A Dependent LP-rounding Approach for the k-Median Problem[⋆]

Moses Charikar[1] and Shi Li[1]

Department of computer science, Princeton University, Princeton NJ 08540, USA

Abstract. In this paper, we revisit the classical k-median problem. Using the standard LP relaxation for k-median, we give an efficient algorithm to construct a probability distribution on sets of k centers that matches the marginals specified by the optimal LP solution. Analyzing the approximation ratio of our algorithm presents significant technical difficulties: we are able to show an upper bound of 3.25. While this is worse than the current best known $3 + \epsilon$ guarantee of [2], because: (1) it leads to 3.25 approximation algorithms for some generalizations of the k-median problem, including the k-facility location problem introduced in [10], (2) our algorithm runs in $\tilde{O}(k^3 n^2/\delta^2)$ time to achieve $3.25(1+\delta)$-approximation compared to the $O(n^8)$ time required by the local search algorithm of [2] to guarantee a 3.25 approximation, and (3) our approach has the potential to beat the decade old bound of $3 + \epsilon$ for k-median.
We also give a 34-approximation for the knapsack median problem, which greatly improves the approximation constant in [13]. Using the same technique, we also give a 9-approximation for matroid median problem introduced in [11], improving on their 16-approximation.

Keywords: Approximation, k-Median Problem, Dependent Rounding

1 Introduction

In this paper, we present a novel LP rounding algorithm for the metric k-median problem which achieves approximation ratio 3.25. For the k-median problem, we are given a finite metric space $(\mathcal{F} \cup \mathcal{C}, d)$ and an integer $k \geq 1$, where $\mathcal{F}$ is a set of facility locations and $\mathcal{C}$ is a set of clients. Our goal is to select k facilities to open, such that the total connection cost for all clients in $\mathcal{C}$ is minimized, where the connection cost of a client is its distance to its nearest open facility. When $\mathcal{F} = \mathcal{C} = X$, the set of points with the same nearest open facility is known as a cluster and thus the sum measures how well X can be partitioned into k clusters. The k-median problem has numerous applications, starting from clustering to data mining [3], to assigning efficient sources of supplies to minimize the transportation cost([12, 16]).

The problem is NP-hard and has received a lot of attention ([15], [6], [7], [10], [1]). The best known approximation factor is $3 + \epsilon$ approximation due to

[⋆] A full version of this paper is available at the authors' web pages

[2]. Jain et al. [9] proved that the k-median problem is $1 + 2/e \approx 1.736$-hard to approximate.

Our algorithm (like several previous ones) for the k-median problem is based on the following natural LP relaxation:

$$\text{LP(1)} \qquad \min \quad \textstyle\sum_{i\in\mathcal{F},j\in\mathcal{C}} d(i,j)x_{i,j} \qquad \text{s.t.}$$

$$\sum_{i\in\mathcal{F}} x_{i,j} = 1, \quad \forall j \in \mathcal{C}; \qquad x_{i,j} \le y_i, \quad \forall i \in \mathcal{F}, j \in \mathcal{C};$$

$$\sum_{i\in\mathcal{F}} y_i \le k; \qquad x_{i,j}, y_i \in [0,1], \quad \forall i \in \mathcal{F}, j \in \mathcal{C}.$$

It is known that the LP has an integrality gap of 2. On the positive side, [1] showed that the integrality gap is at most 3 by giving an exponential time rounding algorithm.

Very recently, Kumar [13] gave a (large) constant-factor approximation algorithm for a generalization of the k-median problem, which is called knapsack median problem. In the problem, each facility $i \in \mathcal{F}$ has an opening cost f_i and we are given a budget M. The goal is to open a set of facilities such that their total opening cost is at most M, and minimize the total connection cost. When $M = k$ and $f_i = 1$ for every facility $i \in \mathcal{F}$, the problem becomes k-median.

Krishnaswamy et al. [11] introduced another generalization of k-median, called matroid-median problem. In the problem, the set of open facilities has to form an independent set of some given matroid. [11] gave a 16-approximation for this problem, assuming there is a separation oracle for the matroid polytope.

1.1 Our Results

We give a simple and efficient rounding procedure. Given an LP solution, we open a set of k facilities from some distribution and connect each client j to its closest open facility, such that the expected connection cost of j is at most 3.25 times its fractional connection cost. This leads to a 3.25 approximation for k-median. Though the provable approximation ratio is worse than that of the current best algorithm, we believe the algorithm (and particularly our approach) is interesting for the following reasons:

Firstly, our algorithm is more efficient than the $3+\epsilon$-approximation algorithm with the same approximation guarantee. The bottleneck of our algorithm is solving the LP, for which we can apply Young's fast algorithm for the k-median LP [17].

Secondly, our approach has the potential to beat the decade old $3 + \epsilon$-approximation algorithm for k-median. In spite of the simplicity of our algorithm, we are unable to exploit its full potential due to technical difficulties in the analysis. Our upper bound of 3.25 is not tight. The algorithm has some parameters which we have instantiated for ease of analysis. It is possible that the algorithm with these specific choices gives an approximation ratio strictly better than 3; further there is additional room for improvement by making a judicious choice of algorithm parameters.

The distribution of solutions produced by the algorithm satisfies marginal conditions and negative correlation. Consequently, the algorithm can be easily extended to solve the k-median problem with facility costs and the k-median problem (called k-facility location problem) with multiple types of facilities, both introduced in [10]. The techniques of this paper yield a factor 3.25 algorithm for the two generalizations.

Based on our techniques for the k-median problem, we give a 34-approximation algorithm for the knapsack median problem, which greatly improves the constant approximation given by [13].(The constant was 2700.) Following the same line of the algorithm, we can give a 9-approximation for the matroid-median problem, improving on the 16-approximation in [11].

2 The Approximation Algorithm for the k-Median Problem

Our algorithm is inspired by the $6\frac{2}{3}$-approximation for k-median by [7] and the clustered rounding approach of Chudak and Shmoys [8] for facility location as well as the analysis of the 1.5-approximation for UFL problem by [4]. In particular, we are able to save the additive factor of 4 that is lost at the beginning of the $6\frac{2}{3}$-approximation algorithm by [7], using some ideas from the rounding approaches for facility location.

We first give with a high level overview of the algorithm. A simple way to match the marginals given by the LP solution is to interpret the y_i variables as probabilities of opening facilities and sample independently for each i. This has the problem that with constant probability, a client j could have no facility opened close to j. In order to address this, we group fractional facilities into bundles, each containing a total fractional of between 1/2 and 1. At most one facility is opened in each bundle and the probability that some facility in a bundle is picked is exactly the volume, i.e. the sum of y_i values for the bundle.

Creating bundles reduces the uncertainty of the sampling process. E.g. if the facilities in a bundle of volume 1/2 are sampled independently, with probability $e^{-1/2}$ in the worst case, no facility will be open; while sampling the bundle as a single entity reduces the probability to 1/2. The idea of creating bundles alone does not reduce the approximation ratio to a constant, since still with some non-zero probability, no nearby facilities are open.

In order to ensure that clients always have an open facility within expected distance comparable to their LP contribution, we pair the bundles. Each pair now has at least a total fraction of 1 facility and we ensure that the rounding procedure always picks one facility in each pair. The randomized rounding procedure makes independent choices for each pair of bundles and for fractional facilities that are not in any bundle. This produces k facilities in expectation. We get exactly k facilities by replacing the independent rounding by a dependent rounding procedure with negative correlation properties so that our analysis need only consider the independent rounding procedure. (The technique of dependent rounding was used in [5] to approximate the fault-tolerant facility location problem.)

Now we proceed to give more details. We solve LP(1) to obtain a fractional solution (x, y). By splitting one facility into many if necessary, we can assume $x_{i,j} \in \{0, y_i\}$. We remove all facilities i from $\mathcal{C}$ with $y_i = 0$. Let $\mathcal{F}_j = \{i \in \mathcal{F} : x_{i,j} > 0\}$. So, instead of using x and y, we shall use $(y, \{\mathcal{F}_j | j \in \mathcal{C}\})$ to denote a solution.

For a subset of facilities $\mathcal{F}' \subseteq \mathcal{F}$, define $\mathsf{vol}(\mathcal{F}') = \sum_{i \in \mathcal{F}'} y_i$ to be the *volume* of $\mathcal{F}'$. So, $\mathsf{vol}(\mathcal{F}_j) = 1, \forall j \in \mathcal{C}$. W.L.O.G, we assume $\mathsf{vol}(\mathcal{F}) = k$. Denote by $d(j, \mathcal{F}')$ the average distance from j to $\mathcal{F}'$ w.r.t weights y, i.e, $d(j, \mathcal{F}') = \sum_{i \in \mathcal{F}'} y_i d(j, i)/\mathsf{vol}(\mathcal{F}')$. Define $d_{av}(j) = \sum_{i \in \mathcal{F}_j} y_i d(i, j)$ to be the connection cost of j in the fractional solution. For a client j, let $B(j, r)$ denote the set of facilities that have distance strictly smaller than r to j.

Our rounding algorithm consists of 4 phases, which we now describe.

2.1 Filtering Phase

We begin our algorithm with a filtering phase, where we select a subset $\mathcal{C}' \subseteq \mathcal{C}$ of clients. $\mathcal{C}'$ has two properties: (1) The clients in $\mathcal{C}'$ are far away from each other. With this property, we can guarantee that each client in $\mathcal{C}'$ can be assigned an exclusive set of facilities with large volume. (2) A client in $\mathcal{C} \backslash \mathcal{C}'$ is close to some client in $\mathcal{C}'$, so that its connection cost is bounded in terms of the connection cost of its neighbour in $\mathcal{C}'$. So, $\mathcal{C}'$ captures the connection requirements of $\mathcal{C}$ and also has a nice structure. After this filtering phase, our algorithm is independent of the clients in $\mathcal{C} \backslash \mathcal{C}'$. Following is the filtering phase.

Initially, $\mathcal{C}' = \emptyset, \mathcal{C}'' = \mathcal{C}$. At each step, we select the client $j \in \mathcal{C}''$ with the minimum $d_{av}(j)$, breaking ties arbitrarily, add j to $\mathcal{C}'$ and remove j and all j's that $d(j, j') \leq 4 d_{av}(j')$ from $\mathcal{C}''$. This operation is repeated until $\mathcal{C}'' = \emptyset$.

Lemma 1. *(1) For any $j, j' \in \mathcal{C}', j \neq j', d(j, j') > 4 \max\{d_{av}(j), d_{av}(j')\}$;*
(2) For any $j' \in \mathcal{C} \backslash \mathcal{C}'$, there is a client $j \in \mathcal{C}'$ such that $d_{av}(j) \leq d_{av}(j'), d(j, j') \leq 4 d_{av}(j')$.

We leave the proof of the lemma to the full version of the paper.

2.2 Bundling Phase

Since clients in $\mathcal{C}'$ are far away from each other, each client $j \in \mathcal{C}'$ can be assigned a set of facilities with large volume. To be more specific, for a client $j \in \mathcal{C}'$, we define a set $\mathcal{U}_j$ as follows. Let $R_j = \frac{1}{2} \min_{j' \in \mathcal{C}', j' \neq j} d(j, j')$ be half the distance of j to its nearest neighbour in $\mathcal{C}'$, and $\mathcal{F}'_j = \mathcal{F}_j \cap B(j, 1.5 R_j)$ to be the set of facilities that serve j and are at most $1.5 R_j$ away.[1] A facility i which belongs to at least one $\mathcal{F}'_j$ is *claimed* by the nearest $j \in \mathcal{C}'$ such that $i \in \mathcal{F}'_j$, breaking ties arbitrarily. Then, $\mathcal{U}_j \subseteq \mathcal{F}_j$ is the set of facilities claimed by j.

[1] It is worthwhile to mention the motivation behind the choice of the scalar 1.5 in the definition of $\mathcal{F}'_j$. If we were only aiming at a constant approximation ratio smaller than 4, we could replace 1.5 with 1, in which case the analysis is simpler. On the other hand, we believe that changing 1.5 to ∞ would give the best approximation,

Lemma 2. *The following two statements are true:*
(1) $1/2 \leq \mathsf{vol}(\mathcal{U}_j) \leq 1, \forall j \in \mathcal{C}'$, *and (2)* $\mathcal{U}_j \cap \mathcal{U}_{j'} = \emptyset, \forall j, j' \in \mathcal{C}', j \neq j'$.

Proof. Statement 2 is trivial; we only consider the first one. Since $\mathcal{U}_j \subseteq \mathcal{F}'_j \subseteq \mathcal{F}_j$, we have $\mathsf{vol}(\mathcal{U}_j) \leq \mathsf{vol}(\mathcal{F}_j) = 1$. For a client $j \in \mathcal{C}'$, the closest client $j' \in \mathcal{C}' \setminus \{j\}$ to j has $d(j, j') > 4d_{av}(j)$ by lemma 1. So, $R_j > 2d_{av}(j)$ and the facilities in $\mathcal{F}_j$ that are at most $2d_{av}(j)$ away must be claimed by j. The set of these facilities has volume at least $1 - d_{av}(j)/(2d_{av}(j)) = 1/2$. Thus, $\mathsf{vol}(\mathcal{U}_j) \geq 1/2$.

The sets $\mathcal{U}_j$'s are called *bundles*. Each bundle $\mathcal{U}_j$ is treated as a single entity in the sense that at most 1 facility from it is open, and the probability that 1 facility is open is exactly $\mathsf{vol}(\mathcal{U}_j)$. From this point, a bundle $\mathcal{U}_j$ can be viewed as a single facility with $y = \mathsf{vol}(\mathcal{U}_j)$, except that it does not have a fixed position. We will use the phrase "opening the bundle $\mathcal{U}_j$" the operation that opens 1 facility randomly from $\mathcal{U}_j$, with probabilities $y_i/\mathsf{vol}(\mathcal{U}_j)$.

2.3 Matching Phase

Next, we construct a matching $\mathcal{M}$ over the bundles (or equivalently, over $\mathcal{C}'$). If two bundles $\mathcal{U}_j$ and $\mathcal{U}_{j'}$ are matched, we sample them using a joint distribution. Since each bundle has volume at least 1/2, we can choose a distribution such that with probability 1, at least 1 bundle is open.

We construct the matching $\mathcal{M}$ using a greedy algorithm. While there are at least 2 unmatched clients in $\mathcal{C}'$, we choose the closest pair of unmatched clients $j, j' \in \mathcal{C}'$ and match them.

2.4 Sampling Phase

Following is our sampling phase.

```
for each pair $(j, j') \in \mathcal{M}$ do
    With probability $1 - \mathsf{vol}(\mathcal{U}_{j'})$, open $\mathcal{U}_j$; with probability $1 - \mathsf{vol}(\mathcal{U}_j)$, open $\mathcal{U}_{j'}$; and with probability $\mathsf{vol}(\mathcal{U}_j) + \mathsf{vol}(\mathcal{U}_{j'}) - 1$, open both $\mathcal{U}_j$ and $\mathcal{U}_{j'}$;
end for
If some $j \in \mathcal{C}'$ is not matched in $\mathcal{M}$, open $\mathcal{U}_j$ randomly and independently with probability $\mathsf{vol}(\mathcal{U}_j)$;
For each facility $i$ not in any bundle $\mathcal{U}_j$, open it independently with probability $y_i$.
```

After we selected the open facilities, we connect each client to its nearest open facility. Let C_j denote the connection cost of a client $j \in \mathcal{C}$. Our sampling process opens k facilities in expectation, since each facility i is open with probability y_i. It does not always open k facilities as we promised. In the full version of the paper, we shall prove the following lemma:

in which case the algorithm also seems cleaner (since $\mathcal{F}'_j = \mathcal{F}_j$). However, if the scalar were ∞, the algorithm is hard to analyze due to some technical reasons. So, the scalar 1.5 is selected so that we don't lose too much in the approximation ratio and yet the analysis is still manageable.

Lemma 3. *There is a rounding procedure in which we always open k facilities and the probability that i is open is exactly y_i. The $\mathbb{E}[C_j]$ in this procedure is at most the $\mathbb{E}[C_j]$ in the rounding procedure we described. Moreover, the events that facilities are open are negatively-correlated; that is, for every set S of facilities,*

$$\Pr[\textit{all facilities in } S \textit{ are open}] \leq \prod_{i \in S} y_i.$$

By Lemma 3, it suffices to consider the rounding procedure we described. We shall outline the proof of the 3.25 approximation ratio for the above algorithm in section 3. As a warmup, we conclude this section with a much weaker result:

Lemma 4. *The algorithm gives a constant approximation for k-median.*

Proof. It is enough to show that the ratio between $\mathbb{E}[C_j]$ and $d_{av}(j)$ is bounded, for any $j \in \mathcal{C}$. Moreover, it suffices to consider a client $j \in \mathcal{C}'$. Indeed, if $j \notin \mathcal{C}'$, there is a client $j_1 \in \mathcal{C}'$ such that $d_{av}(j_1) \leq d_{av}(j), d(j, j_1) \leq 4d_{av}(j)$, by the second property of lemma 1. So $\mathbb{E}[C_j] \leq \mathbb{E}[C_{j_1}] + 4d_{av}(j)$. Thus, the ratio for j is bounded by the ratio for j_1 plus 4. So, it suffices to consider j_1.

W.L.O.G, assume $d_{av}(j_1) = 1$. Let j_2 be the client in $\mathcal{C}' \setminus \{j_1\}$ that is closest to j_1. Consider the case where j_1 is not matched with j_2 (this is worse than the case where they are matched). Then, j_2 must be matched with another client, say $j_3 \in \mathcal{C}'$, before j_1 is matched, and $d(j_2, j_3) \leq d(j_1, j_2)$. The sampling process guarantees that there must be a open facility in $\mathcal{U}_{j_2} \cup \mathcal{U}_{j_3}$. It is true that j_2 and j_3 may be far away from j_1. However, if $d(j_1, j_2) = 2R$ (thus, $d(j_1, j_3) \leq 4R, d_{av}(j_2), d_{av}(j_3) \leq R/2$), the volume of $\mathcal{U}_{j_1}$ is at least $1 - 1/R$. That means with probability at least $1 - 1/R$, j_1 will be connected to a facility that serves it in the fractional solution; only with probability $1/R$, j_1 will be connected to a facility that is $O(R)$ away. This finishes the proof.

3 Outline of the Proof of the 3.25-Approximation Ratio

If we analyze the algorithm as in the proof of lemma 4, an additive factor of 4 is lost at the first step. This additive factor can be avoided,[2] if we notice that there is a set $\mathcal{F}_j$ of facilities of volume 1 around j. Hopefully with some probability, some facility in $\mathcal{F}_j$ is open. It is not hard to show that this probability is at least $1 - 1/e$. So, only with probability $1/e$, we are going to pay the additive factor of 4. Even if there are no open facilities in $\mathcal{F}_j$, the facilities in $\mathcal{F}_{j_1}$ and $\mathcal{F}_{j_2}$ can help to reduce the constant.

A natural style of analysis is: focus on a set of "potential facilities", and consider the expected distance between j and the closest open facility in this set. An obvious candidate for the potential set is $\mathcal{F}_j \cup \mathcal{F}_{j_1} \cup \mathcal{F}_{j_2} \cup \mathcal{F}_{j_3}$. However, we are unable to analyze this complicated system.

Instead, we will consider a different potential set. Observing that $\mathcal{U}_{j_1}, \mathcal{U}_{j_2}, \mathcal{U}_{j_3}$ are disjoint, the potential set $\mathcal{F}_j \cup \mathcal{U}_{j_1} \cup \mathcal{U}_{j_2} \cup \mathcal{U}_{j_3}$ is much more tractable. Even with

[2] this is inspired by the analysis for the facility location problem in [8, 4, 14].

this simplified potential set, we still have to consider the intersection between $\mathcal{F}_j$ and each of $\mathcal{U}_{j_1}$, $\mathcal{U}_{j_2}$ and $\mathcal{U}_{j_3}$. Furthermore, we tried hard to reduce the approximation ratio at the cost of complicating the analysis(recall the argument about the choice of the scalar 1.5). With the potential set $\mathcal{F}_j \cup \mathcal{U}_{j_1} \cup \mathcal{U}_{j_2} \cup \mathcal{U}_{j_3}$, we can only prove a worse approximation ratio. To reduce it to 3.25, different potential sets are considered for different bottleneck cases.

W.L.O.G, we can assume $j \notin \mathcal{C}'$, since we can think of the case $j \in \mathcal{C}'$ as $j \notin \mathcal{C}'$ and there is another client $j_1 \in \mathcal{C}'$ with $d(j, j_1) = 0$. We also assume $d_{av}(j) = 1$. Let $j_1 \in \mathcal{C}'$ be the client such that $d_{av}(j_1) \leq d_{av}(j) = 1, d(j, j_1) \leq 4d_{av}(j) = 4$. Let j_2 be the closest client in $\mathcal{C}' \setminus \{j_1\}$ to j_1, thus $d(j_1, j_2) = 2R_{j_1}$. Then, either j_1 is matched with j_2, or j_2 is matched with a different client $j_3 \in \mathcal{C}'$, in which case we will have $d(j_2, j_3) \leq d(j_1, j_2) = 2R_{j_1}$. We only consider the second case. Readers can verify this is indeed the bottleneck case.

For the ease of notation, define $2R := d(j_1, j_2) = 2R_{j_1}, 2R' := d(j_2, j_3) \leq 2R, d_1 := d(j, j_1), d_2 := d(j, j_2)$ and $d_3 := d(j, j_3)$.

At the top level, we divide the analysis into two cases : the case $2 \leq d_1 \leq 4$ and the case $0 \leq d_1 \leq 2$. (Notice that we assumed $d_{av}(j) = 1$ and thus $0 \leq d_1 \leq 4$.) For some technical reason, we can not include the whole set $\mathcal{F}_j$ in the potential set for the former case. Instead we only include a subset $\mathcal{F}'_j$ (notice that $j \notin \mathcal{C}'$ and thus $\mathcal{F}'_j$ was not defined before). $\mathcal{F}'_j$ is defined as $\mathcal{F}_j \cap B(j, d_1)$.

The case $2 \leq d_1 \leq 4$ is further divided into 2 sub-cases : $\mathcal{F}'_j \cap \mathcal{F}'_{j_1} \subseteq \mathcal{U}_{j_1}$ and $\mathcal{F}'_j \cap \mathcal{F}'_{j_1} \not\subseteq \mathcal{U}_{j_1}$. Thus, we will have 3 cases, and the proof of the approximation ratios appear in the full paper.

1. $2 \leq d_1 \leq 4, \mathcal{F}'_j \cap \mathcal{F}'_{j_1} \subseteq \mathcal{U}_{j_1}$. In this case, we consider the potential set $\mathcal{F}'' = \mathcal{F}'_j \cup \mathcal{F}'_{j_1} \cup \mathcal{U}_{j_2} \cup \mathcal{U}_{j_3}$. Notice that $\mathcal{F}'_j = \mathcal{F}_j \cap B(j, d_1)$, $\mathcal{F}'_{j_1} = \mathcal{F}_{j_1} \cap B(j_1, 1.5R)$. In this case, $E[C_j] \leq 3.243$.
2. $2 \leq d_1 \leq 4, \mathcal{F}'_j \cap \mathcal{F}'_{j_1} \not\subseteq \mathcal{U}_{j_1}$. In this case, some facility i in $\mathcal{F}'_j \cap \mathcal{F}'_{j_1}$ must be claimed by some client $j' \neq j_1$. Since $d(j, i) \leq d_1, d(j_1, i) \leq 1.5R$, we have
$$d(j, j') \leq d(j, i) + d(j', i) \leq d(j, i) + d(j_1, i) \leq d_1 + 1.5R.$$
If $j' \notin \{j_2, j_3\}$, we can include $\mathcal{U}_{j'}$ in the potential set and thus the potential set is $\mathcal{F}'' = \mathcal{F}'_j \cup \mathcal{F}'_{j_1} \cup \mathcal{U}_{j_2} \cup \mathcal{U}_{j_3} \cup \mathcal{U}_{j'}$. If $j \in \{j_2, j_3\}$, then we know j and j_2, j_3 are close. So, we either have a "larger" potential set, or small distances between j and j_2, j_3. Intuitively, this case is unlikely to be the bottleneck case. In this case, we show $E[C_j] \leq 3.189$.
3. $0 \leq d_1 \leq 2$. In this case, we consider the potential set $\mathcal{F}'' = \mathcal{F}_j \cup \mathcal{U}_{j_1} \cup \mathcal{U}_{j_2} \cup \mathcal{U}_{j_3}$. In this case, $E[C_j] \leq 3.25$.

3.1 Running Time of the Algorithm

We now analyze the running time of our algorithm in terms of $n = |\mathcal{F} \cup \mathcal{C}|$. The bottleneck of the algorithm is solving the LP. Indeed, the total running time for rounding is $O(n^2)$.

To solve the LP, we use the $(1+\epsilon)$ approximation algorithm for the fractional k-median problem in [17]. The algorithm gives a fractional solution which opens

$(1+\epsilon)k$ facilities with connection cost at most $1+\epsilon$ times the fractional optimal in time $O(kn^2 \ln(n/\epsilon)/\epsilon^2)$. We set $\epsilon = \delta/k$ for some small constant δ. Then, our rounding procedure will open k facilities with probability $1-\delta$ and $k+1$ facilities with probability δ. The expected connection cost of the integral solution is at most $3.25(1+\delta/k)$ times the fractional optimal. Conditioned on the rounding procedure opening k facilities, the expected connection cost is at most $3.25(1+\delta/k)/(1-\delta) \leq 3.25(1+O(\delta))$ times the optimal fractional value.

Theorem 1. *For any $\delta > 0$, there is a $3.25(1+\delta)$-approximation algorithm for k-median problem that runs in $\tilde{O}\left((1/\delta^2)k^3n^2\right)$ time.*

3.2 Generalization of the Algorithm to Variants of *k*-Median

The distribution of k open facilities produced by our algorithm satisfies marginal conditions. That is, the probability that a facility i is open is exactly y_i. This allows our algorithm to be extended to some variants of the k-median problem.

The first variant is called k-facility location problem, which is a common generalization of k-median and UFL introduced in [10]. In the problem, we are given set $\mathcal{F}$ of facilities, set $\mathcal{C}$ of clients, metric $(d, \mathcal{F} \cup \mathcal{C})$, opening cost f_i for each facility $i \in \mathcal{F}$ and an integer k. The goal is to open at most k facilities and connect each client to its nearest open facility so as to minimize the sum of the opening cost and the connection cost. The best known approximation ratio for the k-facility location problem was $2+\sqrt{3}+\epsilon$, due to Zhang [18]. For this problem, the LP is the same as LP(1), except that we add a term $\sum_{i \in \mathcal{F}} f_i y_i$ to the objective function. After solving the LP, we use our rounding procedure to obtain an integer solution. The expected opening cost of the solution is exactly the fractional opening cost in the LP solution, while the expected connection cost is at most 3.25 times the fractional connection cost. This gives a 3.25-approximation for the problem, improving the $2+\sqrt{3}+\epsilon$-approximation.

Another generalization introduced in [10] is the k-median problem with t types of facilities. The goal of the problem is to open at most k facilities and connect each client to *one facility of each type* so as to minimize the total connection cost. Our techniques yield a 3.25 approximation for this problem as well. We first solve the natural LP for this problem. Then, we apply the rounding procedure to each type of facilities. The only issue is that the number of open facilities of some type in the LP solution might not be an integer. This can be handled using the techniques in the proof of Lemma 3.

4 Approximation Algorithms for Knapsack-Median and Matroid-Median

The LP for knapsack-median is the same as LP (1), except that we change the constraint $\sum_{i \in \mathcal{F}} y_i \leq k$ to the knapsack constraint $\sum_{i \in \mathcal{F}} f_i y_i \leq M$.

As shown in [13], the LP has unbounded integrality gap. To amend this, we do the same trick as in [13]. Suppose we know the optimal cost OPT for the instance. For a client j, let L_j be its connection cost. Then, for some other client j', its connection cost is at least $\max\{0, L_j - d(j,j')\}$. This suggests

$$\sum_{j'\in\mathcal{C}} \max\{0, L_j - d(j,j')\} \leq \mathsf{OPT}. \tag{1}$$

Thus, knowing OPT, we can get an upper bound L_j on the connection cost of j: L_j is the largest number such that the above inequality is true. We solve the LP with the additional constraint that $x_{i,j} = 0$ if $d(i,j) > L_j$. Then, the LP solution, denoted by LP, must be at most OPT. By binary searching, we find the minimum OPT so that $\mathsf{LP} \leq \mathsf{OPT}$. Let $\left(x^{(1)}, y^{(1)}\right)$ be the fractional solution given by the LP. We use $\mathsf{LP}_j = d_{av}(j) = \sum_{i\in\mathcal{F}} d(i,j)x^{(1)}_{i,j}$ to denote the contribution of the client j to LP.

Then we select a set of filtered clients $\mathcal{C}'$ as we did in the algorithm for the k-median problem. For a client $j \in \mathcal{C}$, let $\pi(j)$ be a client $j' \in \mathcal{C}'$ such that $d_{av}(j') \leq d_{av}(j), d(j,j') \leq 4d_{av}(j)$. Notice that for a client $j \in \mathcal{C}'$, we have $\pi(j) = j$. This time, we can not save the additive factor of 4; instead, we move the connection demand on each client $j \notin \mathcal{C}'$ to $\pi(j)$. For a client $j' \in \mathcal{C}'$, let $w_{j'} = \left|\pi^{-1}(j')\right|$ be its connection demand. Let $\mathsf{LP}^{(1)} = \sum_{j'\in\mathcal{C}', i\in\mathcal{F}} w_{j'} x_{i,j'} d(i,j') = \sum_{j'\in\mathcal{C}'} w_{j'} d_{av}(j')$ be the cost of the solution $\left(x^{(1)}, y^{(1)}\right)$ to the new instance. For a client $j \in \mathcal{C}$, let $\mathsf{LP}^{(1)}_j = d_{av}(\pi(j))$ be the contribution of j to $\mathsf{LP}^{(1)}$. (The amount $w_{j'} d_{av}(j')$ is evenly spread among the $w_{j'}$ clients in $\pi^{-1}(j')$.) Since $\mathsf{LP}_j = d_{av}(j) \leq d_{av}(\pi(j)) \leq \mathsf{LP}^{(1)}_j$, we have $\mathsf{LP}^{(1)} \leq \mathsf{LP}$.

For any client $j \in \mathcal{C}'$, let $2R_j = \min_{j'\in\mathcal{C}', j'\neq j} d(j,j')$, if $\mathsf{vol}(B(j,R_j)) \leq 1$; otherwise let R_j be the smallest number such that $\mathsf{vol}(B(j,R_j)) = 1$. ($\mathsf{vol}(S)$ is defined as $\sum_{i\in S} y^{(1)}_i$.) Let $B_j = B(j,R_j)$ for the ease of notation. If $\mathsf{vol}(B_j) = 1$, we call B_j a full ball; otherwise, we call B_j a partial ball. Notice that we always have $\mathsf{vol}(B_j) \geq 1/2$. Notice that $R_j \leq L_j$ since $x^{(1)}_{i,j} = 0$ for all facilities i with $d_{i,j} > L_j$.

We find a matching $\mathcal{M}$ over the partial balls as in Section 2: while there are at least 2 unmatched partial balls, match the two balls B_j and $B_{j'}$ with the smallest $d(j,j')$. Consider the following LP.

$$\text{LP(2)} \qquad \min \quad \textstyle\sum_{j'\in\mathcal{C}'} w_{j'} \left(\sum_{i\in B_{j'}} d(i,j')y_i + \left(1 - \sum_{i\in B_j} y_i\right) R_{j'}\right)$$

$$\sum_{i\in B_{j'}} y_i = 1, \quad \forall j' \in \mathcal{C}', B_{j'} \text{ full}; \qquad \sum_{i\in B_{j'}} y_i \leq 1, \qquad \forall j' \in \mathcal{C}', B_{j'} \text{ partial};$$

$$\sum_{i\in B_j} y_i + \sum_{i\in B_{j'}} y_i \geq 1, \quad \forall (B_j, B_{j'}) \in \mathcal{M}; \qquad \sum_{i\in\mathcal{F}} f_i y_i \leq M;$$

$$y_i \geq 0, \quad \forall i \in \mathcal{F}$$

Let $y^{(2)}$ be an optimal *basic solution* of LP (2) and let $\mathsf{LP}^{(2)}$ be the value of LP(2). For a client $j \in \mathcal{C}$ with $\pi(j) = j'$, let $\mathsf{LP}^{(2)}_j = \sum_{i\in B_{j'}} d(i,j')y_i + \left(1 - \sum_{i\in B_{j'}} y_i\right) R_{j'}$ be the contribution of j to $\mathsf{LP}^{(2)}$. Then we prove

Lemma 5. $\mathsf{LP}^{(2)} \leq \mathsf{LP}^{(1)}$.

Proof. It is easy to see that $y^{(1)}$ is a valid solution for LP(2). By slightly abusing the notations, we can think of $\mathsf{LP}^{(2)}$ is the cost of $y^{(1)}$ to LP(2). We compare the contribution of each client $j \in \mathcal{C}$ with $\pi(j) = j'$ to $\mathsf{LP}^{(2)}$ and to $\mathsf{LP}^{(1)}$. If $B_{j'}$ is a full ball, j' contributes the same to $\mathsf{LP}^{(2)}$ and as to $\mathsf{LP}^{(1)}$. If $B_{j'}$ is a partial ball, j' contributes $\sum_{i \in \mathcal{F}_{j'}} d(i,j')y_i^{(1)}$ to $\mathsf{LP}^{(1)}$ and $\sum_{i \in B_{j'}} d(i,j')y_i^{(1)} + (1 - \sum_{i \in B_{j'}} y_i^{(1)})R_{j'}$ to $\mathsf{LP}^{(2)}$. Since $B_{j'} = B(j', R_{j'}) \subseteq \mathcal{F}_{j'}$ and $\mathsf{vol}(\mathcal{F}_{j'}) = 1$, the contribution of j' to $\mathsf{LP}^{(2)}$ is at most that to $\mathsf{LP}^{(1)}$. So, $\mathsf{LP}^{(2)} \leq \mathsf{LP}^{(1)}$.

Notice that LP(2) only contains y-variables. We show that any basic solution y^* of LP(2) is almost integral. In particular, we prove the following lemma in the full version of the paper:

Lemma 6. *Any basic solution y^* of $LP(2)$ contains at most 2 fractional values. Moreover, if it contains 2 fractional values $y^*_i, y^*_{i'}$, then $y^*_i + y^*_{i'} = 1$ and either there exists some $j \in \mathcal{C}'$ such that $i, i' \in B_j$ or there exists a pair $(B_j, B_{j'}) \in \mathcal{M}$ such that $i \in B_j, i' \in B_{j'}$.*

Let $y^{(3)}$ be the integral solutin obtained from $y^{(2)}$ as follows. If $y^{(2)}$ contains at most 1 fractional value, we zero-out the fractional value. If $y^{(2)}$ contains 2 fractional values $y_i^{(2)}, y_{i'}^{(2)}$, let $y_i^{(3)} = 1, y_{i'}^{(3)} = 0$ if $f_i \leq f_{i'}$ and let $y_i^{(3)} = 0, y_{i'}^{(3)} = 1$ otherwise. Notice that since $y_i^{(2)} + y_{i'}^{(2)} = 1$, this modification does not increase the budget. Let SOL be the cost of the solution $y^{(3)}$ to the original instance.

We leave the proof of the following lemma to the full version of the paper.

Lemma 7. $\sum_{i \in B(j', 5R_{j'})} y_i^{(2)} \geq 1$ *and* $\sum_{i \in B(j', 5R_{j'})} y_i^{(3)} \geq 1$. *i.e, there is an open facility (possibly two facilities whose opening fractions sum up to 1) inside $B(j', 5R_{j'})$ in both the solution $y^{(2)}$ and the solution $y^{(3)}$.*

Lemma 8. $\mathsf{SOL} \leq 34\mathsf{OPT}$.

Proof. Let $\tilde{i}$ be the facility that $y_{\tilde{i}}^{(2)} > 0, y_{\tilde{i}}^{(3)} = 0$, if it exists; let $\tilde{j}$ be the client that $\tilde{i} \in B_{\tilde{j}}$.

Now, we focus on a client $j \in \mathcal{C}$ with $\pi(j) = j'$. Then, $d(j, j') \leq 4d_{av}(j) = 4\mathsf{LP}_j$. Assume that $j' \neq \tilde{j}$. Then, to obtain $y^{(3)}$, we did not move or remove an open facility from $B_{j'}$. In other words, for every $i \in B_{j'}$, $y_i^{(3)} \geq y_i^{(2)}$. In this case, we show

$$\mathsf{SOL}_{j'} \leq \sum_{i \in B_{j'}} d(i, j')y_i^{(2)} + (1 - \sum_{i \in B_{j'}} y_i^{(2)}) \times 5R_{j'}.$$

If there is no open facility in $B_{j'}$ in $y^{(3)}$, then there is also no open facility in $B_{j'}$ in $y^{(2)}$. Then, by Lemma 7, $\mathsf{SOL}_{j'} = 5R_{j'} =$ right-side. Otherwise, there is exactly one open facility in $B_{j'}$ in $y^{(3)}$. In this case, $\mathsf{SOL}_{j'} = \sum_{i \in B_{j'}} d(j', i)y_i^{(3)} \leq$ right-side since $y_i^{(3)} \geq y_i^{(2)}$ and $d(i, j') \leq 5R_{j'}$ for every $i \in B_{j'}$.

Observing that the right side of the inequality is at most $5\mathsf{LP}^{(2)}_j$, we have $\mathsf{SOL}_j \le 4\mathsf{LP}_j + \mathsf{SOL}_{j'} \le 4\mathsf{LP}_j + 5\mathsf{LP}^{(2)}_j$.

Now assume that $j' = \tilde{j}$. Since there is an open facility in $B(j', 5R_{j'})$ by Lemma 7, we have $\mathsf{SOL}_j \le 4\mathsf{LP}_j + 5R_{j'}$. Consider the set $\pi^{-1}(j')$ of clients. Notice that we have $R_{j'} \le L_{j'}$ since $x^{(1)}_{i,j'} = 0$ for facilities i such that $d(i, j') > L_{j'}$. Also by Inequality (1), we have $\sum_{j\in\pi^{-1}(j')}(R_{j'} - d(j,j')) \le \sum_{j\in\pi^{-1}(j')}(L_{j'} - d(j,j')) \le \mathsf{OPT}$. Then, since $d(j,j') \le 4\mathsf{LP}_j$ for every $j \in \pi^{-1}(j')$, we have

$$\begin{aligned}\sum_{j\in\pi^{-1}(j')}\mathsf{SOL}_j &\le \sum_j (4\mathsf{LP}_j + 5R_{j'}) \le 4\sum_j \mathsf{LP}_j + 5\sum_j R_{j'}\\ &\le 4\sum_j \mathsf{LP}_j + 5\Big(\mathsf{OPT} + \sum_j d(j,j')\Big) \le 24\sum_j \mathsf{LP}_j + 5\mathsf{OPT},\end{aligned}$$

where the sums are all over clients $j \in \pi^{-1}(j')$. Summing up all clients $j \in \mathcal{C}$, we have

$$\begin{aligned}\mathsf{SOL} &= \sum_{j\in\mathcal{C}}\mathsf{SOL}_j = \sum_{j\notin\pi^{-1}(\tilde{j})}\mathsf{SOL}_j + \sum_{j\in\pi^{-1}(\tilde{j})}\mathsf{SOL}_j\\ &\le \sum_{j\notin\pi^{-1}(\tilde{j})}(4\mathsf{LP}_j + 5\mathsf{LP}^{(2)}_j) + 24\sum_{j\in\pi^{-1}(\tilde{j})}\mathsf{LP}_j + 5\mathsf{OPT}\\ &\le 24\sum_{j\in\mathcal{C}}\mathsf{LP}_j + 5\sum_{j\in\mathcal{C}}\mathsf{LP}^{(2)}_j + 5\mathsf{OPT} \le 24\mathsf{LP} + 5\mathsf{LP}^{(2)} + 5\mathsf{OPT} \le 34\mathsf{OPT},\end{aligned}$$

where the last inequality follows from the fact that $\mathsf{LP}^{(2)} \le \mathsf{LP}^{(1)} \le \mathsf{LP} \le \mathsf{SOL}$. Thus, we proved

Theorem 2. *There is an efficient 34-approximation algorithm for the knapsack-median problem.*

It is not hard to change our algorithm so that it works for the matroid median problem. The analysis for the matroid median problem is simpler, since $y^{(2)}$ will already be an integral solution. We leave the proof of the following theorem to the full version of the paper.

Theorem 3. *There is an efficient 9-approximation algorithm for the matroid median problem, assuming there is an efficient oracle for the input matroid.*

References

1. Archer, A., Rajagopalan, R., Shmoys, D.B.: Lagrangian relaxation for the k-median problem: new insights and continuity properties. In: In Proceedings of the 11th Annual European Symposium on Algorithms. pp. 31–42 (2003)
2. Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., Pandit, V.: Local search heuristic for k-median and facility location problems. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing. pp. 21–29. STOC '01, ACM, New York, NY, USA (2001), `http://doi.acm.org/10.1145/380752.380755`

3. Bradley, P.S., Fayyad, U.M., Mangasarian, O.L.: Mathematical programming for data mining: Formulations and challenges. INFORMS Journal on Computing 11, 217–238 (1998)
4. Byrka, J.: An optimal bifactor approximation algorithm for the metric uncapacitated facility location problem. In: APPROX '07/RANDOM '07: Proceedings of the 10th International Workshop on Approximation and the 11th International Workshop on Randomization, and Combinatorial Optimization. Algorithms and Techniques. pp. 29–43. Springer-Verlag, Berlin, Heidelberg (2007)
5. Byrka, J., Srinivasan, A., Swamy, C.: Fault-tolerant facility location: A randomized dependent lp-rounding algorithm. In: IPCO. pp. 244–257 (2010)
6. Charikar, M., Guha, S.: Improved combinatorial algorithms for the facility location and k-median problems. In: In Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science. pp. 378–388 (1999)
7. Charikar, M., Guha, S., Tardos, É., Shmoys, D.B.: A constant-factor approximation algorithm for the k-median problem (extended abstract). In: Proceedings of the thirty-first annual ACM symposium on Theory of computing. pp. 1–10. STOC '99, ACM, New York, NY, USA (1999), http://doi.acm.org/10.1145/301250.301257
8. Chudak, F.A., Shmoys, D.B.: Improved approximation algorithms for the uncapacitated facility location problem. SIAM J. Comput. 33(1), 1–25 (2004)
9. Jain, K., Mahdian, M., Saberi, A.: A new greedy approach for facility location problems. In: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing. pp. 731–740. STOC '02, ACM, New York, NY, USA (2002), http://doi.acm.org/10.1145/509907.510012
10. Jain, K., Vazirani, V.V.: Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. J. ACM 48(2), 274–296 (2001)
11. Krishnaswamy, R., Kumar, A., Nagarajan, V., Sabharwal, Y., Saha, B.: The matroid median problem. In: In Proceedings of ACM-SIAM Symposium on Discrete Algorithms. pp. 1117–1130 (2011)
12. Kuehn, A.A., Hamburger, M.J.: A heuristic program for locating warehouses 9(9), 643–666 (Jul 1963)
13. Kumar, A.: Constant factor approximation algorithm for the knapsack median problem. In: Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 824–832. SODA '12, SIAM (2012), http://dl.acm.org/citation.cfm?id=2095116.2095182
14. Li, S.: A 1.488-approximation algorithm for the uncapacitated facility location problem. In: In Proceeding of the 38th International Colloquium on Automata, Languages and Programming (2011)
15. Lin, J.H., Vitter, J.S.: Approximation algorithms for geometric median problems. Inf. Process. Lett. 44, 245–249 (December 1992), http://portal.acm.org/citation.cfm?id=152566.152569
16. Manne, A.: Plant location under economies-of-scale-decentralization and computation. In: Managment Science (1964)
17. Young, N.E.: K-medians, facility location, and the chernoff-wald bound. In: Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms. pp. 86–95. SODA '00, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2000), http://portal.acm.org/citation.cfm?id=338219.338239
18. Zhang, P.: A new approximation algorithm for the k-facility location problem. In: Proceedings of the Third international conference on Theory and Applications of Models of Computation. pp. 217–230. TAMC'06, Springer-Verlag, Berlin, Heidelberg (2006), http://dx.doi.org/10.1007/11750321_21