Coresets for Probabilistic Clustering

Melanie Schmidt

Joint work with
Christiane Lammersen, Christian Sohler

15.12.2011

Clustering Certain Data: Metric k-median

Clustering Certain Data: Metric k-median

Given

Clustering Certain Data: Metric k-median

Given

- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$

Clustering Certain Data: Metric k-median

Given

- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $\mathcal{C} \subseteq X$

Clustering Certain Data: Metric k-median

Given

- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $\mathcal{C} \subseteq X$

Wanted

- A set $C := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and

Clustering Certain Data: Metric k-median

Given

- set of points $P : \{p_1, \ldots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $\mathcal{C} \subseteq X$

Wanted

- A set $C := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and
- an assignment $\rho : P \to C$ minimizing

$$\mathrm{cost}(P, C, \rho) := \sum_{i=1}^{n} \mathrm{D}(p_i, \rho(p_i)).$$

Clustering Certain Data: Metric k-median

Given

- set of points $P : \{p_1, \dots, p_n\}$ from metric space $M = (X, D)$
- set of center candidates $\mathcal{C} \subseteq X$

Wanted

- A set $C := \{c_1, \dots, c_k\} \subseteq \mathcal{C}$ and
- an assignment $\rho : P \to C$ minimizing

$$\text{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} \mathrm{D}(p_i, c).$$

Clustering Certain Data

Clustering Certain Data

General Metrics

Usually $\mathcal{C} = P$.

- no $(1+\varepsilon)$-approximation for $\varepsilon < 0.73$ (Jain et al. 2002)
- first constant factor approximation by Charikar et al. (1999)
- approximation guarantee consecutively improved, 3-approximation by Arya et al. (2001)

Clustering Certain Data

General Metrics

Usually $\mathcal{C} = P$.

- no $(1+\varepsilon)$-approximation for $\varepsilon < 0.73$ (Jain et al. 2002)
- first constant factor approximation by Charikar et al. (1999)
- approximation guarantee consecutively improved, 3-approximation by Arya et al. (2001)

Euclidean Metric

Usually $X = \mathbb{R}^d$, $M = ||\cdot||$ and $\mathcal{C} = \mathbb{R}^d$.

- First $(1+\varepsilon)$-approximation by Arora et al. (1998)
- Several improvements reducing the running time
- Chen: $(1+\varepsilon)$-approximation, pol. in the dimension (2006)

Clustering Certain Data

Coresets

Given a set of points P, a weighted subset $S \subset P$ is a (k, ε)-coreset if for all sets $C \subset \mathcal{C}$ of k centers it holds

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon \, \mathrm{cost}(P, C)$$

where $\mathrm{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p) \mathrm{D}(p, c)$.

Clustering Certain Data

Coresets

Given a set of points P, a weighted subset $S \subset P$ is a (k, ε)-coreset if for all sets $C \subset \mathcal{C}$ of k centers it holds

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon\, \mathrm{cost}(P, C)$$

where $\mathrm{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p) \mathrm{D}(p, c)$.

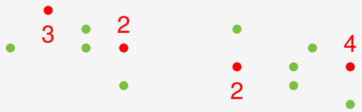
Clustering Certain Data

Coresets

Given a set of points P, a weighted subset $S \subset P$ is a (k, ε)-coreset if for all sets $C \subset \mathcal{C}$ of k centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \,\text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p) \mathrm{D}(p, c)$.



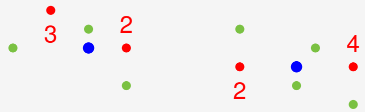
Clustering Certain Data

Coresets

Given a set of points P, a weighted subset $S \subset P$ is a (k, ε)-coreset if for all sets $C \subset \mathcal{C}$ of k centers it holds

$$|\text{cost}_w(S, C) - \text{cost}(P, C)| \leq \varepsilon \, \text{cost}(P, C)$$

where $\text{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p) \mathrm{D}(p, c)$.



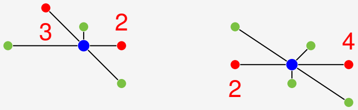
Clustering Certain Data

Coresets

Given a set of points P, a weighted subset $S \subset P$ is a (k, ε)-coreset if for all sets $C \subset \mathcal{C}$ of k centers it holds

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon \, \mathrm{cost}(P, C)$$

where $\mathrm{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p) \mathrm{D}(p, c)$.



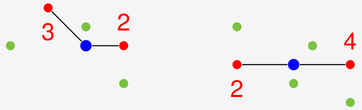
Clustering Certain Data

Coresets

Given a set of points P, a weighted subset $S \subset P$ is a (k, ε)-coreset if for all sets $C \subset \mathcal{C}$ of k centers it holds

$$|\mathrm{cost}_w(S, C) - \mathrm{cost}(P, C)| \leq \varepsilon\, \mathrm{cost}(P, C)$$

where $\mathrm{cost}_w(S, C) = \sum_{p \in S} \min_{c \in C} w(p)\mathrm{D}(p, c)$.



Clustering Certain Data

Clustering Certain Data

Coreset constructions

- '02: Bădoiu, Har-Peled and Indyk: First coreset construction for clustering problems
- '04: Agarwal, Har-Peled and Varadarajan: Definition of coresets as used nowadays
- '04: Har-Peled and Mazumdar, Coreset of size $\mathcal{O}(k\varepsilon^{-d} \log n)$, maintainable in data streams
- '05: Har-Peled, Kushal: Coreset of size $\mathcal{O}(k^2\varepsilon^{-d})$
- '05: Frahling and Sohler: Coreset of size $\mathcal{O}(k\varepsilon^{-d} \log n)$, insertion-deletion data streams
- '06: Chen: Coresets for metric and Euclidean k-median and k-means, polynomial in d,n and ε^{-1}
- '07: Feldman, Monemizadeh, Sohler: weak coresets, poly(k, ε^{-1})

Clustering Uncertain Data

Clustering Uncertain Data

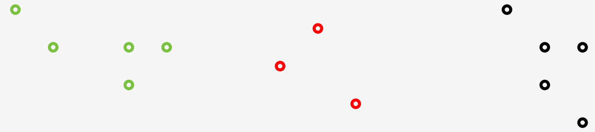
Uncertain Data

Input consisting of (discrete) distributions instead of points

Clustering Uncertain Data

Uncertain Data

Input consisting of (discrete) distributions instead of points



Clustering Uncertain Data

Uncertain Data

Input consisting of (discrete) distributions instead of points

Sources of Uncertain Data

- measurements of sensor networks
- linkage across multiple databases

Metric Assigned Probabilistic k-Median Clustering

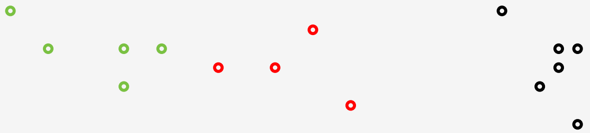
Given

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,

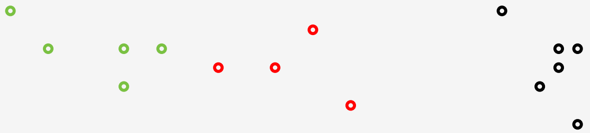
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$



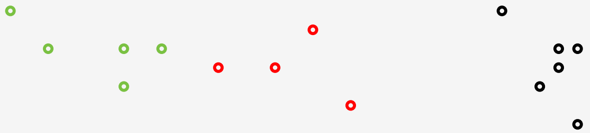
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,



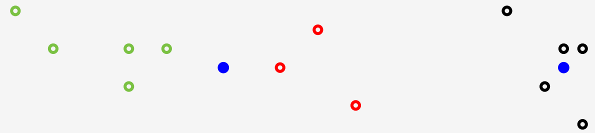
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $\mathcal{C} \subset X$.



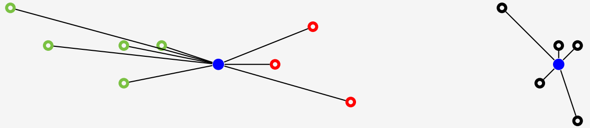
Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $\mathcal{C} \subset X$.



Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $\mathcal{C} \subset X$.



Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $\mathcal{C} \subset X$.

Wanted

- A set $C := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and
- an assignment $\rho : V \to C$ minimizing

$$\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V, C, \rho)\right] := \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)).$$

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,
- set of possible center locations $\mathcal{C} \subset X$.

Wanted

- A set $C := \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and
- an assignment $\rho : V \to C$ minimizing

$$\mathbf{E}_{\mathcal{D}}\left[\mathrm{cost}(V, C)\right] := \min_{\rho: V \to C} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)).$$

Related work

Cormode, McGregor (PODS 2008)

- $(1+\varepsilon)$-approximation for a variant of the above problem
- $(1+\varepsilon)$-approximation for uncertain k-means
- Constant approximation for (assigned) metric k-median
- Bicriteria approximations for uncertain metric k-center

Guha and Munagala (PODS 2009)

- Constant approximation for uncertain metric k-center

Probabilistic Coresets

What should a probabilistic coreset look like?

Probabilistic Coresets

What should a probabilistic coreset look like?

- Consists of probabilistic points (nodes)

Probabilistic Coresets

What should a probabilistic coreset look like?

- Consists of probabilistic points (nodes)
- The probability distributions of the nodes should be sparse

Probabilistic Coresets

What should a probabilistic coreset look like?

- Consists of probabilistic points (nodes)
- The probability distributions of the nodes should be sparse

Coresets

Given a set of uncertain nodes V, a weighted subset U is a (k, ε)-coreset if for all sets C of k centers it holds

$$|\mathbf{E}_{\mathcal{D}'}[\mathrm{cost}_w(U, C)] - \mathbf{E}_{\mathcal{D}}[\mathrm{cost}(V, C)]| \leq \varepsilon \mathbf{E}_{\mathcal{D}}[\mathrm{cost}(V, C)]$$

where $\mathbf{E}_{\mathcal{D}'}[\mathrm{cost}_w(U, C)] := \min_{\rho: U \to C} \sum_{v_i \in U} \sum_{j=1}^{m} p'_{ij} w(v_i) \mathrm{D}(x_j, \rho(v_i)).$

Metric k-median

Idea

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

- Expected distance?

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

- Expected distance?
- Expected distance between two copies of the same probabilistic node is not zero

Metric k-median

Idea

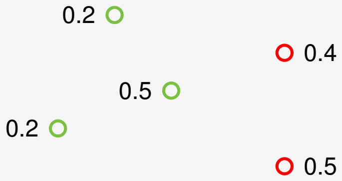
- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

- Expected distance?
- Expected distance between two copies of the same probabilistic node is not zero
- $\rightsquigarrow$ expected distance is not a metric

Metric k-median

Idea

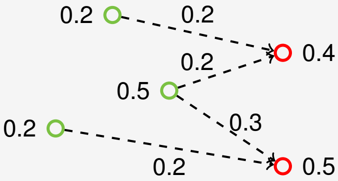
- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?



Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?



Earth Mover Distance

Let $v_{i_1}, v_{i_2} \in V$ and let $p_{i_1} = p_{i_2}$. A mapping $\varrho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ morphs v_{i_1} into v_{i_2} if for all $x_{j_1}, x_{j_2} \in \mathcal{X}$:

$$\sum_{x_j \in \mathcal{X}} \varrho(x_{j_1}, x_j) = p_{i_1 j_1} \quad \text{and} \quad \sum_{x_j \in \mathcal{X}} \varrho(x_j, x_{j_2}) = p_{i_2 j_2}.$$

The cost of ϱ is defined as

$$\sum_{x_{j_1} \in \mathcal{X}} \sum_{x_{j_2} \in \mathcal{X}} \varrho(x_{j_1} x_{j_2}) \cdot \mathrm{D}(x_{j_1}, x_{j_2}).$$

The earth mover distance EMD between v_{i_1} and v_{i_2} is the minimum cost of a mapping that morphs v_{i_1} into v_{i_2}.

Morphing Probability Distributions

- EMD is a metric

- EMD is a metric
- EMD is a generalization of the cost function

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\leadsto \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions
  - for uniform realization probabilities.

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions
  - for uniform realization probabilities.

Fixing Issues

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions
  - for uniform realization probabilities.

Fixing Issues

- for general realization probabilities, group nodes and round to $p_{\min}(1+\varepsilon)^{\ell} \rightarrow$ error is a factor $(1+\varepsilon)$

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions
  - for uniform realization probabilities.

Fixing Issues

- for general realization probabilities, group nodes and round to $p_{\min}(1+\varepsilon)^{\ell} \rightarrow$ error is a factor $(1+\varepsilon)$
- to compute the EMD efficiently, use det. $(1, \varepsilon)$-coresets of the nodes $\rightarrow$ also ensures sparsity of coreset nodes

Theorem

A probabilistic (k, ε)-coreset of size

$$\mathcal{O}(k\varepsilon^{-3} \cdot \text{polylog}(|\mathcal{C}|, n, \delta, 1/p_{\min}))$$

can be computed in time

$$\mathcal{O}(nm + \varepsilon^{-10}kn \cdot \text{polylog}(|\mathcal{C}|, n, m, \delta, 1/p_{\min}))$$

with error probability δ. The probability distributions have size

$$\mathcal{O}(\varepsilon^{-3} \cdot \text{polylog}(|\mathcal{C}|, n, \delta, 1/p_{\min})).$$

Partitioning nodes

Does the same approach work in the Euclidean case?

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

$\leadsto$ Develop coreset construction
$\leadsto$ Use deterministic coreset construction by Chen

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

$\rightsquigarrow$ algorithms for the general case do not work here

$\rightsquigarrow$ Develop coreset construction

$\rightsquigarrow$ Use deterministic coreset construction by Chen

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

$\rightsquigarrow$ algorithms for the general case do not work here

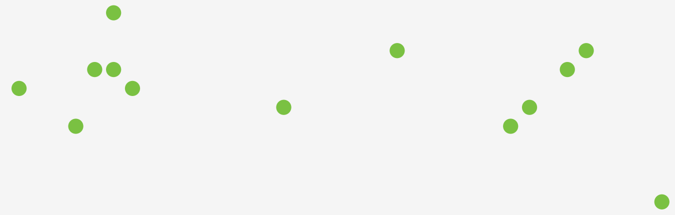
$\rightsquigarrow$ even though probabilistic Euclidean k-median can be seen as deterministic metric k-median, we cannot use deterministic algorithms

$\rightsquigarrow$ Develop coreset construction

$\rightsquigarrow$ Use deterministic coreset construction by Chen

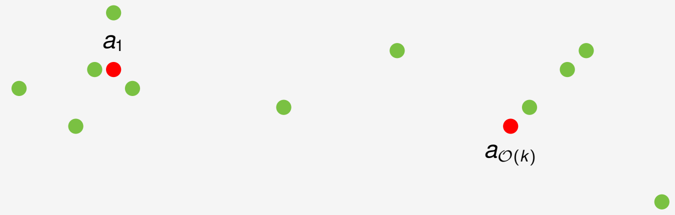
Partitioning nodes

Chen (2006)



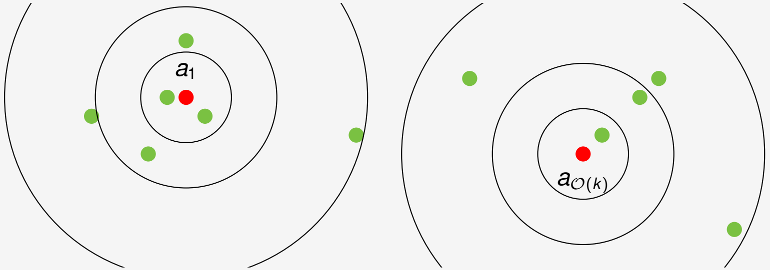
Chen (2006)

- compute bicriteria approximation



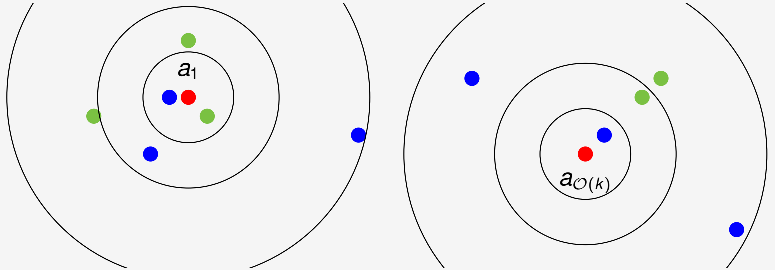
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost



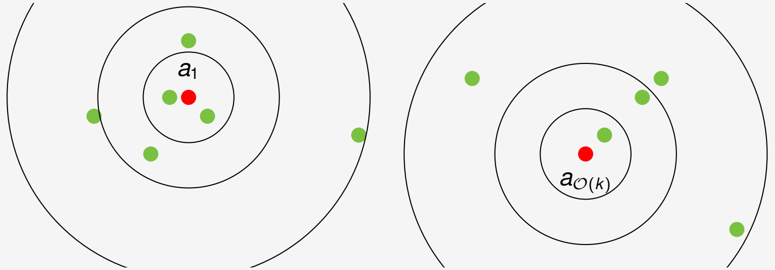
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



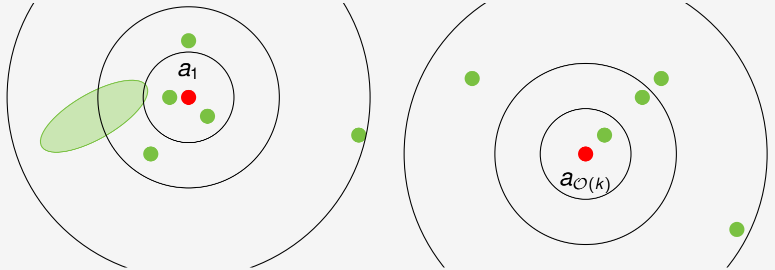
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



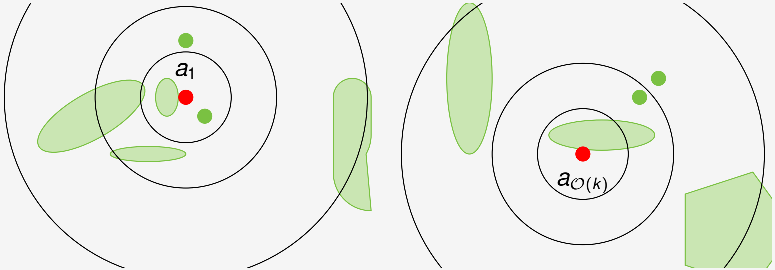
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



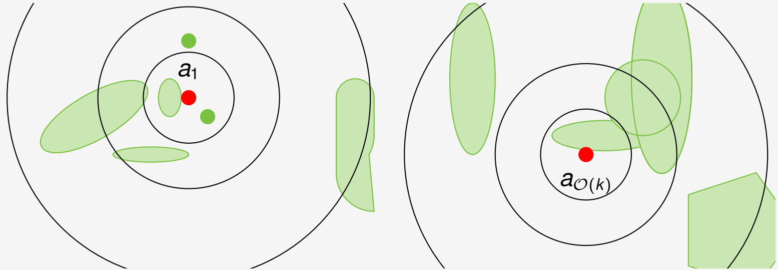
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



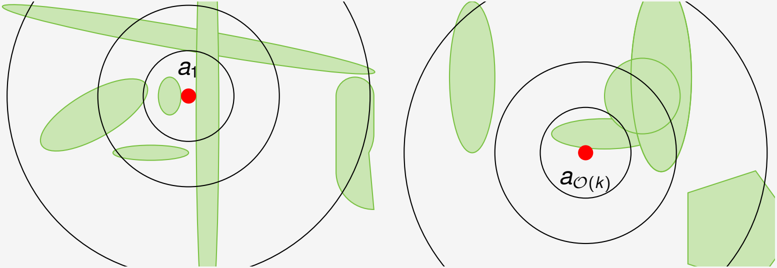
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



Probabilistic Coreset Construction

Probabilistic Coreset Construction

1. Compute 1-medians

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis

- Show that 1. yields a bicriteria approximation for the probabilistic problem

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis

- Show that 1. yields a bicriteria approximation for the probabilistic problem
- Find nice formulation of error in 4.

Probabilistic Coreset Construction

1. Compute 1-medians
2. Partition 1-medians like Chen
3. Refine partitioning according to clustering behaviour
4. Sample sufficiently many points from each partition

Analysis

- Show that 1. yields a bicriteria approximation for the probabilistic problem
- Find nice formulation of error in 4.
- Bound error by analyzing properties of 3.

Result for Euclidean k-median

Theorem

A probabilistic (k, ε)-coreset of size

$$\mathcal{O}(k^2\varepsilon^{-2}d \cdot \operatorname{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min}))$$

can be computed in time

$$\mathcal{O}(knm \cdot \operatorname{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min}))$$

with error probability δ. The probability distributions have size

$$\mathcal{O}(\varepsilon^{-2}d \cdot \operatorname{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min})).$$

Thank you for your attention!