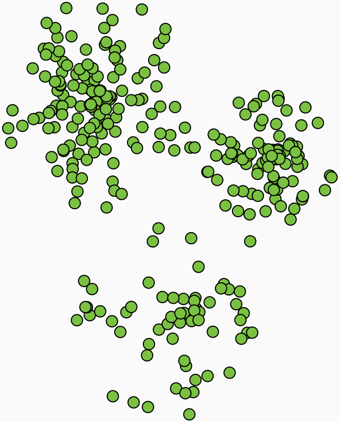
Improved Coresets for Euclidean k-Means

Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic
Chris Schwiegelshohn, Omar Ali Sheikh-Omar

k-Means Clustering

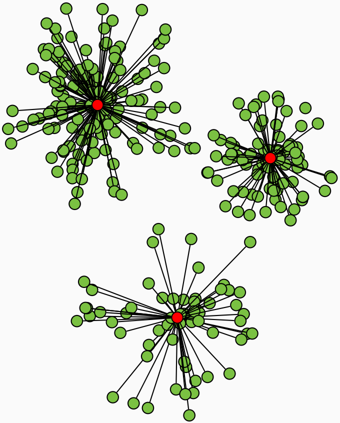


Problem Definition

Let A be a set of n points in d dimensional Euclidean space and let k be a positive integer. The objective consists of finding a set of k centers S minimizing

$$\mathrm{cost}(A, S) := \sum_{p \in A} \min_{c \in S} \|p - c\|^2.$$

k-Means Clustering



Problem Definition

Let A be a set of n points in d dimensional Euclidean space and let k be a positive integer. The objective consists of finding a set of k centers S minimizing

$$\mathrm{cost}(A, S) := \sum_{p \in A} \min_{c \in S} \|p - c\|^2.$$

Coreset Definition

Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

$$|\text{cost}_w(\Omega, S) - \text{cost}(A, S)| \leq \varepsilon \cdot \text{cost}(A, S)$$

Coreset Definition

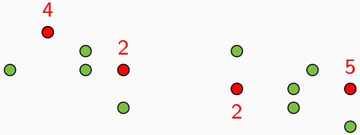
Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

$$|\mathrm{cost}_w(\Omega, S) - \mathrm{cost}(A, S)| \leq \varepsilon \cdot \mathrm{cost}(A, S)$$

Coreset Definition

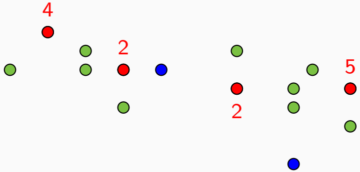
Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

$$|\text{cost}_w(\Omega, S) - \text{cost}(A, S)| \leq \varepsilon \cdot \text{cost}(A, S)$$



Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

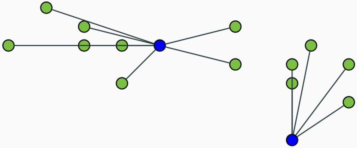
$$|\text{cost}_w(\Omega, S) - \text{cost}(A, S)| \leq \varepsilon \cdot \text{cost}(A, S)$$



Coreset Definition

Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

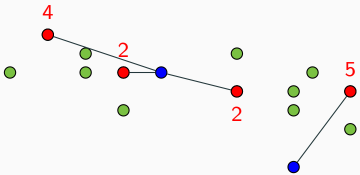
$$|\mathrm{cost}_w(\Omega, S) - \mathrm{cost}(A, S)| \leq \varepsilon \cdot \mathrm{cost}(A, S)$$



Coreset Definition

Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

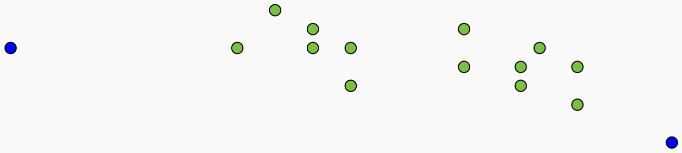
$$|\text{cost}_w(\Omega, S) - \text{cost}(A, S)| \leq \varepsilon \cdot \text{cost}(A, S)$$



Coreset Definition

Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

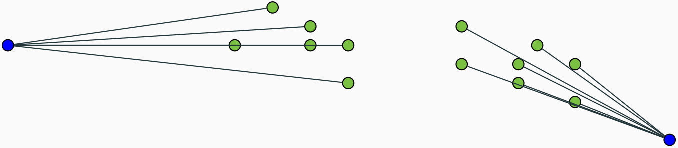
$$|\mathrm{cost}_w(\Omega, S) - \mathrm{cost}(A, S)| \leq \varepsilon \cdot \mathrm{cost}(A, S)$$



Coreset Definition

Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

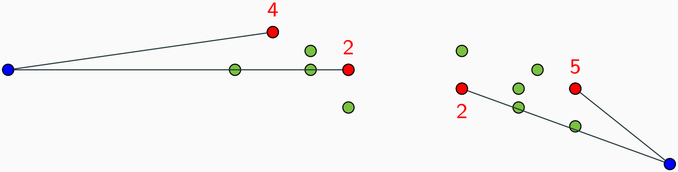
$$|\mathrm{cost}_w(\Omega, S) - \mathrm{cost}(A, S)| \leq \varepsilon \cdot \mathrm{cost}(A, S)$$



Coreset Definition

Given a set of points A, a weighted subset $\Omega \subset A$ is a (k, ε)-coreset if for *all* sets S of k centers it holds

$$|\text{cost}_w(\Omega, S) - \text{cost}(A, S)| \leq \varepsilon \cdot \text{cost}(A, S)$$



Theoretical Results on Coresets for Euclidean k-Means

Upper Bounds

Har-Peled, Mazumdar (STOC'04)	$O(k\epsilon^{-d+2}\log n)$
Chen (Sicomp'09)	$O(dk^2\epsilon^{-2}\log n)$
Langberg, Schulman (SODA'10)	$O(d^2k^3\epsilon^{-2})$
Feldman, Langberg (STOC'11)	$O(dk\epsilon^{-4})$
Feldman, Schmidt, Sohler (Sicomp'20)	$O(k^3\epsilon^{-4})$
Becchetti, Bury, Cohen-Addad, Grandoni, S. (STOC'19)	$O(k\epsilon^{-8})$
Huang, Vishnoi (STOC'20)	$O(k\epsilon^{-6})$
Braverman, Jiang, Krauthgamer, Wu (SODA'21)	$O(k^2\epsilon^{-4})$
Cohen-Addad, Saulpic, S. (STOC'21)	$O(k\epsilon^{-4})$
Cohen-Addad, Larsen, Saulpic, S. (STOC'22)	$O(k^2\epsilon^{-2})$

Lower Bounds

Cohen-Addad, Larsen, Saulpic, S. (STOC'22)	$\Omega(k\epsilon^{-2})$

Theoretical Results on Coresets for Euclidean k-Means

Upper Bounds

Har-Peled, Mazumdar (STOC'04)	$O(k\epsilon^{-d+2}\log n)$
Chen (Sicomp'09)	$O(dk^2\epsilon^{-2}\log n)$
Langberg, Schulman (SODA'10)	$O(d^2k^3\epsilon^{-2})$
Feldman, Langberg (STOC'11)	$O(dk\epsilon^{-4})$
Feldman, Schmidt, Sohler (Sicomp'20)	$O(k^3\epsilon^{-4})$
Becchetti, Bury, Cohen-Addad, Grandoni, S. (STOC'19)	$O(k\epsilon^{-8})$
Huang, Vishnoi (STOC'20)	$O(k\epsilon^{-6})$
Braverman, Jiang, Krauthgamer, Wu (SODA'21)	$O(k^2\epsilon^{-4})$
Cohen-Addad, Saulpic, S. (STOC'21)	$O(k\epsilon^{-4})$
Cohen-Addad, Larsen, Saulpic, S. (STOC'22)	$O(k^2\epsilon^{-2})$
Cohen-Addad, Larsen, Saulpic, S., Sheikh-Omar (NeurIPS'22)	$O(k^{1.5}\epsilon^{-2})$

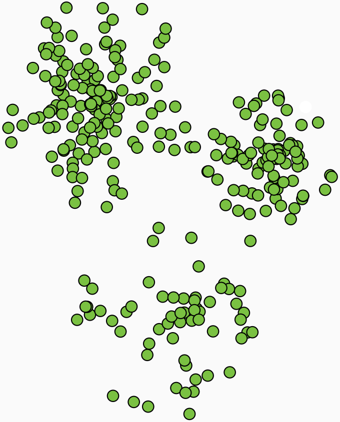
Lower Bounds

Cohen-Addad, Larsen, Saulpic, S. (STOC'22)	$\Omega(k\epsilon^{-2})$

Coreset Algorithms

Algorithm

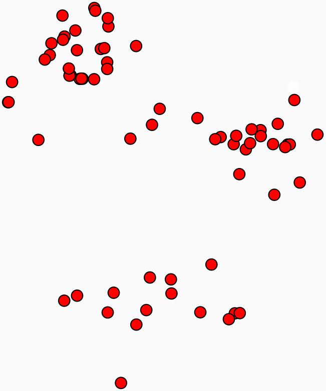
1. Sample S points (typically non unformly)
2. Weigh each point inversely proportionate to the sampling probability



Coreset Algorithms

Algorithm

1. Sample S points (typically non unformly)
2. Weigh each point inversely proportionate to the sampling probability



Analysis Outline

First, we show that the cost of *any arbitrary* solution is approximated.

Second, we show that the cost for *all* solutions is approximated.

Analysis Outline

First, we show that the cost of *any arbitrary* solution is approximated.

Second, we show that the cost for *all* solutions is approximated.

In previous analyses, we have the option of obtaining the following.

First Step	Second Step	Overall
$\Theta(\varepsilon^{-2}\min(k,\varepsilon^{-2}))$	$\Theta(k)$	$O(k\varepsilon^{-2}\min(k,\varepsilon^{-2}))$

Analysis Outline

First, we show that the cost of *any arbitrary* solution is approximated.

Second, we show that the cost for *all* solutions is approximated.

In previous analyses, we have the option of obtaining the following.

We obtain:

First Step	Second Step	Overall
$\Theta(\varepsilon^{-2}\min(k, \varepsilon^{-2}))$	$\Theta(k)$	$O(k\varepsilon^{-2}\min(k, \varepsilon^{-2}))$
$\Theta(\varepsilon^{-2})$	$O(k \cdot \sqrt{k})$	$O(k^{1.5}\varepsilon^{-2})$