Clustering
How hard is it to Classify Data?

Karthik C. S.
(Tel Aviv University)

Joint work(s) with

Vincent Cohen-Addad
(Sorbonne University)

Euiwoong Lee
(New York University)
Spectrum of Computational Problems

Structure
Spectrum of Computational Problems

- Circuit-SAT
- CSP
- Set Cover
- Max Coverage
- Label Cover

Structure
Spectrum of Computational Problems

- Circuit-SAT
- CSP
- Set Cover
- Max Coverage
- Label Cover
- 3-SAT
- 3-LIN
Spectrum of Computational Problems

- Circuit-SAT
- CSP
- Set Cover
- Max Coverage
- 3-SAT
- 3-LIN
- Label Cover
- Max Cut
- Independent Set
- Clique
- Vertex Cover
- Clustering
- TSP
- Steiner Tree
- UGC
- JCH
Spectrum of Computational Problems

- Circuit-SAT
- CSP
- Label Cover
- Set Cover
- Max Coverage
- 3-SAT
- 3-LIN
- Max Cut
- Independent Set
- Clique
- Vertex Cover
- Clustering
- TSP
- Steiner Tree
Spectrum of Computational Problems

- Circuit-SAT
- CSP
- Set Cover
- Max Coverage
- UGC
- Label Cover
- Structure
- Max Cut
- Independent Set
- Clique
- Vertex Cover
- 3-SAT
- 3-LIN
- Clustering
- TSP
- Steiner Tree
Spectrum of Computational Problems

Circuit-SAT
CSP

Set Cover
Max Coverage

Max Cut
Independent Set
Clique
Vertex Cover

Structure

UGC

Label Cover

3-SAT
3-LIN

Clustering
TSP
Steiner Tree
Spectrum of Computational Problems

Circuit-SAT
CSP

Set Cover
Max Coverage

Max Cut
Independent Set
Clique
Vertex Cover

UGC

Label Cover

JCH

Structure

3-SAT
3-LIN

Clustering
TSP
Steiner Tree
Spectrum of Computational Problems

- Circuit-SAT
- CSP
- Set Cover
- Max Coverage
- Max Cut
- Independent Set
- Clique
- Vertex Coverage
- Label Cover
- 3-SAT
- 3-LIN
- 3-LIN
- UGC
- AS'19
- JCH
- Clustering
- TSP
- Steiner Tree
What is Clustering?
What is Clustering?

Task of Classifying Input Data
What is Clustering?

Task of Classifying Input Data
What is Clustering?

Task of Classifying Input Data
What is Clustering?

○ \((\Gamma, \Delta)\) is a metric space
What is Clustering?

- $(\Gamma, \Delta)$ is a metric space
- Input: $X \subseteq \Gamma, k \in \mathbb{N}$
What is Clustering?

- \((\Gamma, \Delta)\) is a metric space
- Input: \(X \subseteq \Gamma, k \in \mathbb{N}\)
- Output: A classification \((C, \sigma)\):
What is Clustering?

- $(\Gamma, \Delta)$ is a metric space
- **Input:** $X \subseteq \Gamma$, $k \in \mathbb{N}$
- **Output:** A classification $(C, \sigma)$:
  - $C \subseteq \Gamma$ and $|C| = k$
What is Clustering?

- $(\Gamma, \Delta)$ is a metric space
- Input: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- Output: A classification $(C, \sigma)$:
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \rightarrow C$
What is Clustering?

- $(\Gamma, \Delta)$ is a metric space
- **Input**: $X \subseteq \Gamma$, $k \in \mathbb{N}$
- **Output**: A classification $(C, \sigma)$:
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \rightarrow C$
  - $\sigma$ is *good*
What is Clustering?

Continuous Version

- $(\Gamma, \Delta)$ is a metric space
- **Input:** $X \subseteq \Gamma, \ k \in \mathbb{N}$
- **Output:** A classification $(C, \sigma)$:
  - $C \subseteq \Gamma$ and $|C| = k$
  - $\sigma : X \rightarrow C$
  - $\sigma$ is *good*
What is Clustering?

**(Γ, Δ)** is a metric space

**Input:** \( X \subseteq \Gamma, \ k \in \mathbb{N} \)

**Output:** A classification \((C, \sigma)\):

- \( C \subseteq \Gamma \) and |C| = \( k \)
- \( \sigma : X \rightarrow C \)
- \( \sigma \) is *good*
What is Clustering?

**Discrete Continuous Version**

- $(\Gamma, \Delta)$ is a metric space

- **Input**: $X \subseteq \Gamma$, $k \in \mathbb{N}$ and $S \subseteq \Gamma$

- **Output**: A classification $(C, \sigma)$:
  - $C \subseteq X$ and $|C| = k$
  - $\sigma : X \rightarrow C$
  - $\sigma$ 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 $(C, \sigma)$

$$\sum_{x \in X} \Delta(x, \sigma(x))$$
What is Good Classification?

- \( k \)-means, \( k \)-median, \( k \)-center, min-sum, etc.

- **\( k \)-median** value of \((C, \sigma)\)

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

- **\( k \)-means** value of \((C, \sigma)\)

\[
\sum_{x \in X} \Delta(x, \sigma(x))^2
\]
What is Good Classification?

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

- **k-median** value of \((C, \sigma)\)
  \[
  \sum_{x \in X} \Delta(x, \sigma(x))
  \]

- **k-means** value of \((C, \sigma)\)
  \[
  \sum_{x \in X} \Delta(x, \sigma(x))^2
  \]

Clustering Problem for objective \(\Lambda\)
What is Good Classification?

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

- $k$-median value of $(C, \sigma)$

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

- $k$-means value of $(C, \sigma)$

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

Clustering Problem for objective $\Lambda$

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

○ \(k\)-means, \(k\)-median, \(k\)-center, min-sum, etc.

○ \(k\)-median value of \((C, \sigma)\)

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

○ \(k\)-means value of \((C, \sigma)\)

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

Clustering Problem for objective \(\Lambda\)

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

No: For all classification \((C, \sigma)\), 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)
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 $\geq 9$
(Ahmadian–Norouzi-Fard–Svensson–Ward’17)
○ **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)
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:
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 $\geq 9$
  (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)
  ○ Fixed **Dimension**: PTAS (Cohen-Addad’18)
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)
  
  - Fixed **Dimension**: PTAS (Cohen-Addad’18)
  
  - Fixed $k$: PTAS (Kumar–Sabharwal–Sen’10)
Hardness of Approximation

**Discrete Version:**

- General metric: $k$-means $\approx \frac{1}{3}$, $k$-median $\approx \frac{1}{7}$ (Guha-Khuller
- $\ell_2$-metric: $k$-means $\ll \frac{1}{3}$, $k$-median $\ll \frac{1}{7}$ (Trevisan
- $\ell_1$-metric: $k$-means $\ll \frac{1}{3}$, $k$-median $\ll \frac{1}{7}$ (Trevisan
- $\ell_\infty$-metric: $k$-means $\ll \frac{1}{3}$, $k$-median $\ll \frac{1}{7}$ (Guruswami-Indyk

**Continuous Version:**

$k$-means in Euclidean metric $< \frac{1}{3}$ (Lee-Schmidt-Wright
Hardness of Approximation

Discrete Version:

- **General metric:** $k$-means $\approx 3.94$, $k$-median $\approx 1.74$
  (Guha-Khuller’99)
Hardness of Approximation

**Discrete Version:**

- **General metric:** $k$-means $\approx 3.94$, $k$-median $\approx 1.74$
  (Guha-Khuller’99)
- **$\ell_2$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_1$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
Hardness of Approximation

**Discrete Version:**

- **General metric:** $k$-means $\approx 3.94$, $k$-median $\approx 1.74$
  (Guha-Khuller’99)
- **$\ell_2$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_1$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_\infty$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Guruswami-Indyk’03)
Discrete Version:

- **General metric:** $k$-means $\approx 3.94$, $k$-median $\approx 1.74$
  (Guha-Khuller’99)
- **$\ell_2$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_1$-metric:** $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_\infty$-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 $\approx 3.94$, $k$-median $\approx 1.74$
  (Guha-Khuller’99)
- **$\ell_2$-metric**: $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_1$-metric**: $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Trevisan’00)
- **$\ell_\infty$-metric**: $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  (Guruswami-Indyk’03)

Continuous Version:

$k$-means in **Euclidean** metric $< 1.36$
(Lee-Schmidt-Wright’17)
Hardness of Approximation

Discrete Version:

- **General metric**: $k$-means $\approx 3.94$, $k$-median $\approx 1.74$
  
  (Guha-Khuller’99)

- **$\ell_2$-metric**: $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  
  (Trevisan’00)

- **$\ell_1$-metric**: $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  
  (Trevisan’00)

- **$\ell_\infty$-metric**: $k$-means $\ll 1.01$, $k$-median $\ll 1.01$
  
  (Guruswami-Indyk’03)

Continuous Version:

$k$-means in Euclidean metric $\ll 1.36$, $1.07$

(Lee-Schmidt-Wright’17)
Discrete Version

<table>
<thead>
<tr>
<th>Metric</th>
<th>$k$-means (JCH)</th>
<th>$k$-median (JCH)</th>
<th>$k$-means (UGC)</th>
<th>$k$-median (UGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>$\ell_2$-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>$\ell_\infty$-metric</td>
<td>3.94*</td>
<td>1.73*</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>
Our Results (Cohen-Addad–K’19, Cohen-Addad–K–Lee)

Discrete Version

<table>
<thead>
<tr>
<th>Metric</th>
<th>$k$-means (JCH)</th>
<th>$k$-median (JCH)</th>
<th>$k$-means (UGC)</th>
<th>$k$-median (UGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>$\ell_2$-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>$\ell_\infty$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>

Continuous Version

$k$-means in $\ell_2$-metric $\approx 1.36$ (JCH), 1.07 (UGC)

$k$-median in $\ell_1$-metric $\approx 1.36$ (JCH), 1.07 (UGC)
Our Results (Cohen-Addad–K’19, Cohen-Addad–K–Lee)

Discrete Version

<table>
<thead>
<tr>
<th>Metric</th>
<th>$k$-means (JCH)</th>
<th>$k$-median (JCH)</th>
<th>$k$-means (UGC)</th>
<th>$k$-median (UGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>$\ell_2$-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>$\ell_\infty$-metric</td>
<td>3.94*</td>
<td>1.73*</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>

Continuous Version

- $k$-means in $\ell_2$-metric $\approx 1.36$ (JCH), $1.07$ (UGC)
- $k$-median in $\ell_1$-metric $\approx 1.36$ (JCH), $1.07$ (UGC)

A New Embedding Framework to potentially get Strong (tight?) Inapproximability results!
Warm up: General Metrics

Max Coverage:

- **Input**: Universe and Collection of Subsets \((U, S, k)\)

- **Objective**: Max Fraction of \(U\) covered by \(k\) subsets in \(S\)

**Theorem (Feige'98)**

- Fix \(\varepsilon > 0\).
- It is NP-hard to distinguish:
  - YES: Max Coverage is at least \(1 - \frac{1}{e} - \varepsilon\)
  - NO: Max Coverage is at most \(1 - \frac{1}{e} + \varepsilon\)
Max Coverage:

- **Input**: Universe and Collection of Subsets \((U, S, k)\)
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:
Max Coverage:

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

**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 \(1 - \frac{1}{e} + \varepsilon\)
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 $1 - \frac{1}{e} + \varepsilon$
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 $1 - \frac{1}{e} + \varepsilon$

**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 $1 - \frac{1}{e} + \varepsilon$

Theorem (Guha-Khuller’99)
Fix $\varepsilon > 0$. Given input $X$. It is NP-hard to distinguish:

YES: There exists $(C^*, \sigma^*)$ such that $\sum_{x \in X} \Delta(x, \sigma^*(x)) \leq |X|$
<table>
<thead>
<tr>
<th>Theorem (Feige’98)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix $\varepsilon &gt; 0$. It is NP-hard to distinguish:</td>
</tr>
<tr>
<td><strong>YES</strong>: Max Coverage is 1</td>
</tr>
<tr>
<td><strong>NO</strong>: Max Coverage is at most $1 - \frac{1}{e} + \varepsilon$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theorem (Guha-Khuller’99)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix $\varepsilon &gt; 0$. Given input $X$. It is NP-hard to distinguish:</td>
</tr>
<tr>
<td><strong>YES</strong>: There exists $(C^<em>, \sigma^</em>)$ such that $\sum_{x \in X} \Delta(x, \sigma^*(x)) \leq</td>
</tr>
<tr>
<td><strong>NO</strong>: For all $(C, \sigma)$ we have $\sum_{x \in X} \Delta(x, \sigma(x)) \geq (1 + \frac{2}{e} - \varepsilon) \cdot</td>
</tr>
</tbody>
</table>
Warm up: General Metrics

universe

clients

sets

candidate centers

$u_1$ in $S$

$u_2$ not in $S$

$S$

sets

candidate centers

1

3
Warm up: General Metrics

Max Coverage

- Cost is 1
- \( u_1 \) in \( S \)
- \( u_2 \) not in \( S \)

K-Median

- Cost is 1

Sets

Candidate centers

Universe

Points to cover

K-Median
Warm up: General Metrics

Max Coverage

universe

points to cover

sets

candidate centers

u_1 in S

u_2 not in S

cost is 1

cost is 3

cost is 3

k-Median
### (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

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

$$\forall \varepsilon > 0, \exists t_\varepsilon \in \mathbb{N} \text{ such that } (1 - \frac{1}{\varepsilon} + \varepsilon, t_\varepsilon)$$-Johnson Coverage problem is NP-hard.
$\bullet \ t = 2$: Vertex Coverage problem
$t = 2$: Vertex Coverage problem

- $\approx 0.9292$ gap is tight!
Johnson Coverage Hypothesis: Discussion

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

- Pick $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
Johnson Coverage Hypothesis: Discussion

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

- Pick $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq (\binom{n}{1})$: Max Coverage problem
  - As $t$ increases, gap approaches $1 - \frac{1}{e}$
Johnson Coverage Hypothesis: Discussion

- $t = 2$: **Vertex Coverage** problem
  - $\approx 0.9292$ gap is tight!

- Pick $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: **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 $C := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: 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
  - $\approx 0.9292$ gap is tight!

- Pick $\mathcal{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: **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 $t = 3$

- Improved **SDP gaps** for Clustering
Theorem (Cohen-Addad–K–Lee)

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

YES: There exists $(C^*, \sigma^*)$ such that

$$\|x - \sigma^*(x)\|_2 \leq n'$$

NO: For all $(C, \sigma)$ we have

$$\|x - \sigma(x)\|_2 \geq (1 + 8e) \cdot n'$$
Theorem (Cohen-Addad–K–Lee)

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

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

\[
\sum_{x \in X} \| (x - \sigma^*(x)) \|_0^2 \leq n',
\]
Theorem (Cohen-Addad–K–Lee)

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

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

\[
\sum_{x \in X} \|x - \sigma^*(x)\|_2^2 \leq n',
\]

**NO**: For all \((C, \sigma)\) we have

\[
\sum_{x \in X} \|x - \sigma(x)\|_2^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.
\]
Theorem (Cohen-Addad–K–Lee)

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

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

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

**NO:** For all \((C, \sigma)\) 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 \((1 - \frac{1}{e}, t)\)-Johnson coverage problem is NP-hard, given input \(X, S \subseteq \{0, 1\}^{O(\log n)}\), it is NP-hard to distinguish:

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

**NO**: For all \((C, \sigma)\) 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 $(\alpha, t)$-Johnson coverage problem is NP-hard, given input $X, S \subseteq \{0, 1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists $(C^*, \sigma^*)$ such that

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

NO: For all $(C, \sigma)$ 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 \(X, S \subseteq \{0, 1\}^{O(\log n)}\), it is NP-hard to distinguish:

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

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

**NO:** For all \((C, \sigma)\) 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 $X, S \subseteq \{0, 1\}^{O(\log n)}$, it is NP-hard to distinguish:

**YES**: There exists $(C^*, \sigma^*)$ such that

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

**NO**: For all $(C, \sigma)$ we have

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

\[
\begin{array}{c}
\binom{n}{t} \\
\binom{n}{t-1}
\end{array}
\]
Points in $\{0, 1\}^d$
Points in $\{0, 1\}^d$
Containment Game
Containment Game

\[ T \in \binom{[n]}{t} \]

\[ (\binom{[n]}{t-1}) \ni S \]
Containment Game

\[
T \in \binom{[n]}{t} \quad (\binom{[n]}{t-1} \ni S)
\]
Containment Game

\[ T \in \binom{[n]}{t} \]

\[ \text{Public Randomness} \]

\[ (\binom{[n]}{t-1}) \ni S \]
Containment Game

\[ T \in \binom{[n]}{t} \]

Public Randomness

GOAL

Determine if \( S \subset T \)
Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0
**Containment Game: Protocols**

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

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

○ Randomized Protocol:
  ○ Message length: $O_{\varepsilon,t}(1)$ bits
  ○ Completeness: 1, Soundness: $\varepsilon$
Let $\mathcal{C} : \mathbb{F}_q^{\log n} \rightarrow \mathbb{F}_q^{c \cdot \log n}$
Let $\mathcal{C} : \mathbb{F}_q^{\log n} \rightarrow \mathbb{F}_q^{c \cdot \log n}$

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

Alice and Bob pick randomly $i \in [c \cdot \log n]$

Bob sends to Alice $S_i := \{\mathcal{C}(u)_i \mid u \in S\}$
Let $\mathcal{C} : \mathbb{F}_q^{\log n} \rightarrow \mathbb{F}_q^{c \cdot \log n}$

Alice and Bob pick randomly $i \in [c \cdot \log n]$.

Bob sends to Alice $S_i := \{\mathcal{C}(u)_i \mid u \in S\}$.

Alice checks if $S_i \subseteq T_i := \{\mathcal{C}(u)_i \mid u \in T\}$. 

Message length: $(t-1) \cdot \log_2 q$

Soundness: $t \cdot (1 - \Delta(\mathcal{C})) \approx O(t \cdot (1/\sqrt{q}))$ (for AG codes)
Let $C : \mathbb{F}_q^{\log n} \rightarrow \mathbb{F}_q^{c \cdot \log n}$

- Alice and Bob pick randomly $i \in [c \cdot \log n]$

- Bob sends to Alice $S_i := \{C(u)_i \mid u \in S\}$

- Alice checks if $S_i \subseteq T_i := \{C(u)_i \mid u \in T\}$

- Message length: $(t - 1) \cdot \log_2 q$
Let $\mathcal{C} : \mathbb{F}_q^{\log n} \rightarrow \mathbb{F}_q^{c \cdot \log n}$

Alice and Bob pick randomly $i \in [c \cdot \log n]$

Bob sends to Alice $S_i := \{\mathcal{C}(u)_i | u \in S\}$

Alice checks if $S_i \subseteq T_i := \{\mathcal{C}(u)_i | u \in T\}$

Message length: $(t - 1) \cdot \log_2 q$

Soundness: $t \cdot (1 - \Delta(\mathcal{C}))$
Containment Game: Randomized Protocol

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

- Alice and Bob pick randomly $i \in [c \cdot \log n]$

- Bob sends to Alice $S_i := \{\mathcal{C}(u)_i \mid u \in S\}$

- Alice checks if $S_i \subseteq T_i := \{\mathcal{C}(u)_i \mid u \in T\}$

- Message length: $(t - 1) \cdot \log_2 q$

- Soundness: $t \cdot (1 - \Delta(\mathcal{C})) \approx O(t/(\sqrt{q}))$ (for AG codes)
Embedding Transcript into Hamming metric

Construct $\tau : 2^{[n]} \rightarrow \{0, 1\}^{q \cdot c \cdot \log n}$
Embedding Transcript into Hamming metric

- Construct $\tau : 2^n \rightarrow \{0, 1\}^{q \cdot c \cdot \log n}$

- Fix $i \in [c \cdot \log n]$ and $S \in 2^n$:
Embedding Transcript into Hamming metric

- Construct \( \tau : 2^n \rightarrow \{0, 1\}^{q \cdot c \cdot \log n} \)

- Fix \( i \in [c \cdot \log n] \) and \( S \in 2^n \):

\[
\tau(S)_i = e_{S_i}, \text{ where } S_i = \{ C(u)_i \mid u \in S \} \subseteq [q]
\]
Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \rightarrow \{0, 1\}^{c \cdot \log n}$

- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$
\tau(S)_i = e_{S_i}, \text{ where } S_i = \{c(u)_i \mid u \in S\} \subseteq [q]
$$

$S=\{1,2,\ldots,t\} \subseteq [n]

$S_i=\{1,2,\ldots,t\} \subseteq [q]$

$S_i=\{1,2,\ldots,t/2,q-t/2+1,\ldots,q\} \subseteq [q]$
Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \rightarrow \{0, 1\}^{q \cdot c \cdot \log n}$

- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:
  $$\tau(S)_i = e_{S_i}, \text{ where } S_i = \{C(u)_i \mid u \in S\} \subseteq [q]$$

- $X = \{\tau(T) \mid T \in E\}$
Embedding Transcript into Hamming metric

- Construct $\tau : 2^n \rightarrow \{0, 1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^n$:
  $\tau(S)_i = e_{S_i}$, where $S_i = \{C(u)_i \mid u \in S\} \subseteq [q]$
- $X = \{\tau(T) \mid T \in E\}$
- $S = \{\tau(S) \mid S \in \binom{[n]}{t-1}\}$
Structural Observations

Suppose $S \subseteq T$

For every block $i$, we have $S_i \subseteq T_i$

$S_i = \{1, 2, \ldots, t/2, q-t/2+1, \ldots, q\} \subseteq [q]$  

$T_i = S_i \cup \{t+1\} \subseteq [q]$  

\[
\begin{array}{cccccc}
1 & 1 \\
\vdots & \vdots \\
1 & 1 \\
0 & 0 \\
\vdots & \vdots \\
0 & 0 \\
0 & \hline
1 \\
0 & \vdots \\
0 & 0 \\
1 & \vdots \\
1 & 1 \\
1 & 1 \\
\end{array}
\]

$|\tau(T_i) - \tau(S_i)| = 1$
Suppose $S \not\subset T$

For most blocks $i$, we have $S_i \not\subset T_i$

<table>
<thead>
<tr>
<th>$S_i \setminus T_i$</th>
<th>$T_i \setminus S_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${q}$</td>
<td>${t+1, t+2}$</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccccccccccc}
1 & 1 \\
\vdots & \vdots \\
1 & 1 \\
0 & 0 \\
\vdots & \vdots \\
0 & 0 \\
0 & 1 \\
0 & 1 \\
0 & 0 \\
\vdots & \vdots \\
0 & 0 \\
1 & 1 \\
\vdots & \vdots \\
1 & 1 \\
1 & 0 \\
\end{array}
\]

$$|\tau(T_i) - \tau(S_i)| \geq 3$$
Completeness of Reduction

\( S' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1} \) be a **cover** of \( E \subseteq \binom{[n]}{t} \)
Completeness of Reduction

○ $S' := \{S_1, \ldots, S_k\} \subseteq \binom{n}{t-1}$ be a cover of $E \subseteq \binom{n}{t}$

○ Build $\sigma : X \rightarrow C \subseteq S$:
Completeness of Reduction

- $S' := \{S_1, \ldots, S_k\} \subseteq ([n]_{t-1})$ be a cover of $E \subseteq ([n]_t)$

- Build $\sigma : X \rightarrow C \subseteq S$

  \[\sigma(\tau(T)) = \tau(S_i), \text{ where } S_i \subset T\]
Completeness of Reduction

- $S' := \{S_1, \ldots, S_k\} \subseteq ([n]_{t-1})$ be a **cover** of $E \subseteq ([n]_t)$

- Build $\sigma : X \rightarrow C \subseteq S$:

  $$\sigma(\tau(T)) = \tau(S_i), \text{ where } S_i \subset T$$

- Fix $T \in E$ and $i \in [c \cdot \log n]$

  Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block $i$ is 1
Completeness of Reduction

- \( S' := \{ S_1, \ldots, S_k \} \subseteq \binom{[n]}{t-1} \) be a cover of \( E \subseteq \binom{[n]}{t} \)

- Build \( \sigma : X \to C \subseteq S \):
  \[
  \sigma(\tau(T)) = \tau(S_i), \text{ where } S_i \subset T
  \]

- Fix \( T \in E \) and \( i \in [c \cdot \log n] \)

Distance between \( \tau(T) \) and \( \sigma(\tau(T)) \) on block \( i \) is 1

- \( k\)-means objective is:
  \[
  \sum_{x \in X} \| x - \sigma(x) \|^2_0 = (c \cdot \log n)^2 \cdot |X|
  \]
Soundness of Reduction

○ $\sigma : X \rightarrow C \subseteq \mathcal{S}$ is some classification
Soundness of Reduction

- \( \sigma : X \rightarrow C \subseteq \mathcal{S} \) is some classification

- Build \( \mathcal{S}' \subseteq \binom{[n]}{t-1} \) of size \( k \):

\[
S \in \mathcal{S}' \iff \tau(S) \in C
\]
Soundness of Reduction

- \( \sigma : X \rightarrow C \subseteq S \) is some classification
- Build \( S' \subseteq \binom{[n]}{t-1} \) of size \( k \):
  \[
  S \in S' \iff \tau(S) \in C
  \]
- \( \exists E' \subseteq E \) s.t. \( \forall T \in E', T \) does not contain any subset in \( S' \)
Soundness of Reduction

- $\sigma : X \rightarrow C \subseteq S$ is some classification
- Build $S' \subseteq \binom{[n]}{t-1}$ of size $k$:
  \[ S \in S' \iff \tau(S) \in C \]
- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, $T$ does not contain any subset in $S'$
- Fix $\tau(T) \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block $i$ is mostly 3
Soundness of Reduction

- \( \sigma : X \to C \subseteq S \) is some classification
- Build \( S' \subseteq \binom{[n]}{t-1} \) of size \( k \):
  \[
  S \in S' \iff \tau(S) \in C
  \]
- \( \exists E' \subseteq E \), s.t. \( \forall T \in E' \), \( T \) does not contain any subset in \( S' \)
- Fix \( \tau(T) \in X_{E'} \) and \( i \in [c \cdot \log n] \)

Distance between \( \tau(T) \) and \( \sigma(\tau(T)) \) on block \( i \) is mostly 3

- \( k \)-means objective is:
  \[
  \sum_{x \in X} \|x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X \setminus X_{E'}| + 9(c \cdot \log n)^2 \cdot |X_{E'}|
  \]
Theorem (Cohen-Addad–K–Lee)

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

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

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

**NO:** For all \((C, \sigma)\) we have

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>(k)-means ((\text{JCH}))</th>
<th>(k)-median ((\text{JCH}))</th>
<th>(k)-means ((\text{UGC}))</th>
<th>(k)-median ((\text{UGC}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ell_1)-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>(\ell_2)-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>(\ell_\infty)-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>

Continuous Version

- \(k\)-means in \(\ell_2\)-metric \(\approx 1.36 \text{ (JCH)}, 1.07 \text{ (UGC)}\)
- \(k\)-median in \(\ell_1\)-metric \(\approx 1.36 \text{ (JCH)}, 1.07 \text{ (UGC)}\)
Our Results (Cohen-Addad–K’19, Cohen-Addad–K–Lee)

Discrete Version

<table>
<thead>
<tr>
<th></th>
<th>$k$-means (JCH)</th>
<th>$k$-median (JCH)</th>
<th>$k$-means (UGC)</th>
<th>$k$-median (UGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>$\ell_2$-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>$\ell_\infty$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>

Continuous Version

$k$-means in $\ell_2$-metric $\approx 1.36$ (JCH), 1.07 (UGC)

$k$-median in $\ell_1$-metric $\approx 1.36$ (JCH), 1.07 (UGC)
### Discrete Version

<table>
<thead>
<tr>
<th>Metric</th>
<th>k-means (JCH)</th>
<th>k-median (JCH)</th>
<th>k-means (UGC)</th>
<th>k-median (UGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ℓ₁-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>ℓ₂-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>ℓ∞-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>

### Continuous Version

- **k-means** in ℓ₂-metric \(\approx 1.36\) (JCH), 1.07 (UGC)
- **k-median** in ℓ₁-metric \(\approx 1.36\) (JCH), 1.07 (UGC)

Use Feige’s Instance
## Our Results (Cohen-Addad–K’19, Cohen-Addad–K–Lee)

### Discrete Version

<table>
<thead>
<tr>
<th>Metric</th>
<th>$k$-means (JCH)</th>
<th>$k$-median (JCH)</th>
<th>$k$-means (UGC)</th>
<th>$k$-median (UGC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$-metric</td>
<td>3.94</td>
<td>1.73</td>
<td>1.56</td>
<td>1.14</td>
</tr>
<tr>
<td>$\ell_2$-metric</td>
<td>1.73</td>
<td>1.27</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>$\ell_\infty$-metric</td>
<td>3.94*</td>
<td>1.73</td>
<td>3.94*</td>
<td>1.73*</td>
</tr>
</tbody>
</table>

### Continuous Version

- **$k$-means** in $\ell_2$-metric $\approx 1.36$ (JCH), 1.07 (UGC)
- **$k$-median** in $\ell_1$-metric $\approx 1.36$ (JCH), 1.07 (UGC)

Use Feige’s Instance

Johnson graph Embedding

Decoding Vertex Cover
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 $\ell_p$-metrics
Key Takeaways

- Improved Inapproximability of
- $k$-means and $k$-median
- In $\ell_p$-metrics
- Using Transcript of Containment Protocol
Key Takeaways

- Improved Inapproximability of
- $k$-means and $k$-median
- In $\ell_p$-metrics
- Using Transcript of Containment Protocol
- And Geometric Realization of Johnson Graphs
Key Takeaways

- Improved Inapproximability of
- $k$-means and $k$-median
- In $\ell_p$-metrics
- Using Transcript of Containment Protocol
- And Geometric Realization of Johnson Graphs

Open: Is JCH true?
THANK YOU!