Inapproximability of Clustering in ℓ_p-metrics

Karthik C. S.

(Weizmann Institute of Science)

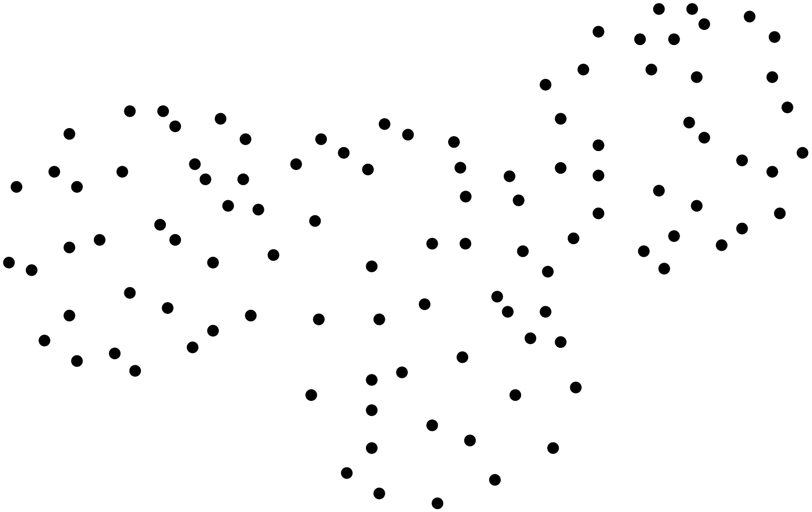
Joint work with

Vincent Cohen-Addad

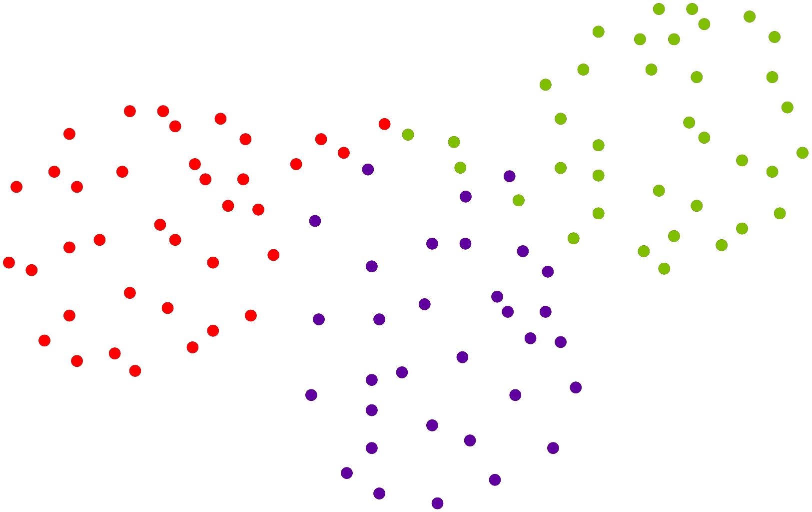
(Sorbonne Université)

What is Clustering?

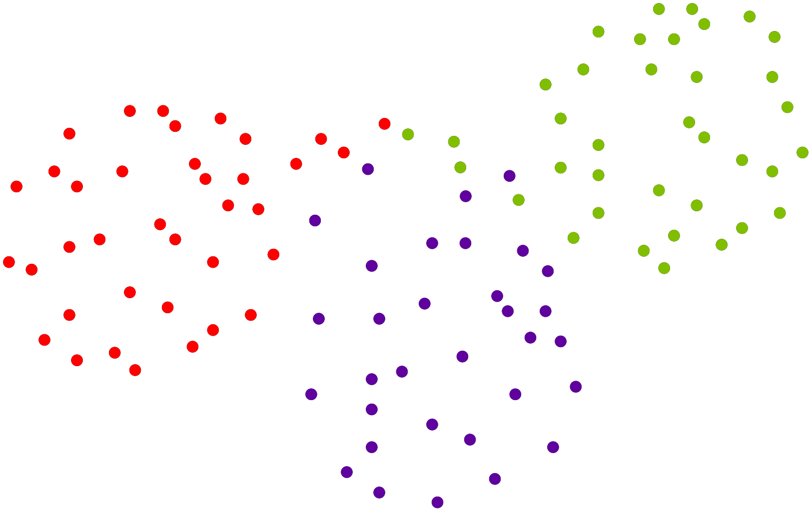
What is Clustering?



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: $X \subseteq \Gamma$

What is Clustering?

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$
- Output: A classification (C, σ):

What is Clustering?

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

What is Clustering?

- (Γ, Δ) is a metric space
- Input: $X \subseteq \Gamma$
- 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: $X \subseteq \Gamma$
- 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: $X \subseteq \Gamma$
- 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: $X \subseteq \Gamma$
- 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: $X \subseteq \Gamma$ and $\mathcal{S} \subseteq \Gamma$
- Output: A classification (C, σ):
  - $C \subseteq \cancel{\Gamma}\ \mathcal{S}$ 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 $\Lambda(X, \sigma) > (1 + \delta) \cdot \beta$

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)

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:

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)

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:

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)

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)

Hardness of Approximation

Discrete Version:

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

Continuous Version:

k-means in Euclidean metric $< \overset{1.07}{\cancel{1.0013}}$
(Lee-Schmidt-Wright'17)

Our Results (Cohen-Addad–K'19)

Discrete Version

	k-means	k-median
ℓ_1-metric	1.56	1.14
ℓ_2-metric	1.17	1.06
ℓ_∞-metric	3.94	1.74

Our Results (Cohen-Addad–K'19)

Discrete Version

	k-means	k-median
ℓ_1-metric	1.56	1.14
ℓ_2-metric	1.17	1.06
ℓ_∞-metric	3.94	1.74

Continuous Version

k-means in ℓ_2-metric ≈ 1.07
k-median in ℓ_1-metric ≈ 1.07

Our Results (Cohen-Addad–K'19)

Discrete Version

	k-means	k-median
ℓ_1-metric	1.56	1.14
ℓ_2-metric	1.17	1.06
ℓ_∞-metric	3.94	1.74

Continuous Version

k-means in ℓ_2-metric ≈ 1.07
k-median in ℓ_1-metric ≈ 1.07

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

Warm up: General Metrics

Warm up: General Metrics

Max Coverage:

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

Max Coverage:

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

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:

Warm up: 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

Warm up: 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 $1 - {}^1\!/_e + \varepsilon$

Warm up: 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 $1 - 1/e + \varepsilon$

$\Downarrow$

Warm up: 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. It is NP-hard to distinguish:

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

Theorem (Cohen-Addad–K'19)

Given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$. It is UG-hard to distinguish:

k-means in Hamming metric

Theorem (Cohen-Addad–K'19)

Given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$. It is UG-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n'$,

k-means in Hamming metric

Theorem (Cohen-Addad–K'19)

Given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$. It is UG-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 $\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq 1.56 \cdot n'$,

where $n' = O(n(\log n)^2)$.

Vertex Coverage

Vertex Coverage:

- Input: Graph (G, k)

Vertex Coverage:

- Input: Graph (G, k)
- Objective: Max Fraction of Edges covered by k Vertices

Vertex Coverage:

- Input: Graph (G, k)
- Objective: Max Fraction of Edges covered by k Vertices

Theorem (Austrin-Khot-Safra'11; Austrin-Stanković'19)

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

Vertex Coverage:

- Input: Graph (G, k)
- Objective: Max Fraction of Edges covered by k Vertices

Theorem (Austrin-Khot-Safra'11; Austrin-Stanković'19)

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

YES: Vertex Coverage is 1

Vertex Coverage:

- Input: Graph (G, k)
- Objective: Max Fraction of Edges covered by k Vertices

Theorem (Austrin-Khot-Safra'11; Austrin-Stanković'19)

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

YES: Vertex Coverage is 1

NO: Vertex Coverage is at most $0.9292 - \varepsilon$

Vertex Coverage

Vertex Coverage:

- Input: Graph (G, k)
- Objective: Max Fraction of Edges covered by k Vertices

Theorem (Austrin-Khot-Safra'11; Austrin-Stanković'19)

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

YES: Vertex Coverage is 1

NO: Vertex Coverage is at most $0.9292 - \varepsilon$

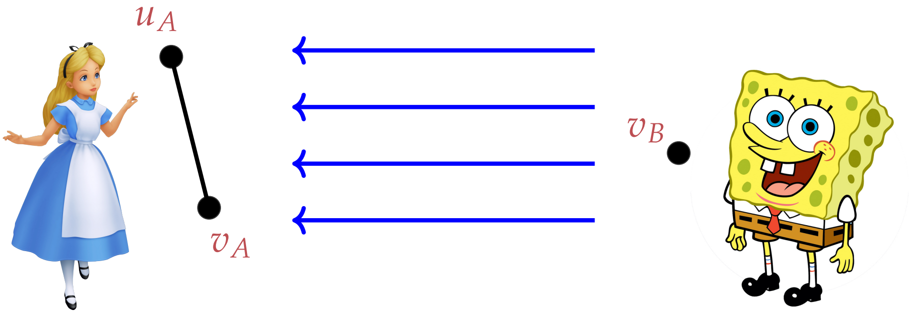
Edges $\rightarrow$ Data Points

Vertices $\rightarrow$ Candidate Centers

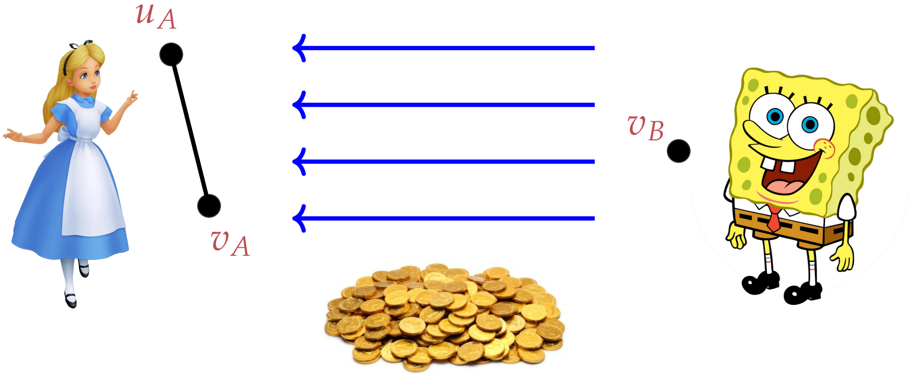
Vertex/Edge Game

Vertex/Edge Game

Vertex/Edge Game

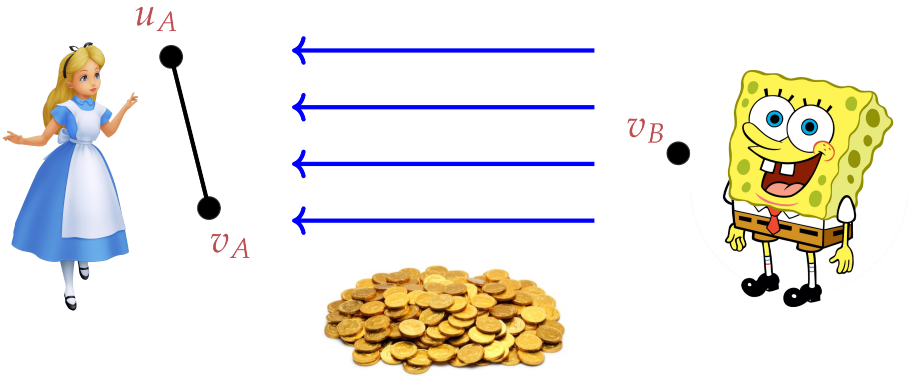


Vertex/Edge Game



Public Randomness

Vertex/Edge Game



Public Randomness

GOAL

$$\boxed{\text{Determine if } v_B \in \{u_A, v_A\}}$$

Vertex/Edge Game: Protocols

- Deterministic Protocol:
  - Message length: $O(\log n)$ bits
  - Completeness: 1, Soundness: 0

Vertex/Edge Game: Protocols

- Deterministic Protocol:
  - Message length: $O(\log n)$ bits
  - Completeness: 1, Soundness: 0
- Randomized Protocol:
  - Message length: $O_\varepsilon(1)$ bits

Vertex/Edge Game: Protocols

- Deterministic Protocol:
  - Message length: $O(\log n)$ bits
  - Completeness: 1, Soundness: 0
- Randomized Protocol:
  - Message length: $O_{\varepsilon}(1)$ bits
  - Completeness: 1, Soundness: ε

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $\mathscr{C}(v_B)_i$

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $\mathscr{C}(v_B)_i$
- Alice checks if $\mathscr{C}(v_B)_i \in \{\mathscr{C}(u_A)_i, \mathscr{C}(v_A)_i\}$

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $\mathscr{C}(v_B)_i$
- Alice checks if $\mathscr{C}(v_B)_i \in \{\mathscr{C}(u_A)_i, \mathscr{C}(v_A)_i\}$
- Message length: $\log_2 q$

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $\mathscr{C}(v_B)_i$
- Alice checks if $\mathscr{C}(v_B)_i \in \{\mathscr{C}(u_A)_i, \mathscr{C}(v_A)_i\}$
- Message length: $\log_2 q$
- Soundness: $1 - O(\Delta(\mathscr{C}))$

Vertex/Edge Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $\mathscr{C}(v_B)_i$
- Alice checks if $\mathscr{C}(v_B)_i \in \{\mathscr{C}(u_A)_i, \mathscr{C}(v_A)_i\}$
- Message length: $\log_2 q$
- Soundness: $1 - O(\Delta(\mathscr{C})) \approx O(1/\sqrt{q})$ (for AG codes)

Embedding Transcript into Hamming metric

- Construct $\tau : V \rightarrow \{0,1\}^{q \cdot c \cdot \log n}$

Embedding Transcript into Hamming metric

- Construct $\tau : V \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$.

Embedding Transcript into Hamming metric

- Construct $\tau : V \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$. For any $t \in [q]$:

$$\tau(v)_{i,t} = 1 \iff \mathscr{C}(v)_i = t$$

Embedding Transcript into Hamming metric

- Construct $\tau : V \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$. For any $t \in [q]$:

$$\tau(v)_{i,t} = 1 \iff \mathscr{C}(v)_i = t$$

- $\mathcal{S} = \{\tau(v) \mid v \in V\}$

Embedding Transcript into Hamming metric

- Construct $\tau : V \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$. For any $t \in [q]$:

$$\tau(v)_{i,t} = 1 \iff \mathscr{C}(v)_i = t$$

- $\mathcal{S} = \{\tau(v) \mid v \in V\}$
- $X = \{\tau(u) \vee \tau(v) \mid (u,v) \in E\}$

Completeness of Reduction

- $V' := \{v_1, \ldots, v_k\} \subseteq V$ be a vertex cover of G

Completeness of Reduction

- V' := $\{v_1, \ldots, v_k\} \subseteq V$ be a vertex cover of G
- Build $\sigma : X \to C \subseteq \mathcal{S}$

Completeness of Reduction

- $V' := \{v_1, \ldots, v_k\} \subseteq V$ be a vertex cover of G
- Build $\sigma : X \to C \subseteq \mathcal{S}$

$$\sigma(x_{u,v}) = \begin{cases} \tau(u) \text{ if } u \in V' \\ \tau(v) \text{ otherwise.} \end{cases}$$

Completeness of Reduction

- $V' := \{v_1, \ldots, v_k\} \subseteq V$ be a vertex cover of G
- Build $\sigma : X \to C \subseteq \mathcal{S}$

$$\sigma(x_{u,v}) = \begin{cases} \tau(u) \text{ if } u \in V' \\ \tau(v) \text{ otherwise.} \end{cases}$$

- Fix $x_{u,v} \in X$ and $i \in [c \cdot \log n]$

Distance between $x_{u,v}$ and $\sigma(x_{u,v})$ on block i is 1

Completeness of Reduction

- $V' := \{v_1, \ldots, v_k\} \subseteq V$ be a vertex cover of G
- Build $\sigma : X \to C \subseteq \mathcal{S}$

$$\sigma(x_{u,v}) = \begin{cases} \tau(u) \text{ if } u \in V' \\ \tau(v) \text{ otherwise.} \end{cases}$$

- Fix $x_{u,v} \in X$ and $i \in [c \cdot \log n]$

Distance between $x_{u,v}$ and $\sigma(x_{u,v})$ on block i is 1

- k-means objective is:

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X|$$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $V' \subseteq V$ of size k:

$$v \in V' \iff \tau(v) \in C$$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $V' \subseteq V$ of size k:

$$v \in V' \iff \tau(v) \in C$$

- $E' \subseteq E$, such that V' does not cover any $e \in E'$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $V' \subseteq V$ of size k:

$$v \in V' \iff \tau(v) \in C$$

- $E' \subseteq E$, such that V' does not cover any $e \in E'$
- Fix $x_{u,v} \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $x_{u,v}$ and $\sigma(x_{u,v})$ on block i is mostly 3

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $V' \subseteq V$ of size k:

$$v \in V' \iff \tau(v) \in C$$

- $E' \subseteq E$, such that V' does not cover any $e \in E'$
- Fix $x_{u,v} \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $x_{u,v}$ and $\sigma(x_{u,v})$ on block i is mostly 3

- k-means objective is:

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X \setminus X_{E'}| + 9 \cdot (c \cdot \log n)^2 \cdot |X_{E'}|$$

Theorem (Cohen-Addad–K'19)

Given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$. It is UG-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 $\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq 1.56 \cdot n'$,

where $n' = O(n(\log n)^2)$.

k-means in Hamming metric: Continuous Case

Theorem (Cohen-Addad–K'19)

Given input $X \subseteq \{0,1\}^{O(\log n)}$. It is UG-hard to distinguish:

k-means in Hamming metric: Continuous Case

Theorem (Cohen-Addad–K'19)

Given input $X \subseteq \{0,1\}^{O(\log n)}$. It is UG-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n'$,

k-means in Hamming metric: Continuous Case

Theorem (Cohen-Addad–K'19)

Given input $X \subseteq \{0,1\}^{O(\log n)}$. It is UG-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \le n'$,

NO: For all (C, σ) we have $\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \ge 1.21 \cdot n'$,

where $n' = O(n(\log n)^2)$.

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u, v) \in E\}$

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u, v) \in E\}$
- Completeness: Choose centers corresponding to vertices

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u, v) \in E\}$
- Completeness: Choose centers corresponding to vertices
- Soundness: $\sigma : X \to C \subseteq \{0, 1\}^{q \cdot c \cdot \log n}$ is some classification

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u,v) \in E\}$
- Completeness: Choose centers corresponding to vertices
- Soundness: $\sigma : X \to C \subseteq \{0,1\}^{q \cdot c \cdot \log n}$ is some classification
- In opt. solution: $\|\sigma(x_{u,v})|_B\|_0 \leq 3$ on every block B

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u, v) \in E\}$
- Completeness: Choose centers corresponding to vertices
- Soundness: $\sigma : X \to C \subseteq \{0, 1\}^{q \cdot c \cdot \log n}$ is some classification
- In opt. solution: $\|\sigma(x_{u,v})|_B\|_0 \leq 3$ on every block B
    - Mostly 3 or 2 $\Rightarrow$ cluster size is small

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u, v) \in E\}$
- Completeness: Choose centers corresponding to vertices
- Soundness: $\sigma : X \to C \subseteq \{0,1\}^{q \cdot c \cdot \log n}$ is some classification
- In opt. solution: $\|\sigma(x_{u,v})|_B\|_0 \leq 3$ on every block B
  - Mostly 3 or 2 $\Rightarrow$ cluster size is small
  - Mostly 0 $\Rightarrow$ pay cost 4 per block

Continuous Case: Analysis

- $X = \{\tau(u) \vee \tau(v) \mid (u, v) \in E\}$
- Completeness: Choose centers corresponding to vertices
- Soundness: $\sigma : X \to C \subseteq \{0,1\}^{q \cdot c \cdot \log n}$ is some classification
- In opt. solution: $\|\sigma(x_{u,v})|_B\|_0 \leq 3$ on every block B
  - Mostly 3 or 2 $\Rightarrow$ cluster size is small
  - Mostly 0 $\Rightarrow$ pay cost 4 per block
  - Mostly 1 $\Rightarrow$ decode vertex

Our Result

Discrete Version

	k-means	k-median
ℓ_1-metric	1.56	1.14
ℓ_2-metric	1.17	1.06
ℓ_∞-metric	3.94	1.74

Continuous Version

k-means in ℓ_2-metric ≈ 1.07

k-median in ℓ_1-metric ≈ 1.07

Other Metrics: More Embedding

Gap Number of ℓ_p-metric

Largest $\alpha > 1$ for which we can realize $V \cup E$ of K_n such that

$$\|u - e\|_p = 1 \text{ if } u \in e \quad \text{and} \quad \|u - e\|_p \geq \alpha \text{ if } u \notin e$$

Other Metrics: More Embedding

Gap Number of ℓ_p-metric

Largest $\alpha > 1$ for which we can realize $V \cup E$ of K_n such that

$$\|u - e\|_p = 1 \text{ if } u \in e \quad \text{and} \quad \|u - e\|_p \geq \alpha \text{ if } u \notin e$$

Replace each block by the embedding realizing gap number

Gap Number of ℓ_p-metrics

- ℓ_0/ℓ_1-metric = 3

Gap Number of ℓ_p-metrics

- ℓ_0/ℓ_1-metric = 3
- ℓ_2-metric > 1.85

Gap Number of ℓ_p-metrics

- ℓ_0/ℓ_1-metric = 3
- ℓ_2-metric > 1.85
- ℓ_∞-metric = 3

Our Result

Discrete Version

	k-means	k-median
ℓ_1-metric	1.56	1.14
ℓ_2-metric	1.17	1.06
ℓ_∞-metric	3.94	1.74

Continuous Version

k-means in ℓ_2-metric ≈ 1.07

k-median in ℓ_1-metric ≈ 1.07

Euclidean k-means: Continuous Case

- k-means cost is sum of all pairwise intra-cluster squared distances

Euclidean k-means: Continuous Case

- k-means cost is sum of all pairwise intra-cluster squared distances
- Look at induced subgraph of each cluster

Euclidean k-means: Continuous Case

- k-means cost is sum of all pairwise intra-cluster squared distances
- Look at induced subgraph of each cluster
  - Adjacent edges squared distance is 2

Euclidean k-means: Continuous Case

- k-means cost is sum of all pairwise intra-cluster squared distances
- Look at induced subgraph of each cluster
  - Adjacent edges squared distance is 2
  - Non-adjacent edges squared distance is 4

Euclidean k-means: Continuous Case

- k-means cost is sum of all pairwise intra-cluster squared distances
- Look at induced subgraph of each cluster
  - Adjacent edges squared distance is 2
  - Non-adjacent edges squared distance is 4
  - Argue that # of edges in cluster $\gg$ max degree of cluster

Our Result

Discrete Version

	k-means	k-median
ℓ_1-metric	1.56	1.14
ℓ_2-metric	1.17	1.06
ℓ_∞-metric	3.94	1.74

Continuous Version

k-means in ℓ_2-metric ≈ 1.07
k-median in ℓ_1-metric ≈ 1.07

Stronger Inapproximability in ℓ_∞-metric

Two ingredients:

Stronger Inapproximability in ℓ_∞-metric

Two ingredients:

Theorem (Essentially Feige'98)

For every $\delta > 0$ there is some $h \in \mathbb{N}$ such that deciding an instance of $(1 - 1/e + \varepsilon)$-hypergraph vertex coverage problem on h-uniform hypergraphs is NP-hard.

Stronger Inapproximability in ℓ_∞-metric

Two ingredients:

Theorem (Essentially Feige'98)

For every $\delta > 0$ there is some $h \in \mathbb{N}$ such that deciding an instance of $(1 - 1/e + \varepsilon)$-hypergraph vertex coverage problem on h-uniform hypergraphs is NP-hard.

Gap hypergraph number in ℓ_∞-metric is 3

Key Takeaways

- Improved Inapproximability of

Key Takeaways

- Improved Inapproximability of
- k-means and k-median

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Membership Protocol

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Membership Protocol
- And Geometric Realization of Complete Graphs

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Membership Protocol
- And Geometric Realization of Complete Graphs
- And Complete Hypergraphs

Open Problem 1

Can we embed vertices and hyperedges
of h-uniform complete hypergraph in Hamming metric
with gap number 3?

Open Problem 1

Can we embed vertices and hyperedges of h-uniform complete hypergraph in Hamming metric with gap number 3?

- Current Reduction gives gap number $1 + {}^{2}/_{(h-1)}$

Open Problem 1

Can we embed vertices and hyperedges
of h-uniform complete hypergraph in Hamming metric
with gap number 3?

- Current Reduction gives gap number $1 + 2/(h-1)$
- Dimension of embedding doesn't matter for ℓ_2-metric
  - Johnson-Lindenstrauss dimension reduction

Open Problem 2

Can we embed vertices and edges of K_n
in Euclidean metric with gap number 2?

Open Problem 2

Can we embed vertices and edges of K_n
in Euclidean metric with gap number 2?

- It holds for $n = 3$

Open Problem 2

Can we embed vertices and edges of K_n
in Euclidean metric with gap number 2?

- It holds for $n = 3$
- Can we prove an upper bound of 2?

Open Problem 3

Can we go beyond Triangle Inequality Barrier?

Open Problem 3

Can we go beyond Triangle Inequality Barrier?

- Can we show $>1 + 8/e$ inapproximability of k-means in any metric?
- Can we show $>1 + 2/e$ inapproximability of k-median in any metric?

THANK
YOU!