An Improved Approximation for $k$-median, and Positive Correlation in Budgeted Optimization

Jaroslaw Byrka $^1$ Thomas Pensyl $^2$ Bartosz Rybicki $^1$
Aravind Srinivasan $^2$ Khoa Trinh $^2$

$^1$University of Wroclaw, Poland
$^2$University of Maryland, USA

ISMP 2015
**k-median**

- **Input:**
  - set of facilities $\mathcal{F}$
  - set of clients $\mathcal{C}$
  - symmetric distance metric $d$ on $\mathcal{C} \cup \mathcal{F}$
  - an integer $k > 0$

- **Goal:** pick $k$ facilities which minimize total connection cost of clients.

- **Example:** $k = 2$, open facilities $A, C$ and pay cost $2 + 2 + 3 = 7$. 
Uncapacitated Facility Location

- **Input:**
  - set of facilities $\mathcal{F}$
  - set of clients $C$
  - symmetric distance metric $d$ on $C \cup \mathcal{F}$
  - each facility has an opening cost

- **Goal:** pick some facilities which minimize total opening cost and connection cost.

- **Example:** open facilities $A, C$ and pay cost $2 + 2 + 3 + 3 + 1 = 11$. 

![Diagram showing facility locations and connections with costs labeled.]
Bi-point solution as middle step

**Bi-point solution:** a feasible solution which is a convex combination of two integral solutions.

- **Step I:** Construct a bi-point solution.
  - Use some special UFL algorithms to get a “cheap” bi-point solution.

- **Step II:** Round bi-point solution to integral one.
## Previous Work

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Bi-point Construction</th>
<th>Bi-point Rounding</th>
<th>k-median Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>’99</td>
<td>Charikar et. al.</td>
<td>(LP Rounding)</td>
<td>6.667</td>
<td></td>
</tr>
<tr>
<td>’99</td>
<td>Jain &amp; Vazirani</td>
<td>3 + ε</td>
<td>2</td>
<td>6 + ε</td>
</tr>
<tr>
<td>’02</td>
<td>Jain, Mahdian &amp; Saberi</td>
<td>2 + ε</td>
<td>2</td>
<td>4 + ε</td>
</tr>
<tr>
<td>’01</td>
<td>Arya et. al.</td>
<td>(Local Search)</td>
<td>3 + ε</td>
<td></td>
</tr>
<tr>
<td>’12</td>
<td>Li &amp; Svensson</td>
<td>2 + ε</td>
<td>1.366 + ε</td>
<td>2.732 + ε</td>
</tr>
<tr>
<td>’15</td>
<td><strong>Our work</strong></td>
<td>2 + ε</td>
<td><strong>1.337 + ε</strong></td>
<td><strong>2.674 + ε</strong></td>
</tr>
</tbody>
</table>
Bi-point Rounding

- Again, bi-point solution is convex combination of two integral solutions, $\mathcal{F}_1$ and $\mathcal{F}_2$.
- $a|\mathcal{F}_1| + b|\mathcal{F}_2| = k$, where $(a, b > 0, a + b = 1)$
Form stars by attaching each facility in $\mathcal{F}_1$ to closest facility in $\mathcal{F}_1$. 

$\mathcal{F}_1$

$\mathcal{F}_2$
Li-Svensson’s rounding algorithm

- Key property: “if some leaf is closed then its root must be open”.
- \( \Pr[i_1 \text{ is open}] \approx a \) for all \( i_1 \in \mathcal{F}_1 \)
- \( \Pr[i_2 \text{ is open}] \approx b \) for all \( i_2 \in \mathcal{F}_2 \)
- approximation factor = \( \frac{1 + \sqrt{3}}{2} \approx 1.366 \)
- \#open facilities = \( a|\mathcal{F}_1| + b|\mathcal{F}_2| + O(1) = k + O(1) \).
There is an open facility in \{i_1, i_2, i_3\}

\[
d_3 \leq d_2 + d(i_2, i_3) \leq d_1 + 2d_2
\]

The total (expected) connection cost is a (non-linear) function of cost (\mathcal{F}_1), cost (\mathcal{F}_2), a, b which can be related to OPT.
Tight Instance

- In tight case, almost all stars have exactly 1 leaf.
- BUT still many clients near 2-leafed stars.
Why not close some 1-leafed stars to open more leaves of the 2-leafed stars?
Consider any 1-leafed star $\langle i_2, i_3 \rangle$ and let $i_4$ be the closest leaf of a 2-leafed star to $i_3$. Then the star $\langle i_2, i_3 \rangle$ is long if

$$d(i_2, i_3) \geq gd(i_3, i_4),$$

for some constant $g > 0$. 

Long 1-leafed stars
Long 1-leafed stars can be closed
Improved Bi-point rounding algorithm:

- Consists of 8 different rounding strategies.
- Analysis uses a non-linear factor revealing program.
- Approximation ratio: 1.337.
Dependent Rounding

Given a vector $P = (p_1, \ldots, p_n)$ with $0 \leq p_i \leq 1$, Dependent Rounding algorithm rounds $P$ into a 0-1 vector $X$ such that

- the sum is preserved: $\sum_{i=1}^n X_i = \sum_{i=1}^n p_i$,
- the marginals are preserved: $\Pr[X_i = 1] = p_i$,
- negative correlation between $X_i$'s: e.g.
  - $\Pr[X_i = 1 \land X_j = 1] \leq p_ip_j$
  - $\Pr[X_i = 0 \land X_j = 0] \leq (1 - p_i)(1 - p_j)$
A Typical Example of Negative Correlation in Facility Location Problems

Assume that we have a client $j$ and a set $S$ of facilities that are close to $j$. Each facility $i \in S$ has an opening variable $y_i$. Applying dependent rounding to opening variables in $S$ gives

$$
\Pr[\text{all facilities in } S \text{ are closed}] = \Pr \left[ \bigwedge_{i \in S} \{i \text{ is closed}\} \right] \\
\leq \prod_{i \in S} (1 - \Pr[i \text{ is open}]) \\
\leq \exp \left( - \sum_{i \in S} y_i \right).
$$
For each root $i \in \mathcal{F}_1$, let $X_i$ be the indicator whether $i$ or its leaves is opened.
Positive Correlation

For each root $i \in \mathcal{F}_1$, let $X_i$ be the indicator whether $i$ or its leaves is opened.

We want to upper-bound $\mathbb{E}[\text{cost}(j)]$:

$$
\mathbb{E}[\text{cost}(j)] \leq \Pr[X_{i_3} = 0]d_2 + \Pr[X_{i_3} = 1 \land X_{i_1} = 1]d_1 \\
+ \Pr[X_{i_3} = 1 \land X_{i_1} = 0]d_3
$$
For each root $i \in \mathcal{F}_1$, let $X_i$ be the indicator whether $i$ or its leaves is opened.

We want to upper-bound $E[\text{cost}(j)]$:

$$E[\text{cost}(j)] \leq \Pr[X_{i_3} = 0]d_2 + \Pr[X_{i_3} = 1 \wedge X_{i_1} = 1]d_1$$
$$+ \Pr[X_{i_3} = 1 \wedge X_{i_1} = 0]d_3$$

$\Pr[X_{i_3} = 1 \wedge X_{i_1} = 1] \leq \Pr[X_{i_3} = 1] \Pr[X_{i_1} = 1] = a^2$, by negative correlation
Positive Correlation

For each root $i \in F_1$, let $X_i$ be the indicator whether $i$ or its leaves is opened.

We want to upper-bound $E[\text{cost}(j)]$:

$$E[\text{cost}(j)] \leq \Pr[X_{i_3} = 0]d_2 + \Pr[X_{i_3} = 1 \land X_{i_1} = 1]d_1$$
$$+ \Pr[X_{i_3} = 1 \land X_{i_1} = 0]d_3$$

- $\Pr[X_{i_3} = 1 \land X_{i_1} = 1] \leq \Pr[X_{i_3} = 1] \Pr[X_{i_1} = 1] = a^2$, by negative correlation
- How about $\Pr[X_{i_3} = 1 \land X_{i_1} = 0]$?
Near Independence

- Independence:
  \[ \Pr[A \land B \land C] = p_A p_B p_C \]

- Near-Independence:
  \[ (1 - \epsilon)p_A p_B p_C \leq \Pr[A \land B \land C] \leq (1 + \epsilon)p_A p_B p_C \]

- Can dependent rounding achieve this property?
  **Answer:** Yes, but with limits...
Extreme Examples

- Observation 1: if $p_1 = p_2 = \ldots = p_n = 1/2$, we have

$$\Pr[X_1 = 1 \land X_2 = 0 \land \ldots \land X_n = 0] = 0.$$  

On the other hand, $\Pr[X_1 = 1] \prod_{i=2}^{n} \Pr[X_i = 0] = 2^{-n} > 0$. 

*Get around this by considering a “small” subset of events.*
Extreme Examples

Observation 1: if \( p_1 = p_2 = \ldots = p_n = 1/2 \), we have

\[
\Pr[X_1 = 1 \land X_2 = 0 \land \ldots \land X_n = 0] = 0.
\]

On the other hand, \( \Pr[X_1 = 1] \prod_{i=2}^{n} \Pr[X_i = 0] = 2^{-n} > 0 \).

Get around this by considering a “small” subset of events.

Observation 2: if \( p_1 = p_2 = \ldots = p_n = 1/n \). Then,

\[
\Pr[X_1 = 1 \land X_2 = 1] = 0.
\]

On the other hand, \( \Pr[X_1 = 1] \Pr[X_2 = 1] = 1/n^2 > 0 \).

Get around this by requiring that the values \( p_i \) are not too close to 0 or 1.
Modified Dependent Rounding

- **Modification:** randomly permute the variables before applying dependent rounding.

Our results: if there is some $\alpha > 0$ that $\alpha \leq p_i \leq 1 - \alpha$ for all $i$, then for any $t$ events of the form $X_i = 0$ or $X_i = 1$, their joint probability lies within $(1 - (1 - t \frac{2}{n} \alpha^2))$ and $(1 + (1 + t \frac{2}{n} \alpha^2))^{t-1}$, of the value it would if these events were independent. In particular, these factors are $(1 + o(1))$ and $(1 - o(1))$ if $t = o(\alpha \sqrt{n})$. 
*Modification:* randomly permute the variables before applying dependent rounding.

*Our results:* if there is some $\alpha > 0$ that $\alpha \leq p_i \leq 1 - \alpha$ for all $i$, then for any $t$ events of the form $X_i = 0$ or $X_i = 1$, their joint probability lies within

$$\left(1 - \left(1 - \frac{t^2}{n\alpha^2}\right)\right) \text{ and } \left(1 + \left(1 + \frac{t}{n\alpha^2}\right)^{t-1}\right),$$

of the value it would if these events were independent.
**Modified Dependent Rounding**

- **Modification:** randomly permute the variables before applying dependent rounding.

- **Our results:** if there is some $\alpha > 0$ that $\alpha \leq p_i \leq 1 - \alpha$ for all $i$, then for any $t$ events of the form $X_i = 0$ or $X_i = 1$, their joint probability lies within

  \[
  \left(1 - \left(1 - \frac{t^2}{n\alpha^2}\right)\right) \quad \text{and} \quad \left(1 + \left(1 + \frac{t}{n\alpha^2}\right)^{t-1}\right),
  \]

  of the value it would if these events were independent.

  - in particular, these factors are $(1 + o(1))$ and $(1 - o(1))$ if $t = o(\alpha\sqrt{n})$.
Modified Dependent Rounding

Application to $k$-median:

- reduce #extra facilities from $O(1/\epsilon)$ to $O(\log(1/\epsilon))$
- reduce run-time from $n^{O(1/\epsilon^2)}$ to $n^{O(1/\epsilon \log(1/\epsilon))}$. 
Open questions

- Closing the gap for k-median?
  - Our algorithm: $(2.674 + \epsilon)$-apx
  - Best known lower bound: $1 + 2/e \approx 1.735$
- Other applications of dependent rounding with near-independence property?
Thank you!