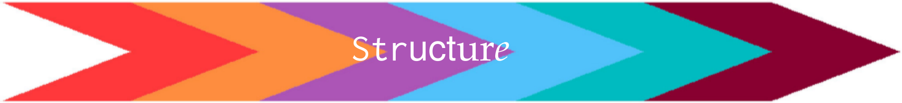
Hardness of Approximation for Metric Clustering

Karthik C. S.
(New York University)

June 22nd 2021

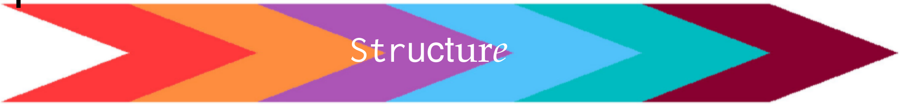
Spectrum of Computational Problems



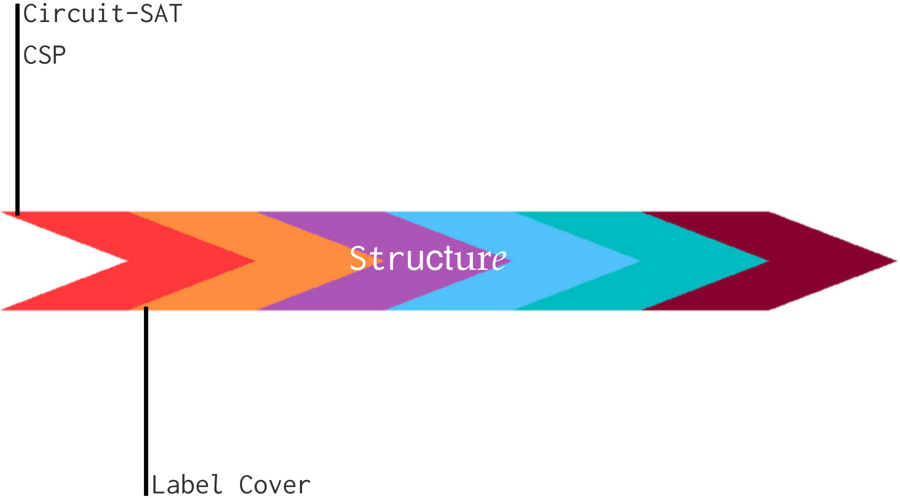
Spectrum of Computational Problems

Circuit-SAT

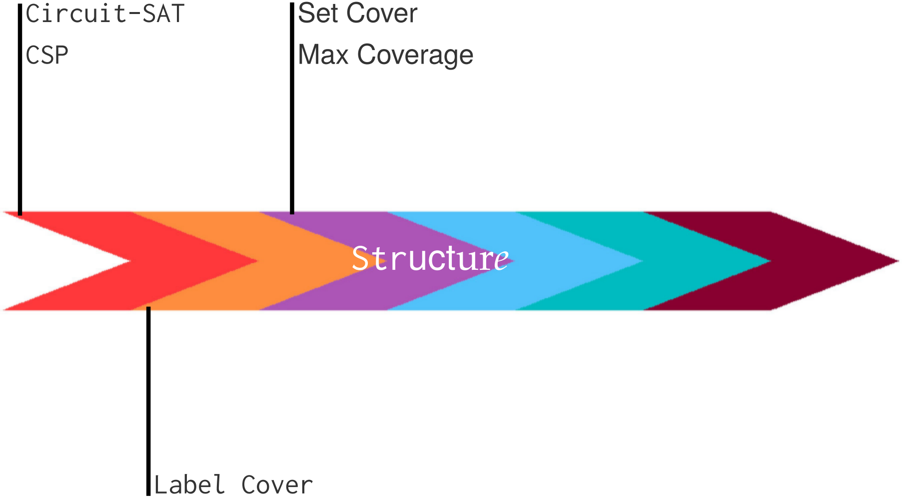
CSP



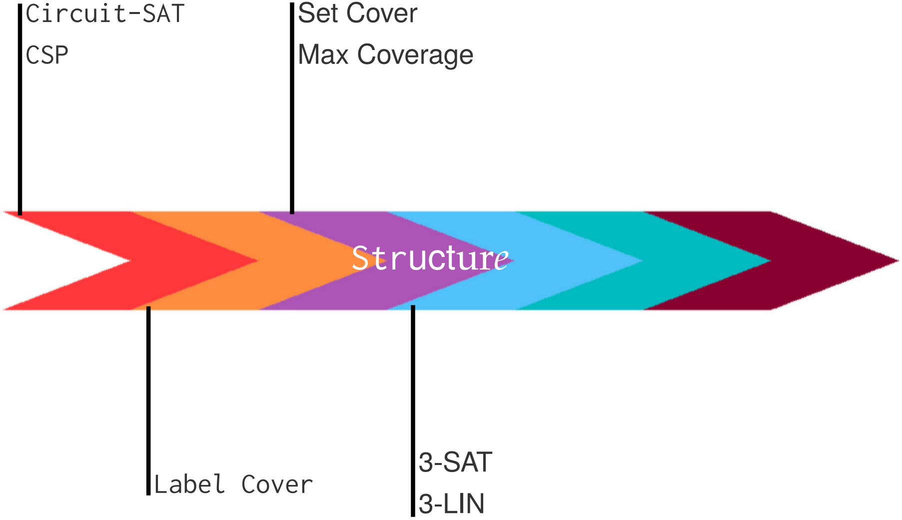
Spectrum of Computational Problems



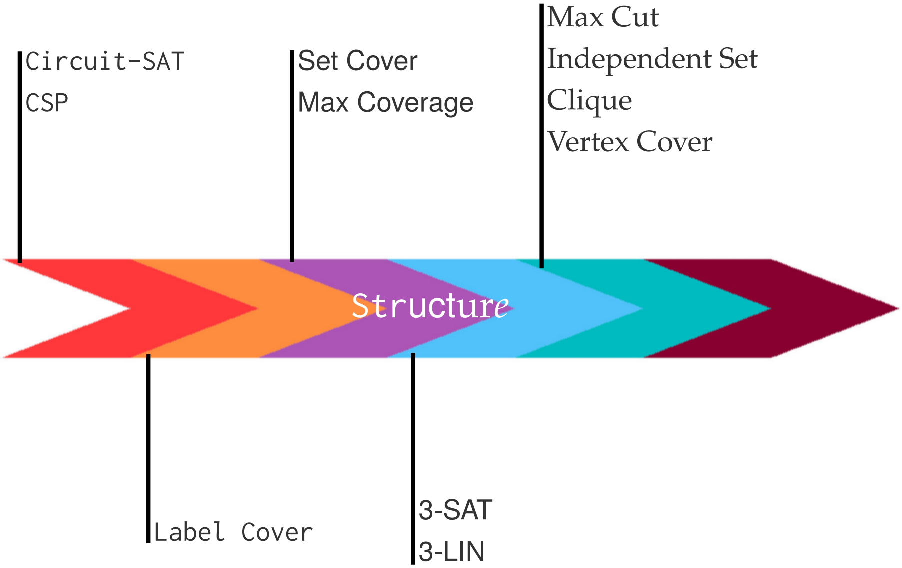
Spectrum of Computational Problems



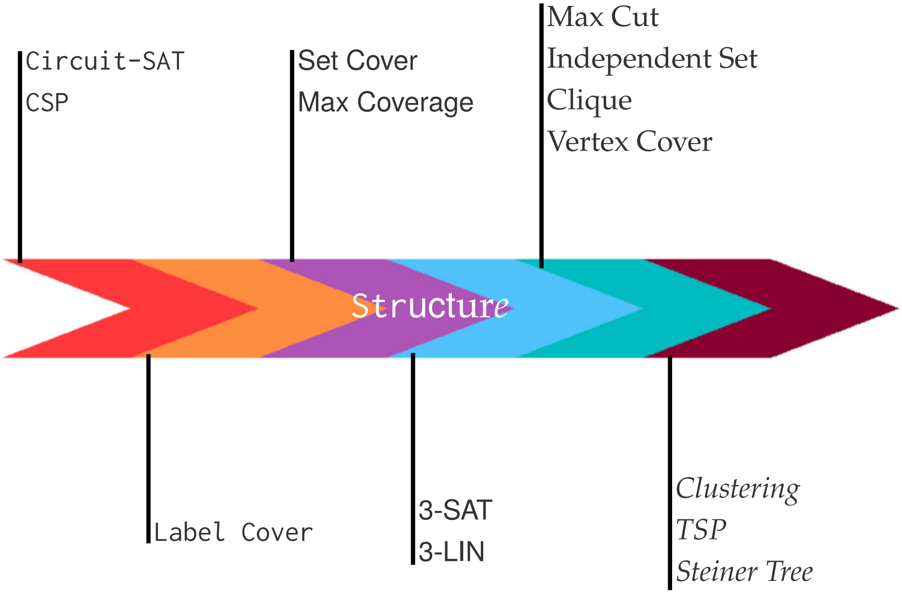
Spectrum of Computational Problems



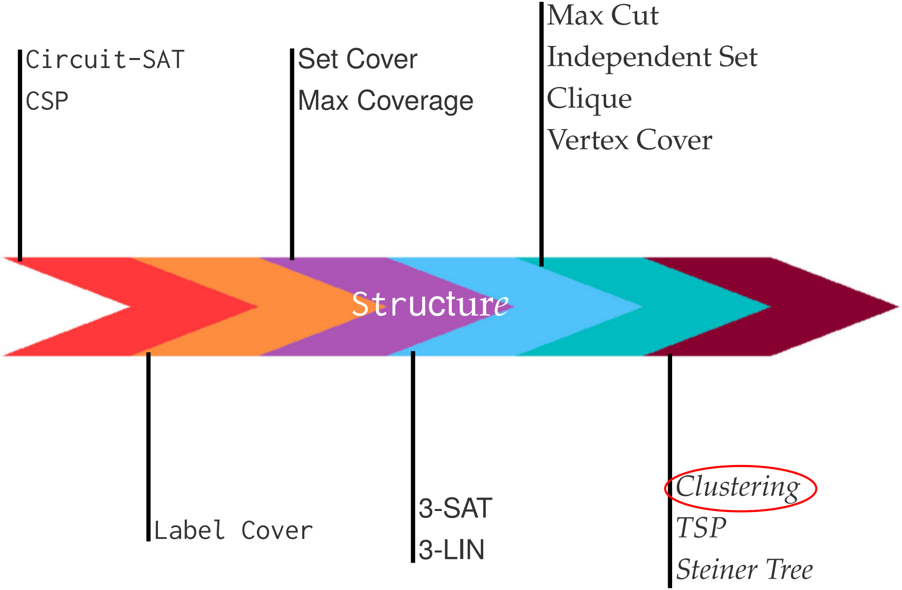
Spectrum of Computational Problems



Spectrum of Computational Problems

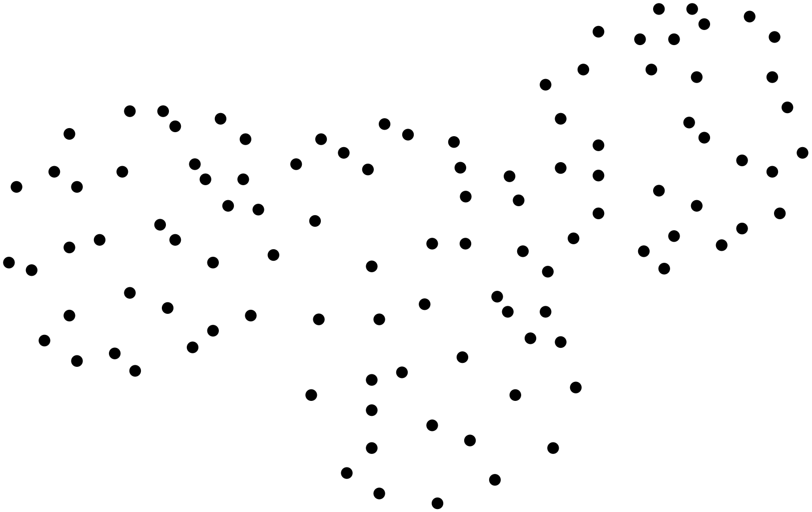


Spectrum of Computational Problems

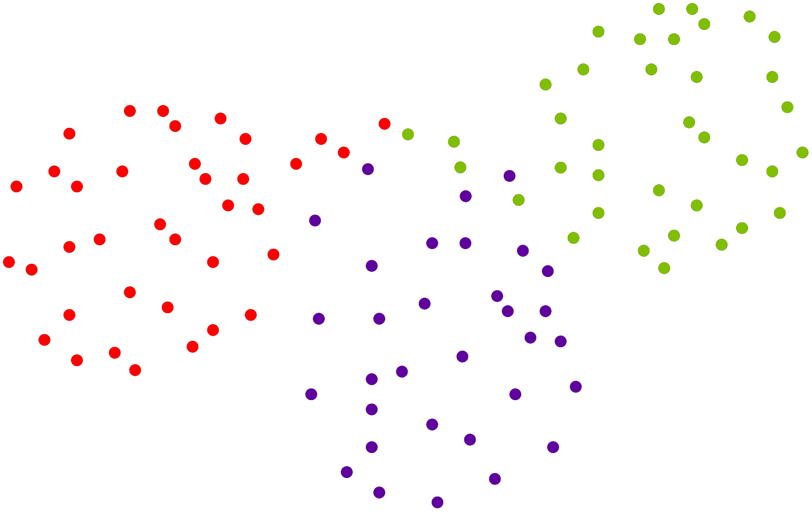


What is Clustering?

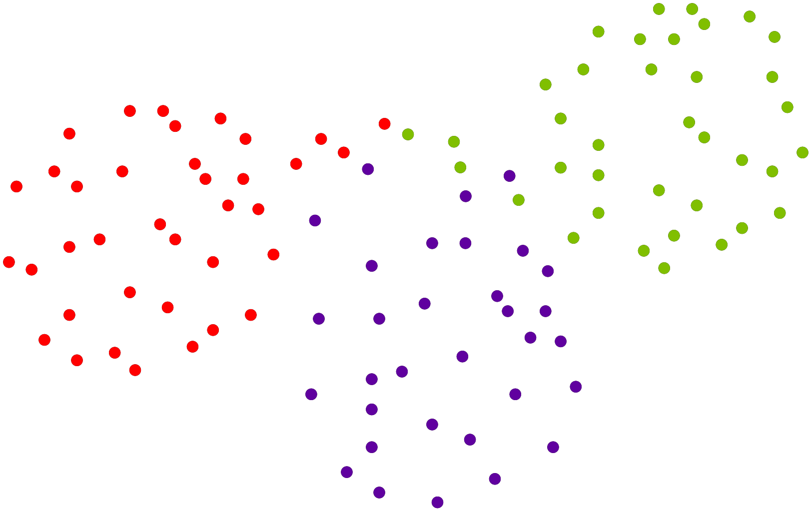
What is Clustering?



What is Clustering?



What is Clustering?



Task of Classifying Input Data

What is Clustering?

- (Γ, Δ) is a metric space

What is Clustering?

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

What is Clustering?

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

What is Clustering?

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

What is Clustering?

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

What is Clustering?

- (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):
  - $C \subseteq \cancel{X}^{\mathcal{S}}$ and $|C| = k$
  - $\sigma : X \to C$
  - σ is *good*

What is Good Classification?

- k-means, k-median, k-center, min-sum, etc.

What is Good Classification?

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

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

- k-means value of (C, σ)

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

Clustering Problem for objective Λ

What is Good Classification?

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

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

- k-means value of (C, σ)

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

Clustering Problem for objective Λ

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

What is Good Classification?

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

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

- k-means value of (C, σ)

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

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 $\Lambda(X, \sigma) > (1 + \delta) \cdot \beta$

What is Good Classification?

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

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

- k-means value of (C, σ)

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

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 $\Lambda(X, \sigma) > (1 + \delta) \cdot \beta$

Exact Computation

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

Exact Computation

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

Exact Computation

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

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

Approximation Algorithms

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

Approximation Algorithms

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

- Euclidean metric k-means:

  - Poly time approximation ≈ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)

Approximation Algorithms

- General metric: k-means ≥ 9
  (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
- Euclidean metric k-means:
  - Poly time approximation ≈ 6.357
    (Ahmadian–Norouzi-Fard–Svensson–Ward'17)
  - Fixed Dimension: PTAS (Cohen-Addad'18)

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

Hardness of Approximation: Before 2019

Discrete Version:

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

Hardness of Approximation: Before 2019

Discrete Version:

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

Hardness of Approximation: Before 2019

Discrete Version:

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

Continuous Version:

- General metric: k-means ≈ 2.47 (Guha-Khuller'99)
- ℓ_2-metric: k-means < 1.0013 (Lee-Schmidt-Wright'17)

Hardness of Approximation: Before 2019

Discrete Version:

I believe is tight!

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

Continuous Version:

- General metric: k-means ≈ 2.47 (Guha-Khuller'99)
- ℓ_2-metric: k-means < 1.0013 (Lee-Schmidt-Wright'17)

Hardness of Approximation: Before 2019

I believe is tight!

Discrete Version:

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

Continuous Version:

- General metric: k-means ≈ 2.47 (Guha-Khuller'99)
- ℓ_2-metric: k-means < 1.0013 (Lee-Schmidt-Wright'17)

Continuous is Computationally Easier than Discrete?

Hardness of Approximation of Metric Clustering since 2019

- Inapproximability of Clustering in ℓ_p-metrics under UGC (Cohen-Addad-K'19)

Hardness of Approximation of Metric Clustering since 2019

- Inapproximability of Clustering in ℓ_p-metrics under UGC (Cohen-Addad-K'19)

- Inapproximability of Continuous k-means, Continuous k-median, and k-minsum in General Metric (Cohen-Addad-K-Lee'21)

Hardness of Approximation of Metric Clustering since 2019

- Inapproximability of Clustering in ℓ_p-metrics under UGC (Cohen-Addad-K'19)

- Inapproximability of Continuous k-means, Continuous k-median, and k-minsum in General Metric (Cohen-Addad-K-Lee'21)

- Tight Inapproximability of Clustering in ℓ_p-metrics under JCH and NP≠P (Cohen-Addad-K-Lee'22?)

State-of-the-art for k-means

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	3.94	1.56	1.38
ℓ_2-metric	1.73	1.17	1.17
ℓ_∞-metric	3.94	3.94	3.94

State-of-the-art for k-means

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	3.94	1.56	1.38
ℓ_2-metric	1.73	1.17	1.17
ℓ_∞-metric	3.94	3.94	3.94

Continuous Version

General metric ≈ 4 (NP≠P)

ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC), 1.06 (NP≠P)

ℓ_1-metric ≈ 2.10 (JCH), 1.16 (NP≠P)

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 $\sum_{x \in X} \Delta(x, \sigma(x))^2 \geq (1 + 8/e - \varepsilon) \cdot |X|$

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 (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

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:

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 $1 - 1/e + \varepsilon$

$$\Downarrow$$

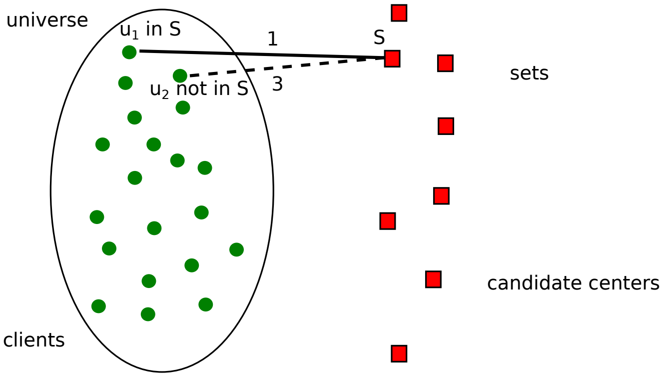
Theorem (Guha-Khuller'99)

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

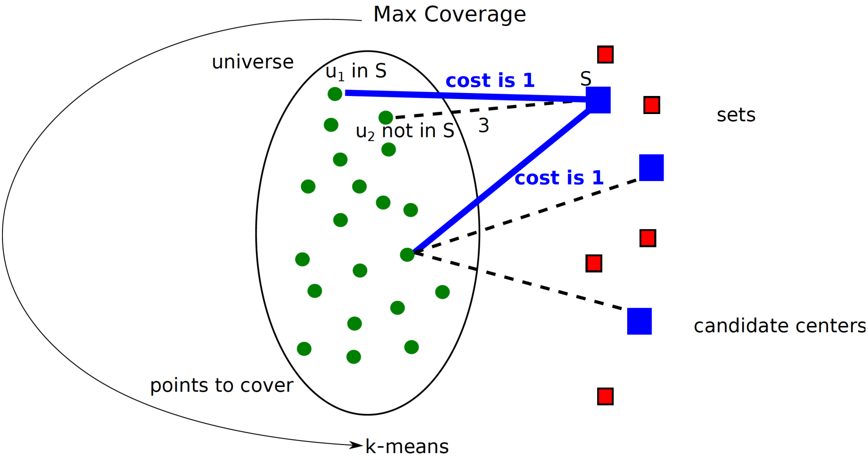
YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

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

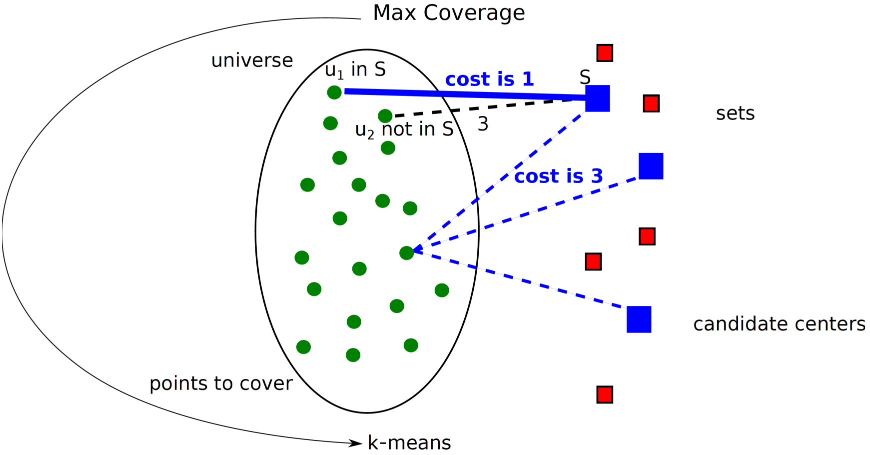
Proof Overview: General Metrics



Proof Overview: General Metrics



Proof Overview: General Metrics



Warm Up: General Metrics

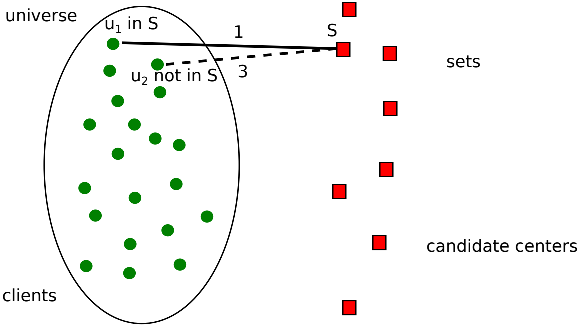
Theorem (Guha-Khuller'99)

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

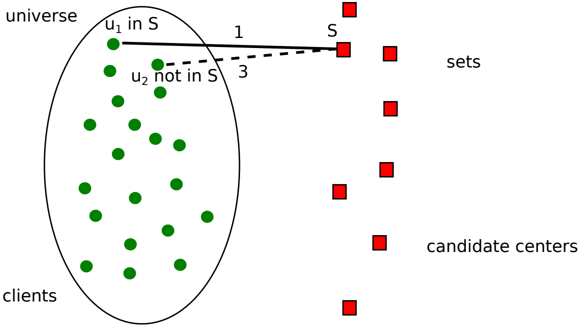
YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

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

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

(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 (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

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

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

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming $(1\text{-}\frac{1}{e}, t)$-Johnson coverage problem is NP-hard, given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Theorem (Cohen-Addad–K–Lee)

Assuming $(1\text{-}\frac{1}{e}, t)$-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq \left(1 + \frac{8}{e}\right) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee)

Assuming (0.93,2)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Theorem (Cohen-Addad–K–Lee)

Assuming (0.93,2)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

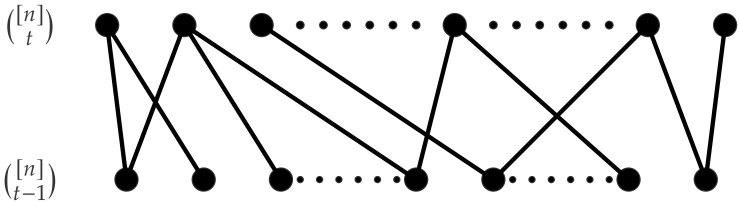
YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

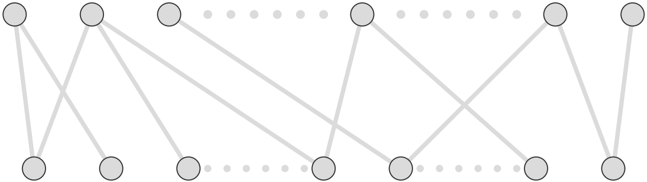
NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq 1.56 \cdot n'.$$

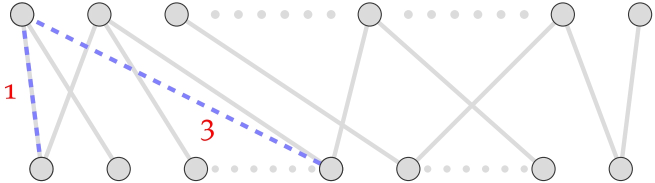
Johnson Graph Embedding



Points in $\{0, 1\}^d$



Points in $\{0, 1\}^d$



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

State-of-the-art for k-means

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	3.94	1.56	1.38
ℓ_2-metric	1.73	1.17	1.17
ℓ_∞-metric	3.94	3.94	3.94

Continuous Version

General metric ≈ 4 (NP≠P)

ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC), 1.06 (NP≠P)

ℓ_1-metric ≈ 2.10 (JCH), 1.16 (NP≠P)

State-of-the-art for k-median

State-of-the-art for k-median

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	1.73	1.14	1.12
ℓ_2-metric	1.27	1.07	1.07
ℓ_∞-metric	1.73	1.73	1.73

State-of-the-art for k-median

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	1.73	1.14	1.12
ℓ_2-metric	1.27	1.07	1.07
ℓ_∞-metric	1.73	1.73	1.73

Continuous Version

General metric ≈ 2 (NP≠P)

ℓ_2-metric ≈ 1.08 (JCH*), 1.015 (NP≠P)

ℓ_1-metric ≈ 1.36 (JCH*), 1.07 (UGC), 1.06 (NP≠P)

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 $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

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 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 $\sum_{x \in X} \|(x - \sigma^*(x)\|_\infty{}^2 \leq n'$,

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

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:

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

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:

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

- 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:

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

- 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

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

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

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

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!

OPEN PROBLEMS

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 $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ 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 (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: Discussion

- $t = 2$: Vertex Coverage problem

Johnson Coverage Hypothesis: Discussion

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

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - $\approx$0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

  - Hypergraph Turán number: Open since 1940s!

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

  - Hypergraph Turán number: Open since 1940s!
  - Recently resolved for $t = 3$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

  - Hypergraph Turán number: Open since 1940s!
  - Recently resolved for $t = 3$

Is JCH true?

State-of-the-art for k-means

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	3.94	1.56	1.38
ℓ_2-metric	1.73	1.17	1.17
ℓ_∞-metric	3.94	3.94	3.94

Continuous Version

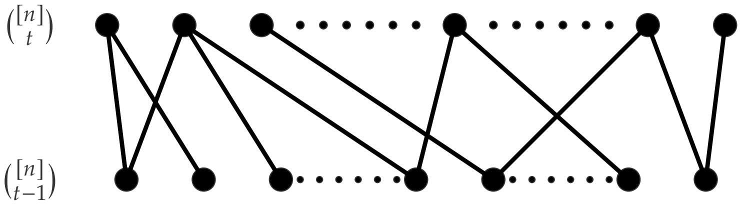
General metric ≈ 4 (NP≠P)

ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC), 1.06 (NP≠P)

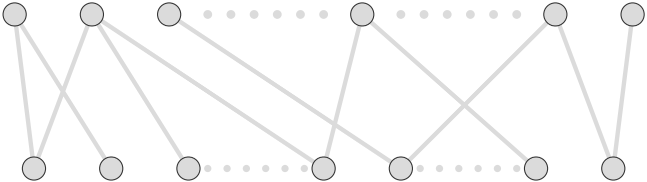
ℓ_1-metric ≈ 2.10 (JCH), 1.16 (NP≠P)

ℓ_∞-metric $\approx$???

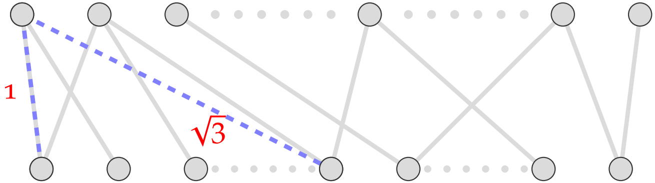
Inapproximability of Clustering in Euclidean metrics



Points in $\{0,1\}^d$



Points in $\{0,1\}^d$



Other Open Problems

◎ k-minsum in ℓ_p-metrics

Other Open Problems

- k-minsum in ℓ_p-metrics
- Capacitated Clustering

Other Open Problems

- k-minsum in ℓ_p-metrics
- Capacitated Clustering
- Fair Clustering

THANK
YOU!