Approximation Algorithms for Geometric Median Problems

Jyh-Han Lin[1]
Jeffrey Scott Vitter[2]

Technical Report No. CS-92-37
August 1992

[1]Motorola, Inc., Paging and Telepoint Systems Lab, 1500 N.W. 22nd Avenue, Boynton Beach, FL 33426-8292
[2]Department of Computer Science, Box 1910, Brown University, Providence, RI 02912

Approximation Algorithms for Geometric Median Problems *

Jyh-Han Lin †
Motorola Inc.
Paging and Telepoint Systems Lab
1500 N.W. 22nd Avenue
Boynton Beach, FL 33426–8292
lin@pts.mot.com

Jeffrey Scott Vitter
Department of Computer Science
Brown University
Providence, R. I. 02912–1910
jsv@cs.brown.edu

Abstract

In this paper we present approximation algorithms for median problems in metric spaces and fixed-dimensional Euclidean space. Our algorithms use a new method for transforming an optimal solution of the linear program relaxation of the s-median problem into a provably good integral solution. This transformation technique is fundamentally different from the methods of randomized and deterministic rounding [7,8] and the methods proposed in [6] in the following way: Previous techniques never set variables with zero values in the fractional solution to 1. This departure from previous methods is crucial for the success of our algorithms.

Keywords: median, clustering, approximation, linear programming, relaxation, rounding, metric space, Euclidean space.

*Support was provided in part by a National Science Foundation Presidential Young Investigator Award CCR–9047466 with matching funds from IBM, by NSF research grant CCR–9007851, and by Army Research Office grant DAAL03–91–G–0035

†The research was conducted while the author was at the Department of Computer Science, Brown University.

1 Introduction

Let us consider a complete (directed or undirected) graph $G = (V, E)$ on n vertices, with vertex set $V = \{1, \ldots, n\}$, edge set $E \subseteq V \times V$, and nonnegative distance c_{ij} associated with edges. We refer to (c_{ij}) as the *distance matrix*. Given a bound $D > 0$, the goal of the (*discrete*) *median problem* is to choose vertices as medians so that the sum of distances from each vertex to its nearest median is no more than D and the number of medians chosen is minimized.[1]

In this paper, we present approximation algorithms for the median problem when the vertices are embedded in metric spaces. That is, we have $c_{ii} = 0$, $c_{ij} = c_{ji}$, and the triangle inequality $c_{ij} \leq c_{ij'} + c_{j'j}$. The main results are stated in the following two theorems, which will be proven in Sections 2 and 3, respectively:

Theorem 1 *There exists a deterministic approximation algorithm for the median problem in metric spaces that, given any $\epsilon > 0$, outputs a median set U satisfying*

$$\sum_{j \in V} \min_{i \in U} c_{ij} \quad \leq \quad 2(1 + 1/\epsilon)D \tag{1}$$

and

$$|U| \quad < \quad (1 + \epsilon)s, \tag{2}$$

where s is the optimal size of median sets with a total distance at most D.

The greedy approximation algorithm in [6] gives a better cost bound using more medians; the right-hand sides of (1) and (2) are replaced by $(1 + 1/\epsilon)D$ and $(1 + \epsilon)s(\ln n + 1)$, respectively. Theorem 1 shows that we can trade off a factor of 2 in the bound on cumulative distance in order to reduce the bound on the number of medians by a logarithmic factor. Furthermore, tradeoffs are available for fixed-dimensional Euclidean spaces:

Theorem 2 *Let $d \geq 2$ be fixed positive integer. There exists a deterministic approximation algorithm for the median problem in d-dimensional Euclidean space that, given any $\epsilon > 0$ and any integer $\rho \geq 1$, outputs a median set U satisfying*

$$\sum_{j \in V} \min_{i \in U} c_{ij} \quad \leq \quad (1 + 1/\rho)(1 + 1/\epsilon)D$$

and

$$|U| < (2\rho - 1)^d (1 + \epsilon)s,$$

where s is the optimal size of median sets with a total distance at most D.

[1] Note that the median problem is distinctly different from the s-center problem of choosing s centers that minimize the worst-case distance from each vertex to its nearest center. For approximation algorithms for the s-center problem, we refer the readers to [1,2,3]

The transformation technique used for the proofs of Theorems 1 and 2 is fundamentally different from those of randomized and deterministic rounding [7,8] and the methods proposed in [6]. The previous techniques never set variables with zero values in the fractional solution to 1. Our algorithms, on the other hand, may set 0-valued variables to 1. This departure from previous methods is crucial for the success of our algorithms.

We show in [6] that the number of medians cannot be approximated within better than logarithmic factors without violating the bound on total distance, unless the dominating set and set cover problems can be approximated within better than logarithmic factors.

2 Metric Spaces

In this section we give the proof of Theorem 1. The median problem can be formulated as a 0-1 integer linear program of minimizing

$$\sum_{j=1}^{n} y_j \tag{3}$$

subject to

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} \leq D \tag{4}$$

$$\sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1,\ldots,n, \tag{5}$$

$$x_{ij} \leq y_j, \qquad i,j = 1,\ldots,n, \tag{6}$$

$$x_{ij}, y_j \in \{0,1\}, \qquad i,j = 1,\ldots,n, \tag{7}$$

where $y_j = 1$ if and only if j is chosen as a median, $x_{ij} = 1$ if and only if $y_j = 1$ and i is assigned to j, and $D > 0$ is a given bound on the total distance.

Our merging algorithm for the median problem works as follows:

Algorithm M

1. Solve the linear program relaxation by linear programming techniques [4,5]; denote the fractional solution by $\widehat{y}, \widehat{x}$.

2. For each i, compute $\widehat{C}_i = \sum_{j \in V} c_{ij}\widehat{x}_{ij}$.

3. Given $\epsilon > 0$, for each vertex i, the *neighborhood* V_i of vertex i consists of all vertices j such that $c_{ij} \leq (1+1/\epsilon)\widehat{C}_i$. A vertex j is in the *extended neighborhood* $\overline{V_i}$ of vertex i if and only if one of the following rules holds:

   (R1) $c_{ij} \leq (1+1/\epsilon)\widehat{C}_i$, that is, vertex j is in the neighborhood V_i of vertex i.

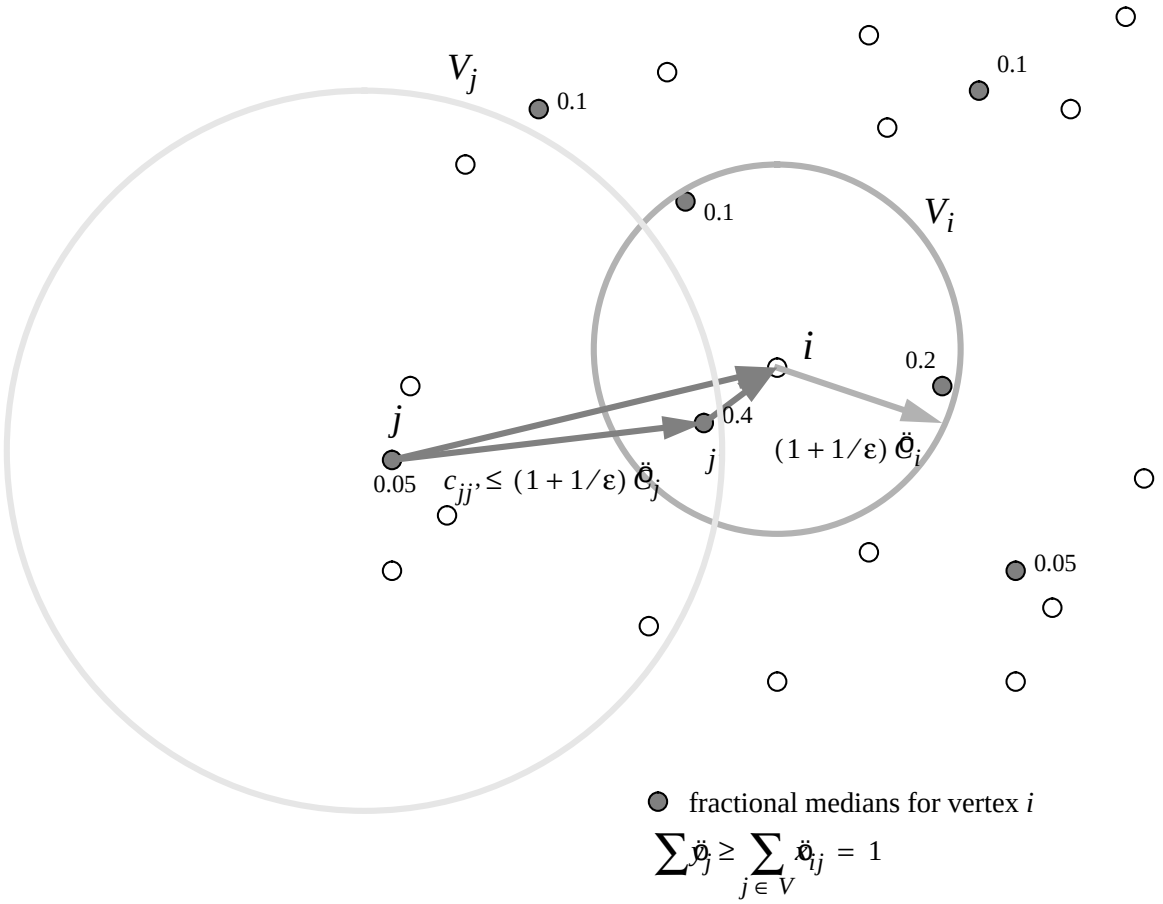

Figure 1: A vertex j is in the extended neighborhood $\overline{V_i}$ of vertex i if and only if $j \in V_i$ or $V_i \cap V_j \neq \emptyset$.

(R2) There exists a vertex j' such that $c_{ij'} \leq (1+1/\epsilon)\widehat{C}_i$ and $c_{jj'} \leq (1+1/\epsilon)\widehat{C}_j$, that is, $V_i \cap V_j \neq \emptyset$.

This construction is illustrated in Figure 1.

4. Sort the collection of sets $\{\overline{V_i}\}_{i \in V}$ by $\widehat{C}_i$ in nondecreasing order.

5. Choose a set $\overline{V_i}$ with smallest $\widehat{C}_i$ and delete any set $\overline{V_j}$ such that $j \in \overline{V_i}$. Repeat this process until no sets remain.

6. Let the median set U consist of all vertices whose extended neighborhood $\overline{V_j}$ are chosen in Step 5.

The proof of Theorem 1 follows from the following lemmas:

Lemma 1 (Symmetry) *For each $\overline{V_i}$, if $j \in \overline{V_i}$ then $i \in \overline{V_j}$.*

Proof: By the definition of $\overline{V_i}$, there are two cases.

Case 1: $c_{ij} \leq (1+1/\epsilon)\widehat{C}_i$. Since $c_{jj} = 0$, by (R2) we have $i \in \overline{V_j}$.

Case 2: Otherwise, there exists a vertex j' such that $c_{ij'} \leq (1+1/\epsilon)\widehat{C}_i$ and $c_{jj'} \leq (1+1/\epsilon)\widehat{C}_j$. We have $i \in \overline{V_j}$ immediately by symmetry and (R2).

□

Lemma 2 *Let $\overline{V_i}$ and $\overline{V_j}$ be two distinct sets selected by Algorithm M in Step 5. Then we have $V_i \cap V_j = \emptyset$.*

Proof: (By contradiction.) Suppose that there exists j' such that $j' \in V_i \cap V_j$. Without loss of generality, we assume $\overline{V_i}$ is selected before $\overline{V_j}$. Since $c_{ij'} \leq (1+1/\epsilon)\widehat{C}_i$ and $c_{jj'} \leq (1+1/\epsilon)\widehat{C}_j$, by (R2) we must have $j \in \overline{V_i}$. Hence, Algorithm M will delete $\overline{V_j}$ after the selection of $\overline{V_i}$, which is a contradiction. □

Lemma 3 *Let U' be a set of medians. If for all j' such that $\widehat{y}_{j'} > 0$, we have $j' \in V_i$ for some $i \in U'$, then $\bigcup_{i \in U'} \overline{V_i} = V$.*

Proof: For each j, there exists at least one $\widehat{y}_{j'} > 0$ such that $c_{jj'} \leq \widehat{C}_j$. Since $c_{ij'} \leq (1+1/\epsilon)\widehat{C}_i$ for some $i \in U'$, by the definition of $\overline{V_i}$ we have $j \in \overline{V_i}$. □

Lemma 4 *For each j, let $\overline{V}_{i(j)}$ be the first set selected by Algorithm M such that $j \in \overline{V}_{i(j)}$, then we have $\widehat{C}_{i(j)} \leq \widehat{C}_j$.*

Proof: (By contradiction.) Suppose that $\widehat{C}_j < \widehat{C}_{i(j)}$. There are two cases, both of which lead to contradiction.

Case 1: $\overline{V_j}$ is selected by Algorithm M before $\overline{V}_{i(j)}$. By Lemma 1, we have $i(j) \in \overline{V_j}$. Hence, we conclude that $\overline{V}_{i(j)}$ must have been deleted already, which is a contradiction.

Case 2: Otherwise, since $\overline{V}_{i(j)}$ is the first set containing j, then immediately before the selection of $\overline{V}_{i(j)}$, $\overline{V_j}$ has not yet been deleted by Algorithm M. This implies $\overline{V}_{i(j)}$ cannot be the next set selected by Algorithm M, which is again a contradiction.

□

Lemma 5 *For each j, let $\overline{V}_{i(j)}$ be the first set selected by Algorithm M such that $j \in \overline{V}_{i(j)}$. Then we have $c_{ji(j)} \leq 2(1+1/\epsilon)\widehat{C}_j$.*

Proof: There are three cases.

Case 1: If $j = i(j)$ then $c_{ji(j)} = 0$.

Case 2: If $c_{i(j)j} \leq (1+1/\epsilon)\widehat{C}_i$ then $c_{ji(j)} = c_{i(j)j} \leq (1+1/\epsilon)\widehat{C}_j$ by symmetry and Lemma 4.

Case 3: Otherwise, we must have $c_{i(j)j'} \leq (1+1/\epsilon)\widehat{C}_{i(j)}$ and $c_{jj'} \leq (1+1/\epsilon)\widehat{C}_j$ for some $j' \in V_{i(j)}$. By symmetry and the triangle inequality, we have

$$c_{ji(j)} \leq c_{jj'} + c_{j'i(j)} \leq (1+1/\epsilon)\widehat{C}_j + (1+1/\epsilon)\widehat{C}_{i(j)} \leq 2(1+1/\epsilon)\widehat{C}_j.$$

The last inequality follows from Lemma 4.

□

Lemma 6 *For each $i \in V$ and $\epsilon > 0$, we have*

$$\sum_{j \in V_i} \widehat{y}_j \geq \sum_{j \in V_i} \widehat{x}_{ij} > \frac{\epsilon}{1+\epsilon},$$

where V_i is the neighborhood of vertex i.

Proof: Suppose $\sum_{j \in V_i} \widehat{x}_{ij} \leq \epsilon/(1+\epsilon)$. Then

$$\begin{aligned}
\widehat{C}_i &= \sum_{j \in V} c_{ij}\widehat{x}_{ij} \\
&\geq \sum_{j \notin V_i} c_{ij}\widehat{x}_{ij} \\
&> (1+\epsilon)\widehat{C}_i \sum_{j \notin V_i} \widehat{x}_{ij} \\
&\geq (1+\epsilon)\widehat{C}_i \left(1 - \frac{\epsilon}{1+\epsilon}\right) \\
&= \widehat{C}_i,
\end{aligned}$$

which is a contradiction. □

We now prove Theorem 1. By Lemma 6, for each set $\overline{V_i}$ selected, the sum of the fractional medians in V_i is greater than $1/(1+\epsilon)$. Lemma 2 implies that each fractional median is covered at most once by some V_i throughout the execution of Algorithm M. Therefore, by Lemma 3 the number of sets (medians) selected is less than

$$\frac{s}{1/(1+\epsilon)} = (1+\epsilon)s.$$

By Lemma 5 we have

$$\sum_{j \in V} \min_{i \in U} c_{ij} \leq 2(1+1/\epsilon) \sum_{j \in V} \widehat{C}_j \leq 2(1+1/\epsilon)D.$$

3 Fixed-Dimensional Euclidean Spaces

In this section we give the proof of Theorem 2. Let vertex i be a median selected by Algorithm M and let $\overline{V_{i,\rho}} \subseteq \overline{V_i}$ be a subset of vertices such that a vertex $j \in \overline{V_i}$ is in $\overline{V_{i,\rho}}$ if and only if $\widehat{C}_j \leq \rho\widehat{C}_i$.

Since $\overline{V_{i,\rho}}$ is bounded by a ball of diameter $2\rho(1+1/\epsilon)\widehat{C}_i$ in d-dimensional Euclidean space, there exists a median set U_i of size at most $(2\rho-1)^d$ such that for all $j \in \overline{V_{i,\rho}}$ we have

$$\min_{\ell \in U_i} c_{j\ell} \leq (1+1/\epsilon)\widehat{C}_i.$$

For each $j \in \overline{V_i} - \overline{V_{i,\rho}}$, there exists $j' \in \overline{V_{i,\rho}}$ such that $c_{jj'} \leq (1+1/\epsilon)\widehat{C}_j$. Therefore, we have

$$\min_{\ell \in U_i} c_{j\ell} \leq c_{jj'} + \min_{\ell \in U_i} c_{j'\ell} \leq (1+1/\epsilon)\widehat{C}_j + (1+1/\epsilon)\widehat{C}_i.$$

Since $\widehat{C}_j > \rho\widehat{C}_i$, we have

$$\min_{\ell \in U_i} c_{j\ell} < (1+1/\epsilon)\widehat{C}_j + \frac{1}{\rho}(1+1/\epsilon)\widehat{C}_j = (1+1/\rho)(1+1/\epsilon)\widehat{C}_j.$$

The rest of the proof follows by replacing each median i selected by Algorithm M by the median set U_i (packing).

4 Conclusions

One interesting long-standing open problem about the Euclidean median problem is the worst-case ratio between the total distance of an optimal integral solution and the total distance of an optimal fractional solution. The results of this paper is a step toward that goal.

References

[1] T. Feder and D. Greene, "Optimal Algorithms for Approximate Clustering," in *Proceedings of the 20th Annual Symposium on the Theory of Computing*, 1988, 434–444.

[2] T. F. Gonzalez, "Clustering to Minimize the Maximum Intercluster Distance," *Theoretical Computer Science* 38 (1985), 293–306.

[3] D. S. Hochbaum and D. B. Shmoys, "A Unified Approach to Approximation Algorithms for Bottleneck Problems," *Journal of the Association for Computing Machinery* 33 (July 1986), 533–550.

[4] N. Karmarkar, "A New Polynomial-Time Algorithm for Linear Programming," *Combinatorica* 4 (1984), 373–395.

[5] L. G. Khachiyan, "A Polynomial Algorithm in Linear Programming," *Soviet Math. Doklady* 20 (1979), 191–194.

[6] J.-H. Lin and J. S. Vitter, "ϵ-Approximations with Minimum Packing Constraint Violation," in *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, Victoria, BC, Canada , May 1992, 771–782.

[7] P. Raghavan, "Probabilistic Construction of Deterministic Algorithms: Approximating Packing Integer Programs," *Journal of Computer and System Science* 37 (1988), 130–143.

[8] P. Raghavan and C. D. Thompson, "Randomized Rounding: A Technique for Provably Good Algorithms and Algorithmic Proofs," *Combinatorics* 7 (1987), 365–374.