Faster Algorithms for the Constrained k-means Problem

Ragesh Jaiswal

CSE, IIT Delhi

June 16, 2015

[Joint work with Anup Bhattacharya (IITD) and Amit Kumar (IITD)]

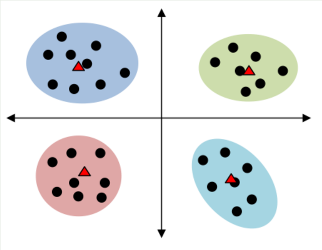
k-means Clustering Problem

Problem (*k*-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k points $C \subset \mathbb{R}^d$ (called centers*) such that the sum of squared Euclidean distance of each point in X to its closest center in C is minimized. That is, the following cost function is minimized:*

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2 \right)$$

Example: $k = 4, d = 2$

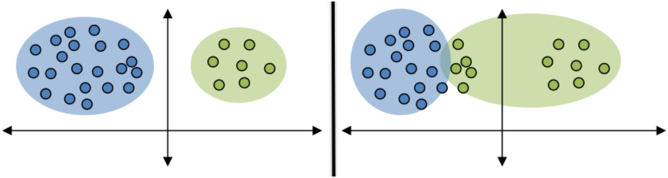


- Lower bounds:
  - The problem is NP-hard when $k \geq 2, d \geq 2$ [Das08, MNV12, Vat09].
  - Theorem [ACKS15]: There is a constant $\epsilon > 0$ such that it is NP-hard to approximate the k-means problem to a factor better than $(1 + \epsilon)$.

- Lower bounds:
  - The problem is NP-hard when $k \geq 2, d \geq 2$ [Das08, MNV12, Vat09].
  - Theorem [ACKS15]: There is a constant $\epsilon > 0$ such that it is NP-hard to approximate the k-means problem to a factor better than $(1 + \epsilon)$.
- Upper bounds: There are various approximation algorithms for the k-means problem.

Citation	**Approx. factor**	**Running Time**
[AV07]	$O(\log k)$	polynomial time
[KMN+02]	$9 + \epsilon$	polynomial time
[KSS10, JKY15, FMS07]	$(1 + \epsilon)$	$O\left(nd \cdot 2^{\tilde{O}(k/\epsilon)}\right)$

- Clustering using the k-means formulation implicitly assumes that the target clustering follows locality property that data points within the same cluster are close to each other in some geometric sense.
- There are clustering problems arising in Machine Learning where locality is not the *only* requirement while clustering.



- Clustering using the k-means formulation implicitly assumes that the target clustering follows locality property that data points within the same cluster are close to each other in some geometric sense.
- There are clustering problems arising in Machine Learning where locality is not the *only* requirement while clustering.
    - *r-gather clustering*: Each cluster should contain at least r points.
    - *Capacitated clustering*: Cluster size is upper bounded.
    - *l-diversity clustering*: Each input point has an associated color and each cluster should not have more that $\frac{1}{l}$ fraction of its points sharing the same color.
    - *Chromatic clustering*: Each input point has an associated color and points with same color should be in different clusters.

- Clustering using the k-means formulation implicitly assumes that the target clustering follows locality property that data points within the same cluster are close to each other in some geometric sense.
- There are clustering problems arising in Machine Learning where locality is not the *only* requirement while clustering.
    - *r-gather clustering*: Each cluster should contain at least r points.
    - *Capacitated clustering*: Cluster size is upper bounded.
    - *l-diversity clustering*: Each input point has an associated color and each cluster should not have more that $\frac{1}{l}$ fraction of its points sharing the same color.
    - *Chromatic clustering*: Each input point has an associated color and points with same color should be in different clusters.
- A unified framework that considers all the above problems would be nice.

- There are clustering problems arising in Machine Learning where locality is not the *only* requirement while clustering.
  - *r-gather clustering*: Each cluster should contain at least r points.
  - *Capacitated clustering*: Cluster size is upper bounded.
  - *l-diversity clustering*: Each input point has an associated color and each cluster should not have more that $\frac{1}{l}$ fraction of its points sharing the same color.
  - *Chromatic clustering*: Each input point has an associated color and points with same color should be in different clusters.
- A unified framework that considers all the above problems would be nice.

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k clusters $X_1, ..., X_k$ such that (i) the clusters satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where}\ \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k centers $C \subset \mathbb{R}^d$ such that the the following cost function is minimized:

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2 \right)$$

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k clusters $X_1, ..., X_k$ such that (i) the clusters satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where } \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k centers $C \subset \mathbb{R}^d$ such that the the following cost function is minimized:

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2\right)$$

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k clusters $X_1, ..., X_k$ such that (i) the clusters satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \;\; where \; \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k centers $C \subset \mathbb{R}^d$ such that the the following cost function is minimized:

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2 \right)$$

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k clusters $X_1, ..., X_k$ such that (i) the clusters satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where } \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k clusters $X_1, ..., X_k$ such that the the following cost function is minimized:

$$\Phi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where } \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k clusters $X_1, ..., X_k$ such that (i) the clusters satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where } \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Fact

For any $X \subset \mathbb{R}^d$ and any point $p \in \mathbb{R}^d$,
$\sum_{x \in X} ||x - p||^2 = \sum_{x \in X} ||x - \Gamma(X)||^2 + |X| \cdot ||\Gamma(X) - p||^2.$

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k centers $C \subset \mathbb{R}^d$ such that the the following cost function is minimized:

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2 \right)$$

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k clusters $X_1, ..., X_k$ such that (i) the clusters satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where } \Gamma(X_i) = \frac{\sum_{x \in X_i} x}{|X_i|}.$$

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k centers $C \subset \mathbb{R}^d$ such that the the following cost function is minimized:

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2 \right)$$

Problem (Attempted formulation in terms of centers)

Given n points $X \subset \mathbb{R}^d$, an integer k, and a set of constraints $\mathbb{D}$, find k centers $C \subset \mathbb{R}^d$ such that...

Constrained k-means

Problem (k-means)

Given n points $X \subset \mathbb{R}^d$, and an integer k, find k centers $C \subset \mathbb{R}^d$ such that the the following cost function is minimized:

$$\Phi_C(X) = \sum_{x \in X} \min_{c \in C} \left(||x - c||^2 \right)$$

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, a set of constraints $\mathbb{D}$, and a partition algorithm $A^{\mathbb{D}}$, find k centers $C \subset \mathbb{R}^d$ such that the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \; \textit{where } (X_1, ..., X_k) \leftarrow A^{\mathbb{D}}(C, X).$$

Partition Algorithm [DX15]

Given a dataset X, constraints $\mathbb{D}$, and centers $C = (c_1, ..., c_k)$, the partition algorithm $A^{\mathbb{D}}(C, X)$ outputs a clustering $(X_1, ..., X_k)$ of X such that (i) all clusters X_i satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$cost(A^{\mathbb{D}}(C, X)) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2.$$

Constrained k-means

Partition Algorithm [DX15]

Given a dataset X, constraints $\mathbb{D}$, and centers $C = (c_1, ..., c_k)$, the partition algorithm $A^{\mathbb{D}}(C, X)$ outputs a clustering $(X_1, ..., X_k)$ of X such that (i) all clusters X_i satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$cost(A^{\mathbb{D}}(C, X)) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2.$$

- What is a partition algorithm for the k-means problem where there are no constraints on the clusters?

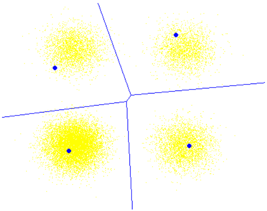
Constrained k-means

Partition Algorithm [DX15]

Given a dataset X, constraints $\mathbb{D}$, and centers $C = (c_1, ..., c_k)$, the partition algorithm $A^{\mathbb{D}}(C, X)$ outputs a clustering $(X_1, ..., X_k)$ of X such that (i) all clusters X_i satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$cost(A^{\mathbb{D}}(C, X)) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2.$$

- What is a partition algorithm for the k-means problem where there are no constraints on the clusters?
    - Voronoi partitioning algorithm.



Constrained k-means

Partition Algorithm [DX15]

Given a dataset X, constraints $\mathbb{D}$, and centers $C = (c_1, ..., c_k)$, the partition algorithm $A^{\mathbb{D}}(C, X)$ outputs a clustering $(X_1, ..., X_k)$ of X such that (i) all clusters X_i satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$cost(A^{\mathbb{D}}(C, X)) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2.$$

- Partition algorithm for r-gather clustering [DX15]:
  - Constraint: Each cluster should have at least r points.

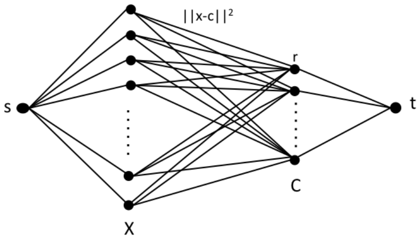


Figure : Partition algorithm: Minimum cost circulation.

Constrained k-means

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, a set of constraints $\mathbb{D}$, and a partition algorithm $A^{\mathbb{D}}$, find k centers $C \subset \mathbb{R}^d$ such that the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where} \ (X_1, ..., X_k) \leftarrow A^{\mathbb{D}}(C, X).$$

Partition Algorithm [DX15]

Given a dataset X, constraints $\mathbb{D}$, and centers $C = (c_1, ..., c_k)$, the partition algorithm $A^{\mathbb{D}}(C, X)$ outputs a clustering $(X_1, ..., X_k)$ of X such that (i) all clusters X_i satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$cost(A^{\mathbb{D}}(C, X)) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2.$$

- Theorem (Main result in [DX15]): There is a $(1 + \epsilon)$-approximation algorithm that runs in time $O(ndL + L \cdot T(A^{\mathbb{D}}))$, where $T(A^{\mathbb{D}})$ denotes running time of $A^{\mathbb{D}}$ and $L = (\log n)^k \cdot 2^{poly(k/\epsilon)}$.

Constrained k-means

Problem (Constrained k-means [DX15])

Given n points $X \subset \mathbb{R}^d$, an integer k, a set of constraints $\mathbb{D}$, and a partition algorithm $A^{\mathbb{D}}$, find k centers $C \subset \mathbb{R}^d$ such that the following cost function is minimized:

$$\Psi(X) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2, \ \textit{where}\ (X_1, ..., X_k) \leftarrow A^{\mathbb{D}}(C, X).$$

Partition Algorithm [DX15]

Given a dataset X, constraints $\mathbb{D}$, and centers $C = (c_1, ..., c_k)$, the partition algorithm $A^{\mathbb{D}}(C, X)$ outputs a clustering $(X_1, ..., X_k)$ of X such that (i) all clusters X_i satisfy $\mathbb{D}$ and (ii) the following cost function is minimized:

$$cost(A^{\mathbb{D}}(C, X)) = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2.$$

- Theorem (Main result in [DX15]): There is a $(1+\epsilon)$-approximation algorithm that runs in time $O(ndL + L \cdot T(A^{\mathbb{D}}))$, where $T(A^{\mathbb{D}})$ denotes running time of $A^{\mathbb{D}}$ and $L = (\log n)^k \cdot 2^{poly(k/\epsilon)}$.
- Theorem (Our Main Result): There is a $(1+\epsilon)$-approximation algorithm that runs in time $O(ndL + L \cdot T(A^{\mathbb{D}}))$, where $T(A^{\mathbb{D}})$ denotes running time of $A^{\mathbb{D}}$ and $L = 2^{\tilde{O}(k/\epsilon)}$.

Constrained k-means

A common theme for all PTAS

- Theorem (Main result in [DX15]): There is a $(1+\epsilon)$-approximation algorithm that runs in time $O(ndL + L \cdot T(A^{\mathbb{D}}))$, where $T(A^{\mathbb{D}})$ denotes running time of $A^{\mathbb{D}}$ and $L = (\log n)^k \cdot 2^{poly(k/\epsilon)}$.
- Theorem (Our Main Result): There is a $(1+\epsilon)$-approximation algorithm that runs in time $O(ndL + L \cdot T(A^{\mathbb{D}}))$, where $T(A^{\mathbb{D}})$ denotes running time of $A^{\mathbb{D}}$ and $L = 2^{\tilde{O}(k/\epsilon)}$.
- Running time of $(1+\epsilon)$-approximation algorithms for k-means:

Citation	**Approx. factor**	**Running Time**
[AV07]	$O(\log k)$	polynomial time
[KMN[+]02]	$9+\epsilon$	polynomial time
[KSS10, JKY15, FMS07]	$(1+\epsilon)$	$O\left(nd \cdot 2^{\tilde{O}(k/\epsilon)}\right)$

- How do these $(1+\epsilon)$-approximation algorithms work?
  - Enumerate a list of k-centers, $C_1, ..., C_l$ and then uses $A^{\mathbb{D}}$ to pick the best one.

List k-means

Problem (List k-means)

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find a list of k-centers, $C_1, ..., C_l$ such that for at least one index $j \in \{1, ..., l\}$, we have

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT,$$

where $C_j = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.

- Observation: Solution to the List k-means problem gives a solution to the constrained k-means problem.

List k-means

Problem (List k-means)

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find a list of k-centers, $C_1, ..., C_l$ such that for at least one index $j \in \{1, ..., l\}$, we have

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT,$$

where $C_j = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.

- Is outputting a list a necessary requirement?

List k-means

Problem (List k-means)

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find a list of k-centers, $C_1, ..., C_l$ such that for at least one index $j \in \{1, ..., l\}$, we have

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1+\epsilon) \cdot OPT,$$

where $C_j = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.

- Is outputting a list a necessary requirement?

Attempted problem definition without list

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find k-centers C such that:

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1+\epsilon) \cdot OPT,$$

where $C = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.

List k-means

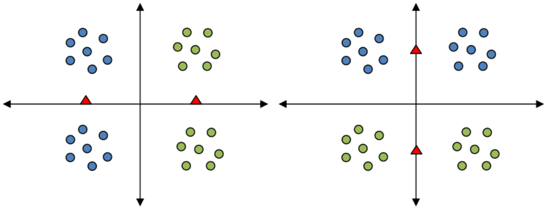
- Is outputting a list a necessary requirement?

Attempted problem definition without list

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find k-centers C such that:

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT,$$

where $C = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.



List k-means

Problem (List k-means)

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find a list of k-centers, $C_1, ..., C_l$ such that for at least one index $j \in \{1, ..., l\}$, we have

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT,$$

where $C_j = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.

- We can formulate an existential question related to the size of such a list.

Question

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Let L be the size of the smallest list of k centers such that there is at least one element $(c_1, ..., c_k)$ in this list such that $\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT$. What is the value of L?

List k-means

Problem (List k-means)

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Given X, k and ϵ, find a list of k-centers, $C_1, ..., C_l$ such that for at least one index $j \in \{1, ..., l\}$, we have

$$\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT,$$

where $C_j = (c_1, ..., c_k)$. Note that $OPT = \sum_{i=1}^{k} \sum_{x \in X_i} ||x - \Gamma(X_i)||^2$.

- We can formulate an existential question related to the size of such a list.

Question

Let $X \subset \mathbb{R}^d$, k be an integer, $\epsilon > 0$ and $X_1, ..., X_k$ be an arbitrary partition of X. Let L be the size of the smallest list of k centers such that there is at least one element $(c_1, ..., c_k)$ in this list such that $\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2 \leq (1 + \epsilon) \cdot OPT$. What is the value of L?

- Our results:
  - Lower bound: $\Omega\left(2^{\tilde{\Omega}\left(\frac{k}{\sqrt{\epsilon}}\right)}\right)$.
  - Upper bound: $O\left(2^{\tilde{O}\left(\frac{k}{\epsilon}\right)}\right)$.

List k-means

- Our results:
  - Lower bound: $\Omega\left(2^{\tilde{\Omega}\left(\frac{k}{\sqrt{\epsilon}}\right)}\right)$.
  - Upper bound: $O\left(2^{\tilde{O}\left(\frac{k}{\epsilon}\right)}\right)$.

Solving k-means via list k-means

Any $(1+\epsilon)$-approximation algorithm that solves k-means or constrained k-means via solving list k-means (which in fact all known algorithms do), then its running time cannot be smaller than $nd \cdot 2^{\tilde{\Omega}(k/\sqrt{\epsilon})}$.

- This explains the common running time expression for all known $(1+\epsilon)$-approximation algorithms.

Citation	Approx. factor	Running Time
[AV07]	$O(\log k)$	polynomial time
[KMN+02]	$9+\epsilon$	polynomial time
[KSS10, JKY15, FMS07]	$(1+\epsilon)$	$O\left(nd \cdot 2^{\tilde{O}(k/\epsilon)}\right)$

Main ideas for upper bound

Lemma ([IKI94])

Let S be a set of s point sampled independently from any given point set $X \subset \mathbb{R}^d$ uniformly at random. Then for any $\delta > 0$, the following holds with probability at least $(1 - \delta)$:

$$\Phi_{\Gamma(S)}(X) \leq \left(1 + \frac{1}{\delta \cdot s}\right) \cdot \Phi_{\Gamma(X)}(X), \;\; \textit{where } \Gamma(X) = \frac{\sum_{x \in X} x}{|X|}$$

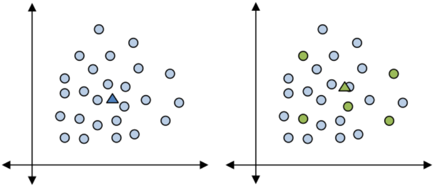
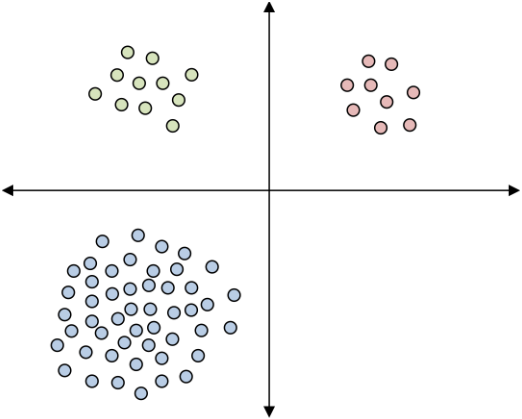


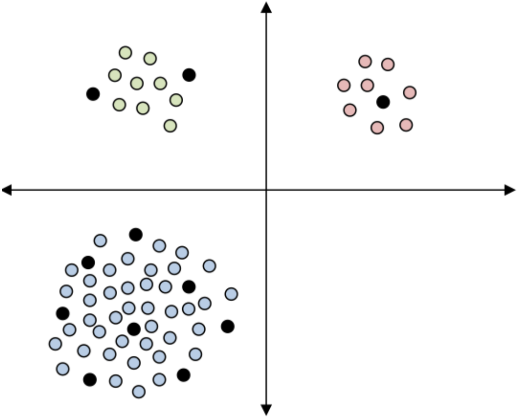
Figure : The cost w.r.t. the centroid (blue triangle) of all points (blue dots) is close to the cost w.r.t. the centroid (green triangle) of a few randomly chosen points (green dots).

- Consider the following simple case where the clusters are separated.

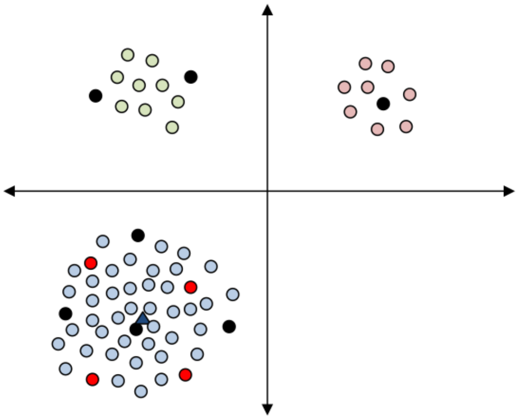


List k-means: upper bound
Main ideas

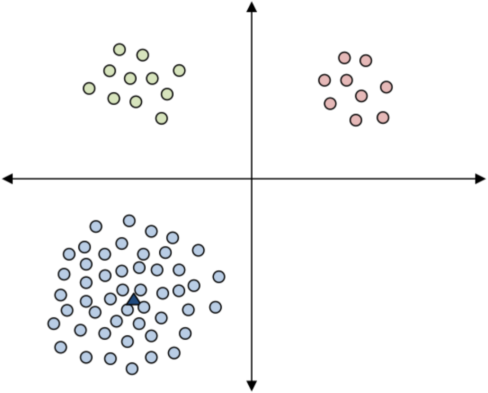
- We randomly sample N points.
- Then consider all possible subsets of the sampled points of size $M < N$.



- One of these subsets represents a uniform sample from the largest cluster.
- The centroid of this subset is a good center for this cluster.

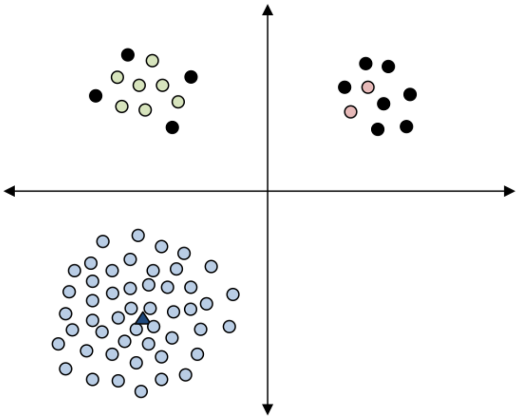


- At this point, we are done with the first cluster and would like to repeat.
- Sampling uniformly at random is not a good idea as other clusters might be small.

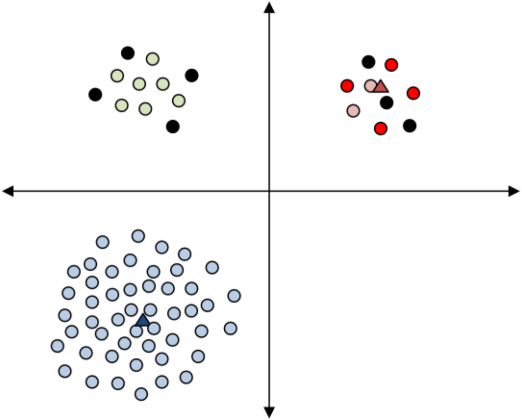


List k-means: upper bound
Main ideas

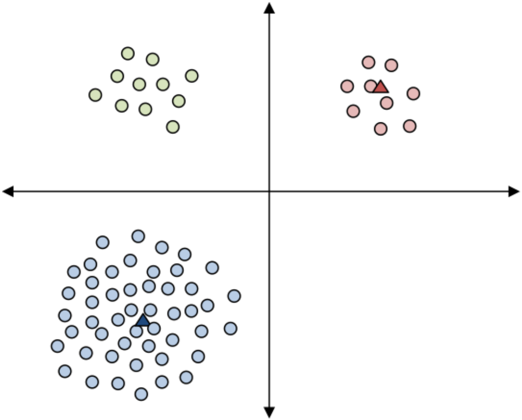
- Solution: We sample using D^2-sampling. That is, we sample using a non-uniform distribution that gives preference to points that are further away from the current centers.



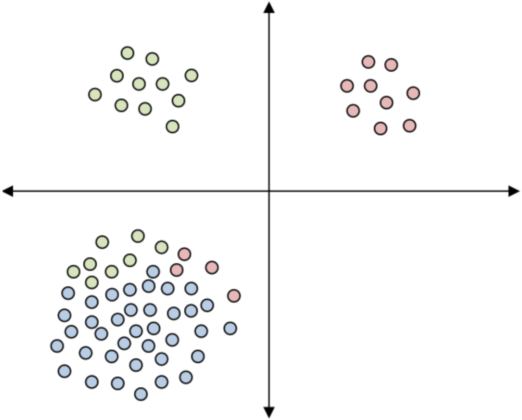
- Again, we consider all possible subsets and one of these subsets behaves like a uniform sample from a target cluster.



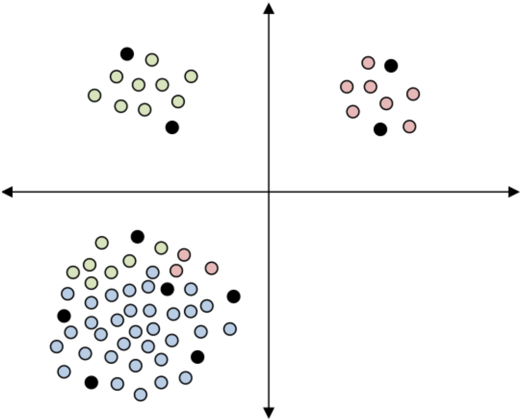
- So, the centroid of this subset if a good center for this cluster.
- Now,we just repeat.



- Consider a more complicated case where the target clusters are not well separated.

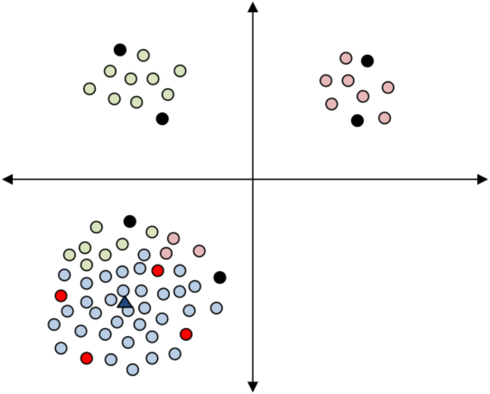


- Again, we start by sampling uniformly at random.

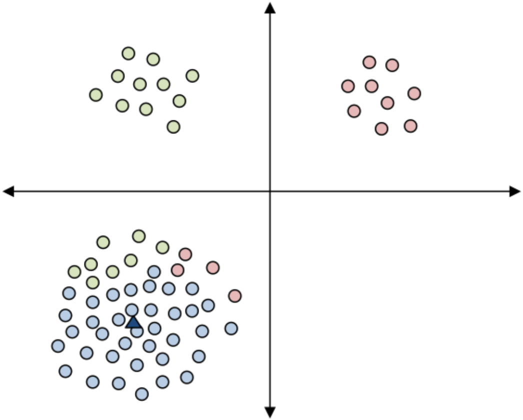


List k-means: upper bound
Main ideas

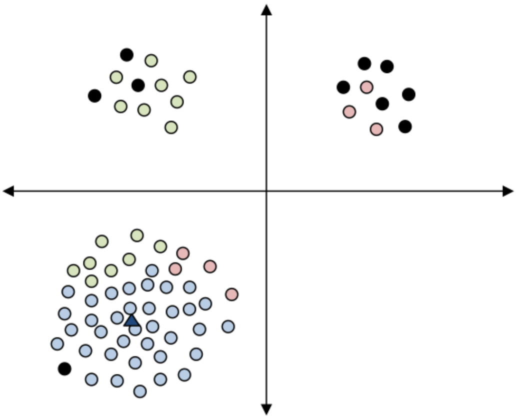
- Again, we start by sampling uniformly at random and considering all possible subsets.
- One of these subsets behave like a uniform sample from the largest cluster and its centroid is good for this cluster.



- Now we are done with the largest cluster and we do a D^2-sampling.

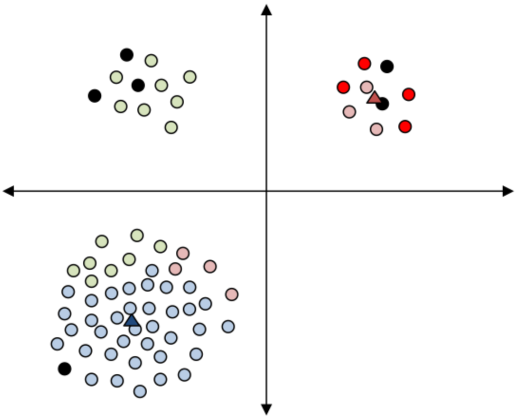


- Now we are done with the largest cluster and we do a D^2-sampling.
- Unfortunately, due to poor separability, none of the subsets behave like a uniform sample from the second cluster.

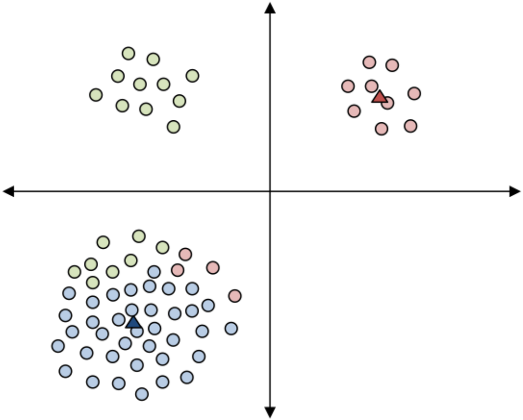


List k-means: upper bound
Main ideas

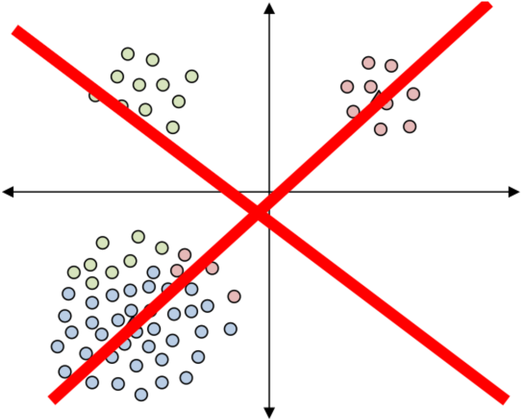
- Unfortunately, due to poor separability, none of the subsets behave like a uniform sample from the second cluster.
- So, we may end up not obtaining a good center for the second cluster.



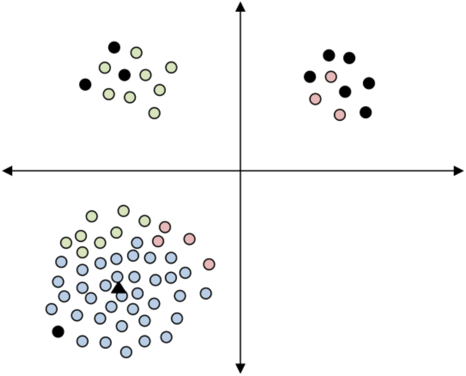
- So, we may end up not obtaining a good center for the second cluster.



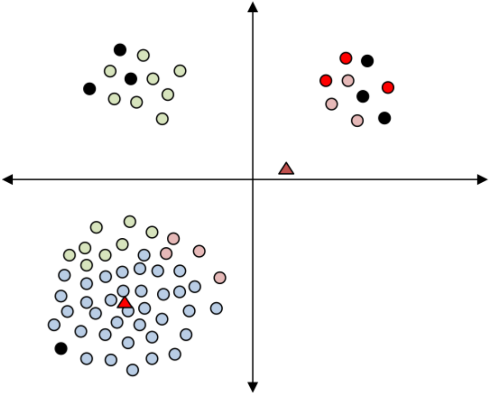
- So, we may end up not obtaining a good center for the second cluster.
- This is an undesirable result.



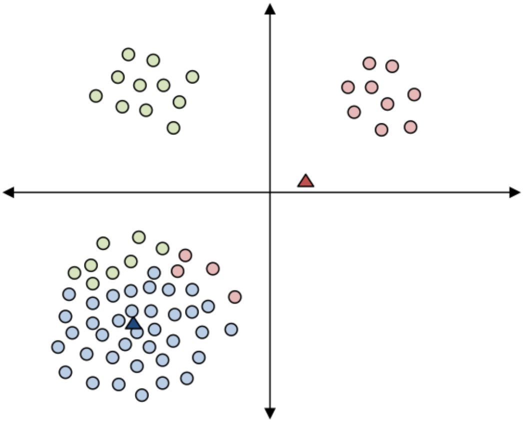
- Let us go back, the reason that D^2-sampling is unable to pick uniform samples from the second cluster is that some points of the cluster is close to the first chosen center.
- What we do is create multiple copies of the first center and add it to the set of points from which all possible subsets are considered.



- These multiple copies act as proxy for the points that are close to the first center.
- Now, one of the subsets behave like a uniform sample and we get a good center.



- And now we just repeat.



Conclusion

- We also get $(1+\epsilon)$-approximation algorithm for the k-median problem with running time $O\left(nd \cdot 2^{\tilde{O}\left(\frac{k}{\epsilon^{O(1)}}\right)}\right)$.
- Our algorithm and analysis easily extends to distance measures that satisfy certain "metric like" properties. This includes:
  - Mahalanobis distance
  - μ-similar Bregman divergence
- Open Problems:
  - Matching upper and lower bounds for list k-median problem.
  - Faster algorithms for specific versions of constrained k-means problem that are designed without going via the list k-means route.

References I

Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop, *The hardness of approximation of euclidean k-means*, CoRR **abs/1502.03316** (2015).

David Arthur and Sergei Vassilvitskii, *k-means++: the advantages of careful seeding*, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (Philadelphia, PA, USA), SODA '07, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

Sanjoy Dasgupta, *The hardness of k-means clustering*, Tech. Report CS2008-0916, Department of Computer Science and Engineering, University of California San Diego, 2008.

Hu Ding and Jinhui Xu, *A unified framework for clustering constrained data without locality property*, SODA'15, pp. 1471–1490, 2015.

Dan Feldman, Morteza Monemizadeh, and Christian Sohler, *A PTAS for k-means clustering based on weak coresets*, Proceedings of the twenty-third annual symposium on Computational geometry (New York, NY, USA), SCG '07, ACM, 2007, pp. 11–18.

Mary Inaba, Naoki Katoh, and Hiroshi Imai, *Applications of weighted voronoi diagrams and randomization to variance-based k-clustering: (extended abstract)*, Proceedings of the tenth annual symposium on Computational geometry (New York, NY, USA), SCG '94, ACM, 1994, pp. 332–339.

Ragesh Jaiswal, Mehul Kumar, and Pulkit Yadav, *Improved analysis of D^2-sampling based PTAS for k-means and other clustering problems*, Information Processing Letters **115** (2015), no. 2.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, *A local search approximation algorithm for k-means clustering*, Proc. 18th Annual Symposium on Computational Geometry, 2002, pp. 10–18.

References II

Amit Kumar, Yogish Sabharwal, and Sandeep Sen, *Linear-time approximation schemes for clustering problems in any dimensions*, J. ACM **57** (2010), no. 2, 5:1–5:32.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan, *The planar k-means problem is NP-hard*, Theoretical Computer Science **442** (2012), no. 0, 13 – 21, Special Issue on the Workshop on Algorithms and Computation (WALCOM 2009).

Andrea Vattani, *The hardness of k-means clustering in the plane*, Tech. report, Department of Computer Science and Engineering, University of California San Diego, 2009.

Thank you