

Improvements on the k-center problem for uncertain data

Sharareh Alipour

Amir Jafari

ABSTRACT

In real applications, there are situations where we need to model some problems based on uncertain data. This leads us to define an uncertain model for some classical geometric optimization problems and propose algorithms to solve them. In this paper, we study the k -center problem, for uncertain input. In our setting, each uncertain point P_i is located independently from other points in one of several possible locations $\{P_{i,1}, \dots, P_{i,z_i}\}$ in a metric space with metric d , with specified probabilities and the goal is to compute k -centers $\{c_1, \dots, c_k\}$ that minimize the following expected cost

$$Ecost(c_1, \dots, c_k) = \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} \min_{j=1, \dots, k} d(\hat{P}_i, c_j)$$

here Ω is the probability space of all realizations

$$R = \{\hat{P}_1, \dots, \hat{P}_n\}$$

of given uncertain points and

$$prob(R) = \prod_{i=1}^n prob(\hat{P}_i).$$

In restricted assigned version of this problem, an assignment $A : \{P_1, \dots, P_n\} \rightarrow \{c_1, \dots, c_k\}$ is given for any choice of centers and the goal is to minimize

$$Ecost_A(c_1, \dots, c_k) = \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, A(P_i)).$$

In unrestricted version, the assignment is not specified and the goal is to compute k centers $\{c_1, \dots, c_k\}$ and an assignment A that minimize the above expected cost.

We give several improved constant approximation factor algorithms for the assigned versions of this problem in a Euclidean space and in a general metric space. Our results significantly improve the results of [14] and generalize the results of [26] to any dimension. Our approach is to replace a certain center point for each uncertain point and study the properties of these certain points. The proposed algorithms are efficient and simple to implement.

Keywords. k -center problem, uncertain points, approximation algorithm.

1. INTRODUCTION

It is not surprising that in many real-world applications, we face uncertainty about the data. Database systems should be able to handle and correctly process

these uncertain data. Most of the time, we need to deal with optimization problems in data bases, such as data integration, streaming, cluster computing and sensor network applications that involve parameters and inputs whose values are known only with some uncertainty[14]. So, an important challenge for database systems is to deal with large amount of data with uncertainty.

In this paper we focus on a classical geometric optimization problem, k -center problem, for uncertain data. First, we introduce the uncertainty models and the previous works, then we propose our algorithms for these models.

Problem Statement

In a metric space X with metric d , the k -center problem for a set of (certain) points $\{P_1, \dots, P_n\}$ in X , asks for k center points $C = \{c_1, \dots, c_k\}$ in X that minimize the following cost

$$cost(c_1, \dots, c_k) = \max_{i=1, \dots, n} d(P_i, C),$$

where $d(P_i, C) = \min_{c \in C} d(P_i, c)$. When the points P_1, \dots, P_n are uncertain, each point has a finite number of possible locations independently from the other points with given probabilities. More precisely, we are given a set $D = \{D_1, \dots, D_n\}$ of n discrete and independent probability distributions. The i -th distribution, D_i is defined over a set of z_i possible locations $P_{i1}, \dots, P_{i,z_i} \in X$. A probability p_{ij} is associated with each location such that $\sum_j p_{ij} = 1$ for every $i \in [n] = \{1, \dots, n\}$ and $j \in \{1, \dots, z_i\}$. Thus, the probabilistic points can be considered to be independent random variables X_i . The locations together with the probabilities specify their distributions $Pr[X_i = P_{ij}] = p_{ij}$ for every $i \in [n]$ and $j \in [z_i]$. A probabilistic set Y , consisting of the probabilistic points, is therefore a random variable. Let $z = \max\{z_1, \dots, z_n\}$ be the maximum number of possibilities for uncertain points.

For simplicity, we use the notation \hat{P}_i for a realization of the uncertain point P_i and the $prob(\hat{P}_i)$ for its probability. We let Ω denote the probability space of all realizations $R = \{P_{1j_1}, \dots, P_{nj_n}\}$ with $prob(R) = \prod_{i=1}^n prob(P_{i,j_i})$.

There are three known versions of the k -center problem for uncertain points based on the definition of the cost function.

- **Unassigned version:**

Here the goal is to find k centers $C = \{c_1, \dots, c_k\}$ that minimize

$$Ecost(c_1, c_2, \dots, c_k) = \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, C).$$

- **Unrestricted assigned version:**

Here, all realizations of an uncertain point P_i are assigned to a center denoted by $A(P_i)$. In fact, all realizations of an uncertain point P_i in the assigned version are in the cluster of the same center. Therefore, the goal is to find k centers $\{c_1, \dots, c_k\}$ and an assignment $A : \{P_1, \dots, P_n\} \rightarrow \{c_1, \dots, c_k\}$ that minimize

$$Ecost_A(c_1, c_2, \dots, c_k) = \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, A(P_i)).$$

- **Restricted assigned version:**

Here for any set of uncertain points $\{P_1, \dots, P_n\}$ and k centers $\{c_1, \dots, c_k\}$ an assignment

$$A : \{P_1, \dots, P_n\} \rightarrow \{c_1, \dots, c_k\}$$

is given. The goal is to find $\{c_1, \dots, c_k\}$ that minimizes $Ecost_A(c_1, \dots, c_k)$

In this paper, we consider three assignments: the expected distance assignment that was first introduced in [26], the 1-center assignments and the expected point assignment for a Euclidean space where both of them are new in this paper as far as we know.

In the expected distance assignment, each uncertain point P_i is assigned to

$$ED(P_i) = \arg. \min_{Q \in \{c_1, \dots, c_k\}} \sum_{\hat{P}_i \in D_i} prob(\hat{P}_i) d(\hat{P}_i, Q).$$

In a Euclidean space, let

$$\bar{P}_i = \sum_{\hat{P}_i \in D_i} prob(\hat{P}_i) \hat{P}_i.$$

In the expected point assignment, each uncertain point P_i is assigned to

$$EP(P_i) = \arg. \min_{Q \in \{c_1, \dots, c_k\}} d(\bar{P}_i, Q).$$

Finally, in the 1-center assignment, let \tilde{P}_i be the 1-center of the single uncertain point P_i . An uncertain point P_i is assigned to

$$OC(P_i) = \arg. \min_{Q \in \{c_1, \dots, c_k\}} d(\tilde{P}_i, Q).$$

Related works

The deterministic k -center problem is a classical problem that has been extensively studied. It is well known that the k -center problem is NP-hard even in the plane [22] and approximation algorithms have been proposed (e.g., see [3, 4, 15]). Efficient algorithms were also given for some special cases, e.g., the smallest enclosing circle and its weighed version and discrete version [9, 20, 21], the Fermat-Weber problem [6], k -center on trees [5, 12, 23]. Refer to [8] for other variations of facility location problems. The deterministic k -center in one-dimensional space is solvable in $O(n \log n)$ time [24]. One of the most elegant approximation algorithms for k -center clustering is the 2-factor approximation algorithm by Gonzalez [13] which can be made to run in $O(n \log k)$ time [11]. One of the fastest methods for k -center clustering in 2 and 3 dimensions is by Aggarwal and Procopiuc [1] which uses a dynamic programming approach to k -center clustering and whose running time is upper bounded by $O(n \log k) + (\frac{k}{\epsilon})^{O(k^{1-\frac{1}{2}})}$. Another elegant solution to the k -center clustering problem was given by Badoiu et.a [4]. This algorithm gives a $(1 + \epsilon)$ -approximation factor algorithm which runs in $2^{O((k \log k)/\epsilon^2)} dn$ in \mathbf{R}^d . Another algorithm based on coresets runs in $O(k^n)$ [19] and it is claimed that the running time is much less than the worst case and thus it's possible to solve some problems when k is small (say $k < 5$).

Several recent works have dealt with clustering problems on probabilistic data. One approach was to generalize well-known heuristic algorithms to the uncertain setting. For example a clustering algorithm called DB-SCAN [10] was also modified to handle probabilistic data by Kriegel and Pfeifle [17, 18] and Xu and Li [27]. Refer to [2] for a survey on data mining of uncertain data.

Cormode and McGregor [7] introduced the study of probabilistic clustering problems. They developed approximation algorithms for the probabilistic settings of k -means, k -median as well as k -center clustering. They described a pair of bicriteria approximation algorithms, for inputs of a particular form; one of which achieves a $(1 + \epsilon)$ -approximation with a large blow up in the number of centers, and the other which achieves a constant factor approximation with only $2k$ centers.

Guha and Muhagala [14] improved upon the previous work. They achieved $O(1)$ -approximations in finite metric space, while preserving the number of centers both for assigned and unassigned version of the k -center problem. More precisely, the approximation factor of their algorithm for unrestricted assigned version is $15(1 + 2\epsilon)$ and the running time of their algorithm is polynomial in input size and $\frac{1}{\epsilon}$.

Munteanu and et.al. presented the first polynomial time $(1 + \epsilon)$ -approximation algorithm for the probabilistic smallest enclosing ball problem with extensions to the streaming setting [25].

Wang and Zhang [26], introduced the restricted as-

signed version under the expected distance assignment. They solved the one-dimensional k -center problem, in $O(zn \log zn + n \log k \log n)$ time. If dimension is one and the z locations of each uncertain point are sorted, they reduced the problem to a linear programming problem and thus solved the problem in $O(zn)$ time by applying a linear time algorithm.

Haung and Li [16] gave a PTAS for unassigned version of the probabilistic k -center problem in \mathbf{R}^d , when both k and d are constants.

2. MAIN RESULTS

In this paper, we propose several approximation algorithms for restricted and unrestricted assigned version of uncertain k -center problem. In this section, we state the main results and in the next section, we give their proofs.

Our main approach is to replace each uncertain point P with its expected point, \bar{P} , in the case of the Euclidean space or its 1-center, \bar{P} , in the case of the general metric space. Next, we compute the k -center for the described certain points and prove that this solution gives an approximation solution for the uncertain points. Note that there are efficient $(1+\epsilon)$ -approximation algorithms for the certain k -center problem in the literature.

1-center in Euclidean space

The first theorem gives a 2-approximation solution for the 1-center problem in the Euclidean space.

THEOREM 2.1. *Let P_1, \dots, P_n be a set of uncertain points in the Euclidean space, and*

$$\bar{P}_1 = \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) \hat{P}_1$$

be the expected point of P_1 . Then \bar{P}_1 is a 2-approximation solution for the 1-center problem for P_1, \dots, P_n .

Note that, we can compute \bar{P}_1 in $O(z)$ time which is independent of n .

Restricted assigned k -center problem in the Euclidean space

For the restricted assigned k -center problem in the Euclidean space, we have the following theorem.

THEOREM 2.2. *For a set of uncertain points P_1, \dots, P_n in a Euclidean space, let c_1, \dots, c_k be $(1+\epsilon)$ -approximation solution for the k -center problem for $\bar{P}_1, \dots, \bar{P}_n$. Let opt_{ED} and opt_{EP} be the minimum expected costs under the expected distance assignment and expected point assignment, respectively. Then,*

$$\text{Ecost}_{ED}(c_1, \dots, c_k) \leq (5 + \epsilon) \text{opt}_{ED}$$

and

$$\text{Ecost}_{EP}(c_1, \dots, c_k) \leq (3 + \epsilon) \text{opt}_{EP}.$$

Unrestricted assigned k -center problem

For unrestricted assigned k -center problem, we prove a stronger approximation algorithm for the Euclidean case and a slightly weaker one for a general metric space.

For the unrestricted version, we present theorems that indicate the relation between the restricted assignment and unrestricted assignment. Note that in the unrestricted version, we have to compute the optimal k centers and also the optimal assignment.

THEOREM 2.3. *For a set of uncertain points P_1, \dots, P_n in a metric space, the minimum expected cost under the expected distance assignment is a 3-approximation for the minimum expected cost for the unrestricted assigned k -center problem.*

So, any algorithm for the restricted assigned version under the expected point assignment gives a 3-approximation solution for the unrestricted assigned version. Since, the restricted assigned version under the expected distance assignment for \mathbf{R}^1 has exact solution [26], so we have a 3-approximation solution for the unrestricted assigned version in \mathbf{R}^1 . For higher dimensions, we present the following theorems.

THEOREM 2.4. *For a set of uncertain points P_1, \dots, P_n in a Euclidean space, let c_1, \dots, c_k be $(1+\epsilon)$ -approximation solution for the k -center problem for $\bar{P}_1, \dots, \bar{P}_n$. Let c_1^*, \dots, c_k^* and an assignment A be the optimal solution for the unrestricted assigned k -center problem for P_1, \dots, P_n . Then,*

$$\text{Ecost}_{ED}(c_1, \dots, c_k) \leq (5 + \epsilon) \text{Ecost}_A(c_1^*, \dots, c_k^*).$$

If, in the above theorem, instead of expected distance assignment we use the expected point assignment, then we get a better approximation factor.

THEOREM 2.5. *For a set of uncertain points P_1, \dots, P_n in a Euclidean space, let c_1, \dots, c_k be $(1+\epsilon)$ -approximation solution for the k -center problem for $\bar{P}_1, \dots, \bar{P}_n$. Let c_1^*, \dots, c_k^* and an assignment A be the optimal solution for the unrestricted assigned k -center problem for P_1, \dots, P_n . Then,*

$$\text{Ecost}_{EP}(c_1, \dots, c_k) \leq (3 + \epsilon) \text{Ecost}_A(c_1^*, \dots, c_k^*).$$

In a general metric space, we do not have the expected point construction and instead we use the 1-center \tilde{P}_i of the single uncertain point P_i .

THEOREM 2.6. *For a set of uncertain points P_1, \dots, P_n in a metric space, let c_1, \dots, c_k be $(1+\epsilon)$ -approximation solution for the k -center problem for $\tilde{P}_1, \dots, \tilde{P}_n$. Let c_1^*, \dots, c_k^* and an assignment A be the optimal solution for the unrestricted assigned k -center problem for P_1, \dots, P_n . Then,*

$$\text{Ecost}_{ED}(c_1, \dots, c_k) \leq (7 + 2\epsilon) \text{Ecost}_A(c_1^*, \dots, c_k^*).$$

If, in the above theorem, instead of expected distance assignment we use the 1-center assignment, then we get a better approximation factor.

THEOREM 2.7. *For a set of uncertain points P_1, \dots, P_n in a metric space, let c_1, \dots, c_k be $(1+\epsilon)$ -approximation solution for the k -center problem for $\bar{P}_1, \dots, \bar{P}_n$. Let c_1^*, \dots, c_k^* and an assignment A be the optimal solution for the unrestricted assigned k -center problem for P_1, \dots, P_n . Then*

$$Ecost_{OC}(c_1, \dots, c_k) \leq (5 + 2\epsilon)Ecost_A(c_1^*, \dots, c_k^*).$$

Note that the best constant approximation factor algorithm for the unrestricted assigned version, was $15(1+2\epsilon)$, with the polynomial running time in input size and $\frac{1}{\epsilon}$ [14].

Our results are summarized in Table 1. Note that the empty places for the running times are due to the fact that they depend on a $(1+\epsilon)$ -approximation algorithm used for the k -center problem of certain points.

3. PROOFS

In this section, we provide the proofs of the theorems stated in the previous section. First, we present two lemmas that are crucial for the rest of this section.

LEMMA 3.1. *For an uncertain point P in a Euclidean space and any point Q , we have*

$$d(\bar{P}, Q) \leq Ed(P, Q) = \sum_{\hat{P} \in D} prob(\hat{P})d(\hat{P}, Q)$$

where $\bar{P} = \sum_{\hat{P} \in D} prob(\hat{P})\hat{P}$ is the expected point of P .

PROOF. Since, $d(\bar{P}, Q)$ can be defined in terms of the Euclidean norm as $\|\bar{P} - Q\|$, using the triangle inequality

$$\begin{aligned} \|\bar{P} - Q\| &= \left\| \sum_{\hat{P} \in D} prob(\hat{P})\hat{P} - Q \right\| \\ &= \left\| \sum_{\hat{P} \in D} prob(\hat{P})(\hat{P} - Q) \right\| \leq \sum_{\hat{P} \in D} prob(\hat{P})\|\hat{P} - Q\| \\ &= Ed(P, Q). \end{aligned}$$

□

LEMMA 3.2. *For uncertain points P_1, \dots, P_n , any k centers c_1, \dots, c_k and any assignment A , we have*

$$Ecost_A(c_1, \dots, c_k) \geq \sum_{\hat{P}_1 \in D_1} prob(\hat{P}_1)d(\hat{P}_1, A(P_1)).$$

PROOF. Let $\Omega(\hat{P}_1)$ be those realizations that P_1 is realized as \hat{P}_1 . Then, $\sum_{R \in \Omega(\hat{P}_1)} prob(R) = prob(\hat{P}_1)$.

We have

$$\begin{aligned} Ecost_A(c_1, \dots, c_k) &= \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, A(P_i)) \\ &= \sum_{\hat{P}_1 \in D_1} \sum_{R \in \Omega(\hat{P}_1)} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, A(P_i)) \\ &\geq \sum_{\hat{P}_1 \in D_1} \sum_{R \in \Omega(\hat{P}_1)} prob(R)d(\hat{P}_1, A(P_1)) \\ &= \sum_{\hat{P}_1 \in D_1} prob(\hat{P}_1)d(\hat{P}_1, A(P_1)). \end{aligned}$$

□

Proof of Theorem 2.1

Let c^* be the optimal 1-center of P_1, \dots, P_n , we need to show that

$$Ecost(\bar{P}_1) \leq 2Ecost(c^*).$$

By the definition of $Ecost$,

$$Ecost(\bar{P}_1) = \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\bar{P}_1, \hat{P}_i).$$

By triangle inequality,

$$\begin{aligned} &\leq \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} (d(\bar{P}_1, c^*) + d(c^*, \hat{P}_i)) \\ &= d(\bar{P}_1, c^*) + \sum_R prob(R) \max_{i=1, \dots, n} d(c^*, \hat{P}_i). \end{aligned}$$

By Lemma 3.1 and definition of $Ecost$,

$$\leq \left(\sum_{\hat{P}_1 \in D_1} prob(\hat{P}_1)d(\hat{P}_1, c^*) \right) + Ecost(c^*).$$

By Lemma 3.2,

$$\leq 2Ecost(c^*).$$

Proof of Theorem 2.2

To prove Theorem 2.2, the following two lemmas are needed.

LEMMA 3.3. *For a set of uncertain points P_1, \dots, P_n in a Euclidean space, let c_1, \dots, c_k be any k centers and $A : \{P_1, \dots, P_n\} \rightarrow \{c_1, \dots, c_k\}$ be any assignment, we have*

$$\sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, \bar{P}_i) \leq 2Ecost_A(c_1, \dots, c_k),$$

Table 1: Our results for various versions of uncertain k -center

Objective	Metric	Running time	Assignment	Approx-Factor
1-center	Euclidean	$O(z)$	-	2
k -center	Euclidean	$O(nz + n \log k)$	restricted assigned version expected distance	6
k -center	Euclidean	-	restricted assigned version expected distance	$5 + \epsilon$
k -center	Euclidean	$O(nz + n \log k)$	restricted assigned version expected point	4
k -center	Euclidean	-	restricted assigned version expected point	$3 + \epsilon$
k -center	Euclidean	$O(nz + n \log k)$	unrestricted assigned version	4
k -center	Euclidean	-	unrestricted assigned version	$3 + \epsilon$
k -center	\mathbf{R}^1	$O(zn \log zn + n \log k \log n)$	unrestricted assigned version	3
k -center	any metric	-	unrestricted assigned version	$5 + \epsilon$

in particular for any $1 \leq i \leq n$,

$$\sum_{\hat{P}_i \in D_i} \text{prob}(\hat{P}_i) d(\hat{P}_i, \bar{P}_i) \leq 2 \text{Ecost}_A(c_1, \dots, c_k).$$

PROOF. We have

$$\begin{aligned} & \sum_{R \in \Omega} \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, \bar{P}_i) \\ & \leq \sum_{R \in \Omega} \text{prob}(R) \max_{i=1, \dots, n} (d(\hat{P}_i, A(P_i)) + d(A(P_i), \bar{P}_i)) \\ & \leq \sum_{R \in \Omega} \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, A(P_i)) + d(A(P_1), \bar{P}_1) \end{aligned}$$

where, we assume $d(A(P_1), \bar{P}_1) = \max_{i=1, \dots, n} d(A(P_i), \bar{P}_i)$, now the above term is

$$= \text{Ecost}_A(c_1, \dots, c_k) + d(A(P_1), \bar{P}_1).$$

It is enough to show $d(A(P_1), \bar{P}_1) \leq \text{Ecost}_A(c_1, \dots, c_k)$. But, according to Lemma 3.2, we have

$$d(A(P_1), \bar{P}_1) \leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(A(P_1), \hat{P}_1).$$

and from Lemma 3.2, it follows that $d(A(P_1), \bar{P}_1) \leq \text{Ecost}_A(c_1, \dots, c_k)$. \square

LEMMA 3.4. Let P_1, \dots, P_n be a set of uncertain points for any k centers c_1, \dots, c_k and assignment A one has

$$\text{cost}(c_1, \dots, c_k) \leq \text{Ecost}_A(c_1, \dots, c_k).$$

where cost is for the certain points $\bar{P}_1, \dots, \bar{P}_n$.

PROOF. One has

$$\text{cost}(c_1, \dots, c_k) = d(c_i, \bar{P}_j).$$

Since, c_i is the closest center to \bar{P}_j and by Lemma 3.1,

$$\leq d(A(P_j), \bar{P}_j) \leq \sum_{\hat{P}_j \in D_j} \text{prob}(\hat{P}_j) d(\hat{P}_j, A(P_j))$$

and by Lemma 3.2,

$$\leq \text{Ecost}_A(c_1, \dots, c_k).$$

\square

Now, we present the proof of Theorem 2.2 for the expected distance assignment. Let c_1^*, \dots, c_k^* be the optimal solution for restricted assigned version of k -center problem with the expected distance assignment. We need to show

$$\text{Ecost}_{ED}(c_1, \dots, c_k) \leq (5 + \epsilon) \text{Ecost}_{ED}(c_1^*, \dots, c_k^*).$$

By definition,

$$\text{Ecost}_{ED}(c_1, \dots, c_k) = \sum_R \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, ED(P_i)).$$

By triangle inequality,

$$\leq \sum_R \text{prob}(R) \max_{i=1, \dots, n} (d(\hat{P}_i, \bar{P}_i) + d(\bar{P}_i, ED(P_i))).$$

If we let $d(\bar{P}_1, ED(P_1)) = \max_{i=1, \dots, n} d(\bar{P}_i, ED(P_i))$ and use Lemma 3.3,

$$\leq 2 \text{Ecost}_{ED}(c_1^*, \dots, c_k^*) + d(\bar{P}_1, ED(P_1))$$

So, we need to show that

$$d(\bar{P}_1, ED(P_1)) \leq (3 + \epsilon) \text{Ecost}_{ED}(c_1^*, \dots, c_k^*).$$

Let c_i be the closest point among $\{c_1, \dots, c_k\}$ to \bar{P}_1 . Then, by Lemma 3.1,

$$d(\bar{P}_1, ED(P_1)) \leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, ED(P_1)).$$

Since, $ED(P_1)$ has the closest expected distance to P_1 among c_1, \dots, c_k ,

$$\leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, c_i).$$

By triangle inequality,

$$\leq (\sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, \bar{P}_1)) + d(\bar{P}_1, c_i).$$

By Lemma 3.3,

$$\sum_{\hat{P}_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, \bar{P}_1) \leq 2\text{Ecost}_{ED}(c_1^*, \dots, c_k^*).$$

So, it remains to show

$$d(\bar{P}_1, c_i) \leq (1 + \epsilon)\text{Ecost}_{ED}(c_1^*, \dots, c_k^*).$$

Since, c_i is the closest center to \bar{P}_1 we have

$$d(\bar{P}_1, c_i) \leq \text{cost}(c_1, \dots, c_k)$$

where cost is for the certain points $\bar{P}_1, \dots, \bar{P}_n$. Since, c_1, \dots, c_k is a $(1 + \epsilon)$ -approximation solution for the k -center problem,

$$\leq (1 + \epsilon)\text{cost}(c_1^*, \dots, c_k^*)$$

and by Lemma 3.4,

$$\leq (1 + \epsilon)\text{Ecost}_{ED}(c_1^*, \dots, c_k^*).$$

So, Theorem 2.2 for the expected distance assignment is proved.

Now, we give the proof of Theorem 2.2 for the the expected point assignment. Let c_1^*, \dots, c_k^* be the optimal solution for the restricted assigned k -center problem for the expected point assignment. We need to show

$$\text{Ecost}_{EP}(c_1, \dots, c_k) \leq (3 + \epsilon)\text{Ecost}_{EP}(c_1^*, \dots, c_k^*).$$

By definition,

$$\text{Ecost}_{EP}(c_1, \dots, c_k) = \sum_R \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, EP(P_i)).$$

By triangle inequality,

$$\leq \sum_R \text{prob}(R) \max_{i=1, \dots, n} (d(\hat{P}_i, \bar{P}_i) + d(\bar{P}_i, EP(P_i))).$$

If we let $d(\bar{P}_1, ED(P_1)) = \max_{i=1, \dots, n} d(\bar{P}_i, ED(P_i))$ and use Lemma 3.3,

$$\leq 2\text{Ecost}_{ED}(c_1^*, \dots, c_k^*) + d(\bar{P}_1, EP(P_1)).$$

So, we need to show that

$$d(\bar{P}_1, EP(P_1)) \leq (1 + \epsilon)\text{Ecost}_{ED}(c_1^*, \dots, c_k^*).$$

By definition of expected point assignment,

$$d(\bar{P}_1, EP(P_1)) = \text{cost}(c_1, \dots, c_k)$$

and since, c_1, \dots, c_k is a $(1 + \epsilon)$ -approximation solution,

$$\text{cost}(c_1, \dots, c_k) \leq (1 + \epsilon)\text{cost}(c_1^*, \dots, c_k^*)$$

and by Lemma 3.4,

$$\leq (1 + \epsilon)\text{Ecost}_{EP}(c_1^*, \dots, c_k^*).$$

So, Theorem 2.2 is completely proved.

REMARK 3.1. *There is a greedy 2-approximation algorithm for deterministic k -center problem of certain points $\bar{P}_1, \dots, \bar{P}_n$ in a metric space given in [13]. It is as follows. First, choose any point, say \bar{P}_1 and then choose the farthest point from \bar{P}_1 , say \bar{P}_2 and then, the farthest point from the set $\{\bar{P}_1, \bar{P}_2\}$, say \bar{P}_3 and continue until finding the farthest point from the set $\{\bar{P}_1, \dots, \bar{P}_{k-1}\}$, say \bar{P}_k . Then, the points $\bar{P}_1, \dots, \bar{P}_k$ is a 2-approximation solution for the deterministic k -center problem. If we use this method, in the first phase of the algorithm, we compute the expected point of each probabilistic point which takes $O(nz)$. Next, we compute $\bar{P}_1, \dots, \bar{P}_k$. The running time of this phase is $O(n \log k)$ [11]. So, the overall running time of algorithm is $O(nz + n \log k)$ and we get respectively a 6 and 4 approximation for the optimal expected cost of the k -center problem for the expected distance and expected point assignments.*

Proof of Theorem 2.3

Let c_1, \dots, c_k be the optimal solution for the restricted assigned k -center problem with expected distance assignment. Let c_1^*, \dots, c_k^* and assignment A be the optimal solution for the unrestricted assigned k -center prob-

lem. Then,

$$\begin{aligned}
Ecost_{ED}(c_1, \dots, c_k) &\leq Ecost_{ED}(c_1^*, \dots, c_k^*) \\
&= \sum_R prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, ED(P_i)) \\
&\leq \sum_R prob(R) \max_{i=1, \dots, n} (d(\hat{P}_i, A(P_i)) + d(A(P_i), ED(P_i))) \\
&\leq Ecost_A(c_1^*, \dots, c_k^*) + d(A(P_1), ED(P_1))
\end{aligned}$$

where $d(A(P_1), ED(P_1)) = \max_{i=1, \dots, n} d(A(P_i), ED(P_i))$.
By triangle inequality,

$$\begin{aligned}
&d(A(P_1), ED(P_1)) \\
&\leq \sum_{\hat{P}_1 \in D_1} prob(\hat{P}_1) (d(A(P_1), \hat{P}_1) + d(\hat{P}_1, ED(P_1)))
\end{aligned}$$

By Lemma 3.2 and the fact that $ED(P_1)$ has the smallest expected distance from P_1 among c_1^*, \dots, c_k^* , we get

$$\begin{aligned}
&\leq Ecost_A(c_1^*, \dots, c_k^*) + \sum_{\hat{P}_1 \in D_1} prob(\hat{P}_1) d(\hat{P}_1, A(P_1)) \\
&\leq 2Ecost_A(c_1^*, \dots, c_k^*).
\end{aligned}$$

So, Theorem 2.3 is proved.

Proof of Theorem 2.4

By definition,

$$Ecost_{ED}(c_1, \dots, c_k) = \sum_R prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, ED(P_i)).$$

By triangle inequality,

$$\begin{aligned}
&\leq \sum_R prob(R) \max_{i=1, \dots, n} (d(\hat{P}_i, A(P_i)) + d(A(P_i), ED(P_i))) \\
&\leq Ecost_A(a_1, \dots, a_k) + d(A(P_1), ED(P_1))
\end{aligned}$$

where $d(A(P_1), ED(P_1)) = \max_{i=1, \dots, n} d(A(P_i), ED(P_i))$.

Now by triangle inequality and Lemma 3.1,

$$\begin{aligned}
&d(A(P_1), ED(P_1)) \leq d(A(P_1), \bar{P}_1) + d(\bar{P}_1, ED(P_1)) \\
&\leq \sum_{\hat{P}_1} prob(\hat{P}_1) d(A(P_1), \hat{P}_1) + \sum_{\hat{P}_1} prob(\hat{P}_1) d(\hat{P}_1, ED(P_1)).
\end{aligned}$$

Let c_1 be a center among c_1, \dots, c_k that is closest to \bar{P}_1 . By Lemma 3.2 and the fact that $ED(P_1)$ has the closest expected distance to P_1 among the centers we get

$$\leq Ecost_A(c_1^*, \dots, c_k^*) + \sum_{\hat{P}_1} prob(\hat{P}_1) d(\hat{P}_1, c_1).$$

If instead of $d(\hat{P}_1, c_1)$, we put $d(\hat{P}_1, A(P_1)) + d(A(P_1), c_1)$ and use Lemma 3.2, we get

$$\leq 2Ecost_A(c_1^*, \dots, c_k^*) + d(A(P_1), c_1).$$

Now,

$$\begin{aligned}
&d(A(P_1), c_1) \leq d(A(P_1), \bar{P}_1) + d(\bar{P}_1, c_1) \\
&\leq \sum_{\hat{P}_1} prob(\hat{P}_1) d(\hat{P}_1, A(P_1)) + cost(c_1, \dots, c_k).
\end{aligned}$$

By Lemma 3.2 and the fact that c_1, \dots, c_k is a $(1 + \epsilon)$ -approximation solution for the k -center problem,

$$\leq Ecost_A(c_1^*, \dots, c_k^*) + (1 + \epsilon)cost(c_1^*, \dots, c_k^*).$$

Finally, by Lemma 3.4,

$$\leq (2 + \epsilon)Ecost_A(c_1^*, \dots, c_k^*).$$

This proves Theorem 2.4.

Proof of Theorem 2.5

By definition,

$$\begin{aligned}
&Ecost_{EP}(c_1, \dots, c_k) \\
&= \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} d(\hat{P}_i, EP(P_i)) \\
&\leq \sum_{R \in \Omega} prob(R) \max_{i=1, \dots, n} (d(\hat{P}_i, \bar{P}_i) + d(\bar{P}_i, EP(P_i))).
\end{aligned}$$

If we let $d(\bar{P}_1, EP(P_1)) = \max_{i=1, \dots, n} d(\bar{P}_i, EP(P_i))$ and use Lemma 3.3 we get

$$\leq 2Ecost_A(c_1^*, \dots, c_k^*) + d(\bar{P}_1, EP(P_1)).$$

Now, by Lemma 3.1,

$$d(\bar{P}_1, EP(P_1)) \leq \sum_{\hat{P}_1} prob(\hat{P}_1) d(\hat{P}_1, EP(P_1)).$$

Since,

$$\begin{aligned}
&d(\hat{P}_1, EP(P_1)) = cost(c_1, \dots, c_k) \\
&\leq (1 + \epsilon)cost(c_1^*, \dots, c_k^*),
\end{aligned}$$

also by Lemma 3.4,

$$\leq (1 + \epsilon)Ecost_A(c_1^*, \dots, c_k^*),$$

this proves Theorem 2.5.

Proofs of Theorem 2.6 and Theorem 2.7

To prove theorems 2.6 and 2.7, we need two lemmas that are analogue of Lemmas 3.3 and 3.4 for a metric space.

LEMMA 3.5. *Let P_1, \dots, P_n be a set of uncertain points in a metric space. Let \tilde{P}_i be the 1-center for the single uncertain point P_i . For any set of centers c_1, \dots, c_k and any assignment $A : \{P_1, \dots, P_n\} \rightarrow \{c_1, \dots, c_k\}$ we have*

$$\sum_R \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, \tilde{P}_i) \leq 3\text{Ecost}_A(c_1, \dots, c_k).$$

PROOF. Let $d(A(P_1), \tilde{P}_1) = \max_{i=1, \dots, n} d(A(P_i), \tilde{P}_i)$. If we use $d(\hat{P}_i, \tilde{P}_i) \leq d(\hat{P}_i, A(P_i)) + d(A(P_i), \tilde{P}_i)$, we get that the left hand side is

$$\leq \text{Ecost}_A(c_1, \dots, c_k) + d(A(P_1), \tilde{P}_1).$$

By triangle inequality,

$$\begin{aligned} & d(A(P_1), \tilde{P}_1) \\ & \leq \sum \text{prob}(\hat{P}_1) d(A(P_1), \hat{P}_1) + \sum \text{prob}(\hat{P}_1) d(\hat{P}_1, \tilde{P}_1). \end{aligned}$$

Since, \tilde{P}_1 is 1-center we get

$$\leq 2 \sum \text{prob}(\hat{P}_1) d(A(P_1), \hat{P}_1),$$

and by Lemma 3.2,

$$\leq 2\text{Ecost}_A(c_1, \dots, c_k).$$

This proves the lemma. \square

LEMMA 3.6. *Let P_1, \dots, P_n be a set of uncertain points in a metric space. For any k centers c_1, \dots, c_k and assignment A one has*

$$\text{cost}(c_1, \dots, c_k) \leq 2\text{Ecost}_A(c_1, \dots, c_k).$$

where cost is for the certain points $\tilde{P}_1, \dots, \tilde{P}_n$, where \tilde{P}_i is the 1-center of the uncertain point P_i .

PROOF. Let

$$\text{cost}(c_1, \dots, c_k) = d(c_i, \tilde{P}_j)$$

Then, since c_i is the closest center to \tilde{P}_j ,

$$\leq d(A(P_j), \tilde{P}_j)$$

by triangle inequality,

$$\leq \sum \text{prob}(\hat{P}_j) d(A(P_j), \hat{P}_j) + \sum \text{prob}(\hat{P}_j) d(\hat{P}_j, \tilde{P}_j).$$

since, \tilde{P}_j is 1-center of P_j ,

$$\begin{aligned} & \leq 2 \sum \text{prob}(\hat{P}_j) d(A(P_j), \hat{P}_j) \\ & \leq 2\text{Ecost}_A(c_1, \dots, c_k) \end{aligned}$$

So, the lemma is proved. \square

We now prove Theorem 2.6. By definition,

$$\begin{aligned} \text{Ecost}_{ED}(c_1, \dots, c_k) &= \sum_{R \in \Omega} \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, ED(P_i)) \\ &\leq \sum_{R \in \Omega} \text{prob}(R) \max_{i=1, \dots, n} d(\hat{P}_i, \tilde{P}_i) + d(\tilde{P}_1, ED(P_1)). \end{aligned}$$

Where $d(\tilde{P}_1, ED(P_1)) = \max_{i=1, \dots, k} d(\tilde{P}_i, ED(P_i))$. Since, by Lemma 3.5, the first term is at most $3\text{Ecost}_A(c_1^*, \dots, c_k^*)$, it is enough to show

$$d(\tilde{P}_1, ED(P_1)) \leq (4 + 2\epsilon)\text{Ecost}_A(c_1^*, \dots, c_k^*).$$

Now by triangle inequality and the fact that \tilde{P}_1 is 1-center of P_1 we get

$$\begin{aligned} & d(\tilde{P}_1, ED(P_1)) \\ & \leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) (d(\tilde{P}_1, \hat{P}_1) + d(\hat{P}_1, ED(P_1))) \\ & \leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) \left(d(A(P_1), \hat{P}_1) + d(\hat{P}_1, ED(P_1)) \right) \\ & \leq \text{Ecost}_A(c_1^*, \dots, c_k^*) + \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, ED(P_1)) \\ & \leq \text{Ecost}_A(c_1^*, \dots, c_k^*) + \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, c_j) \end{aligned}$$

where c_j is the closest among c_i 's to \tilde{P}_1 . Now,

$$\begin{aligned} & \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, c_j) \\ & \leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, \tilde{P}_1) + d(\tilde{P}_1, c_j) \\ & \leq \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1) d(\hat{P}_1, A(P_1)) + d(\tilde{P}_1, c_j) \\ & \leq \text{Ecost}_A(c_1^*, \dots, c_k^*) + d(\tilde{P}_1, c_j). \end{aligned}$$

Now, $d(\tilde{P}_1, c_j) \leq \text{cost}(c_1, \dots, c_k)$. Since, these centers are a $(1 + \epsilon)$ -approximation solution for the k -center problem,

$$\begin{aligned} & \text{cost}(c_1, \dots, c_k) \\ & \leq (1 + \epsilon)\text{cost}(c_1^*, \dots, c_k^*) \end{aligned}$$

by lemma 3.6,

$$\leq (2 + 2\epsilon)Ecost_A(c_1^*, \dots, c_k^*)$$

and this finishes the proof of Theorem 2.6.

Finally, we prove Theorem 2.7. By definition,

$$Ecost_{OC}(c_1, \dots, c_k) = \sum_R prob(R) \max_{i=1, \dots, n} d(P_i, OC(P_i)).$$

By triangle inequality,

$$\leq \sum_R prob(R) \max_{i=1, \dots, n} (d(P_i, \tilde{P}_i) + d(\tilde{P}_i, OC(P_i))).$$

By Lemma 3.5,

$$\leq 3Ecost_A(c_1^*, \dots, c_k^*) + d(\tilde{P}_1, OC(P_1))$$

where $d(\tilde{P}_1, OC(P_1)) = \max_{i=1, \dots, n} d(\tilde{P}_i, OC(P_i))$. Now,

$$d(\tilde{P}_1, OC(P_1)) = cost(c_1, \dots, c_k) \leq (1 + \epsilon)cost(c_1^*, \dots, c_k^*)$$

and by lemma 3.6,

$$\leq (2 + 2\epsilon)Ecost_A(c_1^*, \dots, c_k^*)$$

and this finishes the proof.

4. CONCLUSION

In this paper the k -center problem for uncertain data points have been studied. We have proposed new assignment schemes and obtained improved constant approximation factor algorithms for them. Note that, the new assignments introduced in this paper allowed us to improve the approximation factor for the unrestricted assigned version.

The restricted version with expected distance assignment for \mathbf{R}^1 was studied in [26]. Here we gave approximation algorithms for \mathbf{R}^d and also for any metric space.

The case of unrestricted assigned version which was studied in [14], has been improved. The constant of approximation has been reduced to $5 + \epsilon$ from $15 + \epsilon$. We have also separately studied the case for the metric space and the Euclidean space. In a future work, we intend to use our approach to study the k -median and the k -mean problems.

Also, we intend to give a PTAS for the assigned versions of the uncertain k -center problem.

Acknowledgment

The authors would like to thank Dr. Mohammad Ali Abam who introduced the problem to them and gave valuable comments and helpful suggestions.

5. REFERENCES

- [1] P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002.
- [2] C. C. Aggarwal and P. S. Yu. A survey of uncertain data algorithms and applications. *IEEE Trans. Knowl. Data Eng.*, 21(5):609–623, 2009.
- [3] 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.
- [4] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 250–257, 2002.
- [5] R. Chandrasekaran and A. Tamir. Polynomially bounded algorithms for locating p -centers on a tree. *Math. Program.*, 22(1):304–315, 1982.
- [6] R. Chandrasekaran and A. Tamir. Algebraic optimization: The fermat-weber location problem. *Math. Program.*, 46:219–224, 1990.
- [7] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008, June 9-11, 2008, Vancouver, BC, Canada*, pages 191–200, 2008.
- [8] Z. Drezner and H. W. Hamacher. *Facility location - applications and theory*. Springer, 2002.
- [9] M. E. Dyer. On a multidimensional search technique and its application to the euclidean one-centre problem. *SIAM J. Comput.*, 15(3):725–738, 1986.
- [10] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231, 1996.
- [11] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*, pages 434–444, 1988.
- [12] G. N. Frederickson. Parametric search and locating supply centers in trees. In *Algorithms and Data Structures, 2nd Workshop WADS '91, Ottawa, Canada, August 14-16, 1991, Proceedings*, pages 299–319, 1991.
- [13] T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.
- [14] S. Guha and K. Munagala. Exceeding expectations and clustering uncertain data. In *Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on*

Principles of Database Systems, PODS 2009, June 19 - July 1, 2009, Providence, Rhode Island, USA, pages 269–278, 2009.

Networks and Applications / Other Applications, December 12-14, 2008, Wuhan, China, pages 474–477, 2008.

- [15] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 291–300, 2004.
- [16] L. Huang and J. Li. Stochastic k -center and j -flat-center problems. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 110–129, 2017.
- [17] H. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*, pages 672–677, 2005.
- [18] H. Kriegel and M. Pfeifle. Hierarchical density-based clustering of uncertain data. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA*, pages 689–692, 2005.
- [19] P. Kumar and P. Kumar. Almost optimal solutions to k-clustering problems. *Int. J. Comput. Geometry Appl.*, 20(4):431–447, 2010.
- [20] D. T. Lee and Y. Wu. Complexity of some location problems. *Algorithmica*, 1(2):193–211, 1986.
- [21] N. Megiddo. Linear-time algorithms for linear programming in \mathbb{R}^3 and related problems. *SIAM J. Comput.*, 12(4):759–776, 1983.
- [22] N. Megiddo and K. J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984.
- [23] N. Megiddo and A. Tamir. New results on the complexity of p -center problems. *SIAM J. Comput.*, 12(4):751–758, 1983.
- [24] N. Megiddo, A. Tamir, E. Zemel, and R. Chandrasekaran. An $o(n \log^2 n)$ algorithm for the k -th longest path in a tree with applications to location problems. *SIAM J. Comput.*, 10(2):328–337, 1981.
- [25] A. Munteanu, C. Sohler, and D. Feldman. Smallest enclosing ball for probabilistic data. In *30th Annual Symposium on Computational Geometry, SOCG'14, Kyoto, Japan, June 08 - 11, 2014*, page 214, 2014.
- [26] H. Wang and J. Zhang. One-dimensional k -center on uncertain data. *Theor. Comput. Sci.*, 602:114–124, 2015.
- [27] H. Xu and G. Li. Density-based probabilistic clustering of uncertain data. In *International Conference on Computer Science and Software Engineering, CSSE 2008, Volume 4: Embedded Programming / Database Technology / Neural*