Probabilistic Analysis of a Relaxation for the k-Median Problem

—The Euclidean Model in the Plane—

Sang-Hyung Ahn

《Contents》

1. Introduction

The k-median problem has been widely studied both from the theoretical point of view and for its applications. An interesting theoretical development was the successful probabilistic analysis of several heuristics for this problem (e.g. Fisher and Hochbaum [8] and Papadimitriou[22]). On the other hand, the literature on the k-median problem abounds in exact algorithms. Most are based on the solution of a certain relaxation to be defined later. The computational experience reported in the literature seems to indicate that this particular relaxation yields impressively tight bounds compared to what can usually be expected in integer programming. In this paper we analyze to what extent this relaxation is tight. We perform our analysis for a classical Euclidean model in the plane and show that the relaxation can be expected to provide a bound within one third of one percent of the optimum value of the k-median problem. In addition to the probabilistic analysis, we also report extensive computational

Author: Assistant Professor School of Management, Seoul National University

experiments, based on the solution of thousands of medium-size problems. Some of the results predicted for very large problems by our probabilistic analysis can already be observed on these test problems.

2. Problem Formulation

Consider a set $X=\{X_1, \ldots, X_n\}$ of n points, a positive integer $k \leq n$ and let $d_{ij} \geq 0$ be the distance between X_i and X_j for each $1 \leq i \leq n$ and $1 \leq j \leq n$. (Unless otherwise specified, it is assumed that $d_{ii}=0$, $d_{ij}=d_{ji}$ and $d_{ij} \leq d_{ik}+d_{kj}$, for all i, j, k). The *k-median problem* consists of finding a set $S \subseteq X$, $|S|=k$, that minimizes $\sum_{i=1}^{n} \min_{j \in S} d_{ij}$. (Here $|S|$ denotes the cardinality of the set S). The k-median problem has the following integer programming formulation.

$$Z_{IP}=\min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij} \tag{1}$$

$$\sum_{j=1}^{n} y_{ij}=1 \text{ for } i=1, \ldots, n \tag{2}$$

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

$$0 \leq y_{ij} \leq x_j \leq 1 \text{ for } i, j=1, \ldots, n \tag{4}$$

$$x_j \in \{0, 1\} \text{ for } j=1, \ldots, n. \tag{5}$$

In this formulation $x_j=1$ if $X_j \in S$, 0 otherwise and, for $1 \leq i \leq n$, we can set $y_{ij}=1$ for an index j that achieves $\min_{j \in S} d_{ij}$.

The formulation (1)～(4) is called the linear programming (LP) relaxation of the k-median problem. In other words, the LP relaxation is obtained by ignoring the integrality conditions on x_j, $1 \leq j \leq n$. The optimum value Z_{LP} of this relaxation clearly stisfies $Z_{LP} \leq Z_{IP}$. The bound Z_{LP} has been used extensively in exact algorithms for the k-median problem. (E.g. Marsten[15], Garfinkel Neebe and Rao[10], ReVelle and Swain[23], Diehr[5], Schrage[24], Guignard and Spielberg[11], Narula, Ogbu and Samuelsson[20], Cornuejols, Fisher and

Nemhauser[3], Erlenkotter[6], Galvao[9], Magnanti and Wong[14], Nemhauser and Wolsey[21], Mulvey and Crowder[19], Mavrides[16], Mirchandani, Oudjit and Wong[17], Christofides and Beasley[2], Beasley[1].)

Most of the computational experience has been reported on test problems with $n \leq 100$. For many of these test problems, $Z_{IP}=Z_{LP}$. Recently, Beasley[1] solved forty larger problems (with $100 \leq n \leq 900$) and found a small but positive gap $Z_{IP}-Z_{LP}$ for many of them, The average of $\frac{Z_{IP}-Z_{LP}}{Z_{IP}}$ over these problems was .0024.

In this paper we analyze the ratio $\frac{Z_{IP}-Z_{LP}}{Z_{IP}}$ from a probabilistic point of view as n goes to infinity, under some assumptions on the probability distribution of problem instances. We do not address the worst-case analysis of this ratio except to note that this question was solved by Cornuejols, Fisher and Nemhauser [3] when $d_{ij} \leq 0$. The analysis of [3] does not carry over when the d_{ij}'s are nonnegative and satisfy the distance axioms. In fact, this worst-case analysis is an interesting open question. It would also be interesting to know the worst-case value of $\frac{Z_{IP}-Z_{LP}}{Z_{IP}}$ when the d_{ij}'s are further restricted to represent Euclidean distances. Once again, these questions are not addressed here as we focus on a probabilistic approach.

We will often write statements like $X_n \leq u_n$ *almost surely* (a.s.) for a sequence of random variables (X_n) and real sequence (u_n). This is a well-defined terminology of probability theory and details can be found in Stout [25] for example. We will *invariably* prove that

$$\sum_{n=1}^{\infty} \Pr(X_n > u_n) < \infty$$

which implies the above statement. Non-probabilists will be satisfied that we show $\Pr(X_n > u_n) \to 0$ as $n \to \infty$. If $X_n \leq u_n(1+O(1))$ a.s. and $X_n \geq u_n(1-O(1))$ a.s. then we write $X_n \sim u_n$ a.s.

We study the k-median problem in the plane. When points $X_1, \ldots, X_n$ are uniformly distributed in a unit square and d_{ij} is the Euclidean distance between X_i and X_j, $1 \leq i, j \leq n$, we show that $\frac{Z_{IP}-Z_{LP}}{Z_{IP}} \sim .00284$ almost surely, for any

k such that $\omega \leq k \leq \dfrac{n}{\omega \log n}$ where $\omega = \omega(n) \to \infty$. (In this paper we abbreviate $f(n) \to a$ as $n \to \infty$ by $f(n) \to a$.)

In section 4 we put our probabilistic results in perspective by presenting extensive computational experiments.

In section 5, we show how our results for the the k-median problem relate to the *simple plant location problem* (SPLP). In the SPLP, the data comprise n points $X_1, \ldots, X_n$, distances d_{ij} for $1 \leq i, j \leq n$, and fixed costs f_j associated with each point X_j, $1 \leq j \leq n$. The SPLP consists of finding a nonempty set $S \subseteq X$ that minimizes $\sum_{i=1}^{n} \min_{j \in S} d_{ij} + \sum_{j \in S} f_j$. (Note that, in this problem, $|S|$ is not restricted as in the k-median problem.) An integer programming formulation of

$$Z_{IP} = \min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij} + \sum_{j=1}^{n} f_j x_j$$

SPLP is subject to (2), (4) and (5). The LP relaxation is obtained by relaxing the integrality conditions (5).

In the remainder of this section we state some useful results from the literature. Our proofs use the following lemma (see Hoeffding[12]).

Lemma 1. If $Y_1, \ldots, Y_n$ are independent random variables and $0 \leq Y_i \leq 1$ for $i = 1, \ldots, n$, then, for $0 < \varepsilon < 1$,

$$\Pr(\bar{Y} \geq (1+\varepsilon)\mu \leq e^{-\varepsilon^2 n\mu/3} \text{ and}$$

$$\Pr(\bar{Y} \leq (1-\varepsilon)\mu) \leq e^{-\varepsilon^2 n\mu/2},$$

where $\bar{Y} = \left(\sum_{i=1}^{n} Y_i\right)/n$ and μ is the expected value of $\bar{Y}$.

Given a vector $x = (x_j : j = 1, \ldots, n)$ such that $\sum_j x_j = k$ and $0 \leq x_j \leq 1$ for all j, define

$$Z_{LP}(x) = \min \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} y_{ij}$$

$$\sum_{j=1}^{n} y_{ij} = 1 \text{ for } i = 1, \ldots, n$$

$$0 \leq y_{ij} \leq x_j \text{ for } i, j, = 1, \ldots, n.$$

Note that $Z_{LP} = \min Z_{LP}(x)$

$$\sum_j x_j = k$$

$$0 \leq x_j \leq 1 \text{ for } j = 1, \ldots, n.$$

The following lemma is well-known in the k-median literature and is easy to prove.

Lemma 2. An optimal solution $y = (y_{ij} : i, j = 1, \ldots, n)$ of $Z_{LP}(x)$ is obtained as follows. For each i, sort the values d_{ij}, $j = 1, \ldots, n$, so that

$$d_{ij_1(i)} \leq d_{ij_2(i)} \leq \cdots \leq d_{ij_n(i)},$$

and let p be such that $\sum_{h=j_1(i)}^{j_{p-1}(i)} x_h \leq 1 \leq \sum_{h=j_1(i)}^{j_p(i)} x_h$.

Then

$$y_{ij} = \begin{cases} x_j & \text{for } j = j_1(i), \ldots, j_{p-1}(i) \\ 1 - \sum_{h=j_1(i)}^{j_{p-1}(i)} x_h & \text{for } j = j_p(i) \\ 0 & \text{for } j = j_{p+1}(i), \ldots, j_n(i). \end{cases}$$

Proof. The program $Z_{LP}(x)$ separates for each j into a linear program with upper bounded variables and a single constraint.

Let $d_i = \sum_{j=1}^{n} d_{ij} \; y_{ij}$ where the values of y_{ij} are those defined in Lemma 2. Note that $Z_{LP} \leq \sum_{i=1}^{n} d_i$ since this bound is derived from a primal feasible solution. This bound will be used repeatedly in our proofs where it is computed for the vector x defined by $x_j = k/n$ for $j = 1, \ldots, n$.

The dual of the LP relaxation is

$$\begin{aligned} & Z_{LP} = \max \sum_{i=1}^{n} u_i - \sum_{j=1}^{n} v_j - kw \\ & u_i - t_{ij} \leq d_{ij} \text{ for all } i, j \\ & \sum_{i=1}^{n} t_{ij} - v_j - w \leq 0 \text{ for all } j \\ & t_{ij}, \; v_j \geq 0 \text{ for all } i, j. \end{aligned} \tag{6}$$

For any given vector $u = (u_i : i = 1, \ldots, n)$, define

$$\rho_j(u) = \sum_{i=1}^{n} (u_i - d_{ij})^+ \text{ for } j = 1, \ldots, n,$$

where a^+ denotes max $(0, a)$. Let $Z_D(u)=\sum_{i=1}^{n} u_i-k \max_{j=1,\cdots,n} \rho_j(u)$.

Lemma 3. $Z_{LP}\geq Z_D(u)$ for any vector u.

Proof: It can be checked that, for any given u, a feasible solution of (6) is obtained by setting $t_{ij}=(u_i-d_{ij})^+$, $v_j=0$ and $w=\max_{j=1,\ldots,n} \rho_j(u)$.

3. The Euclidean model in the plane

This section is concerned with the following Euclidean model: n points X_1, ..., X_n are chosen independently and uniformly at random in the unit square $S_0=[0,1]^2$. The distance matrix is given by $d_{ij}=||X_i-X_j||$ for $1\leq i, j\leq n$ where $||\cdot||$ denotes the Euclidean norm. We assume that

$$k\rightarrow\infty \text{ and } n/(\mathrm{klogn})\rightarrow\infty. \tag{7}$$

The following theorem was proved by Papadimitriou [22].

Theorem 1 Under the above conditions,

$$Z_{IP}\sim(.3771967\ldots)n/\sqrt{k} \text{ a.s.}$$

This result was obtained by comparing Z_{IP} to the value Z_C of finding k points in $X=\{X_1, \ldots, X_n\}$ that minimize the sum of the distance to a *continuum* of points in the unit square. Papadimitriou showed that, when (7) holds, $Z_{IP}\sim Z_C$ almost surely. Actually, he used a weaker notion of probabilistic convergence, but Zemel [26] showed that almost sure convergence holds as well. It should be pointed out, however, that the continuous problem yielding Z_C is very different from the LP relaxation. In fact, for the LP relaxaton, we prove

Theorem 2. Under the above conditions,

$$Z_{LP}\sim\frac{2}{3\sqrt{\pi}}\, n/\sqrt{k} \text{ a.s.}$$

where $2/(3\sqrt{\pi})=.3761264\ldots$

Our method of proof consists of conjecturing a near-optimal solution to the LP relaxation and a near-optimal solution to its dual. Then we show that, almost surely, these lower and upper bounds on Z_{LP} are the same, up to small order terms. The probabilistic arguments are based on the estimates of the tails of the binomial distribution given in Lemma 1.

The proof of Theorem 2 will actually provide a constructive way of obtaining an upper bound $Z_{LP}(x)$ and a lower bound $Z_D(u)$ on the optimum value of the LP relaxation of the k-median problem.

Corollary 1. Let $x_j=k/n$ for $j=1, \ldots, n$ and $u_i=\sqrt{k/\pi}$ for $i=1, \ldots, n$. Then $Z_D(u)\leq Z_{LP}\leq Z_{LP}(x)$ and, under condition (7),

$$Z_D(u)\sim Z_{LP} \quad \text{almost surely,}$$

$$Z_{LP}(x)\sim Z_{LP} \quad \text{almost surely.}$$

In addition, in [22], Papadimitriou gives a heuristic which almost surely provides a solution with value $Z_H\sim Z_{IP}$. The complexity of the heuristic is O(nlogn). Combining this result with the fact that $Z_D(u)$ can be computed in linear time, we have a very fast procedure which will almost surely

(i) find a solution with a value close to the optimum,

(ii) prove that the value of this solution is within 0. 3% of the optimum.

Finding the exact optimum is much more expensive as will be shown in Theorem 3. But first we give the proof of Theorem 2.

Proof of Theorem 2. To obtain a probabilistic upper bound on Z_{LP}, we are first going to consider the LP solution

$$x_j = k/n \text{ for } j=1, \ldots, n$$

and the values of y_{ij} as defined in Lemma 2. Let $d_i = \sum_{j=1}^{n} d_{ij}\, y_{ij}$ for $i=1, \ldots, n$.

We must get a probabilistic estimate of d_i for $i=1, \ldots, n$. Let $\varepsilon=\left(\frac{\text{klogn}}{n}\right)^{1/3}$, $r=\left(\frac{1}{k\pi(1-\varepsilon)}\right)^{1/2}$ and let S_r be the square $[r,\ 1-r]^2$. We show first

$$\Pr\left(d_i\geq\frac{2}{3\sqrt{k\pi}}\ (1+O(1))\,|\,X_i\in S_r\right)\leq 2e^{-\frac{\varepsilon^2 n}{9k}} \tag{8}$$

$$\Pr\left(d_i\geq\frac{4}{3\sqrt{k\pi}}\ (1+O(1))\,|\,X_i\in S_r\right.\leq 2e^{-\frac{4\varepsilon^2 n}{9k}} \tag{9}$$

If $X_i\in S_r$, then a circle C_i of radius r centered at X_i is entirely contained in S_0. The number N of points lying in this circle stochastically dominates the binomial B $(n, \pi r^2)$ (since $X_i \in C_i$). We define independent random variables W_j, $j=1, 2, \dots, n$ as follows:

Let

$$W_j = \begin{cases} d_{ij} & \text{if } X_j \in C_i \\ 0 & \text{otherwise.} \end{cases}$$

We note that $E(W_j) = 2\pi r^3/3$ $(j \neq i)$. If $N \geq \lceil \frac{n}{k} \rceil$ then $d_i \leq \frac{k}{n} \sum_{j=1}^{n} W_j$. Now, by Lemma 1,

$$\Pr\left(N < \left\lceil \frac{n}{k} \right\rceil\right) = \Pr\ (N \leq (1-\varepsilon) n\pi r^2) \leq e^{-\frac{\varepsilon^2}{2} n\pi r^2}.$$

Furthermore, if $\hat{W}_j = W_j / r \in [0, 1]$, then by Lemma 1,

$$\Pr\left(\sum_{j=1}^{n} \hat{W}_j \geq (1+\varepsilon)(n-1) \frac{2\pi r^2}{3}\right) \leq e^{-\frac{\varepsilon^2}{3}(n-1)\frac{2\pi r^2}{3}}$$

and (8) follows.

To prove (9), we note that if $X_i \in S_0 - S_r$, we can at worst find a quadrant of a circle centered at X_i with radius $2r$ and contained entirely within S_0. The area of this quadrant is $\pi(2r)^2/4$ and we apply the same method as above with $E(W) = 4\pi r^3/3$.

We are now ready to bound Z_{LP}.

$$Z_{LP} \leq \sum_{i=1}^{n} d_i = \sum_{X_i \in S_r} d_i + \sum_{X_i \in S_0 - S_r} d_i.$$

By Lemma 1,

$$\Pr\{|X \cap S_r| \leq n(1-2r)^2(1-\varepsilon)\} < e^{-\frac{\varepsilon^2}{2} n(1-2r)^2}$$

and thus

$$\Pr\left\{ Z_{LP} \geq (1+O(1)) \left[(1-2r)^2 n \frac{2}{3\sqrt{k\pi}} + (1-(1-2r)^2) n \frac{4}{3\sqrt{k\pi}} \right] \right\}$$

$$\leq (2n+1) e^{-\frac{2\varepsilon^2}{9} n/k}$$

giving

$$Z_{LP} \leq (1+O(1)) \frac{2n}{3\sqrt{k\pi}} \text{ almost surely.} \tag{10}$$

To obtain a probabilistic lower bound on Z_{LP}, we consider the dual problem (6). Let $u_i=r$ for $i=1 \dots n$. Then by Lemma 3

$$Z_{LP} \geq \sum_{i=1}^{n} u_i - k \max_j \left[\sum_{i=1}^{n} (u_i - d_{ij})^+ \right] \tag{11}$$

For fixed j, consider random variables $U_i=(u_i-d_{ij})^+$.

Setting $u_i=r$ we find $E(U_i)=\frac{\pi r^3}{3}$ for $i \neq j$ and $X_j \in S_r$, whereas these values decrease for points $X_j \in S_o - S_r$. Rescaling U to [0, 1] and applying Lemma 1 to $X_j \in S_r$ we find

$$\Pr\left(\sum_{i=1}^{n} U_i \geq (1+\varepsilon)\frac{n\pi r^3}{3}\right) \leq e^{-\frac{\varepsilon^2}{9} n/k}$$

and thus for $k=O\left(\frac{n}{\log n}\right)$ we have

$$\operatorname*{Max}_j \left(\sum_{i=1}^{n} U_i\right) \leq (1+\varepsilon)\frac{n\pi r^3}{3} \quad \text{a.s.}$$

giving

$$Z_{LP} \geq nr - (1+\varepsilon) kn\pi r^3/3 = (1-O(1))\frac{2n}{3\sqrt{k\pi}} \quad \text{a.s.} \tag{12}$$

Combining this with (10) yields the theorem.

One might expect then that an LP-based branch and bound procedure performs well, since Z_{LP} provides a good bound. However, we can prove

Theorem 3. Assume $k/\log n \to \infty$ and $n/k^2 \log n \to \infty$.

Then there exists a constant $\alpha > 0$ such that a branch and bound procedure that branches by fixing a variable x_j to 0 or 1 at each node node of the search tree which is not pruned and uses the LP bound to prune the search tree will almost surely explore at least $n^{\alpha k}$ nodes.

Proof: Each node of the branch and bound tree is associated with two sets J_o and J_1 where $J_t = \{j : x_j \text{ is fixed at } t \text{ in the associated subproblem}\}$ for $t=0, 1$. Let $Z_{LP}(J_0, J_1)$ denote the LP bound computed at this node, i.e. the value of Z_{LP} when we make the restriction $x_j=t$ for $j \in J_t$, $t=0, 1$. We prove the theorem by showing that for some constants $\beta, \gamma > 0$ (to be determined) the following holds almost surely:

For any $J_0, J_1 \subset \{1, \ldots, n\}$ such that (13)
$J_0 \cap J_1 = \phi$, $|J_0| \leq \beta n/k \log n$, $|J_1| \leq \gamma k$, we have
$Z_{LP}(J_0, J_1) \leq .3769 \frac{n}{\sqrt{k}}$

For then we almost surely have to branch at every at every node in which $|J_0| \leq \beta n/k \log n$ and $|J_1| \leq \gamma k$ even if we have an optimal solution of the integer program as our current best solution-by Theorem 1.

This implies that the algorithm must explore at least

$$\binom{\lfloor \beta n/k \log n \rfloor + \lfloor \gamma k \rfloor}{\lfloor \gamma k \rfloor} = n^{\gamma(1-o(1))k} \text{ nodes.} \quad (14)$$

Since β can be chosen arbitrarily close to 1 the theorem will follow. To verify (14) imagine that setting $x_j=0$ means branching to the left and setting $x_j=1$ means branching to the right. (13) implies that our tree contains a copy of all possible paths which make $\lfloor \gamma k \rfloor$ right branches and $\lfloor \beta n/k \log n \rfloor$ left branches. The number of such paths is precisely the left hand side of (14).

Let F denote the family of such pairs J_0, J_1.

Thus let $J_0, J_1 \subset \{1, \cdots, n\}$ be disjoint, $\bar{J} = \{j \notin J_0 \cup J_1\}$, $\bar{n} = |\bar{J}|$, and $\bar{k} = k - |J_1|$. Consider the following solution to the associated linear program.

$$x_j = \begin{cases} 0 & \text{if } j \in J_0 \\ 1 & \text{if } j \in J_1 \\ k/n & \text{if } j \in J. \end{cases}$$

The values of y_{ij} are then defined as in Lemma 2, but only using $j \in J$ to form the sequence $j_1(i), j_2(i), \ldots, j_n(i)$. This choice of y_{ij} is feasible although usually not optimum. However this is sufficient since we only need to compute an upper bound on $Z_{LP}(J_0, J_1)$. We can assume w.l.o.g. that $|J_0| = \lfloor \beta n/k \log n \rfloor$ and $|J_1| = \lfloor \alpha k \rfloor$. Let $\varepsilon > 0$ be small and $r = \sqrt{\frac{1}{(1-\varepsilon)\pi \bar{k}}}$ and proceed as in the proof of Theorem 2, defining variables $W_1, W_2, \ldots, W_{\bar{n}}$ for each i. We find that for $\varepsilon < \frac{1}{2}$ and n large

$$\Pr\left[Z_{LP}(J_0, J_1) > \frac{2n}{3\sqrt{\pi k}}(1+3\varepsilon) \right] \leq (2n+1) e^{-\frac{2\varepsilon^2 \bar{n}}{9\bar{k}}}.$$

Since $|F| \leq n^{\beta n/k \log n + \gamma k}$ we find

$$\Pr\left[\exists (J_0, J_1)\varepsilon F : Z_{LP}(J_0, J_1) > \frac{2n}{3\sqrt{\pi k}}\ (1+3\varepsilon)\right] \leq$$

$$(2n+1)\, n^{\beta n/k \log n + \gamma k} e^{-\frac{2\varepsilon^2 \bar{n}}{9\bar{k}}}.$$

Taking $\beta=\varepsilon^2/5$, $\gamma=\varepsilon$ and ε sufficiently small that $\frac{2(1+3\varepsilon)}{3\sqrt{\pi}(1-\varepsilon)} \leq .3769$ yields

$$\max\{Z_{LP}(J_0, J_1) : (J_0, J_1) \in F\} \leq .3769 \frac{n}{\sqrt{k}} \text{ almost surely.}$$

Any $\alpha < \gamma$ can be used to give the theorem.

4. Computational Experience

The previous section provide asymptotic results as $n \to \infty$ for a classical Euclidean model in the plane. In this section, we report our computational experience with medium-size k-median problems for a Euclidean model This computational experience is based on the solution of about 3, 300 random problems with $n=$ 50 points and an additional 950 random problems with $n=100$ points. The description of these problems is given later.

For each problem we computed Z_{IP} and Z_{LP}. The value of Z_{LP} was obtained by solving a Lagrangian dual by subgradient optimization as explained in [3]. In the process of computing Z_{LP}, this algorithm generates a feasible solution at each subgradient iteration. Of course, if it happens that the value of the best feasible solution generated equals Z_{LP}, the algorithm terminates since, then, $Z_{IP} = Z_{LP}$. For most of the test problems with no gap $Z_{IP} - Z_{LP}$, the algorithm terminated in less than 100 subgradient iterations, due to the above stopping criterion. If, after 100 subgradient iterations, there was still a gap between the best feasible solution (an upper bound on Z_{IP}) and the best Lagrangian relaxation (a lower bound on Z_{LP}), we resorted to branch and bound to find Z_{IP}. When the subgradient algorithm clearly converged to a value different from Z_{IP}, we accepted it as showing that $Z_{IP} \neq Z_{LP}$. In the cases where the subgradient algorithm converged to a value close to Z_{IP} we used the simplex algorithm to compute Z_{LP}. This allowed us to settle cases where was a very small but positive gap $Z_{IP}-Z_{LP}$.

Among the 4250 test problems that we generated we found about 3700 such that $Z_{IP}=Z_{LP}$ and about 550 with a gap $Z_{IP}-Z_{LP}$. Now we give a detailed description of these results.

The first set of experiments involves Euclidean problems. We decided to test whether approximating the Euclidean distances had an influence on the gap $Z_{IP}-Z_{LP}$, since we suspected that data accuracy might be partly responsible for the discrepancy between the computational experience previously reported in the literature, namely few test problems were found to have gaps ([2], [3], [6], [10], [11], [19], [20], [23], [24]), and the results of Section 3 stating the asymptotically most instances should have small but positive gaps. To our surprise, data accuracy had little influence except maybe for the possibility that a very coarse approximation produces harder k-median problems. (These problems are more combinatorial, often have alternate optimal solutions and, in our experience, optimality was harder to prove). We generated 10 problems, each with 50 points occurring at random in the unit square. Then, for $i=1,2,3,4$ and 5, we multiplied each point coordinate by 10^i and rounded it to the closest integer value. The Euclidean distances were then computed and rounded to the closest integer. The k-median problem and its LP relaxation were solved for each $2\leq k\leq 10$ and $1\leq i\leq 5$. For each such pair i, k, Table 1 reports the number of problems (out of 10) with a gap $Z_{IP}-Z_{LP}$.

The same two problems were responsible for all the gaps. The average value of $\dfrac{Z_{IP}-Z_{LP}}{Z_{IP}}$ over the instances that had a gap was approximately 1.5% for

Table 1 Euclidean model with $n=50$. Number of instances with a gap.

i \ k	2	3	4	5	6	7	8	9	10	Total (out of 90)
1	0	2	0	2	2	1	0	0	0	7
2	0	1	0	0	0	2	0	0	1	4
3	0	1	0	1	0	0	2	0	0	4
4	0	1	0	0	1	0	2	0	0	4
5	0	1	0	0	1	0	2	0	0	4
Total (out of 50)	0	6	0	3	4	3	6	0	1	23 (out of 450)

$i=1$, .4% for $i=2$ and .1% for $i=3, 4$ and 5. Overall, the fraction of instances with a gap was about 5%. This is consistentwith the computational experience reported in the literature. Clearly, the asymptotic behavior described in Section 3 is not felt for problems with $n=50$ points. It would be interesting to repeat the computational experiment for Euclidean k-median problems with about $n=1000$ points. Unfortunately our computer budget did not allow to do this.

5. The Simple Plant Location Problem

Although we proved our probabilistic results for the k-median problem, they can also be useful for the SPLP. To define an instance of SPLP, we need fixed costs f_j, $j=1, \dots, n$, in addition to the distances d_{ij}, $1 \leq i, j \leq n$. For simplicity, we assume in this section that the fixed costs f_j are all identical, say $f_j=f$.

Theorem 4 Consider the Euclidean model in the plane and assume that $n^{\varepsilon-1/2} \leq f \leq n^{1-\varepsilon}$ for some fixed $\varepsilon > 0$. Then, for the SPLP,

$$\frac{Z_{IP}-Z_{LP}}{Z_{IP}} \sim .00189255\ldots \text{ almost surely.}$$

Proof. In this proof, Z_{IP} and Z_{LP} denote the optimum values of SPLP and its linear programming relaxation respectively. The solutions of the corresponding k-median problem (with same d_{ij}'s) and its relaxation are denoted by $Z_{IP}(k)$ and $Z_{LP}(k)$ respectively.

By definition $Z_{LP} = \min_k \ (Z_{LP}(k) + kf) = \min \ (Z_1, Z_2, Z_3)$, where

$$Z_1 = \min_{k<\omega} \ (Z_{LP}(k) + kf),$$

$$Z_2 = \min_{\omega \leq k \leq \frac{n}{\omega \log n}} \ (Z_{LP}(k) + kf), \text{ and}$$

$$Z_3 = \min_{k > \frac{n}{\omega \log n}} \ (Z_{LP}(k) + kf).$$

First we compute Z_2. From the proof of Theorem 2,

$$\Pr\left\{ Z_{LP} \notin \left[\frac{2n}{3\sqrt{k\pi}} (1-O(1)),\ \frac{2n}{3\sqrt{k\pi}} (1+O(1)) \right] \right\} = O\ (ne^{-2\omega^{1/3}\log n/q})$$

and so

$$Z_2 \leq \min_{\omega \leq k \leq \frac{n}{\omega \log n}} \left\{ \frac{2n}{3\sqrt{k\pi}} (1+O(1)) + kf \right\} \text{almost surely.}$$

Let $\alpha = \frac{2}{3\sqrt{\pi}}$. The minimum of the function $\frac{\alpha n}{\sqrt{k}} + kf$ is attained when $k = \left(\frac{\alpha n}{2f}\right)^{2/3}$. Note that, given our assumptions on f, this value is in the range $\left[\omega, \frac{n}{\omega \log n}\right]$ for a suitable ω, say $\omega = \log n$. The minimum value of the function is $\left[\frac{27}{4}\alpha^2 n^2 f\right]^{1/3}$. Therefore

$$Z_2 = \left[\frac{27}{4}\alpha^2 n^2 f\right]^{1/3} (1+O(1)) \text{ almost surely.}$$

Now consider Z_3. With our choice of $\omega = \log n$, we have $k > \frac{n}{(\log n)^2}$. Therefore, almost surely,

$$\begin{aligned} Z_3 &\geq \frac{n}{(\log n)^2} f \\ &= \frac{n^{1/3} f^{2/3}}{(\log n)^2} \frac{Z_2}{\left(\frac{27}{4}\alpha^2\right)^{1/3}} (1+O(1)) \geq Z_2. \end{aligned}$$

Finally consider Z_1. For all $k < \log n$, we have $Z_{LP}(k) \geq Z_{LP}(\log n)$. Therefore $Z_1 \geq Z_{LP}(\log n)$. This implies that, almost surely,

$$Z_1 \geq \frac{2n}{3\sqrt{\pi}\log n} (1+O(1)) = c\frac{n^{1/3} f^{-1/3}}{(\log n)^{1/2}} Z_2 (1+O(1)) \geq Z_2,$$

where c is a constant.

We have just proved that

$$Z_{LP} \sim \left[\frac{27}{4}\alpha^2 n^2 f\right]^{1/3} \text{ almost surely.}$$

Similarly, $Z_{IP} = \min_k (Z_{IP}(k) + kf)$. Following the proof of Papadimitriou [22], we can show that

$$Z_{IP} = \min_k \frac{\beta n}{\sqrt{k}} (1+O(1)) + fk \text{ almost surely,} \tag{15}$$

where $\beta = .3771967\ldots$. The minimum in (15) is achieved when $k = \left(\frac{\beta n}{2f}\right)^{2/3}$

and its value is $\left(\frac{27}{4}\beta^2 n^2 f\right)^{1/3}$ $(1+0(1))$.

So $\frac{Z_{IP}-Z_{LP}}{Z_{IP}} \sim \frac{\beta^{2/3}-\alpha^{2/3}}{\beta^{2/3}}$ almost surely.

Similarly, the next result can be shown using the proof of Theorem 8.

Theorem 5 Consider the uniform cost model and assume that $n^{\varepsilon-1} \leq f \leq n^{1-\varepsilon}$ for some fixed $\varepsilon > 0$. Then

$$\frac{Z_{IP}-Z_{LP}}{Z_{IP}} \sim 1-\frac{\sqrt{2}}{2} \text{ almost surely.}$$

6. Conclusion

The LP relaxation (1) − (4) has been widely used in branch and bound algorithms for the k-median problem and has bee reported to provide a tight bound in practice. Our analysis shows that such good results can indeed be expected in a probabilistic sense for some problem instances, but we also identify other instances where the LP relaxation is almost surely not tight. The probabilistic analysis is performed for a classical Euclidean model in location theory. That is, let $\omega = \omega(n) \to \infty$. When $\omega \leq k \leq \frac{n}{\omega \log n}$ in the Euclidean model, $Z_{LP}/Z_{IP} = .99716\ldots + 0(1)$ almost surely.

Our computational experience confirms that only small gaps were observed with a classical Euclidean model.

Another aspect of the probabilistic analysis performed in Section 3 is that, under various assumptions, branch and bound algorithms must almost surely expand a non-polynomial number of nodes to solve k-median problems to optimality.

Finally, we mention as open problems the questions of describing the asymptotic behavior of Z_{LP}/Z_{IP} as $n \to \infty$ when $k \geq \frac{n}{\log n}$ in the Euclidean model.

References

〔1〕 J.E. Beasley "A Note on Solving Large p-Median Problems," Technical Report, Department of Management Science, Imperial College, London, England (September 1984).

〔2〕 N. Christofides and J.E. Beasley "A Tree Search Algorithm for the p-Median Problem," *European Jurnal of Operational Research* 10 (1982), 196-204.

〔3〕 G. Cornuejols, M.L. Fisher and G.L. Nemhauser "Location of Bank Accounts to Optimize Float: An Analytical Study of Exact and Approximate Algorithms," *Management Science* 23 (1977), 789-810.

〔4〕 G. Cornuejols, G.L. Nemhauser and L.A. Wolsey "Worst-Case and Probabilistic Analysis of Algorithms for a Location Problem," *Operations Research* 28 (1980), 847-858.

〔5〕 G. Diehr "An Algorithm for the p-Median Problem," Working Paper No. 191, Western Management Science Institute, University of Calfornia, Los Angeles (1972).

〔6〕 D. Erlenkotter "A Dual-Based Procedure for Uncapacitated Facility Location," *Operations Research* 26 (1978), 992-1009.

〔7〕 S. Even, *Graph Algorithms*, Computer Science Press, Potomac, Maryland (1979).

〔8〕 M.L. Fisher and D.S. Hochbaum "Probabilistic Analysis of the Planar k-Median Problem," *Mathematics of Operations Research* 5 (1980), 27-34.

〔9〕 R.D. Galvão "A Dual-Bounded Algorithm for the p-Median Problem," *Operations Research* 28 (1980), 1112-1121.

〔10〕 R.S. Garfinkel, A.W. Neebe and M.R. Rao "An Algorithm for the m-Median Plant Location Problem," *Transportation Science* 8 (1974), 217-236.

〔11〕 M. Guignard and K. Spielberg "Algorithms for Exploiting the Structure of the Simple Plant Location Problem," *Annals of Discrete Mathematics* 1 (1977), 247-271.

〔12〕 Hoeffding "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association* 58 (1963), 13-30.

〔13〕 A. Kolen "Solving Covering Problems and the Uncapacitated Plant Location Problem on Trees," *European Journal of Operational Research* 12 (1983), 266-278.

〔14〕 T.L. Magnanti and R.T. Wong "Accelerating Benders Decomposition: Algorithmic Enhancement and Model Selection Criteria," *Operations Research* 29 (1981), 464-484.

〔15〕 R.E. Marsten "An Algorithm for Finding Almost All of the Medians of a Network," Discussion Paper No. 23, The Center for Mathematical Studies in Econometrics and Management Science, Northwestern University, Evanston, Illinois (1972).

[16] L.P. Mavrides "An Indirect Method for the Generalized k-Median Problem Applied to Lock-Box Location," *Management Science* 25 (1979), 990-996.

[17] P.B. Mirchandani, A. Oudjit and R.T. Wong "Locational Decisions on Stochastic Multidimensional Networks" (1983).

[18] C. Mukendi "Sur l'implantation d'equipement dans un reseau: le probleme de m-centre", Thesis, University of Grenoble, France (1975).

[19] J.M. Mulvey and H.L. Crowder "Cluster Analysis: An application of Lagrangian Relaxation," *Management Science* 25 (1979), 329-340.

[20] S.C. Narula, U.I. Ogbu and H.M. Samuelsson "An Algorithm for the p-Median Problem," *Operations Research* 25 (1977), 709-713.

[21] G.L. Nemhauser and L.A. Wolsey "Maximizing Submodular Set Functions: Formulations and Analysis of Algorithms," *Annals of Discrete Mathematics* 11 (1981), 279-301.

[22] C.H. Papadimitriou "Worst-Case and Probabilistic Analysis of a Geometric Location Problem," *SIAM Journal on Computing* 10 (1981), 542-557.

[23] Ch. S. ReVelle and R.W. Swain "Control Facilities Location," *Geographical Analysis* 2 (1970), 30-42.

[24] L. Schrage "Implicit Representation of Variable Upper Bounds in Linear Programming," *Mathematical Programming Study* 4 (1975), 118-132.

[25] W.F. Stout, *Almost Sure Convergence*, Academic Press, New York (1974).

[26] E. Zemel "Probabilistic Analysis of Geometric Location problems," *SIAM Journal of Algebraic and Discrete Methods* 6, (1985), 189-200.