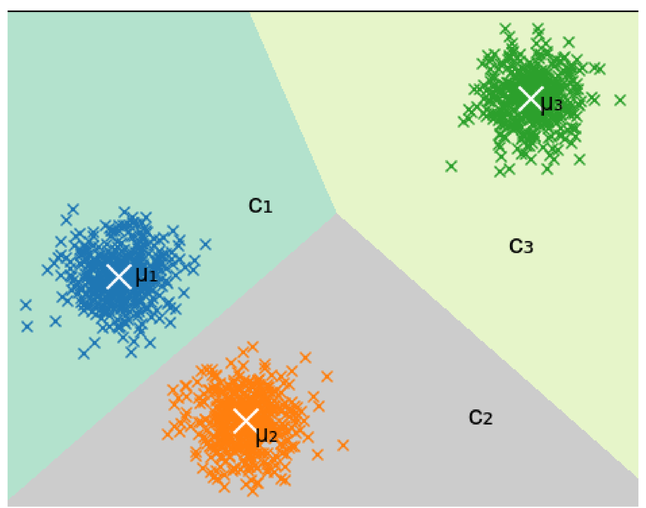
EUCLIDEAN k-MEANS CLUSTERING WITH α-CENTER PROXIMITY

Amit Deshpande (MSR, India), Anand Louis (IISc, India), and Apoorv Vikram Singh (IISc, India)

PROBLEM

k **means:** $\min_{\mu_1,\ldots,\mu_k} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$.



α-Center Proximal Clustering: $\forall i \neq j, x \in C_i$,

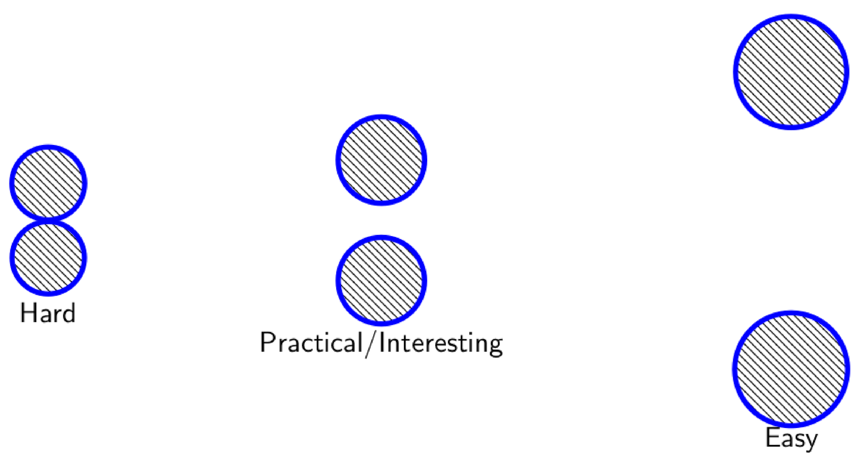
$$\|x - \mu_j\| > \alpha \|x - \mu_i\|$$

Aim: Given $\alpha > 1$, k, n points in $\mathbb{R}^d$, find the minimum cost α-center proximal k-means clustering.

Motivation: Larger the value of α, the more separated are the clusters. A way to get the "*ground-truth*" clustering.

MOTIVATION

- Small k-means cost alone does not imply stable or meaningful clusters in practice.
- Real-world data have α close to 1.
- *Realistic model:* most of the points satisfy α-center proximal clustering, except a small fraction.



Intuition of α-center proximal instances

OUR RESULTS

Best known: For $\alpha \geq 2$, [AMM'17] can find the α-center proximal clustering in polynomial time.

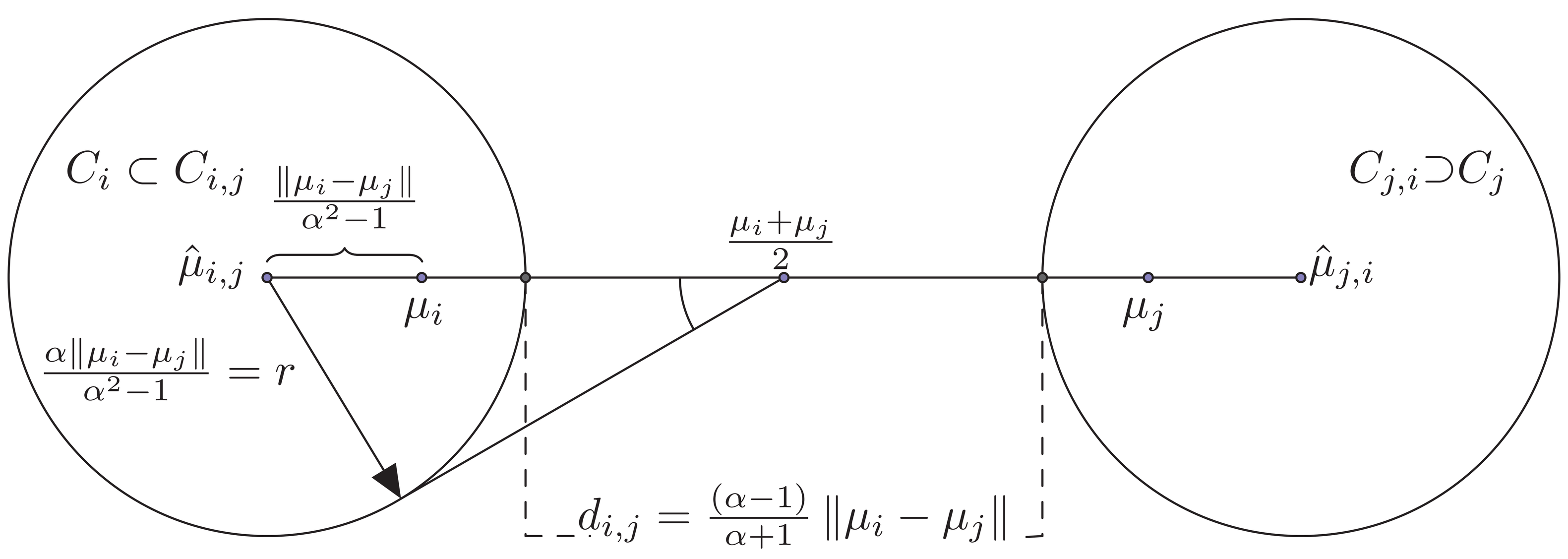
Our Results:

1. **Algorithmic:** If the α-center proximal clustering of the minimum k-means cost has ω-balanced clusters, then our algorithm outputs the minimum cost α-center proximal clustering with a constant probability in time $O\left(nd2^{\mathrm{poly}(k/\omega(\alpha-1))}\right)$, where ω is a balance parameter. This holds for *any* value of $\alpha > 1$. We can also handle some class of outliers.

2. **Hardness:** There exists a value of α and $\varepsilon_0 > 0$ such that it is NP-hard to approximate minimum k-means cost ω-balanced α-center proximal clustering within a factor of $(1 + \varepsilon_0)$. The value of k depends on n in the construction.

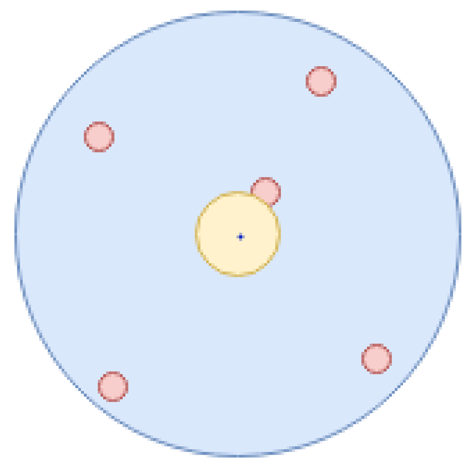
GEOMETRY

This holds for all pairs of α-center proximal clusters. The geometric property was also noted by [TV'10].



MAIN IDEA

Approximate Caratheodory Theorem: Sample points uniformly at random from a cluster of bounded radius. Mean of the sample is close to the cluster mean.

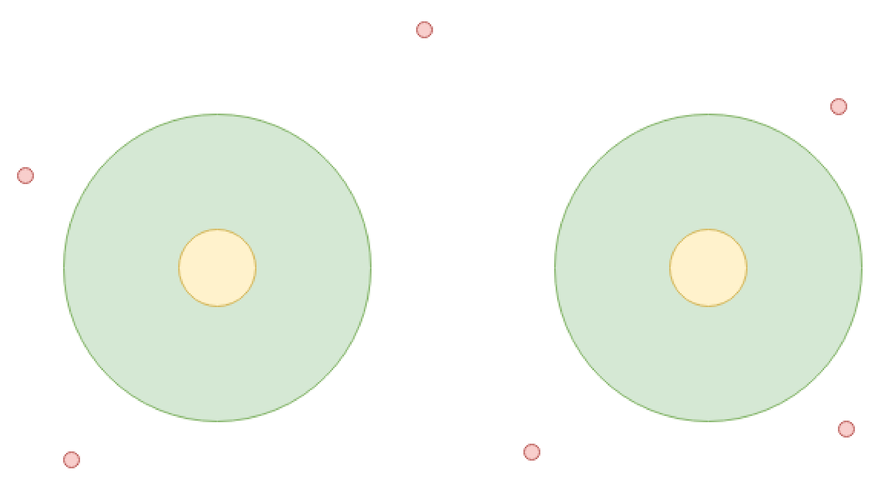


Idea: We can approximate mean of a cluster up to an additive error of $\frac{d_{i,j}}{2}$, and still recover the exact clustering!

OUTLIERS

Model: Let Z denote the set of outliers, and $|Z|$ is known to us. We assume that the set of outliers satisfy the following property. For all $i, j \in [k]$, $x \in C_i$, and $z \in Z$

$$\|z - \mu_j\| > \alpha \|x - \mu_i\|.$$



Idea: Since outliers are "*far-away*", and $|Z|$ is known, we can sample sufficient number of points based on ω and $|Z|$, and get close to the means of each cluster. We use means to get the clusters and then remove the *farthest* $|Z|$ points.

ALGORITHM AND ANALYSIS

Algorithm:

1: Sample poly $\left(\frac{1}{\omega}, k, \frac{\alpha}{(\alpha-1)^2}\right)$ points uniformly at random.
▷ this ensures sufficient points from all clusters
2: Go over all the k partitions of these points ($k+1$ in case of outliers).
▷ at least one of the partitions corresponds to desired clustering
3: Assign all points to nearest center, and output clustering which is α-center proximal with lowest k-means cost.
▷ one of the clusterings will be the desired clustering due to the approximate Caratheodory theorem

Runtime: Step 3 takes time $O(nd)$ for each clustering. Step 2 produces $2^{\mathrm{poly}\left(\frac{1}{\omega}, k, \frac{\alpha}{(\alpha-1)^2}\right)}$ number of clusterings. Therefore, the total running time is $O\left(nd2^{\mathrm{poly}(k/\omega(\alpha-1))}\right)$.

HARDNESS

1. [ACKS'15] showed hardness of approximation of Euclidean k-means clustering using reduction from the vertex cover problem on triangle-free graphs.
2. We show that the instance given by [ACKS'15] is α-center proximal.

REFERENCES

[ACKS'15] Pranjal Awasthi and Moses Charikar and Ravishankar Krishnaswamy and Ali Kemal Sinop. The Hardness of Approximation of Euclidean k-Means. *SoCG*, 2015.

[AMM'17] Haris Angelidakis and Konstantin Makarychev and Yury Makarychev. Algorithms for Stable and Perturbation-Resilient Problems. *STOC*, 2017.

[TV'10] Matus Telgarsky and Andrea Vattani. Hartigan's Method: k-means Clustering without Voronoi. *AISTATS*, 2010.

AL was supported in part by SERB Award ECR/2017/003296. AL is grateful to Microsoft Research for supporting this collaboration.