Improved Algorithms for Clustering with Outliers

Qilong Feng
School of Computer Science and Engineering, Central South University, P.R. China
csufeng@csu.edu.cn

Zhen Zhang
School of Computer Science and Engineering, Central South University, P.R. China
csuzz@foxmail.com

Ziyun Huang
Department of Computer Science and Software Engineering, Penn State Erie, The Behrend College, USA
zxh201@psu.edu

Jinhui Xu
Department of Computer Science and Engineering, State University of New York at Buffalo, USA
jinhui@cse.buffalo.edu

Jianxin Wang
School of Computer Science and Engineering, Central South University, P.R. China
jxwang@csu.edu.cn

Abstract

Clustering is a fundamental problem in unsupervised learning. In many real-world applications, the to-be-clustered data often contains various types of noises and thus needs to be removed from the learning process. To address this issue, we consider in this paper two variants of such clustering problems, called k-median with m outliers and k-means with m outliers. Existing techniques for both problems either incur relatively large approximation ratios or can only efficiently deal with a small number of outliers. In this paper, we present improved solution to each of them for the case where k is a fixed number and m could be quite large. Particularly, we gave the first PTAS for the k-median problem with outliers in Euclidean space $\mathbb{R}^d$ for possibly high m and d. Our algorithm runs in $O(nd(\frac{1}{\epsilon}(k+m))^{(\frac{k}{\epsilon})^{O(1)}})$ time, which considerably improves the previous result (with running time $O(nd(m+k)^{O(m+k)} + (\frac{1}{\epsilon}k\log n)^{O(1)}))$ given by [Feldman and Schulman, SODA 2012]. For the k-means with outliers problem, we introduce a $(6+\epsilon)$-approximation algorithm for general metric space with running time $O(n(\beta\frac{1}{\epsilon}(k+m))^k)$ for some constant $\beta > 1$. Our algorithm first uses the k-means++ technique to sample $O(\frac{1}{\epsilon}(k+m))$ points from input and then select the k centers from them. Compared to the more involving existing techniques, our algorithms are much simpler, *i.e.,* using only random sampling, and achieving better performance ratios.

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

Keywords and phrases Clustering with Outliers, Approximation, Random Sampling

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2019.64

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grants (61672536, 61420106009, 61872450, 61828205), National Science Foundation (CCF-1716400, IIS-1910492), Hunan Provincial Science and Technology Program (2018WK4001).

1 Introduction

Clustering is a fundamental problem in computer science and finds applications in a wide range of domains. Depending on the objective function, it has many different variants. Among them, k-median and k-means are perhaps the two most commonly considered versions. For a given set P of n points in some metric space, the k-median problem aims to identify a set of centers $C = \{c_1 \cdots c_k\}$ that minimizes the objective function $\sum_{x \in P} \min_{c_i \in C} \mathbf{d}(x, c_i)$, where $\mathbf{d}(x, c_i)$ denotes the distance from x to c_i. The k-means problem is very similar to the k-median problem, except that the clustering cost is measured by the squared distance from each point to its corresponding center.

Both the k-median and the k-means problems have shown to be NP-hard [6, 21]. Thus, most of the previous efforts have concentrated on obtaining approximation solutions. In the metric space settings, Charikar, Guha, and Shmoys [9] gave the first constant factor approximation solution to the k-median problem. Arya *et al.* [8] later showed that a simple local search heuristic yields a $(3+\epsilon)$-approximation. Li and Svensson [26] gave a $(1+\sqrt{3}+\epsilon)$-approximation based on a pseudo-approximation. Byrka *et al.* [23] further improved the approximation ratio to $(2.671+\epsilon)$. This is the current best known result for the k-median problem. For the k-means problem, Gupta and Tangwongsan [18] demonstrated that local search can achieve a $(25+\epsilon)$-approximation in metric spaces. The approximation ratio has been recently improved to $(9+\epsilon)$ by Ahmadian *et al.* [3] using a primal-dual algorithm.

All the above results allow the number k of clusters to be any integer between 1 and n. A common way to relax the problem is to assume that k is a fixed number and the space is Euclidean (instead of general metric). For this type of clustering problem, better results have already been achieved. Kumar, Sabharwal, and Sen [25] gave a linear time (*i.e.,* $O(2^{(k/\epsilon)^{O(1)}}nd)$) $(1+\epsilon)$-approximation algorithms for either problem in any dimensions. Chen [11] later improved the running time to $O(ndk + 2^{(k/\epsilon)^{O(1)}}d^2n^{\sigma})$ by using coresets, where σ is an arbitrary positive number. Feldman, Monemzadeh, and Sohler [15] further demonstrated that one can construct a coreset for the k-means problem with size independent of n and d. With this, they showed that a $(1+\epsilon)$-approximation can be obtained in $O(ndk + d\cdot \mathrm{poly}(k,\epsilon) + (\frac{k}{\epsilon})^{O(k/\epsilon)})$ time. Moreover, both the k-median and the k-means problems admit $(1+\epsilon)$-approximations for the case where the dimensionality of the space is a constant [17, 13].

The clustering problem has an implicit assumption that all input points can be clustered into k distinct groups, which may not always hold in real-world applications. Data from such applications are often contaminated with various types of noises, which need to be excluded from the solution. To deal with such noisy data, Charikar *et al.* [10] introduced the clustering with outliers problem. The problem is the same as the ordinary clustering problem, except that a small portion of the input data is allowed to be removed. The removed outlier points are ignored in the objective function. By discarding the set of outliers, one could significantly reduce the clustering cost and thus improve the quality of solution.

For the k-median with outlier problem, Charikar *et al.* [10] gave a $(4(1+\frac{1}{\epsilon}))$-approximation for metric space, which removes a slightly more than m (*i.e.,* $O((1+\epsilon)m)$) outliers. Chen [12] later obtained an algorithm which does not violate either k or m, but has a much large constant approximation ratio. Recently, Krishnaswamy, Li, and Sandeep [24] improved the approximation ratio to $7.08+\epsilon$ [24] using an elegant iterative rounding algorithm. For Euclidean space, better results have also been achieved. Friggstad *et al.* [16] presented a $(1+\epsilon)$-approximation algorithm that uses $(1+\epsilon)k$ centers and runs in $O((nk)^{1/\epsilon^{O(d)}})$ time. Their algorithm is efficient only in fixed dimensional space. Feldman and Schulman [14] gave a $(1+\epsilon)$-approximation algorithm without violating the number of the centers. Their algorithm runs in $O(nd(m+k)^{O(m+k)} + (\frac{1}{\epsilon}k\log n)^{O(1)})$ time, but needs to assume that both k and m are small constants to ensure a polynomial time solution. There has also been work on obtaining coresets for the problem [20].

For the k-means with outliers problem, Friggstad *et al.* [16] designed a bi-criteria algorithm that uses $(1+\epsilon)k$ centers and has an approximation ratio of $(25+\epsilon)$. Krishnaswamy, Li, and Sandeep [24] subsequently presented a $(53+\epsilon)$-approximation algorithm. This is the best existing approximation ratio for the problem.

1.1 Our Contributions

In this paper, we consider two variants of the clustering problem with outliers, k-median with outliers in Euclidean $\mathbb{R}^d$ space (where d could be very high) and k-means with outliers in metric space. For both problems, we assume that k is a fixed number and m is a variable less than n.

For the k-median with m outliers problem, we give the first PTAS for non-constant m, based on a simple random sampling technique. Our algorithm runs in $O(nd(\frac{1}{\epsilon}(k+m))^{(\frac{k}{\epsilon})^{O(1)}})$ time, which significantly improves upon the previously known $(1+\epsilon)$-approximation algorithm for the problem [14, 16].

▶ **Theorem 1.** *Given an instance (P,k,m) of the k-median with m outliers problem, there is a $(1+\epsilon)$-approximation algorithm that runs in $O(nd(\frac{1}{\epsilon}(k+m))^{(\frac{k}{\epsilon})^{O(1)}})$ time.*

For the k-means with m outliers problem, we give a $(6+\epsilon)$-approximation. Our algorithm first uses k-means++ [7] to sample $O(\frac{1}{\epsilon}(k+m))$ points from the input and then select k points from them as centers. k-means++ is an algorithm proposed for resolving the sensitivity issue of Lloyd's k-means algorithm [27] to the locations of its initial centers. Since the k-means with outliers problem needs to discard m outliers, which may cause major changes in the topological structure and clustering cost of the solution, it could greatly deteriorate the performance of many classical clustering algorithms [19, 24]. However, several studies on k-means++ for noisy data seem to suggest that it is an exception and can actually yield quite good solutions [5, 19]. As far as we know, there is no known theoretical analysis that tries to explain the performance of k-means++ on noisy data. The following theorem takes the first step in this direction.

▶ **Theorem 2.** *Given a point set P in a metric space and a parameter $0 < \epsilon \leq 1$, let C be a set of $O(\frac{1}{\epsilon}(k+m))$ points sampled from P using k-means++. Then, C contains a subset of k centers that induces a $(6+\epsilon)$-approximation for k-means with m outliers with constant probability.*

As a corollary to Theorem 2, it is easy to see that $O(\frac{1}{\epsilon}(k+m))^k$ sets of candidate centers for the problem can be generated in $O(n(k+m)\frac{1}{\epsilon})$ time. A $(6+\epsilon)$-approximation can then be obtained by an exhaustive search over the candidate sets.

▶ **Corollary 3.** *Given an instance of the k-means clustering problem with m outliers and a parameter $0 < \epsilon \leq 1$, there is an algorithm that outputs a set C of k centers and a set Z of m outliers, such that $\Phi(P \setminus Z, C) \leq (6+\epsilon)OPT$, where $\Phi(P \setminus Z, C)$ and OPT denote, respectively, the clustering cost induced by the set C and Z and the optimal solution. The algorithm runs in time $O(n(\beta\frac{1}{\epsilon}(k+m))^k)$ for some constant $\beta > 1$.*

1.2 Other Related Work

Most of the aforementioned results are mainly for theoretical purpose. There are also more practical solutions available for clustering. The most popular one for k-means is probably the heuristic technique introduced by Lloyd [27], which iteratively assigns the points to their nearest centers and updates the centers as the means of their corresponding newly generated clusters. It is known that Lloyd's algorithm is sensitive to the locations of the initial centers. An effective remedy for this undesirable issue is the use of an initialization algorithm called

k-means++, which generates an initial set of cluster centers close to the optimal solution. Arthur and Vassilvitskii [7] showed that the k centers generated by k-means++ induce an $O(\log k)$-approximation in an expected sense. Ostrovsky *et al.* [29], Jaiswal and Garg [22], and Agarwal, Jaiswal, and Pal [1] further revealed that these k centers can lead to $O(1)$-approximations under some data separability conditions. Ailon, Jaiswal, and Monteleoni [4] demonstrated that a bi-criteria constant factor approximation can be obtained by sampling $O(k \log k)$ points using k-means++. Aggarwal, Deshpande, and Kannan [2] and Wei [30] later discovered that $O(k)$ points are actually sufficient to ensure a constant factor approximation.

2 Preliminaries

The clustering with outliers problem can be formally defined as follows.

▶ **Definition 4** (k-median/k-means clustering with outliers). *Let P be a set of n points in a metric space $(X, \mathbf{d})$, and $k \geq 1$ and $m \geq 0$ be two non-negative integers. The k-median or k-means clustering problem with outliers is to identify a subset $Z \subseteq P$ of size m and a set $C \subseteq X$ of k centers, such that the clustering cost $\Phi(P \setminus Z, C)$ is minimized among all possible choices of Z and C, where $\Phi(P \setminus Z, C) = \sum_{x \in P \setminus Z} \min_{c \in C} \mathbf{d}(x, c)$ for k-median and $\Phi(P \setminus Z, C) = \sum_{x \in P \setminus Z} \min_{c \in C} \mathbf{d}^2(x, c)$ for k-means.*

In Euclidean space, the clustering with outliers problem is identical, except that the points lie in $\mathbb{R}^d$, and the centers can be k arbitrary points in $\mathbb{R}^d$.

We will use the following result to find the approximate centers, which is known as Chernoff bound.

▶ **Theorem 5** ([28]). *Let $A_1 \dots A_m$ be $0-1$ independent random variables with $Pr(A_i = 1) = p_i$. Let $A = \sum_{i=1}^{m} A_i$ and $u = \sum_{i=1}^{m} E(A_i)$. Let $0 < \alpha < 1$ be an arbitrary real number. Then, $Pr[A \leq (1-\alpha)u] \leq e^{-\frac{\alpha^2 u}{2}}$.*

The following result says that given a cluster $A \subseteq \mathbb{R}^d$, a good approximation to $\Gamma(A)$ can be obtained using a small set of points randomly sampled from A.

▶ **Lemma 6** ([25]). *Given a set R of size $\frac{1}{\lambda^4}$ randomly sampled from a set $A \subseteq \mathbb{R}^d$, where $\lambda > 0$, there exists a procedure* **Construct(R)** *that yields a set of $2^{(1/\epsilon)^{O(1)}}$ points* **core(R)**, *and there exists at least one point $r \in$* **core(R)**, *such that the inequality*

$$\mathbf{d}(r, \Gamma(A)) \leq \lambda \frac{\Delta(A)}{|A|}$$

holds with probability at least $\frac{1}{2}$. The procedure **Construct(R)** *runs in $O(2^{(1/\epsilon)^{O(1)}} \cdot d)$ time.*

3 k-Median Clustering with Outliers in Euclidean Space

In this section, we present a new algorithm for the k-median clustering problem with outliers in the geometric settings. Let $\Phi(x, C) = \min_{c \in C} \mathbf{d}(x, c)$ denote the cost of clustering a point $x \in \mathbb{R}^d$ using a set $C \subseteq \mathbb{R}^d$ of centers. The clustering cost of a point set $P \subseteq \mathbb{R}^d$ induced by C is denoted by $\Phi(P, C) = \sum_{p \in P} \Phi(p, C)$. For a singleton $C = \{c\}$, we also write $\Phi(P, C)$ as $\Phi(P, c)$. The minimum 1-median cost of a set $S \subseteq \mathbb{R}^d$ is denoted by $\Delta(S) = \sum_{x \in S} \mathbf{d}(x, \Gamma(S))$, where $\Gamma(S)$ denotes the optimal center of S.

Algorithm Find-k-centers
Input: a point set P, an integer $k > 0$, and an approximation factor $0 < \epsilon \leq 1$.
Output: a point set $C = \{c_1, \ldots, c_k\}$.

1. let $N = (20k^{10} + 4mk^8)/\epsilon^5$, $M = k^8/\epsilon^4$, $U = \emptyset$;
2. **loop** 2^k times **do**
3.   **Random-sampling(P, k, k, $\emptyset$, ϵ, U)**;
4. **return** the set of centers $C \in U$ with the smallest value of $\Phi(P_m(C), C)$.

Algorithm Random-sampling(Q, g, k, C, ϵ, U)
1. $S = \emptyset$;
2. **if** $g = 0$ **then**
3. $U = U \cup C$;
4.   **return**.
5. sample a set S of size N from Q independently and uniformly;
6. **for** each subset $S' \subseteq S$ of size M **do**
7.   **for** each point $c \in$ **core($\mathbf{S'}$)** **do**
8.   **Random-sampling(Q, $g-1$, k, $C \cup \{c\}$, ϵ, U)**;
9. find the median value β of all values in $\{\Phi(x, C) \mid x \in Q\}$;
10. $Q' = \{x \in Q \mid \Phi(x, C) \leq \beta\}$;
11. **Random-sampling(Q', g, k, C, ϵ, U)**.

Figure 1 The algorithm for k-median with outliers in Euclidean space

3.1 The Algorithm

The general idea of our algorithm solving the k-median clustering problem with outliers is as follows. Assume that $\{P_1, ..., P_k\}$ is the optimal partition of the k-median clustering problem with outliers, where $|P_1| \geq |P_2| \geq \ldots \geq |P_k|$. The objective of our algorithm is to find the approximate centers of $P_i (i = 1, \ldots, k)$. Assume that $P_i (1 \geq i \geq k)$ is the largest cluster whose approximate center has not yet been found. In our algorithm, two strategies are applied to find the approximate center of P_i. It is possible that the points in P_i are far away from the approximate centers already found. For this case, by randomly sampling points in the remaining data set, with large probability, a large portion of P_i is in the sampled set. By enumerating all possible certain size of subsets of the sampled set, there must exist one subset that an approximate center of P_i can be obtained from this set by Lemma 6. On the other hand, if the points in P_i are close to the set of the approximate centers already found (denoted by C), then one center in C can be used to approximate the center of P_i, and the points close to the approximate centers in C can be deleted from the point set. The specific algorithm for the k-median clustering problem with outliers is described in Figure 1. The algorithm Random-sampling has eight parameters $Q, g, k, C, \epsilon, U, N$, and M, where Q is the input dataset, g is the number of centers not yet found, k is the total number of clusters, C is the multi-set of obtained approximate centers, ϵ is a real number $(0 < \epsilon \leq 1)$, U is the collection of candidate solutions, $N = (20k^{10} + 4mk^8)/\epsilon^5$, and $M = k^8/\epsilon^4$.

3.2 Analysis

In this section we show the correctness of Theorem 2. Given an instance of the k-median clustering problem with m outliers (P, k, m), let $Z = \{z_1 \ldots z_m\}$ be the set of outliers in the optimal solution, and $\mathbb{P} = \{P_1 \ldots P_k\}$ be the k-partition of the remaining (inliers) points in P that minimizes the k-median objective function. Without loss of generality, assume that $|P_1| \geq |P_2| \geq \ldots |P_k|$. Denote by $\Delta_k = \sum_{i=1}^{k} \Delta(P_i)$ the clustering cost induced by the optimal solution.

We now give an outline of the proof. In order to prove the correctness of Algorithm Find-k-centers, we need to get that there exists a set of centers in U that achieves the desired approximation for the centers of clusters $P_1, \ldots, P_k$. Assume that a set $C = \{c_1, \ldots, c_{i-1}\}$ of centers has been found. The key point is to prove that the c_i obtained by Algorithm Random-sampling based on C can get a good approximation for P_i. The general idea of proving that c_i is a good approximate center of P_i is as follows. A set B of points that are close to C by a fixed value r can be obtained, where the possible value of r can be enumerated efficiently. The following two cases are discussed: (1) $P_i \cap B \neq \emptyset$, and (2) $P_i \cap B = \emptyset$. For the first case, we show that $\Gamma(P_i)$ is close to a previously sampled point, and there exists a center in C that achieves the desired approximation for $\Gamma(P_i)$. For the second case, we prove that $P \backslash B$ contains a substantial part of P_i. We show that by randomly sampling from $P \backslash B$, a subset of points from P_i can be found, and a good approximate center for P_i is obtained by Lemma 6.

▶ **Lemma 7.** *With a constant probability, there exists a set of approximate centers C^* in the list U generated by the algorithm Find-k-centers, such that for any constant $1 \leq j \leq k$, we have*

$$\mathbf{d}(c_j, \Gamma(P_j)) \leq \frac{\epsilon\Delta(P_j) + 3(j-1)\epsilon\Delta_k}{k^2|P_j|},$$

where c_j denotes the nearest point to $\Gamma(P_j)$ in C^*.

Before proving Lemma 7, we first show its implication. Let C^* denote the center set given in Lemma 7. Given a cluster $P_j \in \mathbb{P}$, we have

$$\begin{aligned}\Phi(P_j, C^*) \leq \Phi(P_j, c_j) &= \sum_{x \in P_j} \mathbf{d}(x, c_j) \leq \sum_{x \in P_j} (\mathbf{d}(x, \Gamma(P_j)) + \mathbf{d}(\Gamma(P_j), c_j) \\ &= \Delta(P_j) + |P_j|\mathbf{d}(c_j, \Gamma(P_j)) \leq \Delta(P_j) + \frac{\epsilon\Delta(P_j) + 3(j-1)\epsilon\Delta_k}{k^2} \\ &\leq \Delta(P_j) + \frac{\epsilon\Delta(P_j)}{k^2} + \frac{3(k-1)\epsilon\Delta_k}{k^2}, \end{aligned} \tag{1}$$

where the third step is due to triangle inequality, and the fifth step follows from the assumption that Lemma 7 is true. Summing both sides of inequality (1) over all $P_j \in \mathbb{P}$, we have

$$\sum_{j=1}^{k} \Phi(P_j, C^*) \leq \Delta_k + \frac{\epsilon\Delta_k}{k^2} + \frac{3(k-1)\epsilon\Delta_k}{k} \leq (1+3\epsilon)\Delta_k. \tag{2}$$

This implies that Lemma 7 is sufficient to ensure the approximation guarantee for our algorithm. We now prove the correctness of Lemma 7.

Proof. (of Lemma 7) We prove the lemma by induction on j. We first consider the case of $j = 1$. Our algorithm initially samples a set of N points from P. Let $S = \{s_1, \ldots, s_N\}$ denote the set of N points sampled from P. Define N random variables $A_1, \ldots, A_N$, such that if $s_i \in P_1$, $A_i = 1$. Otherwise, $A_i = 0$. Since $|P_1| \geq |P_2| \geq \ldots \geq |P_k|$, it is easy to know that for any constant $0 < i \leq N$, we have

$$Pr[A_i = 1] = \frac{|P_1|}{|P|} \geq \frac{|P_1|}{|Z| + k|P_1|} \geq \frac{1}{m+k}.$$

Let $A = \sum_{i=1}^{N} A_i$ and $u = \sum_{i=1}^{N} E(A_i)$. We have $u \geq \frac{N}{m+k} \geq \frac{2k^8}{\epsilon^4}$. Using Lemma 5, we get

$$Pr(A \geq \frac{k^8}{\epsilon^4}) \geq Pr(A \geq \frac{1}{2}u) = 1 - Pr(A \leq \frac{1}{2}u) \geq 1 - e^{-\frac{u}{8}} \geq 1 - e^{-k^8/4\epsilon^4} > \frac{1}{2}.$$

This implies that with probability at least $\frac{1}{2}$, the set of N points sampled from P contains more than $\frac{k^8}{\epsilon^4}$ points from P_i. By feeding $\lambda = \frac{k^2}{\epsilon}$ into Lemma 6, we know that the inequality $\mathbf{d}(c_1, \Gamma(P_1)) \leq \frac{\epsilon\Delta(P_1)}{k^2|P_1|}$ holds with probability at least $\frac{1}{2}$, which implies that Lemma 7 holds for the case $j = 1$.

We now assume that Lemma 7 holds for $j \leq i-1$ and consider the case of $j = i$. Define a multi-set $C^*_{i-1} = \{c_1, \dots c_{i-1}\}$, where c_t is the nearest point to $\Gamma(P_t)$ from C^*_{i-1} for any $1 \leq t \leq i-1$. Define $B_i = \{x \in P \mid \Phi(x, C^*_{i-1}) \leq r_i\}$, where $r_i = \frac{\epsilon\Delta_k}{k^2|P_i|}$. It is easy to see that B_i is a subset of P that consists of the points close to C^*_{i-1}. We distinguish the analysis into the following two cases.

Case (1): $P_i \cap B_i \neq \emptyset$. In this case, P_i contains some points close to C^*_{i-1}. We prove that one center from C^*_{i-1} can be used to approximate $\Gamma(P_i)$.

Case (2): $P_i \cap B_i = \emptyset$. In this case, all the points from P_i are far from the centers in C^*_{i-1}. We prove that P_i contains a substantial part of $P \setminus B$. Thus, a subset of P_i can be randomly sampled from $P \setminus B$ with high probability. By enumerating this subset, a center can be obtained to approximate $\Gamma(P_i)$.

Case (1): $P_i \cap B_i \neq \emptyset$.

Let p be an arbitrary point from $P_i \cap B_i$ and c_f be the nearest point to p in C^*_{i-1}. Let P_f denote the cluster in $\{P_1, \dots, P_{i-1}\}$ such that $\mathbf{d}(c_f, \Gamma(P_f))$ is the smallest value in $\{\mathbf{d}(c_f, \Gamma(P_j)) \mid 1 \leq j \leq i-1\}$. We now prove that c_f can be used to approximate $\Gamma(P_i)$ by triangle inequality and induction assumption. Observe that

$$\begin{aligned}
\mathbf{d}(\Gamma(P_i), c_f) &\leq \mathbf{d}(\Gamma(P_i), p) + \mathbf{d}(p, c_f) \leq \mathbf{d}(\Gamma(P_i), p) + r_i \leq \mathbf{d}(\Gamma(P_f), p) + r_i \\
&\leq \mathbf{d}(\Gamma(P_f), c_f) + \mathbf{d}(c_f, p) + r_i \leq \mathbf{d}(\Gamma(P_f), c_f) + 2r_i \\
&\leq \frac{\epsilon\Delta(P_f) + 3(f-1)\epsilon\Delta_k}{k^2|P_f|} + 2r_i \\
&= \frac{\epsilon\Delta(P_f) + 3(f-1)\epsilon\Delta_k}{k^2|P_f|} + \frac{2\epsilon\Delta_k}{k^2|P_i|}, \qquad (3)
\end{aligned}$$

where the first step and the fourth step are due to triangle inequality, the second step follows from the fact that $p \in B_j$, the third step is derived from the fact that $p \in P_j$, the sixth step follows from the assumption that Lemma 7 holds for $j \leq i-1$, and the last step follows from the definition of r_i. Since $f < i$, we have $|P_f| > |P_i|$. This implies that

$$\begin{aligned}
\frac{\epsilon\Delta(P_f) + 3(f-1)\epsilon\Delta_k}{k^2|P_f|} &= \frac{\epsilon\Delta(P_f)}{k^2|P_f|} + \frac{3(f-1)\epsilon\Delta_k}{k^2|P_f|} \leq \frac{\epsilon\Delta(P_f)}{k^2|P_i|} + \frac{3(f-1)\epsilon\Delta_k}{k^2|P_i|} \\
&\leq \frac{\epsilon\Delta_k}{k^2|P_i|} + \frac{3(i-1)\epsilon\Delta_k}{k^2|P_i|} = \frac{(3i-2)\epsilon\Delta_k}{k^2|P_i|}. \qquad (4)
\end{aligned}$$

Combining inequalities (3) and (4) together, we get

$$\mathbf{d}(\Gamma(P_i), c_f) \leq \frac{(3i-2)\epsilon\Delta_k}{k^2|P_i|} + \frac{2\epsilon\Delta_k}{k^2|P_i|} = \frac{3i\epsilon\Delta_k}{k^2|P_i|} \leq \frac{\epsilon\Delta(P_i) + 3i\epsilon\Delta_k}{k^2|P_i|}.$$

This completes the proof of Lemma 7 in case (1).

Case (2): $P_i \cap B_i = \emptyset$.

For this case, we will show that P_i contains a large fraction of $P \backslash B_i$. Furthermore, algorithm Find-k-centers can find a set Q such that $P \backslash B_i \subseteq Q$ and $|Q| \leq 2|P \backslash B_i|$. Thus, a set S randomly sampled from Q contains a certain number of points from P_i. By enumerating all subsets of S, a subset $S' \subseteq P_i$ of size M, which can be used to find the approximate center for P_i by Lemma 6.

We now show that the proportion of the points of P_i in $P\backslash B_i$ is large. We achieve this by dividing $P\backslash B_i$ into three portions: $Z\backslash B_i$, $\sum_{j=1}^{i-1} P_j \setminus B_j$, and $\sum_{j=i}^{k} P_j\backslash B_i$, and comparing their sizes with $|P_i|$ respectively.

▷ Claim 8. $\frac{|P_i|}{|P\backslash B_i|} \geq \frac{\epsilon}{5k^2+m\epsilon}$.

Proof. It is easy to show that $|Z\backslash B_j| \leq m$. By the definitions of B_i and r_i, we know that $\Phi(P_j, C^*_{i-1}) \geq r_i|P_j\backslash B_i|$ for any $1 \leq j \leq i-1$, which implies that

$$\sum_{j=1}^{i-1} |P_j\backslash B_i| \leq \frac{1}{r_i}\sum_{j=1}^{i-1} \Phi(P_j, C^*_{i-1}) \leq \frac{(1+3\epsilon)\Delta_k}{r_j} = \frac{k^2|P_i|(1+3\epsilon)|}{\epsilon},$$

where the second step is due to our induction assumption and a similar argument in obtaining (2), and he last step is due to the definition of r_i.

By the fact that $|P_1| \geq \ldots \geq |P_k|$, we have $\sum_{j=i}^{k} |P_j\backslash B_i| \leq (k-i)|P_i|$. Thus,

$$\begin{aligned}\frac{|P_i|}{|P\backslash B_i|} &= \frac{|P_i|}{|Z\backslash B_i| + \sum_{j=1}^{i-1}|P_j\backslash B_j| + \sum_{j=i}^{k}|P_j\backslash B_i|}\\ &\geq \frac{|P_i|}{m + \frac{k^2|P_i|(1+3\epsilon)|}{\epsilon} + (k-i)|P_i|}\\ &\geq \frac{1}{m + \frac{k^2(1+3\epsilon)}{\epsilon} + (k-i)} \geq \frac{\epsilon}{5k^2+m\epsilon},\end{aligned} \tag{5}$$

where the last inequality is due to the fact that $0 < \epsilon \leq 1$. ◁

Claim 8 implies that P_i contains a large fraction of $P\backslash B_i$. The algorithm finds the set $P\backslash B_i$ by guessing the number of points from $P\backslash B_i$. Given an integer $1 \leq j \leq \log n$, let β_j denote the $\frac{n}{2^{j-1}}$-th largest value in $\{\Phi(x, C^*_{i-1}) \mid x \in P\}$, and let Q_j denote the set of points $x \in P$ with $\Phi(x, C^*_{i-1}) \leq \beta_j$. We know that there exists a constant l, such that $P\backslash B_i \subseteq Q_l$ and $P\backslash B_i \nsubseteq Q_{l-1}$. It is easy to know that $|P\backslash B_i| \geq \frac{1}{2}|Q_l|$. By Claim 8, we have $\frac{|P_i|}{|Q_l|} \geq \frac{\epsilon}{10k^2+2m\epsilon}$. Using Lemma 5, we know that with probability at least $\frac{1}{2}$, the set of N points randomly sampled from Q_l contains more than $\frac{k^8}{\epsilon^4}$ points from P_j. Using Lemma 6, we can find an approximate center c_i such that $\mathbf{d}(c_i, \Gamma(P_i)) \leq \frac{\epsilon\Delta(P_i)}{k^2|P_i|}$ with probability at least $\frac{1}{2}$. This implies that with probability more than $\frac{1}{2^k}$, Algorithm Random-sampling identifies a set C^* of k centers, such that for any constant $1 \leq j \leq k$, we have

$$\mathbf{d}(c_j, \Gamma(P_j)) \leq \frac{\epsilon\Delta(P_j) + 3(j-1)\epsilon\Delta_k}{k^2|P_j|}.$$

The probability can boosted to a constant by repeatedly running Random-sampling for 2^k times. This completes the proof of Lemma 7. ◀

We are now ready to prove the correctness of Theorem 1.

Theorem 1. Given an instance (P, k, m) of the k-median with m outliers problem, there is a $(1+\epsilon)$-approximation algorithm that runs in $O(nd(\frac{1}{\epsilon}(k+m))^{(\frac{k}{\epsilon})^{O(1)}})$ time.

Proof. Lemma 7 implies that our algorithm gives a $(1+\epsilon)$-approximation for the problem. We focus on the running time of the algorithm. Let $T(n, g)$ be the running time of algorithm Random-sampling on input $(P,\, g,\, k,\, C,\, \epsilon,\, U)$. It is easy to show that $T(n, 0) = O(1)$ and $T(0, g) = 0$. In the algorithm, step 5 takes $(\frac{k+m}{\epsilon})^{O(1)}$ time, step 8 takes $(\frac{k+m}{\epsilon})^{(\frac{k}{\epsilon})^{O(1)}} \cdot d$ time and yield $(\frac{k+m}{\epsilon})^{(\frac{k}{\epsilon})^{O(1)}}$ candidate centers, and step 9 takes $O(ndk)$ time. Thus we get the following recurrence.

$$T(n,g) = (\frac{k+m}{\epsilon})^{O(\frac{k}{\epsilon})} \cdot T(n,g-1) + T(\frac{n}{2},g) + (\frac{k+m}{\epsilon})^{O(1)} + (\frac{k+m}{\epsilon})^{(\frac{k}{\epsilon})^{O(1)}} \cdot d + O(ndk).$$

Choose $\lambda = (\frac{k+m}{\epsilon})^{(\frac{k}{\epsilon})^{O(1)}}$ to be large enough such that

$$T(n,g) \leq \lambda T(n,g-1) + T(\frac{n}{2},g) + \lambda(nd).$$

We claim that $T(n,g) \leq nd\lambda^g \cdot 2^{2g^2}$. This claim holds in the base case. We suppose that the claim holds for $T(n',g')$ $\forall n' < n$, $\forall g' < k$. It is easy to verify that

$$nd\lambda^g \cdot 2^{2g^2} \leq nd\lambda \cdot \lambda^{g-1} \cdot 2^{2(g-1)^2} + \frac{n}{2} d\lambda^g \cdot 2^{2g^2} + \lambda nd,$$

which implies that the claim $T(n,g) \leq nd\lambda^g \cdot 2^{2g^2}$ holds. Thus our algorithm runs in $nd(\frac{1}{\epsilon}(k+m))^{(\frac{k}{\epsilon})^{O(1)}}$ time. ◀

4 k-Means Clustering with Outliers in Metric Space

Our approach for the k-means clustering with m outliers problem first samples a set of $O(\frac{1}{\epsilon}(k+m))$ points with k-means++, enumerates all the subset of size k of the sampled set, and finds the one with the minimal clustering cost. We prove that the subset with minimal clustering cost can achieve $(6+\epsilon)$-approximation to the k-means clustering with m outliers problem. The k-means++ algorithm samples a point with probability proportional to its squared distance to the nearest previously sampled point, as detailed in Figure 2. For t sampled points, the algorithm runs in $O(nt)$ time.

The notations for k-means follows from that of k-median except for a few modifications. We use the squared distances from the points to their corresponding centers to measure the clustering cost. Let $(X, \mathbf{d})$ be a metric space, where $\mathbf{d}$ is the distance function defined over all points in X. Given a point $x \in X$ and a set $C \subseteq X$ of cluster centers, let $\Phi(x, C) = \min_{c \in C} \mathbf{d}(x,c)^2$. Given an instance (P,k,m) of the k-means clustering problem with outliers, let $Z = \{z_1 \ldots z_m\}$ be the set of outliers in the optimal solution, and $\mathbb{P} = \{P_1 \ldots P_k\}$ be the k-partition of the remaining points in P that minimizes the k-means objective function. Given a cluster $P_i \in \mathbb{P}$ and a point c, let $\Gamma(P_i)$ denote its optimal center. The definitions of $\Phi(P_i, C)$, $\Phi(P_i, c)$, and $\Delta(P_i)$ stay no change. Let $\mathbf{b}(P_i, \alpha) = \{x \in P_i \mid \mathbf{d}(x, \Gamma(P_i))^2 \leq \alpha r_i\}$ be the closed ball centered at $\Gamma(P_i)$ of radius αr_i, where $r_i = \frac{\Delta(P_i)}{|P_i|}$.

We first give an outline of the proof of Theorem 3. Given a cluster $P_j \in \mathbb{P}$, it is easy to see that if the value of α is small enough, then any point from $\mathbf{b}(P_j, \alpha)$ can be used to approximate the centroid of P_j. This implies that we can achieve the desired approximation ratio through finding a point from $\mathbf{b}(P_j, \alpha)$ for each cluster $P_j \in \mathbb{P}$. For the the points in P_j, outliers and the set of previously sampled points, there are only two possible relations: either the points in P_j and outliers are far away from the set of previously sampled points, or the points in P_j and outliers are close to the previously sampled points. For the case when the points in P_j and outliers are far away from the set of previously sampled points, by applying k-means++, the points in P_j and outliers can be sampled with high probability, and we prove that $\mathbf{b}(P_j, \alpha)$ contains a substantial portion of the sampled points from P_j. For the case when the points in P_j and outliers are close to the previously sampled points, we prove that the probability of sampling a point from $\mathbf{b}(P_j, \alpha)$ and outliers is small, and a previously sampled point can be used to approximate the centroid of P_j.

The k-means++ algorithm
Input: a point set P and an integer $k > 0$.
Output: a point set $C = \{c_1, \dots, c_k\}$.

1. Sample a point $x \in P$ uniformly at random, initialize C_1 to $\{x\}$;
2. **for** $i = 2$ **to** k **do**:
3. Sample a point $x \in P$ with probability $\frac{\Phi(x,C_i)}{\Phi(P,C_i)}$;
4. $C_i \leftarrow C_{i-1} \cup \{x\}$;
5. $i \leftarrow i + 1$;
6. **return** $C \leftarrow C_i$.

Figure 2 The k-means++ algorithm

Let C_i denote the set of points sampled with k-means++ in the first i iterations. Define $\mathbb{O}_i = \{P_j \in \mathbb{P} \mid cost(P_j, C_i) \leq (6 + \frac{\epsilon}{2})\Delta(P_j)\}$, where $cost(P_j, C_i) = \min_{c \in C_i} \Phi(P_j, c)$. Let T be union of the set of points outside $\mathbb{O}_i$ and Z. The following lemma shows that if the proportion of the cost from the points in T to C_i in $\Phi(P, C_i)$ is small enough, then the points in C_i give the desired approximation for the problem.

▶ **Lemma 9.** *If $\sum_{P_j \in \mathbb{P} \setminus \mathbb{O}_i} \Phi(P_j, C_i) + \Phi(Z, C_i) \leq \frac{\epsilon}{53}\Phi(P, C_i)$, then $\sum_{j=1}^{k} cost(P_j, C_i) \leq (6 + \epsilon)\Delta_k$.*

We now give two useful properties of the closed ball $\mathbf{b}(P_j, \alpha)$. The first property says that any point in such ball is close to $\Gamma(P_j)$, which can be derived from triangle inequality easily. The second property says that the points in the closed ball $\mathbf{b}(P_j, \alpha)$ are quite far from the centers in C_i.

▶ **Lemma 10.** *For any cluster $P_j \in \mathbb{P} \setminus \mathbb{O}_i$, we have*

(i) For any point $c \in \mathbf{b}(P_j, \alpha)$, $\Phi(P_j, c) \leq (2 + 2\alpha)\Delta(P_j)$.

(ii) Let d_j denote the squared distance between $\Gamma(P_j)$ and its nearest point in C_i. Let $\beta = \frac{d_j}{r_j}$ and $1 < \alpha < \beta$. Then $\beta > 2 + \frac{\epsilon}{2}$ and $\frac{\Phi(\mathbf{b}(P_j,\alpha),C_i)}{\Phi(P_j,C_i)} \geq \frac{1}{2(\beta+1)}(4\frac{\sqrt{\beta_j}}{\sqrt{\alpha}} + \beta_j + \ln\alpha - 4\sqrt{\beta_j} - \frac{\beta_j}{\alpha})$.

By feeding $\alpha = 2 + \frac{\epsilon}{4}$ into Lemma 10, we get that any point from $\mathbf{b}(P_j, 2 + \frac{\epsilon}{4})$ can give a $(6 + \frac{\epsilon}{2})$-approximation for the optimal centroid of P_j. Now we show that $\frac{\Phi(\mathbf{b}(P_j,2+\frac{\epsilon}{4}),C_i)}{\Phi(P_j,C_i)}$ is bounded by a constant.

▶ **Lemma 11.** *For any cluster $P_j \in \mathbb{P} \setminus \mathbb{O}_i$, $\frac{\Phi(\mathbf{b}(P_j,2+\frac{\epsilon}{4}),C_i)}{\Phi(P_j,C_i)} \geq \frac{3}{500}$.*

Proof. Define $Q(\alpha, \beta) = \frac{1}{2(\beta+1)}(4\frac{\sqrt{\beta}}{\sqrt{\alpha}} + \beta + \ln\alpha - 4\sqrt{\beta} - \frac{\beta}{\alpha})$. It is easy to verify that $Q(2, \beta)$ increases monotonously with increasing value of β for $\beta \geq 2$. Therefore,

$$\frac{\Delta(C_i, \mathbf{b}(P_j, 2 + \frac{\epsilon}{4}))}{\Delta(C_i, P_j)} \geq \frac{\Delta(C_i, \mathbf{b}(P_j, 2))}{\Delta(C_i, P_j)} \geq Q(2, \beta) > Q(2, 2) > \frac{3}{500},$$

where the first step is derived from the fact that $\mathbf{b}(P_j, 2 + \frac{\epsilon}{4}) \subseteq \mathbf{b}(P_j, 2)$, the second step is due to Lemma 10, and the third step follows from the fact that $\beta_j > 2$, which is derived from Lemma 10. ◀

We now prove the correctness of Theorem 2.

Theorem 2. Given a point set P in a metric space and a parameter $0 < \epsilon \leq 1$, let C be a set of $O(\frac{1}{\epsilon}(k + m))$ points sampled from P using k-means++. Then, C contains a subset of k centers that induces a $(6 + \epsilon)$-approximation for k-means with m outliers with constant probability.

Proof. By Lemma 9, we know that if the current set of the points (sampled with k-means++) does not give the desired approximation ratio, the set of outliers Z or a cluster outside $\mathbb{O}_i$ will be sampled with high probability. In the worst case scenario, we have to pick out k approximate centers for the clusters in $\mathbb{P}$ and all the m outliers.

At each iteration of k-means++, we define a variable A_i. Let $P_l \in \mathbb{P} \setminus \mathbb{O}_i$ be the cluster that maximizes $\Phi(P_l, C_i)$. If $\frac{\Phi(P_l,C_i)}{\Phi(P,C_i)} > \frac{\Phi(Z,C_i)}{\Phi(P,C_i)}$ and the algorithm samples a point from $\mathbf{b}(P_l, 2+\frac{\epsilon}{4})$, or $\frac{\Phi(P_l,C_i)}{\Phi(P,C_i)} \leq \frac{\Phi(Z,C_i)}{\Phi(P,C_i)}$ and the algorithm samples a point from Z, then $A_i = 1$; otherwise, $A_i = 0$. Lemma 10 implies that $E[A_i] \geq \frac{3}{500} \cdot \frac{\epsilon}{53} = \frac{3\epsilon}{26500}$. Let $N = \frac{53000(k+m)}{3\epsilon}$, $A = \sum_{i=1}^{N} A_i$, $u = \sum_{i=1}^{N} E(A_i)$. Using Lemma 5, we have $Pr(A \geq k+m) \geq 1 - Pr(A \leq \frac{1}{2}u) \geq 1 - e^{-k/4} \geq 1 - e^{-1/4}$. This implies that the set of $O(\frac{1}{\epsilon}(k+m))$ points sampled with D^2-sampling contains a subset of k points that induces a $(6+\epsilon)$-approximation with a high constant probability, which completes the proof of Theorem 2. ◀

References

1 Manu Agarwal, Ragesh Jaiswal, and Arindam Pal. k-means++ under approximation stability. *Theoretical Computer Science*, 588:37–51, 2015.

2 Ankit Aggarwal, Amit Deshpande, and Raivi Kannan. Adaptive sampling for k-means clustering. In *Proc. 12nd Int. Workshop and 13rd Int. Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 15–28, 2009.

3 Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and Euclidean k-median by primal-dual algorithms. In *Proc. 58th IEEE Symposium on Foundations of Computer Science*, pages 61–72, 2017.

4 Nir Ailon, Ragesh Jaiswal, and Clairire Monteleoni. Streaming k-means approximation. In *Proc. 23rd Annual Conference on Neural Information Processing Systems*, pages 10–18, 2009.

5 Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J Zaki. Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, 30(11):994–1002, 2009.

6 Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

7 David Arthur and Sergei Vassilvitskii. k-means++: the adavantage of careful seeding. In *Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

8 Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristic for k-median and facility location problems. In *Proc. 33rd Annual ACM Symposium on Theory of Computing*, pages 21–29, 2001.

9 Moses Charikar, Sudipto Guha, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proc. 31st Annual ACM Symposium on Theory of Computing*, pages 1–10, 1999.

10 Moses Charikar, Samir Khuller, David M Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *Proc. 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 642–651, 2001.

11 Ke Chen. On k-median clustering in high dimensions. In *Proc. 17th ACM-SIAM Symposium on Discrete Algorithm*, pages 1177–1185, 2006.

12 Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. In *Proc. 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 8, pages 826–835, 2008.

13 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 *Proc. 57th IEEE Annual Symposium on Foundations of Computer Science*, pages 353–364, 2016.

14 Feldman Dan and Leonard J. Schulman. Data reduction for weighted and outlier-resistant clustering. In *Proc. 31st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1343–1354, 2012.

15 Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proc. 23rd Annual Symposium on Computational Geometry*, pages 11–18, 2007.

16 Zachary Friggstad, Kamyar Khodamoradi, Mohsen Rezapour, and Mohammad R Salavatipour. Approximation schemes for clustering with outliers. In *Proc. 37th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 398–414, 2018.

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

18 Anupam Gupta and Kanat Tangwongsan. Simpler analyses of local search algorithms for facility location. *arXiv:0809.2554*, 2008.

19 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.

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

21 Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proc. 34th Annual ACM Symposium on Theory of Computing*, pages 731–740, 2002.

22 Ragesh Jaiswal and Nitin Garg. Analysis of k-means++ for separable data. In *Proc. 15th Int. Workshop and 16th Int. Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 591–602, 2012.

23 Byrka Jaroslaw, Pensyl Thomas, Rybicki Bartosz, Srinivasan Aravind, and Trinh Khoa. An improved approximation for k-median and positive correlation in budgeted optimization. *ACM Transactions on Algorithms*, 13(2):23, 2017.

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

25 Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2):5:1–5:32, 2010.

26 Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM Journal on Computing*, 45(2):530–547, 2012.

27 Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

28 Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, 1995.

29 Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28:1–28:22, 2013.

30 Dennis Wei. A constant-factor bi-criteria approximation guarantee for k-means++. In *Proc. 30th Annual Conference on Neural Information Processing Systems*, pages 604–612, 2016.