Greedy Strategy Works for Clustering with Outliers and Coresets Construction

Hu Ding
School of Computer Science and Engineering, University of Science and Technology of China, He Fei, China
huding@ustc.edu.cn

Abstract
We study the problems of clustering with outliers in high dimension. Though a number of methods have been developed in the past decades, it is still quite challenging to design quality guaranteed algorithms with low complexities for the problems. Our idea is inspired by the greedy method, Gonzalez’s algorithm, for solving the problem of ordinary $k$-center clustering. Based on some novel observations, we show that this greedy strategy actually can handle $k$-center/median/means clustering with outliers efficiently, in terms of qualities and complexities. We further show that the greedy approach yields small coreset for the problem in doubling metrics, so as to reduce the time complexity significantly. Moreover, a by-product is that the coreset construction can be applied to speedup the popular density-based clustering approach DBSCAN.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems - Geometrical problems and computations

Keywords and phrases minimum enclosing ball, high dimension, outlier recognition, dimension reduction, random sampling

Digital Object Identifier 10.4230/LIPIcs..2019.

1 Introduction
Clustering is one of the most fundamental problems in data analysis [37]. Given a set of elements, the goal of clustering is to partition the set into several groups based on their similarities or dissimilarities. Several clustering models have been extensively studied, such as $k$-center, $k$-median, and $k$-means clustering [6]. In reality, many datasets are noisy and contain outliers. Moreover, outliers could seriously affect the final results in data analysis. The problems of outlier removal have attracted a great amount of attention in the past decades [13]12. Clustering with outliers can be viewed as a generalization of the ordinary clustering problems; however, the existence of outliers makes the problems to be much more challenging.

1.1 Prior Work and Our Contribution
We consider the problem of $k$-center clustering with outliers first. Given $n$ points in $\mathbb{R}^d$ and the number of outliers $z$, the problem is to find $k$ balls to cover at least $n - z$ points and minimize the maximum radius of the balls. A 3-approximation algorithm for the problem in metric graph was proposed by [14]. for the problem in Euclidean space where the cluster centers can appear anywhere in the space, their approximation ratio becomes 4. A following streaming $(4 + \epsilon)$-approximation algorithm was proposed by [17]. Recently, [12] proposed a 2-approximation algorithm for metric $k$-center clustering with outliers (but it is unclear that what the resulting approximation ratio is for the problem in Euclidean space). Existing
algorithms often have high time complexities. For example, the complexities of the algorithms in [14, 47] are $O(kn^2 \log n + n^2 d)$ and $O\left(\frac{1}{2}(kznd + (kz)^2 \log \Phi)\right)$ respectively, where $\Phi$ is the ratio of the optimal radius to the shortest distance between any two distinct input points; the algorithm in [12] needs to solve a complicated model of linear programming and the exact time complexity is not provided. The coreset based idea of [8] needs to enumerate a large number of possible cases and also yields high complexity.

In this paper, we aim to design quality guaranteed algorithms with low complexities in high dimension. Our idea is inspired by the greedy method [29] for solving ordinary $k$-center clustering. Through some novel modifications and reasonable relaxation, we show that this greedy method also works for the problem of $k$-center clustering with outliers. Our approach can achieve the approximation ratio 2 with respect to the clustering cost (i.e., the radius); moreover, the time complexity is linear on the input size $nd$.

We further consider the problem in doubling metrics, motivated by the fact that many real-world datasets often manifest low intrinsic dimensions [9]. For example, image sets usually can be represented in low dimensional manifold though the Euclidean dimension of the image vectors can be very high. “Doubling dimension” is widely used for measuring the intrinsic dimensions of datasets [50]. We adopt the following assumption: the inliers of the given data have a low doubling dimension $\rho > 0$. Note that we do not have any assumption on the outliers; namely, the outliers can scatter arbitrarily in the space. We believe that this assumption captures a large range of high dimensional instances in reality. With this assumption, we show that our approach can further improve the clustering quality. In particular, the greedy approach is able to construct a coreset for the problem of $k$-center clustering with outliers; as a consequence, the time complexity can be significantly reduced if running existing algorithms on the coreset. The size of our coreset is $2z + O((2/\mu)^\rho k)$, where $\mu$ is the small parameter measuring the quality of the coreset (note that $z$ and $k$ are often much smaller than $n$ in practice; the coefficient 2 of $z$ actually can be further reduced to be arbitrarily close to 1, by increasing the coefficient of the second term $(2/\mu)^\rho k$). Moreover, our coreset is a natural “composable coreset” that is an emerging topic for solving the problems in distributed computing [36].

We are aware of some prior work on reducing the data size for $k$-center clustering with outliers. [15] and [35] respectively showed that if more than $z$ outliers are allowed to remove, the random sampling technique can be applied to reduce the data size; but their sample sizes depend on the dimension [15, 35] also provided the coresets construction for $k$-median/means clustering with outliers in doubling metrics, however, their method cannot be extended to the case of $k$-center. [2] considered the coresets construction for ordinary $k$-center clustering without outliers. More detailed discussion is shown in Remark 3.1.

Moreover, our coresets construction can be applied to speedup DBSCAN, a popular density-based clustering approach for outlier recognition [24]. Roughly speaking, DBSCAN groups the points locating in each dense region to be an individual cluster and labels the remaining points as outliers. Despite of its wide applications [51], a major bottleneck is the high time complexity especially when the data size is large. The running time of DBSCAN in $\mathbb{R}^2$ has been improved from $O(n^2)$ to $O(n \log n)$ by [21, 31]. For the case in general $d$-dimensional space, [17] and [28] respectively provided the algorithms achieving the running times lower than $O(n^2)$ ($d$ is assumed to be a constant). If $d$ is high, the straightforward implementation takes $O(n^2 d)$ time (we refer the reader to [28] for more details). It is still
an open problem that whether it is possible to improve the running time in high dimension. In this paper, we show that an approximate result can be obtained by running existing DBSCAN algorithms on the coreset, if the inliers (including core and border points) have a low doubling dimension.

Finally, our greedy strategy can be extended to handle $k$-median/means clustering with outliers. The theoretical algorithms [18,26,43] often have high complexities. Several heuristic algorithms have been studied before [16,48]. By using local search method, [32] provided a 274-approximation of $k$-means clustering with outliers, but the violation on the number of outliers is large (it removes $O(zk\log(n\Delta))$ outliers, where $\Delta$ denotes the diameter of the points).

1.2 Preliminaries

We introduce several important definitions that are used throughout the paper.

- **Definition 1 (k-Center Clustering with Outliers).** Given a set $P$ of $n$ points in $\mathbb{R}^d$ with two positive integers $k$ and $z < n$, the problem of $k$-center clustering with outliers is to find a subset $P'$ of $P$, where $|P'| \geq n - z$, and $k$ cluster centers $\{c_1, \ldots, c_k\} \subset \mathbb{R}^d$, such that $\max_{p \in P'} \min_{1 \leq j \leq k} ||p - c_j||$ is minimized. Here, $||x - y||$ denotes the Euclidean distance between two points $x$ and $y \in \mathbb{R}^d$.

In this paper, we always use $P_{opt}$, a subset of $P$ with size $|P| - z$, to denote the subset yielding the optimal solution. Also, let $\{C_1, \ldots, C_k\}$ be the $k$ clusters forming $P_{opt}$, and the resulting clustering cost be $r_{opt}$; that is, each $C_j$ is covered by an individual ball with radius $r_{opt}$.

Usually, the optimizations relating to outliers are challenging combinatorial problems. Thus we often relax our goal and allow to miss a little more outliers in practice. Actually the same relaxation idea has been adopted by a number of previous work on clustering with outliers [3,15,35].

- **Definition 2 ((k, z)-Center Clustering).** Let $P$ be an instance of $k$-center clustering with $z$ outliers, and $\epsilon \geq 0$. $(k, z)$-center clustering is to find a subset $P'$ of $P$, where $|P'| \geq n - (1 + \epsilon)z$, such that the corresponding clustering cost of Definition 1 on $P'$ is minimized.

  (i) If a solution has the clustering cost, i.e., the radius, at most $\alpha r_{opt}$ with $\alpha \geq 1$, it is called an $\alpha$-approximation. Moreover, if the solution outputs more than $k$ clustering centers, say $\beta k$ with $\beta \geq 1$, it is called an $(\alpha, \beta)$-approximation.

  (ii) Given a set $A$ of cluster centers ($|A|$ could be larger than $k$), the resulting clustering cost is denoted by $\phi_{a}(P, A)$.

Obviously, the problem in Definition 1 is a special case of $(k, z)$-center clustering with $\epsilon = 0$.

Actually, Definition 1 and 2 can be naturally extended to weighted case: each point has a non-negative weight and the number of outliers is replaced by the total weight of outliers. Further, we have the following definition of coresets.

- **Definition 3 (Coreset).** Given a small parameter $\mu \in (0, 1)$ and an instance $P$ of $k$-center clustering with $z$ outliers, a weighted point set $S$ is called the $\mu$-coreset of $P$, if $\phi_{0}(S, H) \in (1 \pm \mu)\phi_{0}(P, H)$ for any set $H$ of $k$ points.

Given a large-scale instance $P$, we can run existing algorithm on its coreset $S$ to compute an approximate solution for $P$; if $|S| \ll n$, the resulting running time can be significantly reduced. Formally, we have the following proposition (see the proof in our supplement).
Proposition 1. If the set $H$ is an $\alpha$-approximation of the $\mu$-coreset $S$, it is an $\alpha \times \frac{1+\mu}{1-\mu}$-approximation of $P$.

As mentioned before, we also consider the case with low doubling dimension. For any $x \in \mathbb{R}^d$ and $r \geq 0$, let $\text{Ball}(x, r)$ be the ball of radius $r$ around $x$.

Definition 4 (Doubling Dimension). The doubling dimension of a point set $P \subset \mathbb{R}^d$ is the smallest number $\rho$, such that for any $x \in P$ and $r \geq 0$, $P \setminus \text{Ball}(x, 2^\rho r)$ is always covered by the union of at most $2^\rho$ balls with radius $r$.

Doubling dimension describes the expansion rate of $P$. For example, a set of points are uniformly distributed inside a $\rho$-dimensional hypercube, and then their doubling dimension is $O(\rho)$ but the Euclidean dimension can be very high. For coresets construction, we adopt the following assumption.

Definition 5 (Low Doubling Dimension Assumption). Given an instance $P$ of $k$-center clustering with outliers, we assume that the inliers $P_{\text{opt}}$ have a constant doubling dimension $\rho > 0$, but the outliers $P \setminus P_{\text{opt}}$ can be scattered arbitrarily in the space.

Other notations. For convenience, we use $\text{dist}(p, Q)$ to denote the shortest distance between a point $p$ and a point set $Q$, i.e., $\min_{q \in Q} ||p - q||$. Further, given two point sets $Q_1$ and $Q_2$, we let $\text{dist}(Q_1, Q_2) = \min_{q_1 \in Q_1, q_2 \in Q_2} ||q_1 - q_2||$.

Other related work. [29,34] provided 2-approximations for ordinary $k$-center clustering, and proved that any approximation ratio lower than 2 implies $N = NP$. For $k$-means/median clustering, several approximation solutions have been proposed [5,40]; if $d$ or $k$ is a constant in $\mathbb{R}^d$, it is able to achieve their PTAS [19,27,38,41,44]. Recent research also focused on distributed clustering with outliers [30,45,46].

In computational geometry, coreset construction is a technique for reducing the data size so as to speedup many optimization problems; we refer the reader to the surveys [7,49] for more details. In particular, the coresets can be used to improve the running times of existing clustering algorithms in Euclidean space and doubling metrics [2,25,35].

The rest of the paper is organized as follows. We study the problem of $k$-center clustering with outliers in Section 2 and consider the coresets construction and application on DBSCAN in Section 3. Due to the space limit, we briefly summarize our results on $k$-median/means clustering with outliers in Section 4 and place the details in our supplement.

2 Algorithms for $(k, z)_{\epsilon}$-Center Clustering

For the sake of completeness, let us briefly introduce the algorithm of [29] first. Initially, it arbitrarily selects a point from $P$, and iteratively selects the following $k - 1$ points, where each $j$-th step ($2 \leq j \leq k$) chooses the point which has the largest minimum distance to the already selected $j - 1$ points; finally, each input point is assigned to its nearest neighbor of these $k$ points. It can be proved that this greedy strategy results in a 2-approximation of $k$-center clustering. In this section, we show that a modified version of Gonzalez’s algorithm yields approximate solutions for the problem of $(k, z)_{\epsilon}$-center clustering.

2.1 $(2, O(1/\epsilon))$-Approximation

Here, we consider the bi-criteria approximation that has more than $k$ cluster centers. The main challenge for implementing Gonzalez’s algorithm is that the outliers and inliers are mixed in $P$; for example, the selected point, which has the largest minimum distance to the
Thus, $\lambda_{1}(E) \geq 1$ by Lemma 6 if $\lambda_{j}(E) = k$, that means $C_{l} \cap E \neq \emptyset$ for any $1 \leq l \leq k$.

Lemma 6 can be easily obtained by Proposition 2, since $|P_{opt}|/|P| \geq 1 - \gamma$.

Recall that $\{C_{1}, C_{2}, \cdots, C_{k}\}$ are the $k$ clusters forming $P_{opt}$. Denote by $\lambda_{j}(E)$ the number of the clusters which have non-empty intersection with $E$ in the $j$-th round. For example, initially $\lambda_{1}(E) \geq 1$ by Lemma 6. Now suppose $\lambda_{j}(E) = k$, that means $C_{j} \cap E \neq \emptyset$ for any $1 \leq j \leq k$.

Lemma 7. In each round of Step 3 of Algorithm 1, with probability $1 - \eta$, either (1) $\text{dist}(Q_{j}, E) \leq 2r_{opt}$ or (2) $\lambda_{j}(E) \geq \lambda_{j-1}(E) + 1$.

Proof. Suppose that (1) is not true, i.e., $\text{dist}(Q_{j}, E) > 2r_{opt}$, and we prove that (2) is true. Let $\mathcal{J}$ include all the indices $l \in \{1, 2, \cdots, k\}$ with $E \cap C_{l} \neq \emptyset$. We claim that $Q_{j} \cap C_{l} = \emptyset$ for each $l \in \mathcal{J}$. Otherwise, let $p \in Q_{j} \cap C_{l}$ and $p' \in E \cap C_{l}$; due to the triangle inequality, we know that $||p - p'|| \leq 2r_{opt}$ which is in contradiction with the assumption $\text{dist}(Q_{j}, E) > 2r_{opt}$. Thus, $Q_{j} \cap P_{opt}$ only contains the points from $C_{l}$ with $l \notin \mathcal{J}$. Moreover, since the number of outliers is $z$, we know that $|Q_{j} \cap P_{opt}| \geq \frac{1}{1 + \epsilon}$. By Proposition 2 if randomly selecting $\frac{1 + \epsilon}{r_{opt}} \log \frac{1}{\eta}$ points from $Q_{j}$, with probability $1 - \eta$, it contains at least one point from $Q_{j} \cap P_{opt}$; also, the point must come from $\cup_{l \notin \mathcal{J}} C_{l}$. Overall, (2) $\lambda_{j}(E) \geq \lambda_{j-1}(E) + 1$ is true.

If (1) of Lemma 7 happens, i.e., $\text{dist}(Q_{j}, E) \leq 2r_{opt}$, then it implies that $\max_{p \in P \setminus Q_{j}} \text{dist}(p, E) \leq 2r_{opt}$; moreover, since $|Q_{j}| = (1 + \epsilon)z$, we have $\phi_{0}(P, E) \leq 2r_{opt}$.

Next, we assume that (1) in Lemma 7 never happens, and prove that $\lambda_{j}(E) = k$ with constant probability when $j = \Theta(k)$. The following idea actually has been used by [1] for

---

2 Some similar heuristics has been studied for other optimization problems. [22][23].
obtaining a bi-criteria approximation for $k$-means clustering. Define a random variable $x_j = 1$ if $\lambda_j(E) = \lambda_{j-1}(E)$, or 0 if $\lambda_j(E) \geq \lambda_{j-1}(E) + 1$, for $j = 1, 2, \ldots$. So $E[x_j] = \eta$ and

$$\sum_{1 \leq s \leq \eta} (1 - x_s) \leq \lambda_j(E).$$

(1)

Also, let $J_j = \sum_{1 \leq s \leq \eta} (x_s - \eta)$ and $J_0 = 0$. Then, $\{J_0, J_1, J_2, \ldots\}$ is a super-martingale with $J_{j+1} - J_j \leq 1$ (more details are shown in our supplement). Through Azuma-Hoeffding inequality, we have $Pr(J_t \geq J_0 + \delta) \leq e^{-\frac{\delta^2}{2}}$ for any $t \in \mathbb{Z}^+$ and $\delta > 0$. Let $t = \frac{k+\sqrt{k}}{1-\eta}$ and $\delta = \sqrt{k}$, the inequality implies that

$$Pr\left( \sum_{1 \leq s \leq \eta} (1 - x_s) \geq \delta \right) \geq 1 - e^{-\frac{\delta^2}{2}}.$$

(2)

Combining (1) and (2), we know that $\lambda(E) = k$ with probability at least $1 - e^{-\frac{\delta^2}{2}}$. Moreover, $\lambda(E) = k$ directly implies that $\text{dist}(p, E) \leq 2r_{\text{opt}}$ for any $p \in P_{\text{opt}}$ based on the triangle inequality. Together with Lemma 6, we have the following theorem.

**Theorem 8.** Let $\epsilon > 0$. If we set $t = \frac{k+\sqrt{k}}{1-\eta}$ for Algorithm 1 with probability $(1 - \eta)(1 - e^{-\frac{\delta^2}{2}})$, $\phi_\epsilon(P, E) \leq 2r_{\text{opt}}$.

**Quality and Running time.** If $\eta$ is a constant, Theorem 8 implies that $E$ is a $(2, O(\frac{1}{\eta}))$-approximation for $(k, z)$-center clustering of $P$ with constant probability. In each round of Step 3, there are $O(\frac{1}{\eta})$ new points added to $E$, thus it takes $O(\frac{1}{\eta}nd)$ time to update the distances from the points of $P$ to $E$; to select the set $Q_j$, we can apply the linear time PICK algorithm 10. Overall, the running time of Algorithm 1 is $O(\frac{1}{\eta}nd)$.

Further, we consider the practical instances under some reasonable assumption, and provide new analysis of Algorithm 1. In reality, the clusters are usually not too small, compared with the number of outliers. For example, it is rare to have a cluster $C_l$ that $|C_l| \ll z$.

**Theorem 9.** If we assume that each optimal cluster $C_l$ has the size at least $\epsilon z$ for $1 \leq l \leq k$, then the set $E$ of Algorithm 7 is a $(4, O(\frac{1}{\eta}))$-approximation for the problem of $(k, z)$-center clustering on $P$ with constant probability.

Compared with Theorem 8, Theorem 9 shows that we can exclude exactly $z$ outliers (rather than $(1+\epsilon)z$), though the approximation ratio with respect to the radius becomes 4. Moreover, our running time $O(\frac{1}{\eta}nd)$ is significantly lower than those of the 4 and $(4+\epsilon)$-approximations by 14,47 in Euclidean space (as discussed in Section 1.1).

**Proof of Theorem 9.** We take a more careful analysis on the proof of Lemma 7. If (1) never happens, eventually $\lambda_j(E)$ will reach $k$ and thus $\phi_\epsilon(P, E) \leq 2r_{\text{opt}}$. So we focus on the case that (1) happens before $\lambda_j(E)$ reaching $k$. Suppose at $j$-th round, $\text{dist}(Q_j, E) \leq 2r_{\text{opt}}$ but $\lambda_j(E) < k$. We consider two cases (i) there exists some $l_0 \notin J$ such that $C_{l_0} \cap (P \setminus Q_j) = \emptyset$ and (ii) otherwise.

For (i), $C_{l_0} \cap (P \setminus Q_j) = \emptyset$ implies that $C_{l_0} \subset Q_j$. Note that we assume $|C_{l_0}| \geq \epsilon z$, i.e., $|C_{l_0}| \geq \frac{\epsilon z}{Q_j}$. Using the same manner in the proof of Lemma 7, we know that (2) $\lambda_j(E) \geq \lambda_{j-1}(E) + 1$ happens with probability $1 - \eta$.

For (ii), $C_{l} \cap (P \setminus Q_j) \neq \emptyset$ for all $l \notin J$. Together with $\text{dist}(Q_j, E) \leq 2r_{\text{opt}}$, we know that there exists $q_i \in C_{l}$ (for each $l \notin J$) such that $\text{dist}(q_i, E) \leq 2r_{\text{opt}}$. Consequently, we
\[
q \in E \\
q_l \in E
\]

\[
\text{Algorithm 2 2-Approximation Algorithm}
\]

\textbf{Input:} An instance \( P \subset \mathbb{R}^d \) of \( k \)-center clustering with \( z \) outliers, and \(|P| = n\); the parameter \( \epsilon > 0 \).

1. Maintain a set \( E \) that is empty at the beginning.
2. Initially, \( j = 1 \); randomly select one point from \( P \) and add it to \( E \).
3. Run the following steps until \( j = k \):
   a. \( j = j + 1 \) and let \( Q_j \) be the farthest \((1 + \epsilon)z\) points to the current \( E \).
   b. Randomly select one point from \( Q_j \) and add it to \( E \).

\textbf{Output} \( E \).

have that \( \forall q \in C_l \),
\[
\text{dist}(q, E) \leq \|q - q_l\| + \text{dist}(q_l, E) \\
\leq 4r_{opt}, \text{ (see the left of Figure 1)}
\]

\[ \Rightarrow \phi_0(P, E) \leq \max_{q \in \bigcup_{l=1}^{k} C_l} \text{dist}(q, E) \leq 4r_{opt}. \quad (3) \]

Overall, after \( t = O(k) \) steps, either \( \lambda_k(E) = k \), i.e., a \((2, O(\frac{1}{\epsilon}))\)-approximation, or a \((4, O(\frac{1}{\epsilon}))\)-approximation of \((k, z)_{0}\)-center clustering is obtained with constant probability. \( \blacktriangleleft \)

\subsection{2.2 2-Approximation}

If \( k \) is a constant, we show that a single-criterion 2-approximation can be obtained. Actually, we use the same strategy as Section 2.1 but run only \( k \) rounds with each round sampling only one point. See Algorithm 2. However, the success probability would be exponentially small on \( k \); hence we need to repeat the process to guarantee a constant success probability.

Denote by \( \{v_1, \ldots, v_k\} \) the \( k \) sampled points of \( E \). Actually, the proof of Theorem 10 is almost identical to that of Lemma 7. The only difference is that the probability of (2) is at least \( 1 - \gamma \) (\( \gamma = z/n \)). If all of these events happen, we either obtain a 2-approximation before \( k \) steps (i.e., \( \text{dist}(E, P \setminus Q_j) \leq 2r_{opt} \) for some \( j < k \)), or \( \{v_1, \ldots, v_k\} \) fall into the \( k \) optimal clusters \( C_1, C_2, \ldots, C_k \) separately (i.e., \( \lambda_k(E) = k \)). No matter which case happens, we always obtain a 2-approximation. So we have Theorem 10.
Greedy Strategy Works for Clustering with Outliers and Coresets Construction

\[ \text{Theorem 10.} \quad \text{With probability at least } (1-\gamma)(\frac{1+\epsilon}{1-\epsilon})^{k-1}, \text{Algorithm 3 yields a 2-approximation for the problem of } (k,z)_0\text{-center clustering on } P. \text{ The running time is } O(kmd). \]

To boost the probability of Theorem 10, we just need to repeat the algorithm. Due to the space limit, please refer to supplement for the proof of Theorem 11.

\[ \text{Theorem 11.} \quad \text{If we run Algorithm 3 } \times O((1+\epsilon)(\frac{1+\epsilon}{1-\epsilon})^{k-1}) \text{ times, with constant probability, at least one time the algorithm yields a 2-approximation.} \]

Similar to Theorem 9, we consider the practical instances. We show that the quality of Theorem 10 can be preserved even only excluding exactly \( z \) outliers, if the optimal clusters are “well separated”. In fact, this property has been widely studied in previous clustering algorithms and is believed to be common for practical instances \([20,39]\). Let \( \{o_1, \cdots, o_k\} \) be the \( k \) cluster (ball) centers of the optimal clusters \( \{C_1, \cdots, C_k\} \).

\[ \text{Theorem 12.} \quad \text{Suppose that each optimal cluster } C_l \text{ has the size at least } \epsilon z \text{ and } ||o_l - o_{l'}|| > 4r_{opt} \text{ for } 1 \leq l \neq l' \leq k. \text{ Then with probability at least } (1-\gamma)(\frac{1+\epsilon}{1-\epsilon})^{k-1}, \text{ the result of Algorithm 2 yields a 2-approximation for the problem of } (k,z)_0\text{-center clustering on } P. \]

Proof. Initially, we know that \( \lambda_1(E) = 1 \) with probability \( 1-\gamma \). Suppose that at the beginning of the \( j \)-th round of Algorithm 2 with \( 2 \leq j \leq k \), \( E \) already has \( j-1 \) points separately falling in \( j-1 \) optimal clusters; also, we still let \( J \) be the set of the indices of these \( j-1 \) clusters. Then we have the following claim.

\[ \text{Claim 1.} \quad ||Q_j \cap (\cup_{l \in J} C_l)|| \geq \epsilon z. \]

Proof. For any \( p \in \cup_{l \in J} C_l \), we have

\[ \text{dist}(p,E) > 4r_{opt} - r_{opt} - r_{opt} = 2r_{opt} \quad (4) \]

from triangle inequality and the assumption \( ||o_l - o_{l'}|| > 4r_{opt} \) for \( 1 \leq l \neq l' \leq k \) (see the right of Figure 1). In addition, for any \( p \in \cup_{l \in J} C_l \), we have

\[ \text{dist}(p,E) \leq 2r_{opt} \quad (5) \]

We consider two cases. If \( \text{dist}(Q_j,E) \leq 2r_{opt} \) at the current round, then \( (4) \) directly implies that \( \cup_{l \in J} C_l \subset Q_j \) and thus \( ||Q_j \cap (\cup_{l \in J} C_l)|| = |\cup_{l \in J} C_l| \geq \epsilon z \) by the assumption that any \( |C_l| \geq \epsilon z \).

Otherwise, \( \text{dist}(Q_j,E) > 2r_{opt} \). Then \( Q_j \cap (\cup_{l \in J} C_l) = \emptyset \) by \( (5) \). Moreover, since there are only \( z \) outliers and \( |Q_j| = (1+\epsilon)z \), we know that \( |Q_j \cap (\cup_{l \in J} C_l)| \geq \epsilon z. \]

Claim 1 reveals that with probability at least \( \frac{1}{1+\epsilon} \), the new added point falls in \( \cup_{l \in J} C_l \), i.e., \( \lambda_j(E) = \lambda_{j-1}(E) + 1 \). Overall, we know that \( \lambda_k(E) = k \), i.e., \( E \) is a 2-approximation of \( (k,z)_0\text{-center clustering} \), with probability at least \( (1-\gamma)(\frac{1+\epsilon}{1-\epsilon})^{k-1}. \]

3 Coresets Construction in Doubling Metrics

In this section, we show some extensions of Algorithm 1 when the inliers \( P_{opt} \) have a constant doubling dimension \( \rho > 0 \). \textbf{Notation:} In the following analysis, we always assume that the assumption of Definition 5 is true by default.

From Definition 4, we directly know that each optimal cluster \( C_l \) of \( P_{opt} \) can be covered by \( 2^\rho \) balls with radius \( r_{opt}/2 \) (see the left figure in Figure 2). Imagine a new instance \( P \) having \( 2^\rho k \) clusters, where the optimal radius is at most \( r_{opt}/2 \). Therefore, we can just replace \( k \) by \( 2^\rho k \) when running Algorithm 1 so as to reduce the approximation ratio on the clustering cost from 2 to 1.
Algorithm 3 The Coreset Construction

Input: An instance $P \in \mathbb{R}^d$ of $k$-center clustering with $z$ outliers, and $|P| = n$; the parameters $\eta$ and $\mu \in (0, 1)$.
1. Let $l = \left(\frac{2}{\mu}\right)^{\rho} k$.
2. Set $\epsilon = 1$ and run Algorithm $1$ $t = \frac{l + \sqrt{l}}{1 - \eta}$ rounds. Record the value $\tilde{r}$ being the maximum distance between $E$ and $P$ by excluding the farthest $2z$ points, after the final round of Algorithm $1$.
3. Let $P_{\tilde{r}} = \{p \mid p \in P \text{ and } \text{dist}(p, E) \leq \tilde{r}\}$.
4. For each point $p \in P_{\tilde{r}}$, assign it to its nearest neighbor in $E$; for each point $q \in E$, let its weight be the number of points assigning to it.
5. Add $P \setminus P_{\tilde{r}}$ to $E$ with each point having weight $1$.

Output $E$ as the coreset.

Theorem 13. If we set $t = \frac{2^\rho \log 2}{1 - \eta}$ for Algorithm 1, with probability $(1 - \eta)(1 - e^{-\frac{1}{1 - \eta} - 1})$, $\phi_\epsilon(P, E) \leq r_{opt}$. So the set $E$ is a $(1, O(\frac{2^\rho}{\epsilon}))$-approximation for the problem of $(k, z)_{\epsilon}$-center clustering on $P$. The running time is $O(2^\rho \epsilon nk)$.

3.1 Coresets for $k$-Center Clustering with Outliers

Inspired by Theorem 13, we can further construct the coreset for the problem of $k$-center clustering with outliers. Let $\mu \in (0, 1)$. If applying Definition 4 recursively, we know that each $C_l$ is covered by $2^\rho \log 2/\mu = \left(\frac{2}{\mu}\right)^{\rho} k$ balls with radius $\frac{3}{2} r_{opt}$, and $P_{opt}$ is covered by $\left(\frac{2}{\mu}\right)^{\rho} k$ such balls in total. See the right figure in Figure 2. We have Algorithm 3 for constructing the $\mu$-coreset.

Theorem 14. With constant probability, Algorithm 3 outputs a $\mu$-coreset of $k$-center clustering with $z$ outliers on $P$. The size of the coreset $E$ is at most $2z + O\left(\left(\frac{2}{\mu}\right)^{\rho} k\right)$, and the construction time is $O\left(\left(\frac{2}{\mu}\right)^{\rho} nk\right)$.

Remark. (1) Comparing with the uniform sampling approaches [15, 35], our coreset size is independent of $d$. Moreover, another benefit is that our coreset works for exactly removing $z$ outliers. Consequently, our coreset can be used for existing algorithms of $k$-center with outliers, such as [14], to reduce their complexities. The previous ideas based on uniform sampling cannot get rid of the violation on the number of outliers, and the sample sizes become infinity if not allowing to remove more than $z$ outliers.

(2) Another feature is that our coreset is a natural composable coreset. If $P$ is partitioned
into $L$ parts, we can run Algorithm 3 for each part, and obtain a coreset with size $\left(2z + O\left(\left(\frac{z}{2}\right)^{\nu}k\right)\right)\frac{L}{n}$ in total (the proof is almost identical to the proof of Theorem 14 below). So our coreset construction can also handle distributed clustering with outliers.

(3) Very recently, [11] also provided a coreset for $k$-center clustering with $z$ outliers in doubling metrics, where their size is $T = O((k + z)\left(\frac{z}{2}\right)^{\nu})$ with $O(nTd)$ construction time. Thus our result in Theorem 14 is a significant improvement in terms of the coreset size and construction time.

(4) The coefficient $2$ of $z$ actually can be further reduced by modifying the value of $\epsilon$ in Step 2 of Algorithm 3. In general, the size of $E$ is $(1 + \epsilon)z + O\left(\frac{z}{2}\right)^{\nu}k$ and the construction time is $O\left(\frac{z}{2}\right)^{\nu}knd$.

Proof of Theorem 14 Similar to Theorem 13 we know that $|P| = n - 2z$ and $\hat{r} \leq 2 \times \frac{z}{2}r_{opt} = \mu r_{opt}$ (the value recorded in Step 2 of Algorithm 3) with constant probability. Thus, the size of $E$ is $|P \setminus P| + O\left(\left(\frac{z}{2}\right)^{\nu}k\right) = 2z + O\left(\left(\frac{z}{2}\right)^{\nu}k\right)$. Moreover, it is easy to see that the running time of Algorithm 3 is $O\left(\left(\frac{z}{2}\right)^{\nu}knd\right)$.

Next, we show that $E$ is a $\mu$-coreset of $P$. For each point $q \in E$, denote by $w(q)$ the weight of $q$; for the sake of convenience in our proof, we view each $q$ as a set of $w(q)$ overlapping points. Thus, from the construction of $E$, we can see that there is a bijective mapping $f$ between $P$ and $E$, where

$$|p - f(p)| \leq \mu r_{opt}, \quad \forall p \in P. \quad (6)$$

Let $H = \{c_1, c_2, \ldots, c_k\}$ be any $k$ points in the space. Suppose that $H$ induces $k$ clusters $\{A_1, A_2, \ldots, A_k\}$ (resp., $\{B_1, B_2, \ldots, B_k\}$) with respect to the problem of $k$-center clustering with $z$ outliers on $E$ (resp., $P$), where each $A_j$ (resp., $B_j$) has the cluster center $c_j$ for $1 \leq j \leq k$. Let $r_E = \phi_0(E, H)$ and $r_P = \phi_0(P, H)$, respectively. Also, let $r'_E$ (resp., $r'_P$) be the smallest value $r$ such that for any $1 \leq j \leq k$, $f(B_j) \subset Ball(c_j, r)$ (resp., $f^{-1}(A_j) \subset Ball(c_j, r)$). We need the following claim.

Claim 2. $|r'_E - r_P| \leq \mu r_{opt}$ and $|r'_P - r_E| \leq \mu r_{opt}$.

Proof. We just need to prove the first inequality since the other one can be obtained by the same manner.

Because each $B_j \subset Ball(c_j, r_P)$ and each point $p$ is moved by a distance at most $\mu r_{opt}$ based on (6), we know that $f(B_j) \subset Ball(c_j, r_P + \mu r_{opt})$, i.e., $r'_E \leq r_P + \mu r_{opt}$.

Let $p_0$ be the point realizing $r_P = \phi_0(P, H)$, that is, there exists some $1 \leq j_0 \leq k$ such that $|c_{j_0} - p_0| = r_P$. The triangle inequality and (6) together imply $|c_{j_0} - f(p_0)| \geq r_P - \mu r_{opt}$. Hence $r'_E \geq r_P - \mu r_{opt}$.

Overall, we have $|r'_E - r_P| \leq \mu r_{opt}$. \hfill \blacksquare

In addition, since $\{f(B_1), \ldots, f(B_k)\}$ also form $k$ clusters for the instance $E$ with respect to the $k$ cluster centers in $H$, we know that

$$r_E \leq r'_E. \quad (7)$$

Similarly, we have

$$r_P \leq r'_P. \quad (8)$$

Combining Claim 2, (7) and (8), we have $r_P - \mu r_{opt} \leq r_E \leq r_P + \mu r_{opt}$, that is

$$\phi_0(E, H) \in (1 \pm \mu)\phi_0(P, H). \quad (9)$$

Consequently, $E$ is a $\mu$-coreset of $P$. \hfill \blacksquare
Though the above obtained coreset can speedup the algorithms of $k$-center with $z$ outliers, we also are wondering that whether it can be used to handle $(k, z)_\epsilon$-center clustering for any $\epsilon > 0$; namely, does $\phi_\epsilon(E, H) \in \phi_\epsilon(P, H) \pm \mu r_{opt}$ (instead of (9)) hold?

Actually this can be proved by almost the same idea of the proof of Theorem 14. We just need to let $\{A_1, A_2, \ldots, A_k\}$ (resp., $\{B_1, B_2, \ldots, B_k\}$) be the induced $k$ clusters of $\phi_\epsilon(E, H)$ (resp., $\phi_\epsilon(P, H)$) instead; correspondingly, $r_E = \phi_\epsilon(E, H)$ and $r_P = \phi_\epsilon(P, H)$. As a consequence, Claim 2, 7, and 8 still hold, and thus $\phi_\epsilon(E, H) \in \phi_\epsilon(P, H) \pm \mu r_{opt}$. So we can run Algorithm 2 on the coreset to compute an approximate solution for $(k, z)_\epsilon$-center clustering.

\textbf{Corollary 15.} Given $\mu \in (0, 1)$ and $\epsilon > 0$, there exists an algorithm yielding a $(2 + 3\mu)$-approximation of $(k, z)_\epsilon$-center clustering in

$$O \left( \frac{2\mu}{\epsilon} k^{d-1} \left( z + \frac{2}{\mu} \right) k d \right)$$

(10)

time, with constant probability.

Comparing with Theorem 10 and 11, we replace the size $n$ by the size of the coreset in (10) while add an extra term $\frac{2}{\mu} k^{d-1} k d$ for constructing the coreset. The proof for the approximation ratio is shown in supplement.

### 3.2 Application for DBSCAN

Given two parameters $r > 0$ and $\text{MinPts} \in \mathbb{Z}^+$, we denote an instance $P$ of DBSCAN as $D(P, r, \text{MinPts})$. DBSCAN divides the points of $P$ into three classes:

1. $p$ is a \textbf{core point}, if $|\text{Ball}(p, r) \cap P| \geq \text{MinPts}$;
2. $p$ is a \textbf{border point}, if $p$ is not a core point but $p \in \text{Ball}(q, r)$ for some core point $q$;
3. the remaining points are all \textbf{outliers}.

To define a cluster of DBSCAN, we need the following definition.

\textbf{Definition 16 (Density-reachable).} We say a point $p \in P$ is density-reachable from a core point $q$, if there exists a sequence of points $p_1, p_2, \ldots, p_t \in P$ such that:

- $p_1 = q$ and $p_t = p$;
- $p_1, \ldots, p_{t-1}$ are all core points;
- $p_{i+1} \in \text{Ball}(p_i, r)$ for each $i = 1, 2, \ldots, t - 1$.

If one arbitrarily picks a core point $q$, then DBSCAN defines the corresponding cluster $= \{p \mid p \in P \text{ and } p \text{ is density-reachable from } q\}$. (11)

Namely, the cluster is the maximal subset containing the points who are density-reachable from $q$. The cluster may contain both core and border points. Also, the cluster is uniquely defined by any of its core points; that is, for any two core point $q$ and $q'$, they define exactly the same cluster if they are density-reachable from each other. Since MinPts is always fixed in our context, we simply use $\text{CL}(P, r)$ and $O(P, r)$ to denote the set of clusters and outliers of $D(P, r, \text{MinPts})$, respectively.

Here, we show that the method of coresets construction in Section 3.1 can be applied to reduce the data size for DBSCAN. We also follow the assumption in Definition 5, where the inliers include the core and border points. Similar to Theorem 14 we run Algorithm 3.
and obtain a much smaller set \( E \), such that existing DBSCAN algorithms can be applied to \( E \) and yield lower time complexities. To realize this idea, we consider the following two questions.

(1) What is the value of \( l \) for Step 2 of Algorithm 3 that is, how many rounds we need for Algorithm 1? Let \( \Delta \) be the diameter of the set of core and border points, i.e., the distance of the farthest two points. Based on the property of doubling dimension, we know that the set of core and border points are covered by \((\frac{4\Delta}{Pr})^p\) balls with radius \( \frac{4}{2}r \). Thus, we set \( l = (\frac{4\Delta}{Pr})^p \).

(2) Unlike Theorem 14, the value of \( \phi \) is unknown for DBSCAN, e.g., the number of outliers \( z \) and the diameter \( \Delta \). So we have to assume that some upper bounds are given in practice. For example, let \( \Delta \leq \tilde{\Delta} \) and \( z \leq \tilde{z} \), where \( \Delta \) and \( \tilde{z} \) are given values.

To state our result clearly, we need to define the relation between two clusterings. Let \( CLP,r,MinPts \) and \( CLP,1+\mu r,MinPts \) be the bijective mapping between \( P \), and \( P,1+\mu r,MinPts \) and \( O(P,1+\mu r) \subset f^{-1}(O(E,r)) \subset O(P,1+\mu r) \). The size of \( E \) is \( O\left(\tilde{z} + (\frac{4\Delta}{Pr})^p \right) \), and the construction time is \( O\left( (\frac{4\Delta}{Pr})^p \right) \).

**Theorem 17.** Given an instance \( \mathcal{D}(P,r,MinPts) \), with constant probability, Algorithm 3 outputs a set \( E \) such that

\[
CL(P,(1-\mu)r) \geq f^{-1}(CL(E,r)) \geq CL(P,(1+\mu)r);
\]

\[
O(P,(1+\mu)r) \subset f^{-1}(O(E,r)) \subset O(P,1+\mu r). \tag{12}
\]

The size of \( E \) is \( O\left(\tilde{z} + (\frac{4\Delta}{Pr})^p \right) \), and the construction time is \( O\left( (\frac{4\Delta}{Pr})^p \right) \).

**Remark.** Theorem 17 reveals that the result returned by \( \mathcal{D}(E,r,MinPts) \) is “bounded” by the results of \( \mathcal{D}(P,(1+\mu)r,MinPts) \) and \( \mathcal{D}(P,(1-\mu)r,MinPts) \). For any cluster \( A \in CL(P,r) \), there exists two clusters \( A^- \in CL(P,(1-\mu)r) \) and \( A^+ \in CL(P,(1+\mu)r) \), such that \( A^- \subset f^{-1}(A) \subset A^+ \). As mentioned in [28], “It is well-known that the clusters of DBSCAN rarely differ considerably when \( r \) changes by just a small factor. In fact, if this really happens, it suggests that the choice of \( r \) is very bad, such that the exact clusters are not stable anyway.”

**Proof of Theorem 17.** Let \( q \in P \) such that \( f(q) \) is a core point of the instance \( \mathcal{D}(E,r,MinPts) \). Thus,

\[
|Ball(f(q),r) \cap E| \geq MinPts. \tag{14}
\]

As mentioned before, the set of \((\frac{4\Delta}{Pr})^p\) balls covering core and border points have the radius \( \frac{4}{2}r \); therefore, the right-hand side of \( (6) \) should be \( \frac{4}{2}r \) instead. Through the triangle inequality, we know that

\[
f^{-1}\left(Ball(f(q),r) \cap E\right) \subset Ball(q,(1+\mu)r) \cap P, \tag{15}
\]

because \( f(q) \) and the points covered by \( Ball(f(q),r) \) all move by a distance at most \( \frac{4}{2}r \). \( (14) \) and \( (15) \) imply that

\[
|Ball(q,(1+\mu)r) \cap P| \geq MinPts. \tag{16}
\]

Therefore, \( q \) is a core point of \( \mathcal{D}(P,(1+\mu)r,MinPts) \). Let \( A \) be the cluster of \( \mathcal{D}(E,r,MinPts) \) that contains \( f(q) \), and \( q' \) be any point such that \( f(q') \in A \). Then \( f(q') \) should be density-reachable from \( f(q) \) with respect to \( \mathcal{D}(E,r,MinPts) \). Using the triangle inequality again,
we know that \( q' \) is density-reachable from \( q \) with respect to \( \mathcal{D}(P, (1 + \mu)r, \text{MinPts}) \). That is, \( q' \) belongs to the cluster of \( \mathcal{D}(P, (1 + \mu)r, \text{MinPts}) \) that contains \( q \), say \( B \). So we have
\[
f^{-1}(A) \subset B.
\]
Consequently, \( f^{-1}(\text{CL}(E, r)) \supseteq \text{CL}(P, (1 + \mu)r) \). The other side of (12) can be proved by the same manner.

Since the outliers are the remaining points by removing the core and border points, together with (12), we know that (13) is true.

\[\text{\ding{111}}\]

4 \((k, z)\)-Median/Means Clustering

Due to the space limit, we overview our main idea and place the details in our supplement. For the problems of \( k \)-median/means clustering with \( z \) outliers, the objective functions should be respectively replaced by \( \sum_{p \in P'} \min_{1 \leq j \leq k} ||p - c_j|| \) and \( \sum_{p \in P'} \min_{1 \leq j \leq k} ||p - c_j||^2 \) in Definition 1. By Markov’s inequality, we know that a large part of \( P_{\text{opt}} \) should be covered by \( k \) balls with some bounded radius. Therefore, we can convert the problems to be the instances of \( k \)-center clustering with outliers. As a consequence, Algorithm 1 and 2 are applicable for the cases of \( k \)-median/means, though the resulting approximation ratios are higher than those in Theorem 8 and 11. Similar to Theorem 13, if the assumption of Definition 5 is true, we can also reduce the approximation ratios with respect to the clustering costs to be 1.

5 Future Work

Following our work, several interesting problems deserve to be studied in future. For example, is there any lower bound on the size of the coresets for \( k \)-center clustering with outliers? In addition, can the coreset construction time of Algorithm 3 be further improved, like the fast net construction method proposed by [33] in doubling metrics? It is also interesting to consider to solve other problems involving outliers by using the greedy strategy studied in this paper.

References

1 Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for \( k \)-means clustering. In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, pages 15–28. Springer, 2009.

2 Sepideh Aghamolaei and Mohammad Ghodsi. A composable coreset for \( k \)-center in doubling metrics. In Proceedings of the 30th Canadian Conference on Computational Geometry, CCCG 2018, August 8-10, 2018, University of Manitoba, Winnipeg, Manitoba, Canada, pages 165–171, 2018.

3 Noga Alon, Seannie Dar, Michal Parnas, and Dana Ron. Testing of clustering. SIAM Journal on Discrete Mathematics, 16(3):393–417, 2003.

4 Noga Alon and Joel H Spencer. The probabilistic method. John Wiley & Sons, 2004.

5 Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for \( k \)-median and facility location problems. SIAM Journal on computing, 33(3):544–562, 2004.

6 Pranjal Awasthi and Maria-Florina Balcan. Center based clustering: A foundational perspective. 2014.

7 Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. arXiv preprint arXiv:1703.06476, 2017.
Greedy Strategy Works for Clustering with Outliers and Coresets Construction

8 Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In Proceedings of the ACM Symposium on Theory of Computing (STOC), pages 250–257, 2002.
9 Mikhail Belkin. Problems of learning on manifolds. The University of Chicago, 2003.
10 Manuel Blum, Robert W. Floyd, Vaughan Pratt, Ronald L. Rivest, and Robert E. Tarjan. Time bounds for selection. Journal of Computer and System Sciences, 7(4):448–461, 1973.
11 Matteo Ceccarello, Andrea Pietracaprina, and Geppino Pucci. Solving k-center clustering (with outliers) in mapreduce and streaming, almost as accurately as sequentially. CoRR, abs/1802.09205, 2018.
12 Deeparnab Chakrabarty, Prachi Goyal, and Ravishankar Krishnaswamy. The non-uniform k-center problem. In 43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy, pages 67:1–67:15, 2016.
13 Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3):15, 2009.
14 Moses Charikar, Samir Khuller, David M Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, pages 642–651. Society for Industrial and Applied Mathematics, 2001.
15 Moses Charikar, Liadan O’Callaghan, and Rina Panigrahy. Better streaming algorithms for clustering problems. In Proceedings of the thirty-fifth annual ACM symposium on Theory of computing, pages 30–39. ACM, 2003.
16 Sanjay Chawla and Aristides Gionis. k-means+: A unified approach to clustering and outlier detection. In Proceedings of the 2013 SIAM International Conference on Data Mining, pages 189–197. SIAM, 2013.
17 Danny Z. Chen, Michiel H. M. Smid, and Bin Xu. Geometric algorithms for density-based data clustering. Int. J. Comput. Geometry Appl., 15(3):239–260, 2005.
18 Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. In Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, pages 826–835. Society for Industrial and Applied Mathematics, 2008.
19 Vincent Cohen-Addad, Philip N Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. In Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on, pages 353–364. IEEE, 2016.
20 Amit Daniely, Nati Linial, and Michael E. Saks. Clustering is difficult only when it does not matter. CoRR, abs/1205.4891, 2012.
21 Mark de Berg, Ade Gunawan, and Marcel Roeloffzen. Faster dbscan and hdbscan in low-dimensional euclidean spaces. In 28th International Symposium on Algorithms and Computation, ISAAC 2017, December 9-12, 2017, Phuket, Thailand, pages 25:1–25:13, 2017.
22 Hu Ding and Jinhui Xu. Random gradient descent tree: A combinatorial approach for SVM with outliers. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., pages 2561–2567, 2015.
23 Hu Ding and Mingquan Ye. Solving minimum enclosing ball with outliers: Algorithm, implementation, and application. CoRR, abs/1804.09653, 2018.
24 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.
25 Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In Proceedings of the forty-third annual ACM symposium on Theory of computing, pages 569–578. ACM, 2011.
Zachary Friggstad, Kamyar Khodamoradi, Mohsen Rezapour, and Mohammad R Salavatipour. Approximation schemes for clustering with outliers. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 398–414. SIAM, 2018.

Zachary Friggstad, Mohsen Rezapour, and Mohammad R Salavatipour. Local search yields a ptas for k-means in doubling metrics. In *Foundations of Computer Science (FOCS)*, 2016 IEEE 57th Annual Symposium on, pages 365–374. IEEE, 2016.

Junhao Gan and Yufei Tao. Dbscan revisited: mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 519–530. ACM, 2015.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

Sudipto Guha, Yi Li, and Qiu Zhang. Distributed partial clustering. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures*, pages 143–152. ACM, 2017.

Ade Gunawan. A faster algorithm for dbscan. *Master’s thesis. Eindhoven University of Technology, the Netherlands*, 2013.

Shalmoli Gupta, Ravi Kumar, Kefu Lu, Benjamin Moseley, and Sergei Vassilvitskii. Local search methods for k-means with outliers. *Proceedings of the VLDB Endowment*, 10(7):757–768, 2017.

Sariel Har-Peled and Manor Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006.

Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

Lingxiao Huang, Shaofeng Jiang, Jian Li, and Xuan Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 814–825. IEEE, 2018.

Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, and Vahab S. Mirrokni. Composable core-sets for diversity and coverage maximization. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS’14, Snowbird, UT, USA, June 22-27, 2014*, pages 100–108, 2014.

Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

Ragesh Jaiswal, Mehul Kumar, and Pulkit Yadav. Improved analysis of d2-sampling based ptas for k-means and other clustering problems. *Information Processing Letters*, 115(2):100–103, 2015.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892, 2002.

Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

Stavros G Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM Journal on Computing*, 37(3):757–782, 2007.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Outlier detection techniques. *Tutorial at PAKDD*, 2009.

Ravishankar Krishnaswamy, Shi Li, and Sai Sandeep. Constant approximation for k-median and k-means with outliers via iterative rounding. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 646–659. ACM, 2018.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM (JACM)*, 57(2):5, 2010.
Greedy Strategy Works for Clustering with Outliers and Coresets Construction

45 Shi Li and Xiangyu Guo. Distributed k-clustering for data with heavy noise. In Advances in Neural Information Processing Systems, pages 7849–7857, 2018.

46 Gustavo Malkomes, Matt J Kusner, Wenlin Chen, Kilian Q Weinberger, and Benjamin Moseley. Fast distributed k-center clustering with outliers on massive data. In Advances in Neural Information Processing Systems, pages 1063–1071, 2015.

47 Richard Matthew McCutchen and Samir Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, pages 165–178. Springer, 2008.

48 Lionel Ott, Linsey Pang, Fabio T Ramos, and Sanjay Chawla. On integrated clustering and outlier detection. In Advances in neural information processing systems, pages 1359–1367, 2014.

49 Jeff M. Phillips. Coresets and sketches. Computing Research Repository, 2016.

50 Kunal Talwar. Bypassing the embedding: algorithms for low dimensional metrics. In Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, pages 281–290, 2004.

51 Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. 2006.

6 Proof of Proposition 1

Suppose $H$ is an $\alpha$-approximation of the instance (coreset) $S$. Let $H_{opt}$ be the set of $k$ cluster centers yielding the optimal solution of $P$. Then we have

$$\phi_0(S, H) \leq \alpha \phi_0(S, H_{opt});$$

$$\phi_0(S, H) \in (1 \pm \mu) \phi_0(P, H);$$

$$\phi_0(S, H_{opt}) \in (1 \pm \mu) \phi_0(P, H_{opt});$$

Combining the above inequalities, we directly have

$$\phi_0(P, H) \leq \frac{1}{1 - \mu} \phi_0(S, H)$$

$$\leq \frac{\alpha}{1 - \mu} \phi_0(S, H_{opt})$$

$$\leq \frac{\alpha(1 + \mu)}{1 - \mu} \phi_0(P, H_{opt}).$$

7 Proof of Proposition 2

Let the number of sampled elements be $h$. Since each sampled element falls in $T$ with probability $\tau$, by taking the union bound, we know that the sample contains at least one element from $T$ with probability $1 - (1 - \tau)^h$. Therefore, if we want $1 - (1 - \tau)^h \geq 1 - \eta$, $h$ should be at least $\log 1/\eta / \log 1/(1 - \tau) \leq \frac{1}{\tau} \log \frac{1}{\eta}$.

8 Proof of Theorem 8

We assume that (1) in Lemma 7 never happens, and prove that $\lambda_j(E) = k$ with constant probability when $j = \Theta(k)$. The idea actually has been used by [1] for obtaining a bi-criteria approximation for $k$-means clustering. Define a random variable $x_j = 1$ if $\lambda_j(E) = \lambda_{j-1}(E)$, or 0 if $\lambda_j(E) \geq \lambda_{j-1}(E) + 1$, for $j = 1, 2, \cdots$. So $E[x_j] = \eta$ and

$$\sum_{1 \leq s \leq j} (1 - x_s) \leq \lambda_j(E).$$
Also, let \( J_j = \sum_{1 \leq s \leq j} (x_s - \eta) \) and \( J_0 = 0 \). Then, \( \{J_0, J_1, J_2, \cdots \} \) is a super-martingale by the following definition.

**Definition 18.** A sequence of real valued random variables \( J_0, J_1, \cdots, J_t \) is called a super-martingale if for every \( j > 1 \), \( \mathbb{E}[J_j | J_0, \cdots, J_{j-1}] \leq J_{j-1} \).

In addition, we know that \( J_{j+1} - J_j \leq 1 \) for each \( j \geq 0 \). Through Azuma-Hoeffding inequality \([4]\), we have

\[
\Pr(J_t \geq J_0 + \delta) \leq e^{-\frac{\delta^2}{2t}}
\]  

(23)

for any \( t \in \mathbb{Z}^+ \) and \( \delta > 0 \). Let \( t = \frac{k+\sqrt{k}}{1-\eta} \) and \( \delta = \sqrt{k} \), \(23\) implies that

\[
\Pr( \sum_{1 \leq s \leq t} (1 - x_s) \geq t(1 - \eta) - \delta) \geq 1 - e^{-\frac{\delta^2}{2t}}
\]

\[
\Rightarrow \Pr( \sum_{1 \leq s \leq t} (1 - x_s) \geq k) \geq 1 - e^{-\frac{k^2}{2t(1-\eta)}}
\]

\[
\Rightarrow \Pr( \sum_{1 \leq s \leq t} (1 - x_s) \geq k) \geq 1 - e^{-\frac{k}{1-\eta}}.
\]  

(24)

Combining \(22\) and \(24\), we know that \( \lambda_t(E) = k \) with probability at least \( 1 - e^{-\frac{k}{1-\eta}} \). Moreover, \( \lambda_t(E) = k \) directly implies that \( \text{dist}(p,E) \leq 2r_{\text{opt}} \) for any \( p \in P_{\text{opt}} \) based on the triangle inequality. Together with Lemma 6, we complete the proof of Theorem 8.

**9 Proof of Theorem 11**

Suppose that we want the success probability to be some constant \( c \in (0,1) \). We can run Algorithm 2 \( N \) times, such that

\[
1 - (1-x)^N = c
\]  

(25)

where \( x = (1-\gamma)(\frac{1}{1+\epsilon})^{k-1} \). Thus, \( N = \log \frac{1}{1-c}/\log \frac{1}{1-x} = O\left(\frac{1+\epsilon}{\epsilon}\right)^{k-1} \) if we assume \( c \) and \( \frac{1}{1-x} \) are both constant values.

**10 The Approximation Ratio in Corollary 15**

Actually, the ratio can be obtained by almost the same manner of Proposition 1.

Suppose \( H \) is a 2-approximation of the instance (coreset) \( S \). Let \( H_{\text{opt}} \) be the set of \( k \) cluster centers yielding the optimal solution of \( P \). Then we have

\[
\phi_\epsilon(S,H) \leq 2\phi_0(S,H_{\text{opt}});
\]  

(26)

\[
\phi_\epsilon(S,H) \in [\phi_\epsilon(P,H) \pm \mu r_{\text{opt}}];
\]  

(27)

\[
\phi_0(S,H_{\text{opt}}) \in (1 \pm \mu)\phi_0(P,H_{\text{opt}});
\]  

(28)

Combining the above inequalities, we directly have

\[
\phi_\epsilon(P,H) \leq \phi_\epsilon(S,H) + \mu r_{\text{opt}}
\leq 2\phi_0(S,H_{\text{opt}}) + \mu r_{\text{opt}}
\leq 2(1 + \mu)\phi_0(P,H_{\text{opt}}) + \mu r_{\text{opt}}
= (2 + 3\mu)r_{\text{opt}}
\]  

(29)
### $(k, z)_{\epsilon}$-Median/Means Clustering

We show that the ideas in Section 2 and 3 can be used to solve the problems of $(k, z)_{\epsilon}$-median/means clustering.

**Definition 19** $(k$-Median/Means Clustering with Outliers). Given a set $P$ of $n$ points in $\mathbb{R}^d$ with two positive integers $k$ and $z < n$, the problem is to find a subset $P'$ of $P$, where $|P'| \geq n - z$, and $k$ cluster centers $\{c_1, \cdots, c_k\} \subset \mathbb{R}^d$, such that the clustering cost on $P'$ is minimized. Specifically, the clustering costs are defined as follows:

- **k-means clustering:** $\sum_{p \in P'} \min_{1 \leq j \leq k} ||p - c_j||^2;
- \text{ (30)}$

- **k-median clustering:** $\sum_{p \in P'} \min_{1 \leq j \leq k} ||p - c_j||.$
- \text{ (31)}

Similar to the case of $k$-center, we also allow a little more outliers than $z$. For any $\epsilon \geq 0$ and a set $A$ of cluster centers, we denote the corresponding clustering costs as $\phi(P, A)$ and $\phi_2^2(P, A)$ for the cases of $k$-median and $k$-means, respectively.

We focus on $(k, z)_{\epsilon}$-median clustering first; the results can be easily extended to $(k, z)_{\epsilon}$-means clustering. We still use $\{C_1, C_2, \cdots, C_k\}$ to denote the $k$ optimal clusters forming $P^*$ with respect to $(k, z)_{\epsilon}$-median clustering. Suppose that each cluster $C_j$ has its cluster center $c_j$ for $1 \leq j \leq k$. With a slight abuse of notations, we let $r_{opt} = \frac{k}{n-r} \phi(P, Opt)$ with $Opt = \{c_1, \cdots, c_k\}$. Also, let $\gamma = z/n$. The following lemma can be easily obtained via Markov’s inequality.

**Lemma 20.** For any $\lambda > 1$, $|P_{opt} \cap \bigcup_{j=1}^{k} \text{Ball}(c_j, \lambda r_{opt})| \geq (1 - \frac{1}{\lambda}) |P_{opt}|$.

Let $\lambda = \frac{3}{e \gamma}$. From Lemma 20, we know that there exists a subset $P'_{opt}$ of $P_{opt}$ covered by $k$ balls with radius $\lambda r_{opt}$, where $|P'_{opt}| \geq (1 - \frac{1}{\lambda}) |P_{opt}| = (1 - \frac{e}{3})(1 - \gamma)n > (1 - (1 + \frac{e}{3})\gamma)n$.

Now, we can run Algorithm 1 or 2 on $P$ with replacing $\epsilon$ and $\gamma$ by $\epsilon' = \frac{e}{3}$ and $\gamma' = (1 + \frac{e}{3})\gamma$, respectively; from Theorem 8 and 10, we know that

$$|P'_{opt} \cap \bigcup_{E \subseteq E} \text{Ball}(c, 2\lambda r_{opt})| \geq (1 - (1 + \epsilon')\gamma')n = (1 - (1 + \epsilon)(1 + \frac{e}{3})\gamma)n \geq (1 - (1 + \epsilon)\gamma)n \text{ (32)}$$

with certain probability, where $E$ is the set outputted by the algorithm. (32) directly implies that

$$\phi_1^1(P, E) \leq (1 - (1 + \epsilon)\gamma)n \times 2\lambda r_{opt} < 2\lambda(1 - \gamma)n r_{opt} = \frac{6}{e \gamma} \phi_1^1(P, Opt). \text{ (33)}$$

**Theorem 21.** Given an instance $P$ of $(k, z)_{\epsilon}$-median clustering, if running Algorithm 1 or 2, the approximation ratio with respect to the clustering cost becomes $\frac{6}{e \gamma}$ rather than 2.

Also, we have similar extensions in doubling metrics. Each of the $k$ balls can be covered by $(2\lambda)^p = \left(\frac{6}{e \gamma}\right)^p$ balls with radius $r_{opt}/2$. 
Theorem 22. Suppose that the assumption of Definition 5 is true. If we run Algorithm 1
\((6/\epsilon\gamma)^\rho k + (6/\epsilon\gamma)^{\rho/2} \sqrt{k}\) steps, with probability \((1 - \eta)(1 - e^{-\frac{\lambda}{\eta}})\), \(\phi_1(P, E) \leq \phi_0(P, Opt)\). That is, the set \(E\) is a \((1, O(\frac{6}{\epsilon\gamma}^\rho))\)-approximation for the problem of \((k, z)\)-median clustering on \(P\). The running time is \(O((\frac{6}{\epsilon\gamma}^\rho)^{\frac{\rho}{2}})\).

For the case of \(k\)-means, we instead let \(r_{opt} = \sqrt{\frac{1}{n-z} \phi_0^2(P, Opt)}\). Consequently, Lemma 20 turns to be \(|\mathcal{P}_{opt} \cap \left( \bigcup_{j=1}^k \text{Ball}(c_j, \sqrt{r_{opt}}) \right)| \geq (1 - \frac{1}{\lambda})|\mathcal{P}_{opt}|\) for any \(\lambda > 1\). With similar calculation, we have the following results.

Theorem 23. Given an instance \(P\) of \((k, z)\)-means clustering, if running Algorithm 1 or 2, the approximation ratio with respect to the clustering cost becomes \(\frac{12}{\epsilon\gamma}\) rather than 2.

Theorem 24. Suppose that the assumption of Definition 5 is true. If we run Algorithm 1
\((2\sqrt{3/\epsilon\gamma})^\rho k + (2\sqrt{3/\epsilon\gamma})^{\rho/2} \sqrt{k}\) steps, with probability \((1 - \eta)(1 - e^{-\frac{\lambda}{\eta}})\), \(\phi_2(P, E) \leq \phi_0(P, Opt)\).
That is, the set \(E\) is a \((1, O(\frac{2\sqrt{3}}{\epsilon\gamma}^\rho))\)-approximation for the problem of \((k, z)\)-means clustering on \(P\). The running time is \(O((2\sqrt{3}/\epsilon\gamma)^{\rho/2})\).