k-means++ under Approximation Stability

Manu Agarwal[1]*, Ragesh Jaiswal[2], and Arindam Pal[3]**

[1] IIT Rajasthan
[2] IIT Delhi
[3] TCS Innovations Lab

Abstract. The Lloyd's algorithm, also known as the k-means algorithm, is one of the most popular algorithms for solving the k-means clustering problem in practice. However, it does not give any performance guarantees. This means that there are datasets on which this algorithm can behave very badly. One reason for poor performance on certain datasets is bad initialization. The following simple sampling based seeding algorithm tends to fix this problem: pick the first center randomly from among the given points and then for $i \geq 2$, pick a point to be the i^{th} center with probability proportional to the squared distance of this point from the previously chosen centers. This algorithm is more popularly known as the k-means++ seeding algorithm and is known to exhibit some nice properties. These have been studied in a number of previous works [AV07,AJM09,ADK09,BR11]. The algorithm tends to perform well when the optimal clusters are *separated* in some sense. This is because the algorithm gives preference to further away points when picking centers. Ostrovsky et al.[ORSS06] discuss one such separation condition on the data. Jaiswal and Garg [JG12] show that if the dataset satisfies the separation condition of [ORSS06], then the sampling algorithm gives a constant approximation with probability $\Omega(1/k)$. Another separation condition that is strictly weaker than [ORSS06] is the approximation stability condition discussed by Balcan et al.[BBG09]. In this work, we show that the sampling algorithm gives a constant approximation with probability $\Omega(1/k)$ if the dataset satisfies the separation condition of [BBG09] and the optimal clusters are not too small. We give a negative result for datasets that have small optimal clusters.

1 Introduction

The k-means clustering problem is defined as follows:

> Given n points $\mathcal{X} = \{x_1, ..., x_n\} \in \mathbb{R}^d$, find k points $\{c_1, ..., c_k\} \in \mathbb{R}^d$ (these are called centers) such that the following objective function is minimized:
>
> $$\phi_{\{c_1,...,c_k\}}(\mathcal{X}) = \sum_{x \in \mathcal{X}} \min_{c \in \{c_1,...,c_k\}} D(x, c)$$

* This work was done while the author was visiting IIT Delhi and was supported by the Summer Research Fellowship programme at IIT Delhi.

** Major part of this work was done while the author was at IIT Delhi.

where $D(x, c)$ denotes the square of the Euclidean distance between points x and c.

Note that the k centers define an implicit clustering of the points in $\mathcal{X}$ as all the points that have the same closest center are in the same cluster. This problem is known to be an NP-hard problem when $k \geq 2$. We can generalize the problem for any distance measure by defining the distance function D accordingly. Such generalized version of the problem is known as the k-median problem with respect to a given distance measure. Here, we will talk about the k-means problem and then generalize our results for the k-median problem with respect to distance measures that are metrics in an approximate sense.

As discussed in the abstract, the most popular algorithm for solving the k-means problem is the Lloyd's algorithm that can be described as follows: (i) Pick k centers arbitrarily (ii) consider the implicit clustering induced by these centers (iii) move the centers to the respective centroids of these induced clusters and then repeat (ii) and (iii) until the solution does not improve. Even though this algorithm works extremely well in practice, it does not have any performance guarantees, the main problem being arbitrary initialization. This means that the algorithm takes a very long time to converge or the final solution is arbitrarily bad compared to the optimal. The following simple sampling algorithm that is more popularly known as the k-means++ seeding algorithm seems to fix the problem to some extent:

(**SampAlg**) Pick the first center uniformly at random from $\mathcal{X}$. Choose a point $x \in \mathcal{X}$ to be the i^{th} center for $i \geq 2$ with probability proportional to the squared distance of x from the nearest previously chosen centers, i.e., with probability $\frac{\min_{c \in \{c_1, \ldots, c_{i-1}\}} D(x,c)}{\phi_{\{c_1, \ldots, c_{i-1}\}}(\mathcal{X})}$.

In this work, we study some properties of this simple sampling algorithm. First, let us look at the previous works.

Previous work The above algorithm, apart from being simple, easy-to-implement, and quick, exhibits some very nice theoretical properties. Arthur and Vassilvitskii [AV07] show that **SampAlg** gives $O(\log k)$ approximation in expectation. They also give an example where the algorithm gives solution with approximation factor $\Omega(\log k)$ in expectation. Ailon et al. [AJM09] and Aggarwal et al. [ADK09] show that this algorithm is a constant factor pseudo-approximation algorithm. This means that **SampAlg** gives a solution that is within a constant factor of the optimal (w.r.t. k centers) if it is allowed to output more than k centers. Brunsch and Röglin [BR11] gave an example where **SampAlg** gives an approximation factor of $(2/3 - \epsilon) \log k$ with probability exponentially small in k thus closing the open question regarding whether the sampling algorithm gives a constant approximation with not-too-small probability. Jaiswal and Garg [JG12] observe that **SampAlg** behaves well for datasets that satisfy the separation condition $\frac{\Delta_{k-1}(\mathcal{X})}{\Delta_k(\mathcal{X})} \geq c$, where $\Delta_i(\mathcal{X})$ denotes the optimal value of the cost for the i-means problem on data $\mathcal{X}$. They show that under this separation condition,

the algorithm gives a constant approximation factor with probability $\Omega(1/k)$. This separation condition was discussed by Ostrovsky et al. [ORSS06] who also observe that **SampAlg** behaves well under such separation and construct a PTAS for the k-means problem using a variant of **SampAlg** in their algorithm. Balcan et al. discuss a strictly weaker separation condition than [ORSS06]. This separation condition has gained prominence and a number of followup works has been done. In this work, we show that **SampAlg** behaves well even under this weaker separation property. Next, we discuss our results in more detail.

Our results Let us first discuss the [BBG09] separation condition. This is known as the $(1+\alpha, \epsilon)$-approximation stability condition.

Definition 1 ($(1+\alpha, \epsilon)$-approximation stability). *Let $\alpha > 0, 1 \geq \epsilon > 0$. Let $\mathcal{X} \in \mathbb{R}^d$ be a point set and let $C_1^*, ..., C_k^*$ denote the optimal k clusters of $\mathcal{X}$ with respect to the k-means objective. $\mathcal{X}$ is said to satisfy $(1+\alpha, \epsilon)$-approximation stability if any $(1+\alpha)$-approximate clustering $C_1, ..., C_k$ is ϵ-close to $C_1^*, ..., C_k^*$. ϵ-closeness means that at most ϵ fraction of points have to be reassigned in $C_1, ..., C_k$ to be able to match $C_1^*, ..., C_k^*$.*

Note that for a fixed value of ϵ, the larger the value of α the stronger is the separation between the optimal clusters. Our techniques easily generalize for large values of α. The above condition captures how stable the optimal clustering is under approximate clustering solutions. This separation condition has been shown to be strictly weaker than the [ORSS06] separation condition. More specifically, it has been shown (see Section 6 in [BBG09] and Lemma 5.1 in [ORSS06]) if a dataset $\mathcal{X}$ satisfies the separation condition $\frac{\Delta_k(\mathcal{X})}{\Delta_{k-1}(\mathcal{X})} \leq \epsilon$, then any near-optimal k-means solution is ϵ-close to the optimal k-means solution. They also give an example that shows that the other direction does not hold.

Main Theorem for k-means The next theorem gives our main result for the k-means problem. Here the distance measure is square of the Euclidean distance.

Theorem 1 (Main Theorem). *Let $0 < \epsilon, \alpha \leq 1$. Let $\mathcal{X} \in \mathbb{R}^d$ be a dataset that satisfies the $(1+\alpha, \epsilon)$-approximation stability and each optimal cluster has size at least $(60\epsilon n/\alpha^2)$. Then the sampling algorithm* **SampAlg** *gives an 8-approximation to the k-means objective with probability $\Omega(1/k)$.*

When $\alpha > 1$, we get the following result.

Theorem 2 (Main Theorem, large α). *Let $0 < \epsilon \leq 1$ and $\alpha > 1$. Let $\mathcal{X} \in \mathbb{R}^d$ be a dataset that satisfies the $(1+\alpha, \epsilon)$-approximation stability and each optimal cluster has size at least $70\epsilon n$. Then the sampling algorithm* **SampAlg** *gives an 8-approximation to the k-means objective with probability $\Omega(1/k)$.*

Generalization to k-median w.r.t. approximate metrics The above result can be generalized for the k-median problem with respect to distance measures that are approximately metric. This means that the distance measure D satisfies the following two properties:

Definition 2 (γ-approximate symmetry). *Let $0 < \gamma \leq 1$. Let $\mathcal{X}$ be some data domain and D be a distance measure with respect to $\mathcal{X}$. D is said to satisfy the γ-approximate symmetry property if the following holds:*

$$\forall x, y \in \mathcal{X}, \gamma \cdot D(y, x) \leq D(x, y) \leq (1/\gamma) \cdot D(y, x). \tag{1}$$

Definition 3 (δ-approximate triangle inequality). *Let $0 < \delta \leq 1$. Let $\mathcal{X}$ be some data domain and D be a distance measure with respect to $\mathcal{X}$. D is said to satisfy the δ-approximate triangle inequality if the following holds:*

$$\forall x, y, z \in \mathcal{X}, D(x, z) \leq (1/\delta) \cdot (D(x, y) + D(y, z)). \tag{2}$$

Here is our main theorem for the general k-median problem.

Theorem 3 (k-median). *Let $0 < \epsilon, \gamma, \delta, \alpha \leq 1$. Consider the k-median problem with respect to a distance measure that satisfies γ-symmetry and δ-approximate triangle inequality. Let $\mathcal{X} \in \mathbb{R}^d$ be a dataset that satisfies the $(1+\alpha, \epsilon)$-approximation stability and each optimal cluster has size at least $(20\epsilon n/\delta^2\alpha^2)$. Then the sampling algorithm* **SampAlg** *gives an $\frac{8}{(\gamma\delta)^2}$-approximation to the k-median objective with probability $\Omega(1/k)$.*

When $\alpha > 1$, we get the following result.

Theorem 4 (k-median, large α). *Let $0 < \epsilon, \gamma, \delta \leq 1$ and $\alpha > 1$. Consider the k-median problem with respect to a distance measure that satisfies γ-symmetry and δ-approximate triangle inequality. Let $\mathcal{X} \in \mathbb{R}^d$ be a dataset that satisfies the $(1 + \alpha, \epsilon)$-approximation stability and each optimal cluster has size at least $(20\epsilon n/\delta^2)$. Then the sampling algorithm* **SampAlg** *gives an $\frac{8}{(\gamma\delta)^2}$-approximation to the k-median objective with probability $\Omega(1/k)$.*

Negative result for small clusters The above two Theorems show that the sampling algorithm behaves well when the data satisfies the Approximation-stability property and the optimal clusters are large. This leaves open the question as to what happens when the clusters are small. The next Theorem shows a negative result if the clusters are small. We show that if the clusters are small, then in the worst case, **SampAlg** gives $O(\log k)$ approximation with probability exponentially small in k.

Theorem 5. *Let $0 < \epsilon, \alpha \leq 1$. Consider the k-means problem. There exists a dataset $\mathcal{X} \in \mathbb{R}^d$ such that the following holds:*

- *$\mathcal{X}$ satisfies the $(1 + \alpha, \epsilon)$ approximation stability property, and*
- **SampAlg** *achieves an approximation factor of $\left(\frac{1}{2} \cdot \log k\right)$ with probability at most $e^{-\sqrt{k}-o(1)}$.*

2 Proof of Theorems 1 and 2

We follow the framework of Jaiswal and Garg [JG12]. We denote the dataset by $\mathcal{X} = \{x_1, ..., x_n\} \in \mathbb{R}^d$. Let $C_1^*, ..., C_k^*$ denote the optimal k clusters with respect to the k-means objective function and let $c_1^*, ..., c_k^*$ denote the centroids of these optimal clusters. We denote the optimal cost with OPT, i.e.,

$$OPT = \sum_{x \in \mathcal{X}} \min_{c \in \{c_1^*, ..., c_k^*\}} D^2(c, x),$$

where $D(.,.)$ denotes the Euclidean distance between any pair of points. For any point $x \in \mathcal{X}$, we denote the distance of this point to the closest center in $\{c_1^*, ..., c_k^*\}$ with $w(x)$ and the distance of this point to the second closest center with $w_2(x)$.

The following Lemma from [BBG09] will be crucial in our analysis.

Lemma 1 (Lemma 4.1 in [BBG09]). *If the dataset satisfies $(1 + \alpha, \epsilon)$-approximation-stability for the k-means objective, then*

(a) If $\forall i, |C_i^*| \geq 2\epsilon n$, then less than ϵn points have $w_2^2(x) - w^2(x) \leq \frac{\alpha \cdot OPT}{\epsilon n}$.
(b) For any $t > 0$, at most $t\epsilon n$ points have $w^2(x) \geq \frac{OPT}{t\epsilon n}$.

Let $c_1, ..., c_i$ denote the centers that are chosen by the first i iterations of **SampAlg** and let $j_1, ..., j_i$ denote the indices of the optimal clusters to which these centers belong, i.e., if $c_p \in C_q^*$, then $j_p = q$. Let $J_i = \{j_1\} \cup ... \cup \{j_i\}$ and let $\bar{J}_i = \{1, ..., k\} \setminus J_i$. So, J_i denotes the clusters that are *covered* and $\bar{J}_i$ denotes the clusters that are not covered by the end of the i^{th} iteration. An optimal cluster being *covered* means that a point has been chosen as a center from the cluster. Let $\mathcal{X}_i = \cup_{j \in J_i} C_j^*$ and let $\bar{\mathcal{X}}_i = \cup_{j \in \bar{J}_i} C_j^*$.

Let B_1 be the subset of points in $\bar{\mathcal{X}}_i$ such that for any point $x \in B_1$, $w_2^2(x) - w^2(x) \leq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n}$. Let B_2 denote the subset of points in $\bar{\mathcal{X}}_i$ such that for every point $x \in B_2$, $w^2(x) \geq \frac{\alpha^2 \cdot \mathrm{OPT}}{6\epsilon n}$. Note that from Lemma 1, we have that $|B_1| \leq \epsilon n$ and $|B_2| \leq 6\epsilon n/\alpha^2$. Let $B = B_1 \cup B_2$ and $\bar{B} = \bar{\mathcal{X}}_i \setminus B$. We have $|B| \leq 7\epsilon n/\alpha^2$.

Lemma 2. *Let $\beta = \frac{1-\alpha/2}{6+\alpha}$. For any $x \in \bar{B}$ we have, we have $D^2(x, c_t) \geq \beta \cdot D^2(x, c_{j_t}^*)$.*

Proof. Let j be the index of the optimal cluster to which x belongs. Note that $w^2(x) = D^2(x, c_j^*)$ and $w_2^2(x) \leq D^2(x, c_{j_t}^*)$. Figure 1 shows this arrangement. For any $x \in \bar{B}$, we have:

$$\begin{aligned} w_2^2(x) - w^2(x) &\geq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n} \geq 6 \cdot w^2(x)/\alpha \\ \Rightarrow w_2^2(x) &\geq (1 + 6/\alpha) \cdot w^2(x) \end{aligned} \tag{3}$$

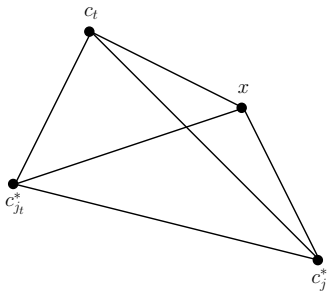

Fig. 1. x belongs to the uncovered cluster j.

We will now argue that $D^2(x, c_t) \geq \beta \cdot D^2(x, c^*_{j_t})$. For the sake of contradiction, assume that $D^2(x, c_t) < \beta \cdot D^2(x, c^*_{j_t})$. Then we observe the following inequalities.

$$
\begin{aligned}
& 2 \cdot D^2(x, c^*_j) + 2 \cdot D^2(x, c_t) \geq D^2(c_t, c^*_j) \quad \text{(triangle inequality)} \\
\Rightarrow\ & 2 \cdot D^2(x, c^*_j) + 2 \cdot D^2(x, c_t) \geq D^2(c_t, c^*_{j_t}) \quad (\text{since } D^2(c_t, c^*_j) \geq D^2(c_t, c^*_{j_t})) \\
\Rightarrow\ & 2 \cdot D^2(x, c^*_j) + 2 \cdot D^2(x, c_t) \geq \frac{1}{2} \cdot D^2(x, c^*_{j_t}) - D^2(x, c_t) \quad \text{(triangle inequality)} \\
\Rightarrow\ & 3 \cdot D^2(x, c_t) \geq \frac{1}{2} \cdot D^2(x, c^*_{j_t}) - 2 \cdot D^2(x, c^*_j) \\
\Rightarrow\ & 3\beta \cdot D^2(x, c^*_{j_t}) > \frac{1}{2} \cdot D^2(x, c^*_{j_t}) - 2 \cdot D^2(x, c^*_j) \\
& \text{(using assumption } D^2(x, c_t) < \beta \cdot D^2(x, c^*_{j_t})) \\
\Rightarrow\ & D^2(x, c^*_j) > \frac{(1 - 6\beta)}{4} \cdot D^2(x, c^*_{j_t}) \\
\Rightarrow\ & w^2(x) > \frac{1}{1 + 6/\alpha} \cdot w_2^2(x) \quad (\text{since } D^2(x, c^*_{j_t}) \geq w_2^2(x) \text{ and } \beta = \tfrac{1-\alpha/2}{6+\alpha})
\end{aligned}
$$

This contradicts with Equation (3). Hence, we get that for any $x \in \bar{B}$ and any $t \in \{1, ..., i\}$, we have $D^2(x, c_t) \geq \beta \cdot D^2(x, c^*_{j_t})$. This proves the Lemma.

Let $W_{min} = \min_{t \in [k]} \left(\sum_{x \in C^*_t, x \in \bar{B}} w_2^2(x) \right)$. Let C_i denote the set of centers $\{c_1, ..., c_i\}$ that are chosen in the first i iterations of **SampAlg**. Let $\mathcal{X}_i = \cup_{t \in J_i} C^*_t$ and $\bar{\mathcal{X}}_i = \mathcal{X} \setminus \mathcal{X}_i$. So, in some sense, $\mathcal{X}_i$ denote the points that are covered by the algorithm after step i and $\bar{\mathcal{X}}_i$ are the uncovered points. For any subset of points $Y \in \mathcal{X}$, $\phi_{C_i}(Y)$ is the cost of the points in Y with respect to the centers C_i, i.e., $\phi_{C_i}(Y) = \sum_{x \in Y} \min_{c \in C_i} D^2(x, c)$. We can now present our next useful lemma which says that the cost of the uncovered points is significant. Note that this implies that the probability of a point being picked from an uncovered clusters in step $(i + 1)$ is significant.

Lemma 3. *Let* $\beta = \frac{1-\alpha/2}{6+\alpha}$. $\phi_{\{c_1, ..., c_i\}}(\bar{\mathcal{X}}_i) \geq \beta \cdot (k - i) \cdot W_{min}$.

Proof. This Lemma follows from the definition of W_{min} and Lemma 2.

We will need a few more definitions. The remaining analysis will be on the lines of a similar analysis in [JG12]. Let E_i denote the event that the set J_i contains i distinct indices from $\{1, ..., k\}$. This means that the first i sampled centers cover i optimal clusters. The next Lemma is from [AV07] and shows that given that event E_i happens, the expected cost of points in $\mathcal{X}_i$ with respect to C_i is at most some constant times the optimal cost of $\mathcal{X}_i$ with respect to $\{c_1^*, ..., c_k^*\}$.

Lemma 4 (Lemma 3.1 and 3.2 in [AV07]). $\forall i, \mathbf{E}[\phi_{\{c_1,...,c_i\}}(\mathcal{X}_i)|E_i] \leq 4 \cdot \phi_{\{c_1^*,...,c_k^*\}}(\mathcal{X}_i)$.

The next Lemma (this is Lemma 4 in [JG12]) shows that the probability that **SampAlg** returns a good solution depends on the probability of the event E_k, i.e., the event that all the clusters get covered.

Lemma 5. $\Pr\left[\phi_{\{c_1,...,c_k\}}(\mathcal{X}) \leq 8 \cdot \phi_{\{c_1^*,...,c_k^*\}}(\mathcal{X})\right] \geq (1/2) \cdot \Pr[E_k]$

Proof. From the previous Lemma, we know that $\mathbf{E}[\phi_{\{c_1,...,c_k\}}(\mathcal{X})|E_k] \leq 4 \cdot \phi_{\{c_1^*,...,c_k^*\}}(\mathcal{X})$. Using Markov, we get that $\Pr[\phi_{\{c_1,...,c_k\}}(\mathcal{X}) > 8 \cdot \phi_{\{c_1^*,...,c_k^*\}}(\mathcal{X})|E_k] \leq 1/2$. Removing the conditioning on E_k, we get the desired Lemma.

We will now argue in the remaining discussion that $\Pr[E_k] \geq 1/k$. This follows from the next Lemma that shows that $\Pr[E_{i+1}|E_i] \geq \frac{k-i}{k-i+1}$.

Lemma 6. $\Pr[E_{i+1}|E_i] \geq \frac{k-i}{k-i+1}$.

Proof. $\mathbf{Pr}[E_{i+1} \mid E_i]$ is just the conditional probability that the $(i+1)^{th}$ center is chosen from the set $\bar{\mathcal{X}}_i$ given that the first i centers are chosen from i different optimal clusters. This probability can be expressed as

$$\mathbf{Pr}[E_{i+1} \mid E_i] = \mathbf{E}\left[\frac{\phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)}{\phi_{\{c_1,...,c_i\}}(\mathcal{X})} \mid E_i\right] \tag{4}$$

For the sake of contradiction, let us assume that

$$\mathbf{E}\left[\frac{\phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)}{\phi_{\{c_1,...,c_i\}}(\mathcal{X})} \mid E_i\right] = \mathbf{Pr}[E_{i+1} \mid E_i] < \frac{k-i}{k-i+1} \tag{5}$$

Applying Jensen's inequality, we get the following:

$$\frac{1}{\mathbf{E}\left[\frac{\phi_{\{c_1,...,c_i\}}(\mathcal{X})}{\phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)} \mid E_i\right]} \leq \mathbf{E}\left[\frac{\phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)}{\phi_{\{c_1,...,c_i\}}(\mathcal{X})} \mid E_i\right] < \frac{k-i}{k-i+1}$$

This gives the following:

$$\begin{aligned} 1 + \frac{1}{k-i} &< \mathbf{E}\left[\frac{\phi_{\{c_1,...,c_i\}}(\mathcal{X})}{\phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)} \mid E_i\right] \\ &= \mathbf{E}\left[\frac{\phi_{\{c_1,...,c_i\}}(\mathcal{X}_i) + \phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)}{\phi_{\{c_1,...,c_i\}}(\bar{\mathcal{X}}_i)} \mid E_i\right] \end{aligned}$$

$$
\begin{aligned}
&= 1 + \mathbf{E}\left[\frac{\phi_{\{c_1,\ldots,c_i\}}(\mathcal{X}_i)}{\phi_{\{c_1,\ldots,c_i\}}(\bar{\mathcal{X}_i})} \mid E_i\right] \\
\Rightarrow \frac{1}{k-i} &\leq \mathbf{E}\left[\frac{\phi_{\{c_1,\ldots,c_i\}}(\mathcal{X}_i)}{\beta\cdot(k-i)\cdot W_{min}} \mid E_i\right] \quad \text{(using Lemma 3)} \\
&\leq \frac{\mathbf{E}[\phi_{\{c_1,\ldots,c_i\}}(X_i) \mid E_i]}{\beta\cdot(k-i)\cdot W_{min}} \\
&\leq \frac{4\cdot\phi_{\{c_1^*,\ldots,c_k^*\}}(\mathcal{X})}{\beta\cdot(k-i)\cdot W_{min}} \quad \text{(using Lemma 4)} \\
\Rightarrow \frac{W_{min}}{\mathrm{OPT}} &\leq \frac{4}{\beta} = 4\cdot\frac{6+\alpha}{1-\alpha/2} \qquad (6)
\end{aligned}
$$

The above gives us an upper bound on W_{min}. Next, we get a lower bound on W_{min} that contradicts with the above bound. Let j be the index of the optimal cluster such that $\sum_{x\in C_j^*, x\in\bar{B}} w_2^2(x)$ is minimized. Note that $W_{min} = \sum_{x\in C_j^*, x\in\bar{B}} w_2^2(x)$. We note that for any $x \notin B_1$, we have $w_2^2(x) - w^2(x) \geq \frac{\alpha\cdot\mathrm{OPT}}{\epsilon n}$. This gives us the following:

$$
\begin{aligned}
&\forall x \notin B_1, x \in C_j^*, w_2^2(x) \geq \frac{\alpha\cdot\mathrm{OPT}}{\epsilon n} \\
\Rightarrow W_{min} &= \sum_{x\in C_j^*, x\in\bar{B}} w_2^2(x) \geq \frac{\alpha\cdot\mathrm{OPT}}{\epsilon n}\cdot\frac{52\epsilon n}{\alpha^2} = \frac{52}{\alpha}\cdot\mathrm{OPT} \qquad (7)
\end{aligned}
$$

The above being true since all clusters are of size at least $\frac{60\epsilon n}{\alpha^2}$. Note that this contradicts with equation (6) since $\alpha \leq 1$.

This concludes the proof of Theorem 1.

Proof (Proof of Theorem 2). We run through the same proof as discussed above with the following quantities redefined as follows: Let B_1 be the subset of points in $\bar{\mathcal{X}_i}$ such that for any point $x \in B_1$, $w_2^2(x) - w^2(x) \leq \frac{\alpha\cdot\mathrm{OPT}}{\epsilon n}$. Let B_2 denote the subset of points in $\bar{\mathcal{X}_i}$ such that for every point $x \in B_2$, $w^2(x) \geq \frac{\mathrm{OPT}}{6\epsilon n}$. Note that from Lemma 1, we have that $|B_1| \leq \epsilon n$ and $|B_2| \leq 6\epsilon n$. Let $B = B_1 \cup B_2$ and $\bar{B} = \bar{\mathcal{X}_i} \setminus B$. We have $|B| \leq 7\epsilon n$. Now, we note that Lemma 3 works for $\beta = \frac{\alpha-1/2}{6+\alpha}$. This changes equation (6) as follows:

$$
\frac{W_{min}}{\mathrm{OPT}} \leq \frac{4}{\beta} = 4\cdot\frac{6+\alpha}{\alpha-1/2} \qquad (8)
$$

Furthermore, equation (7) gets modified to the following:

$$
W_{min} = \sum_{x\in C_j^*, x\in\bar{B}} w_2^2(x) \geq \frac{\alpha\cdot\mathrm{OPT}}{\epsilon n}\cdot(56\epsilon n) = 56\alpha\cdot\mathrm{OPT} \qquad (9)
$$

The above being true since all clusters are of size at least $70\epsilon n$. Note that this contradicts with equation (8) since $\alpha > 1$.

3 Small Cluster

In the previous section, we saw a positive result on datasets that have large optimal clusters. In this section, we show that if the dataset have optimal clusters that are small in size, then **SampAlg** may have a bad behavior. More formally, we will prove Theorem 5 in this Section. We will need the following result from [BR11] for proving this Theorem.

Theorem 6 (Theorem 1 from [BR11]). *Let $r : \mathbb{N} \to \mathbb{R}$ be a real function. If $r(k) = \delta^* \log k$ for a fixed real $\delta^* \in (0, 2/3)$, then there is a class of instances on which* **SampAlg** *achieves an $r(k)$-approximation with probability at most $e^{1-(3/2)\delta^* - o(1)}$.*

Let $\mathcal{X}_{BR}$ denote the dataset on which **SampAlg** gives an approximation factor of $((1/3)\log k')$ with probability at most $e^{-\sqrt{k'}-o(1)}$ when solving the k'-means problem. We will construct another dataset using $\mathcal{X}_{BR}$ and show that **SampAlg** behaves poorly on this dataset. We will need the following fact from [BR11] for our analysis:

Fact 1 ([BR11]) $OPT(k', \mathcal{X}_{BR}) = \frac{k'(k'-1)}{2}$.

Consider the dataset $\mathcal{X} = \mathcal{X}_{far} \cup \mathcal{X}_{BR}$ where $\mathcal{X}_{far}$ has the following properties:

1. $\mathcal{X}_{BR} \cap \mathcal{X}_{far} = \phi$,
2. $|\mathcal{X}_{far}| = |\mathcal{X}_{BR}| \cdot \left(\frac{1}{\epsilon} - 1\right)$.
3. All points in $\mathcal{X}_{far}$ are located at a point c such that the distance of every point $x \in \mathcal{X}_{BR}$ from c is at least $4 \cdot \sqrt{\frac{(1+\alpha)(k-1)(k-2)}{2 \cdot |\mathcal{X}_{far}|}}$.

We solve the k-means problem for $k = k' + 1$ on the dataset $\mathcal{X}$ that has $n = \frac{|\mathcal{X}_{BR}|}{\epsilon}$ points. Note that the size of the smallest optimal cluster for this dataset is of size $\epsilon n/k$. We first observe cost of the optimal solution of $\mathcal{X}$.

Lemma 7. $OPT(k, \mathcal{X}) = k'(k'-1)/2$.

Proof. This is simple using the Fact 1.

We now show that $\mathcal{X}$ has the $(1+\alpha, \epsilon)$-approximation stability property.

Lemma 8. *$\mathcal{X}$ satisfies the $(1+\alpha, \epsilon)$-approximation stability property.*

Proof. Consider any $(1+\alpha)$-approximate solution for the dataset $\mathcal{X}$. Let $c_1, ..., c_k$ be the centers with respect to this approximate solution. We have $\phi_{\{c_1,...,c_k\}}(\mathcal{X}) \leq (1+\alpha) \cdot (k-1)(k-2)/2$. Consider the center in $\{c_1, ..., c_k\}$ that is closest to the point c. Let this center be c_j. Then we note that:

$$D^2(c, c_j) \leq \frac{(1+\alpha)(k-1)(k-2)}{2 \cdot |\mathcal{X}_{far}|}$$

Since the distance of every point in $\mathcal{X}_{BR}$ from point c is at least $4 \cdot \sqrt{\frac{(1+\alpha)(k-1)(k-2)}{2 \cdot |\mathcal{X}_{far}|}}$, we get that all points in $\mathcal{X}_{far}$ are correctly classified. Furthermore, since the number of points in $\mathcal{X}_{BR}$ is at most ϵ fraction of total points, we get that the total number of mis-classified points cannot be more than ϵn and hence the data $\mathcal{X}$ satisfies the $(1+\alpha, \epsilon)$ approximation stability property.

4 Proof of Theorems 3 and 4

Consider the k-median problem with respect to a distance measure $D(.,.)$ that satisfies the γ-symmetry and δ-approximate triangle inequality. The following Lemma is a generalized version of the Lemma in [BBG09] for any given distance measure. The proof remains the same as the proof of Lemma 3.1 in [BBG09].

Lemma 9 (Generalization of Lemma 3.1 in [BBG09]). *If the dataset satisfies $(1+\epsilon, \alpha)$-approximation-stability for the k-median objective, then*

(a) If $\forall i, |C_i^*| \geq 2\epsilon n$, then less than ϵn points have $w_2(x) - w(x) \leq \frac{\alpha \cdot OPT}{\epsilon n}$.
(b) For any $t > 0$, at most $t\epsilon n$ points have $w(x) \geq \frac{OPT}{t\epsilon n}$.

where $w(x)$ denotes the distance of the point x to the closest optimal center as per the distance measure D and $w_2(x)$ is the distance to the second closest center.

We now prove a generalized version of Lemma 2 for distance measures that satisfy γ-symmetry and δ-approximate triangle inequality. We can redefine some of the previous quantities for this case. Let B_1 be the subset of points in $\bar{\mathcal{X}}_i$ such that for any point $x \in B_1$, $w_2(x) - w(x) \leq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n}$. Let B_2 denote the subset of points in $\bar{\mathcal{X}}_i$ such that for every point $x \in B_2$, $w(x) \geq \frac{\delta^2 \alpha^2 \cdot \mathrm{OPT}}{\epsilon n}$. Note that from Lemma 9, we have that $|B_1| \leq \epsilon n$ and $|B_2| \leq \frac{\epsilon n}{\delta^2 \alpha^2}$. Let $B = B_1 \cup B_2$ and we have $|B| \leq \frac{2\epsilon n}{\delta^2 \alpha^2}$. Let $\bar{B} = \bar{\mathcal{X}}_i$.

Lemma 10. *Let $\beta = \frac{\delta^2 + \frac{1}{\alpha} - 1}{(1+\frac{1}{\delta^2 \alpha})(1+\delta)}$. For any $x \in \bar{B}$, we have $D(x, c_t) \geq \beta \cdot D(x, c_{j_t}^*)$.*

Proof. Consider any point $x \in \bar{B}$. Let $x \in C_j^*$. In other words, j is the index of the optimal cluster to which x belongs. Note that $w(x) = D(x, c_j^*)$ and $w_2(x) \leq D(x, c_{j_t}^*)$. Please refer Figure 1 that shows this arrangement. For any $x \in \bar{B}$, we have:

$$\begin{aligned} w_2(x) - w(x) &\geq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n} \geq \frac{1}{\delta^2 \alpha} \cdot w(x) \\ \Rightarrow w_2(x) &\geq \left(1 + \frac{1}{\delta^2 \alpha}\right) \cdot w(x) \end{aligned} \tag{10}$$

We will now argue that $D(x, c_t) \geq \beta \cdot D(x, c_{j_t}^*)$. For the sake of contradiction, assume that $D(x, c_t) < \beta \cdot D(x, c_{j_t}^*)$. Then we observe the following inequalities.

$$\begin{aligned} & D(x, c_t) + D(x, c_j^*) \geq \delta \cdot D(c_t, c_j^*) \\ & \quad (\delta\text{-approximate triangle inequality}) \\ \Rightarrow\ & D(x, c_t) + D(x, c_j^*) \geq \delta \cdot D(c_t, c_{j_t}^*) \\ & \quad (\text{since } D(c_t, c_j^*) \geq D(c_t, c_{j_t}^*)) \\ \Rightarrow\ & D(x, c_t) + D(x, c_j^*) \geq \delta \cdot (\delta \cdot D(x, c_{j_t}^*) - D(x, c_t)) \\ & \quad (\delta\text{-approximate triangle inequality}) \end{aligned}$$

$$
\begin{aligned}
&\Rightarrow (1+\delta)\cdot D(x,c_t) \geq \delta^2 \cdot D(x,c^*_{j_t}) - D(x,c^*_j)\\
&\Rightarrow (1+\delta)\cdot \beta \cdot D(x,c^*_{j_t}) > \delta^2 \cdot D(x,c^*_{j_t}) - D(x,c^*_j)\\
&\quad \text{(using assumption } D(x,c_t) < \beta \cdot D(x,c^*_{j_t}))\\
&\Rightarrow D(x,c^*_j) > (\delta^2 - \beta(1+\delta)) \cdot D(x,c^*_{j_t})\\
&\Rightarrow w(x) > \frac{1}{\left(1+\frac{1}{\delta^2\alpha}\right)} \cdot w_2(x)\\
&\quad \text{(since } D(x,c^*_{j_t}) \geq w_2(x) \text{ and } \beta = \tfrac{\delta^2+\frac{1}{\alpha}-1}{(1+\frac{1}{\delta^2\alpha})(1+\delta)})
\end{aligned}
$$

This contradicts with Equation (10). Hence, we get that for any $x \in \bar{B}$ and any $t \in \{1,...,i\}$, we have $D(x,c_t) \geq \beta \cdot D(x,c^*_{j_t})$. This proves the Lemma.

The rest of the proof remains the same as that for the k-means problem of the previous section. The main difference that arises due to the generalization is that instead of using Lemma 4 we will have to use the following generalized version. This is Lemma 3 in [JG12].

Lemma 11. $\forall i, \mathbf{E}[\phi_{\{c_1,...,c_i\}}(\mathcal{X}_i)|E_i] \leq \frac{4}{(\gamma\delta)^2} \cdot \phi_{\{c^*_1,...,c^*_k\}}(\mathcal{X}_i)$.

So the approximation factor changes from 8 to $8/(\gamma\delta)^2$ due to this generalization. Finally, equation (6) changes as follows:

$$\frac{W_{min}}{\mathrm{OPT}} \leq \frac{4}{\beta} = 4 \cdot \frac{(1+\frac{1}{\delta^2\alpha})(1+\delta)}{\delta^2+\frac{1}{\alpha}-1} \tag{11}$$

Furthermore, equation (7) gets modified to the following:

$$W_{min} = \sum_{x \in C^*_j, x \in \bar{B}} w_2^2(x) \geq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n} \cdot \frac{18\epsilon n}{\delta^2\alpha^2} = \frac{18}{\delta^2\alpha} \cdot \mathrm{OPT} \tag{12}$$

The above being true since all clusters are of size at least $\frac{20\epsilon n}{\delta^2\alpha^2}$. Note that this contradicts with equation (11) since $\alpha \leq 1$.

Proof (Proof of Theorem 4). We run through the same proof as discussed above with the following quantities redefined as follows: Let B_1 be the subset of points in $\bar{\mathcal{X}}_i$ such that for any point $x \in B_1$, $w_2^2(x) - w^2(x) \leq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n}$. Let B_2 denote the subset of points in $\bar{\mathcal{X}}_i$ such that for every point $x \in B_2$, $w^2(x) \geq \frac{\delta^2 \cdot \mathrm{OPT}}{\epsilon n}$. Note that from Lemma 1, we have that $|B_1| \leq \epsilon n$ and $|B_2| \leq \epsilon n/\delta^2$. Let $B = B_1 \cup B_2$ and $\bar{B} = \bar{\mathcal{X}}_i$. We have $|B| \leq 2\epsilon n/\delta^2$. Now, we note that Lemma 3 works for $\beta = \frac{\delta^2+\alpha-1}{(1+\alpha/\delta^2)(1+\delta)}$. This changes equation (6) as follows:

$$\frac{W_{min}}{\mathrm{OPT}} \leq \frac{4}{\beta} = 4 \cdot \frac{(1+\alpha/\delta^2)(1+\delta)}{\delta^2+\alpha-1} \tag{13}$$

Furthermore, equation (7) gets modified to the following:

$$W_{min} = \sum_{x \in C^*_j, x \in \bar{B}} w_2^2(x) \geq \frac{\alpha \cdot \mathrm{OPT}}{\epsilon n} \cdot (18\epsilon n/\delta^2) = \frac{18\alpha}{\delta^2} \cdot \mathrm{OPT} \tag{14}$$

The above being true since all clusters are of size at least $20\epsilon n/\delta^2$. Note that this contradicts with equation (13) since $\alpha > 1$.

5 Acknowledgements

Ragesh Jaiswal would like to thank the anonymous referee of [JG12] for initiating the questions discussed in this paper.

References

[ADK09] Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. *In APPROX-RANDOM*, pages 15–28, 2009.

[AJM09] Nir Ailon, Ragesh Jaiswal and Claire Monteleoni. Streaming k-means approximation. *In Advances in Neural Information Processing Systems (NIPS'09)*, pp. 10–18. 2009.

[AV07] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. *In Proceedings of the 18th annual ACM-SIAM symposium on Discrete Algorithms (SODA'07)*, pp. 1027–1035, 2007.

[ABS10] P. Awasthi, A. Blum, and O. Sheffet. Stability yields a PTAS for k-median and k-means clustering. em In Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS'10), pp. 309–318, 2010.

[BBG09] Maria-Florina Balcan, Avrim Blum and Anupam Gupta. Approximate clustering without the approximation *In Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'09)*, pp. 1068–1077, 2009.

[BR11] Tobias Brunsch and Heiko Röglin. A bad instance for k-means++ *In Proceedings of the 8th annual conference on Theory and applications of models of computation*, pp. 344–352, 2011.

[JG12] Ragesh Jaiswal and Nitin Garg. Analysis of k-means++ for separable data. *In Proceedings of the 16th International Workshop on Randomization and Computation*, pp. 591–602, 2012.

[JKS12] Ragesh Jaiswal, Amit Kumar and Sandeep Sen. A Simple D^2-sampling based PTAS for k-means and other Clustering Problems. *In Proceedings of the 18th Annual International Conference on Computing and Combinatorics*, pp. 13–24, 2012.

[ORSS06] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *In Proceedings of the 47th IEEE FOCS*, pages 165–176, 2006.