Improved constant approximation factor algorithms for $k$-center problem for uncertain data

Sharareh Alipour
School of computer science, Institute for research in fundamental sciences(IPM)
alipour@ipm.ir
https://orcid.org/0000-0002-3626-8960

Abstract

In real applications, database systems should be able to manage and process data with uncertainty. Any real dataset may have missing or rounded values, also the values of data may change by time. So, it becomes important to handle these uncertain data. An important problem in database technology is to cluster these uncertain data.

In this paper, we study the $k$-center problem for uncertain points in a general metric space. First we present a greedy approximation algorithm that builds $k$ centers using a farthest-first traversal in $k$ iterations. This algorithm improves the approximation factor of the unrestricted assigned $k$-center problem from 10 to 6. Next we restrict the centers to be selected from a finite set of points and we show that the optimal solution for this restricted setting is a 2-approximation factor solution for the optimal solution of the assigned $k$-center problem. Using this idea we improve the approximation factor of the unrestricted assigned $k$-center problem to 4 by increasing the running time mildly.

2012 ACM Subject Classification Theory of computation → Database theory, Theory of computation → Facility location and clustering

Keywords and phrases $k$-center, Uncertain points, Approximation algorithms, Greedy algorithms

1 Introduction

Uncertainty about the data appears in many real-world applications and an important issue for database systems is handling and correctly processing 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. So, an important challenge for database systems is to deal with large amount of data with uncertainty.

In this paper we study $k$-center problem for uncertain data. The $k$-center problem for certain points is defined as follows. Suppose we are given a set of $n$ clients that are located in a metric space. The goal is to build $k$ servers so that the maximum distance of each client to its closest server is minimized.

In real world, the location of the clients is not fixed, they can be at home, at work, or at some other locations and each of these possible locations have certain probability. This leads to a definition of uncertain version of the $k$-center problem, where the goal is to build $k$ servers, so that the expected travel cost is minimized.

This expected cost can be defined according to the practical needs in different ways. The easiest way, is to find the cost of each realization of our uncertain clients and then take the weighted average of these costs according to the probability of each realization. This is called the unassigned uncertain $k$-center problem. Another method is to assign to each client (regardless of its stochastic location) a fixed server according to a certain rule.
One reasonable rule (but not the only one) is to assign the server with minimum expected distance to that client. Now, for each realization we calculate the cost to be the maximum distance of clients with their assigned server and take a weighted average as before. This is the assigned uncertain \( k \)-center problem. The optimal solution that minimizes the cost with a given assignment or among all possible assignments, is the solution of the restricted or unrestricted assigned uncertain \( k \)-center problem, respectively.

The classical \( k \)-center problem, is shown to be NP-hard [23] and there are approximation algorithms for this problem for both certain and uncertain version. Therefore, our goal is to present fast and robust algorithms that improve the approximation factor of the previous algorithms.

**Problem Statement**

In the classical \( k \)-center problem we are given a set of (certain) points \( \{P_1, \ldots, P_n\} \) in a metric space \( X \) with metric \( d \). The \( k \)-center problem asks for \( k \) center points \( C = \{c_1, \ldots, c_k\} \) in \( X \) that minimize the following cost

\[
\text{cost}(c_1, \ldots, c_k) = \max_{i=1,\ldots,n} d(P_i, C),
\]

where \( d(P_i, C) = \min_{c \in C} d(P_i, c) \).

In the uncertain \( k \)-center problem each point has a finite number of possible locations independently from the other points with given probabilities. Precisely, we are given a set \( D = \{D_1, \ldots, D_n\} \) of \( n \) discrete and independent probability distributions. The i-th distribution, \( D_i \), is defined over a set of \( z_i \) possible locations \( P_{i_1}, \ldots, P_{i_{z_i}} \in X \). A probability \( p_{ij} \) is associated to each location such that \( \sum_{j=1}^{z_i} p_{ij} = 1 \) for every \( i \in [n] = \{1, \ldots, n\} \). Thus, the probabilistic points can be considered to be independent random variables \( X_i \). The locations together with the probabilities specify their distributions \( \text{Pr}[X_i = P_{ij}] = p_{ij} \) 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 \( z = \max\{z_1, \ldots, z_n\} \) be the maximum number of possibilities for uncertain points.

For simplicity, we use the notation \( \hat{P}_i \) for a realization of an uncertain point \( P_i \) and \( \text{prob}(\hat{P}_i) \) for its probability. We define \( \Omega \) as the probability space of all realizations \( R = \{P_{i_1}, \ldots, P_{i_{z_n}}\} \) with \( \text{prob}(R) = \prod_{i=1}^{n} \text{prob}(P_{ij}) \).

There are three known versions of the \( k \)-center problem for uncertain points based on the definition of the cost function. To motivate, imagine a city where its citizens are randomly in several locations (home, work, etc.) and our goal is to build \( k \) hospitals to make the expected travel cost of the citizens minimum. There are two plausible scenarios. First, any citizen can go to a closest hospital, depending on his location, this corresponds to the unassigned version of the \( k \)-center problem. Second, that a citizen based on his insurance policy or health issue has to go to an assigned hospital regardless of his location. This is corresponding to the assigned version of the \( k \)-center problem. If the assignment is an output of the optimization problem, then we have the unrestricted assigned version and if there is a assignment rule given as an input of the problem, we have the restricted assigned version. These are made precise below.

- **Unassigned version:**
  
  Here the goal is to find \( k \) centers \( C = \{c_1, \ldots, c_k\} \) that minimize
  
  \[
  E\text{cost}(c_1, c_2, \ldots, c_k) = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1,\ldots,n} d(\hat{P}_i, C).
  \]

- **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 $\{c_1, \ldots, c_k\}$ and an assignment $A : \{P_1, \ldots, P_n\} \rightarrow \{c_1, \ldots, c_k\}$ that minimize

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

**Restricted assigned version:**
Here for any set of uncertain points $\{P_1, \ldots, P_n\}$ and $k$ centers $\{c_1, \ldots, c_k\}$ an assignment $A : \{P_1, \ldots, P_n\} \rightarrow \{c_1, \ldots, c_k\}$ is given. The goal is to find $\{c_1, \ldots, c_k\}$ that minimizes the value of $Ecost_A(c_1, \ldots, c_k)$. In this paper, we consider the expected distance assignment that was first introduced in [27].

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

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

If several centers minimize the above expected distance, we arbitrarily assign one of them to $P_i$. This comment applies to other cases where we take $\arg\min$. There are other assignments such as expected point assignment and 1-center assignment that have been introduced in [3].

**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 [24] and approximation algorithms have been proposed (e.g., see [4, 5, 16]). Efficient algorithms were also given for some special cases, e.g., the smallest enclosing circle and its weighed version and discrete version [10, 21, 22], the Fermat-Weber problem [7], $k$-center on trees [6, 13, 24]. Refer to [9] for other variations of facility location problems. The deterministic $k$-center in one-dimensional space is solvable in $O(n \log n)$ time [25]. One of the most elegant approximation algorithms for $k$-center clustering is the 2-factor approximation algorithm by Gonzalez [14] which can be made to run in $O(n \log k)$ time [12]. One of the fastest methods for $k$-center clustering in 2 and 3 dimensions is by Agarwal and Procopiuc [11] which uses a dynamic programming approach to $k$-center clustering and whose running time is upper bounded by $O(n \log k) + (\frac{1}{k})O(k^{1.5})$. Another elegant solution to the $k$-center clustering problem was given by Badoiu et al. [13]. This algorithm gives a $(1 + \epsilon)$-approximation factor algorithm which runs in $2^{O((k \log k)/\epsilon^2)}dn$ in $\mathbb{R}^d$. Another algorithm based on coresets runs in $O(kn)$ [20] 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 DBSCAN [11] was also modified to handle probabilistic data by Kriegel and Pfeifle [18, 19] and Xu and Li [28]. Refer to [2] for a survey on data mining of uncertain data.

Cormode and McGregor [8] 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 Munagala [15] 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 $15(1 + 2\varepsilon)$ and the running time of their algorithm is polynomial in input size and $\frac{1}{\varepsilon}$.

Munteanu and et.al. presented the first polynomial time $(1 + \varepsilon)$-approximation algorithm for the probabilistic smallest enclosing ball problem with extensions to the streaming setting [20].

Wang and Zhang [27], introduced the restricted assigned 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.

Huang and Li [17] gave a PTAS for unassigned version of the probabilistic $k$-center problem in $\mathbb{R}^d$, when both $k$ and $d$ are constants. However for the assigned version, no such PTAS solutions are found.

Alipour and Jafari [3], introduced two other assignments for the restricted version of the problem. They presented a fast $10$-approximation factor algorithm for the unrestricted $k$-center problem. They also provide fast algorithms for the restricted $k$-center problem. The approximation factor of their algorithms for the expected distance assignment was $(5 + \varepsilon)$.

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

| Objective | Metric | Running time | Assignment | Approx-factor | Theorem |
|-----------|--------|--------------|------------|---------------|---------|
| $k$-center | General | $O(n^{\frac{3}{k}}z^3)$ | - | 2 | Corollary 9 |
| $k$-center | $\mathbb{R}^d$ | $O(nzk^2)$ | expected distance | 4 | Theorem 2 |
| $k$-center | General | $O(nzk^2 + nz^2)$ | unrestricted | 6 | Theorem 4 |
| $k$-center | General | $O(\left(\frac{n^2}{k}\right)n^3z^3)$ | unrestricted | 4 | Corollary 8 |

Our results

In this paper we have presented some algorithms for both unrestricted and restricted version of the problem. Our results are summarized in Table 1.

The main approaches of the paper are as follows. We present a greedy algorithm that is an extension of the well known greedy algorithm for the classical $k$-center problem presented in [14] and the algorithms presented in [3]. The greedy algorithm for the deterministic $k$-center problem simply builds the $k$ centers among the given points by choosing the point farthest away from the current set of centers in each iteration as the new center. We use the same idea for the uncertain $k$-center problem but the definition of the farthest point and the location of the centers are different. Then we analyze the approximation factor of this algorithm.

We also restrict the centers to be selected from a finite set of points and we show that the optimal solution in this restriction is a $2$-approximation factor solution for the assigned $k$-center problem. This enables us to improve the approximation factor for the restricted assigned $k$-center problem.
2. Expected distance assignment in $\mathbb{R}^d$

In this section we propose our algorithm for the expected distance assignment in $\mathbb{R}^d$. First we need some definitions. The expected point of each uncertain point $P_i$ is defined as

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

**Lemma 1.** For an uncertain point $P$ in a Euclidean space and any point $Q$, we have

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

For a given set $C$ of points, we define the distance of an uncertain point $P$ from $C$ as follows:

$$d(C, P) = \min_{c \in C} \sum_{\hat{P} \in D} \text{prob}(\hat{P}) d(\hat{P}, c).$$

Now we describe our algorithm.

- For each uncertain point $P_i$ compute its expected point, $\bar{P}_i$.
- Let $C = \{\bar{P}_1\}$.
- For each uncertain point $P_j$, compute $d(C, P_j)$.
- Pick the point $P_j$ with the greatest expected distance from $C$.
- Add $P_j$ to $C$ ($P_j$ may be added repeatedly). Continue this for $k$ iterations.
- Let $C = \{c_1, \ldots, c_k\}$.

**Theorem 2.** The $k$-centers in the above algorithm give us a 4-approximation solution for the expected distance assignment of the uncertain $k$-center problem.

**Proof.** Let $c_1^*, \ldots, c_k^*$ be the optimal solution. So we have $\text{OPT} = E\text{cost}_{ED}(c_1^*, \ldots, c_k^*) = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^n d(\hat{P}_i, c^*_P)$. Where $P_i$ is assigned to $c^*_P \in \{c_1^*, \ldots, c_k^*\}$ under the expected distance assignment.

There are two cases for the elements of $C = \{c_1, c_2, \ldots, c_k\}$.

- Case 1: suppose that for each $c_i^*$, there is exactly one uncertain point $P_j$ such that $\bar{P}_j = c_i \in C$ and $P_j$ is assigned to $c_i^*$ in the optimal solution. In this case we have

$$E\text{cost}_{ED}(c_1, \ldots, c_k) = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^n d(\bar{P}_i, c_P) \leq \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^n (d(\bar{P}_i, c^*_P) + d(c^*_P, c_P))$$

Suppose that $d(c^*_P, c_P) = \max_{i=1}^n d(c^*_P, c_P)$

$$\leq \text{OPT} + d(c^*_P, c_P)$$

So, it is enough to show that $d(c^*_P, c_P) \leq 3\text{OPT}$. We have

$$d(c^*_P, c_P) \leq \sum_{\hat{P}_i \in D_i} \text{prob}(\hat{P}_i) d(\hat{P}_i, c^*_P) \leq \text{OPT} + \sum_{\hat{P}_i \in D_i} \text{prob}(\hat{P}_i) d(\hat{P}_i, P_j)$$
According to our assumption there is a point $P_j$ that $\bar{P}_j \in C$ and $P_j$ is assigned to $c^*_P$. Note that since $P_1$ is assigned to $c_{P_1}$, the expected distance of $P_1$ from $c_{P_1}$ is less than the expected distance of $P_1$ from $\bar{P}_j$. So we have

$$\leq OPT + \sum_{\hat{P}_1 \in D_1} \text{prob}(\hat{P}_1)d(\hat{P}_1, c_{P_1}) + d(c^*_{P_1}, \bar{P}_j)$$

$$\leq OPT + OPT + d(c^*_{P_1}, \bar{P}_j)$$

$$\leq 2OPT + \sum_{\hat{P}_j \in D_j} \text{prob}(\hat{P}_j)d(\hat{P}_j, c_{P_1}) \leq 3OPT$$

Case 2: Suppose that there are two points $P_u$ and $P_v$ such that $\bar{P}_u$ and $\bar{P}_v$ are in $C$ and both $P_u$ and $P_v$ are assigned to the same center (for example $c^*_2$) in the optimal solution. Suppose that $\bar{P}_u$ is added to $C$ before $\bar{P}_v$ and $\bar{P}_v$ is added in the $i$th step. So, we have

$$Ecost_{ED}(c_1, \ldots, c_k) \leq Ecost_{ED}(c_1, \ldots, c_{i-1}) = \sum_{R \in \Omega} \text{prob}(R) \max_{j=1}^n d(\hat{P}_j, c_{P_j})$$

$$\leq \sum_{R \in \Omega} \text{prob}(R) \max_{j=1}^n (d(\hat{P}_j, c_{P_j}) + d(c^*_{P_j}, c_{P_j}))$$

Note that $c_{P_j} \in \{c_1, \ldots, c_{i-1}\}$. Suppose that $d(c^*_{P_j}, c_{P_j}) = \max_{i=1}^n d(c^*_{P_i}, c_{P_i})$, then we have

$$\leq OPT + d(c^*_{P_j}, c_{P_j})$$

$$\leq OPT + \sum_{\hat{P}_j \in D_j} \text{prob}(\hat{P}_j)d(\hat{P}_j, c^*_{P_j}) + d(\hat{P}_j, c_{P_j})$$

$$\leq OPT + OPT + \sum_{\hat{P}_j \in D_j} \text{prob}(\hat{P}_j)d(\hat{P}_j, c_{P_j})$$

Now we show that $\sum_{\hat{P}_j \in D_j} \text{prob}(\hat{P}_j)d(\hat{P}_j, c_{P_j}) \leq 2OPT$. According to the algorithm, $P_v$ is the farthest point to $c_1, \ldots, c_{i-1}$, so

$$\sum_{\hat{P}_j \in D_j} \text{prob}(\hat{P}_j)d(\hat{P}_j, c_{P_j}) \leq \sum_{\hat{P}_v \in D_v} \text{prob}(\hat{P}_v)d(\hat{P}_v, c_{P_j})$$

Also we know that the expected distance of $P_v$ from $\bar{P}_u$ is greater than the expected distance of $P_v$ from $c_{P_v}$ which means

$$\leq \sum_{\hat{P}_v \in D_v} \text{prob}(\hat{P}_v)d(\hat{P}_v, \bar{P}_u)$$

$$\leq \sum_{\hat{P}_v \in D_v} \text{prob}(\hat{P}_v)d(\hat{P}_v, c_{P_v}) + d(c^*_{P_v}, \bar{P}_u)$$

$$\leq OPT + d(c^*_{P_v}, \bar{P}_u)$$

$$\leq OPT + \sum_{\hat{P}_u \in D_u} \text{prob}(\hat{P}_u)d(\hat{P}_u, c_{P_v})$$
Since $P_u$ and $P_v$ are assigned to the same center we have
\[ \leq 2OPT \]

First, we compute the expected point of each uncertain point which takes $O(z)$ for each uncertain point and $O(nz)$ for all of them. In each iteration, we compute the expected distance of each point from $C$. There are at most $k$ centers and computing the expected distance of each uncertain point from each center takes $O(z)$ time. Since there are $k$ centers and $n$ uncertain points, each iteration takes $O(nzk)$ time. So, the overall running time is $O(nzk^2)$.

### 3 Unrestricted $k$-center in general metric

In this section, we propose our algorithm for unrestricted $k$-center problem in general metric space.

**Theorem 3.** Let $D$ be the set of possible locations of an uncertain point $P$ and let
\[ O' = \arg \min_{Q \in D} \sum_{\hat{P} \in D} \text{prob}(\hat{P})d(\hat{P}, Q) \]
then $O'$ is a 2-approximation for the 1-center of $P$, i.e. for any point $Q$,
\[ \sum_{\hat{P} \in D} \text{prob}(\hat{P})d(\hat{P}, O') \leq 2 \sum_{\hat{P} \in D} \text{prob}(\hat{P})d(\hat{P}, Q). \]

Note that if $P$ has $z$ locations, then we can compute its approximate 1-center in $O(z^2)$ time.

As in the previous section, we define the distance of an uncertain point $P_j$ from a point set $C$ as follows:
\[ d(C, P_j) = \min_{O_i' \in C} \sum_{\hat{P}_j \in D} \text{prob}(\hat{P}_j)d(\hat{P}_j, O_i'). \]

Now our algorithm
- For each uncertain point $P_i$, compute its approximated 1-center, $O_i'$.
- Let $C = \{O_i'\}$.
- For each uncertain point $P_j$, compute $d(C, P_j)$.
- Pick the point $P_j$ with the greatest distance from $C$.
- Add $O_j'$ to $C$ ($O_j'$ may be added repeatedly). Continue this till $k$ iterations.
Let $C = \{c_1, \ldots, c_k\}$.

**Theorem 4.** Suppose that we have a $(1 + \alpha)$-approximation algorithm for the 1-center of an uncertain point $P$. $Ecost_{E,D}(\{c_1, \ldots, c_k\})$ which means the cost is computed under the expected distance assignment gives us a $(5 + \alpha)$-approximation solution for the unrestricted $k$-center problem.

**Proof.** Let $c_1', \ldots, c_k'$ and assignment $A$ be the optimal solution for the unrestricted $k$-center problem. So, we have $OPT = Ecost_A(c_1', \ldots, c_k') = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^n d(\hat{P}_i, c_{p_i})$. Where $P_i$ is assigned to $c_{p_i} \in \{c_1', \ldots, c_k'\}$ in assignment $A$.

There are two cases for the elements of $C = \{c_0, c_1, \ldots, c_n\}$.  

\[ CVIT \; 2016 \]
Improved constant approximation factor algorithms for $k$-center problem for uncertain data

Case 1: Suppose that for each $c^*_i$, there is exactly one uncertain point $P_j$ such that $O'_j = c_i \in C$ and $P_j$ is assigned to $c^*_i$ in the optimal solution. In this case we have

$$E_{cost}(c_1, \ldots, c_k) = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^{n} d(\hat{P}_i, c^*_i)$$

Without loss of generality suppose that

$$d(c^*_p, c_p) = \max_{i=1}^{n} d(c^*_p, c_p).$$

$$\leq OPT + d(c^*_p, c_p)$$

$$\leq OPT + \sum_{\hat{P}_i \in D_1} \text{prob}(\hat{P}_i)(d(\hat{P}_i, c_p) + d(\hat{P}_i, c^*_p))$$

$$\leq OPT + OPT + \sum_{\hat{P}_i \in D_1} \text{prob}(\hat{P}_i)(d(\hat{P}_i, c_p))$$

The expected distance of $P_i$ from $c_p$ is less than the expected distance of $P_i$ from any point $c \in C$. According to our assumption there is a point $P_j$ such that $O'_{p_j} \in C$ and $P_j$ is assigned to $c_p$ in the optimal solution, so we have

$$\sum_{\hat{P}_i \in D_1} \text{prob}(\hat{P}_i)d(\hat{P}_i, c_p) \leq \sum_{\hat{P}_i \in D_1} \text{prob}(\hat{P}_i)d(\hat{P}_i, O'_{p_j})$$

$$\leq \sum_{\hat{P}_i \in D_1} \text{prob}(\hat{P}_i)d(\hat{P}_i, c^*_p) + d(c^*_p, O'_{p_j})$$

$$\leq OPT + \sum_{\hat{P}_i \in D_1} \text{prob}(\hat{P}_i)d(\hat{P}_i, c^*_p) + d(\hat{P}_i, O'_{p_j})$$

$$\leq OPT + OPT + (1 + \alpha)OPT$$

So, we have

$$\sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^{n} d(\hat{P}_i, c_p) \leq (5 + \alpha)OPT$$

Case 2: Suppose that there are two points $P_u$ and $P_v$ such that $O'_u$ and $O'_v$ are in $C$ and both $P_u$ and $P_v$ are assigned to the same center (for example $c^*_2$) in the optimal solution. Suppose that $O'_u$ is added to $C$ before $O'_v$ and $O'_v$ is added in the $i$th step. So, we have

$$E_{cost}(c_1, \ldots, c_k) \leq E_{cost}(c_1, \ldots, c_{i-1}) = \sum_{R \in \Omega} \text{prob}(R) \max_{j=1}^{n} d(\hat{P}_j, c_p)$$

Note that $c_{P_j} \in \{c_1, \ldots, c_{i-1}\}$ and $c_{P_u} = O'_u$.

$$\leq \sum_{R \in \Omega} \text{prob}(R) \max_{j=1}^{n} d(\hat{P}_j, c^*_p) + d(c^*_p, c_p)$$
Without loss of generality let \( d(c^*_P, c_{P1}) = \max_{j=1}^{n} d(c^*_P, c_{Pj}). \)
\[
\leq OPT + d(c^*_P, c_{P1}) \\
\leq OPT + \sum_{\hat{p}_1 \in D_1} \text{prob}(\hat{p}_1)(d(\hat{p}_1, c^*_P) + d(\hat{p}_1, c_{P1})) \\
\leq 2OPT + \sum_{\hat{p}_1 \in D_1} \text{prob}(\hat{p}_1)d(\hat{p}_1, c_{P1})
\]
We know that \( \sum_{\hat{p}_v \in D_v} \text{prob}(\hat{p}_v)d(\hat{p}_v, c_{Pv}) \) has the maximum distance from \( \{c_1, \ldots, c_{i-1}\} \), so
\[
\leq 2OPT + \sum_{\hat{p}_v \in D_v} \text{prob}(\hat{p}_v)d(\hat{p}_v, c_{Pv}) \\
\leq 2OPT + \sum_{\hat{p}_v \in D_v} \text{prob}(\hat{p}_v)d(\hat{p}_v, c^*_P) + d(c^*_P, c_{Pv}) \\
\leq 2OPT + OPT + d(c^*_P, c_{Pv}) \\
\leq 3OPT + \sum_{\hat{P}_u \in D_u} \text{prob}((\hat{P}_u)d(c^*_P, \hat{P}_u) + d(\hat{P}_u, c_{Pv})
\]
Since both \( P_u \) and \( P_v \) are assigned to the same center and \( c_{Pv} \) is a \((1 + \alpha)\)-approximation solution for the 1-center of \( P_u \) we have
\[
\leq 4OPT + (1 + \alpha)OPT = (5 + \alpha)OPT
\]

The running time for computing a 2-approximation solution for the 1-center of each uncertain point is \( O(z^2) \) [3]. In each iteration we compute the expected distance of each uncertain point from the centers which takes \( O(kz) \) for each uncertain point. Since there are \( n \) uncertain points each iteration takes \( O(nzk) \) and for \( k \) iterations the running time is \( O(nzk^2) \). So the overall running time is \( O(nzk^2) + O(nz^2) \).

4 Another approach for \( k \)-center problem

In this section we present a different approach for the problem. Suppose that we restrict the centers to be selected from the expected points of the uncertain points in \( \mathbb{R}^d \) or the 1-center of the uncertain points in a general metric or the possible locations of the uncertain points in a general metric. We denote this version of the problem by restricted centers. We have the following theorem.

\textbf{Theorem 5.} If the centers are restricted to be selected from the expected points of the uncertain points, then the optimal solution in this setting is a 2-approximation factor of the original unrestricted assigned \( k \)-center problem in \( \mathbb{R}^d \).

\textbf{Proof.} Let \( \{c'_1, \ldots, c'_k\} \subseteq \{\hat{P}_1, \ldots, \hat{P}_k\} \) and assignment \( A' \) be the optimal solution in the unrestricted \( k \)-center problem when the centers are restricted to be selected from the expected points of the centers. Suppose that \( \{c^*_1, \ldots, c^*_k\} \subseteq \mathbb{R}^d \) and assignment \( A \) are the optimal
Improved constant approximation factor algorithms for $k$-center problem for uncertain data

solution for the unrestricted $k$-center problem. Note that $P_i$ is assigned to $c'_{P_i}$ in the optimal solution of the restricted centers and $c^*_{P_i}$ in the original unrestricted $k$-center problem. For each point $c_i^*$ let $f(c_i^*) \in \{\hat{P}_1, \ldots, \hat{P}_n\}$ be the closest point to $c_i^*$. We have

$$Ecost_{A'}(c'_1, \ldots, c'_k) = \sum_{R \in \Omega} \frac{n}{\text{prob}(R)} \max_{i=1}^{\text{n}} d(\hat{P}_i, c'_{P_i})$$

Since $A'$ and $c'_1, c'_2, \ldots, c'_k$ are the optimal solution for the case where centers are selected from the expected points of $P_i$'s, if we assign each $P_i$ to another point rather than $c'_{P_i}$ then the expected cost will be greater than $Ecost_{A'}(c'_1, \ldots, c'_k)$, so

$$\leq \sum_{R \in \Omega} \frac{n}{\text{prob}(R)} \max_{i=1}^{n} d(\hat{P}_i, f(c^*_{P_i}))$$

$$\leq \sum_{R \in \Omega} \frac{n}{\text{prob}(R)} \max_{i=1}^{n} (d(\hat{P}_i, c^*_{P_i}) + d(c'_{P_i}, f(c^*_{P_i})))$$

Assume that $d(c'_{P_i}, f(c^*_{P_i})) = \max_{i=1}^{n} d(c'_{P_i}, f(c^*_{P_i}))$.

$$\leq OPT + d(c'_{P_i}, f(c^*_{P_i}))$$

Since $f(c^*_{P_i}) \in \{\hat{P}_1, \ldots, \hat{P}_n\}$ is the closest point to $c^*_{P_i}$, then $d(c'_{P_i}, f(c^*_{P_i})) \leq d(c'_{P_i}, \hat{P}_1)$ so,

$$\leq OPT + d(c^*_{P_i}, \hat{P}_1)$$

$$\leq OPT + \sum_{P_i \in D_i} \text{prob}(\hat{P}_i) d(\hat{P}_i, c^*_{P_i})$$

$$\leq 2OPT$$

Theorem 6. If the centers are restricted to be selected from the possible locations of $P_i$'s, $P_{i,j}$'s, then the optimal solution for unrestricted assigned $k$-center problem in this setting is a 2-approximation factor of the original unrestricted assigned $k$-center problem in a general metric space.

Since the proof is similar to the proof of Theorem we omit the proof.

Theorem 7. If the centers are restricted to be selected from the possible locations of $P_i$'s, $P_{i,j}$'s, then the optimal solution for expected distance assignment in this setting is a 4-approximation factor of the original unrestricted assigned $k$-center problem in a general metric space.

Proof. Suppose that $c'_1, c'_2, \ldots, c'_k \subseteq \{P_{1,1}, P_{1,2}, \ldots, P_{n,z}\}$ be the optimal solution for expected point assignment when the centers are restricted to be selected from $P_{i,j}$'s. Let $c^*_1, c^*_2, \ldots, c^*_k$ and assignment $A$ be the optimal solution for the unrestricted assigned $k$-center problem. For each point $c_i^*$, let $c''_i \in \{P_{1,1}, P_{1,2}, \ldots, P_{n,z}\}$ be the closest point to $c_i^*$. Since
\( c^1, c^2, \ldots, c^k \subseteq \{ P_{1,1}, P_{1,2}, \ldots, P_{n,z} \} \) is the optimal solution for expected point assignment when the centers are restricted to be selected from \( P_{i,j} \)'s then we have

\[
E_{\text{cost}}(c^1, \ldots, c^k) \leq E_{\text{cost}}(c''^1, \ldots, c''^k) = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^{n} d(\hat{P}_i, c''^i) \\
\sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^{n} (d(\hat{P}_i, c''^i) + d(c''^i, c''^i))
\]

Suppose that \( \max_{i=1}^{n} d(c''^i, c''^i) = d(c''^i, c''^i) \]

\[
\leq \text{OPT} + d(c''^i, c''^i) \\
\leq \text{OPT} + \sum_{\hat{P}_i \in \mathcal{D}_1} \text{prob}(\hat{P}_i) d(\hat{P}_i, c''^i) + d(\hat{P}_i, c''^i)
\]

Suppose that the closest point to \( c''^i \) is \( c''^j \). Since each point is assigned to the center that minimizes its expected distance from its center, so the expected distance of \( P_1 \) from any center \( c''^j \), is greater than the expected distance of \( P_1 \) from \( c''^j \).

\[
\leq \text{OPT} + \text{OPT} + \sum_{\hat{P}_i \in \mathcal{D}_1} \text{prob}(\hat{P}_i) d(\hat{P}_i, c''^j) \\
\leq 2\text{OPT} + \sum_{\hat{P}_i \in \mathcal{D}_1} \text{prob}(\hat{P}_i) d(\hat{P}_i, c''^j) + d(c''^j, c''^j) \\
\leq 3\text{OPT} + \sum_{\hat{P}_i \in \mathcal{D}_1} \text{prob}(\hat{P}_i) d(c''^j, c''^j)
\]

Since \( c''^j \) is the closest point to \( c''^i \) so for any \( \hat{P}_i \in \mathcal{D}_1 \) we have \( d(c''^i, c''^j) \leq d(\hat{P}_i, c''^i) \).

\[
\leq 3\text{OPT} + \sum_{\hat{P}_i \in \mathcal{D}_1} \text{prob}(\hat{P}_i) d(\hat{P}_i, c''^j) \\
\leq 4\text{OPT}
\]

Now, we show that for a given assignment \( A \) and a given set of centers \( \{ c_1, \ldots, c_k \} \) how to compute \( E_{\text{cost}}(A, c_1, \ldots, c_n) \). Let \( \text{prob}(P_{i,j} = \text{max}) \) be the probability that \( P_{i,j} \) realizes the \( \max_{i=1}^{n} d(\hat{P}_i, c_{P_i}) \) that is

\[
\text{prob}(P_{i,j} = \text{max}) = \text{prob}(\hat{P}_i = P_{i,j} \cap \bigcap_{k=1, k \neq i}^{n} \text{prob}(d(P_k, c_{P_i}))
\].

For each \( P_k \) we compute the probability that \( d(\hat{P}_k, c_{P_k}) < d(P_{i,j}, c_{P_i}) \), which takes \( O(z) \) for each \( P_k \). Then we have

\[
\text{prob}(P_{i,j} = \text{max}) = \text{prob}(P_i = P_{i,j} \cap \bigcap_{k=1, k \neq i}^{n} \text{prob}(d(P_k, c_{P_i}) < d(P_{i,j}, c_{P_i})
\].

So, the overall running time for each \( P_{i,j} \) is \( O(nz) \), thus in \( O(n^2z^2) \) time we can compute \( \text{prob}(P_{i,j} = \text{max}) \).
Improved constant approximation factor algorithms for $k$-center problem for uncertain data

We know that

$$E_{\text{cost}}(c_1, \ldots, c_k) = \sum_{R \in \Omega} \text{prob}(R) \max_{i=1}^{n} d(\hat{P}_i, c_{P_i})$$

which is equal to the following

$$\sum_{i=1}^{n} \sum_{j=1}^{z} \text{prob}(P_{i,j} = \max) d(P_{i,j}, c_{P_i}).$$

For each $P_{i,j}$ we compute $\text{prob}(P_{i,j} = \max)$ in $O(n^2 z^2)$ so $E_{\text{cost}}(c_1, \ldots, c_k)$ can be computed in $O(n^3 z^3)$.

Now we give a trivial 4-approximation solution for the unrestricted assigned $k$-center problem. Our algorithm is as follows. For each $k$-subset $c_1, c_2, \ldots, c_k$ of the $P_{i,j}$’s compute $E_{\text{cost}}(c_1, \ldots, c_k)$ and choose $\min\{c_1, \ldots, c_k\} \subseteq P_{i,j}$ since there are $\binom{n^2}{k}$ subset and we can compute $E_{\text{cost}}(c_1, \ldots, c_k)$ in $O(n^3 z^3)$, a 4-approximation solution can be computed in $O(\binom{n^2}{k} n^3 z^3)$.

- **Corollary 8.** In a general metric space, we can compute a 4-approximation solution for unrestricted assigned $k$-center problem in $O(\binom{n^2}{k} n^3 z^3)$.

By Theorem [5], we conclude the following.

- **Corollary 9.** In a general metric space, we can compute a 2-approximation solution for the probabilistic 1-center problem in $O(n^4 z^4)$.

- **Remark.** Note that we can not use Theorem [5] for the unrestricted assigned $k$-center problem, because for each $k$-subset $c_1, c_2, \ldots, c_k$ we have to give the optimal assignment $A$ to compute $E_{\text{cost}}(c_1, \ldots, c_k)$, and we do not know the optimal assignment even when the centers are fixed unless we consider all the assignments (since each point can be assigned to any of the $k$ centers, for $c_1, c_2, \ldots, c_k$ we have $k^n$ assignments) that takes many time. But by Theorem [7], since the assignment is defined to be the expected distance assignment we can compute a 4-approximation solution.

### 5 conclusion

In this paper we have studied the $k$-center problem for uncertain data points. In the first part we have given improved constant approximation factor algorithms for both restricted and unrestricted assigned version of the problem.

Then we show that if we restrict the centers to be chosen from some certain points then the optimal solution with this restriction gives a 2-approximation solution for the general problem. This leads us to give improved approximation factor algorithms for the unrestricted assigned $k$-center problem.

### References

1. Pankaj K. Agarwal and Cecilia Magdalena Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002. URL: [http://dx.doi.org/10.1007/s00453-001-0110-y](http://dx.doi.org/10.1007/s00453-001-0110-y)

2. Charu C. Aggarwal and Philip S. Yu. A survey of uncertain data algorithms and applications. *IEEE Trans. Knowl. Data Eng.*, 21(5):609–623, 2009. URL: [http://dx.doi.org/10.1109/TKDE.2008.190](http://dx.doi.org/10.1109/TKDE.2008.190)
Sharareh Alipour and Amir Jafari. Improvements on the k-center problem for uncertain data. In Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018, pages 425–433, 2018. URL: http://doi.acm.org/10.1145/3196959.3196969

Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. SIAM J. Comput., 33(3):544–562, 2004. URL: http://dx.doi.org/10.1137/S0097539702416402

Mihai Badoiu, Sariel Har-Peled, and Piotr 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. URL: http://doi.acm.org/10.1145/509907.509947

R. Chandrasekaran and Arie Tamir. Polynomially bounded algorithms for locating p-centers on a tree. Math. Program., 22(1):304–315, 1982. URL: http://dx.doi.org/10.1007/BF01581045

R. Chandrasekaran and Arie Tamir. Algebraic optimization: The fermat-weber location problem. Math. Program., 46:219–224, 1990. URL: http://dx.doi.org/10.1007/BF01585739

Graham Cormode and Andrew 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. URL: http://doi.acm.org/10.1145/1376916.1376944

Zvi Drezner and Horst W. Hamacher. Facility location - applications and theory. Springer, 2002. URL: http://www.springer.com/computer/swe/book/978-3-540-42172-6

Martin E. Dyer. On a multidimensional search technique and its application to the euclidean one-centre problem. SIAM J. Comput., 15(3):725–738, 1986. URL: http://dx.doi.org/10.1137/0215052

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei 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. URL: http://www.aaai.org/Library/KDD/1996/kdd96-037.php

Tomás Feder and Daniel 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. URL: http://doi.acm.org/10.1145/622126.62255 doi:10.1145/622126.62255

Greg 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. URL: http://dx.doi.org/10.1007/BFb0028271 doi:10.1007/BFb0028271

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. Theor. Comput. Sci., 38:293–306, 1985. URL: http://dx.doi.org/10.1016/0304-3975(85)90224-5 doi:10.1016/0304-3975(85)90224-5

Sudipto Guha and Kamesh 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. URL: http://doi.acm.org/10.1145/1559795.1559836 doi:10.1145/1559795.1559836
Improved constant approximation factor algorithms for k-center problem for uncertain data

16 Sariel Har-Peled and Soham 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 15-16, 2004, pages 291–300, 2004. URL: [http://doi.acm.org/10.1145/1007352.1007400](http://doi.acm.org/10.1145/1007352.1007400)

17 Lingxiao Huang and Jian 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. URL: [http://dx.doi.org/10.1137/1.9781611974782.8](http://dx.doi.org/10.1137/1.9781611974782.8)

18 Hans-Peter Kriegel and Martin 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. URL: [http://doi.acm.org/10.1145/1081870.1081955](http://doi.acm.org/10.1145/1081870.1081955)

19 Hans-Peter Kriegel and Martin 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. URL: [http://dx.doi.org/10.1109/ICDM.2005.75](http://dx.doi.org/10.1109/ICDM.2005.75)

20 Pankaj Kumar and Piyush Kumar. Almost optimal solutions to k-clustering problems. *Int. J. Comput. Geometry Appl.*, 20(4):431–447, 2010. URL: [http://dx.doi.org/10.1142/S0218195910003372](http://dx.doi.org/10.1142/S0218195910003372)

21 D. T. Lee and Ying-Fung Wu. Complexity of some laction problems. *Algorithmica*, 1(2):193–211, 1986. URL: [http://dx.doi.org/10.1007/BF01840442](http://dx.doi.org/10.1007/BF01840442)

22 Nimrod Megiddo. Linear-time algorithms for linear programming in r^3 and related problems. *SIAM J. Comput.*, 12(4):759–776, 1983. URL: [http://dx.doi.org/10.1137/0213014](http://dx.doi.org/10.1137/0213014)

23 Nimrod Megiddo and Arie Tamir. New results on the complexity of p-center problems. *SIAM J. Comput.*, 12(4):751–758, 1983. URL: [http://dx.doi.org/10.1137/0212051](http://dx.doi.org/10.1137/0212051)

24 Nimrod Megiddo and Arie Tamir. New results on the complexity of p-center problems. *SIAM J. Comput.*, 12(4):751–758, 1983. URL: [http://dx.doi.org/10.1137/0212051](http://dx.doi.org/10.1137/0212051)

25 Nimrod Megiddo, Arie Tamir, Eitan 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. URL: [http://dx.doi.org/10.1137/0210023](http://dx.doi.org/10.1137/0210023)

26 Alexander Munteanu, Christian Sohler, and Dan 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. URL: [http://doi.acm.org/10.1145/2582112.2582114](http://doi.acm.org/10.1145/2582112.2582114)

27 Haitao Wang and Jingru Zhang. One-dimensional k-center on uncertain data. *Theor. Comput. Sci.*, 602:114–124, 2015. URL: [http://dx.doi.org/10.1016/j.tcs.2015.08.017](http://dx.doi.org/10.1016/j.tcs.2015.08.017)

28 Huajie Xu and Guohui 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 Networks and Applications / Other Applications, December 12-14, 2008, Wuhu, China, pages 474–477, 2008. URL: [http://dx.doi.org/10.1109/CSSE.2008.968](http://dx.doi.org/10.1109/CSSE.2008.968)