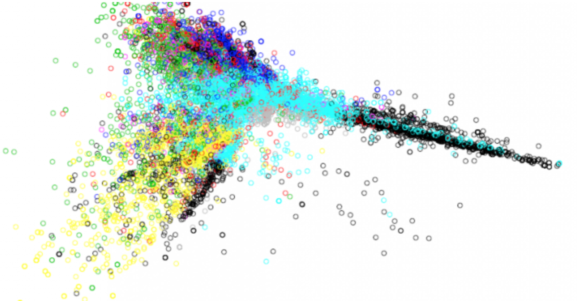
Hardness of Approximation for Metric Clustering

Karthik C. S.
(Rutgers University)

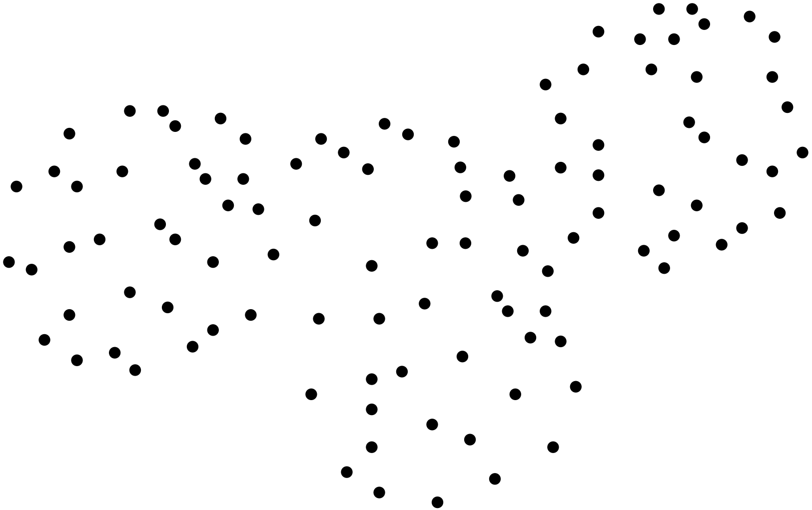
March 4th 2022



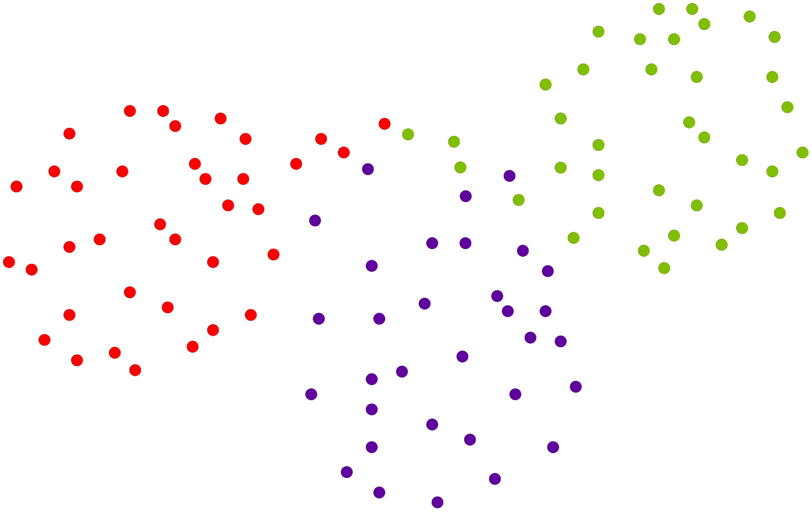
Classifying Handwritten Digits

28×28
grayscale image

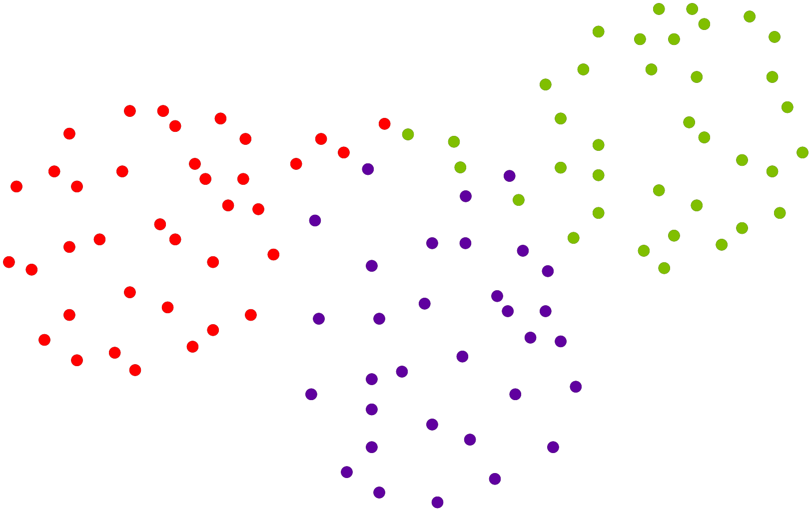
Clustering: Abstraction



Clustering: Abstraction



Clustering: Abstraction



Task of Classifying Input Data

Clustering: Applications

- Reveal internal structure of data
    - Clustering gene expression

Clustering: Applications

- Reveal internal structure of data
  - Clustering gene expression
- Partition data
  - Market segmentation

Clustering: Applications

- Reveal internal structure of data
  - Clustering gene expression
- Partition data
  - Market segmentation
- Data Preparation
  - Summarize news

Clustering: Applications

- Reveal internal structure of data
  - Clustering gene expression
- Partition data
  - Market segmentation
- Data Preparation
  - Summarize news
- Data Exploration
  - Underlying rules and Reoccurring patterns

Clustering: Modeling

- (Γ, Δ) is a metric space

Clustering: Modeling

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$

Clustering: Modeling

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$

Clustering: Modeling

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$

Clustering: Modeling

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

Continuous Version

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

Discrete
~~Continuous~~ Version

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

Discrete ~~Continuous~~ Version

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$ and $\mathcal{S} \subseteq \Gamma$
- Output: A classification (C, σ):
  - $C \subseteq$ ~~X~~ $\mathcal{S}$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

What is Good Classification?

- k-means, k-median, k-center, min-sum, correlation clustering . . .

What is Good Classification?

- k-means, k-median, k-center, min-sum, correlation clustering ...
- k-center value of (C, σ)

$$\max_{x \in X} \Delta(x, \sigma(x))$$

What is Good Classification?

- k-means, k-median, k-center, min-sum, correlation clustering ...
- k-center value of (C, σ)

$$\max_{x \in X} \Delta(x, \sigma(x))$$

- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

What is Good Classification?

- k-means, k-median, k-center, min-sum, correlation clustering ...
- k-center value of (C, σ)

$$\max_{x \in X} \Delta(x, \sigma(x))$$

- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

- k-means value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))^2$$

What is Good Classification?

- k-means, k-median, k-center, min-sum, correlation clustering ...
- k-center value of (C, σ)

$$\max_{x \in X} \Delta(x, \sigma(x))$$

- k-median value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))$$

- k-means value of (C, σ)

$$\sum_{x \in X} \Delta(x, \sigma(x))^2$$

- Don't fit: Facility Location, Hierarchical Clustering ...

Computational Question

Given (X, S, k) as input find a classification (C, σ) that minimizes the Clustering objective

Computational Question

Given (X, S, k) as input find a classification (C, σ) that minimizes the Clustering objective

Clustering Problem for objective Λ

Computational Question

Given (X, S, k) as input find a classification (C, σ) that minimizes the Clustering objective

Clustering Problem for objective Λ

Yes: There is classification (C^*, σ^*), such that $\Lambda(X, \sigma^*) \leq \beta$

Computational Question

Given (X, S, k) as input find a classification (C, σ) that minimizes the Clustering objective

Clustering Problem for objective Λ

Yes: There is classification (C^*, σ^*), such that $\Lambda(X, \sigma^*) \leq \beta$

No: For all classification (C, σ), we have $\Lambda(X, \sigma) > \beta$

NP-Hard

Efficient Approximation

NP-Hard to Approximate

Hardness of Approximation

- Many important problems are not tractable

Hardness of Approximation

- Many important problems are not tractable
- Need to cope with the intractability

Hardness of Approximation

- Many important problems are not tractable
- Need to cope with the intractability
- Design algorithms that find solutions whose cost or value is close to the optimum

Hardness of Approximation

- Many important problems are not tractable
- Need to cope with the intractability
- Design algorithms that find solutions whose cost or value is close to the optimum
- For some fundamental problems finding good approximate solutions is as hard as finding optimal solutions

Hardness of Approximation

- Many important problems are not tractable
- Need to cope with the intractability
- Design algorithms that find solutions whose cost or value is close to the optimum
- For some fundamental problems finding good approximate solutions is as hard as finding optimal solutions
- Area studying such results: Hardness of Approximation

New Computational Question

Given (X, S, k) as input find a classification (C, σ) that approximately minimizes the Clustering objective

New Computational Question

Given (X, S, k) as input find a classification (C, σ) that approximately minimizes the Clustering objective

Clustering Problem for objective Λ

New Computational Question

Given (X, S, k) as input find a classification (C, σ) that approximately minimizes the Clustering objective

Clustering Problem for objective Λ

Yes: There is classification (C^*, σ^*), such that $\Lambda(X, \sigma^*) \leq \beta$

New Computational Question

Given (X, S, k) as input find a classification (C, σ) that approximately minimizes the Clustering objective

Clustering Problem for objective Λ

Yes: There is classification (C^*, σ^*), such that $\Lambda(X, \sigma^*) \leq \beta$

No: For all classification (C, σ), we have $\Lambda(X, \sigma) > (1 + \delta) \cdot \beta$

k-center

k-center modeling

- Input: $X, S \subseteq \mathbb{R}^d, k \in \mathbb{N}$

k-center modeling

- Input: $X, S \subseteq \mathbb{R}^d, k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq S$ and $|C| = k$
  - $\sigma : X \to C$
  - (C, σ) minimizes $\max_{x \in X} \|x - \sigma(x)\|_p$

State-of-the-art: General Metrics

- NP-hard [FPT81]

State-of-the-art: General Metrics

- NP-hard [FPT81]
- Poly Time 3-approximation (Gonzalez Algorithm)

State-of-the-art: General Metrics

- NP-hard [FPT81]
- Poly Time 3-approximation (Gonzalez Algorithm)
- NP-Hard to approximate to $3 - o(1)$ factor! [FPT81]

State-of-the-art: ℓ_p Metrics

- ℓ_1 and ℓ_∞ metrics
  - Poly Time 3-approximation

State-of-the-art: ℓ_p Metrics

- ℓ_1 and ℓ_∞ metrics
  - Poly Time 3-approximation
  - NP-Hard to approximate to $3 - o(1)$ factor! [FG88]
- Euclidean metric
  - Poly Time 2.74-approximation! [NSS13]

State-of-the-art: ℓ_p Metrics

- ℓ_1 and ℓ_∞ metrics
  - Poly Time 3-approximation
  - NP-Hard to approximate to $3 - o(1)$ factor! [FG88]
- Euclidean metric
  - Poly Time 2.74-approximation! [NSS13]
  - NP-Hard to approximate to 2.65 factor [FG88]

Proof Overview: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$

Proof Overview: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Proof Overview: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Karp'72)

It is NP-hard to distinguish:

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Max Coverage is 1

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is < 1

Proof Overview: General Metrics

Theorem (Karp'72)

It is NP-hard to distinguish:

Proof Overview: General Metrics

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Max Coverage is 1

Proof Overview: General Metrics

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is < 1

$\Downarrow$

Proof Overview: General Metrics

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is < 1

$$\Downarrow$$

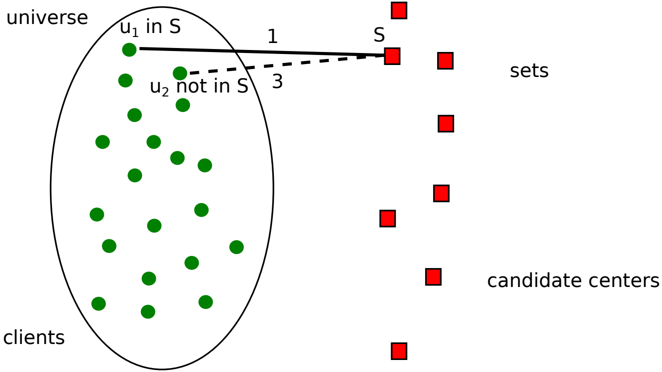
Theorem (Fowler-Paterson-Tanimoto'81)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\max_{x \in X} \Delta(x, \sigma^*(x)) \leq 1$

NO: For all (C, σ) we have $\max_{x \in X} \Delta(x, \sigma(x)) \geq 3$

Proof Overview: General Metrics



Proof Overview: General Metrics

Theorem (Fowler-Paterson-Tanimoto'81)

Given input (X, S, k). It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\max_{x \in X} \Delta(x, \sigma^*(x)) \leq 1$

NO: For all (C, σ) we have $\max_{x \in X} \Delta(x, \sigma(x)) \geq 3$

k-means & k-median

k-means and k-median modeling

- Input: $X, S \subseteq \mathbb{R}^d$, $k \in \mathbb{N}$

k-means and k-median modeling

- Input: $X, S \subseteq \mathbb{R}^d$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq S$ and $|C| = k$
  - $\sigma : X \to C$
  - k-means: (C, σ) minimizes $\sum_{x \in X} \|x - \sigma(x)\|_p^2$
  - k-median: (C, σ) minimizes $\sum_{x \in X} \|x - \sigma(x)\|_p$

Exact Computation

- NP-hard when $k = 2$ (Dasgupta'07)

Exact Computation

- NP-hard when $k = 2$ (Dasgupta'07)
- NP-hard in Euclidean plane (Megiddo–Supowit'84, Mahajan–Nimbhorkar–Varadarajan'12)

Exact Computation

- NP-hard when $k = 2$ (Dasgupta'07)
- NP-hard in Euclidean plane
  (Megiddo–Supowit'84,
  Mahajan–Nimbhorkar–Varadarajan'12)
- W[2]-hard in general metric (Guha-Khuller'99)

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)

- Euclidean metric k-means:

Approximation Algorithms

- General metric: k-means $\geq$ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

- General metric: k-median $\geq$ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)

- Euclidean metric k-means:
  - Poly time approximation $\approx$ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

- General metric: k-median ≥ 2.67
  (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)

- Euclidean metric k-means:

  - Poly time approximation ≈ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
  - Fixed Dimension: PTAS (Cohen-Addad'18)

Approximation Algorithms

- General metric: k-means ≥ 9 (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
- General metric: k-median ≥ 2.67 (Byrka–Pensyl–Rybicki–Srinivasan–Trinh'17)
- Euclidean metric k-means:
  - Poly time approximation ≈ 6.357 (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
  - Fixed Dimension: PTAS (Cohen-Addad'18)
  - Fixed k: PTAS (Kumar–Sabharwal–Sen'10)

Discrete Version:

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_1-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_1-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_∞-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Guruswami-Indyk'03)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_1-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Trevisan'00)
- ℓ_∞-metric: k-means $\ll 1.01$, k-median $\ll 1.01$ (Guruswami-Indyk'03)

Continuous Version:

k-means in Euclidean metric < 1.0013
(Lee-Schmidt-Wright'17)

Discrete Version:

- **General metric**: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll$ ~~1.01,~~ 1.73 k-median $\ll$ ~~1.01~~ 1.27 (Trevisan'00)
- ℓ_1-metric: k-means $\ll$ ~~1.01,~~ 3.94 k-median $\ll$ ~~1.01~~ 1.73 (Trevisan'00)
- ℓ_∞-metric: k-means $\ll$ ~~1.01,~~ 3.94 k-median $\ll$ ~~1.01~~ 1.73 (Guruswami-Indyk'03)

Continuous Version:

k-means in Euclidean metric $<$ ~~1.0013~~ 1.36
(Lee-Schmidt-Wright'17)

Hardness of Approximation

Discrete Version:

- General metric: k-means ≈ 3.94, k-median ≈ 1.74 (Guha-Khuller'99)
- ℓ_2-metric: k-means $\ll$ ~~1.01~~ 1.73 , 1.17, k-median $\ll$ ~~1.01~~ 1.27 , 1.06 (Trevisan'00)
- ℓ_1-metric: k-means $\ll$ ~~1.01~~ 3.94 , 1.56, k-median $\ll$ ~~1.01~~ 1.73 , 1.14 (Trevisan'00)
- ℓ_∞-metric: k-means $\ll$ ~~1.01~~ 3.94, 3.94, k-median $\ll$ ~~1.01~~ 1.73, 1.73 (Guruswami-Indyk'03)

Continuous Version:

k-means in Euclidean metric $<$ ~~1.0013~~ 1.36 , 1.07
(Lee-Schmidt-Wright'17)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Continuous Version

k-means in ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric ≈ 1.36 (JCH), 1.07 (UGC)

A New Embedding Framework to potentially get Strong (tight?) Inapproximability results!

Warm Up: General Metrics

Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

Warm Up: General Metrics

Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

Warm Up: General Metrics

Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

NO: For all (C, σ) we have $\sum_{x \in X} \Delta(x, \sigma(x))^2 \geq (1 + 8/e - \varepsilon) \cdot |X|$

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$

Proof Overview: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Proof Overview: General Metrics

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

Max Coverage:

- Input: Universe and Collection of Subsets $(U, \mathcal{S}, k)$
- Objective: Max Fraction of U covered by k subsets in $\mathcal{S}$

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

Proof Overview: General Metrics

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

$\Downarrow$

Proof Overview: General Metrics

Theorem (Feige'98)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

$\Downarrow$

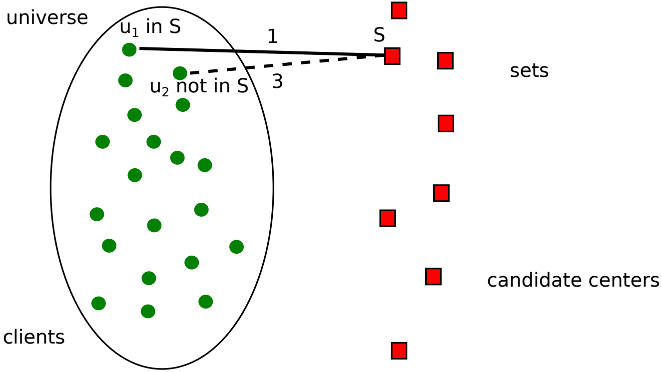
Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

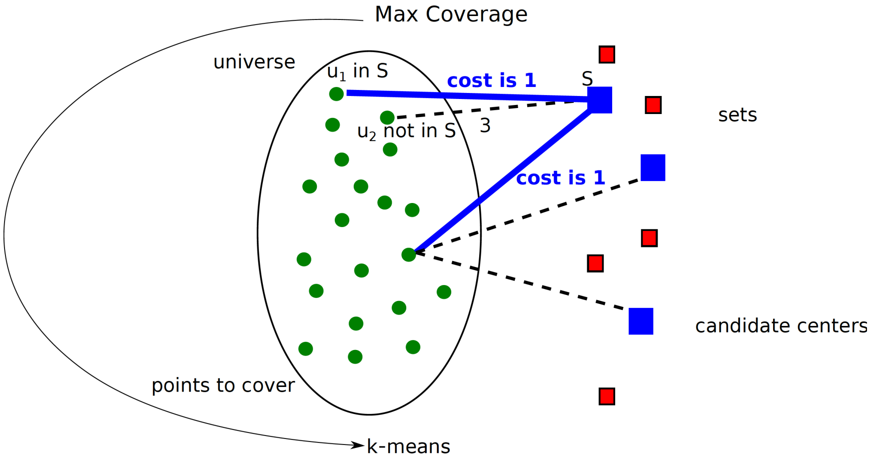
YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

NO: For all (C, σ) we have $\sum_{x \in X} \Delta(x, \sigma(x))^2 \geq (1 + 8/e - \varepsilon) \cdot |X|$

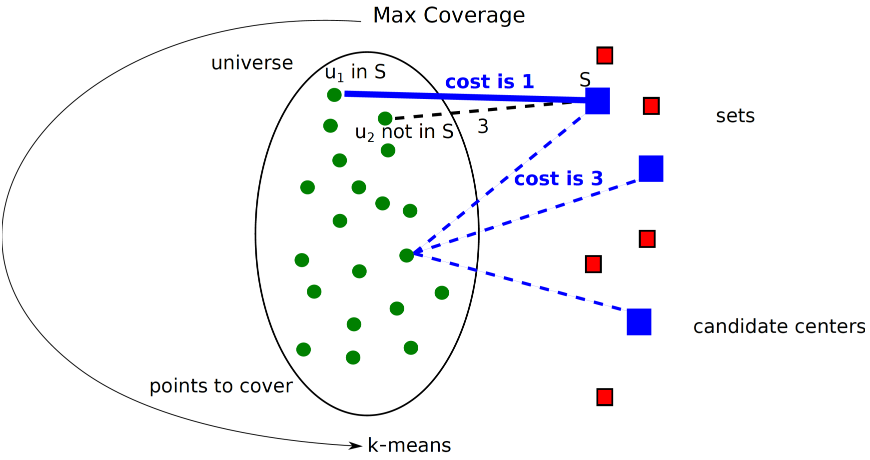
Proof Overview: General Metrics



Proof Overview: General Metrics



Proof Overview: General Metrics



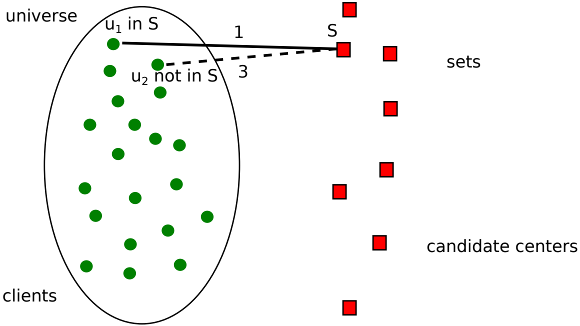
Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

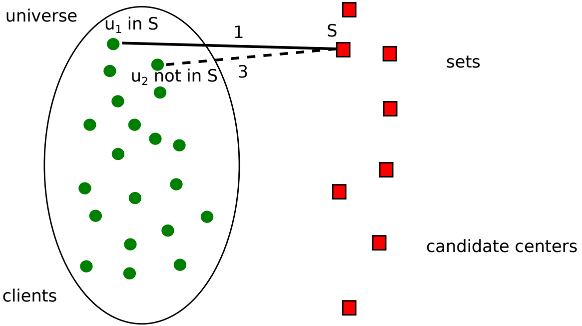
YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

NO: For all (C, σ) we have $\sum_{x \in X} \Delta(x, \sigma(x))^2 \geq (1 + 8/e - \varepsilon) \cdot |X|$

Johnson Coverage Hypothesis



Johnson Coverage Hypothesis



Johnson Coverage Hypothesis (Cohen-Addad–K–Lee)

Fix $\varepsilon > 0$. It is NP-hard to distinguish:

YES: Max Coverage is 1

NO: Max Coverage is at most $1 - 1/e + \varepsilon$

even when set system is induced subgraph of **Johnson graph**.

Johnson Coverage Hypothesis

(α, t)-Johnson Coverage Problem

Given $E \subseteq \binom{[n]}{t}$, and k as input, distinguish between:

Completeness: There exists $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ such that

$$\forall T \in E,\ \exists S_i \in \mathscr{C},\ S_i \subset T.$$

Soundness: For every $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ we have

$$\Pr_{T \sim E}[\exists S_i,\ S_i \subset T] \leq \alpha.$$

Johnson Coverage Hypothesis

(α, t)-Johnson Coverage Problem

Given $E \subseteq \binom{[n]}{t}$, and k as input, distinguish between:

Completeness: There exists $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ such that

$$\forall T \in E,\ \exists S_i \in \mathcal{C},\ S_i \subset T.$$

Soundness: For every $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ we have

$$\Pr_{T \sim E}[\exists S_i,\ S_i \subset T] \leq \alpha.$$

Johnson Coverage Hypothesis (Cohen-Addad–K–Lee)

$\forall \varepsilon > 0, \exists t_\varepsilon \in \mathbb{N}$ such that $(1 - \frac{1}{e} + \varepsilon, t_\varepsilon)$-Johnson Coverage problem is NP-hard.

Johnson Coverage Hypothesis: What can we show?

- $t = 2$: Vertex Coverage problem

Johnson Coverage Hypothesis: What can we show?

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!

Johnson Coverage Hypothesis: What can we show?

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!

- 3-Hypergraph Vertex Coverage problem is NP-Hard to approximate to a factor of $7/8$

Proof Framework

3 ingredients

3 ingredients

- JCH instance

3 ingredients

- JCH instance
- Dimensionality reduction for all ℓ_p-metrics
  - Works only for JCH instances
  - Arises from transcript of a communication game

Proof Framework

3 ingredients

- JCH instance
- Dimensionality reduction for all ℓ_p-metrics
  - Works only for JCH instances
  - Arises from transcript of a communication game
- Johnson Graph Embedding into ℓ_p-metrics

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94^*	1.73^*

Continuous Version

k-means in ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric ≈ 1.36 (JCH), 1.07 (UGC)

Continuous k-means and k-median

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \le n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \ge 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \le n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \ge 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

- Constant Bicriteria inapproximability

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

- Constant Bicriteria inapproximability
- Assuming UGC, hardness for $k = 2$!

Continuous k-means and k-median

Theorem (Cohen-Addad–K–Lee'21)

Given input $X \subseteq \mathbb{R}^{O(n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_\infty{}^2 \geq 4 \cdot n'$.

- k-median: 2 inapproximability

Continuous is harder than Discrete!

- Constant Bicriteria inapproximability
- Assuming UGC, hardness for $k = 2$!
- Dependency on d, k, and ℓ_∞ tight

Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$

Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot{\cup} X_2 \dot{\cup} \cdots \dot{\cup} X_k$ that minimizes:

Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ that minimizes:

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y)$$

Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot{\cup} X_2 \dot{\cup} \cdots \dot{\cup} X_k$ that minimizes:

$$\sum_{i \in [k]} \sum_{x, y \in X_i} \Delta(x, y)$$

- Approximation: $O(\log n)$ [Behsaz et al.'15]

Minsum (Definition)

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ that minimizes:

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y)$$

- Approximation: $O(\log n)$ [Behsaz et al.'15]
- Hardness: $1 + \varepsilon$ [Guruswami-Indyk'03]

Minsum (Result)

Theorem (Cohen-Addad–K–Lee'21)

Given input (X, k), it is NP-hard to distinguish:

Minsum (Result)

Theorem (Cohen-Addad–K–Lee'21)

Given input (X, k), it is NP-hard to distinguish:

YES: There exists a partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ such that

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \leq n',$$

Minsum (Result)

Theorem (Cohen-Addad–K–Lee'21)

Given input (X, k), it is NP-hard to distinguish:

YES: There exists a partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ such that

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \le n',$$

NO: For every partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ we have

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \ge 1.41 \cdot n'.$$

Minsum (Result)

Theorem (Cohen-Addad–K–Lee'21)

Given input (X, k), it is NP-hard to distinguish:

YES: There exists a partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ such that

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \leq n',$$

NO: For every partition $X := X_1 \dot\cup X_2 \dot\cup \cdots \dot\cup X_k$ we have

$$\sum_{i \in [k]} \sum_{x,y \in X_i} \Delta(x, y) \geq 1.41 \cdot n'.$$

Key Ingredient: Hard Instances of Max-Coverage with large girth

Key Takeaways

Improved Inapproximability of

Key Takeaways

Improved Inapproximability of

- k-means and k-median in ℓ_p-metric using JCH framework

Key Takeaways

Improved Inapproximability of

- k-means and k-median in ℓ_p-metric using JCH framework
- Continuous versions of k-means and k-median in General metric

Key Takeaways

Improved Inapproximability of

- k-means and k-median in ℓ_p-metric using JCH framework
- Continuous versions of k-means and k-median in General metric
- k-minsum in General metric

THANK
YOU!