Faster Algorithms for Constrained k-Means Problem

Anup Bhattacharya (IIT Delhi)

Joint work with
Amit Kumar (IIT Delhi) & Ragesh Jaiswal (IIT Delhi)

k-means Clustering

- Given n points in $\mathbb{R}^d$ and integer k, find k centers in $\mathbb{R}^d$.
- Minimize objective function $\sum_{x \in X} \min_{c \in C} ||x - c||^2$.

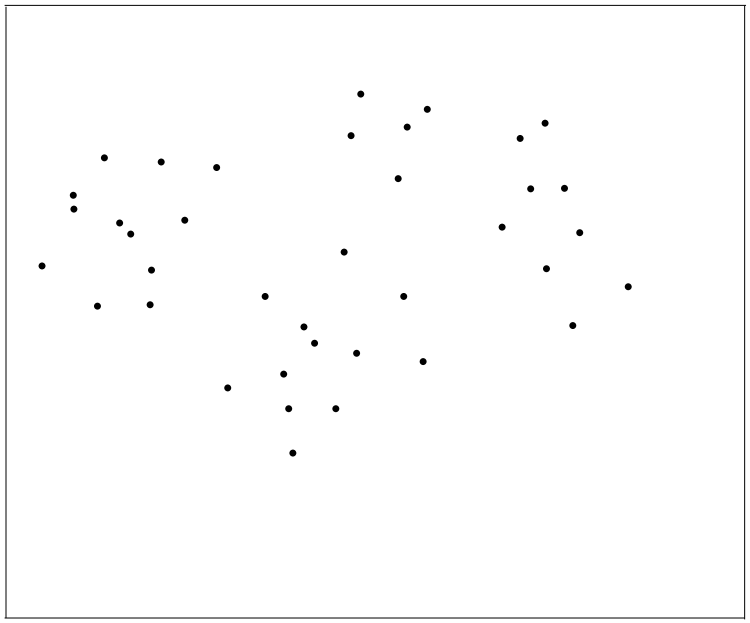

Figure : Input points in $\mathbb{R}^2$, $k = 4$

k-means Clustering

- Given n points in $\mathbb{R}^d$ and integer k, find k centers in $\mathbb{R}^d$.
- Minimize objective function $\sum_{x \in X} \min_{c \in C} ||x - c||^2$.

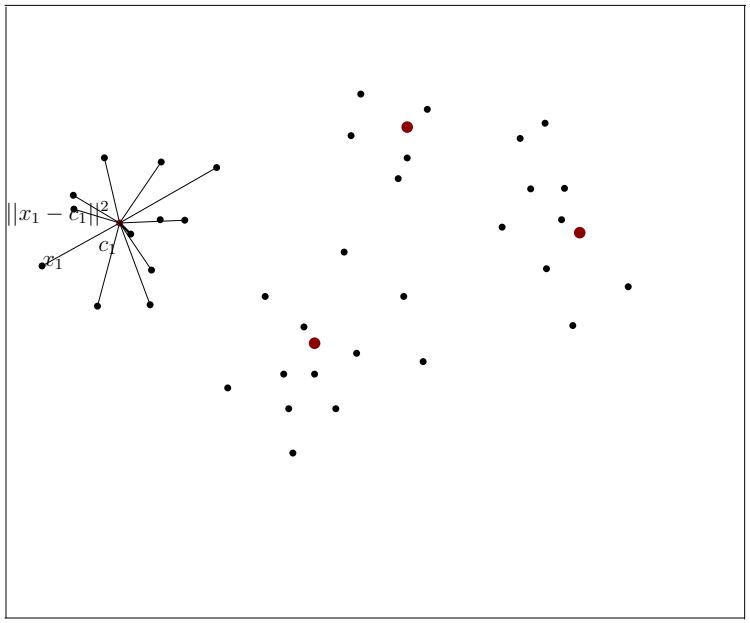

Figure : Input points in $\mathbb{R}^2$, $k = 4$

k-means Clustering

- Given n points in $\mathbb{R}^d$ and integer k, find k centers in $\mathbb{R}^d$.
- Minimize objective function $\sum_{x \in X} \min_{c \in C} ||x - c||^2$.
- Given k centers, clusters are formed using Voronoi partitioning.

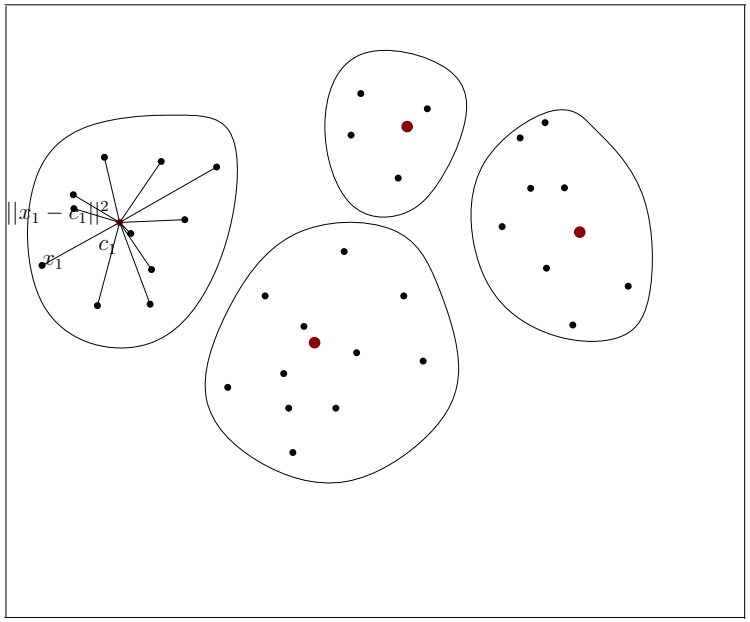

Figure : Input points in $\mathbb{R}^2$, $k = 4$

k-means Clustering: Known Bounds

- NP-hard for $k \geq 2$ [Dasgupta2008], $d \geq 2$ [Mahajan et al 2012].
- Approximation ratio: Algorithm ALG gives α-approximation if for all instances I, $ALG(I) \leq \alpha OPT(I)$.
- Upper Bound: Many $(1+\epsilon)$-approximation algorithms known. $(1+\epsilon)$-approximation in time $O(nd2^{\tilde{O}(\frac{k}{\epsilon})})$ [Jaiswal et al. 2015]
- Lower Bound: $\exists \epsilon > 0$ NP-hard to approximate by factor $(1+\epsilon)$ [Awasthi et al 2015].
- Results for unconstrained k-means problem.

k-means Clustering: Known Bounds

- NP-hard for $k \geq 2$ [Dasgupta2008], $d \geq 2$ [Mahajan et al 2012].
- Approximation ratio: Algorithm ALG gives α-approximation if for all instances I, $ALG(I) \leq \alpha OPT(I)$.
- Upper Bound: Many $(1+\epsilon)$-approximation algorithms known. $(1+\epsilon)$-approximation in time $O(nd2^{\tilde{O}(\frac{k}{\epsilon})})$ [Jaiswal et al. 2015]
- Lower Bound: $\exists \epsilon > 0$ NP-hard to approximate by factor $(1+\epsilon)$ [Awasthi et al 2015].
- Results for unconstrained k-means problem.

k-means Clustering: Known Bounds

- NP-hard for $k \geq 2$ [Dasgupta2008], $d \geq 2$ [Mahajan et al 2012].
- Approximation ratio: Algorithm ALG gives α-approximation if for all instances I, $ALG(I) \leq \alpha OPT(I)$.
- Upper Bound: Many $(1+\epsilon)$-approximation algorithms known.
  $(1+\epsilon)$-approximation in time $O(nd2^{\tilde{O}(\frac{k}{\epsilon})})$ [Jaiswal et al. 2015]
- Lower Bound: $\exists \epsilon > 0$ NP-hard to approximate by factor $(1+\epsilon)$ [Awasthi et al 2015].
- Results for unconstrained k-means problem.

k-means Clustering: Known Bounds

- NP-hard for $k \geq 2$ [Dasgupta2008], $d \geq 2$ [Mahajan et al 2012].
- Approximation ratio: Algorithm ALG gives α-approximation if for all instances I, $ALG(I) \leq \alpha OPT(I)$.
- Upper Bound: Many $(1+\epsilon)$-approximation algorithms known. $(1+\epsilon)$-approximation in time $O(nd2^{\tilde{O}(\frac{k}{\epsilon})})$ [Jaiswal et al. 2015]
- Lower Bound: $\exists \epsilon > 0$ NP-hard to approximate by factor $(1+\epsilon)$ [Awasthi et al 2015].
- Results for unconstrained k-means problem.

k-means Clustering: Known Bounds

- NP-hard for $k \geq 2$ [Dasgupta2008], $d \geq 2$ [Mahajan et al 2012].
- Approximation ratio: Algorithm ALG gives α-approximation if for all instances I, $ALG(I) \leq \alpha OPT(I)$.
- Upper Bound: Many $(1+\epsilon)$-approximation algorithms known. $(1+\epsilon)$-approximation in time $O(nd2^{\tilde{O}(\frac{k}{\epsilon})})$ [Jaiswal et al. 2015]
- Lower Bound: $\exists \epsilon > 0$ NP-hard to approximate by factor $(1+\epsilon)$ [Awasthi et al 2015].
- Results for unconstrained k-means problem.

Constrained Clustering: Examples

- Given n points in $\mathbb{R}^d$, and integer k.
- Minimize objective while obeying additional constraints.
- Examples of constraints:
  - r-gather clustering: Each cluster has size at least r.
  - Capacitated clustering: Cluster sizes have upper bounds.
  - Chromatic clustering: No two points in cluster with same color.

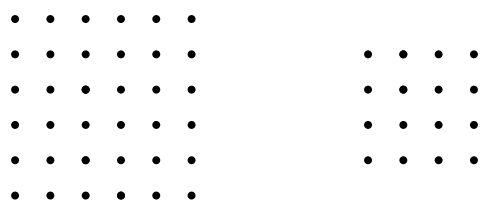

Figure : r-gather clustering: Input points in $\mathbb{R}^2$, $k = 2$, $r = 20$

Constrained Clustering: Examples

- r-gather clustering: Each cluster has size at least r.
- Unconstrained k-means clustering on the input instance.

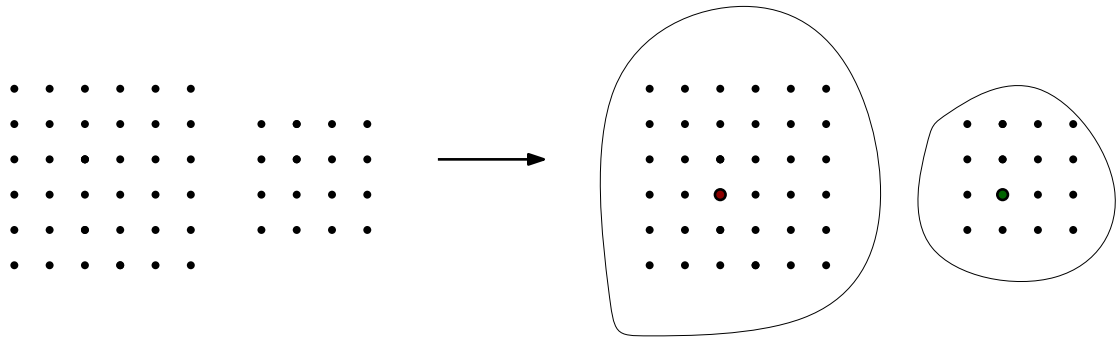

Figure : Solution for Unconstrained clustering

Constrained Clustering: Examples

- r-gather clustering: Each cluster has size at least r.

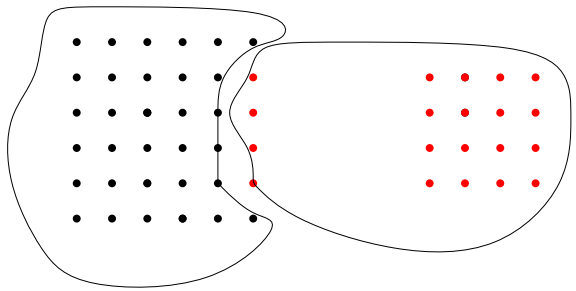

Figure : r-gather clustering: Input points in $\mathbb{R}^2$, $k = 2$, $r = 20$

Constrained k-means Problem

- Unconstrained k-means: Find clustering which optimizes objective.
- Constraints restrict the set of possible clusterings of an input dataset.
- Constrained k-means: Find clustering which optimizes objective while satisfying constraint.

Constrained k-means Problem

- Unconstrained k-means: Find clustering which optimizes objective.
- Constraints restrict the set of possible clusterings of an input dataset.
- Constrained k-means: Find clustering which optimizes objective while satisfying constraint.

Constrained k-means Problem

- Unconstrained k-means: Find clustering which optimizes objective.
- Constraints restrict the set of possible clusterings of an input dataset.
- Constrained k-means: Find clustering which optimizes objective while satisfying constraint.

Constrained k-means Problem

- Constrained k-means [Ding & Xu 2015]: Given n points in $\mathbb{R}^d$, integer k, and set of constraints, find k clusters which minimize objective function.
- $(1+\epsilon)$-approximation for constrained k-means [Ding & Xu 2015].

Constrained k-means Problem

- Locality property: Points in the same cluster are closer to each other compared to points in different clusters.
- True for unconstrained clustering.
- Locality not valid for constrained clustering.

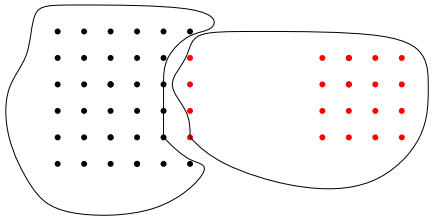

Figure : r-gather Clustering: Input points in $\mathbb{R}^2$, $k = 2$, $r = 20$

Constrained k-means Problem

- Locality property: Points in the same cluster are closer to each other compared to points in different clusters.
- True for unconstrained clustering.
- Locality not valid for constrained clustering.

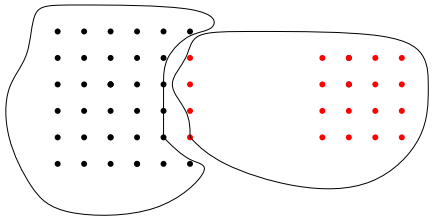

Figure : r-gather Clustering: Input points in $\mathbb{R}^2$, $k = 2$, $r = 20$

Constrained k-means Problem

- Locality property: Points in the same cluster are closer to each other compared to points in different clusters.
- True for unconstrained clustering.
- Locality not valid for constrained clustering.

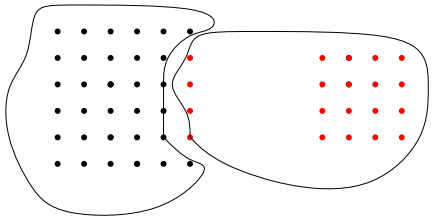

Figure : r-gather Clustering: Input points in $\mathbb{R}^2$, $k = 2$, $r = 20$

Constrained k-means: Find Clusters from Centers

- Given k centers, find clustering.
- Unconstrained clustering: Voronoi partitioning works.
- Constrained clustering: Partition algorithm [Ding & Xu 2015].
- Designed polynomial time partition algorithms for various constrained k-means problems.

Constrained k-means: Find Clusters from Centers

- Given k centers, find clustering.
- Unconstrained clustering: Voronoi partitioning works.
- Constrained clustering: Partition algorithm [Ding & Xu 2015].
- Designed polynomial time partition algorithms for various constrained k-means problems.

Constrained k-means: Find Clusters from Centers

- Given k centers, find clustering.
- Unconstrained clustering: Voronoi partitioning works.
- Constrained clustering: Partition algorithm [Ding & Xu 2015].
- Designed polynomial time partition algorithms for various constrained k-means problems.

Constrained k-means: Find Clusters from Centers

- Given k centers, find clustering.
- Unconstrained clustering: Voronoi partitioning works.
- Constrained clustering: Partition algorithm [Ding & Xu 2015].
- Designed polynomial time partition algorithms for various constrained k-means problems.

Partition Algorithm

- Given k centers, find clustering which minimizes objective while satisfying constraints.
- Partition algorithm for r-gather clustering [Ding & Xu 2015]
- Reduces to min-cost circulation problem.

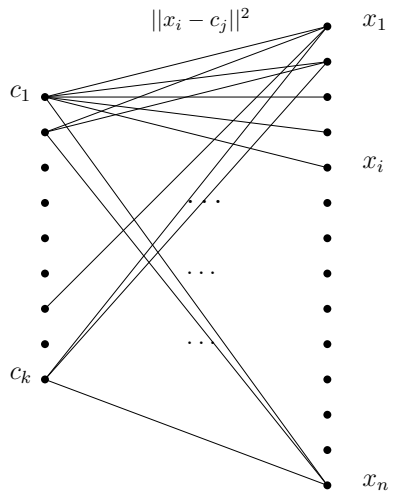

Figure : Partition algorithm for r-gather Clustering

Constrained k-means: Known Results

- Ding & Xu give $(1+\epsilon)$-approximation running in time $O(ndL + P(X) \cdot L)$.
- Let L be number of candidate centers, $L \leq O((\log n)^k 2^{\text{poly}(\frac{k}{\epsilon})})$.
- Let $P(X)$ be time required by Partition to find clustering cost for each k-center in L.

Constrained k-means Problem

- Given constraints, let $X = X_1 \uplus \ldots \uplus X_k$ be OPT clusters.
- Given a dataset, different constraints would result in different OPT clusterings.
- Can a single set of k centers may give good approximation for all possible OPT clusterings?
- Is it possible to return a list of such candidate k-centers one of which is good?

Constrained k-means Problem

- Given constraints, let $X = X_1 \uplus \ldots \uplus X_k$ be OPT clusters.
- Given a dataset, different constraints would result in different OPT clusterings.
- Can a single set of k centers may give good approximation for all possible OPT clusterings?
- Is it possible to return a list of such candidate k-centers one of which is good?

Constrained k-means Problem

- Given constraints, let $X = X_1 \uplus \ldots \uplus X_k$ be OPT clusters.
- Given a dataset, different constraints would result in different OPT clusterings.
- Can a single set of k centers may give good approximation for all possible OPT clusterings?
- Is it possible to return a list of such candidate k-centers one of which is good?

Constrained k-means Problem

- Given constraints, let $X = X_1 \uplus \ldots \uplus X_k$ be OPT clusters.
- Given a dataset, different constraints would result in different OPT clusterings.
- Can a single set of k centers may give good approximation for all possible OPT clusterings?
- Is it possible to return a list of such candidate k-centers one of which is good?

Constrained k-means Problem

- Given constraints, let $X = X_1 \uplus \ldots \uplus X_k$ be OPT clusters.
- Given a dataset, different constraints would result in different OPT clusterings.
- Can a single set of k centers may give good approximation for all possible OPT clusterings?
- Is it possible to return a list of such candidate k-centers one of which is good?

List k-means Problem

- Given $X \subseteq \mathbb{R}^d$, integer k, $\epsilon > 0$ and some implicit OPT partition $X_1, \ldots, X_k$.
- List k-means finds a set $C = \{C_1, \ldots, C_L\}$ of k-centers such that $\exists j \in [1, L]$, C_j gives $(1 + \epsilon)$-approximation wrt $X_1, \ldots, X_k$.

List k-means to Constrained k-means

- List k-means outputs a list of candidate k-centers.
- For each k-center, compute clustering using partition algorithm.
- The clustering with minimum cost would be the solution for constrained k-means.

Our Results for Constrained k-means

- List size in [Ding & Xu] is $L \leq O((\log n)^k 2^{\text{poly}(\frac{k}{\epsilon})})$
- Our algorithm has list size $L \leq 2^{\tilde{O}(\frac{k}{\epsilon})}$
- Notice that list size is independent of n.
- Almost matching lower bound: $L \geq 2^{\tilde{\Omega}(\frac{k}{\sqrt{\epsilon}})}$
- Running time: $O(ndL + P(X) \cdot L)$
- Can be extended for List k-median problem.

Our Results for Constrained k-means

- List size in [Ding & Xu] is $L \leq O((\log n)^k 2^{\text{poly}(\frac{k}{\epsilon})})$
- Our algorithm has list size $L \leq 2^{\tilde{O}(\frac{k}{\epsilon})}$
- Notice that list size is independent of n.
- Almost matching lower bound: $L \geq 2^{\tilde{\Omega}(\frac{k}{\sqrt{\epsilon}})}$
- Running time: $O(ndL + P(X) \cdot L)$
- Can be extended for List k-median problem.

Our Results for Constrained k-means

- List size in [Ding & Xu] is $L \leq O((\log n)^k 2^{\text{poly}(\frac{k}{\epsilon})})$
- Our algorithm has list size $L \leq 2^{\tilde{O}(\frac{k}{\epsilon})}$
- Notice that list size is independent of n.
- Almost matching lower bound: $L \geq 2^{\tilde{\Omega}(\frac{k}{\sqrt{\epsilon}})}$
- Running time: $O(ndL + P(X) \cdot L)$
- Can be extended for List k-median problem.

Our Results for Constrained k-means

- List size in [Ding & Xu] is $L \leq O((\log n)^k 2^{\text{poly}(\frac{k}{\epsilon})})$
- Our algorithm has list size $L \leq 2^{\tilde{O}(\frac{k}{\epsilon})}$
- Notice that list size is independent of n.
- Almost matching lower bound: $L \geq 2^{\tilde{\Omega}(\frac{k}{\sqrt{\epsilon}})}$
- Running time: $O(ndL + P(X) \cdot L)$
- Can be extended for List k-median problem.

Our Results for Constrained k-means

- List size in [Ding & Xu] is $L \leq O((\log n)^k 2^{\text{poly}(\frac{k}{\epsilon})})$
- Our algorithm has list size $L \leq 2^{\tilde{O}(\frac{k}{\epsilon})}$
- Notice that list size is independent of n.
- Almost matching lower bound: $L \geq 2^{\tilde{\Omega}(\frac{k}{\sqrt{\epsilon}})}$
- Running time: $O(ndL + P(X) \cdot L)$
- Can be extended for List k-median problem.

Upper Bound Sketch: Approximate Largest Optimum Cluster

- Let S be set of $O(\frac{1}{\epsilon})$ randomly sampled points. Then $\delta > 0$, wp at least $(1 - \delta)$ [Inaba et al]

$$\sum_{x \in X} ||x - \psi(S)||^2 \leq (1 + O(\epsilon)) \sum_{x \in X} ||x - \psi(X)||^2$$

- Fact: For any $X \subset \mathbb{R}^d$ and any point $p \in \mathbb{R}^d$

$$\sum_{x \in X} ||x - p||^2 = \sum_{x \in X} ||x - \psi(X)||^2 + |X| \; ||p - \psi(X)||^2$$

- Sample $O(\frac{k}{\epsilon})$ points and consider all subsets of size $O(\frac{1}{\epsilon})$.

Upper Bound Sketch: Approximate Largest Optimum Cluster

- Sample $O(\frac{k}{\epsilon})$ points randomly.
- Consider all subsets of size $O(\frac{1}{\epsilon})$ and compute their mean.

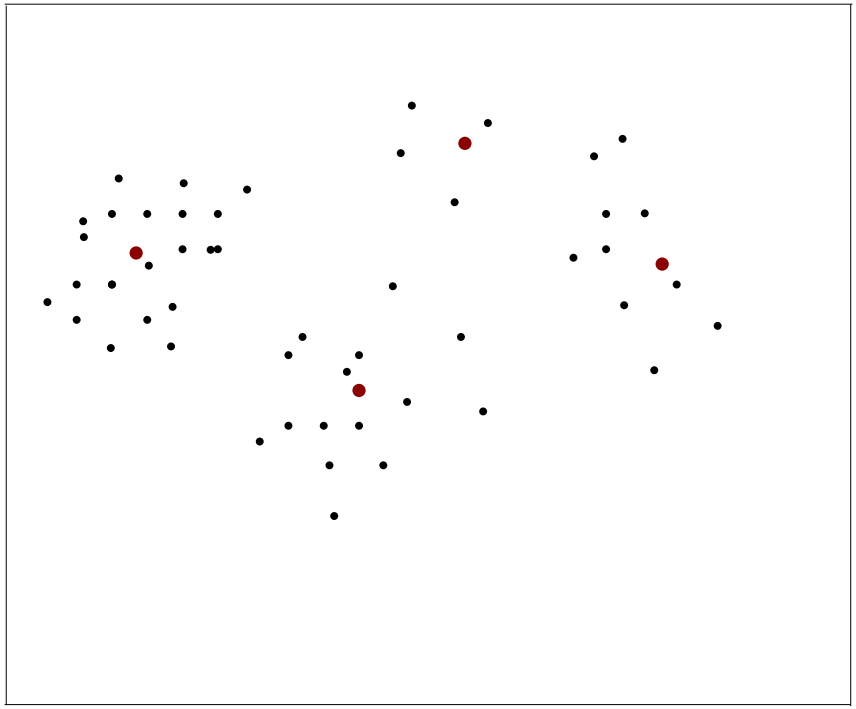

Figure : Approximate Largest OPT Cluster: $k = 4$

Upper Bound Sketch: Approximate Largest Optimum Cluster

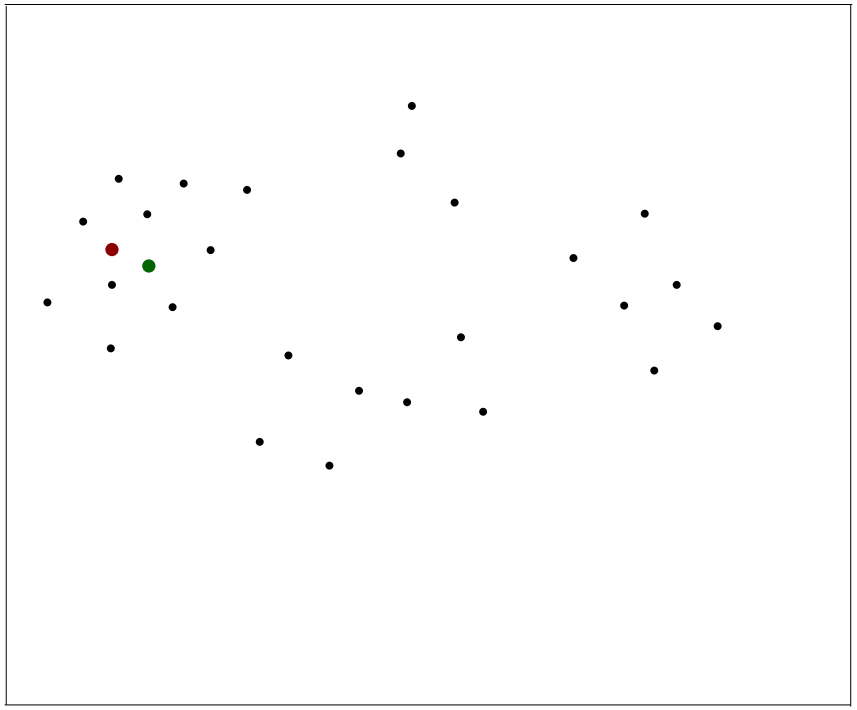

Figure : Approximate Largest OPT Cluster: $k = 4$

- What about other smaller OPT clusters?
- Random sampling does not work: OPT clusters may be very small.

Approximate Other Clusters

- Use D^2-sampling:
  - Let C be set of already chosen centers.
  - D^2-sampling chooses point p wp proportional to $\min_{c \in C} ||p - c||^2$.
- Sample $O(\frac{k}{\epsilon^3})$ points using D^2-sampling and try subsets of size $O(\frac{1}{\epsilon})$.
- Centroid of at least one subset approximates the OPT center.
- This process gives $(1 + \epsilon)$-approximation for unconstrained k-means.

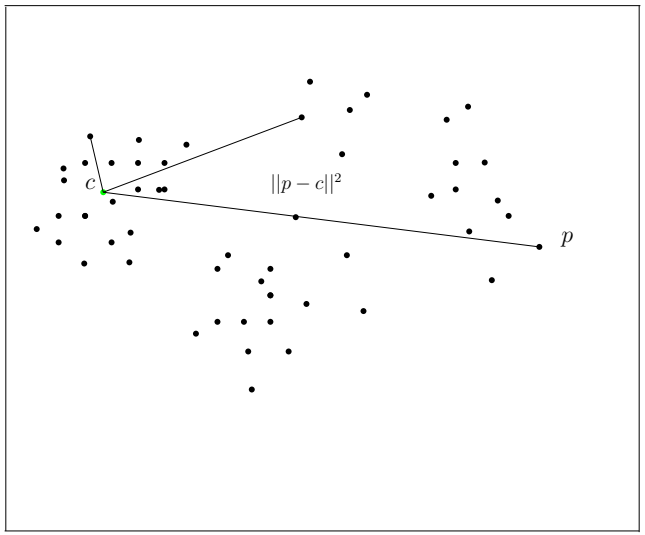

Figure : D^2-sampling points

Constrained Clustering

- For the largest OPT cluster things are fine.
- D^2-sampling based scheme does not work for constrained clustering.

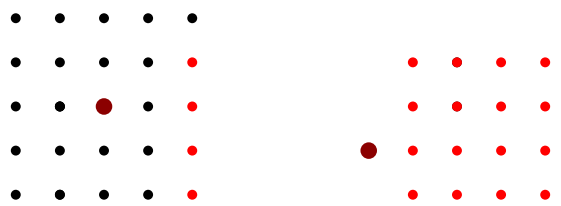

Figure : D^2-sampling points, $k = 2$

Constrained Clustering

- Centroid of none of the subsets may be good.

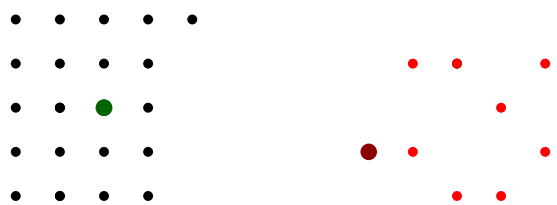

Figure : D^2-sampling points, $k = 2$

Idea: Constrained Clustering

- Cluster misses representation if portions of it close to covered clusters.
- Idea: Add $O(\frac{1}{\epsilon})$ copies of centers in C to the set of sampled points.
- Trying all subsets of this new set works.
- We obtain $(1+\epsilon)$-approximation for List k-means with $L = 2^{\tilde{O}(\frac{k}{\epsilon})}$.

Conclusions

- $(1+\epsilon)$-approximation for List k-means problem.
- Almost tight lower bound for List k-means problem.
- $(1+\epsilon)$-approximation for List k-median in time $O(nd2^{\tilde{O}(\frac{k}{\epsilon^{O(1)}})})$.
- We don't have matching lower bound for List k-median.
- Analysis extends to distance measures having 'metric like' properties, includes Mahalanobis distance, μ-similar Bregman divergence.

Thanks

Questions