

A Constant Factor Approximation Algorithm for k -Median Clustering with Outliers*

Ke Chen[†]

October 26, 2007

Abstract

We consider the *k-median clustering with outliers problem*: Given a finite point set in a metric space and parameters k and m , we want to remove m points (called outliers), such that the cost of the optimal k -median clustering of the remaining points is minimized. We present the first polynomial time constant factor approximation algorithm for this problem.

1 Introduction

Clustering is the process of classifying a set of objects into groups such that objects in each group are similar. One widely studied variant of clustering is the *k-median* problem. Here we are given a set of points (in a metric space), and we wish to choose at most k points as *medians* (or facilities), so as to minimize the total distance of connecting each point to its closest median.

We consider the *k-median with outliers problem* (MO for short): Given parameters k and m , we wish to remove a set of at most m points (called *outliers*) from the data set, such that the cost of the optimal k -median clustering of the remaining data is minimized. The problem was considered by Charikar *et al.* [CKMN01], and they presented a bi-criteria approximation algorithm for this problem. In particular, their algorithm computes a solution with at most $(1 + \lambda)m$ outliers that costs at most $4(1 + 1/\lambda)OPT$, where OPT is the cost of the optimal solution and $\lambda > 0$ is an arbitrary parameter specified in advance.

This problem arises naturally in situations where noise and errors contained in the data may exert a strong influence over the optimal clustering cost. By removing outliers, one can dramatically reduce the clustering cost and improve the quality of the clustering. In some circumstances, the discovered outliers do not fit the rest of the data, and they are worthy of further investigation. In particular, once identified, they can be used to discover anomalies in the data [RRPS04].

Besides the practical considerations mentioned above, the problem is theoretically interesting. Since the first constant factor approximation algorithm for the k -median problem in metric spaces [CGTS02], there have been numerous developments on this problem and its variants. However, it remains elusive how to design constant factor approximation algorithms for k -median variants that have more than one global constraint. (Indeed, sometimes adding one more global constraint to an optimization problem makes it considerably harder than the original problem. For example, computing the minimum spanning tree is easy, while finding efficient approximation algorithms

*A preliminary version of this paper is to appear in *Proc. 19th ACM-SIAM Sympos. Discrete Algorithms*.

[†]Department of Computer Science; University of Illinois at Urbana-Champaign; 201 N. Goodwin Avenue; Urbana, IL, 61801, USA; kechen@uiuc.edu; <http://www.uiuc.edu/~kechen/>. Work on this paper was partially supported by a NSF award CCR-0132901.

for k -MST took decades, and the status of bounded degree MST problem has yet to be completely settled [Goe06, SL07].) In particular, as a notable example of such problems, the k -median with outliers problem (MO) has two global constraints imposed by k and m . The MO problem has received considerable interest recently, and coming up with a constant-factor approximation algorithm is a well known open problem, see the discussions in [JV01, CKMN01, Khu05].

Related work. We focus on the most related work here, for further information see [CR05] and references therein. The k -median problem is shown to be NP-hard by a reduction from dominating set [LV92]. The first constant factor approximation algorithm for the k -median problem is based on filtering and LP-rounding ideas [CGTS02]. Jain and Vazirani [JV01] gave an algorithm based on primal-dual schema and Lagrange-relaxation technique. The current best approximation guarantee for this problem is $(3 + \varepsilon)$, and it is based on local search [AGK⁺04].

The *facility location with outliers problem* (FLO for short) is the Lagrangian relaxation of the k -median with outliers problem (MO). Several algorithms [CKMN01, JMM⁺03, Mah04] were developed for FLO. Other variants on clustering with outliers include the work of Aboud and Rabani [AR06], which provides an approximation algorithm for a variant of correlation clustering with outliers.

The (uniform) capacitated k -median problem is a k -median variant which have two global constraints. Here we are allowed to open k medians but there is an upper bound on the number of data points each median can serve. There are several bi-criteria approximation algorithms for this problem [CGTS02, BCR01, CR05].

Local search is a popular technique for solving combinatorial optimization problems in practice. Despite their conceptual simplicity, local search algorithms tend to be hard to analyse. It is successfully applied in various facility location problems [KPR00, CG99, AGK⁺04, ST06] and to the k -median problem [AGK⁺04]. For some other approximation algorithms that use local search, see [AH98, KR00, KBP03].

Our contribution. We present the first efficient constant factor approximation algorithm for the k -median with outliers problem (MO). Our algorithm is built upon the Lagrangian relaxation framework outlined in [JV01]. It first computes two solutions C_- and C_+ for the *facility location with outliers problem* (FLO), which is the Lagrangian relaxation of MO. Here, C_- has at most k centers and C_+ has at least $k + 1$ centers.

In Section 3.1, we combine C_- and C_+ into the required approximate solution, when C_+ uses at least $k + 2$ facilities. The challenge is to merge a solution with few centers (C_-) which might be too expensive and a solution (C_+) that has too many facilities but is relatively cheap. To confound the difficulty in this “merging” stage, the outliers in these two solutions are not necessarily the same. To perform this “merge”, we employ a different greedy algorithm, rather than using the augmentation approaches used in previous approximation algorithms for the k -median problem [JV01, CG99].

We use *successive local search*, in Section 3.2, to obtain a constant factor approximation algorithm for MO when C_+ uses $k + 1$ facilities. In this case, the cost of C_- cannot be bounded directly by the cost of the optimal solution, and as a result, combining C_- and C_+ into a single solution (as done in previous works [JV01, CG99] and in Section 3.1) is no longer viable. To circumvent this difficulty, we use a local search algorithm for the *penalty k -median with outliers problem* (PMO for short) as a subroutine, with gradually increased penalty parameters. Instead of directly bounding the cost of a locally optimal solution for PMO, we bound the number of points that receive penalty in the solution.

The use of successive local search, in Section 3.2, is new and we consider the introduction of this technique and its analysis to be the main technical contribution of this paper. Interestingly,

neither PMO nor MO can be solved by applying the standard local search methods directly (see Appendix A). Thus, the new technique seems to be required if one wants to use local search paradigm to solve this problem. Those structural difficulties might explain the challenge in solving this problem, and the complexity of the analysis of our algorithm.

The rest of the paper is organized as follows. In Section 3, we present the algorithm. In Section 4, we provide the intuition why the algorithm works, and prove some key properties. In Section 5, we prove the correctness of the algorithm for the case $|C_+| \geq k + 2$. In Section 6, we prove the correctness for the case $|C_-| = k + 1$. We conclude in Section 7.

2 Preliminaries

We slightly abuse notations and refer to multisets as sets. Given a set X , the notation $|X|$ refers to the total size of X (i.e., an element with weight w in X contributes w to $|X|$).

We are given a metric space with a distance function $d(\cdot, \cdot)$ defined over it. We make the standard assumption that we can compute $d(p, q)$, for any p and q , in constant time. A point may be selected to be a *facility*, which *serves* the points that are connected to it. The cost of assigning (or, connecting) a set of points V to a facility q is $\nu(q, V) = \sum_{p \in V} d(q, p)$. The cost of assigning a set V to a set C of facilities is $\nu(C, V) = \sum_{p \in V} d(C, p)$, where $d(C, p) = \min_{q \in C} d(q, p)$.

Given a set V of n points and a set C of facilities, let $\mathbf{N}_{n-m}(C, V)$ be the set of $n - m$ points in V nearest to C . Let

$$A_m(C, V) = \nu(C, \mathbf{N}_{n-m}(C, V))$$

be the cost of connecting V to C while excluding the most “expensive” m points from consideration (those m excluded points are the outliers).

Definition 2.1 (k -median with m outliers.) Let $\text{MO}(k, V, m)$ be an instance of the k -median with m outliers problem, consisting of an integer $k \geq 1$, a set V of n points, and $m \geq 0$. The objective of $\text{MO}(k, V, m)$ is to compute a set $C \subseteq V$ of k points minimizing the cost $A_m(C, V)$. Let $\text{opt}_{\text{mo}}(k, V, m)$ denote the cost of the optimal solution.

In the remainder of the paper, we consider the problem instance $\text{MO}(k, P, m)$, where P is a given set of n points. For technical reasons, we assume that the distances between all pairs of points in P are distinct, and the spread of P is polynomially bounded, in particular, $d_{\max}/d_{\min} = O(n^2)$, where d_{\max} and d_{\min} are the maximal and minimal inter-point distances in P , respectively. One can slightly perturb the distance function d so that it fulfills those requirements. See Appendix D for details.

2.1 The Lagrangian approach

The following is a Lagrangian relaxation of the k -median with m outliers problem (MO).

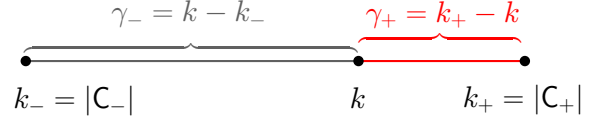
Definition 2.2 (Facility location with m outliers.) Let $\text{FLO}(z, V, m)$ be an instance of facility location with m outliers, consisting of a parameter $z \geq 0$, a set V of points, and an integer $m \geq 0$. The objective of $\text{FLO}(z, V, m)$ is to compute a set $C \subseteq V$ minimizing the cost $A_m(C, V) + z|C|$. Let $\text{opt}_{\text{flo}}(z, V, m)$ denote the cost of the optimal solution.

Theorem 2.3 ([CKMN01, Mah04]) *Given a set V of points and $z \geq 0$, one can compute a facility set $C \subseteq V$ such that $A_m(C, V) + 3z(|C| - 1) \leq 3\text{opt}_{\text{flo}}(z, V, m)$.*

Let FLOALG denote the algorithm provided for FLO by Charikar *et al.* [CKMN01]. (In fact, the constant approximation factors provided by the algorithm of Mahdian [Mah04] are slightly better, but this does not affect our results substantially.)

Consider $\text{FLO}(z, \mathbf{P}, m)$. When $z = 0$, the algorithm FLOALG opens all the facilities, and when $z = nd_{max}$, it opens only a single facility. We perform a binary search on the interval $[0, nd_{max}]$ to find z_- and z_+ such that the algorithm opens $k_- \leq k$ and $k_+ \geq k + 1$ facilities for $\text{FLO}(z_-, \mathbf{P}, m)$ and $\text{FLO}(z_+, \mathbf{P}, m)$, respectively, and moreover, $|z_- - z_+| \leq d_{min}/n^2$ (this can be done in $O(\log n)$ steps, since the spread of \mathbf{P} is polynomially bounded). Let C_- and C_+ be the facility sets computed by the algorithm for z_- and z_+ , respectively. Here $|C_-| = k_-$ and $|C_+| = k_+$.

Let $\gamma_- = k - k_-$ and $\gamma_+ = k_+ - k$. We have $\gamma_- \geq 0$ and $\gamma_+ \geq 1$, since $k_- \leq k$ and $k_+ \geq k + 1$. Also, we have $\gamma_- + \gamma_+ = k_+ - k_-$.



2.2 A modified point set \mathbf{P}^w

Let $M_+ = \mathbf{N}_{n-m}(C_+, \mathbf{P})$ be the set of the $n - m$ points in \mathbf{P} closest to C_+ .

Definition 2.4 (The weight function w .) A point p is *heavy* if $p \in C_+$. Its weight, denoted by $w(p)$, is the number of points in M_+ served by p . Given two heavy points p and q , if $w(p) > w(q)$ then p is *heavier* than q . A point p is *light* if $p \in \mathbf{P} \setminus M_+$ (that is, p is one of the m outliers in the solution induced by C_+). A light point has weight one. The points in $M_+ \setminus C_+$ have weight zero, and they are neither heavy nor light.

For a set $Q \subseteq \mathbf{P}$ of points, let $w(Q) = \sum_{p \in Q} w(p)$.

Definition 2.5 (The multiset \mathbf{P}^w .) Given a point set $Q \subseteq \mathbf{P}$, the notation Q^w refers to a *multiset* where each point $p \in Q$ occurs $w(p)$ times (note that $p \notin Q^w$ if $w(p) = 0$). We will abuse our notation slightly, and for a given point $q \in C_+$, denote by q^w the set $\{q\}^w$, which is a multiset where q appears $w(q)$ times.

In particular, the multiset \mathbf{P}^w consists of all the heavy points and the m light points, where the weight of each heavy point q is $w(q)$ and the weight of each light point is one. Observe that the *size* of \mathbf{P}^w is $|\mathbf{P}^w| = w(\mathbf{P}) = n$, and the number of distinct points in \mathbf{P}^w is $k_+ + m$.

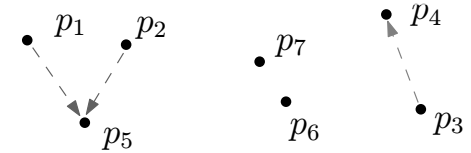


Figure 1: $\mathbf{P} = \{p_1, \dots, p_7\}$, $m = 2$, and $C_+ = \{p_4, p_5\}$. We have $w(p_4) = 2$, $w(p_5) = 3$, $w(p_6) = w(p_7) = 1$, and $w(p_1) = w(p_2) = w(p_3) = 0$. Therefore, $\mathbf{P}^w = \{p_4, p_4, p_5, p_5, p_5, p_6, p_7\}$. The points p_4 and p_5 are heavy, and p_6 and p_7 are light.

Note that \mathbf{P}^w is the set $C_+ \cup (\mathbf{P} \setminus M_+)$ with appropriate weights associated with the points. The multiset \mathbf{P}^w can be thought of as a *coreset* of \mathbf{P} , which is roughly a coarse representation of the original set \mathbf{P} . (The interested reader is referred to [HM04] for definition.)

Definition 2.6 (Include, exclude, and partly-include.) Given a heavy point p and a set $Q \subseteq \mathbf{P}^w$, if p occurs $w(p)$ times in Q then it *includes* p , if p does not appear in Q then it *excludes* p , and otherwise, it *partly-includes* p .

Definition 2.7 (The set \mathcal{X} .) Let \mathcal{X}' be the set of $n - m$ points in \mathbf{P}^w closest to C_- . Since all distances (between distinct points) in \mathbf{P}^w are distinct, there might be (only) one heavy point, say

q , which is partly-included in \mathcal{X}' . In this case, we remove all copies of q from \mathcal{X}' and let \mathcal{X} be the resulting set, otherwise, set $\mathcal{X} = \mathcal{X}'$.

For a set $B \subseteq \mathbb{P}^w$, let $h_w(B)$ denote the number of distinct heavy points in B , and $l_w(B)$ denote the number of light points in B (note that each light point appears exactly once in \mathbb{P}^w , and as such the light points are distinct).

Definition 2.8 (Mass, cost, and benefit.) If $l_w(\mathcal{X}) = 0$ then let $\xi = 0$. Otherwise, let

$$\xi = \frac{k_+ - h_w(\mathcal{X}) - 1}{l_w(\mathcal{X})}. \quad (1)$$

For a point $p \in \mathcal{X}$, let $\text{cost}(p) = \nu(\mathbb{C}_-, p)$. The *mass* of p , denoted by $\text{mass}(p)$, is ξ if p is light, and $1/w(p)$ otherwise. For a set $B \subseteq \mathcal{X}$ of points, let $\text{mass}(B) = \sum_{p \in B} \text{mass}(p)$ and $\text{cost}(B) = \sum_{p \in B} \text{cost}(p)$, and the *benefit* of B is $\text{ben}(B) = \text{mass}(B) - 1$.

2.3 The local search method

We shall reduce MO to the penalty k -median with m outliers problem (PMO), which is defined below, and apply the local search method to PMO.

In the PMO problem, we are allowed to exclude more than m outliers, but every such additional outlier incurs a penalty. Equivalently, given a set V of n points, a set C of facilities, and a penalty parameter $\varrho > 0$, let

$$\mathcal{A}_m(C, V, \varrho) = \sum_{p \in \mathbf{N}_{n-m}(C, V)} \min(\text{d}(C, p), \varrho)$$

denote the cost of PMO clustering V with m outliers and penalty ϱ , where $\mathbf{N}_{n-m}(C, V)$ is the set of $n - m$ points in V closest to C . Namely, we assign $\mathbf{N}_{n-m}(C, V)$ to C . Every point $p \in \mathbf{N}_{n-m}(C, V)$ pays a connection cost, which is the distance $\text{d}(C, p)$ capped by the *penalty* ϱ .

Definition 2.9 (Penalty k -median with m outliers.) Let $\text{PMO}(k, V, \varrho, m)$ denote an instance of penalty k -median with m outliers, consisting of an integer $k \geq 1$, a set V of points, a penalty parameter $\varrho > 0$, and $m \geq 0$. The objective is to compute a set $C \subseteq V$ of k facilities minimizing the cost $\mathcal{A}_m(C, V, \varrho)$. Let $\text{opt}_{\text{pmo}}(k, V, \varrho, m)$ denote the cost of the optimal solution.

Observe that the problem $\text{PMO}(k, V, \varrho, m)$ is a relaxation of $\text{MO}(k, V, m)$. In particular, for $\varrho = \infty$, we have $\mathcal{A}_m(C, V, \varrho) = \mathbf{A}_m(C, V)$.

Definition 2.10 (Neighbor facility sets.) Given a set $C \subseteq \mathbb{P}^w$ of k facilities, let

$$\mathbf{N}(C) = \{C\} \cup \{C - q' + q'' \mid q' \in C, q'' \in \mathbb{P}^w \setminus C\}$$

denote the *neighbor facility sets* of C , where $C - q' + q'' = (C \setminus \{q'\}) \cup \{q''\}$.

Definition 2.11 (The sets \mathbf{H} and \mathcal{H} .) Recall that there are $|\mathbb{C}_+| = k_+ \geq k + 1$ heavy points in \mathbb{P}^w . Let \mathbf{H} consists of the k heaviest among them, and

$$\mathcal{H} = \{C \mid C \subseteq \mathbb{P}^w, C \text{ contains at least } k - 1 \text{ heavy points, and } |C| = k\}.$$

$\mathbf{N}_{n-m}(C, V)$	The set of $n - m$ points in V closest to C .
$\nu(C, V)$	Cost of connecting the points in V to their nearest facilities in C .
$A_m(C, V)$	$\nu(C, M)$, where M consists of the $n - m$ points in V closest to C .
$\text{MO}(k, V, m)$	An instance of k -median with m outliers, with objective to compute $C \subseteq V$ minimizing $A_m(C, V)$.
$\mathcal{A}_m(C, V, \varrho)$	Cost of connecting M to C , where M consists of the $n - m$ points in V closest to C , and each point $p \in M$ pays a cost of $\min(\varrho, d(C, p))$.
$\text{PMO}(k, V, \varrho, m)$	An instance of penalty k -median with m outliers, with objective to compute $C \subseteq V$ minimizing $\mathcal{A}_m(C, V, \varrho)$.
$\text{opt}_{\text{mo}}(k, V, m)$	the cost of the optimal solution to $\text{MO}(k, V, m)$.
$\text{opt}_{\text{pmo}}(k, V, \varrho, m)$	the cost of the optimal solution to $\text{PMO}(k, V, \varrho, m)$.
opt	$\text{opt}_{\text{mo}}(k, \mathbf{P}, m)$, the cost of the optimal solution to $\text{MO}(k, \mathbf{P}, m)$.
opt^{w}	$\text{opt}_{\text{mo}}(k, \mathbf{P}^{\text{w}}, m)$, the cost of the optimal solution to $\text{MO}(k, \mathbf{P}^{\text{w}}, m)$.

Figure 2: Notations.

3 The algorithm

The input is the set \mathbf{P} and parameters k and m . The algorithm uses binary search over the range $[0, nd_{\max}]$ to find z_- and z_+ such that $|z_- - z_+| \leq d_{\min}/n^2$, and the sets $\mathbf{C}_- = \text{FLOALG}(z_-, \mathbf{P}, m)$ and $\mathbf{C}_+ = \text{FLOALG}(z_+, \mathbf{P}, m)$ satisfy $|\mathbf{C}_-| \leq k$ and $|\mathbf{C}_+| \geq k + 1$. (Here, FLOALG is used to make the decision in the binary search.) See Section 2.1 for details. Next, it computes a multiset \mathbf{P}^{w} by collapsing the clusters (of \mathbf{P}) induced by \mathbf{C}_+ into their respective facilities, see Section 2.2. The algorithm checks if $\gamma_+ = k_+ - k \geq 2$, and if so, it uses GREEDYMERGE, described below in Section 3.1, to compute the desired solution \mathbf{C} . Otherwise, $\gamma_+ = 1$, and the algorithm uses SUCCESSIVELS, described in Section 3.2.

3.1 The algorithm GreedyMerge for the case $\gamma_+ \geq 2$

We shall compute a set $\mathbf{C} \subseteq \mathbf{C}_- \cup \mathbf{C}_+$ such that $|\mathbf{C}| = k$ and it is the required solution.

Suppose that $\mathbf{C}_- = \{f_1, \dots, f_{k_-}\}$, and let \mathcal{X}_i be the set of points of \mathcal{X} that are nearest to f_i , for $i = 1, \dots, k_-$. Assume, without loss of generality, that $\text{ben}(\mathcal{X}_1), \dots, \text{ben}(\mathcal{X}_\alpha) > 0$ and $\text{ben}(\mathcal{X}_{\alpha+1}), \dots, \text{ben}(\mathcal{X}_{k_-}) \leq 0$, for some $1 \leq \alpha \leq k_-$, and furthermore,

$$\frac{\text{cost}(\mathcal{X}_1)}{\text{ben}(\mathcal{X}_1)} \leq \dots \leq \frac{\text{cost}(\mathcal{X}_\alpha)}{\text{ben}(\mathcal{X}_\alpha)}.$$

Let k' be the index satisfying $\sum_{t=1}^{k'-1} \text{ben}(\mathcal{X}_t) < \gamma_+ \leq \sum_{t=1}^{k'} \text{ben}(\mathcal{X}_t)$, where $k' \leq \alpha$. Construct a set \mathbf{C} of k facilities as follows.

- (i) Let $\bar{\mathcal{Y}} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_{k'-1}, \mathcal{Y}_{k'}\}$. The set $\mathcal{Y}_{k'}$ is generated greedily from $\mathcal{X}_{k'}$ by repeatedly picking the point p in $\mathcal{X}_{k'}$ (that has not been added yet) with the smallest $\text{cost}(p)/\text{mass}(p)$ value. Here, if p is heavy, we add in all its copies. We repeat this till

$$\text{BEN}(\bar{\mathcal{Y}}) = \sum_{B \in \bar{\mathcal{Y}}} \text{ben}(B) \in [\gamma_+, \gamma_+ + 1) \quad (2)$$

for the first time.

```

Algorithm SUCCESSIVELS( $k, P^w, m$ )
 $i \leftarrow 0$ 
 $\varrho_0 \leftarrow d_{min}/10$ 
 $B_0 \leftarrow H$ 
 $\Delta_0 \leftarrow \Delta(B_0, P^w, \varrho_0, m)$ .
while  $\Delta_i > 0$  do
     $i \leftarrow i + 1$ 
     $\varrho_i \leftarrow 3\varrho_{i-1}$ 
     $B_i \leftarrow \text{LOCALSEARCH}(B_{i-1}, P^w, \varrho_i)$ 
     $\Delta_i \leftarrow \Delta(B_i, P^w, \varrho_i, m)$ .
 $\mathcal{X} \leftarrow H \cup \bigcup_{t=0}^i N(B_t)$ 
return  $\text{argmin}_{C \in \mathcal{X}} A_m(C, P^w)$ .

```

(a)

```

Algorithm LOCALSEARCH( $B, P^w, \varrho$ )
while  $\exists B' \in N(B) \cup \{H\}$  such that
     $\mathcal{Z}(B') < \mathcal{Z}(B) - \frac{\varrho}{3}$  do
     $B \leftarrow B'$ 
return  $B$ 

```

(b)

Figure 3: (a) A successive local search algorithm for $\text{MO}(k, P^w, m)$. Here, $\Delta(B_i, P^w, \varrho_i, m)$ is the number of points that pay the penalty ϱ_i in the PMO clustering induced by B_i . Formally, this is the number of points in $N_{n-m}(B_i, P^w)$ that are in distance $> \varrho_i$ from B_i , see Section 2.3. (b) Here, $\mathcal{Z}(B')$ refers to $A_m(B', P^w, \varrho)$, and $\mathcal{Z}(B)$ refers to $A_m(B, P^w, \varrho)$.

(ii) Let $J \subseteq C_+$ be the set of $k - |\bar{y}| = k - k'$ heaviest points not included in $\mathcal{Y} = \bigcup \bar{y}$.

(iii) Return $C = \{f_1, \dots, f_{k'}\} \cup J$.

3.2 The algorithm SuccessiveLS for the case $\gamma_+ = 1$

The algorithm $\text{SUCCESSIVELS}(k, P^w, m)$ is presented in Figure 3 (a). Its input consists of the point set P^w , C_+ , and integers k and m , and it returns the desired approximation. The procedure SUCCESSIVELS uses LOCALSEARCH , depicted in Figure 3 (b). Here, the set C_- is not used by the algorithm, and C_+ is used to derive the sets H and \mathcal{H} , see Definition 2.11.

Intuitively, SUCCESSIVELS works by generating a set of candidate facility sets, among which at least one is more expensive than the optimal solution by only a constant factor. Therefore, the cheapest solution among the candidates generated provides the required approximation.

3.3 The result

We have the following result.

Theorem 3.1 *Given a set P of n points, integral parameters $k \geq 1$ and $m \geq 0$, one can compute, in $O(k^2(k+m)^2n^3 \log n)$ time, a set $C \subseteq P$ of k facilities such that $A_m(C, P) = O(\text{opt})$, where $\text{opt} = \text{opt}_{\text{mo}}(k, P, m)$.*

The rest of the paper is dedicated to proving Theorem 3.1. In particular, it is implied by Lemma 5.10 and Lemma 6.8.

4 Intuition and Correctness

4.1 Intuition

We handle the two cases $\gamma_+ \geq 2$ and $\gamma_+ = 1$ separately, because a key claim (see Claim 4.4) used in bounding the cost of C_- works only for the case $\gamma_+ \geq 2$, see also Remark 4.5. Moreover, the analysis of the local search method does not hold in the case $\gamma_+ \geq 2$, see Lemma 6.5.

Intuition for GreedyMerge ($\gamma_+ \geq 2$). In the clustering of P^w induced by C_+ , every heavy point itself is a cluster (recall that the total weight of heavy points is $n - m$). GREEDYMERGE needs to “pack” these k_+ clusters (i.e., heavy points) into k clusters, with the help of C_- . Note that $\mathcal{X}_1, \dots, \mathcal{X}_{k_-}$ are the k_- clusters in the clustering of \mathcal{X} induced by C_- , and intuitively, consider $\mathcal{X}_1, \dots, \mathcal{X}_{k_-}$ as a MO clustering of P^w (recall that $|\mathcal{X}|$ is roughly $n - m$). To do the packing, we assign a mass of one to (all copies of) each heavy point. Intuitively, the mass of \mathcal{X}_i is the (fractional) number of heavy points in \mathcal{X}_i . The mass of \mathcal{X}_i may be fractional, since it might contain light points. The mass of a light point p (i.e., ξ) is the fraction of the heavy points that are “ejected” from \mathcal{X} because of p (if p is included in \mathcal{X} , then some heavy points must have been excluded by \mathcal{X}). Naturally, we would like to use \mathcal{X}_i with maximum mass, since it packs the largest number of (fractional) heavy points into a single new cluster. In fact, a cluster \mathcal{X}_i with mass one or less does not help us in this merging process (since \mathcal{X}_i would use one facility on its own). In particular, we are mainly interested in the (added) benefit of \mathcal{X}_i , namely $\text{ben}(\mathcal{X}_i) = \text{mass}(\mathcal{X}_i) - 1$. Furthermore, great benefit with prohibitive cost is of little use for us. As such, we sort the \mathcal{X}_i s by their *return*, namely $\text{cost}(\mathcal{X}_i) / \text{ben}(\mathcal{X}_i)$. Next, we pick as many of them as necessary so that we can add the remaining (uncovered) heavy points as clusters to the solution, and still use only k facilities.

Intuition for SuccessiveLS ($\gamma_+ = 1$). Here, we reduce the k -median with m outliers problem (MO) to the penalty k -median with m outliers (PMO). The objective of MO is to compute C minimizing $A_m(C, P^w)$, while PMO aims to minimize $\mathcal{A}_m(C, P^w, \varrho)$. Observe that those two cost functions are the same when the penalty parameter ϱ is sufficiently large. Therefore, if we can obtain a constant factor approximation solution for PMO (with a large penalty parameter), then we are done (because it is also a constant factor approximation for MO). Furthermore, when the penalty is small enough (i.e., less than the minimal inter-point distance), the optimal solution to PMO is easy to compute — it is just H , the set of the k heaviest points in P^w . Now, we start with a (very) small penalty parameter, and gradually increase the penalty parameter by “doubling” it in each round. Because the penalty parameter increases “slowly”, and the solution computed from each round is used as the starting point for the next round, we argue that the solution of LOCALSEARCH tracks the optimal solution cost. This implies that, when the penalty parameter becomes large enough, we have the required approximation. More formally, let $\bar{\omega}_i$ be the cost of the optimal solution to PMO in the i th round, and let ω_i be the cost of the corresponding LOCALSEARCH solution (in the same round). Roughly, since $\omega_i - \omega_{i-1} = O(\bar{\omega}_i - \bar{\omega}_{i-1})$, for every $i \geq 1$, we have $\omega_i = O(\bar{\omega}_i)$. In particular, for i sufficiently large, we obtain the required approximation.

4.2 Correctness

Observation 4.1 *Let V be a set of n points, $C \subseteq V$ be a set of facilities, and M' be a set of at least $n - m$ points in V . It holds that $A_m(C, V) \leq \nu(C, M')$.*

Lemma 4.2 *Given a set V of points and non-negative parameters m and z , let C be the facility set computed by FLOALG for $FLO(z, V, m)$. It holds that, for any $k \geq 1$,*

$$A_m(C, V) \leq 3\text{opt}_{\text{mo}}(k, V, m) + 3z(k - |C| + 1).$$

Proof: We have $\text{opt}_{\text{flo}}(z, V, m) \leq \text{opt}_{\text{mo}}(k, V, m) + zk$, for any $k \geq 1$, as $\text{opt}_{\text{mo}}(k, V, m) + zk$ is the FLO cost of serving V using the k optimal facilities realizing $\text{opt}_{\text{mo}}(k, V, m)$. Now, it follows from Theorem 2.3 that

$$\begin{aligned} A_m(C, V) &\leq 3\text{opt}_{\text{flo}}(z, V, m) - 3z(|C| - 1) \leq 3(\text{opt}_{\text{mo}}(k, V, m) + zk) - 3z(|C| - 1) \\ &= 3\text{opt}_{\text{mo}}(k, V, m) + 3z(k - |C| + 1). \quad \blacksquare \end{aligned}$$

The following is motivated by the work of Jain and Vazirani [JV01] on k -median clustering. Conceptually, they merge C_- and C_+ by using the fractional solution

$$C^* = \frac{\gamma_+}{\gamma_- + \gamma_+} C_- + \frac{\gamma_-}{\gamma_- + \gamma_+} C_+. \quad (3)$$

Here, a facility in C^* is now assigned a fractional weight and the total weight of C^* is k . This provides a convex combination of the two solutions into a single solution. Next, Jain and Vazirani use a random merging procedure to realize an integral facility having (roughly) the cost of C^* (in expectation). Furthermore, the cost of C_+ is $O(OPT)$ and the cost of C_- can be bounded by $O\left(\frac{\gamma_- + \gamma_+}{\gamma_+} OPT\right)$, where OPT is the cost of the optimal solution. Plugging this into Eq. (3) yields the required approximation.

However, our situation here is more subtle, since we have different outlier sets associated with the two solutions that we need to merge. In particular, there does not seem to be an easy way to adapt their algorithm to this problem.

Claim 4.3 *We have $A_m(C_+, P) \leq 3\text{opt}$, where $\text{opt} = \text{opt}_{\text{mo}}(k, P, m)$.*

Proof: Since $\gamma_+ = k_+ - k = |C_+| - k$, it holds that, by Lemma 4.2,

$$A_m(C_+, P) \leq 3\text{opt} + 3z_+(k - |C_+| + 1) = 3\text{opt} + 3z_+(1 - \gamma_+). \quad (4)$$

Note that $z_+ \geq 0$ and $\gamma_+ \geq 1$, as such, we have $A_m(C_+, P) \leq 3\text{opt}$. ■

Claim 4.4 *If $\gamma_+ \geq 2$, then $A_m(C_-, P) \leq 9 \frac{\gamma_- + \gamma_+}{\gamma_+} \text{opt}$.*

Proof: We first bound z_+ . By Eq. (4), we have

$$3z_+(\gamma_+ - 1) \leq 3\text{opt} - A_m(C_+, P) \leq 3\text{opt},$$

which implies $z_+ \leq \frac{\text{opt}}{\gamma_+ - 1}$. Since $z_- \leq z_+ + \frac{d_{\min}}{n^2}$ and $d_{\min} \leq \text{opt}$, it follows that

$$\begin{aligned} z_-(\gamma_- + 1) &\leq \left(z_+ + \frac{d_{\min}}{n^2}\right) (\gamma_- + 1) \leq \left(\frac{\text{opt}}{\gamma_+ - 1} + \frac{\text{opt}}{n^2}\right) (\gamma_- + 1) \\ &= \left(\frac{\gamma_- + 1}{\gamma_+ - 1} + \frac{\gamma_- + 1}{n^2}\right) \text{opt} \leq \frac{\gamma_- + 2}{\gamma_+ - 1} \text{opt}, \end{aligned}$$

since $\frac{\gamma_- + 1}{n^2} \leq \frac{1}{\gamma_+ - 1}$. Now, by Lemma 4.2, we obtain

$$\begin{aligned} A_m(C_-, P) &\leq 3\text{opt} + 3z_-(k - |C_-| + 1) = 3\text{opt} + 3z_-(\gamma_- + 1) \\ &\leq \left(3 + 3\frac{\gamma_- + 2}{\gamma_+ - 1}\right)\text{opt} = 3\frac{\gamma_+ + \gamma_- + 1}{\gamma_+ - 1}\text{opt}. \end{aligned}$$

We have $\gamma_+ - 1 \geq \frac{\gamma_+}{2}$ since $\gamma_+ \geq 2$, and $\gamma_+ + \gamma_- + 1 \leq \frac{3}{2}(\gamma_+ + \gamma_-)$ since $\gamma_+ + \gamma_- \geq \gamma_+ \geq 2$. As such, $\frac{\gamma_+ + \gamma_- + 1}{\gamma_+ - 1} \leq 3\frac{\gamma_+ + \gamma_-}{\gamma_+}$, implying the claim. \blacksquare

Remark 4.5 If $\gamma_+ = 1$ then z_+ cannot be bounded by using Lemma 4.2, as done in Claim 4.4. In fact, z_+ may be arbitrarily large compared to opt in this case. As such, a similar claim to Claim 4.4 does not hold here, and the convex combination in Eq. (3)_{p9} is not necessarily a constant approximation for MO. This is the reason why we cannot apply GREEDYMERGE in this case.

If $\gamma_+ \geq 2$ and $\frac{\gamma_- + \gamma_+}{\gamma_+} = O(1)$ then, by Claim 4.4, the set C_- is the required approximation (since $|C_-| = k_- \leq k$). For example, if $k_+ \geq 2k$, then $\frac{\gamma_- + \gamma_+}{\gamma_+} \leq 2$, and as such $A_m(C_-, P) \leq 18\text{opt}$. If $\gamma_+ \geq 2$ and $\gamma_- \leq u$, for some $u \geq 0$, then we have $\frac{\gamma_- + \gamma_+}{\gamma_+} \leq 1 + u$ and as such $A_m(C_-, P) \leq (9 + 9u)\text{opt}$. In particular, for a fixed u , the solution C_- yields the required constant factor approximation. Henceforth, we assume that $k_+ < 2k$. Furthermore, if $\gamma_+ \geq 2$, then we assume that $\gamma_- > 3$.

The easy proof of the following lemma (which is implied by Claim 4.3) is delegated to Appendix B.

Lemma 4.6 (i) For $C \subseteq P$, we have that $|A_m(C, P^w) - A_m(C, P)| \leq 3\text{opt}$.

(ii) If $A_m(C, P^w) \leq \gamma \text{opt}^w$, for some $\gamma \geq 1$, then $A_m(C, P) \leq (4\gamma + 3)\text{opt}$.

The following corollary is implied by Claim 4.4 and Lemma 4.6 (i).

Corollary 4.7 If $\gamma_+ \geq 2$ then $A_m(C_-, P^w) \leq \left(3 + 9\frac{\gamma_- + \gamma_+}{\gamma_+}\right)\text{opt}$.

5 Correctness of GreedyMerge ($\gamma_+ \geq 2$)

In this section, we show that, for the case $\gamma_+ \geq 2$, GREEDYMERGE computes a solution C such that $|C| = k$ and $A_m(C, P) \leq 39\text{opt}$. Here, we assume that $\gamma_- \geq 3$, see Remark 4.5.

Let $Z = \mathcal{Y} \cup J^w$, where \mathcal{Y} and J are the sets constructed in the step (i) and step (ii) of GREEDYMERGE, respectively. The cost $\nu(C, Z)$ is equal to $\text{cost}(\mathcal{Y})$, and it is in turn $O\left(\frac{\gamma_+}{\gamma_- + \gamma_+}\text{cost}(\mathcal{X})\right)$, see Lemma 5.7 below. Moreover, Corollary 4.7 implies that $\text{cost}(\mathcal{X}) = O\left(\frac{\gamma_- + \gamma_+}{\gamma_+}\text{opt}\right)$, and combining these inequalities yields

$$\nu(C, Z) = \text{cost}(\mathcal{Y}) = O\left(\frac{\gamma_+}{\gamma_- + \gamma_+}\text{cost}(\mathcal{X})\right) = O\left(\frac{\gamma_+}{\gamma_- + \gamma_+} \cdot \frac{\gamma_- + \gamma_+}{\gamma_+}\text{opt}\right) = O(\text{opt}).$$

We are not quite done yet, as we have to argue that the size of Z is at least $n - m$, see Lemma 5.9. This claim is intuitively implied by $\text{BEN}(\bar{\mathcal{Y}}) \geq \gamma_+$ (see Eq. (2)_{p6}) but the proof is tedious, and we defer it to Appendix C.

5.1 GreedyMerge is well defined

In this section, we show that all the steps of the algorithm succeed. Indeed, Claim 5.3 below proves that k' , used in step (i) of `GREEDYMERGE`, does exist. Also, in step (i), we always have $\text{mass}(p) > 0$, as the mass of any point in \mathcal{X} is positive. In step (ii) of `GREEDYMERGE`, we have $k' \leq k_- < k$, and furthermore, Claim 5.4 below implies that at least $k - k'$ heavy points are excluded by \mathcal{Y} , thus guaranteeing that step (ii) succeeds.

Observation 5.1 (i) All heavy points are either included or excluded by \mathcal{X} .

(ii) If $l_{\mathbb{w}}(\mathcal{X}) = 0$ then $h_{\mathbb{w}}(\mathcal{X}) = k_+$, and if $l_{\mathbb{w}}(\mathcal{X}) > 0$ then $h_{\mathbb{w}}(\mathcal{X}) \leq k_+ - 1$.

(iii) We have $\xi \geq 0$, see Eq. (1)_{p5}. Moreover, for a set $B \subseteq \mathcal{X}$, we have

$$\text{mass}(B) = h_{\mathbb{w}}(B) + \xi \cdot l_{\mathbb{w}}(B). \quad (5)$$

Claim 5.2 (i) If $l_{\mathbb{w}}(\mathcal{X}) = 0$ then $\text{mass}(\mathcal{X}) = k_+$, and if $l_{\mathbb{w}}(\mathcal{X}) > 0$ then $\text{mass}(\mathcal{X}) = k_+ - 1$.

(ii) $\sum_{i=1}^{\alpha} \text{ben}(\mathcal{X}_i) \geq \sum_{i=1}^{k_-} \text{ben}(\mathcal{X}_i) \geq \gamma_- + \gamma_+ - 1$.

Proof: (i) If $l_{\mathbb{w}}(\mathcal{X}) = 0$ then, by Eq. (5), the total mass of all the points in \mathcal{X} is $\text{mass}(\mathcal{X}) = h_{\mathbb{w}}(\mathcal{X}) + \xi \cdot l_{\mathbb{w}}(\mathcal{X}) = k_+ + 0 = k_+$, by Observation 5.1 (ii). Otherwise, we have $l_{\mathbb{w}}(\mathcal{X}) > 0$, and as such,

$$\text{mass}(\mathcal{X}) = h_{\mathbb{w}}(\mathcal{X}) + \xi \cdot l_{\mathbb{w}}(\mathcal{X}) = h_{\mathbb{w}}(\mathcal{X}) + \frac{k_+ - h_{\mathbb{w}}(\mathcal{X}) - 1}{l_{\mathbb{w}}(\mathcal{X})} \cdot l_{\mathbb{w}}(\mathcal{X}) = k_+ - 1.$$

(ii) We have $\text{mass}(\mathcal{X}) \geq k_+ - 1$, by (i), and $k_+ - k_- = \gamma_- + \gamma_+$, by definition. As such,

$$\sum_{i=1}^{k_-} \text{ben}(\mathcal{X}_i) = \sum_{i=1}^{k_-} (\text{mass}(\mathcal{X}_i) - 1) = \text{mass}(\mathcal{X}) - k_- \geq k_+ - 1 - k_- = \gamma_- + \gamma_+ - 1.$$

Furthermore, since $\text{ben}(\mathcal{X}_i) \leq 0$, for $i = \alpha + 1, \dots, k_-$, we have

$$\sum_{i=1}^{\alpha} \text{ben}(\mathcal{X}_i) \geq \sum_{i=1}^{\alpha} \text{ben}(\mathcal{X}_i) + \sum_{i=\alpha+1}^{k_-} \text{ben}(\mathcal{X}_i) = \sum_{i=1}^{k_-} \text{ben}(\mathcal{X}_i). \quad \blacksquare$$

Claim 5.3 (i) There exists $k' \leq \alpha$ such that $\sum_{t=1}^{k'-1} \text{ben}(\mathcal{X}_t) < \gamma_+ \leq \sum_{t=1}^{k'} \text{ben}(\mathcal{X}_t)$.

(ii) Step (i) of `GREEDYMERGE` succeeds in computing $\mathcal{Y}_{k'}$ such that Eq. (2)_{p6} holds.

Proof: (i) By assumption, we have $\gamma_- \geq 3$, and as such, $\gamma_+ \leq \gamma_- + \gamma_+ - 1 \leq \sum_{t=1}^{\alpha} \text{ben}(\mathcal{X}_t)$, by Claim 5.2 (ii). Therefore, k' is the first index for which this sum exceeds γ_+ .

(ii) In step (i) of `GREEDYMERGE`, adding each point to $\mathcal{Y}_{k'}$ can increase the benefit of $\mathcal{Y}_{k'}$ by at most 1. This implies, by (i), that at some point, $\text{BEN}(\bar{\mathcal{Y}}) = \sum_{i=1}^{k'-1} \text{ben}(\mathcal{X}_i) + \text{ben}(\mathcal{Y}_{k'})$ will fall inside the interval $[\gamma_+, \gamma_+ + 1)$, since $\mathcal{Y}_{k'} \subseteq \mathcal{X}_{k'}$. \blacksquare

Claim 5.4 *At least $k - k'$ heavy points are not included in \mathcal{Y} . Thus, in step (ii) of GREEDYMERGE, there are enough heavy points to be included in J , namely, $h_w(J^w) = k - k'$.*

Proof: Since, by definition, $\text{mass}(B) = \text{ben}(B) + 1$, for $B \subseteq \mathcal{X}$, we have

$$\text{mass}(\mathcal{Y}) = \sum_{i=1}^{k'-1} \text{mass}(\mathcal{X}_i) + \text{mass}(\mathcal{Y}_{k'}) = \sum_{i=1}^{k'-1} \text{ben}(\mathcal{X}_i) + \text{ben}(\mathcal{Y}_{k'}) + k' = \text{BEN}(\bar{\mathcal{Y}}) + k'.$$

Now, by Eq. (2)_{p6}, which holds by Claim 5.3 (ii), this implies

$$\gamma_+ + k' \leq \text{mass}(\mathcal{Y}) < \gamma_+ + 1 + k'. \quad (6)$$

Since the mass of (all the copies) of a heavy point is one, it follows that the number of heavy points in \mathcal{Y} is strictly smaller than $\gamma_+ + 1 + k'$ (or equivalently, it is at most $\gamma_+ + k'$). Now, since the total number of heavy points is $|\mathcal{C}_+| = k_+$, it follows that at least $k_+ - (\gamma_+ + k') = k - k'$ heavy points are not included in \mathcal{Y} , as the set \mathcal{Y} does not partly-include any heavy point. ■

5.2 Bounding $\text{cost}(\mathcal{Y})$

In this section, we prove that $\text{cost}(\mathcal{Y}) = \nu(\mathcal{C}_-, \mathcal{Y}) = O(\text{cost}(\mathcal{X}))$. The following technical lemma holds, since for any four real numbers $x, y \geq 0$ and $u, v > 0$ satisfying $\frac{x}{u} \leq \frac{y}{v}$, we have $\frac{x}{u} \leq \frac{x+y}{u+v} \leq \frac{y}{v}$.

Lemma 5.5 *Given $x_1, \dots, x_c \geq 0$ and $y_1, \dots, y_c > 0$ such that $x_1/y_1 \leq \dots \leq x_c/y_c$, we have that for any $1 \leq b \leq c$ and $0 < \beta \leq 1$, it holds*

$$\frac{\sum_{t=1}^{b-1} x_t + \beta x_b}{\sum_{t=1}^{b-1} y_t + \beta y_b} \leq \frac{\sum_{t=1}^c x_t}{\sum_{t=1}^c y_t}.$$

Claim 5.6 *We have that $\text{cost}(\mathcal{Y}_{k'}) \leq \beta \text{cost}(\mathcal{X}_{k'})$, where $\beta = \frac{\text{mass}(\mathcal{Y}_{k'})}{\text{mass}(\mathcal{X}_{k'})}$.*

Proof: Observe that $0 < \beta \leq 1$. Suppose that the set $\mathcal{X}_{k'}$ consists of u distinct points, p_1, \dots, p_u , and furthermore, $\frac{\text{cost}(p_i)}{\text{mass}(p_i)} \leq \frac{\text{cost}(p_{i+1})}{\text{mass}(p_{i+1})}$, for $i = 1, \dots, u-1$. As such, $\mathcal{Y}_{k'}$ consists of $p_1, \dots, p_{u'}$, for some $u' \leq u$. By Lemma 5.5, we have

$$\frac{\text{cost}(\mathcal{Y}_{k'})}{\text{mass}(\mathcal{Y}_{k'})} = \frac{\sum_{i=1}^{u'} w(p_i) \cdot \text{cost}(p_i)}{\sum_{i=1}^{u'} w(p_i) \cdot \text{mass}(p_i)} \leq \frac{\sum_{i=1}^u w(p_i) \cdot \text{cost}(p_i)}{\sum_{i=1}^u w(p_i) \cdot \text{mass}(p_i)} = \frac{\text{cost}(\mathcal{X}_{k'})}{\text{mass}(\mathcal{X}_{k'})},$$

implying that $\text{cost}(\mathcal{Y}_{k'}) \leq \frac{\text{mass}(\mathcal{Y}_{k'})}{\text{mass}(\mathcal{X}_{k'})} \text{cost}(\mathcal{X}_{k'}) = \beta \text{cost}(\mathcal{X}_{k'})$. ■

Lemma 5.7 *We have that $\text{cost}(\mathcal{Y}) \leq 3 \frac{\gamma_+}{\gamma_- + \gamma_+} \text{cost}(\mathcal{X}) \leq 36 \text{opt}$.*

Proof: Let $\Delta = \sum_{t=1}^{k'-1} \text{ben}(\mathcal{X}_t) + \beta \text{ben}(\mathcal{X}_{k'})$ and $\Gamma = \sum_{t=1}^{k'-1} \text{cost}(\mathcal{X}_t) + \beta \text{cost}(\mathcal{X}_{k'})$, where $\beta = \frac{\text{mass}(\mathcal{Y}_{k'})}{\text{mass}(\mathcal{X}_{k'})}$. We have

$$\begin{aligned} \beta \text{ben}(\mathcal{X}_{k'}) &= \beta(\text{mass}(\mathcal{X}_{k'}) - 1) = \text{mass}(\mathcal{Y}_{k'}) - \beta = (\text{mass}(\mathcal{Y}_{k'}) - 1) + (1 - \beta) \\ &= \text{ben}(\mathcal{Y}_{k'}) + 1 - \beta \leq \text{ben}(\mathcal{Y}_{k'}) + 1. \end{aligned}$$

Therefore,

$$\Delta = \sum_{t=1}^{k'-1} \text{ben}(\mathcal{X}_t) + \beta \text{ben}(\mathcal{X}_{k'}) \leq \sum_{t=1}^{k'-1} \text{ben}(\mathcal{X}_t) + \text{ben}(\mathcal{Y}_{k'}) + 1 = \text{BEN}(\bar{\mathcal{Y}}) + 1 < \gamma_+ + 2, \quad (7)$$

by the construction of $\bar{\mathcal{Y}}$, see Eq. (2)_{p6}. Since $\frac{\text{cost}(\mathcal{X}_1)}{\text{ben}(\mathcal{X}_1)} \leq \dots \leq \frac{\text{cost}(\mathcal{X}_\alpha)}{\text{ben}(\mathcal{X}_\alpha)}$ and $1 \leq k' \leq \alpha$, we have, by Lemma 5.5, that

$$\frac{\Gamma}{\Delta} = \frac{\sum_{t=1}^{k'-1} \text{cost}(\mathcal{X}_t) + \beta \text{cost}(\mathcal{X}_{k'})}{\sum_{t=1}^{k'-1} \text{ben}(\mathcal{X}_t) + \beta \text{ben}(\mathcal{X}_{k'})} \leq \frac{\sum_{t=1}^{\alpha} \text{cost}(\mathcal{X}_t)}{\sum_{t=1}^{\alpha} \text{ben}(\mathcal{X}_t)} \leq \frac{\text{cost}(\mathcal{X})}{\gamma_- + \gamma_+ - 1},$$

since $\sum_{t=1}^{\alpha} \text{ben}(\mathcal{X}_t) \geq \gamma_- + \gamma_+ - 1$, by Claim 5.2 (ii), and $\sum_{t=1}^{\alpha} \text{cost}(\mathcal{X}_t) \leq \text{cost}(\mathcal{X})$. This implies that

$$\Gamma \leq \frac{\text{cost}(\mathcal{X})}{\gamma_- + \gamma_+ - 1} \Delta < \frac{\gamma_+ + 2}{\gamma_- + \gamma_+ - 1} \text{cost}(\mathcal{X}),$$

since $\Delta < \gamma_+ + 2$, see Eq. (7)_{p13}. By Claim 5.6, $\text{cost}(\mathcal{Y}_{k'}) \leq \beta \text{cost}(\mathcal{X}_{k'})$, and as such,

$$\begin{aligned} \text{cost}(\mathcal{Y}) &= \sum_{t=1}^{k'-1} \text{cost}(\mathcal{X}_t) + \text{cost}(\mathcal{Y}_{k'}) \leq \sum_{t=1}^{k'-1} \text{cost}(\mathcal{X}_t) + \beta \text{cost}(\mathcal{X}_{k'}) = \Gamma \\ &\leq \frac{\gamma_+ + 2}{\gamma_- + \gamma_+ - 1} \text{cost}(\mathcal{X}) \leq 3 \frac{\gamma_+}{\gamma_- + \gamma_+} \text{cost}(\mathcal{X}), \end{aligned}$$

since $\frac{\gamma_+ + 2}{\gamma_- + \gamma_+ - 1} \leq \frac{\gamma_+ + 3}{\gamma_- + \gamma_+} \leq 3 \frac{\gamma_+}{\gamma_- + \gamma_+}$ (implied by $\gamma_+ \geq 2$). Now, since $|\mathcal{X}| \leq n - m$, by the construction of \mathcal{X} , it holds that

$$\text{cost}(\mathcal{X}) \leq A_m(\mathbf{C}_-, \mathbf{P}^w) \leq \left(3 + 9 \frac{\gamma_- + \gamma_+}{\gamma_+} \right) \text{opt},$$

by Corollary 4.7. Putting above two inequalities together, we obtain

$$\text{cost}(\mathcal{Y}) \leq 3 \frac{\gamma_+}{\gamma_- + \gamma_+} \left(3 + 9 \frac{\gamma_- + \gamma_+}{\gamma_+} \right) \text{opt} \leq 36 \text{opt}. \quad \blacksquare$$

5.3 Putting things together

Lemma 5.8 *We have that $\nu(\mathbf{C}, Z) \leq 36 \text{opt}$.*

Proof: Since $Z = \mathcal{Y} \cup J^w$ and $\mathbf{C} = \{f_1, \dots, f_{k'}\} \cup J$, we have, by Lemma 5.7, that

$$\nu(\mathbf{C}, Z) \leq \nu(\{f_1, \dots, f_{k'}\}, \mathcal{Y}) + \nu(J, J^w) = \text{cost}(\mathcal{Y}) + 0 \leq 36 \text{opt},$$

as $\mathcal{Y} \subseteq \bigcup_{i=1}^{k'} \mathcal{X}_i$, and f_i is the (nearest) facility of \mathcal{X}_i in \mathbf{C}_- . \blacksquare

The proof of the following lemma can be found in Appendix C.

Lemma 5.9 *We have $|Z| \geq n - m$.*

Lemma 5.10 *If $\gamma_+ \geq 2$, then one can compute, in $O(n^2 \log^3 n)$ time, a set \mathbf{C} of k facilities such that $A_m(\mathbf{C}, \mathbf{P}) \leq 39 \text{opt}$.*

Proof: The algorithm is GREEDYMERGE, presented in Section 3.1. By Lemma 5.9, it holds $|Z| \geq n - m$. Since $Z \subseteq P^w$, by Observation 4.1, we have $A_m(C, P^w) \leq \nu(C, Z)$, which is at most 36opt , by Lemma 5.8. Now, Lemma 4.6 (i) implies that $A_m(C, P) \leq 3\text{opt} + A_m(C, P^w) \leq 39\text{opt}$. The overall running time is dominated by computing C_- and C_+ , which takes $O(n^2 \log^3 n)$ time [CKMN01]. \blacksquare

6 Correctness of SuccessiveLS ($\gamma_+ = 1$)

In this section, we show that, for the case $\gamma_+ = 1$, SUCCESSIVELS computes a solution C such that $|C| = k$ and C is the desired approximation.

Definition 6.1 (Acceptable solution.) A facility set C of size k is an *acceptable solution* if $A_m(C, P^w) \leq b'\text{opt}^w$, where b' is an appropriate fixed constant.

We shall prove that $C = \text{SUCCESSIVELS}(k, P^w, m)$ is an acceptable solution, which implies, by Lemma 4.6 (ii), that $A_m(C, P) = O(\text{opt})$. We remind the reader that in the penalty k -median with outliers problem (PMO), we are allowed to have more than m outliers, but every such extra outlier incurs an additional penalty ϱ .

Observation 6.2 Let V be a set of n points, $C \subseteq V$, and $\varrho > 0$ be a penalty parameter.

- (i) $\mathcal{A}_m(C, V, \varrho) \leq \nu(C, M) + \varrho(n - m - |M|)$, for any $M \subseteq V$ such that $|M| \leq n - m$.
- (ii) $\mathcal{A}_m(C, V, \varrho) \leq \nu(C, M)$, for any $M \subseteq V$ such that $|M| \geq n - m$.
- (iii) $\text{opt}_{\text{pmo}}(k, V, \varrho, m) \leq \text{opt}_{\text{mo}}(k, V, m)$.

6.1 The analysis of SuccessiveLS

Consider the algorithm LOCALSEARCH depicted in Figure 3. In the i th iteration, the facility set B_i is computed for the problem $\text{PMO}(k, P^w, \varrho_i, m)$. Let \overline{B}_i be the *optimal* solution for the same instance. The notations used in this section are summarized in the table on the

$\varrho_0 = d_{\min}/10$	$\varrho_i = 3^i \varrho_0$
$\Theta_i = \Theta(B_i, P^w, \varrho_i, m)$	$\overline{\Theta}_i = \Theta(\overline{B}_i, P^w, \varrho_i, m)$
$\Delta_i = n - m - \Theta_i $	$\overline{\Delta}_i = n - m - \overline{\Theta}_i $
$\eta_i = \nu(B_i, \Theta_i)$	$\overline{\eta}_i = \nu(\overline{B}_i, \overline{\Theta}_i)$
$\omega_i = \mathcal{A}_m(B_i, P^w, \varrho_i)$	$\overline{\omega}_i = \mathcal{A}_m(\overline{B}_i, P^w, \varrho_i)$ $= \text{opt}_{\text{pmo}}(k, P^w, \varrho_i, m)$

right. Here, $\Theta_i = \Theta(B_i, P^w, \varrho_i, m)$ denotes the set of points of $\mathbf{N}_{n-m}(B_i, P^w)$ in distance $\leq \varrho_i$ from B_i , namely, these are the points that contribute their true distances (from B_i) to $\mathcal{A}_m(B_i, P^w, \varrho_i)$ (note that a point in $\mathbf{N}_{n-m}(C, V) \setminus \Theta_i$ pays only the penalty, as its distance to B_i is strictly larger than ϱ_i). As such, Δ_i is the number of points that pay the penalty in the PMO clustering induced by B_i . By definition, we have

$$\omega_i = \nu(B_i, \Theta_i) + (n - m - |B_i|)\varrho_i = \eta_i + \Delta_i \varrho_i$$

and

$$\overline{\omega}_i = \nu(\overline{B}_i, \overline{\Theta}_i) + (n - m - |\overline{B}_i|)\varrho_i = \overline{\eta}_i + \overline{\Delta}_i \varrho_i = \text{opt}_{\text{pmo}}(k, P^w, \varrho_i, m),$$

as \overline{B}_i is the optimal solution.

The quantity $\overline{\Delta}$ is “dual” to the penalty parameter ϱ . In particular, $\overline{\Delta}$ is monotone decreasing as a function of ϱ^1 .

¹We sketch the proof here for $\overline{\Delta}_{i+1} \leq \overline{\Delta}_i$. Indeed, by Observation 6.2 (i), it is not hard to verify that $\overline{\eta}_i + \overline{\Delta}_i \varrho_i \leq \overline{\eta}_{i+1} + \overline{\Delta}_{i+1} \varrho_i$ and $\overline{\eta}_{i+1} + \overline{\Delta}_{i+1} \varrho_{i+1} \leq \overline{\eta}_i + \overline{\Delta}_i \varrho_{i+1}$. Adding these two inequalities together, we obtain $\overline{\Delta}_{i+1}(\varrho_{i+1} - \varrho_i) \leq \overline{\Delta}_i(\varrho_{i+1} - \varrho_i)$. Since $\varrho_{i+1} - \varrho_i = 3\varrho_i - \varrho_i > 0$, this implies that $\overline{\Delta}_{i+1} \leq \overline{\Delta}_i$.

Claim 6.3 *It holds that $\omega_0 = \bar{\omega}_0$, $\omega_1 = 3\bar{\omega}_0$, and $\omega_2 = 9\bar{\omega}_0$.*

Proof: It is easy to verify, by construction of \mathbf{P}^w , that any k points of \mathbf{P}^w have total weight at most $n - m$. As such, when $j = 0, 1, 2$, it holds that $\Theta(C, \mathbf{P}^w, \varrho_j, m) = C^w$, for any $C \subseteq \mathbf{P}^w$ satisfying $|C| = k$, since $\varrho_j \leq 9d_{\min}/10 < d_{\min}$ (which implies that no point in $\mathbf{P}^w \setminus C^w$ is in distance smaller than ϱ_j to C). Therefore, when $j = 0, 1, 2$, we have

$$\mathcal{A}_m(C, \mathbf{P}^w, \varrho_j) = \nu(C, C^w) + (n - m - |C^w|)\varrho_j = (n - m - |C^w|)\varrho_j.$$

This implies that $B_0 = \bar{B}_0 = B_1 = B_2 = \mathbf{H}$, because \mathbf{H} is the set of the k heaviest points. Now the claim follows, since $\varrho_2 = 9\varrho_0$ and $\varrho_1 = 3\varrho_0$. \blacksquare

Claim 6.4 *For $i \geq 0$, it holds that (i) $\omega_{i+1} - \omega_i \leq 2\Delta_i \varrho_i$ and (ii) $2\bar{\Delta}_{i+1} \varrho_i \leq \bar{\omega}_{i+1} - \bar{\omega}_i$.*

Proof: (i) We have $\omega_{i+1} = \mathcal{A}_m(B_{i+1}, \mathbf{P}^w, \varrho_{i+1}) \leq \mathcal{A}_m(B_i, \mathbf{P}^w, \varrho_{i+1})$, since B_{i+1} is computed by a local search starting from B_i . In addition, by Observation 6.2 (i), we have

$$\mathcal{A}_m(B_i, \mathbf{P}^w, \varrho_{i+1}) \leq \nu(B_i, \Theta_i) + (n - m - |\Theta_i|)\varrho_{i+1} = \eta_i + \Delta_i \varrho_{i+1}.$$

It follows that

$$\omega_{i+1} \leq \eta_i + \Delta_i \varrho_{i+1} = \eta_i + 3\Delta_i \varrho_i = \omega_i + 2\Delta_i \varrho_i,$$

since $\varrho_{i+1} = 3\varrho_i$ and $\omega_i = \eta_i + \Delta_i \varrho_i$.

(ii) We have $\bar{\omega}_i = \mathcal{A}_m(\bar{B}_i, \mathbf{P}^w, \varrho_i) \leq \mathcal{A}_m(\bar{B}_{i+1}, \mathbf{P}^w, \varrho_i)$, since \bar{B}_i is the optimal solution for $\text{PMO}(k, \mathbf{P}^w, \varrho_i, m)$. By Observation 6.2 (i), we have

$$\mathcal{A}_m(\bar{B}_i, \mathbf{P}^w, \varrho_{i+1}) \leq \nu(\bar{B}_i, \Theta_i) + (n - m - |\Theta_i|)\varrho_{i+1} = \eta_i + \Delta_i \varrho_{i+1}.$$

It follows that

$$\bar{\omega}_i \leq \bar{\eta}_{i+1} + \bar{\Delta}_{i+1} \varrho_i = (\bar{\eta}_{i+1} + 3\bar{\Delta}_{i+1} \varrho_i) - 2\bar{\Delta}_{i+1} \varrho_i = \bar{\omega}_{i+1} - 2\bar{\Delta}_{i+1} \varrho_i,$$

since $\bar{\omega}_{i+1} = \bar{\eta}_{i+1} + \bar{\Delta}_{i+1} \varrho_{i+1} = \bar{\eta}_{i+1} + 3\bar{\Delta}_{i+1} \varrho_i$. \blacksquare

The proof of the following lemma can be found in Section 6.2.

Lemma 6.5 *If $\omega_i \leq 9\text{opt}^w$ and there is no acceptable solution in $\mathcal{H} \cup \mathbf{N}(B_i)$, then $\Delta_i \leq \bar{\Delta}_{i-1}$.*

Naturally, when the penalty parameter exceeds d_{\max} , no point would pay the penalty in the solution computed by SUCCESSIVELS. As such, before $\varrho_i > 3d_{\max}$, we would have $\Delta_i = 0$ and thus, SUCCESSIVELS terminates. Since $\varrho_0 = d_{\min}/10$ and $d_{\max}/d_{\min} = O(n^2)$, this implies that it terminates after $O(\log n)$ calls to LOCALSEARCH (with gradually increasing penalty parameters).

Lemma 6.6 *If there is no acceptable solution in $\mathcal{H} \cup \bigcup_{t=0}^I \mathbf{N}(B_t)$, then $\omega_j \leq 9\text{opt}^w$, for $j = 0, \dots, I$, where I is the smallest index such that $\Delta_I = 0$.*

Proof: By induction on j . For the base cases $j = 0, 1$, and 2 , Claim 6.3 implies that $\omega_j \leq 9\bar{\omega}_0 = 9\text{opt}_{\text{pmo}}(k, \mathbf{P}^w, \varrho_0, m) \leq 9\text{opt}^w$, by Observation 6.2 (iii). Thus, assume that the claim holds when $0 \leq j \leq i - 1$, where $3 \leq i \leq I$. We need to show that $\omega_i \leq 9\text{opt}^w$.

By Lemma 6.5, we have that $\Delta_t \leq \bar{\Delta}_{t-1}$, for $1 \leq t \leq i - 1$, since $\omega_t \leq 9\text{opt}^w$ by the induction hypothesis. Therefore, since $\varrho_t = 9\varrho_{t-2}$, for $2 \leq t \leq i - 1$, we have

$$\omega_{t+1} - \omega_t \leq 2\Delta_t \varrho_t \leq 2\bar{\Delta}_{t-1} \varrho_t = 18\bar{\Delta}_{t-1} \varrho_{t-2} \leq 9(\bar{\omega}_{t-1} - \bar{\omega}_{t-2}),$$

by Claim 6.4. Summing this inequality, for $t = 2, \dots, i-1$, we obtain $\omega_i - \omega_2 \leq 9(\bar{\omega}_{i-2} - \bar{\omega}_0)$. This implies $\omega_i \leq 9(\bar{\omega}_{i-2} - \bar{\omega}_0) + \omega_2 = 9\bar{\omega}_{i-2} \leq 9\text{opt}^w$ since $\omega_2 = 9\bar{\omega}_0$, by Claim 6.3, and $\bar{\omega}_{i-2} = \text{opt}_{\text{pmo}}(k, \mathbf{P}^w, \varrho_{i-2}, m) \leq \text{opt}^w$, by Observation 6.2 (iii). ■

Claim 6.7 *The set $\mathcal{H} \cup \bigcup_{t=0}^I \mathcal{N}(B_t)$ contains an acceptable solution, where I is the smallest index such that $\Delta_I = 0$.*

Proof: Assume for the sake of contradiction that $\mathcal{H} \cup \bigcup_{t=0}^I \mathcal{N}(B_t)$ does not contain an acceptable solution. Since $\Delta_I = 0$, it follows that $|\Theta_I| = n - m - \Delta_I = n - m$ and $\omega_I = \nu(B_I, \Theta_I) + \varrho_I \Delta_I = \nu(B_I, \Theta_I)$. Therefore, by Observation 4.1, $A_m(B_I, \mathbf{P}^w) \leq \nu(B_I, \Theta_I) = \omega_I \leq 9\text{opt}^w$, by Lemma 6.6. However, by definition, this implies that B_I is an acceptable solution. A contradiction. ■

Lemma 6.8 *If $\gamma_+ = 1$, then one can compute, in $O(k^2(k+m)^2 n^3 \log n)$ time, a set \mathcal{C} of k points such $A_m(\mathcal{C}, \mathbf{P}) \leq (4b' + 3)\text{opt}$, where b' is the constant in Definition 6.1.*

Proof: The algorithm is SUCCESSIVELS, described in Section 3.2. By Claim 6.7, we have $A_m(\mathcal{C}, \mathbf{P}^w) \leq b' \text{opt}^w$, where \mathcal{C} is the solution computed by SUCCESSIVELS. Now, Lemma 4.6 (ii) implies $A_m(\mathcal{C}, \mathbf{P}) \leq (4b' + 3)\text{opt}$.

The overall running time of SUCCESSIVELS is dominated by the calls to LOCALSEARCH. As discussed above, SUCCESSIVELS terminates after $O(\log n)$ calls of LOCALSEARCH. There are $O(nd_{\max}/(d_{\min}/30)) = O(n^3)$ local search steps done by LOCALSEARCH, because nd_{\max} is an upper bound of the cost for any valid solution for $\text{MO}(k, \mathbf{P}^w, m)$ and $\varrho_0/3 = d_{\min}/30$ is a lower bound on the improvement a local search step makes. Each local search step in LOCALSEARCH needs to check $O(k(k+m))$ neighbors and each check (namely, to see if a neighbor facility set is better than the current solution) takes $O(k(k+m))$ time, since there are only $k_+ + m = O(k+m)$ distinct points in \mathbf{P}^w , see Remark 4.5. Hence, the total running time is $O(k^2(k+m)^2 n^3 \log n)$. ■

6.2 Proof of Lemma 6.5

6.2.1 Notations and assumptions

Given a parameter $\varrho \geq 0$ and an arbitrary facility set B satisfying $|B| = k$, let $F = \text{LOCALSEARCH}(B, \mathbf{P}^w, 3\varrho)$. And let \bar{F} be the optimal solution for $\text{PMO}(k, \mathbf{P}^w, \varrho, m)$. The notations used in this section are summarized in the table on the right.

In the remainder of this section, we prove that $\Delta \leq \bar{\Delta}$ under the following assumptions:

$F = \text{LOCALSEARCH}(B, \mathbf{P}^w, 3\varrho)$, where B is an arbitrary set of k facilities. $U = \Theta(F, \mathbf{P}^w, 3\varrho, m)$. $\Delta = n - m - U $. U_v : the points of U served by v , for $v \in F$.
\bar{F} : Optimal solution for $\text{PMO}(k, \mathbf{P}^w, \varrho, m)$. $\bar{U} = \Theta(\bar{F}, \mathbf{P}^w, \varrho, m)$. $\bar{\Delta} = n - m - \bar{U} $. $\bar{U}_{\bar{x}}$: the points of \bar{U} served by \bar{x} , for $\bar{x} \in \bar{F}$.

(A1): $A_m(F, \mathbf{P}^w, 3\varrho) \leq 9\text{opt}^w$.

(A2): $\mathcal{H} \cup \mathcal{N}(F)$ does not contain an acceptable solution.

(A3): $\Delta > 0$, that is, $|U| < n - m$. (If $\Delta = 0$, then the claim trivially holds, since $\bar{\Delta} \geq 0$.)

Specifically, the claim is that the LOCALSEARCH solution (with penalty parameter 3ϱ) penalizes no more points than the optimal solution (with penalty parameter ϱ). In other words, the balls of radius 3ϱ centered at the facilities of the LOCALSEARCH solution cover no less points than the balls of radius ϱ centered at the facilities of the optimal solution.

6.2.2 Proof of Lemma 6.5

Our proof is remotely similar to the approach used by Arya *et al.* [AGK⁺04]. We establish a bijection $\pi : F \rightarrow \bar{F}$ such that $|U_v \setminus \bar{U}| \geq |\bar{U}_{\pi(v)} \setminus U|$ holds for all $v \in F$. The quantity $|U_v \setminus \bar{U}|$ quantifies by how much \bar{U} would grow (in size) if the cluster U_v is added to \bar{U} , and $|\bar{U}_{\pi(v)} \setminus U|$ quantifies by how much U would grow if $\bar{U}_{\pi(v)}$ is added to U . Therefore, $|U_v \setminus \bar{U}| \geq |\bar{U}_{\pi(v)} \setminus U|$ implies, in some sense, that U_v is more “valuable” than $\bar{U}_{\pi(v)}$. In particular, if π has this property, then

$$|U| - |\bar{U}| = |U \setminus \bar{U}| - |\bar{U} \setminus U| = \sum_{v \in F} (|U_v \setminus \bar{U}| - |\bar{U}_{\pi(v)} \setminus U|) \geq 0, \quad (8)$$

and thus, $-|U| \leq -|\bar{U}|$. This implies $\Delta = n - m - |U| \leq n - m - |\bar{U}| = \bar{\Delta}$, by definition.

Lemma 6.9 *Under the assumptions of Section 6.2.1, we have that*

- (i) $\nu(F, U) \leq \mathcal{A}_m(F, \mathbf{P}^w, 3\varrho) \leq 9\text{opt}^w$, and
- (ii) $\nu(\bar{F}, \bar{U}) \leq \mathcal{A}_m(\bar{F}, \mathbf{P}^w, \varrho) \leq \text{opt}^w$.

Proof: (i) The first inequality holds because $\mathcal{A}_m(F, \mathbf{P}^w, 3\varrho) = \nu(F, U) + 3\varrho\Delta$ and $\varrho, \Delta \geq 0$. The second inequality holds by assumption (A1).

(ii) The first inequality holds by the same argument as (i). As for the second inequality, since \bar{F} is optimal for $\text{PMO}(k, \mathbf{P}^w, \varrho, m)$, we have $\mathcal{A}_m(\bar{F}, \mathbf{P}^w, \varrho) = \text{opt}_{\text{pmo}}(k, \mathbf{P}^w, \varrho, m) \leq \text{opt}^w$, by Observation 6.2 (iii). ■

The proof of the following lemma can be found in Section 6.2.3. Intuitively, it holds because $|\mathbf{C}_+| = k + 1$ and $w(\mathbf{C}_+) = n - m$. Indeed, assume that such \bar{x}, \bar{y} , and q satisfying $\nu(q, \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}) = O(\text{opt}^w)$ exists, namely, we can use one single facility (i.e., q) to serve $\bar{U}_{\bar{x}}$ and $\bar{U}_{\bar{y}}$ together “cheaply”. It is not hard to argue that the size of $\bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}$ is larger than two heavy points, say h_1 and h_2 . Since there are $k + 1$ heavy points in total, and their total weight is $n - m$, we can use the $k - 1$ heavy points (other than h_1 and h_2) as the $k - 1$ clusters. These $k - 1$ clusters together with q (which serves $\bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}$) would be an acceptable solution, contradicting assumption (A2).

Lemma 6.10 *Under the assumptions of Section 6.2.1, for any $\bar{x}, \bar{y} \in \bar{F}$ and $q \in \mathbf{P}$, we have $\nu(q, \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}) \geq 15\text{opt}^w$.*

Definition 6.11 (Match, capture, and prisoner.) Two facilities $v \in F$ and $\bar{x} \in \bar{F}$ overlap if $U_v \cap \bar{U}_{\bar{x}} \neq \emptyset$. We construct a graph $\mathcal{G} = (F \cup \bar{F}, \mathcal{E})$, where the edge $v\bar{x} \in \mathcal{E}$ if v and \bar{x} overlap. The degree of $u \in F \cup \bar{F}$ is denoted by $\deg(u)$.

A facility $v \in F$ and a facility $\bar{x} \in \bar{F}$ *match*, if $v\bar{x} \in \mathcal{E}$ and $\deg(v) = \deg(\bar{x}) = 1$.

A facility $v \in F$ *captures* a facility $\bar{x} \in \bar{F}$, if v is the nearest neighbor to \bar{x} in F and $d(v, \bar{x}) < 2\varrho$. In this case, \bar{x} is a *prisoner* of v .

Observation 6.12 *Under the assumptions of Section 6.2.1, we have $|U| < n - m$, and as such, all the points of \mathbf{P}^w in distance at most 3ϱ from F are in U .*

Claim 6.13 *Under the assumptions of Section 6.2.1, if v captures \bar{x} then $v\bar{x} \in \mathcal{E}$.*

Proof: Since $d(F, \bar{x}) \leq d(v, \bar{x}) < 2\varrho$, we have, by Observation 6.12, that $\bar{x} \in U$. Now, since the nearest neighbor to \bar{x} in F is v , it follows that \bar{x} is in the cluster of v , namely $\bar{x} \in U_v$. Therefore, we have $\bar{x} \in \bar{U}_{\bar{x}} \cap U_v$, which implies the claim. ■

Claim 6.14 For $v \in F$ and $\bar{x} \in \bar{F}$, if \bar{x} is a prisoner of v , then

- (i) $\bar{U}_{\bar{x}} \subseteq U$ (that is, $|\bar{U}_{\bar{x}} \setminus U| = 0$), and
- (ii) for any $p \in \bar{U}_{\bar{x}}$, it holds that $d(v, p) \leq d(F, p) + 2d(\bar{F}, p)$.

Proof: (i) For a point $p \in \bar{U}_{\bar{x}}$, it holds, by the triangle inequality, that $d(F, p) \leq d(v, p) \leq d(v, \bar{x}) + d(\bar{x}, p) \leq 2\rho + \rho = 3\rho$. Thus, by Observation 6.12, we have $p \in U$.

(ii) Fix a point $p \in \bar{U}_{\bar{x}}$, and let s be the nearest neighbor to p in F . Since v captures \bar{x} , the nearest neighbor to \bar{x} in F is v , and as such $d(v, \bar{x}) \leq d(s, \bar{x})$. Therefore, by the triangle inequality,

$$\begin{aligned} d(v, p) &\leq d(v, \bar{x}) + d(\bar{x}, p) \leq d(s, \bar{x}) + d(\bar{x}, p) \leq (d(s, p) + d(p, \bar{x})) + d(\bar{x}, p) \\ &= d(F, p) + 2d(\bar{x}, p) = d(F, p) + 2d(\bar{F}, p). \end{aligned} \quad \blacksquare$$

Claim 6.15 Under the assumptions of Section 6.2.1, any facility in \bar{F} can be a prisoner of at most one facility in F , and any facility of F can capture at most one facility of \bar{F} .

Proof: The first assertion follows from the definition, since a prisoner always belong to its nearest neighbor in F (which is distinct, as all distances are distinct). As for the second claim, let $v \in F$, and assume, for the sake of contradiction, that v captures two facilities $\bar{x}, \bar{y} \in \bar{F}$. By Claim 6.14 (ii), we have

$$\begin{aligned} \nu(v, \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}) &= \sum_{p \in \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}} d(v, p) \leq \sum_{p \in \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}} (d(F, p) + 2d(\bar{F}, p)) \\ &\leq \sum_{p \in U} d(F, p) + \sum_{p \in \bar{U}} 2d(\bar{F}, p) = \nu(F, U) + 2\nu(\bar{F}, \bar{U}), \end{aligned}$$

since $\bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}} \subseteq \bar{U}$ and $\bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}} \subseteq U$, by Claim 6.14 (i). Now, By Lemma 6.9, we have

$$\nu(v, \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}) \leq \nu(F, U) + 2\nu(\bar{F}, \bar{U}) \leq 9\text{opt}^w + 2\text{opt}^w \leq 11\text{opt}^w,$$

contradicting Lemma 6.10. \blacksquare

Definition 6.16 Let $F_C \subseteq F$ be the set of facilities that capture some facilities of \bar{F} , and let $\bar{F}_C \subseteq \bar{F}$ be the corresponding set of prisoners. By Claim 6.15, there exists a bijection $\pi_C : F_C \rightarrow \bar{F}_C$ such that v captures $\pi_C(v)$ for each $v \in F_C$.

Let $F_M \subseteq F \setminus F_C$ be the set of facilities which match some facilities in $\bar{F} \setminus \bar{F}_C$, and let $\bar{F}_M \subseteq \bar{F} \setminus \bar{F}_C$ be the set of facilities which match some facilities in F_M . It follows from the definition that there exists a bijection $\pi_P : F_M \rightarrow \bar{F}_M$ such that v and $\pi_P(v)$ matches each other, for every $v \in F_M$.

Let $F_L = F \setminus (F_C \cup F_M)$ and $\bar{F}_L = \bar{F} \setminus (\bar{F}_C \cup \bar{F}_M)$. Let $\pi_L : F_L \rightarrow \bar{F}_L$ be an arbitrary bijection.

Let $\pi : F \rightarrow \bar{F}$ be the bijection formed together by π_C, π_P , and π_L . See Figure 4.

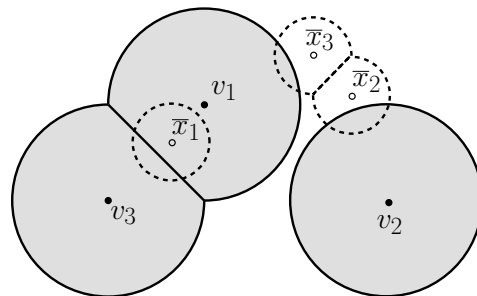


Figure 4: $F = \{v_1, v_2, v_3\}$ and $\bar{F} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$. The area inside the circles represents the points in U , the area inside the dashed circles represents the points in \bar{U} . Here, v_1 captures \bar{x}_1 , and v_2 matches \bar{x}_2 but does not capture \bar{x}_2 . We have $\pi(v_1) = \bar{x}_1$, $\pi(v_2) = \bar{x}_2$, and $\pi(v_3) = \bar{x}_3$.

We next establish that, for all $v \in F$, it holds $|U_v \setminus \bar{U}| \geq |\bar{U}_{\pi(v)} \setminus U|$, which proves Eq. (8)_{p17} and thus implies Lemma 6.5. In fact, since π_L is an arbitrary bijection (between F_L and \bar{F}_L), our proof would imply the stronger property that $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U|$, for any $v \in F_L$ and $\bar{x} \in \bar{F}_L$. (However, our proof actually does not require this stronger property.)

The following lemma is implied immediately by Claim 6.14 (i).

Lemma 6.17 *If $v \in F_C$ and $\bar{x} = \pi(v)$ then $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U| = 0$.*

Lemma 6.18 *Under the assumptions of Section 6.2.1, there does not exist a multiset $M \subseteq \mathbf{P}^w$ and a set $C \in \mathcal{H} \cup \mathcal{N}(F)$ such that $|M| \geq n - m$ and $\nu(C, M) \leq \mathbf{b}'\text{opt}^w$, where \mathbf{b}' is the constant in Definition 6.1.*

Proof: Assume for the sake of contradiction that such a set exists. Then, by Observation 4.1, it holds $\mathcal{A}_m(C, \mathbf{P}^w) \leq \nu(C, M) \leq \mathbf{b}'\text{opt}^w$, which implies that C is an acceptable solution. This contradicts the assumption (A2) that $\mathcal{H} \cup \mathcal{N}(F)$ does not contain such a solution. ■

Let $U_{v \rightarrow \bar{x}} = (U \setminus U_v) \cup \bar{U}_{\bar{x}}$ and $\bar{U}_{\bar{x} \rightarrow v} = (\bar{U} \setminus \bar{U}_{\bar{x}}) \cup U_v$.

Lemma 6.19 *If $|\bar{U}_{\bar{x} \rightarrow v}| - |\bar{U}| \geq |U| - |U_{v \rightarrow \bar{x}}|$, then $|\bar{U}_{\bar{x} \rightarrow v}| \geq |\bar{U}|$.*

Proof: Assume for the sake of contradiction that $|\bar{U}_{\bar{x} \rightarrow v}| < |\bar{U}|$. Let $F_{v \rightarrow \bar{x}} = F - v + \bar{x}$, where the notation $F - v + \bar{x}$ refers to $(F \setminus \{v\}) \cup \{\bar{x}\}$. We have

$$\begin{aligned} \nu(F_{v \rightarrow \bar{x}}, U_{v \rightarrow \bar{x}}) - \nu(F, U) &\leq \left(\nu(F - v, U \setminus U_v) + \nu(\bar{x}, \bar{U}_{\bar{x}}) \right) - \left(\nu(F - v, U \setminus U_v) + \nu(v, U_v) \right) \\ &= \nu(\bar{x}, \bar{U}_{\bar{x}}) - \nu(v, U_v). \end{aligned} \quad (9)$$

This implies that

$$\begin{aligned} \nu(F_{v \rightarrow \bar{x}}, U_{v \rightarrow \bar{x}}) &\leq \nu(F, U) - \nu(v, U_v) + \nu(\bar{x}, \bar{U}_{\bar{x}}) \leq \nu(F, U) + \nu(\bar{F}, \bar{U}) \\ &\leq 9\text{opt}^w + \text{opt}^w \leq 10\text{opt}^w, \end{aligned} \quad (10)$$

by Lemma 6.9. If $|U_{v \rightarrow \bar{x}}| \geq n - m$ then $F_{v \rightarrow \bar{x}}$ is an acceptable solution, contradicting Lemma 6.18. Thus, we have $|U_{v \rightarrow \bar{x}}| < n - m$. Now, by Observation 6.2 (i), we have $\mathcal{A}_m(F_{v \rightarrow \bar{x}}, \mathbf{P}^w, 3\varrho) \leq \nu(F_{v \rightarrow \bar{x}}, U_{v \rightarrow \bar{x}}) + 3\varrho(n - m - |U_{v \rightarrow \bar{x}}|)$. Therefore,

$$\begin{aligned} D &= \mathcal{A}_m(F_{v \rightarrow \bar{x}}, \mathbf{P}^w, 3\varrho) - \mathcal{A}_m(F, \mathbf{P}^w, 3\varrho) \\ &\leq \left(\nu(F_{v \rightarrow \bar{x}}, U_{v \rightarrow \bar{x}}) + 3\varrho \cdot (n - m - |U_{v \rightarrow \bar{x}}|) \right) - \left(\nu(F, U) + 3\varrho \cdot (n - m - |U|) \right) \\ &= \nu(F_{v \rightarrow \bar{x}}, U_{v \rightarrow \bar{x}}) - \nu(F, U) + 3\varrho \cdot (|U| - |U_{v \rightarrow \bar{x}}|) \\ &\leq \nu(\bar{x}, \bar{U}_{\bar{x}}) - \nu(v, U_v) + 3\varrho \cdot (|U| - |U_{v \rightarrow \bar{x}}|), \end{aligned}$$

by Eq. (9). Moreover, since F is the solution computed by LOCALSEARCH, we have that $D \geq -(3\varrho)/3 = -\varrho$ and as such,

$$\nu(\bar{x}, \bar{U}_{\bar{x}}) - \nu(v, U_v) + 3\varrho \cdot (|U| - |U_{v \rightarrow \bar{x}}|) \geq D \geq -\varrho. \quad (11)$$

Let $\bar{F}_{\bar{x} \rightarrow v} = \bar{F} - \bar{x} + v$. Since $|\bar{U}_{\bar{x} \rightarrow v}| < |\bar{U}| \leq n - m$ (by assumption), arguing as above, we have

$$\bar{D} = \mathcal{A}_m(\bar{F}_{\bar{x} \rightarrow v}, \mathbf{P}^w, \varrho) - \mathcal{A}_m(\bar{F}, \mathbf{P}^w, \varrho) \leq \nu(v, U_v) - \nu(\bar{x}, \bar{U}_{\bar{x}}) + \varrho \cdot (|\bar{U}| - |\bar{U}_{\bar{x} \rightarrow v}|).$$

Moreover, since \bar{F} is the optimal solution for $\text{PMO}(k, \mathbf{P}^w, \varrho, m)$, we have $\bar{D} \geq 0$. It follows

$$\nu(v, U_v) - \nu(\bar{x}, \bar{U}_{\bar{x}}) + \varrho \cdot (|\bar{U}| - |\bar{U}_{\bar{x} \rightarrow v}|) \geq 0. \quad (12)$$

Now, adding Eq. (11) and Eq. (12) together, we obtain

$$3\varrho \cdot (|U| - |U_{v \rightarrow \bar{x}}|) + \varrho \cdot (|\bar{U}| - |\bar{U}_{\bar{x} \rightarrow v}|) \geq -\varrho.$$

Since $|\bar{U}_{\bar{x} \rightarrow v}| - |\bar{U}| \geq |U| - |U_{v \rightarrow \bar{x}}|$, it follows that

$$3\varrho \cdot (|\bar{U}_{\bar{x} \rightarrow v}| - |\bar{U}|) + \varrho \cdot (|\bar{U}| - |\bar{U}_{\bar{x} \rightarrow v}|) \geq 3\varrho \cdot (|U| - |U_{v \rightarrow \bar{x}}|) + \varrho \cdot (|\bar{U}| - |\bar{U}_{\bar{x} \rightarrow v}|) \geq -\varrho,$$

or equivalently, $|\bar{U}_{\bar{x} \rightarrow v}| - |\bar{U}| \geq -1/2$. This implies that $|\bar{U}_{\bar{x} \rightarrow v}| \geq |\bar{U}|$, contradicting our assumption $|\bar{U}_{\bar{x} \rightarrow v}| < |\bar{U}|$. \blacksquare

Lemma 6.20 *Under the assumptions of Section 6.2.1, if $v \in F_M$ and $\bar{x} = \pi(v)$ then $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U|$.*

Proof: Since v and \bar{x} match each other, \bar{x} is the only facility in \bar{F} that overlaps with v , and as such, $|U_{v \rightarrow \bar{x}}| = |U \setminus U_v| + |\bar{U}_{\bar{x}}| = |U| - |U_v| + |\bar{U}_{\bar{x}}|$. Similarly, we have $|\bar{U}_{\bar{x} \rightarrow v}| = |\bar{U}| - |\bar{U}_{\bar{x}}| + |U_v|$. It thus follows that $|\bar{U}_{\bar{x} \rightarrow v}| - |\bar{U}| = |U_v| - |\bar{U}_{\bar{x}}| = |U| - |U_{v \rightarrow \bar{x}}|$. Now, by Lemma 6.19, we have $|\bar{U}_{\bar{x} \rightarrow v}| \geq |\bar{U}|$, which implies $|U_v| \geq |\bar{U}_{\bar{x}}|$. Therefore, we have $|U_v \setminus \bar{U}| = |U_v| - |U_v \cap \bar{U}| \geq |\bar{U}_{\bar{x}}| - |U_v \cap \bar{U}_{\bar{x}}| = |\bar{U}_{\bar{x}} \setminus U|$. \blacksquare

The proof of the following claim is similar to the proof of Lemma 6.20, and is thus omitted.

Claim 6.21 *Let $v \in F_L$ and $\bar{x} = \pi(v)$. Under the assumptions of Section 6.2.1, if $\deg(v) = \deg(\bar{x}) = 0$ then $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U|$.*

Lemma 6.22 *Under the assumptions of Section 6.2.1, there does not exist a multiset $G \subseteq \mathbf{P}^w$ of size Δ , such that $G \cap U = \emptyset$ and for all $p \in G$, it holds $d(p, F) \leq 5\varrho$.*

Proof: Assume for the sake of contradiction that G exists. Then, we have $|U \cup G| = |U| + \Delta = n - m$, and moreover,

$$\begin{aligned} \nu(F, U \cup G) &\leq \nu(F, U) + \nu(F, G) \leq \nu(F, U) + 5\varrho|G| = \nu(F, U) + 5\varrho\Delta \\ &\leq \frac{5}{3}(\nu(F, U) + 3\varrho\Delta) = \frac{5}{3}\mathcal{A}_m(F, \mathbf{P}^w, 3\varrho) \leq 15\text{opt}^w, \end{aligned}$$

since $\mathcal{A}_m(F, \mathbf{P}^w, 3\varrho) \leq 9\text{opt}^w$, by Lemma 6.9. Namely, F is an acceptable solution, which contradicts Lemma 6.18. \blacksquare

Lemma 6.23 *Let $v \in F$ and $\bar{x} \in \bar{F}$. If $v\bar{x} \in \mathcal{E}$, then $d(v, \bar{x}) \leq 4\varrho$, and furthermore, for all $p \in \bar{U}_{\bar{x}}$, it holds $d(p, v) \leq 5\varrho$.*

Proof: Since $v\bar{x} \in \mathcal{E}$, there is a point q that is in both U_v and $\bar{U}_{\bar{x}}$. Therefore, we have $d(q, v) \leq 3\varrho$ and $d(q, \bar{x}) \leq \varrho$. By the triangle inequality, it holds $d(v, \bar{x}) \leq d(v, q) + d(q, \bar{x}) \leq 4\varrho$. For an arbitrary point $p \in \bar{U}_{\bar{x}}$, we have that $d(p, \bar{x}) \leq \varrho$, and as such, again by the triangle inequality, $d(p, v) \leq d(p, \bar{x}) + d(\bar{x}, v) \leq \varrho + 4\varrho = 5\varrho$. \blacksquare

Lemma 6.24 *Under the assumptions of Section 6.2.1, there exists a heavy point $h \in \mathbf{P}^w$ such that $h \notin U$ and furthermore, for all $v \in F$, it holds $\Delta \leq w(h) \leq |U_v|$.*

Proof: Consider an arbitrary heavy point h' . Since $|U| < n - m$, it follows by the definition of U that h' appears either $w(h')$ times or not at all in U . Note that the total weight of all the heavy points is $n - m$, namely $w(C_+) = n - m$. This implies (since $|U| < n - m$) that there exists at least one heavy point that does not appear in U , and let h denote this point.

Assume, for the sake of contradiction, that $w(h) \leq \Delta - 1$. Recall that $H \subseteq C_+$ is the set of k heaviest points and $|C_+| = k + 1$. Thus, $n - m - |H|$ is the weight of the heavy point with the least weight in C_+ . As such, it holds that $n - m - |H| \leq w(h) \leq \Delta - 1$. By Observation 6.2 (i), we have

$$\begin{aligned} \mathcal{A}_m(H, P^w, 3\varrho) &\leq \nu(H, H^w) + 3\varrho(n - m - |H^w|) \leq 0 + 3\varrho(\Delta - 1) \\ &\leq \nu(F, U) + 3\varrho\Delta - 3\varrho = \mathcal{A}_m(F, P^w, 3\varrho) - 3\varrho. \end{aligned} \quad (13)$$

On the other hand, since $F = \text{LOCALSEARCH}(B, P^w, 3\varrho)$, where B is an arbitrary set of k facilities, it holds that $\mathcal{A}_m(F, P^w, 3\varrho) - \varrho \leq \mathcal{A}_m(H, P^w, 3\varrho)$, see Figure 3 (note that H is one of the candidate solutions considered by LOCALSEARCH). Combining this inequality with Eq. (13), we obtain

$$\mathcal{A}_m(F, P^w, 3\varrho) - \varrho \leq \mathcal{A}_m(H, P^w, 3\varrho) \leq \mathcal{A}_m(F, P^w, 3\varrho) - 3\varrho,$$

which is a contradiction.

Next, we prove the other inequality $w(h) \leq |U_v|$, for every $v \in F$. Assume for the sake of contradiction that $w(h) > |U_v|$. Let $M = (U \setminus U_v) \cup h^w$. We have

$$|M| = |U \setminus U_v| + w(h) = |U| - |U_v| + w(h) \geq |U| + 1,$$

since $h \notin U$ and $w(h) > |U_v|$. Now, note that

$$\nu(F - v + h, M) \leq \nu(F - v, U \setminus U_v) + \nu(\{h\}, h^w) \leq \nu(F, U) + 0 = \nu(F, U).$$

If $|M| \leq n - m$ then by Observation 6.2 (i), it holds that

$$\begin{aligned} \mathcal{A}_m(F - v + h, P^w, 3\varrho) &\leq \nu(F - v + h, M) + 3\varrho(n - m - |M|) \\ &\leq \nu(F, U) + 3\varrho(n - m - |U| - 1) = \mathcal{A}_m(F, P^w, 3\varrho) - 3\varrho, \end{aligned}$$

since $|M| \geq |U| + 1$. Now, arguing as above, this contradicts the local optimality of F , as $F - v + h$ is one of the possible solutions considered by LOCALSEARCH, see Figure 3.

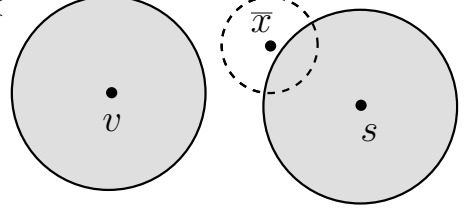
Otherwise, we have $|M| > n - m$, and by Observation 6.2 (ii), it holds that

$$\begin{aligned} \mathcal{A}_m(F - v + h, P^w, 3\varrho) &\leq \nu(F - v + h, M) \leq \nu(F, U) \\ &\leq \nu(F, U) + 3\varrho(n - m - |U| - 1) = \mathcal{A}_m(F, P^w, 3\varrho) - 3\varrho, \end{aligned}$$

since $n - m - |U| - 1 \geq 0$ (implied by $|U| < n - m$). Again, this is a contradiction to the local optimality of F . \blacksquare

Claim 6.25 *Let $v \in F_L$ and $\bar{x} = \pi(v)$. Under the assumptions of Section 6.2.1, if $\deg(v) = 0$, and there is a facility $s \in F$ such that $s \neq v$ and $s\bar{x} \in \mathcal{E}$, then $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U|$.*

Proof: Assume for the sake of contradiction that $|U_v \setminus \bar{U}| < |\bar{U}_{\bar{x}} \setminus U|$. Since the degree of v is zero, we have $|U_v| = |U_v \setminus \bar{U}| < |\bar{U}_{\bar{x}} \setminus U|$. By Lemma 6.24, we have $\Delta \leq |U_v|$. It thus follows $\Delta < |\bar{U}_{\bar{x}} \setminus U|$, and as such, there exists a subset $G \subseteq \bar{U}_{\bar{x}} \setminus U$ such that $|G| = \Delta$. Furthermore, by Lemma 6.23, each point of $\bar{U}_{\bar{x}} \setminus U$ is within distance 5ρ to s . Since $G \subseteq \bar{U}_{\bar{x}} \setminus U$, this implies that $d(p, F) \leq d(p, s) \leq 5\rho$, for each $p \in G$. However, this contradicts Lemma 6.22. \blacksquare



The following claim will be useful in proving Claim 6.27 below.

Claim 6.26 For $v \in F_L$ and $s \in F - v$ such that $d(v, s) \leq 8\rho$, it holds $\nu(F - v, U_v \cap \bar{U}) \leq 11\nu(v, U_v) + 2\nu(\bar{F}, \bar{U})$.

Proof: Consider $\bar{y} \in \bar{F}$ such that $U_v \cap \bar{U}_{\bar{y}}$ is not empty. For an arbitrary point $p \in U_v \cap \bar{U}_{\bar{y}}$, it holds $d(v, p) \leq 3\rho$ and $d(\bar{y}, p) \leq \rho$.

If $d(\bar{y}, v) \geq 2\rho$, then by the triangle inequality, we have that $d(p, s) \leq d(p, v) + d(v, s) \leq 3\rho + 8\rho = 11\rho$ and $d(p, v) \geq d(\bar{y}, v) - d(\bar{y}, p) \geq 2\rho - \rho = \rho$. In particular, for $p \in U_v \cap \bar{U}_{\bar{y}}$, we have $\nu(v, p) \geq \rho$, and as such $\nu(v, U_v \cap \bar{U}_{\bar{y}}) \geq \rho |U_v \cap \bar{U}_{\bar{y}}|$. Since $s \in F - v$, we have

$$\nu(F - v, U_v \cap \bar{U}_{\bar{y}}) \leq \nu(s, U_v \cap \bar{U}_{\bar{y}}) \leq 11\rho |U_v \cap \bar{U}_{\bar{y}}| \leq 11\nu(v, U_v \cap \bar{U}_{\bar{y}}) = 11 \sum_{p \in U_v \cap \bar{U}_{\bar{y}}} d(F, p).$$

If $d(\bar{y}, v) < 2\rho$, then the distance between \bar{y} and its nearest neighbor in F is less than 2ρ , and as such, \bar{y} is a prisoner of $\pi^{-1}(\bar{y})$. Note that $\pi^{-1}(\bar{y}) \neq v$, since otherwise, v captures \bar{y} , contradicting that $v \in F_L$. Claim 6.14 (ii) implies that

$$\nu(F - v, U_v \cap \bar{U}_{\bar{y}}) \leq \nu(\pi^{-1}(\bar{y}), U_v \cap \bar{U}_{\bar{y}}) \leq \sum_{p \in U_v \cap \bar{U}_{\bar{y}}} (d(F, p) + 2d(\bar{F}, p)).$$

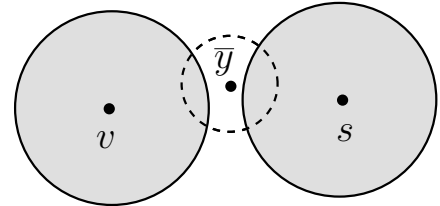
Combining these two cases, we obtain $\nu(F - v, U_v \cap \bar{U}_{\bar{y}}) \leq \sum_{p \in U_v \cap \bar{U}_{\bar{y}}} (11d(F, p) + 2d(\bar{F}, p))$. Sum-

ming the inequality over all facilities $\bar{y} \in \bar{F}$, we have that

$$\nu(F - v, U_v \cap \bar{U}) \leq \sum_{p \in U_v \cap \bar{U}} (11d(F, p) + 2d(\bar{F}, p)) \leq 11\nu(v, U_v) + 2\nu(\bar{F}, \bar{U}). \quad \blacksquare$$

Claim 6.27 Let $v \in F_L$ and $\bar{x} = \pi(v)$. Under the assumptions of Section 6.2.1, if there exists $\bar{y} \in \bar{F}$ and $s \in F$ such that $v\bar{y}, s\bar{y} \in \mathcal{E}$, then $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U|$. Note that $s \neq v$, but it is possible that $\bar{y} = \bar{x}$.

Proof: Assume for the sake of contradiction that $|U_v \setminus \bar{U}| < |\bar{U}_{\bar{x}} \setminus U|$. By the triangle inequality and Lemma 6.23, it follows that $d(v, s) \leq d(v, \bar{y}) + d(\bar{y}, s) \leq 4\rho + 4\rho = 8\rho$. And for any $q \in U_v$, we have $d(v, q) \leq d(q, v) + d(v, s) \leq 3\rho + 8\rho = 11\rho$.



(i) If $|U_v \setminus \bar{U}| \geq \Delta$, then there exists a multiset $G \subseteq U_v \setminus \bar{U}$ such that $|G| = \Delta$. Let $M' = (U \setminus U_v) \cup (U_v \cap \bar{U}) \cup G \cup (\bar{U}_{\bar{x}} \setminus U)$. Observe that M' is the

union of the four disjoint sets. Indeed, $\overline{U_x} \setminus U$ is disjoint from U , which contains the other three sets. It is easy to verify the other pairs of sets are also disjoint, using similar arguments. Therefore,

$$\begin{aligned} |M'| &= |U \setminus U_v| + |U_v \cap \overline{U}| + |G| + |\overline{U_x} \setminus U| \\ &= (|U| - |U_v|) + (|U_v| - |U_v \setminus \overline{U}|) + \Delta + |\overline{U_x} \setminus U| \\ &= |U| + \Delta - |U_v \setminus \overline{U}| + |\overline{U_x} \setminus U| > n - m, \end{aligned}$$

since $|U| + \Delta = n - m$ and $|U_v \setminus \overline{U}| < |\overline{U_x} \setminus U|$ (by assumption). Furthermore, for all $q \in G \subseteq U_v$, it holds that $d(q, s) \leq 11\rho$ and as such, $\nu(F - v, G) \leq \nu(s, G) \leq 11\rho|G| = 11\rho\Delta$. Let $X = (U \setminus U_v) \cup (U_v \cap \overline{U}) \cup G$, and by Claim 6.26, we have

$$\begin{aligned} \Gamma &= \nu(F - v, X) = \nu(F - v, U \setminus U_v) + \nu(F - v, U_v \cap \overline{U}) + \nu(F - v, G) \\ &\leq \nu(F - v, U \setminus U_v) + (11\nu(v, U_v) + 2\nu(\overline{F}, \overline{U})) + 11\rho\Delta \\ &\leq 11\nu(F - v, U \setminus U_v) + 11\nu(v, U_v) + 11\rho\Delta + 2\nu(\overline{F}, \overline{U}) \\ &= 11\nu(F, U) + 11\rho\Delta + 2\nu(\overline{F}, \overline{U}) \leq 11\mathcal{A}_m(F, P^w, 3\rho) + 2\mathcal{A}_m(\overline{F}, P^w, \rho) \\ &\leq 11 \cdot 9\text{opt}^w + 2\text{opt}^w = 101\text{opt}^w, \end{aligned}$$

since $\mathcal{A}_m(F, P^w, 3\rho) \leq 9\text{opt}^w$ and $\mathcal{A}_m(\overline{F}, P^w, \rho) \leq \text{opt}^w$ by Lemma 6.9. Let $F' = F - v + \overline{x}$, we have

$$\begin{aligned} \nu(F', M') &= \nu(F - v + \overline{x}, X \cup (\overline{U_x} \setminus U)) \leq \Gamma + \nu(\overline{x}, \overline{U_x} \setminus U) \leq 101\text{opt}^w + \nu(\overline{F}, \overline{U}) \\ &\leq 101\text{opt}^w + \mathcal{A}_m(\overline{F}, P^w, \rho) \leq 102\text{opt}^w. \end{aligned}$$

Namely, F' is an acceptable solution, which contradicts Lemma 6.18.

(ii) Consider the case that $|U_v \setminus \overline{U}| < \Delta$. By Lemma 6.24, there exists a heavy point h such that $h \notin U$ and $w(h) \geq \Delta$. Let $M'' = U \cup h^w$, and let $F'' = F - v + h$. It holds that $|M''| = |U| + w(h) \geq |U| + \Delta = n - m$. Set $G = U_v \setminus \overline{U}$, arguing as above, we have

$$\Gamma = \nu(F - v, U) = \nu(F - v, (U \setminus U_v) \cup (U_v \cap \overline{U}) \cup G) \leq 101\text{opt}^w,$$

and as such, $\nu(F'', M'') \leq \Gamma + \nu(h, h^w) \leq 101\text{opt}^w$. Again, this implies that F'' is an acceptable solution, and this contradicts Lemma 6.18. \blacksquare

Claim 6.28 *Let $v \in F_L$ and $\overline{x} = \pi(v)$. Under the assumptions of Section 6.2.1, if there exists $\overline{y}, \overline{z} \in \overline{F}$ such that $v\overline{y}, v\overline{z} \in \mathcal{E}$ and $\deg(\overline{y}) = \deg(\overline{z}) = 1$ (namely, both \overline{y} and \overline{z} overlap only v), then $|U_v \setminus \overline{U}| \geq |\overline{U_x} \setminus U|$. Note that $\overline{y} \neq \overline{z}$, but it is possible that $\overline{y} = \overline{x}$ or $\overline{z} = \overline{x}$.*

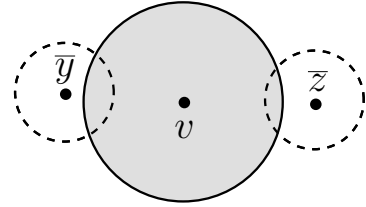
Proof: Assume for the sake of contradiction that $|U_v \setminus \overline{U}| < |\overline{U_x} \setminus U|$. Since both \overline{y} and \overline{z} overlap with only v , we have $(\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U = (\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U_v$. Also note that, by Lemma 6.23, every point in $\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}$ is within distance 5ρ to v .

If $|(\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U| \geq \Delta$ then there exists a subset $G \subseteq (\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U$ such that $|G| = \Delta$. Since each point in G is within distance 5ρ to v , this contradicts Lemma 6.22.

If $|(\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U| < \Delta$ then we have $\nu(v, (\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U) \leq 5\rho|(\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U| < 5\rho\Delta$, since every point in $\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}$ is within distance 5ρ to v . Now, by Lemma 6.9, we have that $\mathcal{A}_m(F, P^w, 3\rho) \leq 9\text{opt}^w$, and as such,

$$\begin{aligned} \nu(v, \overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) &= \nu(v, (\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \cap U_v) + \nu(v, (\overline{U_{\overline{y}}} \cup \overline{U_{\overline{z}}}) \setminus U_v) < \nu(v, U_v) + 5\rho\Delta \\ &\leq \nu(F, U) + 5\rho\Delta \leq \frac{5}{3}\mathcal{A}_m(F, P^w, 3\rho) \leq 15\text{opt}^w. \end{aligned}$$

However, this contradicts Lemma 6.10. \blacksquare



Lemma 6.29 *Under the assumptions of Section 6.2.1, if $v \in F_L$ and $\bar{x} = \pi(v)$ then $|U_v \setminus \bar{U}| \geq |\bar{U}_{\bar{x}} \setminus U|$.*

Proof: Consider the degrees of v and \bar{x} . There are six cases.

- (i) If $\deg(v) = \deg(\bar{x}) = 0$, then the lemma holds by Claim 6.21.
- (ii) If $\deg(v) = 0$ and $\deg(\bar{x}) \geq 1$, then the lemma holds by Claim 6.25.
- (iii) If $\deg(v) = 1$, $\exists v\bar{y} \in \mathcal{E}$, and $\deg(\bar{y}) = 1$, then by definition, they match, which contradicts $v \in F_L$.
- (iv) If $\deg(v) = 1$, $\exists v\bar{y} \in \mathcal{E}$, and $\deg(\bar{y}) > 1$, then the lemma holds by Claim 6.27.
- (v) If $\deg(v) \geq 2$, $\exists v\bar{y}, v\bar{z} \in \mathcal{E}$, and $\deg(\bar{y}) = \deg(\bar{z}) = 1$, then the lemma holds by Claim 6.28.
- (vi) If $\deg(v) \geq 2$, $\exists v\bar{y} \in \mathcal{E}$, and $\deg(\bar{y}) > 1$, then the lemma holds by Claim 6.27. \blacksquare

Lemma 6.17, Lemma 6.20, and Lemma 6.29 imply that $|U_v \setminus \bar{U}| \geq |\bar{U}_{\pi(v)} \setminus U|$ holds for every facility $v \in F$. As discussed, in Section 6.2.1, this implies Lemma 6.5.

6.2.3 Proof of Lemma 6.10

The proofs in this section depends only on the claims and lemmas preceeding Lemma 6.10.

Lemma 6.30 *Under the assumptions of Section 6.2.1, there does not exist two heavy points h and h' , a multiset $G \subseteq \mathbf{P}^w$, and a facility $q \in \mathbf{P}^w$, such that (i) $|G| \geq w(h) + w(h')$, (ii) the multiset G excludes every heavy point in $\mathbf{C}_+ - h - h'$, and (iii) $\nu(q, G) \leq 15\text{opt}^w$.*

Proof: Assume for the sake of contradiction that they do exist. Let $B = \mathbf{C}_+ - h - h'$. Since $|\mathbf{C}_+| = k_+ = k+1$, we have $|B| = k-1$. It holds that $|B^w \cup G| = w(B) + |G| \geq w(B) + w(h) + w(h') = w(\mathbf{C}_+) = n - m$. Furthermore,

$$\nu(B + q, B^w \cup G) \leq \nu(B, B^w) + \nu(q, G) \leq 0 + 15\text{opt}^w.$$

Since $B + q \in \mathcal{H}$ and it is an acceptable solution, this contradicts Lemma 6.18. \blacksquare

Claim 6.31 *Under the assumptions of Section 6.2.1, the following holds:*

- (i) *There is at most one facility \bar{x} in \bar{F} such that $\bar{U}_{\bar{x}}$ partly-include a heavy point.*
- (ii) *There is no facility \bar{x} in \bar{F} such that $\bar{U}_{\bar{x}}$ includes two or more heavy points. (However, $\bar{U}_{\bar{x}}$ may include one heavy point and partly-include another heavy point.)*

Proof: (i) Since $\bar{U} \subseteq \mathbf{P}^w$ is the set of $n - m - \bar{\Delta}$ closest points to \bar{F} , and the inter-point distances of \mathbf{P} are distinct, it follows that at most one heavy point can be “shattered” by \bar{F} .

(ii) Assume for the sake of contradiction that $\bar{U}_{\bar{x}}$ includes two heavy points h and h' . Let $G = \{h, h'\}^w$. Since $\bar{U}_{\bar{x}}$ includes h and h' , we have $G \subseteq \bar{U}_{\bar{x}}$, and as such $\nu(\bar{x}, G) \leq \nu(\bar{x}, \bar{U}_{\bar{x}}) \leq \nu(\bar{F}, \bar{U}) \leq \text{opt}^w$, by Lemma 6.9. But this contradicts Lemma 6.30. \blacksquare

Lemma 6.32 *Let $\bar{x}, \bar{y} \in \bar{F}$ be two facilities. And let $\bar{U}_{\bar{x}, \bar{y}} = \bar{U}_{\bar{x}} \cup \bar{U}_{\bar{y}}$ and $\bar{U}_{-\bar{x}-\bar{y}} = \bar{U} \setminus \bar{U}_{\bar{x}, \bar{y}}$. Under the assumptions of Section 6.2.1, the following holds:*

- (i) *$\bar{U}_{-\bar{x}-\bar{y}}$ excludes at least two heavy points.*
- (ii) *If h and h' are two heavy points excluded by $\bar{U}_{-\bar{x}-\bar{y}}$, then $|\bar{U}_{\bar{x}, \bar{y}}| \geq w(h) + w(h')$.*

Proof: (i) By Claim 6.31, $\overline{U}_{-\overline{x}-\overline{y}}$ can only include at most $k - 2$ heavy points, and may partly-include another heavy point. Since there are $k + 1$ heavy points in total, there must be at least $(k + 1) - (k - 2) - 1 = 2$ heavy points excluded by $\overline{U}_{-\overline{x}-\overline{y}}$.

(ii) Assume, for the sake of contradiction, that $|\overline{U}_{\overline{x},\overline{y}}| < \mathbf{w}(h) + \mathbf{w}(h')$. Let $M = \overline{U}_{-\overline{x}-\overline{y}} \cup \{h, h'\}^{\mathbf{w}}$, and $\overline{F}' = \overline{F} - \overline{x} - \overline{y} + h + h'$. We have

$$\nu(\overline{F}', M) \leq \nu(\overline{F} - \overline{x} - \overline{y}, \overline{U}_{-\overline{x}-\overline{y}}) + \nu(\{h, h'\}, h^{\mathbf{w}} \cup h'^{\mathbf{w}}) \leq \nu(\overline{F}, \overline{U}) + 0 = \nu(\overline{F}, \overline{U}). \quad (14)$$

Furthermore, since $|\overline{U}_{\overline{x},\overline{y}}| < \mathbf{w}(h) + \mathbf{w}(h')$, we have

$$|M| = |\overline{U}_{-\overline{x}-\overline{y}}| + \mathbf{w}(h) + \mathbf{w}(h') = |\overline{U}| - |\overline{U}_{\overline{x},\overline{y}}| + \mathbf{w}(h) + \mathbf{w}(h') > |\overline{U}|.$$

If $|M| \leq n - m$ then by Observation 6.2 (i) and Eq. (14), we have

$$\begin{aligned} \mathcal{A}_m(\overline{F}', \mathbf{P}^{\mathbf{w}}, \varrho) &\leq \nu(\overline{F}', M) + (n - m - |M|)\varrho \\ &< \nu(\overline{F}, \overline{U}) + (n - m - |\overline{U}|)\varrho = \mathcal{A}_m(\overline{F}, \mathbf{P}^{\mathbf{w}}, \varrho), \end{aligned}$$

since $|M| > |\overline{U}|$. This contradicts the optimality of \overline{F} .

If $|M| > n - m$ then let $M' = \mathbf{N}_{n-m}(\overline{F}', M)$. Now, apply the above argument to \overline{F}' and M' , we similarly get a contradiction. \blacksquare

Lemma 6.10 (restatement) Under the assumptions of Section 6.2.1, for any $\overline{x}, \overline{y} \in \overline{F}$ and $q \in \mathbf{P}$, we have $\nu(q, \overline{U}_{\overline{x}} \cup \overline{U}_{\overline{y}}) \geq 15\text{opt}^{\mathbf{w}}$.

Proof: Assume, for the sake of contradiction, that $\nu(q, \overline{U}_{\overline{x}} \cup \overline{U}_{\overline{y}}) < 15\text{opt}^{\mathbf{w}}$. Let $\overline{U}_{\overline{x},\overline{y}} = \overline{U}_{\overline{x}} \cup \overline{U}_{\overline{y}}$ and $\overline{U}_{-\overline{x}-\overline{y}} = \overline{U} \setminus \overline{U}_{\overline{x},\overline{y}}$. There are several possibilities.

- (i) $\overline{U}_{\overline{x},\overline{y}}$ includes two heavy points, h and h' . Let $G = h^{\mathbf{w}} \cup h'^{\mathbf{w}}$. Since $G \subseteq \overline{U}_{\overline{x},\overline{y}}$, we have $\nu(q, G) \leq \nu(q, \overline{U}_{\overline{x}} \cup \overline{U}_{\overline{y}}) \leq 15\text{opt}^{\mathbf{w}}$ in this case. However, this is impossible, by Lemma 6.30.
- (ii) $\overline{U}_{\overline{x},\overline{y}}$ includes one heavy point h , partly-include another heavy point h' , and excludes every other. In this case, h and h' are excluded by $\overline{U}_{-\overline{x}-\overline{y}}$, and as such, $|\overline{U}_{\overline{x},\overline{y}}| \geq \mathbf{w}(h) + \mathbf{w}(h')$, by Lemma 6.32 (ii). Now, setting $G = \overline{U}_{\overline{x},\overline{y}}$, we have a contradiction, by Lemma 6.30.
- (iii) $\overline{U}_{\overline{x},\overline{y}}$ excludes every heavy point except for h . In this case, h is excluded by $\overline{U}_{-\overline{x}-\overline{y}}$. In addition, By Lemma 6.32 (i), at least two heavy points are excluded by $\overline{U}_{-\overline{x}-\overline{y}}$, and as such, there must be another heavy point, say h' , excluded by $\overline{U}_{-\overline{x}-\overline{y}}$. Now, by Lemma 6.32 (ii), we have $|\overline{U}_{\overline{x},\overline{y}}| \geq \mathbf{w}(h) + \mathbf{w}(h')$. Now, setting $G = \overline{U}_{\overline{x},\overline{y}}$, we have a contradiction, by Lemma 6.30.
- (iv) $\overline{U}_{\overline{x},\overline{y}}$ excludes every heavy point. In this case, by Lemma 6.32 (i), at least two heavy points, say h and h' , are excluded by $\overline{U}_{-\overline{x}-\overline{y}}$, and as such, by Lemma 6.32 (ii), we have $|\overline{U}_{\overline{x},\overline{y}}| \geq \mathbf{w}(h) + \mathbf{w}(h')$. Now, setting $G = \overline{U}_{\overline{x},\overline{y}}$, we have a contradiction, by Lemma 6.30. \blacksquare

7 Conclusions

In this paper, we present the first efficient (i.e., polynomial time) constant-factor approximation algorithm for the k -median with outliers problem. A natural direction for future research is to extend the techniques used to other optimization problems with non-trivial global constraints, such as the capacitated k -median problem.

The new *successive local search* method, used in Section 3.2, is fairly general and should be applicable to other problems, since many combinatorial optimization problems can be reduced to their corresponding penalty versions. To use this method, however, it is crucial to bound the number of points that receive penalty. This is not easy and depends on the problem at hand.

8 Acknowledgments

The author thanks Chandra Chekuri and Sariel Har-Peled for their helpful comments on the manuscript. The example presented in Appendix A is due to Yusu Wang.

References

- [AGK⁺04] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k -median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.
- [AH98] E. Arkin and R. Hassin. On local search for weighted k -set packing. *Math. of Oper. Res.*, 23:640–648, 1998.
- [AR06] A. Aboud and Y. Rabani. Correlation clustering with penalties. manuscript, 2006.
- [BCR01] Y. Bartal, M. Charikar, and D. Raz. Approximating min-sum k -clustering in metric spaces. In *Proc. 33rd Annu. ACM Sympos. Theory Comput.*, pages 11–20, 2001.
- [CG99] M. Charikar and S. Guha. Improved combinatorial algorithms for the facility location and k -median problems. In *Proc. 40th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 378–388, 1999.
- [CGTS02] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k -median problem. *J. Comput. Sys. Sci.*, 65(1):129–149, 2002.
- [CKMN01] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *Proc. 12th ACM-SIAM Sympos. Discrete Algorithms*, pages 642–651, 2001.
- [CR05] J. Chuzhoy and Y. Rabani. Approximating k -median with non-uniform capacities. In *Proc. 16th ACM-SIAM Sympos. Discrete Algorithms*, pages 952–958, 2005.
- [Goe06] M. X. Goemans. Minimum bounded degree spanning trees. In *Proc. 47th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 273–282, 2006.
- [HM04] S. Har-Peled and S. Mazumdar. Coresets for k -means and k -median clustering and their applications. In *Proc. 36th Annu. ACM Sympos. Theory Comput.*, pages 291–300, 2004.
- [JMM⁺03] K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V. V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP. *J. Assoc. Comput. Mach.*, 50(6):795–824, 2003.
- [JV01] K. Jain and V. V. Vazirani. Approximation algorithms for metric facility location and k -median problems using the primal-dual schema and Lagrangian relaxation. *J. Assoc. Comput. Mach.*, 48(2):274–296, 2001.

- [KBP03] S. Khuller, R. Bhatia, and R. Pless. On local search and placement of meters in networks. *SIAM J. Comput.*, pages 470–487, 2003.
- [Khu05] S. Khuller. Problems column. *ACM Trans. Algorithms*, 1(1):157–159, 2005.
- [KPR00] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman. Analysis of a local search heuristic for facility location problems. *J. Algorithms*, 37(1):146–188, 2000.
- [KR00] J. Konemann and R. Ravi. A matter of degree: improved approximation algorithms for degree-bounded minimum spanning trees. In *Proc. 32nd Annu. ACM Sympos. Theory Comput.*, pages 537–546, 2000.
- [LV92] J-H Lin and J. S. Vitter. ε -approximations with minimum packing constraint violation (extended abstract). In *Proc. 24th Annu. ACM Sympos. Theory Comput.*, pages 771–782, 1992.
- [Mah04] M. Mahdian. *Facility Location and the Analysis of Algorithms through Factor-Revealing Programs*. Ph.D. dissertation, MIT, Department of Computer Science, 2004.
- [RRPS04] D. Ren, I. Rahal, W. Perrizo, and K. Scott. A vertical distance-based outlier detection method with local pruning. In *Proc. 13th ACM Conf. Information and Knowledge Management*, pages 279–284, 2004.
- [SL07] M. Singh and L. C. Lau. Approximating minimum bounded degree spanning trees to within one of optimal. In *Proc. 39th Annu. ACM Sympos. Theory Comput.*, pages 661–670, 2007.
- [ST06] Z. Svitkina and E. Tardos. Approximation algorithm for facility location with hierarchical facility costs. In *Proc. 17th ACM-SIAM Sympos. Discrete Algorithms*, pages 1088–1097, 2006.

A A counter example for the standard local search algorithm

The *local search* method uses the concept of neighborhood. Specifically, it starts with an initial feasible solution and then repeatedly searches the neighborhood of the current solution for a better solution until it cannot be improved any further (that is, it reaches a *locally optimal* solution). In our problem, given a solution S , which is a set of k facilities, a *neighbor* X of S is a set of k facilities such that $|X \cap S| \geq k - b$, for some constant b . Namely, X is obtained by swapping at most b facilities of S with facilities outside S .

Let $N(S)$ denote the set of all neighbors of S . The local search algorithm repeatedly replace S by a better center set in $N(S)$ as long as such center set exists. In what follows, we present an example demonstrating that a locally optimal solution yielded by this standard local search algorithm may have arbitrarily bad performance, compared to the globally optimal solution (namely, the *locality gap* can be arbitrarily large).

Suppose that $n \gg m \gg k > 1$, and $u = m/(k-1)$ is an integer. Consider an input V as follows: the set V is partitioned into disjoint subsets $B, C_1, \dots, C_{k-1}, D_1, \dots, D_{k-2}$, and E , such that the distance between any pair of points belonging to different subsets is very large. Suppose that

- (i) $|B| = n - 2m$ and $d(p, q) = 0$, for any $p, q \in B$.
- (ii) For each $i = 1, \dots, k-1$, we have $|C_i| = u$ and $d(p, q) = \beta$, for any $p, q \in C_i$.

(iii) For each $j = 1, \dots, k-2$, we have $|D_j| = u-1$ and $d(p, q) = 0$, for any $p, q \in D_j$.

(iv) $|E| = u+k-2$ and $d(p, q) = \gamma$, for any $p, q \in E$.

We further assume that $u \gg k$ and $\gamma < (u-1)\beta < 2\gamma$.

For $Y \in \{B, C_1, \dots, C_{k-1}, D_1, \dots, D_{k-2}, E\}$, let $f(Y)$ denote an arbitrary point in Y . Consider a solution $\mathcal{S} = \{f(B), f(D_1), \dots, f(D_{k-2}), f(E)\}$, namely, we place a facility in each of the subsets $B, D_1, \dots, D_{k-2}, E$. The $m = (k-1)u$ outliers in this solution are the points in C_1, \dots, C_{k-1} .

Claim A.1 *The solution \mathcal{S} is locally optimal, incurring a cost of $(u+k-3)\gamma$, if $\mathbf{b} < k-1$.*

Proof: In the solution \mathcal{S} , serving B by $f(B)$ costs 0, serving D_j by $f(D_j)$ costs 0, for $j = 1, \dots, k-2$, and serving E by $f(E)$ costs $(u+k-3)\gamma$. Therefore, the cost of \mathcal{S} is $(u+k-3)\gamma$.

To see that \mathcal{S} is locally optimal, observe that we cannot swap $f(B)$ out, because otherwise we need to serve points in B by a facility not in B (recall that $|B| = n-2m$ and $n \gg m$), which is very costly. For the same reason, we cannot swap $f(E)$ out. Suppose that we swap $\mathbf{b}' \leq \mathbf{b}$ facilities, say $f(D_1), \dots, f(D_{\mathbf{b}'})$, with $f(C_1), \dots, f(C_{\mathbf{b}'})$, and let \mathcal{S}' denote the resulting solution. That is, $\mathcal{S}' = \{f(B), f(C_1), \dots, f(C_{\mathbf{b}'}), f(D_{\mathbf{b}'+1}), \dots, f(D_{k-2}), f(E)\}$. It is easy to verify that the cost of \mathcal{S}' is $\mathbf{b}' \cdot (u-1)\beta + (u+k-3-\mathbf{b}')\gamma$, which is greater than $(u+k-3)\gamma$ since $(u-1)\beta > \gamma$. This implies that we cannot improve \mathcal{S} by swapping $\leq \mathbf{b}$ facilities. \blacksquare

It is easy to verify that the optimal solution is $\bar{\mathcal{S}} = (B, C_1, \dots, C_{k-1})$, which has a cost of $(k-1)(u-1)\beta$. Since $u \gg k$ and $(u-1)\beta$ is only slightly larger than γ , it follows that the locality gap

$$\frac{(u+k-3)\gamma}{(k-1)(u-1)\beta} > \frac{u+k-3}{2(k-1)},$$

since $(u-1)\beta < 2\gamma$ (by assumption). This may be arbitrarily large, depending on the ratio u/k .

Now, note that $\text{MO}(k, V, m)$ can be reduced to $\text{PMO}(k, V, \varrho, m)$, by setting $\varrho = \infty$. Since $\text{MO}(k, V, m)$ cannot be solved by the above local search algorithm, neither can $\text{PMO}(k, V, \varrho, m)$.

B Proof of Lemma 4.6

Proof: (i) We will prove that $|A_m(C, \mathbf{P}^w) - A_m(C, \mathbf{P})| \leq A_m(C_+, \mathbf{P})$. Because, by Claim 4.3, $A_m(C_+, \mathbf{P}) \leq 3\text{opt}$, this implies the claim. In the following, we focus on the case when $A_m(C, \mathbf{P}^w) \leq A_m(C, \mathbf{P})$, since the other case is similar.

Let Q and Q' be the (multi)sets of $n-m$ nearest points to C in \mathbf{P}^w and \mathbf{P} , respectively. By the definition of ϕ and w (see Section 2.2), there exists a set $Q'' \subseteq \mathbf{P}$ of $n-m$ points such that $\{\phi(p) \mid p \in Q''\} = Q$. Therefore,

$$\begin{aligned} \nu(C, Q'') - \nu(C, Q) &= \sum_{p \in Q''} (d(p, C) - d(\phi(p), C)) \leq \sum_{p \in \mathbf{P}} |d(p, C) - d(\phi(p), C)| \\ &\leq \sum_{p \in \mathbf{P}} |d(p, \phi(p))| = A_m(C_+, \mathbf{P}). \end{aligned}$$

In addition, we have $\nu(C, Q') - \nu(C, Q'') \leq 0$, since Q' is the set of $n-m$ nearest points to C in \mathbf{P} . It thus follows that

$$\begin{aligned} |A_m(C, \mathbf{P}^w) - A_m(C, \mathbf{P})| &= A_m(C, \mathbf{P}) - A_m(C, \mathbf{P}^w) = \nu(C, Q') - \nu(C, Q) \\ &= (\nu(C, Q') - \nu(C, Q'')) + (\nu(C, Q'') - \nu(C, Q)) \\ &\leq 0 + A_m(C_+, \mathbf{P}). \end{aligned}$$

(ii) Suppose that C_o is an optimal solution for $\text{MO}(k, P, m)$, namely $|C_o| = k$ and $A_m(C_o, P) = \text{opt}$. Then, by (i), we have

$$\text{opt}^w \leq A_m(C_o, P^w) \leq 3\text{opt} + A_m(C_o, P) = 4\text{opt}.$$

It follows that $A_m(C, P^w) \leq \gamma \text{opt}^w \leq 4\gamma \text{opt}$, which implies by (i) the claim. \blacksquare

C Proof of Lemma 5.9

Recall that the set J consists of the (distinct) $k - k'$ heaviest points excluded by \mathcal{Y} , and $Z = \mathcal{Y} \cup J^w$ is the set of points clustered by the solution \mathbf{C} output by GREEDYMERGE .

Claim C.1 *If \mathcal{X} does not contain any light point (namely, $l_w(\mathcal{X}) = 0$), then $|Z| \geq n - m$.*

Proof: Since $l_w(\mathcal{X}) = 0$ and $\mathcal{Y} \subseteq \mathcal{X}$, it follows that $l_w(\mathcal{Y}) = 0$. As such, by Eq. (5)_{p11}, we have $\text{mass}(\mathcal{Y}) = h_w(\mathcal{Y}) + \xi \cdot l_w(\mathcal{Y}) = h_w(\mathcal{Y})$. On the other hand, by Eq. (6)_{p12}, we have $\text{mass}(\mathcal{Y}) \geq \gamma_+ + k'$, implying $h_w(\mathcal{Y}) \geq \gamma_+ + k'$. Now, by the way GREEDYMERGE works, the set Z contains $h_w(\mathcal{Y}) + h_w(J^w) \geq (\gamma_+ + k') + (k - k') = \gamma_+ + k = k_+$ heavy points, since $Z = \mathcal{Y} \cup J^w$ and $h_w(J^w) = k - k'$, by Claim 5.4. That is, Z contains all the heavy points of \mathbf{C}_+ , which implies that $|Z| \geq w(\mathbf{C}_+) = n - m$. \blacksquare

As such, in the following, we assume that $l_w(\mathcal{X}) > 0$. Recall that $h_w(P^w \setminus Z)$ is the number of distinct heavy points in $P^w \setminus Z$. (Note that $P^w \setminus Z$ is the set of outliers for the clustering of Z computed by GREEDYMERGE .)

Lemma C.2 *If \mathcal{X} contains a light point, then $l_w(\mathcal{Y}) \geq h_w(P^w \setminus Z) / \xi$ (see Eq. (1)_{p5}).*

Proof: Since $Z = \mathcal{Y} \cup J^w$ and $h_w(J^w) = k - k'$, by Claim 5.4, we have

$$\begin{aligned} h_w(\mathcal{Y}) &= h_w(Z) - h_w(J^w) = h_w(P^w) - h_w(P^w \setminus Z) - (k - k') \\ &= k_+ - h_w(P^w \setminus Z) - k + k' = \gamma_+ + k' - h_w(P^w \setminus Z), \end{aligned} \quad (15)$$

since $h_w(P^w) = k_+$ and $k_+ - k = \gamma_+$. It follows that

$$\left(\gamma_+ + k' - h_w(P^w \setminus Z) \right) + \xi \cdot l_w(\mathcal{Y}) = h_w(\mathcal{Y}) + \xi \cdot l_w(\mathcal{Y}) = \text{mass}(\mathcal{Y}) \geq \gamma_+ + k',$$

by Eq. (5)_{p11} and Eq. (6)_{p12}. This implies that $l_w(\mathcal{Y}) \geq h_w(P^w \setminus Z) / \xi$. \blacksquare

Claim C.3 *If \mathcal{X} contains a light point, then $h_w(P^w \setminus Z) \leq h_w(P^w \setminus \mathcal{X}) - 1$.*

Proof: We have

$$\begin{aligned} \text{mass}(\mathcal{X}) - \text{mass}(\mathcal{Y}) &= \left(h_w(\mathcal{X}) + \xi \cdot l_w(\mathcal{X}) \right) - \left(h_w(\mathcal{Y}) + \xi \cdot l_w(\mathcal{Y}) \right) \\ &= h_w(\mathcal{X}) - h_w(\mathcal{Y}) + \xi(l_w(\mathcal{X}) - l_w(\mathcal{Y})) \\ &\geq h_w(\mathcal{X}) - h_w(\mathcal{Y}), \end{aligned}$$

since $l_w(\mathcal{X}) \geq l_w(\mathcal{Y})$ (implied by $\mathcal{X} \supseteq \mathcal{Y}$) and $\xi \geq 0$, by Observation 5.1 (iii). As such,

$$\begin{aligned} h_w(\mathcal{X}) &\leq \text{mass}(\mathcal{X}) - \text{mass}(\mathcal{Y}) + h_w(\mathcal{Y}) \\ &\leq (k_+ - 1) - (\gamma_+ + k') + (\gamma_+ + k' - h_w(P^w \setminus Z)) \\ &= k_+ - h_w(P^w \setminus Z) - 1, \end{aligned}$$

since $\text{mass}(\mathcal{X}) = k_+ - 1$ (by Claim 5.2), $\text{mass}(\mathcal{Y}) \geq \gamma_+ + k'$ (by Eq. (6)_{p12}), and Eq. (15). It follows that

$$h_w(\mathbf{P}^w \setminus Z) \leq k_+ - h_w(\mathcal{X}) - 1 = h_w(\mathbf{P}^w) - h_w(\mathcal{X}) - 1 = h_w(\mathbf{P}^w \setminus \mathcal{X}) - 1.$$

since $h_w(\mathbf{P}^w) = k_+$. ■

Given a set $Q \subseteq \mathbf{P}^w$ of heavy points, the *average weight* of Q is $|Q|/h_w(Q)$.

Observation C.4 *Let Q and Q' be two sets of heavy points of \mathbf{P}^w , where $Q \subseteq Q'$. Let S be a subset of Q' , consisting of the $h_w(S)$ lightest points in Q' . If $h_w(S) \leq h_w(Q)$ then*

$$\frac{|S|}{h_w(S)} \leq \frac{|Q|}{h_w(Q)}.$$

Given a set $Q \subseteq \mathbf{P}^w$, let $H_w(Q)$ be the multiset of all the heavy points in Q , and $L_w(Q)$ be the set of all the light points in Q .

Lemma C.5 *If \mathcal{X} contains a light point, then $|H_w(\mathbf{P}^w \setminus Z)| \leq h_w(\mathbf{P}^w \setminus Z)/\xi$.*

Proof: By the construction of \mathcal{X} , there exists a point $p \in \mathbf{P}^w \setminus \mathcal{X}$ such that $|\mathcal{X}| + w(p) > n - m$. Let q be the heaviest point in $\mathbf{P}^w \setminus \mathcal{X}$. Clearly, $w(q) \geq w(p)$, and as such,

$$|H_w(\mathcal{X})| + |L_w(\mathcal{X})| + w(q) = |\mathcal{X}| + w(q) \geq |\mathcal{X}| + w(p) > n - m.$$

On the other hand, we have $|H_w(\mathcal{X})| + |H_w(\mathbf{P}^w \setminus \mathcal{X})| = |H_w(\mathbf{P}^w)| = w(\mathbf{C}_+) = n - m$. It thus follows that $|L_w(\mathcal{X})| + w(q) > n - m - |H_w(\mathcal{X})| = |H_w(\mathbf{P}^w \setminus \mathcal{X})|$, or equivalently,

$$|H_w(\mathbf{P}^w \setminus \mathcal{X})| - w(q) < |L_w(\mathcal{X})|.$$

Let Q be the set of all the heavy points in $\mathbf{P}^w \setminus \mathcal{X}$ except for q . As such, we have $|Q| = |H_w(\mathbf{P}^w \setminus \mathcal{X})| - w(q) < |L_w(\mathcal{X})| = l_w(\mathcal{X})$ and $h_w(Q) = h_w(\mathbf{P}^w \setminus \mathcal{X}) - 1$. Therefore, the average weight of Q is

$$\frac{|Q|}{h_w(Q)} < \frac{l_w(\mathcal{X})}{h_w(\mathbf{P}^w \setminus \mathcal{X}) - 1} = \frac{l_w(\mathcal{X})}{h_w(\mathbf{P}^w) - h_w(\mathcal{X}) - 1} = \frac{l_w(\mathcal{X})}{k_+ - h_w(\mathcal{X}) - 1} = \frac{1}{\xi},$$

see Eq. (1)_{p5}. Note that $Q \subseteq H_w(\mathbf{P}^w \setminus \mathcal{X}) \subseteq H_w(\mathbf{P}^w \setminus \mathcal{Y})$, since $\mathcal{Y} \subseteq \mathcal{X}$. By the way GREEDYMERGE works, $H_w(\mathbf{P}^w \setminus Z)$ is a subset of $H_w(\mathbf{P}^w \setminus \mathcal{Y})$, consisting of the $h_w(\mathbf{P}^w \setminus Z)$ lightest points in $H_w(\mathbf{P}^w \setminus \mathcal{Y})$. Furthermore, we have

$$h_w(\mathbf{P}^w \setminus Z) \leq h_w(\mathbf{P}^w \setminus \mathcal{X}) - 1 = h_w(Q),$$

by Claim C.3. Therefore, by Observation C.4, we have

$$\frac{|H_w(\mathbf{P}^w \setminus Z)|}{h_w(\mathbf{P}^w \setminus Z)} \leq \frac{|Q|}{h_w(Q)} < \frac{1}{\xi}. \quad \blacksquare$$

Lemma 5.9 (*restatement*) $|Z| \geq n - m$.

Proof: Claim C.1 handles the case $l_w(\mathcal{X}) = 0$. So, consider the case when $l_w(\mathcal{X}) > 0$. The total weight of Z is the number of light points in \mathcal{Y} (note that there is no light points in J^w) plus the total weight of the heavy points in Z , namely,

$$\begin{aligned} |Z| &= l_w(Z) + |H_w(Z)| = l_w(\mathcal{Y}) + |H_w(Z)| = l_w(\mathcal{Y}) + |H_w(\mathbf{P}^w)| - |H_w(\mathbf{P}^w \setminus Z)| \\ &\geq \frac{h_w(\mathbf{P}^w \setminus Z)}{\xi} + w(\mathbf{C}_+) - \frac{h_w(\mathbf{P}^w \setminus Z)}{\xi} = n - m, \end{aligned}$$

by Lemma C.2 and Lemma C.5. ■

D Perturbation of the distance function d

We first compute a real number α as an estimate of opt .

Lemma D.1 *One can compute in polynomial time a real number α such that $\text{opt}/(3n) \leq \alpha \leq \text{opt}$.*

Proof: The problem of k -center with m outliers (CO for short) is to compute a set of m outliers so as to minimize the cost of the k -center clustering of the remaining points. Let $\text{opt}_{\text{co}}(\mathbf{P}, k, m)$ be the cost of the optimal solution for the CO instance with input point set \mathbf{P} . It is easy to verify that $\text{opt}/n \leq \text{opt}_{\text{co}}(\mathbf{P}, k, m) \leq \text{opt}$. We use the algorithm for CO presented in [CKMN01] to compute β such that $\beta/3 \leq \text{opt}_{\text{co}}(\mathbf{P}, k, m) \leq \beta$. The claim now follows by setting $\alpha = \beta/3$. ■

Given parameters $0 < \varepsilon < 1$ and $1 \leq \gamma$, we shall perturb the distance function d , and denote the resulting new distance function by d' . We claim that if one can compute a set \mathbf{C} of k facilities such that $A_m(\mathbf{C}, \mathbf{P}) \leq \gamma \text{opt}$ under the distance function d' , then it holds $A_m(\mathbf{C}, \mathbf{P}) \leq (1 + \varepsilon)\gamma \text{opt}$ under the original distance function d . We omit the easy proof here, and only specify the perturbation scheme in the following.

Let $\Delta = \varepsilon\alpha/(2n)$, and let $\tau(p, q)$ be a small random real number such that $0 \leq \tau(p, q) \leq \Delta/2$, for all $p, q \in \mathbf{P}$ (note that $\tau(p, q)$ is independent for every $p, q \in \mathbf{P}$). For each pair $p, q \in \mathbf{P}$, if $d(p, q) > 6\gamma n\alpha$ then let $d'(p, q) = 6\gamma n\alpha + \Delta + \tau(p, q)$, otherwise, let $d'(p, q) = d(p, q) + \Delta + \tau(p, q)$. It is easy to verify that, under the distance function d' , the inter-point distances are distinct (with high probability), the ratio between the maximum inter-point distance and the minimum inter-point distance is $O(\gamma n^2/\varepsilon)$, and $d'(x, y) + d'(y, z) \geq d'(x, z)$ holds for every $x, y, z \in \mathbf{P}$.