A Constant Factor Approximation for Lower-Bounded k-Median

Yutian Guo, Junyu Huang, and Zhen Zhang(✉)

School of Computer Science and Engineering, Central South University, Changsha 410083, People's Republic of China
csuzz@foxmail.com

Abstract. The lower-bounded k-median problem considers a set C of clients, a set F of facilities, and a parameter B, the goal is to open k facilities and connect each client to an opened facility, such that each opened facility is connected with at least B clients and the total connection cost is minimized. The problem is known to admit an $O(1)$-approximation algorithm, while the constant is implicit and seems to be a very large constant. In this paper, we give an approach that converts the lower-bounded k-median problem to the capacitated facility location problem, which yields a $(516+\epsilon)$-approximation for the lower-bounded k-median problem.

Keywords: Approximation algorithm · k-median

1 Introduction

k-median is a widely studied clustering problem and has applications in many fields related to computer science. Given a set C of clients and a set F of facilities located in a metric space, the k-median problem aims to open a set $S \subseteq F$ of at most k facilities, such that the objective function $\sum_{j \in C} d(j, S)$ is minimized, where $d(j, S)$ denotes the distance from $j \in C$ to its nearest facility in S. This problem is known to be NP-hard, which leads to a lot of efforts devoted to obtaining its approximation algorithms [5,7,11,13,16]. The current best approximation guarantee for the problem is $2.675 + \epsilon$ [5], which is obtained Li and Svensson [16].

The clustering problem has an inherent assumption that each client can be optionally connected to any opened facility. However, many real world scenarios associate a notion of *lower bound* with each facility, and the number of the clients connected to each facility should not be less than the lower bound associated with it. For example, in the design of buy-at-bulk network, a set of demands needs to be connected to a set of servers, and each server is required to have

This work was supported by National Natural Science Foundation of China (61672536, 61872450, 61828205, and 61802441), Hunan Provincial Key Lab on Bioinformatics, and Hunan Provincial Science and Technology Program (2018WK4001).

J. Chen et al. (Eds.): TAMC 2020, LNCS 12337, pp. 119–131, 2020.
https://doi.org/10.1007/978-3-030-59267-7_11

a minimum amount of demand assigned to it. Karger and Minkoff [12] and Guha *et al.* [10] introduced the lower-bounded facility location problem to deal with such constraints. They presented constant-factor bi-criteria approximation algorithms for the problem, which violate the lower bound of the facilities. On the basis of the techniques given in [12] and [10], Svitkina [17] gave a $(448+\epsilon)$-approximation without violating the lower bound constraint. The ratio was later improved to $82.6+\epsilon$ by Ahmadian and Swamy [1]. Most recently, Li [15] gave a $(3926+\epsilon)$-approximation for the facility location problem with non-uniform lower bounds.

In this paper, we consider the lower-bounded k-median problem.

Definition 1 (lower-bounded k-median). *Given a set C of clients and a set F of facilities located in a metric space, an integer k, and a parameter B, the lower-bounded k-median problem is to open a set S of at most k facilities, and identify a connection function σ, such that the number of the clients connected to each facility is no less than B, and the objective cost$(S,\sigma)=\sum_{j\in C} d(j,\sigma(j))$ is minimized, where $\sigma(j)=i$ denotes client j is connected to facility $i\in S$ for each $j\in C$, and $d(j,\sigma(j))$ denotes the distance from j to $\sigma(j)$.*

In Euclidean space, Ding and Xu [9] gave a $(1+\epsilon)$-approximation for the lower-bounded k-median problem with running time $O(n^2d\cdot(\log n)^{k+2}\cdot 2^{poly(k/\epsilon)})$. Bhattacharya *et al.* [4] later improved the running time of the algorithm in [9] to $O(n^2d\cdot(\log n)^2\cdot(\frac{k}{\epsilon})^{O(k/\epsilon)})$. Ahmadian and Swamy [2] gave that the problem admits an $O(1)$-approximation algorithm that runs in polynomial time. The approximation ratio is implicit but seems to be a very large number.

1.1 Our Results

In this paper, we obtain the following result for the lower-bounded k-median problem.

Theorem 1. *There exists a $(516+\epsilon)$-approximation algorithm for the lower-bounded k-median problem that runs in polynomial time.*

We now give the high level idea of our approach. Given an instance of the lower-bounded k-median problem, it can be seen that the instance has a feasible solution if $|C|\geq B$. We present a bi-criteria approximation algorithm for the problem. The algorithm yields a constant factor approximation solution which violates the lower bound of the facilities. To convert such a bi-criteria approximation solution to a feasible solution, we reconnect some clients and close some facilities so that the lower bound constraint of each facility can be satisfied. We consider an instance of the capacitated facility location problem to minimize the loss in the cost induced by the converting. This instance is constructed by interchanging the roles of the clients and the facilities. A set of clients connected to a same location is now viewed as a facility whose capacity is the same as the size of the client set, and a facility whose lower bound is violated now becomes a

set T of clients with $|T| = (1-\lambda)B$, where λB is the number of the clients connected to this facility in the original bi-criteria approximation solution. Based on a known $O(1)$-approximation algorithm for the capacitated facility location problem, we convert the bi-criteria solution to a solution satisfying the lower bound, which induces a constant factor loss in the approximation ratio.

1.2 Other Related Work

A commonly studied extension of the clustering problem is the capacitated clustering problem, which can be viewed as the opposite of the lower-bounded clustering problem in some sense. In this problem, each facility is associated with a capacity, and the number of the clients connected to a facility should be less than its capacity. The capacitated facility location problem can be formally defined as follows.

Definition 2 (Capacitated facility location). *Given a set C of clients and a set F of facilities located in a metric space, an opening cost f_i and a capacity u_i associated with each $i \in F$, the capacitated facility location problem is to open a set S of facilities, and identify a connection function σ, such that the number of the clients connected to each facility $i \in S$ is no more than u_i, and the objective $cost(S,\sigma) = \sum_{j \in C} d(j, \sigma(j)) + \sum_{i \in S} f_i$ is minimized, where $\sigma(j) = i$ denotes client j is connected to facility $i \in S$ for each $j \in C$, and $d(j, \sigma(j))$ denotes the distance from j to $\sigma(j)$.*

The capacitated clustering problem is significantly harder than the ordinary clustering problem. There are several known $O(1)$-approximation algorithms for the capacitated facility location problem. The current best approximation guarantee for the problem is $5 + \epsilon$ [3], which was obtained based on a local search algorithm. However, constant factor approximation algorithms for the capacitated k-median problem only exist for the case where the capacity constraint or the number of clusters can be violated [8,14].

2 A Bi-criteria Approximation

In this section, we give a constant factor bi-criteria approximation for the lower-bounded k-median problem by constructing an instance of the k-facility location problem. This problem can be formally defined as follows.

Definition 3 (k-facility location). *Given a set C of clients and a set F of facilities located in a metric space, an integer k, and an opening cost f_i associated with each $i \in F$, the k-facility location problem aims to open a set S of at most k facilities, such that the objective $\sum_{j \in C} d(j, S) + \sum_{i \in S} f_i$ is minimized, where $d(j, S)$ denotes the distance from j to its nearest facility in S.*

Let $I = (C, F, k, B)$ be an instance of the lower-bounded k-median problem. Given a facility $i \in F$, let J_i denote the set of the B clients in C closest to i. Given $\beta \in (0, 1)$ and a solution (S, σ) of I, we call (S, σ) a β-covered solution if it

connects no less than βB clients to each $i \in F$, and let $cost_I(S, \sigma)$ denote its cost of the problem. We construct an instance $I' = (C, F, f)$ for the k-facility location problem, where $f_i = \frac{2\beta}{1-\beta} \sum_{j \in J_i} d(i, j)$ for each $i \in F$. We solve the constructed instance by the algorithm given in [6], which gives a 3.25-approximation solution for the k-facility location problem. Let S' be the resulted set of the open facilities. Let $cost_{I'}(S')$ denote the cost of S' for the k-facility location problem.

We now show that any λ-approximation solution of I' can be converted to an $O(\lambda)$-approximation solution of I, which induces a constant factor violation in the lower bound of the facilities.

Lemma 1. *For an arbitrary solution (S, σ) of I, the solution is also a feasible for I', and we have*

$$\sum_{i \in S} f_i \leq \frac{2\beta}{1-\beta} cost(S, \sigma).$$

Proof. Since (S, σ) is a feasible solution of I, for each $i \in S$, i is connected with at least B clients. This implies that $\sum_{j \in J_i} d(i, j) \leq \sum_{j \in \sigma^{-1}(i)} d(i, j)$. Thus, we have

$$\begin{aligned}
\sum_{i \in S} f_i &= \sum_{i \in S} \left(\frac{2\beta}{1-\beta} \sum_{j \in J_i} d(i, j)\right) \\
&\leq \frac{2\beta}{1-\beta} \sum_{i \in S, j \in \sigma^{-1}(i)} d(i, j) \\
&= \frac{2\beta}{1-\beta} \sum_{j \in C} d(j, \sigma(j)) \\
&= \frac{2\beta}{1-\beta} cost_I(S, \sigma),
\end{aligned}$$

where the first step follows from the the definition of f_i. □

Lemma 1 implies that for any solution (S, σ) of instance I, we have $cost_{I'}(S) \leq \frac{1+\beta}{1-\beta} cost_I(S)$. We proceed by showing that a solution S' to I' can be converted to a β-covered solution (S, σ) of I.

Lemma 2. *Given a solution S' of I', we can find a β-covered solution (S, σ) of I such that $cost_I(S, \sigma) \leq cost_{I'}(S')$.*

Proof. We prove the lemma by giving an algorithm that yields the desired β-covered solution. Based on instance I', we construct a new instance I^k for the k-median problem by removing all facility open costs. Let $S = S'$ initially. While there exists some $i \in S$ such that $cost_{I'}(S \backslash \{i\}) \leq cost_{I'}(S)$, let $S = S \backslash \{i\}$. The final solution (S, σ) is called a minimal feasible solution of instance I^k, where each $j \in C$ is assigned to its nearest facility in S by function σ. It is easy to show that $cost_{I^k}(S, \sigma) \leq cost_{I'}(S)$, which implies that $cost_{I^k}(S, \sigma) \leq cost_{I'}(S) \leq cost_{I'}(S')$.

Now we want to show that in the solution (S, σ), each facility $i \in S$ is connected by at least βB clients. For the sake of contradiction, assume that there exists a facility $i \in S$ such that $|\sigma^{-1}(i)| \leq \beta B$. This implies that $|J_i \backslash \sigma^{-1}(i)| \geq (1-\beta)B$. We know that there exists a client $j' \in J_i \backslash \sigma^{-1}(i)$, such that

$$d(i, j') \leq \frac{1}{(1-\beta)B} \sum_{j \in J_i \backslash \sigma^{-1}(i)} d(i, j) \leq \frac{1}{(1-\beta)B} \sum_{j \in J_i} d(i, j).$$

Since j' is not connected to i in the solution (S, σ), it is connected to some other facility $i' \in S$ with $d(j', i') \leq d(j', i)$. Thus, we have

$$\begin{aligned}
\sum_{j \in \sigma^{-1}(i)} d(j, i') &\leq \sum_{j \in \sigma^{-1}(i)} (d(j, i) + d(i, j') + d(j', i')) \\
&\leq \sum_{j \in \sigma^{-1}(i)} d(j, i) + |\sigma^{-1}(i)| \times 2d(i, j') \\
&\leq \sum_{j \in \sigma^{-1}(i)} d(j, i) + \beta B \times \frac{2}{(1-\beta)B} \sum_{j \in J_i} d(i, j) \\
&= \sum_{j \in \sigma^{-1}(i)} d(j, i) + \frac{2\beta}{(1-\beta)} \sum_{j \in J_i} d(i, j).
\end{aligned}$$

If we close i and reconnect each client from i to i', then the increment in the connection cost is no more than $\frac{2\beta}{(1-\beta)} \sum_{j \in J_i} d(i, j)$, which is bounded by f'_i. We have $cost_{I'}(S \backslash \{i\}) \leq cost_{I'}(S)$, contradicting that (S, σ) is a minimal feasible solution of instance I^k. Thus (S, σ) is a β-covered solution of I. □

Based on Lemma 1 and Lemma 2, we get the following approximation guarantee.

Theorem 2. *There exists a* $3.25\frac{1+\beta}{1-\beta}$*-approximation algorithm for the lower-bounded* k*-median problem which violates the lower bound by a factor* β*.*

Proof. We first get a solution S' of I' using the 3.25-approximation algorithm for the k-facility location problem. We denote the set of the opened facilities in an optimal solution of I by S^*. We have $cost_{I'}(S') \leq 3.25 cost_{I'}(S^*)$. Thus, we get

$$\begin{aligned}
cost_{I'}(S') &\leq 3.25\Big(\sum_{i \in S^*} f_i + \sum_{j \in C} d(j, S^*)\Big) \\
&\leq 3.25\Big(\frac{2\beta}{1-\beta} cost_I(S^*, \sigma^*) + \sum_{j \in C} d(j, S^*)\Big) \\
&\leq 3.25\Big(\frac{2\beta}{1-\beta} cost_I(S^*, \sigma^*) + cost_I(S^*, \sigma^*)\Big) \\
&= 3.25 \frac{1+\beta}{1-\beta} cost_I(S^*, \sigma^*),
\end{aligned}$$

where the second step follows from Lemma 1. By Lemma 2, we can obtain a β-covered solution (S^o, σ^o) of I such that

$$cost_I(S^o, \sigma^o) \leq cost_{I'}(S') \leq 3.25(\frac{1+\beta}{1-\beta})cost_I(\sigma^*).$$

□

By Theorem 2, we can get a pseudo-solution for the lower-bounded k-median problem, which violates the lower bound restriction by the a constant $\beta \in (0, 1)$. This implies that we find a solution where each opened facility is connected with at least βB clients instead of B clients. In the following we will show how to make such a solution feasible for the lower-bounded k-median problem.

3 The Approximation Algorithm

3.1 Aggregating Clients

Given an instance $I = (C, F, B, k)$ of the lower-bounded k-median problem, by Lemma 1 and Lemma 2, we can obtain a bi-criteria approximation solution (S^o, σ^o) which violates the constraint of lower bound by a factor β. We construct a new instance I^1 for the lower-bounded k-median problem, where C, F, and B are the same as that of I, but the metric is different from I. In instance I^1, each client $j \in C$ is moved to $\sigma^o(j)$. Then, for each $i_1, i_2 \in F$ and $j_1, j_2 \in C$, we have $d^1(i_1, i_2) = d(i_1, i_2)$, $d^1(i_1, j_1) = d(i_1, \sigma^o(j_1))$, and $d^1(j_1, j_2) = d(\sigma^o(j_1), \sigma^o(j_2))$. For arbitrary $i \in F$ and $j \in C$, using triangle inequality, we get

$$d^1(i, j) = d(i, \sigma^o(j)) \leq d(i, j) + d(j, \sigma^o(j)). \tag{1}$$

For an optimal solution (S^*, σ^*) of instance I, we have

$$\begin{aligned} cost_{I^1}(S^*, \sigma^*) &\leq cost_I(S^*, \sigma^*) + cost_I(S^o, \sigma^o) \\ &\leq cost_I(S^*, \sigma^*) + 3.25\frac{1+\beta}{1-\beta}cost_I(S^*, \sigma^*) \\ &= \left(1 + 3.25\frac{1+\beta}{1-\beta}\right) cost_I(S^*, \sigma^*), \end{aligned} \tag{2}$$

where the first step follows from inequality (1), the second step follows from Theorem 2. We have the following result based on the methods in [17].

Theorem 3. *If there is an α_1-approximation solution of I^1, we can efficiently find an α-approximation solution of I, where $\alpha = \alpha_1(1 + 3.25\frac{1+\beta}{1-\beta}) + 3.25\frac{1+\beta}{1-\beta}$.*

3.2 Contracting Facility Set

We now focus on instance $I^1 = (C, F, B, k)$. For each $i \in S^o$, define $\gamma_i = \{j|\sigma^{o-1}(i)\}$ as the set of clients connected to i. We have $|\gamma_i| \geq \beta B$ for each $i \in S^o$. An instance $I^2 = (C, S^o, B, k)$ is constructed by removing each facility in $F \backslash S^o$ from I^1.

Lemma 3. *If there is a solution (S^1, σ^1) of I^1, then we can efficiently find a solution (S^2, σ^2) of I^2 such that $cost_{I^2}(S^2, \sigma^2) \leq 2cost_{I^1}(S^1, \sigma^1)$.*

Proof. For each $i \in F \backslash S^o$, let i' denote the facility in S^o nearest to i. We construct a solution (S^2, σ^2) of I^2 by opening each $i \in S^1 \cap S^o$ and facility i' for each $i \in F \cap (S^1 \backslash S^o)$. For a facility $i \in F \cap (S^1 \backslash S^o)$, the clients connected to i in solution (S^1, σ^1) are reconnected to i'. By triangle inequality and the definition of $d^1(*)$, the increased cost induced by a client j is bounded by $d^1(i, i') = d(i, i') \leq d(i, \sigma^{o-1}(j)) = d^1(i, j)$. Summing the inequality over each $j \in C$, we get that the total increased cost is no more than $cost_{I^1}(S^1, \sigma^1)$, which implies that $cost_{I^2}(S^2, \sigma^2) \leq 2cost_{I^1}(S^1, \sigma^1)$. □

Lemma 3 implies that a solution of instance I^1 can be converted to a feasible solution of instance I^2. It is easy to see that a solution (S, σ) of instance I^2 is also feasible of instance I^1 and satisfies $cost_{I^1}(S, \sigma) = cost_{I^2}(S, \sigma)$. Thus, we get the following result for I^1 and I^2.

Theorem 4. *Given an α_2-approximation solution of I^2, we can find an α_1-approximation solution of I^1, where $\alpha_1 = 2\alpha_2$.*

Proof. Let (S^{*1}, σ^{*1}) be an optimal solution of instance I^1. Using Lemma 3, we can get a solution (S^2, σ^2) of I^2 that satisfies $cost_{I^2}(S^2, \sigma^2) \leq 2cost_{I^1}(S^{*1}, \sigma^{*1})$. Let (S, σ) denote an α_2-approximation solution of I^2, we have $cost_{I^2}(S, \sigma) \leq 2\alpha_2 cost_{I^1}(S^{*1}, \sigma^{*1})$. Recall that (S, σ) is also feasible for I^1 and $cost_{I^1}(S, \sigma) = cost_{I^2}(S, \sigma) \leq 2\alpha_2 cost_{I^1}(S^{*1}, \sigma^{*1})$. □

Now we focus on instance I^2. We only consider one facility for each position in I^2.

3.3 Adding Penalties to Instance I^2

Based on instance I^2, we construct a new instance I^3 by considering penalties for closing the facilities from S^o. For each $i \in S^o$, if i is closed in the solution, then a penalty cost $Pc_{I^3}(i) = \frac{2\beta-1}{\beta}|\gamma_i|\ell_i$ should be paid, where ℓ_i denotes the distance from i to its nearest facility in $S^o \backslash \{i\}$. For a solution (S, σ) of I^3, define $Pc_{I^3}(S, \sigma) = \sum_{i \in S^o \backslash S} Pc_{I^3}(i)$ as the total penalty cost of (S, σ).

Lemma 4. *For any solution (S, σ) of I^2 and I^3, we have*

$$cost_{I^2}(S, \sigma) \leq cost_{I^3}(S, \sigma) \leq \frac{3\beta - 1}{\beta} cost_{I^2}(S, \sigma).$$

Proof. The cost of solution (S, σ) of I^3 consists of the connection cost and the penalty cost of the closed facilities, where the connection cost is equal to $cost_{I^2}(S, \sigma)$. Thus, $cost_{I^2}(S, \sigma) \leq cost_{I^3}(S, \sigma)$. We have

$$\begin{aligned}\sum_{i \in S^o \setminus S} Pc_{I^3}(i) &= \sum_{i \in S^o \setminus S} \frac{2\beta - 1}{\beta} |\gamma_i| \ell_i \\ &\leq \frac{2\beta - 1}{\beta} \sum_{i \in S^o \setminus S} \sum_{j \in \gamma_i} d^1(j, \sigma(j)) \\ &\leq \frac{2\beta - 1}{\beta} \sum_{j \in C} d^1(j, \sigma(j)) \\ &= \frac{2\beta - 1}{\beta} cost_{I^2}(S, \sigma).\end{aligned}$$

This implies that

$$cost_{I^3}(S, \sigma) = cost_{I^2}(S, \sigma) + Pc_{I^3}(S, \sigma) \leq \frac{3\beta - 1}{\beta} cost_{I^2}(S, \sigma).$$

□

Lemma 4 implies that I^2 can be converted to I^3 with a constant factor loss in the approximation ratio.

Theorem 5. *Given an α_3-approximation solution of I^3, we can find an α_2-approximation solution of I^2, where $\alpha_2 = \frac{3\beta-1}{\beta}\alpha_3$.*

Proof. Let (S^{*2}, σ^{*2}) be an optimal solution of instance I^2. By Lemma 4, there exists a solution (S^3, σ^3) of I^3 such that $cost_{I^3}(S^3, \sigma^3) \leq \frac{3\beta-1}{\beta} cost_{I^2}(S^{*2}, \sigma^{*2})$. Let (S, σ) denote an α_3-approximation solution of I^3, we have $cost_{I^2}(S, \sigma) \leq cost_{I^3}(S, \sigma) \leq \frac{3\beta-1}{\beta}\alpha_3 cost_{I^2}(S^{*2}, \sigma^{*2})$. □

3.4 Constructing an Instance of Capacitated Facility Location

In this section, we show how to convert I^3 to an instance of the capacitated facility location problem (CFL). Recall that we only consider one facility for each position in the lower bounded k-median problem. For each $i \in S^o$, let $\Delta_i^1 = |\gamma_i|$ and $\Delta_i^2 = |\gamma_i| - B$, we define a variable $\Delta_i \in \{\Delta_i^1, \Delta_i^2\}$. In addition, we define P_i as the position of i for any $i \in S^o$. If $\Delta_i = \Delta_i^1$, then we close facility i. In such case, Δ_i^1 clients should be reconnected. For the case where $\Delta_i = \Delta_i^2$, we open facility i, if $\Delta_i^2 > 0$ then $|\Delta_i^2|$ clients from γ_i can be reconnected without violating the lower bound of i. Otherwise, $|\Delta_i^2|$ clients should be reconnected to i. It can be seen that instance I^3 is to identify the value of Δ_i for each P_i where $i \in S^o$ such that $\sum_{i \in S^o} \Delta_i \geq 0$.

We now show how to construct an instance I^4 of CFL based on I^3. To avoid confusion, each facility in the CFL instance is called a C-facility, and each

client in CFL instance is called a C-client. The total cost of an instance of CFL is the sum of the open cost of C-facilities and connection cost of C-clients. For each P_i where $i \in S^o$, we construct a C-facility with open cost $\frac{2\beta-1}{\beta}|\gamma_i|\ell_i$ and capacity $\Delta_i^1 - \Delta_i^2$. Moreover, a set of C-clients or a C-facility in P_i are constructed depending on the value of Δ_i^2. If $\Delta_2^i < 0$, then we construct a set of $|\Delta_i^2|$ C-clients. If $\Delta_i^2 > 0$, then we construct a C-facility with open cost 0 and capacity Δ_i^2. Note that some locations may have more than one C-facilities in I^4. Given a solution (S, σ) of I^4, let $f_{I^4}(S, \sigma)$ denote the open cost of C-facilities and $\theta_{I^4}(S, \sigma)$ denote the connection cost of C-clients.

Lemma 5. *Given any solution (S, σ) of instance I^3, we can find a solution (S^c, σ^c) of instance I^4 of CFL such that $cost_{I^4}(S^c, \sigma^c) \leq cost_{I^3}(S, \sigma)$.*

Proof. We identify the value of Δ_i for each P_i where $i \in S^o$ based on solution (S, σ). We have $\sum_{i \in S^o} \Delta_i \geq 0$ due to (S, σ) is a feasible solution of I^3. As mentioned above, in instance I^4, for each P_i where $i \in S^o$, there are $|\Delta_i^2|$ C-clients in P_i which need to be connected if $\Delta_i^2 < 0$. Otherwise we open the C-facility with open cost 0 and capacity Δ_i^2 at this position. Moreover, for each P_i where $i \in S^o \setminus S$, we open the C-facility at this position, whose open cost and capacity are $\frac{2\beta-1}{\beta}|\gamma_i|\ell_i$ and $\Delta_i^1 - \Delta_i^2$, respectively.

Now, we get a set S^c of opened C-facility for instance I^4. Note that if a position P_i where $i \in S$ is located two opened C-facilities, then in this position the total capacity is Δ_i^1. If there are $|\Delta_i^2|$ C-clients in P_i where $i \in S^o \setminus S$, then the C-facility in the same position has the priority of connecting these clients. Recall that the capacity of such a C-facility is $\Delta_i^1 - \Delta_i^2$, which implies that the $|\Delta_i^2|$ C-clients can be connected to it without violating the capacity. Thus, in instance I^4, we have $\Delta_i = \Delta_i^2$ for each P_i where $i \in S$ and $\Delta_i = \Delta_i^1$ for each P_i where $i \in S^o \setminus S$. Recall that $\sum_{i \in S^o} \Delta_i \geq 0$. Thus we can find a feasible solution for I^4 based on S^c.

Let (S^c, σ^c) denote the constructed solution of I^4, where σ^c is obtained by "switching" the direction of σ. Assume that η clients located in position P_1 are connected to a facility located in P_2 by σ for some $\eta > 0$. For the instance I^4, if there exists C-clients in P_2, then η C-clients in P_2 are connected to the C-facilities in P_1 by σ^c. Otherwise we will do nothing and this is feasible for I^4. It can be seen that the connection cost of solution (S^c, σ^c) on instance I^4 is no more than the connection cost of (S, σ) on I^3.

In solution (S, σ) to instance I^3, if a facility $i \in S^o$ is not opened, then a penalty cost $\frac{2\beta-1}{\beta}|\gamma_i|\ell_i$ should be paid. The penalty cost is equal to the open cost of a C-facility with capacity $\Delta_i^1 - \Delta_i^2$ in instance I^4. Thus, we get

$$\begin{aligned} cost_{I^4}(S^c, \sigma^c) &= \theta_{I^4}(S^c, \sigma^c) + f_{I^4}(S^c, \sigma^c) \\ &\leq \theta_{I^3}(S, \sigma) + Pc_{I^3}(S, \sigma) \\ &= cost_{I^3}(S, \sigma). \end{aligned}$$

□

Lemma 5 implies that a solution (S, σ) of I^3 can be converted to a solution (S^c, σ^c) of I^4. We now show how a solution of I^4 can be converted to a solution of I^3.

Lemma 6. *Given a solution (S^c, σ^c) of instance CFL, we can find a solution (S, σ) of instance I^3 such that $cost_{I^3}(S, \sigma) \leq \frac{2\beta}{2\beta-1} cost_{I^4}(S^c, \sigma^c)$.*

Proof. We construct a solution (S, σ) of I^3 based on solution (S^c, σ^c). Given a position, if a C-facility with open cost $\frac{2\beta-1}{\beta}|\gamma_i|\ell_i$ and capacity $\Delta_i^1 - \Delta_i^2$ in the position is opened in I^4, then no facility in the position is opened in I^3. Otherwise, the facility in the position is opened in I^3. For a position where a C-facility i with open cost $\frac{2\beta-1}{\beta}|\gamma_i|\ell_i$ and capacity $\Delta_i^1 - \Delta_i^2$ is opened in instance I^4, we have the following two cases: (1) $\Delta_i^2 > 0$, and (2) $\Delta_i^2 < 0$. For case (1), a C-facility with open cost 0 and capacity Δ_i^2 is located in the position. For case (2), there are $|\Delta_i^2|$ C-clients in the position that can be connected with the C-facilities located in the same position. In both cases, the C-facilities in the position can still be connected with Δ_i^1 C-clients. For each position where no facility is opened in instance I^3, then Δ_i^1 clients in the position should be reconnected, and the total capacity in instance I^4 is Δ_i^1. For other positions, we have $\Delta_i = \Delta_i^2$ in both instances I^3 and I^4. Since (S^c, σ^c) is feasible for I^4, we have $\sum_{i \in S^o} \Delta_i \geq 0$, which implies that we can find a feasible solution for I^3 based on S.

We now find the connection function σ for instance I^3. Such a function cannot be simply identified by "switching" the direction of σ^c. Indeed, the number of the C-clients connected with each C-facility is not guaranteed to be equal to the capacity of the C-facility. However, all the clients need to be connected in instance I^3. This implies that connecting the clients in I^3 by "switching" the direction of σ^c may cause some clients unconnected. It can be seen that such unconnected clients are located in the positions where no facility is opened in solution (S, σ) and the number of the C-clients which are connected to this position is not equal to the sum of the capacities of the C-facilities located in the same position.

For each P_i where $i \in S^o$, let δ_{P_i} be the set of the unconnected clients located in P_i. For the case where $\delta_{P_i} > 0$, we first attempt to connect each unconnected client to the nearest facility i' to i. If i' is opened, then we connect each client in δ_{P_i} to i', and the connection cost is at most $B\ell_i \leq \frac{|\gamma_i|}{\beta}\ell_i$. If i' is not opened, we further consider the following two cases: (1) $|\delta_{P_i}| + |\delta_{P_{i'}}| \geq B$, and (2) $|\delta_{P_i}| + |\delta_{P_{i'}}| < B$. For case (1), we open i' and connect each client in δ_{P_i} to i'. The condition of case (1) implies that i' can be opened without violating its lower bound. For case (2), we move each client in δ_i to i' and let $\delta_{P_{i'}} = \delta_{P_i} \cup \delta_{P_{i'}}$. We now perform the same operation described above on $P_{i'}$.

The challenge is that the procedure may be caught in several facilities, and we cannot open a facility to satisfy the lower bound. For instance, it may be the case that the clients are moved to a facility i' for more than one time, and the walk forms a cycle. Let i denote the facility in the previous position of i' in the walk. Our approach to deal with this issue is to connect these clients to the

opened facility i^o that minimizes the connection cost. We have $|\delta_{P_i}| < B$ and $|\gamma_{i'}|+|\gamma_i| \geq 2\beta B$. So there are at least $|\gamma_{i'}|+|\gamma_{i^*}|-|\delta_{P_{i^*}}| \geq (2\beta-1)B$ connected clients and we can bound the connection cost as

$$\sum_{j\in\gamma_{i'}\cup\gamma_{i\setminus\delta_{P_i}}} d^1(j,\sigma(j)) \geq (2\beta-1)Bd^1(i^o,\{i',i\}). \tag{3}$$

Using triangle inequality and inequality (3), we have

$$\begin{aligned}\sum_{j\in\delta_{P_i}} d^1(j,i^o) &\leq Bd^1(i,i^o) \leq B(d^1(i^o,\{i',i\})+\ell_i)\\ &\leq \frac{1}{2\beta-1}\sum_{j\in\gamma_{i'}\cup\gamma_{i\setminus\delta_{P_i}}} d^1(j,\sigma(j)) + \frac{\ell_i|\gamma_i|}{\beta},\end{aligned}$$

which implies that the total increased cost induced by the unconnected clients is no more than

$$\frac{1}{2\beta-1}Pc_{I^3}(S,\sigma)+\frac{1}{2\beta-1}\theta_{I^3}(S,\sigma).$$

Thus, we have

$$\begin{aligned}cost_{I^3}(S,\sigma) &\leq \theta_{I^3}(S,\sigma)+Pc_{I^3}(S,\sigma)+\frac{1}{2\beta-1}Pc_{I^3}(S,\sigma)+\frac{1}{2\beta-1}\theta_{I^3}(S,\sigma)\\ &= \frac{2\beta}{2\beta-1}Pc_{I^3}(S,\sigma)+\frac{2\beta}{2\beta-1}\theta_{I^3}(S,\sigma)\\ &= \frac{2\beta}{2\beta-1}cost_{CFL}(S^c,\sigma^c).\end{aligned}$$

□

Theorem 6. *Given an α_4-approximation solution of CFL, we can find an α_3-approximation solution of I^3, where $\alpha_3 = \frac{2\beta}{2\beta-1}\alpha_4$.*

Proof. Let (S^{*3},σ^{*3}) be an optimal solution of instance I^3. Using Lemma 5, there exists a solution (S^c,σ^c) of I^4 such that $cost_{I^4}(S^c,\sigma^c) \leq cost_{I^3}(S^{*3},\sigma^{*3})$. Given an α_4-approximation solution (S',σ') of I^4, we have $cost_{I^4}(S',\sigma') \leq \alpha_{I^4}cost_{I^3}(S^{*3},\sigma^{*3})$. By Lemma 6, we can get a solution (S,σ) of I^3 that satisfies $cost_{I^3}(S,\sigma) \leq \frac{2\beta}{2\beta-1}cost_{I^4}(S',\sigma') \leq \frac{2\beta}{2\beta-1}\alpha_4 cost_{I^3}(S^{*3},\sigma^{*3})$.

3.5 Combining Everything

Using the algorithm for the capacitated facility location problem given in [1], we get a $(1+\sqrt{2})$-approximation solution of I^4. Let $\beta = \frac{2}{3}$. By Theorem 6, we get $\alpha_3 = \frac{2\beta}{2\beta-1}(1+\sqrt{2}) = 4(1+\sqrt{2})$. By Theorems 5, 4, and 3, we get $\alpha_2 = \frac{3\beta-1}{\beta}\alpha_3 = \frac{3\beta-1}{\beta}\times 4(1+\sqrt{2}) = 6(1+\sqrt{2})$, $\alpha_1 = 2\alpha_2 = 2\times 6(1+\sqrt{2}) = 12(1+\sqrt{2})$, $\alpha = \alpha_1(1+3.25\frac{1+\beta}{1-\beta})+3.25\frac{1+\beta}{1-\beta} = 12(1+\sqrt{2})\times(1+3.25\frac{1+\beta}{1-\beta})+3.25\frac{1+\beta}{1-\beta} \approx 516$.

References

1. Ahmadian, S., Swamy, C.: Improved approximation guarantees for lower-bounded facility location. In: Erlebach, T., Persiano, G. (eds.) WAOA 2012. LNCS, vol. 7846, pp. 257–271. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38016-7_21
2. Ahmadian, S., Swamy, C.: Approximation algorithms for clustering problems with lower bounds and outliers. In: Proceedings of the 43rd International Colloquium on Automata, Languages, and Programming, pp. 69:1–69:15 (2016)
3. Bansal, M., Garg, N., Gupta, N.: A 5-approximation for capacitated facility location. In: Epstein, L., Ferragina, P. (eds.) ESA 2012. LNCS, vol. 7501, pp. 133–144. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33090-2_13
4. Bhattacharya, A., Jaiswal, R., Kumar, A.: Faster Algorithms for the Constrained k-means Problem. Theory of Comput. Syst. **62**(1), 93–115 (2017). https://doi.org/10.1007/s00224-017-9820-7
5. Byrka, J., Pensyl, T., Rybicki, B., Srinivasan, A., Trinh, K.: An improved approximation for k-median and positive correlation in budgeted optimization. ACM Trans. Algorithms **13**(2), 23:1–23:31 (2017)
6. Charikar, M., Li, S.: A dependent LP-rounding approach for the k-median problem. In: Czumaj, A., Mehlhorn, K., Pitts, A., Wattenhofer, R. (eds.) ICALP 2012. LNCS, vol. 7391, pp. 194–205. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31594-7_17
7. Cohen-Addad, V., Klein, P.N., Mathieu, C.: Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics. In: Proceedings of the 57th IEEE Symposium on Foundations of Computer Science, pp. 353–364 (2016)
8. Demirci, H.G., Li, S.: Constant approximation for capacitated k-median with $(1+\epsilon)$-capacity violation. In: Proceedings of the 43rd International Colloquium on Automata, Languages, and Programming, pp. 73:1–73:14 (2016)
9. Ding, H., Xu, J.: A unified framework for clustering constrained data without locality property. In: Proc. 26th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1471–1490 (2015)
10. Guha, S., Meyerson, A., Munagala, K.: Hierarchical placement and network design problems. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pp. 603–612 (2000)
11. Jain, K., Vazirani, V.V.: Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. J. ACM **48**(2), 274–296 (2001)
12. Karger, D.R., Minkoff, M.: Building Steiner trees with incomplete global knowledge. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pp. 613–623 (2000)
13. Kumar, A., Sabharwal, Y., Sen, S.: Linear-time approximation schemes for clustering problems in any dimensions. J. ACM **57**(2), 1–32 (2010)
14. Li, S.: Approximating capacitated k-median with $(1+\epsilon)k$ open facilities. In: Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 786–796 (2016)

15. Li, S.: On facility location with general lower bounds. In: Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 2279–2290 (2019)
16. Li, S., Svensson, O.: Approximating k-median via pseudo-approximation. SIAM J. Comput. **45**(2), 530–547 (2016)
17. Svitkina, Z.: Lower-bounded facility location. ACM Trans. Algorithms **6**(4), 69:1–69:16 (2010)