# Analysis of $k$-means++ for Separable Data

Ragesh Jaiswal[1] and Nitin Garg[1]

Department of Computer Science and Engineering,
IIT Delhi, New Delhi, India.
{cs5070222 , rjaiswal}@cse.iitd.ac.in.

**Abstract.** $k$-means++ [5] seeding procedure is a simple sampling based algorithm that is used to quickly find $k$ centers which may then be used to start the Lloyd's method. There has been some progress recently on understanding this sampling algorithm. Ostrovsky et al. [10] showed that if the data satisfies the separation condition that $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq c$ ($\Delta_i(P)$ is the optimal cost w.r.t. $i$ centers, $c > 1$ is a constant, and $P$ is the point set), then the sampling algorithm gives an $O(1)$-approximation for the $k$-means problem with probability that is exponentially small in $k$. Here, the distance measure is the squared Euclidean distance. Ackermann and Blömer [2] showed the same result when the distance measure is any $\mu$-similar Bregman divergence. Arthur and Vassilvitskii [5] showed that the $k$-means++ seeding gives an $O(\log k)$ approximation in expectation for the $k$-means problem. They also give an instance where $k$-means++ seeding gives $\Omega(\log k)$ approximation in expectation. However, it was unresolved whether the seeding procedure gives an $O(1)$ approximation with probability $\Omega\left(\frac{1}{poly(k)}\right)$, even when the data satisfies the above-mentioned separation condition. Brunsch and Röglin [8] addressed this question and gave an instances on which $k$-means++ achieves an approximation ratio of $(2/3-\epsilon)\cdot\log k$ only with exponentially small probability. However, the instances that they give satisfy $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} = 1+o(1)$. In this work, we show that the sampling algorithm gives an $O(1)$ approximation with probability $\Omega\left(\frac{1}{k}\right)$ for any $k$-means problem instance where the point set satisfy separation condition $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq 1+\gamma$, for some fixed constant $\gamma$. Our results hold for any distance measure that is a metric in an approximate sense. For point sets that do not satisfy the above separation condition, we show $O(1)$ approximation with probability $\Omega(2^{-2k})$.

## 1 Introduction

The $k$-median problem with respect to a point domain $\mathcal{X}$ and a distance measure $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, is defined as follows:

> Given a set $P \subseteq \mathcal{X}$ of $n$ points, find a subset $C \subseteq \mathcal{X}$ of $k$ points (these are called *centers*) such that the objective function
>
> $$\phi_C(P) = \sum_{p \in P} \min_{c \in C} D(p, c)$$

is minimized. For $\mathcal{X} = \mathbb{R}^d$ and $D(x, y) = ||x - y||^2$, the problem is called the $k$-means problem.

$k$-means++ seeding is a simple sampling algorithm that is used to quickly find $k$ centers that is then used to start the Lloyd's method. This sampling procedure is extremely simple and can be described as follows:

(**SampAlg**) Pick the first center uniformly at random from $P$. Choose a point $p \in P$ to be the $i^{th}$ center for $i > 1$ with probability proportional to the distance of $p$ from the nearest previously chosen centers, i.e., with probability $\frac{\min_{c \in C} D(p,c)}{\phi_C(P)}$.

There has been some recent progress in understanding the above sampling procedure. However, even this simple procedure is not fully understood. There are a number of important questions that are unresolved. Next, we give the current state of understanding and discuss some of the unresolved questions.

*Previous work* The non-uniform sampling technique defined above was first analysed by Ostrovsky et al. [10] for the $k$-means problem. They showed that if the given data is separable in the sense that $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq c > 1$, for some fixed constant $c$, then the sampling algorithm gives an $O(1)$ approximation with probability exponentially small in $k$. After this, Arthur and Vassilvitskii [5] showed that the algorithm gives an $O(\log k)$ approximation in expectation for *any* data set. They also give a problem instance where the algorithm gives an approximation of $\Omega(\log k)$ in expectation. However, for the instance that they construct, the sampling algorithm gives an $O(1)$ approximation with constant probability. The sampling algorithm may be regarded as useful as long as we can show that it gives an $O(1)$ approximation with probability $\Omega\left(\frac{1}{poly(k)}\right)$. This is because we may repeat $O(poly(k))$ times and take the best answer. Some initial progress towards this question was by Aggarwal [3] et al. and Ailon et al. [4] who showed that sampling more than $k$ centers gives an $O(1)$ *pseudo*-approximation with constant probability. However, the basic question whether we can get an $O(1)$ approximation with probability $\Omega\left(\frac{1}{poly(k)}\right)$ remained unresolved. In a recent paper, Brunsch and Röglin [8] gave a problem instance where the sampling algorithm gives a $(2/3 - \epsilon) \cdot \log k$ approximation with probability exponentially small in $k$. This resolves the question for the case when the data is not assumed to be separable in the sense of Ostrovsky et al. [10]. However, the example that they construct satisfies $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \leq 1 + o(1)$ and hence does not satisfy the separability condition in the spirit of Ostrovsky et al.

Most of the above-mentioned results are for the $k$-means problem where the data set consists of points in $\mathbb{R}^d$ and the distance measure is the squared Euclidean distance. There are multiple instances in Machine Learning where the goal is to solve the problem with respect to other distance measures. Some examples include the Kullback-Leibler divergence, Mahalanobis distance, Itakura-Saito divergence. We can ask the same questions for the $k$-median problem with respect

to these distance measures. Ackermann and Blömer [2] analysed the sampling algorithm, **SampAlg**, with respect to a general class of distance measures called the $\mu$-similar Bregman divergences. They show that if the data set satisfies the separation condition in the spirit of Ostrovsky et al., (that is $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq c > 1$), then **SampAlg** gives an $O(1)$-approximation with probability $\Omega(2^{-2k})$.

In our work, we analyse the sampling algorithm for the case that the data is separable, i.e., $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq c$ for some constant $c > 1$. This separability condition has been argued to be reasonable when using $k$-means objective to cluster data since the condition implies that the data is "well-clusterable".

*Our contribution* We show that given a data set that is separable, i.e., $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq 1 + \gamma$, for some constant $\gamma$, **SampAlg** gives an $O(1)$ approximation with probability $\Omega(1/k)$. Our analysis works for the $k$-median problem with respect to *any* distance measure that is a metric in some approximate sense. We will look at some conditions that the distance measure needs to satisfy below.

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

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

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

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

**Definition 3 (Centroid property).** *A distance measure $D$ over space $\mathcal{X}$ is said to satisfy the centroid property if for any subset $P \subseteq \mathcal{X}$ and any point $c \in \mathcal{X}$, we have:*

$$\sum_{p \in P} D(p, c) = \Delta_1(P) + |P| \cdot D(m(P), c),$$

*where $m(P) = \frac{\sum_{p \in P} p}{|P|}$ denotes the mean of the points in $P$. Also, as mentioned earlier, $\Delta_1(P)$ denote the optimal cost with respect to 1 center.*

Note that in the $k$-means problem, $\mathcal{X} = \mathbb{R}^d$ and $D(x, y) = ||x - y||^2$. This distance measure satisfies $\alpha$-approximate symmetry and $\beta$-approximate triangle inequality for $\alpha = 1$ and $\beta = 1/2$. The squared Euclidean distance also satisfies the Centroid property. Note that the squared Euclidean distance is not the only distance measure, used for clustering in practice, that satisfies these properties. *Mahalanobis distance* also satisfies the above properties. A class of distance measures called *Bregman divergences* that are used frequently in Machine Learning is known to satisfy the Centroid property. Furthermore, an important sub-class of Bregman divergences, called $\mu$-similar Bregman divergences, is known to satisfy all of the above properties (see [1] for an overview of Bregman divergences). We can now give our main result using the above definitions:

**Theorem 1 (Main Theorem).** *Let $0 < \alpha, \beta \leq 1$ and $\gamma = \frac{32}{(\alpha\beta)^4}$ be constants. Let $D$ be a distance measure over space $\mathcal{X}$ such that $D$ satisfies $\alpha$-approximate symmetry, $\beta$-approximate trianlge inequality, and the Centroid property. Let $P \subseteq \mathcal{X}$ be any set of $n$ points from the space $\mathcal{X}$ such that the following holds:*

$$\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq 1 + \gamma, \tag{3}$$

*where $\Delta_i(P)$ is defined to be the optimal value of the objective function with $i$ centers, i.e., $\Delta_i(P) = \min_{C,|C|=i} \left[\sum_{p\in P} \min_{c\in C} D(p,c)\right]$. Then* **SampAlg** *gives an $O(1)$-approximation with probability $\Omega(1/k)$.*

We also show that when the data is not given to be separable, then **SampAlg** gives an $O(1)$ approximation with probability $\Omega(2^{-2k})$. Note that this is for any $k$-median instance with respect to any distance measure that satisfy $\alpha$-symmetry and $\beta$-triangle inequality [1]. This is an improvement over the result by Ackermann Blömer [2] who get a similar result though for separable data, i.e. $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq c$ for some fixed constant $c$. We discuss this result in Section 3.

*Techniques* Here is an outline of the proof of our Main Theorem. Let $\{A_1, ..., A_k\}$ denote the points in the optimal clustering. From the Centroid property, we know that the centroids $\{c_1, ..., c_k\}$ of $\{A_1, ..., A_k\}$ are the optimal centers. Let $d_{ij} = D(c_i, c_j)$ and let $T_{min} = \min_{i\neq j} [|A_i| \cdot d_{ij}]$. Let $C'$ denote any set of $i$ points chosen by the first $i$ iterations of the algorithm. Let $j$ be the index of an optimal cluster such that no point in $C'$ belongs to $A_j$. We will first argue that $\phi_{C'}(A_j) \geq d \cdot T_{min}$ (for some constant $d$) by showing that if this were not the case, then the separability condition is violated. Let $X_i$ denote the points in those optimal clusters such that $C'$ has a point from that cluster and let $\bar{X}_i$ denote the remaining points. From the previous argument, we know that $\phi_{C'}(\bar{X}_i) \geq (k-i) \cdot d \cdot T_{min}$. On the other hand, we can argue that the expected cost of the centers $C'$ w.r.t. $X_i$ is at most $d' \cdot \Delta_k(P)$ (for some constant $d'$). Then we show that the probability of picking the $(i+1)^{th}$ point from $\bar{X}_i$ is at least $\frac{k-i}{k-i+1}$. Note that this probability is proportional to $\frac{\phi_{C'}(\bar{X}_i)}{\phi_{C'}(P)}$ and if this were smaller than $\frac{k-i}{k-i+1}$, then $\frac{T_{min}}{\Delta_k(P)} \leq d''$ (for some constant $d''$) but this contradicts with the separability condition. So, the probability that we pick points from each optimal cluster is $\Omega(1/k)$ (using telescoping product). Conditioned on this event, we will argue that the expected cost is at most some constant times the optimal.

We now focus on the proof of our Main Theorem

## 2 Proof of Theorem 1

Let $A_1, ..., A_k$ denote the optimal clusters, i.e., the point set $P$ is partitioned into subsets $A_1, ..., A_k$ such that all points in $A_i$ are in the $i^{th}$ cluster as per

[1] the Centroid property is not required for this result

the optimal $k$-median clustering. Let $C_{OPT} = \{c_1, ..., c_k\}$ be the optimal cluster centers. So, $\forall i \neq j, p \in A_j, D(p, c_i) \geq D(p, c_j)$. For any set of centers $C$, we denote the distance of a point $p$ to it nearest center in $C$ with $D(p, C)$. For any optimal cluster $A_i$, let $r_i = \frac{\sum_{p \in A_i} D(p, c_i)}{|A_i|}$.

We will need the following two basic lemmas. These are generalizations of Lemmas 3.1 and 3.2 in [5].

**Lemma 1.** *Consider any optimal cluster $A_i$. Let $c$ be a point chosen from $A_i$ uniformly at random. Then we have $\mathbf{Exp}[\phi_{\{c\}}(A_i)] \leq \frac{2}{\alpha\beta} \cdot \phi_{\{c_i\}}(A_i)$.*

*Proof.* The expected cost may be written as:

$$\begin{aligned}
\mathbf{Exp}[\phi_{\{c\}}(A_i)] &= \sum_{p \in A_i} \frac{1}{|A_i|} \cdot \sum_{q \in A_i} D(q, p) \\
&\leq \sum_{p \in A_i} \frac{1}{|A_i|} \cdot \sum_{q \in A_i} (1/\beta) \cdot (D(q, c_i) + D(c_i, p)) \\
&\leq \sum_{p \in A_i} \frac{1}{|A_i|} \cdot \sum_{q \in A_i} (1/\beta) \cdot (D(q, c_i) + (1/\alpha) \cdot D(p, c_i)) \\
&= \sum_{p \in A_i} \frac{1}{|A_i|} \cdot \left[ \frac{\phi_{\{c_i\}}(A_i)}{\beta} + \frac{|A_i|}{\alpha\beta} \cdot D(p, c_i)) \right] \leq \frac{2}{\alpha\beta} \cdot \phi_{\{c_i\}}(A_i)
\end{aligned}$$

□

**Lemma 2.** *Let $C$ be any set of centers. Consider any optimal cluster $A_i$. Let $c$ be a center chosen using non-uniform sampling with respect to the set $C$ and let $C' = C \cup \{c\}$. The we have $\mathbf{Exp}[\phi_{C'}(A_i) | c \in A_i] \leq \frac{4}{(\alpha\beta)^2} \cdot \phi_{\{c_i\}}(A_i)$.*

*Proof.* The probability that we choose a point $p \in A_i$ to be $c$, conditioned on the fact that $c$ is chosen from $A_i$ is given by $\frac{D(p,C)}{\sum_{q \in A_i} D(q,C)}$. Once we choose $p$ to be $c$, then any point $q' \in A_i$ contributes $\min(D(q', C), D(q', c))$ to the cost. Using these two observations, we get the following:

$$\mathbf{Exp}[\phi_{C'}(A_i) | c \in A_i] = \sum_{p \in A_i} \frac{D(p, C)}{\sum_{q \in A_i} D(q, C)} \cdot \sum_{q' \in A_i} \min(D(q', C), D(q', p)) \quad (4)$$

From $\beta$-approximate triangle inequality, we have that $D(p, C) \leq (1/\beta) \cdot (D(p, q'') + D(q'', C))$ for all $q'' \in A_i$. So, we have

$$D(p, C) \leq \frac{1}{\beta |A_i|} \cdot \left( \sum_{q'' \in A_i} D(p, q'') + \sum_{q'' \in A_i} D(q'', C) \right) \quad (5)$$

Using above in (4), we get the following:

$$\begin{aligned}
\mathbf{Exp}[\phi_{C'}(A_i)|c \in A_i] \leq & \frac{1}{\beta|A_i|} \cdot \sum_{p \in A_i} \frac{\sum_{q'' \in A_i} D(p,q'')}{\sum_{q \in A_i} D(q,C)} \cdot \sum_{q' \in A_i} D(q',C) + \\
& \frac{1}{\beta|A_i|} \cdot \sum_{p \in A_i} \frac{\sum_{q'' \in A_i} D(q'',C)}{\sum_{q \in A_i} D(q,C)} \cdot \sum_{q' \in A_i} D(q',p) \\
= & \frac{1}{\beta|A_i|} \cdot \sum_{p \in A_i} \sum_{q'' \in A_i} D(p,q'') + \frac{1}{\beta|A_i|} \cdot \sum_{p \in A_i} \sum_{q' \in A_i} D(q',p) \\
\leq & \frac{1}{\beta|A_i|} \cdot \sum_{p \in A_i} \sum_{q'' \in A_i} D(p,q'') + \frac{1}{\alpha\beta|A_i|} \cdot \sum_{p \in A_i} \sum_{q' \in A_i} D(p,q') \quad \text{(using } \alpha\text{-symmetry)} \\
\leq & \frac{2}{\alpha\beta} \cdot \frac{1}{|A_i|} \sum_{p \in A_i} \sum_{q \in A_i} D(p,q) \\
\leq & \frac{2}{\alpha\beta^2} \cdot \frac{1}{|A_i|} \sum_{p \in A_i} \sum_{q \in A_i} (D(p,c_i) + D(c_i,q)) \quad \text{(using } \beta\text{-triangle inequality)} \\
\leq & \frac{2}{(\alpha\beta)^2} \cdot \frac{1}{|A_i|} \sum_{p \in A_i} \sum_{q \in A_i} (D(p,c_i) + D(q,c_i)) \quad \text{(using } \alpha\text{-symmetry)} \\
= & \frac{4}{(\alpha\beta)^2} \cdot \phi_{\{c_i\}}(A_i)
\end{aligned}$$

□

The above lemma says that conditioned on picking the next center from a cluster $A_i$, the expected cost of this cluster with respect to the currently chosen centers is within $O(1)$ factor of the optimal. So, in general once we pick a center from an optimal cluster, there is good chance that we may be able to "forget" about this cluster in the future as we already have a constant approximation with respect to this cluster. The issue might be that the given a current set of centers $C$, the probability of sampling the next center from a given cluster might be very small. We show that if this happens, then the separation condition is violated.

Let $C_i = \{c'_{j_1}, ..., c'_{j_i}\}$ be the centers chosen in the first $i$ steps of the sampling algorithm, where $J_i = \{j_1, ..., j_i\}$ denotes the subset of indices of the optimal cluster to which the centers belongs. Let $X_i = \cup_{j \in J_i} A_j$. Let $E_i$ be the event that $J_i$ contains $i$ distinct indices, i.e., the cardinality of $J_i$ is $i$. We will later show that $\forall i, Pr[E_i] \geq \frac{k-i+1}{k}$.

First, we show that the expected cost of $C_i$ with respect to the point set $X_i$ is at most some constant times the cost of $C_{OPT}$ with respect to $X_i$.

**Lemma 3.** $\forall i, \mathbf{Exp}[\phi_{C_i}(X_i) \mid E_i] \leq \frac{4}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(X_i)$.

*Proof.* The proof follows from Lemmas 1 and 2. □

In the next Lemma, we get a lower bound on the probability that the cost of the solution given by the sampling algorithm is at most some constant times the cost of the optimal solution.

**Lemma 4.** $\mathbf{Pr}\left[\phi_{C_k}(P) \leq \frac{8}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(P)\right] \geq (1/2) \cdot \mathbf{Pr}[E_k]$.

*Proof.* Given that event $E_k$ happens, we have $X_k = P$ and from Lemma 3, we get that $\mathbf{Exp}[\phi_{C_k}(P) \mid E_k] \leq \frac{4}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(P)$. By Markov, we get that

$$\mathbf{Pr}\left[\phi_{C_k}(P) > (8/(\alpha\beta)^2) \cdot \phi_{C_{OPT}}(P) \mid E_k\right] \leq 1/2.$$

Removing the conditioning on $E_k$ we get the desired Lemma. □

Now, all we need to show is that $\mathbf{Pr}[E_k] \geq 1/k$. This trivially follows from Lemma 6 that shows that $\mathbf{Pr}[E_{i+1} \mid E_i] \geq \frac{k-i}{k-i+1}$.

We will need the some additional definitions. Let $\bar{X}_i = P \setminus X_i$. Let $\bar{J}_i = [k] \setminus J_i$. Note that conditioned on $E_i$ happening, $|\bar{J}_i| = k - i$. For any $s \in \bar{J}_i$ let $I_s$ denote the index $t \in J_i$ such that $D(c_s, c'_t)$ is minimized. Let $V_s = D(c_s, c_{I_s})$. We know that

$$D(c'_{I_s}, c_{I_s}) \leq D(c'_{I_s}, c_s) \quad \text{and} \quad V_s \leq (1/\beta) \cdot (D(c_s, c'_{I_s}) + D(c'_{I_s}, c_{I_s}))$$

The first inequality is due to the fact that $c'_{I_s} \in A_{I_s}$ (hence is $c'_{I_s}$ is closer to the center of $A_{I_s}$ than of $A_s$). The above inequality gives us the following:

$$V_s \leq (1/\beta) \cdot (D(c_s, c'_{I_s}) + (1/\alpha) \cdot D(c_s, c'_{I_s})) \leq \frac{2}{\alpha\beta} \cdot D(c_s, c'_{I_s}) \tag{6}$$

Let $T_s = |A_s| \cdot V_s$. Let $T_{min} = \min_{i \neq j} |A_i| \cdot D(c_i, c_j)$. Note that

$$\forall s \in \bar{J}_i, T_s \geq T_{min}. \tag{7}$$

Using the above definitions we can show the following Lemma.

**Lemma 5.** $\phi_{C_i}(\bar{X}_i) \geq (k - i) \cdot \frac{(\alpha\beta)^2}{8} \cdot T_{min}$

*Proof.* For any $s$, let $A_s^{in}$ denote those data points that are closer to the center $c_s$ than any data point that does not belong to $A_s$, i.e.,

$$A_s^{in} = \{p | p \in A_s \text{ and } \forall q \notin A_s, D(p, c_s) \leq D(p, q)\}$$

Let the remaining points in $A_s$ be denoted by $A_s^{out}$, i.e., $A_s^{out} = A_s \setminus A_s^{in}$. Next, we will argue that if the data is separable,i.e., $\Delta_{k-1}(P)/\Delta_k(P) \geq 1 + \gamma$, then $|A_s^{in}| \geq |A_s^{out}|$.

*Claim.* Let $\gamma = \frac{32}{(\alpha\beta)^4}$. If $\Delta_{k-1}(P)/\Delta_k(P) > 1 + \gamma$, then $\forall s, |A_s^{in}| \geq |A_s^{out}|$.

*Proof.* Consider any point $p \in A_s^{out}$. Let $N[p]$ denote the point $\notin A_s$ that is nearest to $p$ and let $I[p]$ denote the index of the cluster to which $N[p]$ belongs. We note that the following inequalities hold:

$$\begin{aligned}
D(p, c_{I[p]}) &\leq \frac{1}{\beta} \cdot \big(D(p, N[p]) + D(N[p], c_{I[p]})\big) \quad \text{(using triangle property)} \\
&\leq \frac{1}{\beta} \cdot \big(D(p, c_s) + D(N[p], c_{I[p]})\big) \quad \text{(since } D(p, N[p]) \leq D(p, c_s)\text{)} \\
&\leq \frac{1}{\beta} \cdot (D(p, c_s) + D(N[p], c_s)) \quad \text{(since } D(N[p], c_{I[p]}) \leq D(N[p], c_s)\text{)} \\
&\leq \frac{1}{\beta} \cdot \left(D(p, c_s) + \frac{1}{\beta} \cdot (D(N[p], p) + D(p, c_s))\right) \quad \text{(using triangle property)} \\
&\leq \frac{1}{\beta} \cdot \left((1 + \frac{1}{\beta}) \cdot D(p, c_s) + \frac{1}{\alpha\beta} \cdot D(p, N[p])\right) \quad \text{(using symmetry property)} \\
&\leq \frac{1}{\beta} \cdot \left((1 + \frac{1}{\beta}) \cdot D(p, c_s) + \frac{1}{\alpha\beta} \cdot D(p, c_s)\right) \quad \text{(since } D(p, N[p]) \leq D(p, c_s)\text{)} \\
&\leq \frac{3}{\alpha\beta^2} \cdot D(p, c_s) \qquad (8)
\end{aligned}$$

For the sake of contradiction, let us assume that $|A_s^{in}| < |A_s^{out}|$. Let $f$ be any one-one function that maps data points in $A_s^{in}$ to data points in $A_s^{out}$.

For any point $p \in A_s^{in}$, the following inequalities hold:

$$\begin{aligned}
D(p, c_{I[f(p)]}) &\leq \frac{1}{\beta} \cdot \big(D(p, f(p)) + D(f(p), c_{I[f(p)]})\big) \quad \text{(using triangle property)} \\
&\leq \frac{1}{\beta} \cdot \left(\frac{1}{\beta} \cdot (D(p, c_s) + D(c_s, f(p))) + D(f(p), c_{I[f(p)]})\right) \quad \text{(using triangle property)} \\
&\leq \frac{1}{\beta} \cdot \left(\frac{1}{\beta} \cdot \left(D(p, c_s) + \frac{1}{\alpha} \cdot D(f(p), c_s)\right) + D(f(p), c_{I[f(p)]})\right) \quad \text{(using symmetry property)} \\
&\leq \frac{1}{\beta} \cdot \left(\frac{1}{\beta} \cdot \left(D(p, c_s) + \frac{1}{\alpha} \cdot D(f(p), c_s)\right) + \frac{3}{\alpha\beta^2} \cdot D(f(p), c_s)\right) \quad \text{(using (8))} \\
&= \left(\frac{1}{\beta^2} \cdot D(p, c_s) + \frac{4}{\alpha\beta^2} \cdot D(f(p), c_s)\right) \qquad (9)
\end{aligned}$$

Using (8) and (9), we get the following:

$$\begin{aligned}
\sum_{p \in A_s^{in}} D(p, c_{I[f(p)]}) + \sum_{p \in A_s^{out}} D(p, c_{I[p]}) &\leq \frac{1}{\beta^2} \cdot \sum_{p \in A_s^{in}} D(p, c_s) + \frac{4}{\alpha\beta^2} \cdot \sum_{p \in A_s^{in}} D(f(p), c_s) + \\
&\qquad \frac{3}{\alpha\beta^2} \cdot \sum_{p \in A_s^{out}} D(p, c_s) \\
&\leq \frac{1}{\beta^2} \cdot \sum_{p \in A_s^{in}} D(p, c_s) + \frac{7}{\alpha\beta^2} \cdot \sum_{p \in A_s^{out}} D(p, c_s) \quad \text{(since } f \text{ is one-one)} \\
&\leq \frac{8}{\alpha\beta^2} \cdot \sum_{p \in A_s} D(p, c_s) = \frac{8}{\alpha\beta^2} \cdot |A_s| \cdot r_s \qquad (10)
\end{aligned}$$

Using 10, we get that

$$\frac{\phi_{\{c_1,\ldots,c_k\}\setminus c_s}(P)}{\phi_{\{c_1,\ldots,c_k\}}(P)} = \frac{\sum_{t\in[k]\setminus s}|A_t|\cdot r_t + \frac{8}{\alpha\beta^2}\cdot|A_s|\cdot r_s}{\sum_{t\in[k]}|A_t|\cdot r_t} \leq \frac{8}{\alpha\beta^2}$$

This contradicts with the fact that $\Delta_{k-1}(P)/\Delta_k(P) \geq 1+\gamma = 1+\frac{32}{(\alpha\beta)^4}$. This concludes the proof of the claim. □

We use the above claim to prove the Lemma. For any $s \in \bar{J}_i$ and $p \in A_s^{in}$ we have

$$\begin{aligned}
&\frac{1}{\beta}\cdot(D(p,C_i)+D(c_s,p)) \geq D(c_s,C_i) \quad \text{(using triangle property)}\\
&\Rightarrow \frac{1}{\beta}\cdot\left(D(p,C_i)+\frac{1}{\alpha}\cdot D(p,c_s)\right) \geq D(c_s,C_i) \quad \text{(using symmetry property)}\\
&\Rightarrow \frac{1}{\beta}\cdot\left(D(p,C_i)+\frac{1}{\alpha}\cdot D(p,C_i)\right) \geq D(c_s,C_i) \quad \text{(using definition of } A_s^{in})\\
&\Rightarrow \frac{2}{\alpha\beta}\cdot D(p,C_i) \geq D(c_s,C_i)\\
&\Rightarrow D(p,C_i) \geq \frac{\alpha\beta}{2}\cdot D(c_s,C_i)\\
&\Rightarrow D(p,C_i) \geq \frac{\alpha\beta}{2}\cdot D(c_s,c'_{I_s})\\
&\Rightarrow D(p,C_i) \geq \frac{(\alpha\beta)^2}{4}\cdot D(c_s,c_{I_s}) \quad \text{(using (6))}\\
&\Rightarrow D(p,C_i) \geq \frac{(\alpha\beta)^2}{4}\cdot V_s
\end{aligned}$$

From this we get the following:

$$\begin{aligned}
&\sum_{p\in A_s^{in}} D(p,C_i) \geq \frac{(\alpha\beta)^2}{4}\cdot\frac{|A_s|}{2}\cdot V_s \quad \text{(since } |A_s^{in}| \geq |A_s|/2 \text{ from previous claim)}\\
&\Rightarrow \sum_{p\in A_s} D(p,C_i) \geq \frac{(\alpha\beta)^2}{8}\cdot T_{min} \quad \text{(using (7))}\\
&\Rightarrow \sum_{s\in\bar{J}_i}\sum_{p\in A_s} D(p,C_i) \geq (k-i)\cdot\frac{(\alpha\beta)^2}{8}\cdot T_{min} \quad \text{(since } |\bar{J}_i| \geq (k-i))\\
&\Rightarrow \phi_{C_i}(\bar{X}_i) \geq (k-i)\cdot\frac{(\alpha\beta)^2}{8}\cdot T_{min}
\end{aligned}$$

This concludes the proof of Lemma 5. □

**Lemma 6.** $\forall i, \mathbf{Pr}[E_{i+1} \mid E_i] \geq \frac{k-i}{k-i+1}$

*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 $\bar{X}_i$ given that the first $i$ centers are chosen from $i$ different optimal clusters. This probability can be expressed as

$$\mathbf{Pr}[E_{i+1} \mid E_i] = \mathbf{Exp}\left[\frac{\phi_{C_i}(\bar{X}_i)}{\phi_{C_i}(P)} \mid E_i\right] \tag{11}$$

For the sake of contradiction, let us assume that

$$\mathbf{Exp}\left[\frac{\phi_{C_i}(\bar{X}_i)}{\phi_{C_i}(P)} \mid E_i\right] = \mathbf{Pr}[E_{i+1} \mid E_i] < \frac{k-i}{k-i+1} \tag{12}$$

Applying Jensen's inequality, we get the following:

$$\frac{1}{\mathbf{Exp}\left[\frac{\phi_{C_i}(P)}{\phi_{C_i}(\bar{X}_i)} \mid E_i\right]} \leq \mathbf{Exp}\left[\frac{\phi_{C_i}(\bar{X}_i)}{\phi_{C_i}(P)} \mid E_i\right] < \frac{k-i}{k-i+1}$$

This gives the following:

$$\begin{aligned}
1 + \frac{1}{k-i} &< \mathbf{Exp}\left[\frac{\phi_{C_i}(P)}{\phi_{C_i}(\bar{X}_i)} \mid E_i\right] \\
&= \mathbf{Exp}\left[\frac{\phi_{C_i}(X_i) + \phi_{C_i}(\bar{X}_i)}{\phi_{C_i}(\bar{X}_i)} \mid E_i\right] \\
&= 1 + \mathbf{Exp}\left[\frac{\phi_{C_i}(X_i)}{\phi_{C_i}(\bar{X}_i)} \mid E_i\right] \\
\Rightarrow \frac{1}{k-i} &\leq \mathbf{Exp}\left[\frac{\phi_{C_i}(X_i)}{\frac{(\alpha\beta)^2}{8} \cdot (k-i) \cdot T_{min}} \mid E_i\right] \quad \text{(using Lemma 5)} \\
&\leq \frac{\mathbf{Exp}[\phi_{C_i}(X_i) \mid E_i]}{\frac{(\alpha\beta)^2}{8} \cdot (k-i) \cdot T_{min}} \\
&\leq \frac{\frac{4}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(P)}{\frac{(\alpha\beta)^2}{8} \cdot (k-i) \cdot T_{min}} \quad \text{(using Lemma 3)} \\
\Rightarrow \frac{T_{min}}{\phi_{C_{OPT}}(P)} &\leq \frac{32}{(\alpha\beta)^4}
\end{aligned} \tag{13}$$

Let $I_{min}$ be the index for which $\min_{j \neq I_{min}}(|A_{I_{min}}| \cdot D(c_{I_{min}}, c_j))$ is minimized. Note that $T_{min} = \min_{j \neq I_{min}}(|A_{I_{min}}| \cdot D(c_{I_{min}}, c_j))$. Consider the set $C' = \cup_{s \neq I_{min}}\{c_s\}$, i.e., all centers except the center of the $I_{min}^{th}$ cluster. We will compute the cost of $C'$ with respect to $P$:

$$\begin{aligned}
\frac{\phi_{C'}(P)}{\phi_{C_{OPT}}(P)} &\leq \frac{\phi_{C_{OPT}}(P) + T_{min}}{\phi_{C_{OPT}}(P)} \quad \text{(using Centroid property)} \\
&\leq 1 + \frac{32}{(\alpha\beta)^4} \quad \text{(using (13))}
\end{aligned}$$

This contradicts with the fact that $P$ satisfies $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} > 1 + \frac{32}{(\alpha\beta)^4}$. □

## 3 Analysis of SampAlg without separation condition

In this section, we will show that **SampAlg** gives an $O(1)$ approximation with probability $\Omega(2^{-2k})$ for any data set. This holds with respect to any distance measure that satisfies the $\alpha$-symmetry and $\beta$-triangle inequality. Note that the Centroid property is not required. This is stated more formally in the next Theorem.

**Theorem 2.** *Let $0 < \alpha, \beta \leq 1$ be constants. Let $D$ be a distance measure over space $\mathcal{X}$ such that $D$ satisfies $\alpha$-approximate symmetry and $\beta$-approximate triangle inequality. Let $P \subseteq \mathcal{X}$ be any set of $n$ points from the space $\mathcal{X}$. Then* **SampAlg** *gives an $O(1)$-approximation with probability $\Omega(2^{-2k})$.*

*Proof.* We will use the definitions and notations from the previous Section. Given a set of centers $C_i$, we say that an optimal cluster $A_j$ is "covered" if there exists a center $c' \in C$ such that $\phi_{\{c'\}}(A_j) \leq \frac{8}{(\alpha\beta)^2} \cdot \phi_{\{c_j\}}(A_j)$. Note that if there is a set of centers $C'$ such that all the optimal clusters are covered, then $\phi_{C'}(P) \leq \frac{8}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(P)$. We will show that, with probability $\Omega(2^{-2k})$, either $C_k$ covers all the optimal clusters or gives a constant approximation. Recall that $C_i$ denotes the set of centers after $i$ centers are picked. Let $R_i$ denote the set of indices of optimal clusters that are covered by $C_i$. Let $Y_i = \cup_{j \in R_i} A_j$ and $\bar{Y}_i = P \setminus Y_i$. The probability that $(i+1)^{th}$ chosen center covers a previously uncovered cluster is given by $\frac{\phi_{C_i}(\bar{Y}_i)}{\phi_{C_i}(P)}$. Suppose that $\frac{\phi_{C_i}(\bar{Y}_i)}{\phi_{C_i}(P)} < 1/2$. This implies that $\phi_{C_i}(\bar{Y}_i) < \phi_{C_i}(Y_i)$. This further implies that

$$\phi_{C_i}(P) = \phi_{C_i}(\bar{Y}_i) + \phi_{C_i}(Y_i) < 2 \cdot \phi_{C_i}(Y_i) \leq \frac{16}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(Y_i) \leq \frac{16}{(\alpha\beta)^2} \cdot \phi_{C_{OPT}}(P).$$

The above basically means that the current set of centers already gives a constant approximation with respect to the entire point set $P$. Choosing more centers will only lower the cost. On the other hand, if $\frac{\phi_{C_i}(\bar{Y}_i)}{\phi_{C_i}(P)} \geq 1/2$, then this implies that with probability at least $1/2$ the $(i+1)^{th}$ center is from one of the uncovered clusters. Conditioned on this, from Lemma 2 we know that with probability at least $1/2$, the newly chosen center covers a previously uncovered cluster. So, with probability at least $1/4$, a new cluster gets covered in step $(i+1)$.

So, either the set of chosen centers $C_k$ gives an approximation factor of $\frac{16}{(\alpha\beta)^2}$ or with probability at least $2^{-2k}$ covers all optimal clusters. The latter implies that $C_k$ gives $\frac{8}{(\alpha\beta)^2}$ approximation. So, in summary, **SampAlg** gives an $\frac{16}{(\alpha\beta)^2}$-approximation with probability at least $2^{-2k}$. □

## 4 Conclusions and Open Problems

In this paper, we have shown that given that the data is separable in the spirit of Ostrovsky et al. [10], i.e., $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} \geq 1 + \gamma_1$ (for some fixed constant $\gamma_1$), then the $k$-means++ based sampling algorithm **SampAlg** gives an $O(1)$ approximation with probability $\Omega(1/k)$. On the other hand, Brunsch and Röglin

[8] gave an instance where **SampAlg** gives $(2/3 - \epsilon) \cdot \log k$ approximation with probability exponentially small in $k$. However, their instance is not separable, i.e., $\frac{\Delta_{k-1}(P)}{\Delta_k(P)} = 1 + \gamma_2$, where $\gamma_2 = o(1)$ and use high dimension. Some interesting open questions are:

- How does **SampAlg** behave when $1 + \gamma_2 \leq \frac{\Delta_{k-1}(P)}{\Delta_k(P)} \leq 1 + \gamma_1$?
- How does **SampAlg** behave for planar $k$-median instances (or in general low dimensional instances)?

The planar (dimension = 2) $k$-means problem was shown to be NP-hard by Mahajan et al. [9]. The lower-bound instances constructed by Arthur and Vassilvitskii [5], Aggarwal et al. [3], and Brunsch and Röglin [8] use high dimension. So, it may be possible that **SampAlg** gives $O(1)$ with high probability for any planar $k$-means instances. Another interesting direction is to explore the behavior of **SampAlg** when the data satisfies $(c, \epsilon)$-closeness property of Balcan et al. [6]. This property was argued to be weaker than the separability condition of Ostrovsky et al. [10].

## References


1. M. R. Ackermann. Algorithms for the Bregman $k$-Median Problem. PhD thesis, University of Paderborn, Department of Computer Science (2009).
2. Marcel R. Ackermann and Johannes Blömer. Bregman Clustering for Separable Instances. In Proceedings of the 12th Scandinavian Symposium and Workshop on Algorithm Theory (SWAT '10), Lecture Notes in Computer Science, vol. 6139, pp. 212-223, Springer, 2010.
3. Ankit Aggarwal, Amit Deshpande, and Ravi Kannan. Adaptive sampling for k-means clustering. In *APPROX-RANDOM*, pages 15–28, 2009.
4. Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni. Streaming k-means approximation. In *Advances in Neural Information Processing Systems 22*, pages 10–18. 2009.
5. David Arthur and Sergei Vassilvitskii. $k$-means++: the advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete Algorithms (SODA'07), pp. 1027–1035, 2007.
6. Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '09), pp. 1068-1077, 2009.
7. A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh. Clustering with Bregman divergences. Journal of Machine Learning Research 6 (October 2005), pp. 1705-1749, 2005.
8. 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 (TAMC'11), pp. 344-352, 2011.
9. Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The Planar k-Means Problem is NP-Hard. In Proceedings of the 3rd International Workshop on Algorithms and Computation (WALCOM '09), pp. 274-285. 2009.
10. Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *Proc. 47th IEEE FOCS*, pages 165–176, 2006.