G\left(C_{H_k}; \mu_{1:k}\right) + \sum_{t=0}^{k-1} \Phi_{t+1} - \Phi_t$$

$\leq 8\mathrm{OPT}$ in expectation

k-Means++

Lemma (Good iteration). Let $t \leq k-1$. Conditioned on $\mu_{1:t}$ and on the event that μ_{t+1} lands in an uncovered cluster, $\mathbb{E}[\Phi_{t+1}] \leq \Phi_t$

Proof.

- Suppose $\mu_{t+1} \in C_I^\star$ with $I \in U_t$. Then, $W_{t+1} = W_t$ and $U_{t+1} = U_t \setminus \{I\}$

$$\Phi_{t+1} = W_{t+1} \frac{G\left(C_{U_{t+1}}; \mu_{1:t+1}\right)}{|U_{t+1}|} \leq W_t \frac{G\left(C_{U_{t+1}}; \mu_{1:t}\right)}{|U_{t+1}|} = W_t \frac{G\left(C_{U_t}; \mu_{1:t}\right) - G(C_I^\star; \mu_{1:t})}{|U_t| - 1}$$

- Conditioned on landing in an uncovered cluster, we have for any $i \in U_t$,

$$\Pr[I = i | I \in U_t] = \frac{G(C_i^\star; \mu_{1:t})}{G\left(C_{U_t}; \mu_{1:t}\right)}$$

$$\mathbb{E}[G(C_I^\star; \mu_{1:t}) | \mu_{1:t}, I \in U_t] = \sum_{i \in U_t} \frac{G(C_i^\star; \mu_{1:t})}{G\left(C_{U_t}; \mu_{1:t}\right)} G(C_i^\star; \mu_{1:t}) \geq \frac{G\left(C_{U_t}; \mu_{1:t}\right)^2 / m}{G\left(C_{U_t}; \mu_{1:t}\right)} = \frac{G\left(C_{U_t}; \mu_{1:t}\right)}{m}$$

k-Means++

Lemma (Good iteration). Let $t \leq k-1$. Conditioned on $\mu_{1:t}$ and on the event that μ_{t+1} lands in an uncovered cluster, $\mathbb{E}[\Phi_{t+1}] \leq \Phi_t$

Proof.

- Thus, we get that

$$\begin{aligned}\mathbb{E}[\Phi_{t+1}|\mu_{1:t}, I \in U_t] &\leq W_t \frac{G\left(C_{U_t}; \mu_{1:t}\right) - \mathbb{E}[G(C_I^\star; \mu_{1:t})|\mu_{1:t}, I \in U_t]}{|U_t| - 1} \\ &\leq W_t G\left(C_{U_t}; \mu_{1:t}\right) \frac{1 - 1/|U_t|}{|U_t| - 1} \\ &= \Phi_t\end{aligned}$$

■

k-Means++

Lemma (Bad iteration). Let $t \leq k-1$. If μ_{t+1} lands in an already-hit cluster, then

$$\Phi_{t+1} - \Phi_t \leq \frac{G\left(C_{U_t}; \mu_{1:t}\right)}{|U_t|}$$

Proof.

- In this case, $W_{t+1} = W_t + 1$, $U_{t+1} = U_t$
- Thus,

$$\Phi_{t+1} = W_{t+1} \frac{G\left(C_{U_{t+1}}; \mu_{1:t+1}\right)}{|U_{t+1}|} = (W_t + 1) \frac{G\left(C_{U_t}; \mu_{1:t+1}\right)}{|U_t|} \leq (W_t + 1) \frac{G\left(C_{U_t}; \mu_{1:t}\right)}{|U_t|}$$

■

k-Means++

Lemma. For every $0 \le t \le k-1$,

$$\mathbb{E}[\Phi_{t+1} - \Phi_t | \mu_{1:t}] \le \frac{G(C_{H_t}; \mu_{1:t})}{k-t}$$

Proof.

- The probability that the $(t+1)$-th iteration is wasted is

$$p = \Pr[\mu_{t+1} \in C_{H_t} | \mu_{1:t}] = \frac{G(C_{H_t}; \mu_{1:t})}{G(\mathcal{X}; \mu_{1:t})}$$

- Combining the good iteration and bad iteration lemmas, we have

$$\mathbb{E}[\Phi_{t+1} - \Phi_t | \mu_{1:t}] \le 0 \cdot (1-p) + \frac{G(C_{U_t}; \mu_{1:t})}{|U_t|} \cdot p = \frac{G(C_{U_t}; \mu_{1:t}) G(C_{H_t}; \mu_{1:t})}{|U_t| G(\mathcal{X}; \mu_{1:t})}$$

$$\le \frac{G(C_{H_t}; \mu_{1:t})}{|U_t|} \le \frac{G(C_{H_t}; \mu_{1:t})}{k-t}$$

■

k-Means++

Theorem (Arthur-Vassilvitskii '07**).** Let $\mu_1, \dots, \mu_k \in \mathcal{X}$ be the output of k-means++ and $\mu_1^\star, \dots, \mu_k^\star$ be the optimal centers. Then,

$$\mathbb{E}[G(\mu_1, \dots, \mu_k)] \leq O(\log k) \cdot \mathrm{OPT}$$

Proof.

$$\mathbb{E}[G(\mu_1, \dots, \mu_k)] = \mathbb{E}\left[G\left(C_{H_k}; \mu_{1:k}\right)\right] + \sum_{t=0}^{k-1} \mathbb{E}[\Phi_{t+1} - \Phi_t] \leq 8\mathrm{OPT} + \sum_{t=0}^{k-1} \mathbb{E}[\Phi_{t+1} - \Phi_t]$$

- By the total expectation, we have

$$\mathbb{E}[\Phi_{t+1} - \Phi_t] = \mathbb{E}\left[\mathbb{E}[\Phi_{t+1} - \Phi_t | \mu_{1:t}]\right] \leq \frac{\mathbb{E}\left[G\left(C_{H_t}; \mu_{1:t}\right)\right]}{k-t} \leq \frac{8\mathrm{OPT}}{k-t}$$

- Thus,

$$\sum_{t=0}^{k-1} \mathbb{E}[\Phi_{t+1} - \Phi_t] \leq 8\mathrm{OPT} \cdot \sum_{t=1}^{k} \frac{1}{t} = O(\log k) \cdot \mathrm{OPT}$$

■

k-Means++

k-means++ initialization:

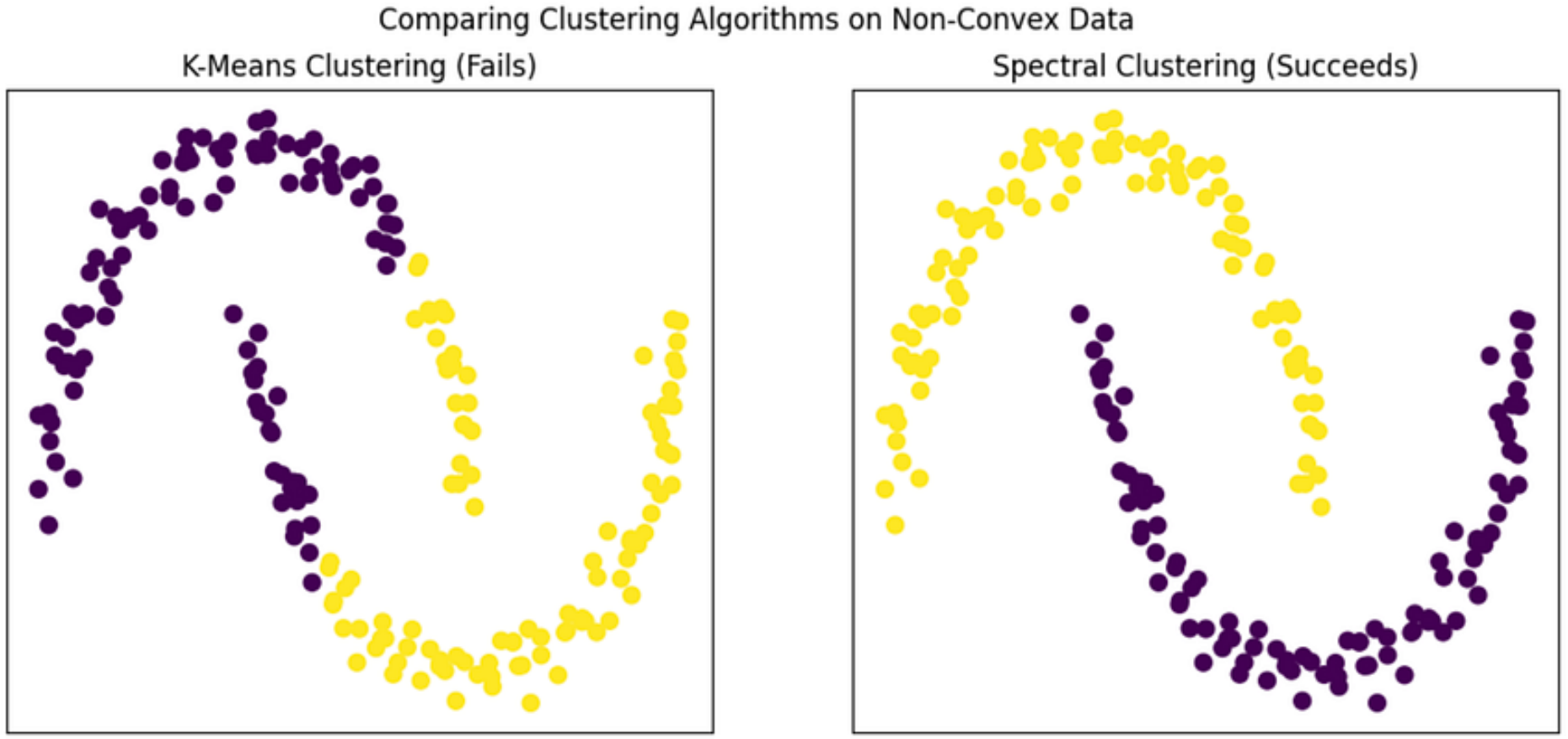
1. Sample $\mu_1 \in \mathcal{X}$ uniformly at random
2. For each $j \in \{2, \dots, k\}$:
   i. Sample $\mu_j \in \mathcal{X}$ with probability $\propto G(\mu_j; \mu_{1:j-1}) = \min_{i<j} \|\mu_j - \mu_i\|^2$

Theorem (Arthur-Vassilvitskii '07**).** Let $\mu_1, \dots, \mu_k \in \mathcal{X}$ be the output of k-means++ and $\mu_1^\star, \dots, \mu_k^\star$ be the optimal centers. Then,

$$\mathbb{E}[G(\mu_1, \dots, \mu_k)] \leq O(\log k) \cdot G(\mu_1^\star, \dots, \mu_k^\star)$$

Arthur-Vassilvitskii also proved that $O(\log k)$-approximation of k-means++ is tight

Spectral Clustering



Based on “closeness” to a center point

Based on “connectivity” or “similarity”

https://dilipkumar.medium.com/spectral-clustering-algorithm-03a62854d19b

Spectral Clustering