mathematics

Article

Stochastic Approximate Algorithms for Uncertain Constrained *K*-Means Problem

Jianguang Lu [1,2], Juan Tang [3,4,*], Bin Xing [2,5] and Xianghong Tang [1]

1 State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China; jglu@gzu.edu.cn (J.L.); xhtang@gzu.edu.cn (X.T.)
2 Chongqing Innovation Center of Industrial Big-Data Co., Ltd., Chongqing 400707, China; bingxcq@outlook.com
3 School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China
4 Institute of Computing Science and Technology, Guangzhou University, Guangzhou 510006, China
5 National Engineering Laboratory for Industrial Big-Data Application Technology, Chongqing 400707, China
* Correspondence: tangjn16@gzhu.edu.cn

Abstract: The k-means problem has been paid much attention for many applications. In this paper, we define the uncertain constrained k-means problem and propose a $(1+\epsilon)$-approximate algorithm for the problem. First, a general mathematical model of the uncertain constrained k-means problem is proposed. Second, the random sampling properties of the uncertain constrained k-means problem are studied. This paper mainly studies the gap between the center of random sampling and the real center, which should be controlled within a given range with a large probability, so as to obtain the important sampling properties to solve this kind of problem. Finally, using mathematical induction, we assume that the first $j-1$ cluster centers are obtained, so we only need to solve the j-th center. The algorithm has the elapsed time $O((\frac{1891ek}{\epsilon^2})^{8k/\epsilon}nd)$, and outputs a collection of size $O((\frac{1891ek}{\epsilon^2})^{8k/\epsilon}n)$ of candidate sets including approximation centers.

Keywords: stochastic approximate algorithms; uncertain constrained k-means; approximation centers

check for updates

Citation: Lu, J.; Tang, J.; Xing, B.; Tang, X. Stochastic Approximate Algorithms for Uncertain Constrained *k*-Means Problem. *Mathematics* **2022**, *10*, 144. https://doi.org/10.3390/math10010144

Academic Editors: Ximeng Liu, Yinbin Miao and Zuobin Ying

Received: 25 November 2021
Accepted: 27 December 2021
Published: 4 January 2022

1. Introduction

The k-means problem has received much attention in the past several decades. The k-means problems consists of partitioning a set P of points in d-dimensional space $\mathbb{R}^d$ into k subsets $P_1, \ldots, P_k$ such that $\sum_{i=1}^{k}\sum_{p\in P_i}||p-c_i||^2$ is minimized, where c_i is the center of P_i, and $||p-q||$ is the distance between two points of p and q. The k-means problem is one of the classical NP-hard problems, and has been paid much attention in the literature [1–3].

For many applications, each cluster of the point set may satisfy some additional constraints, such as chromatic clustering [4], r-capacity clustering [5], r-gather clustering [6], fault tolerant clustering [7], uncertain data clustering [8], semi-supervised clustering [9], and l-diversity clustering [10]. The constrained clustering problems was studied by Ding and Xu, who presented the first unified framework in [11]. Given a point set $P \subseteq \mathbb{R}^d$, and a positive integer k, a list of constraints $\mathbb{L}$, the constrained k-means problem is to partition P into k clusters $\mathbb{P} = \{P_1, \ldots, P_k\}$, such that all constraints in $\mathbb{L}$ are satisfied and $\sum_{P_i\in\mathbb{P}}\sum_{x\in P_i}||x-c(P_i)||^2$ is minimized, where $c(P_i) = \frac{1}{|P_i|}\sum_{x\in P_i} x$ denotes the centroid of P_i.

In recent years, particular research has been focused on the constrained k-means problem. Ding and Xu [11] showed the first polynomial time approximation scheme with running time $O(2^{poly(k/\epsilon)}(\log n)^k nd)$ for the constrained k-means problem, and obtained a collection of size $O(2^{poly(k/\epsilon)}(\log n)^{k+1})$ of candidate approximate centers. The existing fastest approximation schemes for the constrained k-means problem takes $O(2^{O(k/\epsilon)}nd)$ time [12,13], which was first shown by Bhattacharya, Jaiswai, and Kumar [12]. Their algorithm gives a collection of size $O(2^{O(k/\epsilon)})$ of candidate approximate centers. In this paper, we propose the uncertain constrained k-means problem, which supposes that all points are random variables with probabilistic distributions. We present a stochastic approximate algorithm for the uncertain constrained k-means problem. The uncertain constrained k-means problem can be regarded as a generalization of the constrained k-means problem. We prove the random sampling properties of the uncertain constrained k-means problem, which are fundamental for our proposed algorithm. By applying random sampling and mathematical induction, we propose a stochastic approximate algorithm with lower complexity for the uncertain constrained k-means problem.

This paper is organized as follows. Some basic notations are given in Section 2. Section 3 provides an overview of the new algorithm for the uncertain constrained k-means problem. In Section 4, we discuss the detailed algorithm for the uncertain constrained k-means problem. In Section 5, we investigate the correctness, success probability, and running time analysis of the algorithm. Section 6 concludes this paper and gives possible directions for future research.

2. Preliminaries

Definition 1 (Uncertain constrained k-means problem)**.** *Given a random variable set $\mathcal{X} \subseteq \mathbb{R}^d$, the probability density function $f_X(s)$ for every random variable $X \in \mathcal{X}$, a list of constraints $\mathbb{L}$, and a positive integer k, the uncertain constrained k-means problem is to partition $\mathcal{X}$ into k clusters $\mathbb{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$, such that all constraints in $\mathbb{L}$ are satisfied and $\sum_{\mathcal{X}_i \in \mathbb{X}} \sum_{X \in \mathcal{X}_i} \int^{\mathbb{R}^d} ||s - c(\mathcal{X}_i)||^2 f_X(s)ds$ is minimized, where $c(\mathcal{X}_i) = \frac{1}{|\mathcal{X}_i|} \sum_{X \in \mathcal{X}_i} \int^{\mathbb{R}^d} s f_X(s)ds$ denotes the centroid of $\mathcal{X}_i$.*

Definition 2 ([13])**.** *Let $\mathcal{X}$ be a set of random variables in $\mathbb{R}^d$, $f_X(s)$ be probability density function for every random variable $X \in \mathcal{X}$, and $q \in \mathbb{R}^d$ and P be a set of points in $\mathbb{R}^d$, $p \in P$.*

- *Define $f_2(q, \mathcal{X}) = \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||s - q||^2 f_X(s)ds$.*
- *Define $c(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} s f_X(s)ds$.*
- *Define $dist(X, P) = min_{p \in P} \int^{\mathbb{R}^d} ||s - p|| f_X(s)ds$.*

Definition 3 ([13])**.** *Let $\mathcal{X}$ be a set of random variables in $\mathbb{R}^d$, $f_X(s)$ be the probability density function for every random variable $X \in \mathcal{X}$, and $\mathcal{X}_1, \ldots, \mathcal{X}_k$ be a partition of $\mathcal{X}$.*

- *Define $m_j = c(\mathcal{X}_j)$.*
- *$\beta_j = \frac{|\mathcal{X}_j|}{|\mathcal{X}|}$.*
- *Define $\sigma_j = \sqrt{\frac{f_2(m_j, \mathcal{X}_j)}{|\mathcal{X}_j|}}$.*
- *Define*
  $OPT_k(\mathcal{X}) = \sum_{j=1}^{k} \sum_{X \in \mathcal{X}_j} \int^{\mathbb{R}^d} ||s - c(\mathcal{X}_j)||^2 f_X(s)ds = \sum_{j=1}^{k} f_2(m_j, \mathcal{X}_j)$.
- *Define $\sigma_{opt} = \sqrt{\frac{OPT_k(\mathcal{X})}{|\mathcal{X}|}} = \sqrt{\sum_{i=1}^{k} \beta_i \sigma_i^2}$.*

Lemma 1. *For any point $x \in \mathbb{R}^d$ and a random variable set $\mathcal{X} \subseteq \mathbb{R}^d$, $f_2(x, \mathcal{X}) = f_2(c(\mathcal{X}), \mathcal{X}) + |\mathcal{X}|||c(\mathcal{X}) - x||^2$.*

Proof. Let $f_X(s)$ be the probability density function for every random variable $X \in \mathcal{X}$.

$$f_2(x, \mathcal{X}) = \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||s - x||^2 f_X(s) ds \tag{1}$$

$$= \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||s - c(\mathcal{X}) + c(\mathcal{X}) - x||^2 f_X(s) ds \tag{2}$$

$$= \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||s - c(\mathcal{X})||^2 f_X(s) ds + \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||c(\mathcal{X}) - x||^2 f_X(s) ds \tag{3}$$

$$= f_2(c(\mathcal{X}), \mathcal{X}) + ||c(\mathcal{X}) - x||^2 \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} f_X(s) ds \tag{4}$$

$$= f_2(c(\mathcal{X}), \mathcal{X}) + |\mathcal{X}| ||c(\mathcal{X}) - x||^2. \tag{5}$$

The (3) equality follows from the fact that $\sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} (s - c(\mathcal{X})) f_X(s) ds = 0$. □

Lemma 2. *Let $\mathcal{X}$ be a set of random variables in $\mathbb{R}^d$ and $f_X(s)$ be the probability density function for every random variable $X \in \mathcal{X}$. Assume that $\mathcal{T}$ is a set of random variables obtained by sampling random variables from $\mathcal{X}$ uniformly and independently. For $\forall\ \delta > 0$, we have:*

$$Pr(||c(\mathcal{T}) - c(\mathcal{X})||^2 > \frac{1}{\delta |\mathcal{T}|} \sigma^2) < \delta, \tag{6}$$

where $\sigma^2 = \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||s - c(\mathcal{X})||^2 f_X(s) ds$.

Proof. First, observe that

$$E(c(\mathcal{T})) = c(\mathcal{X}),\ \ E(||c(\mathcal{T}) - c(\mathcal{X})||^2) = \frac{1}{|\mathcal{T}|} \sigma^2 \tag{7}$$

where $\sigma^2 = \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} \int^{\mathbb{R}^d} ||s - c(\mathcal{X})||^2 f_X(s) ds$. Then apply the Markov inequality to obtain the following.

$$Pr(||c(\mathcal{T}) - c(\mathcal{X})||^2 > \frac{1}{\delta |\mathcal{T}|} \sigma^2) < \delta. \tag{8}$$

□

Lemma 3. *Let $\mathcal{Q}$ be a set of random variables in $\mathbb{R}^d$, $f_X(s)$ be the probability density function for every random variable $X \in \mathcal{Q}$, and $\mathcal{Q}_1$ be an arbitrary subset of $\mathcal{Q}$ with $\alpha |\mathcal{Q}|$ random variables for some $0 < \alpha \le 1$. Then $||c(\mathcal{Q}) - c(\mathcal{Q}_1)|| \le \sqrt{\frac{1-\alpha}{\alpha}} \sigma$, where $\sigma^2 = \frac{1}{|\mathcal{Q}|} \sum_{X \in \mathcal{Q}} \int^{\mathbb{R}^d} ||s - c(\mathcal{Q})||^2 f_X(s) ds$.*

Proof. Let $\mathcal{Q}_2 = \mathcal{Q} \setminus \mathcal{Q}_1$. By Lemma 1, we have the following two equalities.

$$f_2(c(\mathcal{Q}), \mathcal{Q}_1) = f_2(c(\mathcal{Q}_1), \mathcal{Q}_1) + |\mathcal{Q}_1| ||c(\mathcal{Q}_1) - c(\mathcal{Q}||^2, \tag{9}$$

$$f_2(c(\mathcal{Q}), \mathcal{Q}_2) = f_2(c(\mathcal{Q}_2), \mathcal{Q}_2) + |\mathcal{Q}_2| ||c(\mathcal{Q}_2) - c(\mathcal{Q}||^2. \tag{10}$$

Let $L = ||c(\mathcal{Q}_1) - c(\mathcal{Q}_2)||$. By the definition of the mean point, we have:

$$c(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{X \in \mathcal{Q}} \int^{\mathbb{R}^d} s f_X(s) ds = \frac{1}{|\mathcal{Q}|} (|\mathcal{Q}_1| c(\mathcal{Q}_1) + |\mathcal{Q}_2| c(\mathcal{Q}_2)). \tag{11}$$

Thus, the three points $\{c(\mathcal{Q}), c(\mathcal{Q}_1), c(\mathcal{Q}_2)\}$ are collinear, while $||c(\mathcal{Q}_1) - c(\mathcal{Q})|| = (1 - \alpha)L$ and $||c(\mathcal{Q}_2) - c(\mathcal{Q})|| = \alpha L$. Meanwhile, by the definition of σ, we have $\sigma^2 =$

$\frac{1}{|\mathcal{Q}|}(\sum_{X\in\mathcal{Q}_1}\int^{\mathbb{R}^d}||s-c(\mathcal{Q})||^2 f_X(s)ds+\sum_{X\in\mathcal{Q}_2}\int^{\mathbb{R}^d}||s-c(\mathcal{Q})||^2 f_X(s)ds)$. Combining Equality (9) and Equality (10), we have:

$$\sigma^2 \geq \frac{1}{|\mathcal{Q}|}(|\mathcal{Q}_1|||c(\mathcal{Q}_1)-c(\mathcal{Q}||^2+|\mathcal{Q}_2|||c(\mathcal{Q}_2)-c(\mathcal{Q}||^2) \quad (12)$$
$$=\alpha((1-\alpha)L)^2+(1-\alpha)(\alpha L)^2 \quad (13)$$
$$=\alpha(1-\alpha)L^2. \quad (14)$$

Thus, we have $L \leq \frac{\sigma}{\sqrt{\alpha(1-\alpha)}}$, which means that $||c(\mathcal{Q})-c(\mathcal{Q}_1)||=(1-\alpha)L\leq\sqrt{\frac{1-\alpha}{\alpha}}\sigma$. □

Lemma 4 ([12])**.** *For any $x,y,z\in\mathbb{R}^d$, then $||x-z||^2\leq 2||x-y||^2+2||y-z||^2$.*

Theorem 1 ([14])**.** *Let $X_1,\ldots,X_s$ be s, an independent random $0-1$ variable, where X_i takes 1 with a probability of at least p for $i=1,\ldots,s$. Let $X=\sum_{i=1}^{s}X_i$. Then, for any $\delta>0$, $Pr(X<(1-\delta)ps)<e^{-\frac{1}{2}\delta^2 ps}$.*

3. Overview of Our Method

In this section, we first introduce the main idea of our methodology to solve the uncertain constrained k-means problem.

Considering the optimal partition $\mathbb{X}=\{\mathcal{X}_1,\ldots,\mathcal{X}_k\}(|\mathcal{X}_1|\geq\ldots\geq|\mathcal{X}_k|)$ of $\mathcal{X}$, since $|\mathcal{X}_1|/|\mathcal{X}|\geq 1/k$, if we could sample a set $\mathcal{S}$ of size $O(k/\epsilon)$ from $\mathcal{X}$ uniformly and independently, then at least $O(1/\epsilon)$ random variables in $\mathcal{S}$ are from $\mathcal{X}_1$ with a certain probability. All subsets of $\mathcal{S}$ of size $O(1/\epsilon)$ could be enumerated to discover the approximate center of $\mathcal{X}_1$.

We assume that $C_{j-1}=\{c_1,\ldots,c_{j-1}\}$ is the set including approximate centers of the $\mathcal{X}_1,\ldots,\mathcal{X}_j$. Let $\mathcal{B}_j=\{X\in\mathcal{X}|dist(X,C_{j-1})=min_{c\in C_{j-1}}\int^{\mathbb{R}^d}||s-c||f_X(s)ds\leq r_j\}$, where $r_j=\sqrt{\frac{\epsilon}{40\beta_j k}}\sigma_{opt}$. The set $\mathcal{X}_j$ is divided into two parts: $\mathcal{X}_j^{out}$ and $\mathcal{X}_j^{in}$, where $\mathcal{X}_j^{out}=\mathcal{X}_j\setminus\mathcal{B}_j$ and $\mathcal{X}_j^{in}=\mathcal{X}_j\cap\mathcal{B}_j$. For each random variable X, let $\widetilde{X}$ be the nearest point (particular random variable) in C_{j-1} to X. Let $\widetilde{\mathcal{X}}_j^{in}=\{\widetilde{X}|X\in\mathcal{X}_j^{in}\}$, and $\widetilde{\mathcal{X}}_j=\widetilde{\mathcal{X}}_j^{in}\cup\mathcal{X}_j^{out}$.

If most of the random variables of $\mathcal{X}_j$ are in $\mathcal{X}_j^{in}$, our idea is to use the center of $\widetilde{\mathcal{X}}_j^{in}$ to approximate the center of $\mathcal{X}_j$. The center of $\widetilde{\mathcal{X}}_j^{in}$ is found based on C_{j-1}. If most of the random variables of $\mathcal{X}_j$ are in $\mathcal{X}_j^{out}$, our ideal is to replace the center of $\mathcal{X}_j$ with the center of $\widetilde{\mathcal{X}}_j$. For seeking out the approximate center of $\widetilde{\mathcal{X}}_j$, we should find out a subset $\mathcal{S}'$ by uniformly sampling from $\widetilde{\mathcal{X}}_j$. However, the set $\mathcal{X}_j^{out}$ is unknown. We need to find the set $\mathcal{S}'\cap\mathcal{X}_j^{out}$. We apply a branching strategy to find a set $\mathcal{Q}$ such that $\mathcal{X}\setminus\mathcal{B}_j\subseteq\mathcal{Q}$, and $|\mathcal{Q}|<2|\mathcal{X}\setminus\mathcal{B}_j|$. Then, a random variables set $\mathcal{S}$ is obtained by sampling random variables from $\mathcal{Q}$ independently and uniformly. And the set $\mathcal{X}\setminus\mathcal{B}_j\subseteq\mathcal{Q}$ can be replaced by a subset $\mathcal{S}^*$ of $\mathcal{S}$ from $\mathcal{X}_j^{out}$. Based on $\mathcal{S}^*$ and $\widetilde{\mathcal{X}}_j^{in}$, the approximation center of $\widetilde{\mathcal{X}}_j$ could be obtained. Therefore, the algorithm presented in this paper outputs a collection of size $O((\frac{1891ek}{\epsilon^2})^{8k/\epsilon}n)$ of candidate sets containing approximation centers, and has the running time $O((\frac{1891ek}{\epsilon^2})^{8k/\epsilon}nd)$.

4. Our Algorithm cMeans

Given an instance $(\mathcal{X},k,\mathbb{L})$ of the uncertain constrained k-means problem, $\mathbb{X}=\{\mathcal{X}_1,\ldots,\mathcal{X}_k\}$ denotes an optimal partition of $(\mathcal{X},k,\mathbb{L})$. There exist six parameters (ϵ, $\mathcal{Q}$, g, k, C, U) in our **cMeans**, where $\epsilon\in(0,1]$ is the approximate factor, $\mathcal{Q}$ is the input random variable set, g is the number of centers, k is the number of the clusters, C is the set of approximate cluster centers, and U is a collection of candidate sets including the approximate center. Let $M=\frac{6}{\epsilon}$, $N=\frac{79{,}380k}{\epsilon^3}$, where M is the size of subsets of the sampling set and N is the size of the sampling set. Without loss of generality, assume that values of M and N are integers.

We use the branching strategy to seek out the approximate centers of clusters in $\mathbb{X}$. There exist two branches in our algorithm **cMeans**, which can be seen in Figure 1. On one branch, a size N set $\mathcal{S}_1$ is obtained by sampling from $\mathcal{Q}$ uniformly and independently; $\mathcal{S}_2$ is constructed by $\mathcal{S}_1$ and M copies of each point in C. Moreover, we consider each subset $\mathcal{S}'$ of size M of $\mathcal{S}_2$, and the centroid c of $\mathcal{S}'$ is solved to represent the approximate center of $\mathcal{X}_{k-g+1}$, and our algorithm **cMeans**$(\epsilon, \mathcal{Q}, g-1, k, C \cup \{c\}, U)$ is used to obtain the remaining $g-1$ cluster centers.

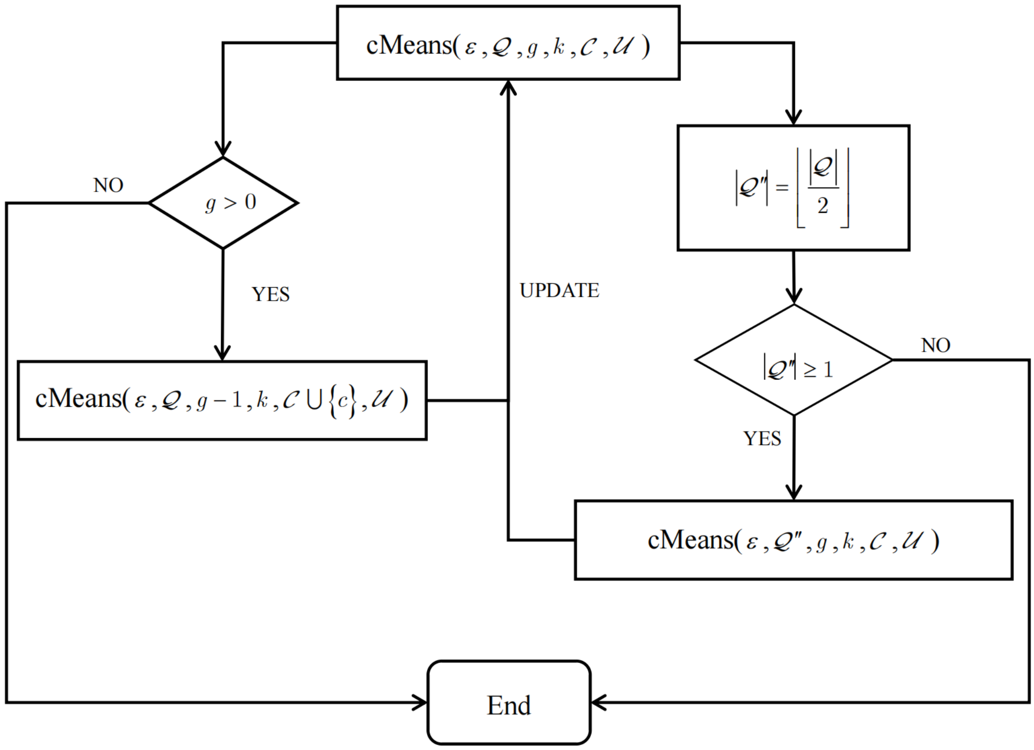

Figure 1. Flow chart of our algorithm **cMeans**.

On the other branch, for each random variable $X \in \mathcal{Q}$, we calculate the distance between X and C first. H denotes the set of all distances of random variables in $\mathcal{X}$ to C, where H is a multi-set. We should obtain the median value m for all values in H, which is the $\lfloor |H|/2 \rfloor$-th element if all of the values in H are sorted. In the second branch, $\mathcal{Q}$ is divided into two parts, $\mathcal{Q}'$ and $\mathcal{Q}''$, based on m such that for $\forall X' \in \mathcal{Q}'$, $X'' \in \mathcal{Q}''$, $dist(X', C) \leq dist(X'', C)$, where $|\mathcal{Q}'| = \lceil \frac{|\mathcal{Q}|}{2} \rceil$, $|\mathcal{Q}''| = \lfloor \frac{|\mathcal{Q}|}{2} \rfloor$. Subroutine **cMeans**$(\epsilon, \mathcal{Q}'', g, k, C, U)$ is used to obtain the remaining g cluster centers. Therefore, we present the specific algorithm for seeking out a collection of candidate sets in the Algorithm 1.

Algorithm 1: cMeans$(\epsilon, \mathcal{Q}, g, k, C, U)$

Input: $(\epsilon, \mathcal{Q}, g, k, C, U)$
Output: a collection of candidate sets
$M = \frac{6}{\epsilon}, N = \frac{79380k}{\epsilon^3}, \mathcal{S}_1 = \mathcal{S}_2 = H = \emptyset$;
if $g = 0$ **then**
  add C to the collection U;
end
sample a set $\mathcal{S}_1$ of size N from $\mathcal{Q}$ independently and uniformly;
if $C = \emptyset$ **then**
  $\mathcal{S}_2 = \mathcal{S}_1$;
end
else
  $\mathcal{S}_2 = \mathcal{S}_1 \cup \{M$ copies of each point in $C\}$;
end
for *each subset* $\mathcal{S}'$ *of size* M *of* $\mathcal{S}_2$ **do**
  compute the centroid c of $\mathcal{S}'$;
  cMeans$(\epsilon, \mathcal{Q}, g-1, k, C \cup \{c\}, U)$;
end
for *each random variable* $X \in \mathcal{Q}$ **do**
  compute $dist(X, C)$, and add $dist(X, C)$ to H;
  obtain the median value m of all values in H, which is the $\lfloor \frac{|H|}{2} \rfloor$-th element if all the values in H are sorted;
  divide $\mathcal{Q}$ into $\mathcal{Q}'$ and $\mathcal{Q}''$ by m such that for $\forall X' \in \mathcal{Q}'$, $X'' \in \mathcal{Q}''$, $dist(X', C) \leq dist(X'', C)$, where $|\mathcal{Q}'| = \lceil \frac{|\mathcal{Q}|}{2} \rceil$,$|\mathcal{Q}''| = \lfloor \frac{|\mathcal{Q}|}{2} \rfloor$;
  if $|\mathcal{Q}''| \geq 1$ **then**
    cMeans$(\epsilon, \mathcal{Q}'', g, k, C, U)$;
  end
end

5. Analysis of Our Algorithm cMeans

We investigate the success probability, correctness, and time complexity analysis of the algorithm **cMeans** in this section.

Lemma 5. *There exists a candidate set, with a probability of at least* $1/12^k$*, including the approximate center* $C_k = \{c_1, \ldots, c_k\}$ *in* U *satisfying* $||m_j - c_j||^2 \leq \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2 (1 \leq j \leq k)$.

The following Lemmas from Lemma 6 to 16 are used to prove Lemma 5. We prove Lemma 5 via induction on j. For $j = 1$, we can obtain $\beta_1 \geq 1/k$ easily, and prove the success probability first.

Lemma 6. *In the process of finding* c_1 *in our algorithm* ***cMeans****, by sampling a set of* $79{,}380k/\epsilon^3$ *random variables from* $\mathcal{X}$ *independently and uniformly, denoted by* $\mathcal{S}_1$*, the probability that at least* $6/\epsilon$ *random variables in* $\mathcal{S}_2$ *are from* $\mathcal{X}_1$ *is at least* $1/2$.

Proof. In our algorithm **cMeans**, we assume that $\mathcal{S}_1 = S_1, \ldots, S_N$, where $N = 79{,}380k/\epsilon^3$. Let $x_1^{'}, \ldots, x_N^{'}$ be the corresponding random variables of elements in $\mathcal{S}_1$. If $S_i \in \mathcal{X}_1$, then

$x_i' = 1$. Otherwise $x_i' = 0$. It is known easily that $Pr[S_i \in \mathcal{X}_1] \geq \frac{1}{k}$. Let $x = \sum_{i=1}^{N} x_i'$, $u = \sum_{i=1}^{N} E(x_i')$. We obtain that $u \geq 79{,}380k/\epsilon^3$. Then,

$$Pr[x > \frac{6}{\epsilon}] = 1 - Pr[x \leq \frac{6}{\epsilon}] \tag{15}$$

$$= 1 - Pr[x \leq \frac{6\epsilon^2}{79{,}380} \frac{79{,}380}{\epsilon^3}] \tag{16}$$

$$\geq 1 - Pr[x \leq \frac{\epsilon^2}{13{,}230} u] \tag{17}$$

$$\geq 1 - e^{-\frac{(1-\frac{\epsilon^2}{13{,}230})^2 u}{2}} \tag{18}$$

$$\geq 1 - e^{-\frac{(1-\frac{\epsilon^2}{13{,}230})^2 \frac{79{,}380}{\epsilon^3}}{2}} \tag{19}$$

$$\geq 1 - e^{-\frac{(1-\frac{1}{13{,}230})^2 \cdot 79{,}380}{2}} \tag{20}$$

$$\geq \frac{1}{2}. \tag{21}$$

□

From Lemma 6, an $\mathcal{S}^*$ with size $6/\epsilon$ of $\mathcal{S}_2$ can be obtained, and the probability that all points in $\mathcal{S}^*$ are from $\mathcal{X}_1$ is at least $1/2$. Let c_1 denote the centroid of $\mathcal{S}^*$, and $\delta = 5/6$. For $|\mathcal{S}^*| = 6/\epsilon$, by Lemma 2, we conclude that $||m_1 - c_1||^2 \leq \frac{1}{5}\epsilon\sigma_1^2$ holds with a probability of at least $1/6$. Then, the probability that a subset $\mathcal{S}^*$ of size $6/\epsilon$ of $\mathcal{S}_2$ can be found such that $||m_1 - c_1||^2 \leq \frac{1}{5}\epsilon\sigma_1^2 \leq \frac{9}{10}\epsilon\sigma_1^2 + \frac{1}{10\beta_1 k}\epsilon\sigma_{opt}^2$ holds is at least $1/12$. Therefore, we conclude that Lemma 5 holds for $j = 1$.

Moreover, we assume that for $j \leq j_0 (1 \leq j_0)$, Lemma 5 holds with a probability of at least $1/12^j$. Considering the case $j = j_0 + 1$, we prove Lemma 5 by the following two cases: (1)$|\mathcal{X}_j^{out}| \leq \frac{\epsilon}{49}\beta_j n$; (2)$|\mathcal{X}_j^{out}| > \frac{\epsilon}{49}\beta_j n$.

5.1. Analysis for Case 1: $|\mathcal{X}_j^{out}| \leq \frac{\epsilon}{49}\beta_j n$

Since $|\mathcal{X}_j^{out}| \leq \frac{\epsilon}{49}\beta_j n$, most of the random variables of $\mathcal{X}_j$ are in $\mathcal{B}_j$. Our idea is to replace the center of $\mathcal{X}_j$ with the center of $\widetilde{\mathcal{X}}_j^{in}$. Thus, we need to find the approximate center c_j of $\widetilde{\mathcal{X}}_j^{in}$ and the bound distance $||m_j - c_j||$. We divide the distance $||m_j - c_j||$ into the following three parts: $||m_j - m_j^{in}||$, $||m_j^{in} - \widetilde{m}_j^{in}||$, and $||\widetilde{m}_j^{in} - c_j||$. We first study the distance between m_j and m_j^{in}.

Lemma 7. $||m_j - m_j^{in}|| \leq \sqrt{\frac{\epsilon}{48}}\sigma_j$.

Proof. Since $|\mathcal{X}_j| = \beta_j n$ and $|\mathcal{X}_j^{out}| \leq \frac{\epsilon}{49}\beta_j n$, the proportion of $\mathcal{X}_j^{in}$ in $\mathcal{X}_j$ is at least $1 - \frac{\epsilon}{49}$. By Lemma 3, $||m_j - m_j^{in}|| \leq \sqrt{\frac{\epsilon/49}{1-\epsilon/49}}\sigma_j \leq \sqrt{\frac{\epsilon}{48}}\sigma_j$. □

Lemma 8. $||m_j^{in} - \widetilde{m}_j^{in}|| \leq r_j$.

Proof. Since $m_j^{in} = \frac{1}{|\mathcal{X}_j^{in}|}\sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d} s f_X(s)ds$, and $\widetilde{m}_j^{in} = \frac{1}{|\mathcal{X}_j^{in}|}\sum_{X\in\mathcal{X}_j^{in}}\widetilde{X}$, we can obtain the following:

$$||m_j^{in} - \widetilde{m}_j^{in}|| = ||\frac{1}{|\mathcal{X}_j^{in}|}\sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d} s f_X(s)ds - \frac{1}{|\mathcal{X}_j^{in}|}\sum_{X\in\mathcal{X}_j^{in}}\widetilde{X}|| \tag{22}$$

$$= \frac{1}{|\mathcal{X}_j^{in}|}||\sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d}(s-\widetilde{X})f_X(s)ds|| \tag{23}$$

$$\leq \frac{1}{|\mathcal{X}_j^{in}|}\sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d}||s-\widetilde{X}||f_X(s)ds \tag{24}$$

$$\leq \frac{1}{|\mathcal{X}_j^{in}|}\sum_{X\in\mathcal{X}_j^{in}} r_j \tag{25}$$

$$= r_j. \tag{26}$$

□

Lemma 9. $f_2(\widetilde{m}_j^{in}, \widetilde{\mathcal{X}}_j^{in}) \leq 2|\mathcal{X}_j^{in}|r_j^2 + 2f_2(m_j, \mathcal{X}_j^{in}) - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2$.

Proof. Since $|\widetilde{\mathcal{X}}_j^{in}| = |\mathcal{X}_j^{in}|$, by 1, we have $f_2(m_j, \widetilde{\mathcal{X}}_j^{in}) = f_2(\widetilde{m}_j^{in}, \widetilde{\mathcal{X}}_j^{in}) + |\mathcal{X}_j^{in}|||\widetilde{m}_j^{in} - m_j||$. Then,

$$f_2(\widetilde{m}_j^{in}, \widetilde{\mathcal{X}}_j^{in}) = f_2(m_j, \widetilde{\mathcal{X}}_j^{in}) - |\mathcal{X}_j^{in}|||\widetilde{m}_j^{in} - m_j||^2 \tag{27}$$

$$= \sum_{X\in\mathcal{X}_j^{in}}||\widetilde{X} - m_j||^2 - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2 \tag{28}$$

$$= \sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d}||\widetilde{X} - m_j||^2 f_X(s)ds - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2 \tag{29}$$

$$= \sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d}||\widetilde{X} - s + s - m_j||^2 f_X(s)ds - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2 \tag{30}$$

$$\leq \sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d}(2||\widetilde{X} - s||^2 + 2||s - m_j||^2)f_X(s)ds - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2 \tag{31}$$

$$\leq 2|\mathcal{X}_j^{in}|r_j^2 + 2\sum_{X\in\mathcal{X}_j^{in}}\int^{\mathbb{R}^d}||s - m_j||^2 f_X(s)ds - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2 \tag{32}$$

$$= 2|\mathcal{X}_j^{in}|r_j^2 + 2f_2(m_j, \mathcal{X}_j^{in}) - |\mathcal{X}_j^{in}|||m_j - \widetilde{m}_j^{in}||^2 \tag{33}$$

□

Lemma 10. *In the process of finding c_j in our algorithm **cMeans**, for the set $\mathcal{S}_2$ in step 5, a subset $\mathcal{S}^*$ of size $6/\epsilon$ of $\mathcal{S}_2$ can be obtained such that all random variables in $\mathcal{S}^*$ are from $\widetilde{\mathcal{X}}_j^{in}$. Let c_j be the centroid of $\mathcal{S}^*$. Then, the inequality $||\widetilde{m}_j^{in} - c_j||^2 \leq \frac{2}{5}\epsilon r_j^2 + \frac{49}{120}\epsilon\sigma_j^2 - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2$ holds with a probability of at least 1/6.*

Proof. For each point $p \in C_{j-1}$, $6/\epsilon$ copies of p are added to $\mathcal{S}_2$ in step 9 in our algorithm **cMeans**. Thus, a subset $\mathcal{S}^*$ of size $6/\epsilon$ of $\mathcal{S}_2$ can be obtained such that all random variables in $\mathcal{S}^*$ are from $\widetilde{\mathcal{X}}_j^{in}$. Let $\delta = 5/6$. Since $|\mathcal{S}^*| = 6/\epsilon$, by Lemma 2, $||\widetilde{m}_j^{in} - c_j||^2 \leq \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j^{in}, \widetilde{\mathcal{X}}_j^{in})}{|\mathcal{X}_j^{in}|}$ holds with a probability of at least 1/6. Assume that $||\widetilde{m}_j^{in} - c_j||^2 \leq \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j^{in}, \widetilde{\mathcal{X}}_j^{in})}{|\mathcal{X}_j^{in}|}$. Then,

$$
\begin{aligned}
||\widetilde{m}_j^{in} - c_j||^2 &\leq \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j^{in}, \widetilde{\mathcal{X}}_j^{in})}{|\mathcal{X}_j^{in}|} && (34)\\
&\leq \frac{1}{5}\epsilon\frac{2|\mathcal{X}_j^{in}|r_j^2 + 2f_2(m_j, \mathcal{X}_j^{in}) - |\mathcal{X}_j^{in}|\,||m_j - \widetilde{m}_j^{in}||^2}{|\mathcal{X}_j^{in}|} && (35)\\
&= \frac{2}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\frac{f_2(m_j, \mathcal{X}_j^{in})}{|\mathcal{X}_j^{in}|} - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2 && (36)\\
&\leq \frac{2}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\frac{f_2(m_j, \mathcal{X}_j)}{|\mathcal{X}_j| - |\mathcal{X}_j^{out}|} - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2 && (37)\\
&\leq \frac{2}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\frac{\beta_j n\sigma_j^2}{(1-\epsilon/49)\beta_j n} - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2 && (38)\\
&\leq \frac{2}{5}\epsilon r_j^2 + \frac{49}{120}\epsilon\sigma_j^2 - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2. && (39)
\end{aligned}
$$

□

Lemma 11. *If c_j satisfies $||\widetilde{m}_j^{in} - c_j||^2 \leq \frac{2}{5}\epsilon r_j^2 + \frac{49}{120}\epsilon\sigma_j^2 - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2$, then $||m_j - c_j||^2 \leq \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2$.*

Proof. Assume that c_j satisfies $||\widetilde{m}_j^{in} - c_j||^2 \leq \frac{2}{5}\epsilon r_j^2 + \frac{49}{120}\epsilon\sigma_j^2 - \frac{1}{5}\epsilon||m_j - \widetilde{m}_j^{in}||^2$. Then,

$$
\begin{aligned}
||m_j - c_j||^2 &= ||m_j - \widetilde{m}_j^{in} + \widetilde{m}_j^{in} - c_j||^2 && (40)\\
&\leq 2||m_j - \widetilde{m}_j^{in}||^2 + 2||\widetilde{m}_j^{in} - c_j||^2 && (41)\\
&\leq (2 - \frac{2}{5}\epsilon)||m_j - \widetilde{m}_j^{in}||^2 + \frac{4}{5}\epsilon r_j^2 + \frac{49}{60}\epsilon\sigma_j^2 && (42)\\
&\leq (2 - \frac{2}{5}\epsilon)||m_j - m_j^{in} + m_j^{in} - \widetilde{m}_j^{in}||^2 + \frac{4}{5}\epsilon r_j^2 + \frac{49}{60}\epsilon\sigma_j^2 && (43)\\
&\leq (2 - \frac{2}{5}\epsilon)(2||m_j - m_j^{in}||^2 + 2||m_j^{in} - \widetilde{m}_j^{in}||^2) + \frac{4}{5}\epsilon r_j^2 + \frac{49}{60}\epsilon\sigma_j^2 && (44)\\
&\leq (2 - \frac{2}{5}\epsilon)(\frac{1}{24}\epsilon\sigma_j^2 + 2r_j^2) + \frac{4}{5}\epsilon r_j^2 + \frac{49}{60}\epsilon\sigma_j^2 && (45)\\
&\leq \frac{9}{10}\epsilon\sigma_j^2 + 4r_j^2 && (46)\\
&= \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2. && (47)
\end{aligned}
$$

□

5.2. Analysis for Case 2: $|\mathcal{X}_j^{out}| > \frac{\epsilon}{49}\beta_j n$

Let $\widetilde{\mathcal{X}}_j = \widetilde{\mathcal{X}}_j^{in} \cup \mathcal{X}_j^{out}$, and $\widetilde{m}_j$ denote the centroid of $\widetilde{\mathcal{X}}_j$. Our idea is to replace the center of $\mathcal{X}_j$ with the center of $\widetilde{\mathcal{X}}_j$. But it is difficult to seek out the center of $\widetilde{\mathcal{X}}_j$. Thus, we try to find an approximate center c_j of $\widetilde{\mathcal{X}}_j$.

Lemma 12. $\frac{|\mathcal{X}_j^{out}|}{|\mathcal{X} \backslash \mathcal{B}_j|} \geq \frac{\epsilon^2}{3969k}$.

Proof.

$$\frac{|\mathcal{X}_j^{out}|}{|\mathcal{X} \setminus \mathcal{B}_j|} = \frac{|\mathcal{X}_j^{out}|}{\sum_{i=1}^{j-1} |\mathcal{X}_i \setminus \mathcal{B}_j| + |\mathcal{X}_j^{out}| + \sum_{i=j+1}^{k} |\mathcal{X}_i \setminus \mathcal{B}_j|} \tag{48}$$

$$\geq \frac{|\mathcal{X}_j^{out}|}{\sum_{i=1}^{j-1} \frac{f_2(c_i,\mathcal{X}_i)}{r_j^2} + |\mathcal{X}_j^{out}| + \sum_{i=j+1}^{k} |\mathcal{X}_i|} \tag{49}$$

$$\geq \frac{|\mathcal{X}_j^{out}|}{\sum_{i=1}^{j-1} \frac{f_2(m_i,\mathcal{X}_i)+|\mathcal{X}_i|||m_i-c_i||^2}{r_j^2} + |\mathcal{X}_j^{out}| + \sum_{i=j+1}^{k} |\mathcal{X}_i|} \tag{50}$$

$$\geq \frac{|\mathcal{X}_j^{out}|}{\frac{(1+\epsilon)n\sigma_{opt}^2}{r_j^2} + |\mathcal{X}_j^{out}| + \sum_{i=j+1}^{k} |\mathcal{X}_i|} \tag{51}$$

$$\geq \frac{|\mathcal{X}_j^{out}|}{\frac{40(1+\epsilon)k\beta_j n}{\epsilon} + |\mathcal{X}_j^{out}| + (k-j)\beta_j n} \tag{52}$$

$$\geq \frac{\frac{\epsilon}{49}\beta_j n}{\frac{40(1+\epsilon)k\beta_j n}{\epsilon} + \frac{\epsilon}{49}\beta_j n + (k-j)\beta_j n} \tag{53}$$

$$\geq \frac{\epsilon^2}{(80k+k)49 + (\epsilon - 49j)\epsilon} \tag{54}$$

$$\geq \frac{\epsilon^2}{3969k} \tag{55}$$

□

Lemma 13. $||m_j - \widetilde{m}_j|| \leq r_j$.

Proof.

$$||m_j - \widetilde{m}_j|| = ||\frac{1}{|\mathcal{X}_j|} \sum_{X \in \mathcal{X}_j} \int^{\mathbb{R}^d} s f_X(s) ds - \frac{1}{|\mathcal{X}_j|} (\sum_{X \in \mathcal{X}_j^{in}} \widetilde{X} + \sum_{X \in \mathcal{X}_j^{out}} \int^{\mathbb{R}^d} s f_X(s) ds)|| \tag{56}$$

$$= \frac{1}{|\mathcal{X}_j|} || \sum_{X \in \mathcal{X}_j^{in}} \int^{\mathbb{R}^d} (s - \widetilde{X}) f_X(s) ds|| \tag{57}$$

$$= \frac{1}{|\mathcal{X}_j|} \sum_{X \in \mathcal{X}_j^{in}} \int^{\mathbb{R}^d} ||s - \widetilde{X}|| f_X(s) ds \tag{58}$$

$$\leq \frac{1}{|\mathcal{X}_j|} \sum_{X \in \mathcal{X}_j^{in}} r_j \tag{59}$$

$$= \frac{|\mathcal{X}_j^{in}|}{|\mathcal{X}_j|} r_j \tag{60}$$

$$\leq r_j \tag{61}$$

□

Lemma 14. $f_2(\widetilde{m}_j, \widetilde{\mathcal{X}}_j) \leq 2 f_2(m_j, \mathcal{X}_j) + 4\beta_j n r_j^2$.

Proof.

$$f_2(\widetilde{m}_j, \widetilde{\mathcal{X}}_j) = \sum_{X\in\mathcal{X}_j^{in}} ||\widetilde{X} - \widetilde{m}_j||^2 + \sum_{X\in\mathcal{X}_j^{out}} \int^{\mathbb{R}^d} ||s - \widetilde{m}_j||^2 f_X(s)ds \quad (62)$$

$$= \sum_{X\in\mathcal{X}_j^{in}} \int^{\mathbb{R}^d} ||\widetilde{X} - \widetilde{m}_j||^2 f_X(s)ds + \sum_{X\in\mathcal{X}_j^{out}} \int^{\mathbb{R}^d} ||s - \widetilde{m}_j||^2 f_X(s)ds \quad (63)$$

$$= \sum_{X\in\mathcal{X}_j^{in}} \int^{\mathbb{R}^d} ||\widetilde{X} - s + s - \widetilde{m}_j||^2 f_X(s)ds + \sum_{X\in\mathcal{X}_j^{out}} \int^{\mathbb{R}^d} ||s - \widetilde{m}_j||^2 f_X(s)ds \quad (64)$$

$$\leq \sum_{X\in\mathcal{X}_j^{in}} \int^{\mathbb{R}^d} (2||\widetilde{X} - s||^2 + 2||s - \widetilde{m}_j||^2) f_X(s)ds + \sum_{X\in\mathcal{X}_j^{out}} \int^{\mathbb{R}^d} ||s - \widetilde{m}_j||^2 f_X(s)ds \quad (65)$$

$$\leq 2\sum_{X\in\mathcal{X}_j^{in}} \int^{\mathbb{R}^d} ||\widetilde{X} - s||^2 f_X(s)ds + 2\sum_{X\in\mathcal{X}_j^{out}} \int^{\mathbb{R}^d} ||s - \widetilde{m}_j||^2 f_X(s)ds \quad (66)$$

$$\leq 2|\mathcal{X}_j^{in}|r_j^2 + 2f_2(\widetilde{m}_j, \mathcal{X}_j) \quad (67)$$

$$= 2|\mathcal{X}_j^{in}|r_j^2 + 2f_2(m_j, \mathcal{X}_j) + 2|\mathcal{X}_j|||m_j - \widetilde{m}_j||^2 \quad (68)$$

$$\leq 2f_2(m_j, \mathcal{X}_j) + 4\beta_j n r_j^2 \quad (69)$$

□

Lemma 15. *In the process of finding c_j in our algorithm **cMeans**, we assume that $\mathcal{Q}$ satisfies $\mathcal{X} \setminus \mathcal{B}_j \subseteq \mathcal{Q}$ and $|\mathcal{Q}| < 2|\mathcal{X} \setminus \mathcal{B}_j|$. For the set $\mathcal{S}_2$ in step 5, a subset $\mathcal{S}^*$ of size $6/\epsilon$ of $\mathcal{S}_2$ can be obtained such that all random variables in $\mathcal{S}^*$ are from $\widetilde{\mathcal{X}}_j^{in}$ with a probability of 1/2. Let c_j denotes the centroid of $\mathcal{S}^*$. Then, the inequality $||\widetilde{m}_j - c_j||^2 \leq \frac{4}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\sigma_j^2$ holds with a probability of at least 1/6.*

Proof. In our algorithm **cMeans**, we assume that $\mathcal{S}_1 = S_1, \ldots, S_N$, where $N = 79380k/\epsilon^3$. Let $x_1^{'}, \ldots, x_N^{'}$ be the corresponding random variables of elements in $\mathcal{S}_1$. If $S_i \in \mathcal{X}_j^{out}$, obtain $x_i^{'} = 1$, or else $x_i^{'} = 0$. It is known easily that $Pr[S_i \in \mathcal{X}_j^{out}] \geq \frac{\epsilon^2}{7938k}$ by Lemma 12. Let $x = \sum_{i=1}^N x_i^{'}, u = \sum_{i=1}^N E(x_i^{'})$. We obtain that $u \geq 10/\epsilon$, and

$$Pr[x > \frac{6}{\epsilon}] = 1 - Pr[x \leq \frac{6}{\epsilon}] \quad (70)$$

$$\geq 1 - Pr[x \leq \frac{3}{5}u] \quad (71)$$

$$\geq 1 - e^{-\frac{(1-\frac{3}{5})^2 u}{2}} \quad (72)$$

$$\geq 1 - e^{-\frac{(1-\frac{3}{5})^2 \frac{10}{\epsilon}}{2}} \quad (73)$$

$$\geq 1 - e^{-\frac{4}{5}} \quad (74)$$

$$\geq \frac{1}{2}. \quad (75)$$

Then, the probability that at least $6/\epsilon$ random variables in $\mathcal{S}_1$ are from $\mathcal{X}_j^{out}$ is at least 1/2. Since $\mathcal{S}_2 = \mathcal{S}_1 \cup \{6/\epsilon$ copies of each point in $C\}$, a subset $\mathcal{S}^*$ of size $6/\epsilon$ of $\mathcal{S}_2$ can be obtained, and the probability that all random variables in $\mathcal{S}^*$ are from $\widetilde{\mathcal{X}}_j^{in}$ is at least 1/2. Let c_j denote the centroid of $\mathcal{S}^*$ and $\delta = 5/6$. For $|\mathcal{S}^*| = 6/\epsilon$ and $|widetilde\mathcal{X}_j| = |\mathcal{X}_j|$, by Lemma 2, $||\widetilde{m}_j - c_j||^2 \leq \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j,\widetilde{\mathcal{X}}_j)}{|\widetilde{\mathcal{X}}_j|} = \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j,\widetilde{\mathcal{X}}_j)}{|\mathcal{X}_j|}$ holds with a probability of at least 1/6. Assume that $||\widetilde{m}_j - c_j||^2 \leq \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j,\widetilde{\mathcal{X}}_j)}{|\mathcal{X}_j|}$. Then,

$$||\widetilde{m}_j - c_j||^2 \leq \frac{\epsilon}{5}\frac{f_2(\widetilde{m}_j,\widetilde{\mathcal{X}}_j)}{|\mathcal{X}_j|} \leq \frac{\epsilon}{5}\frac{2f_2(m_j,\mathcal{X}_j)+4\beta_j n r_j^2}{|\mathcal{X}_j|} \leq \frac{4}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\sigma_j^2. \quad (76)$$

□

Lemma 16. *If c_j satisfies $||\widetilde{m}_j - c_j||^2 \leq \frac{4}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\sigma_j^2$, then $||m_j - c_j||^2 \leq \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2$.*

Proof. Assume that c_j satisfies $||\widetilde{m}_j - c_j||^2 \leq \frac{4}{5}\epsilon r_j^2 + \frac{2}{5}\epsilon\sigma_j^2$. Then,

$$\begin{aligned}
||m_j - c_j||^2 &= ||m_j - \widetilde{m}_j + \widetilde{m}_j - c_j||^2 && (77)\\
&\leq 2||m_j - \widetilde{m}_j||^2 + 2||\widetilde{m}_j - c_j||^2 && (78)\\
&\leq 2r_j^2 + \frac{8}{5}\epsilon r_j^2 + \frac{4}{5}\epsilon\sigma_j^2 && (79)\\
&= \frac{4}{5}\epsilon\sigma_j^2 + (2 + \frac{8}{5}\epsilon) r_j^2 && (80)\\
&\leq \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2. && (81)
\end{aligned}$$

□

Lemma 17. *Given an instance $(\mathcal{X}, k, \mathbb{L})$ of the uncertain constrained k-means problem, where the size of $\mathcal{X}$ is n, for $\forall\epsilon \in (0,1], k \geq 2$, we assume that by using our algorithm **cMeans**(ϵ, $\mathcal{X}$, k, C,U) (C and U are initialized as empty sets), a collection U of candidate sets including approximate centers is obtained. If there exists a set $C_k = \{c_1, \ldots, c_k\}$ in U satisfying that $||m_j - c_j||^2 \leq \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2 (1 \leq j \leq k)$, then C_k is a $(1+\epsilon)$-approximation for the uncertain constrained k-means problem.*

Proof. Assume that $C_k = c_1, \ldots, c_k$ is a set in U satisfying that $||m_j - c_j||^2 \leq \frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2 (1 \leq j \leq k)$. Then,

$$\begin{aligned}
\sum_{j=1}^{k} f_2(c_j, \mathcal{X}_j) &= \sum_{j=1}^{k}(f_2(m_j, \mathcal{X}_j) + |\mathcal{X}_j| ||m_j - c_j||^2) && (82)\\
&\leq \sum_{j=1}^{k}(f_2(m_j, \mathcal{X}_j) + \beta_j n(\frac{9}{10}\epsilon\sigma_j^2 + \frac{1}{10\beta_j k}\epsilon\sigma_{opt}^2)) && (83)\\
&\leq \sum_{j=1}^{k}(f_2(m_j, \mathcal{X}_j) + \frac{9}{10}\epsilon n \sum_{j=1}^{k}\beta_j\sigma_j^2 + \frac{1}{10}\epsilon n\sigma_{opt}^2 && (84)\\
&\leq \sum_{j=1}^{k}(f_2(m_j, \mathcal{X}_j) + \frac{9}{10}\epsilon n\sigma_{opt}^2 + \frac{1}{10}\epsilon n\sigma_{opt}^2 && (85)\\
&= (1+\epsilon)\cdot OPT_k(P). && (86)
\end{aligned}$$

□

5.3. Time Complexity Analysis

We analyze the time complexity for our algorithm **cMeans** in this section.

Lemma 18. *The time complexity of our algorithm **cMeans** is $O(4^k(\frac{13231ek}{\epsilon^2})^{6k/\epsilon}\frac{1}{\epsilon}nd)$.*

Proof. Let $a = C_{N+kM}^{M}$, which $N = \frac{79380k}{\epsilon^3}$, $M = \frac{6}{\epsilon}$. By the Stirling formula,

$$C_{N+kM}^{M} \leq \frac{(N+kM)^M}{M!} \approx O((e\frac{N+kM}{M})^M) = O((\frac{13231ek}{\epsilon^2})^{\frac{6}{\epsilon}}).$$

In our algorithm **cMeans**, steps 5–9 have a run time of $O(k/\epsilon^3)$, step 11 have a run time of $O(d/\epsilon)$, and steps 13–16 have a run time of $O(knd)$. Let $T(n,g)$ denote the time complexity of algorithm **cMeans**, where g is the number of cluster centers, and n is the size of $\mathcal{Q}$.

If $g = 0$, $T(n,0) = O(1)$. When $n = 1$, $T(1,g) = a(T(1,g-1) + O(d/\epsilon)) + O(k/\epsilon^3)$. Because $a > k/\epsilon^3$, $T(1,g) = a(T(1,g-1) + O(d/\epsilon)) \leq a^g \cdot T(1,0) + g \cdot a^g \cdot O(d/\epsilon) = O(g \cdot a^g \cdot d/\epsilon)$. Therefore, $T(1,g) \leq O(4^g(\frac{13231ek}{\epsilon^2})^{6g/\epsilon})\frac{1}{\epsilon}d$, where $e = 2.7183$.

For $\forall n \geq 2$ and $g \geq 1$, the recurrence of $T(n,g)$ could be obtained as follows:

$$T(n,g) = a \cdot T(n,g-1) + T(\lfloor\frac{n}{2}\rfloor, g) + a \cdot O(\frac{d}{\epsilon}) + O(\frac{k}{\epsilon^3}) + O(knd).$$

Because $a > k/\epsilon^3$, two constants b_1 and b_2 with $b_1 \geq 1$ and $b_2 \geq 1$ could be obtained to arrive at the following recurrence.

$$T(n,g) \leq a \cdot T(n,g-1) + T(\lfloor\frac{n}{2}\rfloor, g) + a \cdot b_1 \cdot \frac{d}{\epsilon} + b_2 \cdot knd.$$

Now we claim that $T(n,g) \leq b_1 \cdot b_2 \cdot \frac{1}{\epsilon} \cdot a^g \cdot 2^{2g} \cdot nd - b_1 \cdot \frac{d}{\epsilon}$. If $g = 0$, then $T(n,0) = O(1)$. If $g \geq 1, n = 1$, then $T(1,g) \leq O(4^g(\frac{13231ek}{\epsilon^2})^{6g/\epsilon})\frac{1}{\epsilon}d$, and the claim holds. Suppose that if $\forall n_1 \geq 0, \forall g > g_1$, the claim holds for $T(n_1, g_1)$, and if $\forall 0 < n_2 < n, \forall g_2$, the claim holds for $T(n_2, g_2)$. We need to prove that:

$$\begin{aligned} b_1 \cdot b_2 \cdot \frac{1}{\epsilon} \cdot a^g \cdot 2^{2g} \cdot nd - b_1 \cdot \frac{d}{\epsilon} \geq a(b_1 \cdot b_2 \cdot \frac{1}{\epsilon} \cdot a^(g-1) \cdot 2^{2(g-1)} \cdot nd - b_1 \cdot \frac{d}{\epsilon}) \\ + b_1 \cdot b_2 \cdot \frac{1}{\epsilon} \cdot a^g \cdot 2^{2g} \cdot \lfloor\frac{n}{2}\rfloor d - b_1 \cdot \frac{d}{\epsilon} + a \cdot b_1 \cdot \frac{d}{\epsilon} + b_2 \cdot knd. \end{aligned}$$

The above formula can be simplified as $\frac{1}{4\epsilon} \cdot b_1 \cdot a^g 2^{2g} \geq k$, which holds for $\forall g \geq 1$. For $a = (\frac{13231ek}{\epsilon^2})^{6/\epsilon}$, $T(n,k) = O(4^k(\frac{13231ek}{\epsilon^2})^{6k/\epsilon}\frac{1}{\epsilon}nd)$. □

Thus, we can obtain the following Theorem 2.

Theorem 2. *Given an instance $(\mathcal{X}, k, \mathbb{L})$ of the uncertain constrained k-means problem, where the size of $\mathcal{X}$ is n, for $\forall \epsilon \in (0,1], k \geq 2$, by using our algorithm* ***cMeans***$(\epsilon, \mathcal{X}, k, C, U)$*, a collection U of candidate sets including approximate centers can be obtained with a probability of at least* $1/12^2$ *such that U includes at least one candidate set including approximate centers that is a* $(1+\epsilon)$*-approximation for the uncertain constrained k-means problem, and the time complexity of our algorithm* ***cMeans*** *is* $O(4^k(\frac{13231ek}{\epsilon^2})^{6k/\epsilon}\frac{1}{\epsilon}nd)$.

6. Conclusions

In this paper, we defined the uncertain constrained k-means problem first, and then presented a stochastic approximate algorithm for the problem in detail. We proposed a general mathematical model of the uncertain constrained k-means problem, and studied the random sampling properties, which are very important to deal with the uncertain constrained k-means problem. By applying a random sampling technique, we obtained a $(1+\epsilon)$-approximate algorithm for the problem. Then, we investigated the success probability, correctness and time complexity analysis of our algorithm **cMeans**, whose running time is $O(4^k(\frac{13231ek}{\epsilon^2})^{6k/\epsilon}\frac{1}{\epsilon}nd)$. However, there also exists a big gap between the current algorithms for the uncertain constrained k-means problem and the practical algorithms for the problem, which has been mentioned in [13] similarly.

We will try to explore a much more practical algorithm for the uncertain constrained k-means problem in future. It is known that the 2-means problem is the smallest version of the k-means problem, and remains NP-hard. The approximation schemes for the 2-means problem can be generalized to solve the k-means problem. Due to the particularity of the uncertain constrained 2-means problem, we will study approximation schemes for the uncertain constrained 2-means problem and reduce the algorithm complexity of approximation schemes for the uncertain constrained k-means problem through approximation schemes of the uncertain constrained 2-means problem. Additionally, we will apply the proposed algorithm to some practical problems in the future.

Author Contributions: J.L. and J.T. contributed to supervision, methodology, validation and project administration. B.X. and X.T. contributed to review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Science and Technology Foundation of Guizhou Province ([2021]015), in part by the Open Fund of Guizhou Provincial Public Big Data Key Laboratory (2017BDKFJJ019), in part by the Guizhou University Foundation for the introduction of talent ((2016) No. 13), in part by the GuangDong Basic and Applied Basic Research Foundation (No. 2020A1515110554), and in part by the Science and Technology Program of Guangzhou (No. 202002030138), China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Feldman, D.; Monemizadeh, M.; Sohler, C. A PTAS for k-means clustering based on weak coresets. In Proceedings of the 23rd ACM Symposium on Computational Geometry, SoCG, Gyeongju, Korea, 6–8 June 2007; pp. 11–18.
2. Ostrovsky, R.; Rabani, Y.; Schulman, L.J.; Swamy, C. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM* **2012**, *59*, 28:1–28:22. [CrossRef]
3. Jaiswal, R.; Kumar, A.; Sen, S. A simple D^2-sampling based PTAS for k-means and other clustering problems. *Algorithmica* **2014**, *71*, 22–46. [CrossRef]
4. Arkin, E.M.; Diaz-Banez, J.M.; Hurtado, F.; Kumar, P.; Mitchell, J.S.; Palop, B.; Perez-Lantero, P.; Saumell, M.; Silveira, R.I. Bichromatic 2-center of pairs of points. *Comput. Geom.* **2015**, *48*, 94–107. [CrossRef]
5. Yhuller, S.; Sussmann, Y.J. The capacitated k-center problem. *SIAM J. Discrete Math.* **2000**, *13*, 403–418.
6. Har-Peled, S.; Raichel, B. Net and prune: A linear time algorithm for Euclidean distance problems. *J. ACM* **2015**, *62*, 4401–4435. [CrossRef]
7. Swamy, C.; Shmoys, D.B. Fault-tolerant facility location. *ACM Trans. Algorithms* **2008**, *4*, 1–27. [CrossRef]
8. Xu, G.; Xu, J. Efficient approximation algorithms for clustering point-sets. *Comput. Geom.* **2010**, *43*, 59–66. [CrossRef]
9. Valls, A.; Batet, M.; Lopez, E.M. Using expert's rules as background knowledge in the clusdm methodology. *Eur. J. Oper. Res.* **2009**, *195*, 864–875. [CrossRef]
10. Li, J.; Yi, K.; Zhang, Q. Clustering with deversity. In Proceedings of the 37th International Colloquium on Automata, Languages and Programming, ICALP, Bordeaux, France, 6–10 July 2010; pp. 188–200.
11. Ding, H.; Xu, J. A unified framework for clustering constrained data without locality property. In Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, San Diego, CA, USA, 4–6 January 2015; pp. 1471–1490.
12. Bhattacharya, A.; Jaiswal, R.; Kumar, A. Faster algorithms for the constrained k-means problem. *Theory Comput. Syst.* **2018**, *62*, 93–115. [CrossRef]
13. Feng, Q.; Hu, J.; Huang, N.; Wang, J. Improved PTAS for the constrained k-means problem. *J. Comb. Optim.* **2019**, *37*, 1091–1110. [CrossRef]
14. Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **1963**, *58*, 13–30. [CrossRef]