# On Euclidean $k$-Means Clustering with $\alpha$-Center Proximity

**Amit Deshpande**
Microsoft Research, India

**Anand Louis**
Indian Institute of Science

**Apoorv Singh**
Indian Institute of Science

## Abstract

$k$-means clustering is NP-hard in the worst case but previous work has shown efficient algorithms assuming the optimal $k$-means clusters are *stable* under additive or multiplicative perturbation of data. This has two caveats. First, we do not know how to efficiently verify this property of optimal solutions that are NP-hard to compute in the first place. Second, the stability assumptions required for polynomial time $k$-means algorithms are often unreasonable when compared to the ground-truth clusters in real-world data. A consequence of multiplicative perturbation resilience is *center proximity*, that is, every point is closer to the center of its own cluster than the center of any other cluster, by some multiplicative factor $\alpha > 1$.

We study the problem of minimizing the Euclidean $k$-means objective only over clusterings that satisfy $\alpha$-center proximity. We give a simple algorithm to find the optimal $\alpha$-center-proximal $k$-means clustering in running time exponential in $k$ and $1/(\alpha - 1)$ but linear in the number of points and the dimension. We define an analogous $\alpha$-center proximity condition for outliers, and give similar algorithmic guarantees for $k$-means with outliers and $\alpha$-center proximity. On the hardness side we show that for any $\alpha' > 1$, there exists an $\alpha \leqslant \alpha'$, $(\alpha > 1)$, and an $\varepsilon_0 > 0$ such that minimizing the $k$-means objective over clusterings that satisfy $\alpha$-center proximity is NP-hard to approximate within a multiplicative $(1 + \varepsilon_0)$ factor.

## 1 INTRODUCTION

Popular clustering algorithms optimize objectives such as $k$-means, $k$-median, $k$-center etc. under the implicit assumption that this optimization would recover the *ground truth* clustering. However, the exact optimization of these objectives is NP-hard, and the optimal clusters need not be *balanced* in size or *stable* under small additive or multiplicative perturbation of distances between input points. Previous work by Balcan and Liang (2016), Kumar and Kannan (2010) has shown that assuming the optimal solutions are *balanced* and *stable*, such NP-hard objectives can be optimized exactly in polynomial time . However, there are two caveats. First, when the assumption is about the optimal solution of an NP-hard objective, we do not know how to test this property efficiently for an input instance. Second, the stability assumptions like that of Balcan and Liang (2016), required to get these exact polynomial time clustering algorithms are often unreasonable, when compared to the ground truth in practice, but at the same time, these assumptions cannot be relaxed due to almost-matching NP-hardness results as shown by Ben-David and Reyzin (2014).

Given a set of $n$ points $X = \{x_1, x_2, \ldots, x_n\}$ in a metric space with the underlying metric $\mathrm{dist}(\cdot, \cdot)$, and a positive integer $k$, the $k$-means objective is to find centers $\mu_1, \mu_2, \ldots, \mu_k$ in the given metric space so as to minimize the sum of squared distances of all the points to their nearest centers, respectively, i.e., minimize $\sum_{i=1}^{n} \min_{1 \leqslant j \leqslant k} \mathrm{dist}(x_i, \mu_j)^2$. This results in clusters $C_1, C_2, \ldots, C_k$, where the cluster $C_j$ consists of all the points $x_i$ whose nearest center is $\mu_j$. Aloise et al. (2009); Dasgupta and Freund (2009); Mahajan et al. (2012) showed that optimization of the $k$-means objective is NP-hard in the worst case, even for Euclidean $k$-means with $k = 2$ or $d = 2$ . Euclidean $k$-means is also known to be NP-hard to approximate within some fixed constant $c > 1$, shown by Awasthi et al. (2015). Therefore, it difficult to efficiently verify any non-trivial property such as *stability* of the optimal $k$-means solution.

For a given $\alpha > 1$, an instance of $k$-means clustering is called *$\alpha$-multiplicative perturbation resilient*, if perturbing the inter-point distances within a multiplicative factor of $\alpha$ does not change the optimal $k$-means clusters. Awasthi et al. (2012) showed that for center-based objectives like $k$-means and $k$-median, 3-multiplicative perturbation resilient instances can be solved exactly in polynomial time. Angelidakis et al. (2017) improved this result to 2-multiplicative

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

perturbation resilient instances. Angelidakis et al. (2017) show that $\alpha$-multiplicative perturbation resilience implies a geometric property called $\alpha$-center proximity, which they exploit in their algorithm. Independently, recent work by Friggstad et al. (2018) gives a local-search algorithm for solving $\alpha$-multiplicative perturbation resilient instances of discrete $k$-means; in these instances, the optimal $k$-means solution satisfies $\alpha$-center-proximity. The algorithm due to Friggstad et al. has running time $O\left(n^{d^{O(d)}(\alpha-1)^{-O(d/(\alpha-1))}}k\right)$, where the polynomial in $n$ has a large exponent that is itself exponential in the dimension $d$.

**Definition 1.1** ($\alpha$-Center Proximity)**.** Let $C_1, C_2, \ldots, C_k$ be a clustering of $X$ with the centers $\mu_1, \mu_2, \ldots, \mu_k$ and the underlying metric $\text{dist}(\cdot, \cdot)$. We say that the clustering $C_1, C_2, \ldots, C_k$ of $X$ satisfies $\alpha$-center proximity if for all $i \neq j$ and $x \in C_i$, we have $\text{dist}(x, \mu_j) > \alpha\ \text{dist}(x, \mu_i)$. We say that a clustering with its corresponding centers is *$\alpha$-center proximal* if it satisfies the $\alpha$-center proximity property.

Ben-David and Reyzin (2014) point out that when $\alpha$ is not close to 1, $\alpha$-multiplicative perturbation resilience is an unreasonable assumption compared to the ground truth in practice. At the same time, hardness results shown by Ben-David and Reyzin (2014); Balcan et al. (2016) are a barrier in obtaining polynomial time algorithms even for $\alpha < 2$. Ackerman and Ben-David (2009) introduce the notion of additive perturbation resilience. Assuming that the optimal $k$-means clusters do not change when the inter-point distances are perturbed within an additive $\varepsilon\ \text{diam}\,(X)$, their algorithm computes the optimal $k$-means solution in $n^{O(k/\varepsilon^2)}$ time. The dependence on diameter is undesirable because it can grow arbitrarily even in the presence of a single outlier.

To avoid any dependence on diameter, Vijayaraghavan et al. (2017) extend the definition of $\varepsilon$-additive perturbation resilience to assume that the optimal $k$-means clusters do not change if we move the points by at most $\varepsilon \cdot \max_{i,j} \left\|\mu_i - \mu_j\right\|$. For these instances, they show a geometric property called *angular separation* between the optimal clusters, and use a modified perceptron algorithm to solve 2-means optimally on these instances in time $dn^{\text{poly}(1/\varepsilon)}$. We observe (Section 2 and Figure 1) that $\alpha$-multiplicative perturbation resilience for Euclidean $k$-means instances implies a similar *angular separation*. For $k$-means on these instances, the running time of Vijayaraghavan et al. (2017) is $dn^{O(k^2/\varepsilon^8)}$. To get a faster algorithm, they define a stronger notion of stability called $(\rho, \Delta, \varepsilon)$-separation, where in addition to the *angular separation* there is a margin of $\rho$ between any pair of clusters. Their algorithm optimally solves any $(\rho, \Delta, \varepsilon)$-separated instance of Euclidean $k$-means with $\beta$-balanced clusters in time $\tilde{O}(n^2kd)$, whenever $\rho = \Omega(\Delta/\varepsilon^2 + \beta\Delta/\varepsilon)$. They also show that their algorithm is robust to outliers as long as the fraction $\eta$ of outliers satisfies the following equation

$$\rho = \Omega\left(\frac{\Delta}{\varepsilon^2}\left(\frac{w_{\max} + \eta}{w_{\min} - \eta}\right)\right),$$

where $w_{\max}$ and $w_{\min}$ are the fraction of points in the largest and the smallest optimal cluster, respectively.

Given any set of cluster centers, the outliers are the last few points when we order all the points in non-decreasing order of their distances to the nearest centers, respectively. For the set of outliers to be unambiguous and stable under small perturbations to the input data, intuitively one needs a multiplicative gap between the distances of inliers and outliers to their respective centers. So we define an analogous center proximity property for clustering with outliers as follows.

**Definition 1.2** ($\alpha$-Center Proximity with Outliers)**.** Consider a $k$-means instance on the set of points $X$ with underlying metric $\text{dist}(\cdot, \cdot)$, and an integer parameter $z$. We define the distance of a point $x$ to a set or a tuple $(\mu_1, \ldots, \mu_n)$ as $\min_i \|x - \mu_i\|$. Given any centers $\mu_1, \mu_2, \ldots, \mu_k$, let $Z \subseteq X$ be the subset of the farthest $z$ points in $X$ based on their distances to the nearest center, and let $C_1, C_2, \ldots, C_k$ be the clustering of $X \setminus Z$ where $C_j$ consists of the points in $X \setminus Z$ that have $\mu_j$ as their nearest center. Such a clustering $C_1, C_2, \ldots, C_k$ of $X \setminus Z$ with $Z$ as outliers satisfies $\alpha$-center proximity, if for all $i \neq j$ and $x \in C_i$, we have $\text{dist}(x, \mu_j) > \alpha\ \text{dist}(x, \mu_i)$, and moreover, for all $i, j \in [k]$, $x \in C_i$ and $y \in Z$ we have $\text{dist}(y, \mu_j) > \alpha\ \text{dist}(x, \mu_i)$.

Center proximity is a desirable property for the output of any clustering algorithm used in practice. Balanced clusters of size $\Omega(n/k)$ is another desirable property to avoid small, meaningless clusters. Motivated by this, we study the problem of minimizing the $k$-means objective for given center-proximity and balance parameters.

There are many relevant work on stability including Ben-David (2015) that we discuss in the supplementary material due to the space constraints.

### 1.1 Our results

Unlike previous work, we do not assume that an optimal solution for the $k$-means objective on the given input satisfies $\alpha$-center proximity. There is no easy way to algorithmically verify this promise. In fact, as we show later in this paper, there are instances where the optimal $k$-means solutions satisfy $\alpha$-center proximity, for a small constant $\alpha > 1$, but the $k$-means problem still remains NP-hard. Therefore, we define our problem as finding a clustering of the smallest $k$-means cost among all the clusterings that satisfy $\alpha$-center proximity.

We show that there exists a constant $c > 1$ such that $k$-means remains NP-hard to approximate within a multiplicative factor $c$, even on the instances where an optimal $k$-means solution satisfies $\alpha$-center proximity and is balanced, i.e., each optimal cluster has size $\Omega(n/k)$. Moreover, for $\alpha$ close to 1, there may not be a unique optimal $k$-means solution that satisfies $\alpha$-center proximity and is

balanced. In fact, given any $\alpha > 1$, there exists an instance with $2^{\Omega(k/(\alpha-1))}$ such optimal $k$-means solutions that satisfy $\alpha$-center proximity and are balanced.

For any $\alpha > 1$, we show an interesting geometric property (Proposition 2.2 and Figure 1) for clusterings that satisfy $\alpha$-center proximity, namely, any pair of disjoint clusters must lie inside two disjoint balls. The centers of these balls need not be at the means of the clusters, allowing the clusters to be arbitrarily large (see Figure 1). The degenerate case for $\alpha = 1$ is two balls of infinite radii touching at their separating hyperplane. We note that a similar geometric property in a different context was observed by Telgarsky and Vattani (2010). Telgarsky and Vattani (2010) defined a version of Harting's method, which does updates based on the value $\alpha$. The value of $\alpha$ in their case is defined in terms of the size of the clusters. Although in a different context, the geometric insight they achieved is the same as in Proposition 2.2.

We show the following algorithmic result for minimizing the $k$-means objective over the clusterings that are balanced and satisfy $\alpha$-center proximity.

**Theorem 1.3.** *For any $\alpha > 1$, a balance parameter $\omega > 0$, and given any set of $n$ points in $\mathbb{R}^d$ , we can* exactly *find a clustering of the least $k$-means cost among all solutions that satisfy $\alpha$-center proximity and are balanced, i.e., each cluster has size at least $\omega n/k$. Our algorithm finds such an optimal clustering in time $O(2^{poly(k/\omega(\alpha-1))}\, nd)$, with constant probability.*

**Remark 1.4.** Our algorithm as stated requires $\alpha$ as an input. In practice, when the value of $\alpha$ is not available, we can invoke our algorithm with a sequence of decreasing values of $\alpha$ till a "satisfactory" clustering is found.

Since $k$-means is hard to approximate within some fixed constant $c > 1$, even on instances where the optimal solutions are balanced and satisfy $\alpha$-center proximity, the exponential running time in our algorithm is unavoidable. We show the following hardness result:

**Theorem 1.5.** *For any $2 > \alpha' > 1$ there exists an $\alpha \leqslant \alpha'$, $(\alpha > 1)$, constants $\varepsilon > 0$, and $\omega > 0$, such that it is NP-hard to approximate the optimal $\alpha$-center proximal Euclidean $k$-means, where the size of each cluster is at least $\omega n/k$, to a factor better than $(1 + \varepsilon)$.*

The running time of our algorithm is exponential only in the number of clusters $k$, the balance parameter $\omega$ and the center proximity parameter $\alpha$ but it is linear in the number of points $n$ and the dimension $d$.

We show a similar exact algorithm for minimizing the $k$-means objective with $z$ outliers, where the minimization is only over clusterings that satisfy the $\alpha$-center proximity with outliers and are balanced.

**Theorem 1.6.** *For any $\alpha > 1$, a balance parameter $\omega > 0$, given any set of $n$ points in $\mathbb{R}^d$ and an outlier parameter $z \in [n]$, we can* exactly *find a clustering of the least $k$-means cost among all solutions that satisfy $\alpha$-center proximity with $z$ outliers and are balanced, i.e., each cluster has size at least $\omega n/k$. Our algorithm finds such an optimal clustering in time $O(2^{poly(k/\omega(\alpha-1))}\, nd)$, with constant probability.*

In fact, Theorem 1.3 and Theorem 1.6 hold for any clustering objective as long as the centers used to define $\alpha$-center proximity are the *means* or *centroids* of the clusters.

In the case when most points satisfy $\alpha$-center proximity and form balanced clusters, we show an algorithm which outputs a list of clusterings, such that one of the clusterings corresponds to the case where the points which satisfy $\alpha$-perturbation resilience are correctly clustered.

**Theorem 1.7.** *For any $\alpha > 1$, a balance parameter $\omega > 0$, given any set of $n$ points in $\mathbb{R}^d$ and a parameter $\beta \in [n]$, we can output a list, of $k$-means clustering, of size $O\left(2^{poly(k/\omega(\alpha-1))}\right)$ such that one of them is the minimum cost clustering among all solutions that satisfy $\alpha$-center proximity without $\beta$ points and are balanced, i.e., each cluster has size at least $\omega n/k$. Our algorithm finds such an optimal clustering in time $O(2^{poly(k/\omega(\alpha-1))}\, nd)$, with constant probability.*

We also show exact algorithm for minimizing the $k$-means objective over clustering that satisfy $\alpha$-center proximity but no balance requirement. However, the running time of our algorithm depends exponentially on the ratio of the distances between the farthest and the closest pair of means.

**Theorem 1.8.** *For any $\alpha > 1$ and given any set of $n$ points in $\mathbb{R}^d$, and a parameter $\gamma$, where*

$$\gamma \geqslant \frac{\max_{i,j} \|\mu_i - \mu_j\|}{\min_{i \neq j} \|\mu_i - \mu_j\|},$$

*for the means $\mu_1, \mu_2, \ldots, \mu_k$ of the optimal solution, we can* exactly *find a clustering of the least $k$-means cost among all solutions that satisfy $\alpha$-center proximity. Our algorithm finds such an optimal clustering in time $O(2^{poly(k\gamma/(\alpha-1))}\, nd)$, with constant probability.*

We also show that the optimal $\alpha$-center proximal clustering with balanced cluster need not be unique. In fact, the number of possible optimal $\alpha$-center proximal clustering with balanced cluster can be exponential in $k$ and $1/(\alpha - 1)$. We show the following result:

**Proposition 1.9.** *For any $2 > \alpha' > 1$, and any $k \in \mathbb{Z}$, there exists $\alpha \leqslant \alpha'$ (and $\alpha > 1$), $n, d$ and a set of points $X \in \mathbb{R}^d$ such that such that the number of possible optimal $\alpha$-center proximal clusterings, where the size of each cluster is $n/k$ $(\omega = 1)$ is $2^{\tilde{\Omega}(k/(\alpha^2-1))}$.*

### 1.2 Notation

We use $d \in \mathbb{Z}$ to denote the dimension of the ambient space. For a vector $v \in \mathbb{R}^d$, we use $\|v\|$ to denote its Euclidean

norm. We use $x_1, \ldots, x_n \in \mathbb{R}^d$ to denote the points in the instance, and we use $X$ to denote the set of these points. For any set of points $S$, we define its *diameter* as $\mathsf{diam}(S) \stackrel{\text{def}}{=} \max_{u,v \in S} \|u - v\|$.

## 2 GEOMETRIC PROPERTIES OF $\alpha$-CENTER PROXIMAL INSTANCES

We will assume Euclidean metric throughout the paper.

**Definition 2.1.** Suppose there exists a clustering $C_1, ..., C_k$ of $X$ that is $\alpha$-center proximal. Let $C_i$ and $C_j$ be two clusters with $\mu_i$ and $\mu_j$ as their respective means. We define the vector $\hat{\mu}_{i,j} \stackrel{\text{def}}{=} \frac{\alpha^2\mu_i - \mu_j}{\alpha^2 - 1}$ and $r_{i,j} \stackrel{\text{def}}{=} \frac{\alpha}{\alpha^2-1}\left\|\mu_i - \mu_j\right\|$. Let $C_{i,j}$ denote the ball centered at $\hat{\mu}_{i,j}$ with radius $r_{i,j}$. Let $p_{i,j} \stackrel{\text{def}}{=} (\hat{\mu}_{i,j} + \hat{\mu}_{j,i})/2$ and $D_{i,j} \stackrel{\text{def}}{=} \left\|\hat{\mu}_{i,j} - \hat{\mu}_{j,i}\right\|$. Let $u \stackrel{\text{def}}{=} \frac{(\hat{\mu}_{i,j} - \hat{\mu}_{j,i})}{\|\hat{\mu}_{i,j} - \hat{\mu}_{j,i}\|}$ denote the unit vector along the line joining $\hat{\mu}_{i,j}$ and $\hat{\mu}_{j,i}$, and let $d_{i,j}$ denote the distance between the closest points in $C_{i,j}$ and $C_{j,i}$.

We note that Balcan and Liang (2016) show in their Lemma 3.3 that for a pair of clusters $C_i, C_j$ in a $\alpha$-center proximal clustering, and $\alpha \geqslant 1 + \sqrt{2}$, the clusters are contained in disjoint balls around their centers $\mu_i$ and $\mu_j$ respectively. We show in the following proposition that for any $\alpha > 1$, and for a pair of clusters $C_i, C_j$ in a $\alpha$-center proximal clustering, the points in the clusters are contained in disjoint balls of radius $r_{ij}$ and $r_{ji}$ centered around $\hat{\mu}_{i,j}$ and $\hat{\mu}_{j,i}$ respectively. See Figure 1 and the supplementary material for more details.

**Proposition 2.2** (Geometric implication of $\alpha$-center proximity property)**.** *Let $X$ satisfy the $\alpha$-center proximity property (for any $\alpha > 1$) and let $C_i$ and $C_j$ be two clusters in its optimal solution. Any point $x \in C_i$, satisfies*

$$\forall x \in C_i, \left\|x - \hat{\mu}_{i,j}\right\| < r_{i,j}\,. \tag{1}$$

**Remark 2.3.** The line joining the respective centers of the ball passes through the mean of the clusters as well. Refer to the Figure 1 for getting an insight into the geometric structure defined by the equation 1. We refer the reader to figure 1a of Vijayaraghavan et al. (2017) for insight into the geometry of $\varepsilon$-additive perturbation resilient instance, and to notice the similarity between the geometric structures of instances satisfying the two stability properties.

### 2.1 Estimating Means Suffices for Cluster Recovery

In this section we assume that we can get access to a set of points $\{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}$ such that, $\|\tilde{\mu}_i - \mu_i\| \leqslant 2\delta r_{i,j}$ for all $j \in [k]$ and for each $i \in [k]$, with probability some constant probability. We will show how to obtain such points in Section 3.

**Algorithm 1:** $\alpha$-center proximal $k$-means clustering with balanced clusters

**Input:** a list $\mathcal{L}$ of $k$-tuples, where there exists at least one tuple $(\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ such that for any $i \in [k]$, $\|\tilde{\mu}_i - \mu_i\| \leqslant 2\delta r_{i,j}$ for all $j \in [k]$, a number $\alpha > 1$, $\omega > 0$

**Output:** An $\alpha$-center proximal clustering of minimum cost where the size of each cluster is at least $\omega n/k$.

```
Initialize (μ̃_1^(p), μ̃_2^(p), ..., μ̃_k^(p)) ← (0̄, 0̄, ..., 0̄)
cost ← Σ_{i=1}^n min_{1⩽j⩽k} ||x_i − μ̃_j^(0)||^2
for a tuple μ̃^(q) ∈ L do
    for j = 1 to k do
        C_j^(q) = {x_i : μ̃_j^(q) is the nearest center to x_i}
    end for
    if Σ_{i=1}^n min_{1⩽j⩽k} ||x_i − μ̃_j^(q)||^2 < cost then
        if C_1^(q), ..., C_k^(q) all have size at least ωn/k
         and satisfy α-center proximity then
            (μ̃_1^(p), μ̃_2^(p), ..., μ̃_k^(p)) ← (μ̃_1^(q), μ̃_2^(q), ..., μ̃_k^(q))
            cost = Σ_{i=1}^n min_{1⩽j⩽k} ||x_i − μ̃_j^(q)||^2
        end if
    end if
end for
for i = 1 to n do
    Label(i) ← argmin_{1⩽j⩽k} ||x_i − μ̃_j^(p)||
end for
```

The following proposition shows that to cluster a point to their desired cluster, an additive approximation to the center also works (deciding based on the proximity of the data points to the approximate center).

**Proposition 2.4.** *Suppose we have a set of points $\{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}$ such that for any $i \in [k]$, $\|\tilde{\mu}_i - \mu_i\| \leqslant 2\delta r_{i,j}$ for all $j \in [k]$, then we can find the optimal clustering $C_1, \ldots, C_k$.*

For appropriate values of $\delta$, i.e., $\delta \leqslant \frac{(\alpha-1)^2}{8\alpha}$, we show in the supplementary material that a point belonging to a cluster is closer to its approximate mean than the approximate mean of some other cluster. This observation follows from the Figure 1.

The proof of Theorem 1.3 and Theorem 1.8 follows from the above observation combined with the fact that we can indeed get $\{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}$ as above, which we show in the Section 3. For the complete details of the proof, refer to the supplementary material.

### 2.2 $\alpha$-Outliers Center Proximity

In this section we assume that we can get access to a set of points $\{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}$ such that, $\|\tilde{\mu}_i - \mu_i\| \leqslant \delta\,\mathsf{diam}(C_i)$ for all $i \in [k]$, with probability some constant probability. We will

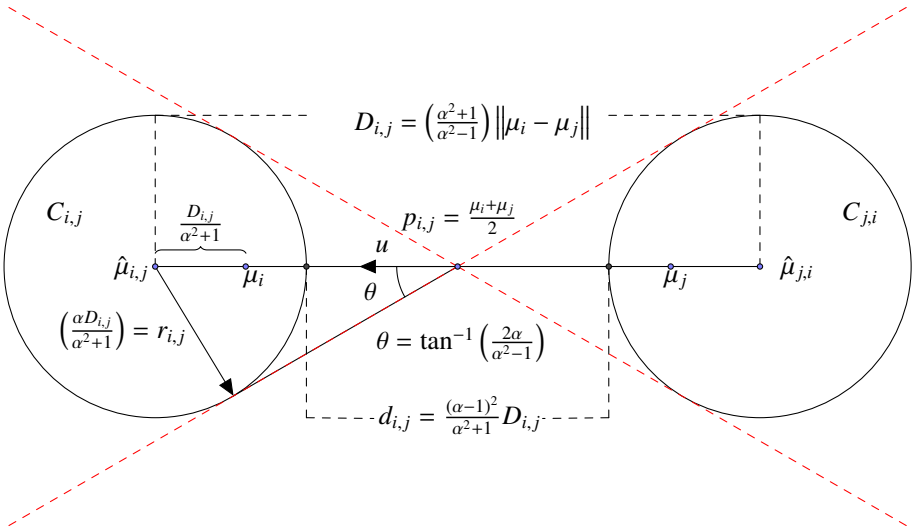


Figure 1: Geometric implication of $\alpha$-center proximity property.

show how to obtain such points in Section 3.1

We first note that, by our Definition 1.2, outliers are the farthest most points in the data-set, where distance is measured from the set of optimal centers $\mu_1, \ldots, \mu_k$.

**Lemma 2.5.** *The Definition 1.2 implies that for any $q \in Z$*

$$\|q - \mu_j\| \geqslant \alpha \frac{\operatorname{diam}(C_i) + \operatorname{diam}(C_j)}{4}, \qquad (\forall i, j \in [k]) \ . \tag{2}$$

We show that with respect to the approximate means as well, the outliers are the farthest set of points for values of $\delta$ as in the previous subsection.

**Lemma 2.6.** *For any $q \in Z$ and $x \in C_i$, we have that $\|x - \tilde{\mu}_i\| < \|q - \tilde{\mu}_j\|$, for $i, j \in [k]$.*

Therefore the proof of the Theorem 1.6 is similar to the proof of Theorem 1.3 and Theorem 1.8 combined with the above observation. The case where most points satisfy $\alpha$-center proximity and form balanced clusters, Proposition 3.1 suggests an algorithm which outputs a list of clusterings, such that one of the clusterings corresponds to the case where the points which satisfy $\alpha$-perturbation resilience are correctly clustered. From this observation the proof of Theorem 1.7 follow. See the supplementary material for full details of the proof.

## 3 SAMPLING

### 3.1 Balanced Cluster Assumption

The following proposition works for general $A_1, \ldots, A_k, Z$. For our case, $A_i$ corresponds to cluster $C_i$, and $Z$ corresponds to the set of outliers.

**Proposition 3.1.** *Fix $\delta, \beta_1 \in (0,1)$. Let $X$ be a set of points in $\mathbb{R}^d$, partitioned into $k+1$ sets $A_1, \ldots, A_k, Z$ such that for each $i$, $|A_i| \geqslant \omega |X|/k$. Let $\mu_i$ denote the mean of $A_i$, i.e., $\mu_i \stackrel{\text{def}}{=} \left(\sum_{v \in A_i} v\right)/|A_i|$. There exists a randomized algorithm which outputs the set $\left\{(\tilde{\mu}_1^{(p)}, \ldots, \tilde{\mu}_k^{(p)}) : p \in [T]\right\}$ satisfying*

$$\mathbb{P}\left[\exists p \in [T] \text{ s.t. } \left\|\tilde{\mu}_i^{(p)} - \mu_i\right\| \leqslant \delta \operatorname{diam}(A_i) \ \forall i \in [k]\right] \geqslant 1 - \beta_1 \ ,$$

*in $O(T)$ iterations where $T < (k+1)^{\frac{16k^2}{\delta^2\beta_1}\log\left(\frac{2k}{\beta_1}\right)}$.*

The proof of the above proposition proceeds using the observation from Barman (2015), which says that with constant probability, one can get close to the mean of a set of points with bounded diameter, by randomly sampling a constant number of points. Since we assume that the desired clustering is balanced (with the balance parameter $\omega$), we can uniformly sample a constant number of points and with a constant probability we will get some points from all the clusters. Using this, we can get close to the actual means, thus proving the above proposition. The proposition directly gives us an algorithm for obtaining such point in time $O(2^{\text{poly}(k/\omega(\alpha-1))}\, nd)$. We pick a constant number of points, and go over all $k$ partitions of those points ($k+1$ partitions in case of outliers) and one of the partitions is the desired clustering. For the complete details of the proof, see the supplementary material.

### 3.2 Balanced Mean Distance Assumption

Assume that the unknown $\alpha$-center proximal $k$-means clustering of lowest cost is $\{C_1, \ldots, C_k\}$, with means $\{\mu_1, \ldots, \mu_k\}$ respectively. In this section we assume that the ratio of the maximum pairwise distance between the means to the minimum pairwise distance between the means is

bounded by a factor $\gamma^*$. We assume that we are given an upper bound $\gamma$ on $\gamma^*$. More formally, we are given a $\gamma$, such that

$$\gamma \geqslant \frac{\max_{i,j} \left\|\mu_i - \mu_j\right\|}{\min_{i \neq j} \left\|\mu_i - \mu_j\right\|}.$$

**Proposition 3.2.** *Fix $\delta \in (0,1)$. Let $X$ be a set of points in $\mathbb{R}^d$, partitioned into $k$ sets $C_1, \ldots, C_k$, and let $\mu_i$ denote the mean of $C_i$, i.e., $\mu_i \stackrel{\text{def}}{=} \left(\sum_{v \in C_i} v\right) / |C_i|$, such that $\gamma = \max_{i,j} \left\|\mu_i - \mu_j\right\| \Big/ \min_{i \neq j} \left\|\mu_i - \mu_j\right\|$. Let $\varepsilon \leqslant \left(\frac{2\delta}{\gamma k}\right)^2$. The Algorithm 2 constructs $O\left(2^{\mathsf{poly}\left(\frac{k}{\varepsilon}\right)}\right)$ $k$-tuples. With constant probability, there exists at least one $k$-tuple $(\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ satisfying*

$$\|\tilde{\mu}_i - \mu_i\| \leqslant 2\delta r_{i,j} \text{ for all } j \in [k] \text{ and for each } i \in [k].$$

*Moreover, the algorithm runs in time $O\left(2^{\mathsf{poly}\left(\frac{k}{\varepsilon}\right)} nd\right)$.*

The proof of the Proposition 3.2 is similar to the proof of Lemma 3.3 of Ding and Xu (2015), with minor modifications, keeping our application in mind.

Using the geometric properties of $\alpha$-center proximal instances (Figure 1) we define $\mathsf{rad}_{\mathsf{min}}(X) \stackrel{\text{def}}{=} \frac{\alpha}{\alpha^2 - 1} \min_{i \neq j} \left\|\mu_i - \mu_j\right\|$. The proposition implies that the radius of the largest $C_{i,j}$ ball (for some $i, j \in [k]$) is at most $\gamma\, \mathsf{rad}_{\mathsf{min}}(X)$, ie.,

$$\frac{r_{i,j}}{\gamma} \leqslant \mathsf{rad}_{\mathsf{min}}(X) \leqslant r_{i,j} \qquad \text{for } i, j \in [k]. \tag{3}$$

**Proposition 3.3.** *Let $R \stackrel{\text{def}}{=} \max_{i,j \in [n]} \left\|x_i - x_j\right\|$, and let $\mathsf{rad}_{\mathsf{min}}(X) \stackrel{\text{def}}{=} \frac{\alpha}{\alpha^2 - 1} \min_{i \neq j} \left\|\mu_i - \mu_j\right\|$. Then,*

$$\mathsf{rad}_{\mathsf{min}}(X) \in \left[\left(\frac{\alpha}{(\alpha+1)^2}\right)\frac{R}{\gamma}, \left(\frac{\alpha}{\alpha^2 - 1}\right)\frac{R}{\gamma}\right]. \tag{4}$$

The proof of the proposition follows from the Figure 1. Refer to the supplementary material for complete details.

The Algorithm 3 is almost same as the Algorithm Peeling-and-Enclosing-Tree of Ding and Xu (2015), with a minor variation in step 2(b).

We note a set of preliminary lemmas and definition which we will need for the proof of Proposition 3.2.

**Definition 3.4** (Simplex)**.** A $k$-simplex is a $k$-dimensional polytope which is the convex hull of its $k+1$ vertices. More formally, suppose the $k+1$ points $u_0, \ldots, u_k \in \mathbb{R}^k$ are affinely independent, Then, the simplex determined by them is the set of points

$$\mathcal{V} = \left\{\theta_0 u_0 + \cdots + \theta_k u_k \;\middle|\; \sum_{i=0}^{k} \theta_i = 1, \theta_i \geqslant 0 \quad \forall i \in [k]\right\}.$$

**Lemma 3.5** (Lemma 1, Inaba et al. (1994))**.** *Let $S$ be a set of $n$ points in $\mathbb{R}^d$, $T$ be a randomly selected subset of size $t$ from $S$, and $\mu(S), \mu(T)$ be the mean points of $S$ and $T$ respectively. With probability $1-\eta$, $\|\mu(S) - \mu(T)\|^2 \leqslant \frac{1}{\eta t}\sigma^2$, where $\sigma^2 = \frac{1}{n}\sum_{s \in S} \|s - \mu(s)\|^2$ and $0 \leqslant \eta \leqslant 1$.*

**Algorithm 2:** Algorithm Peeling-and-Enclosing

**Input:** $X = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$, $k \geqslant 2$, $\alpha$, $\gamma$.
**Output:** A list $\mathcal{L}$ containing $k$-tuples, where a $k$-tuple contains $k$ mean points.

1: Set $\varepsilon = \left(\frac{(\alpha-1)^2 k}{4\alpha\gamma(k+1)}\right)^2$.
2: **for** $i = 0$ to $\log_{(1+\varepsilon)}\left(\frac{\alpha+1}{\alpha-1}\right)$ **do**
3:
3: $\quad \zeta = (1+\varepsilon)^i \left(\frac{\alpha}{(\alpha+1)^2}\right)\frac{R}{\gamma}$.
4: $\quad$ Run **Algorithm Peeling-and-Enclosing-Tree**.
5: $\quad$ Let $\mathcal{T}_i$ be the output tree.
6: **end for**
7: For each root-to-leaf path of every $\mathcal{T}_i$, build a $k$-tuple candidate using the $k$ points associated with the path.
8: Append the $k$-tuple to the list $\mathcal{L}$.

**Lemma 3.6** (Lemma 4, Ding and Xu (2014))**.** *Let $\Gamma$ be a set of elements, and $S$ be a subset of $\Gamma$ with $\frac{|S|}{|\Gamma|} = \alpha$ for some $\alpha \in (0,1)$. If we randomly select $\frac{t \ln \frac{t}{\eta}}{\ln(1+\alpha)} = O\left(\frac{t}{\alpha} \ln \frac{t}{\eta}\right)$ elements from $\Gamma$, then with probability at least $1-\eta$, the sample contains $t$ or more elements from $S$ for $0 < \eta < 1$ and $t \in \mathbb{Z}^+$.*

**Lemma 3.7** (Lemma 2.3 (Simplex Lemma II), Ding and Xu (2015))**.** *Let $Q$ be a set of points in $\mathbb{R}^d$ with a partition of $Q = \cup_{l=1}^{j} Q_l$ and $Q_{l_1} \cap Q_{l_2} = \emptyset$ for any $l_1 \neq l_2$. Let $o$ be the mean point of $Q$, and $o_l$ be the mean point of $Q_l$ for $1 \leqslant l \leqslant j$. Let $\sigma^2 = \frac{1}{|Q|}\sum_{q \in Q} \|q - o\|^2$. Let $\{o'_1, \ldots, o'_j\}$ be $j$ points in $\mathbb{R}^d$ such that $\left\|o_l - o'_l\right\| \leqslant L$ for $1 \leqslant l \leqslant j, L > 0$, and $\mathcal{V}'$ be the simplex determined by $\{o'_1, \ldots, o'_j\}$. Then for any $0 < \varepsilon \leqslant 1$, it is possible to construct a grid of size $O\left((8j/\varepsilon)^j\right)$ inside $\mathcal{V}'$ such that at least one grid point $\tau$ satisfies the inequality $\|\tau - o\| \leqslant \sqrt{\varepsilon}\sigma + (1+\varepsilon)L$.*

**Lemma 3.8** (Lemma 2.2, Ding and Xu (2015))**.** *Let $Q$ be a set of points in $\mathbb{R}^d$, and $Q_1$ be its subset containing $\alpha|Q|$ points for some $0 < \alpha \leqslant 1$. Let $o$ and $o_1$ be the mean points of $Q$ and $Q_1$, respectively. Then $\|o - o_1\| \leqslant \sqrt{\frac{1-\alpha}{\alpha}}\sigma$, where $\sigma^2 = \frac{1}{|Q|}\sum_{q \in Q} \|q - o\|^2$.*

**Notations:** Let $\mathcal{OPT} = \{C_1, \ldots, C_k\}$ be the $k$ unknown optimal clusters for the lowest cost $\alpha$-center proximal $k$-means objective, with means $\mu_j$. W.l.o.g. we assume that $|C_1| \geqslant \ldots \geqslant |C_k|$.

The following lemma is similar to the Lemma 3.3 of Ding and Xu (2015) with minor modifications.

**Lemma 3.9.** *Among all the points generated by the Algorithm 3, with constant probability, there exists at least one tree, $\mathcal{T}_i$, which has a root-to-leaf path with each of its nodes $v_j$ at level $j, (1 \leqslant j \leqslant k)$ associating with a point $p_{v_j}$ and*

**Algorithm 3:** Algorithm Peeling-and-Enclosing-Tree

**Input:** $\zeta$ and an instance of $k$-means $X$.
**Output:** a tree $\mathcal{T}$.

1: Initialize $\mathcal{T}$ as a single root node $v$ associated with no point.
2: Recursively grow each node $v$ in the following way
3: **if** the height of $v$ is already $k$ **then**
4: $v$ is a leaf.
5: **else**
6: Let $j$ be the height of $v$. Build the radius candidate set $\mathcal{R} = \left\{ \frac{1+l\frac{\varepsilon}{2}}{2(1+\varepsilon)} j 2^{1/2} \sqrt{\varepsilon}\zeta\gamma \middle| 0 \leqslant l \leqslant 4 + \frac{2}{\varepsilon} \right\}$.
7: **for** each $r \in \mathcal{R}$ **do**
8: Let $\left\{p_{v_1}, \ldots, p_{v_j}\right\}$ be the $j$ points associated with nodes on the root-to-v path.
9: For each $p_{v_l}, 1 \leqslant l \leqslant j$, construct a ball $\mathcal{B}_{j+1,l}$ centered at $p_{v_l}$ and with radius $r$.
10: Sample $s = \frac{8k^3}{\varepsilon^9} \ln \frac{k^2}{\varepsilon^6}$ points uniformly from $\left(X \setminus \cup_{l=1}^{j} \mathcal{B}_{j+1,l}\right)$. Compute the means of all subsets of this sample, denote them by $\Pi = \{\pi_1, \ldots, \pi_{2^s-1}\}$.
11: For each $\pi_i \in \Pi$, construct two simplices with vertices $\left\{p_{v_1}, \ldots, p_{v_j}, \pi_i\right\}$ and $\left\{p_{v_1}, \ldots, p_{v_j}\right\}$, respectively. For each simplex, build a grid with size $O\left((32j/\varepsilon)^j\right)$ inside it and each of its $2^j$ possibly-degenerated sub-simplices.
12: Take all the above $2^{s+j}(32j/\varepsilon)^j$ grid points inside $2^s$ simplices. For each grid point, add one child to $v$, associate it with the grid point.
13: **end for**
14: **end if**
15: Output $\mathcal{T}$.

*satisfying the inequality*

$$\left\|p_{v_j} - \mu_j\right\| \leqslant \varepsilon\gamma \,\mathsf{rad}_{\mathsf{min}}(X) + (1+\varepsilon) j \sqrt{\varepsilon}\gamma \,\mathsf{rad}_{\mathsf{min}}(X) \ . \quad (5)$$

**Algorithm and Proof Overview:** We will give a high level idea of the algorithm and the proof. At each searching step, the algorithm performs a 'sphere peeling' and 'simplex enclosing' step, to generate $k$ approximate mean points for the clusters. Initially the algorithm uses a random sampling technique to find an approximate mean $p_{v_1}$ for $C_1$. This can be done as $\frac{|C_1|}{n} \geqslant 1/k$, and hence we can sample. Suppose we have found the $j$ approximate means for some $j \in \{1, 2, ..., k-1\}$. Therefore, at the $(j+1)^{th}$ iteration, the algorithm already has approximate mean points mean points $p_{v_1}, \ldots, p_{v_j}$ for $C_1, \ldots, C_j$. It is not clear how to distinguish points which belong to $C_1, \ldots, C_j$ from those which belong to $C_{j+1}$. Also, the number of points in the cluster $C_{j+1}$ could be small, it is tough to obtain a significant fraction of such points using random sampling. Therefore, the idea used is to seperate the points in $C_{j+1}$ using $j$ peeling spheres, $\mathcal{B}_{j+1,1}, \ldots, \mathcal{B}_{j+1,j}$, centered at the $j$ approximate mean points respectively and with radius approximately being $\mathsf{rad}_{\mathsf{min}}(X)$. Note that $\mathcal{B}_{j+1,1}, \ldots, \mathcal{B}_{j+1,j}$ can have some points from $C_{j+1}$. Let $P_{j+1}$ be the set of unknown points in $C_{j+1} \setminus \left(\cup_{l=1}^{j} \mathcal{B}_{j+1,l}\right)$. The algorithm considers two cases, a) $\left|P_{j+1}\right|$ is large and b) $\left|P_{j+1}\right|$ is small. For the case a) when $\left|P_{j+1}\right|$ is large, we can sample points from $P_{j+1}$ using random sampling, and get an approximate mean $\pi$ of $P_{j+1}$, and then construct a simplex determined by $\pi, p_{v_1}, \ldots, p_{v_j}$ to contain the $(j+1)^{th}$ mean point, using Lemma 3.7. This is because, $C_{j+1} \cap \mathcal{B}_{j+1,l}$, $l \in [j]$ can be seen as a partition of $C_{j+1}$ whose approximate mean is $p_{v_l}$, thus the simplex lemma II applies. For case b) where $\left|P_{j+1}\right|$ is small, it directly constructs the simplex determined by $p_{v_1}, \ldots, p_{v_j}$, and searches for the approximate mean point of $C_{j+1}$ in the grid. We can establish that $C_{j+1} \cap \mathcal{B}_{j+1,l}$, $l \in [j]$ can be seen as a partition of $C_{j+1}$ whose approximate mean is $p_{v_l}$, and from the Lemma 3.8, which roughly says that even if we remove a small number of points from a cluster, its new mean remains close to the original mean.

Setting $\varepsilon \leqslant \left(\frac{2\delta k}{\gamma(k+1)}\right)^2$ in the equation (5), we get statement of the Proposition 3.2. For complete details of the proof, refer to the supplementary material.

# 4 LOWER BOUND

## 4.1 Hardness Result

Our hardness result immediately follows from Awasthi et al. (2015). We will show a reduction from Vertex-Cover problem to the $\alpha$-center proximal $k$-means clustering with balanced clusters. The Vertex-Cover problem can be stated as follows: Given an undirected graph $G = (V, E)$, choose a subset $S$ of vertices with minimum $|S|$, such that $S$ is incident on every edge of the graph. Awasthi et al. (2015) show the following lemma:

**Lemma 4.1** (Corollary 5.3, Awasthi et al. (2015))**.** *Given any unweighted triangle-free graph $G$ with bounded degrees, it is NP-hard to approximate Vertex-Cover within any factor smaller than* 1.36.

**Theorem 4.2.** *There exists constants $\alpha > 1$, $\omega > 0$, $\varepsilon > 0$, such that there is an efficient reduction from instances of Vertex-Cover on triangle-free graphs of bounded degree to those of $\alpha$-center proximal instances of Euclidean $k$-means clustering, where the size of each cluster is at least $\omega n/k$, that satisfies the following properties:*

*(i) if the Vertex-Cover instance has value $k$, the optimal $\alpha$-center proximal $k$-means clustering where the size of each cluster is at least $\omega n/k$, has cost at most $m-k$.*

*(ii) if the Vertex-Cover instance has value at least $k(1+\varepsilon)$, then the optimal $\alpha$-center proximal $k$-means clustering, where the size of each cluster is at least $\omega n/k$, has a cost at least $m - (1 - \Omega(\varepsilon))k$.*

Table 1: This table shows the fraction of points satisfying $\alpha >$ threshold value, for k++, GT, k++-pruned, and GT-pruned clustering, for various values of the threshold.

| Dataset | $\alpha \geqslant 1.04$ | $\alpha \geqslant 1.06$ | $\alpha \geqslant 1.08$ | $\alpha \geqslant 1.1$ | $\alpha \geqslant 1.12$ |
|---|---|---|---|---|---|
| Wine (k++) | 1 | 0.994 | 0.989 | 0.989 | 0.978 |
| Wine (k++ - pruned) | 1 | 1 | 1 | 1 | 1 |
| Wine (GT) | 1 | 0.994 | 0.989 | 0.989 | 0.978 |
| Wine (GT - pruned) | 1 | 1 | 1 | 1 | 1 |
| Iris (k++) | 0.993 | 0.993 | 0.993 | 0.98 | 0.98 |
| Iris (k++ - pruned) | 1 | 1 | 1 | 1 | 1 |
| Iris (GT) | 0.993 | 0.993 | 0.987 | 0.987 | 0.98 |
| Iris (GT - pruned) | 1 | 1 | 1 | 1 | 1 |
| Banknote Auth. (k++) | 0.989 | 0.985 | 0.98 | 0.976 | 0.97 |
| Banknote Auth. (k++ - pruned) | 0.999 | 0.999 | 0.998 | 0.997 | 0.992 |
| Banknote Auth. (GT) | 0.989 | 0.985 | 0.98 | 0.976 | 0.97 |
| Banknote Auth. (GT - pruned) | 0.999 | 0.999 | 0.998 | 0.997 | 0.992 |

Table 2: This table shows the fraction of points satisfying $\alpha >$ threshold value, for k++, GT, k++-pruned, and GT-pruned clustering, for various values of the threshold.

| Dataset | $\alpha \geqslant 1.017$ | $\alpha \geqslant 1.019$ | $\alpha \geqslant 1.021$ | $\alpha \geqslant 1.023$ | $\alpha \geqslant 1.025$ |
|---|---|---|---|---|---|
| Letter Rec. (k++) | 0.966 | 0.962 | 0.957 | 0.952 | 0.948 |
| Letter Rec. (k++ - pruned) | 0.995 | 0.994 | 0.994 | 0.994 | 0.994 |
| Letter Rec. (GT) | 0.964 | 0.96 | 0.954 | 0.949 | 0.945 |
| Letter Rec. (GT - pruned) | 0.995 | 0.994 | 0.994 | 0.994 | 0.993 |

The proof of the Theorem 1.5 follows immediately from Theorem 4.2 and Lemma 4.1 combined together. See the supplementary material for a more detailed proof.

### 4.2 On the Size of Possible Clustering

Our construction is similar to the instance constructed by Bhattacharya et al. (2018) in their Theorem 2, thereby proving the Proposition 1.9. Refer to the supplementary material for complete details.

## 5 EXPERIMENTS

We try to calculate the value of $\alpha$ for multiple real-world labeled datasets. We consider two possible clusterings of a dataset: 'k++', and 'GT'. The k++ clustering of a dataset corresponds to the cluster returned by the smallest $k$-means cost of 1000 trials of Lloyd's algorithms on the unlabeled (by removing the labels) dataset using $k$-mean++ initialization. The GT clustering of a dataset corresponds to finding the mean of a cluster using the labels, and then run Lloyd's algorithm to get the clustering.

We also look at the pruned versions of the dataset ('k++ - pruned', 'GT - pruned'). The pruning is done by fixing a value for $\alpha$. We remove points from the dataset which do not satisfy the $\alpha$ proximity condition and recompute the means of the pruned clusters. There will still be some fraction of points not satisfying the $\alpha$ proximity condition in the pruned dataset, however we observe that this fraction has decreased. We report the fraction of the points satisfying the $\alpha$ proximity condition in the Table 1 and Table 2.

The experiments were run on four labeled datasets from the UCI Machine Learning Repository: Wine ($n = 178, k = 3, d = 13$), Iris ($n = 150, k = 3, d = 4$), Bank Authentication ($n = 1372, k = 2, d = 4$), and Letter Recognition ($n = 20,000, k = 26, d = 16$).

### Acknowledgements

AL is grateful to Microsoft Research for supporting this collaboration. AL was supported in part by SERB Award ECR/2017/003296. We thank Ravishankar Krishnaswamy for helpful pointers.

### References

Ackerman, M. and Ben-David, S. (2009). Clusterability: A theoretical study. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 1–8. PMLR. 2

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248. 1

Angelidakis, H., Makarychev, K., and Makarychev, Y. (2017). Algorithms for stable and perturbation-resilient problems. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 438–451. ACM. 1, 2

Awasthi, P., Blum, A., and Sheffet, O. (2012). Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54. 1

Awasthi, P., Charikar, M., Krishnaswamy, R., and Sinop, A. K. (2015). The Hardness of Approximation of Euclidean k-Means. In *31st International Symposium on Computational Geometry (SoCG 2015)*, volume 34 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 754–767. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 1, 7

Balcan, M. and Liang, Y. (2016). Clustering under perturbation resilience. *SIAM J. Comput.*, 45(1):102–155. 1, 4

Balcan, M.-F., Haghtalab, N., and White, C. (2016). k-Center Clustering Under Perturbation Resilience. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 68:1–68:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 2

Barman, S. (2015). Approximating nash equilibria and dense bipartite subgraphs via an approximate version of caratheodory's theorem. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 361–369. ACM. 5

Ben-David, S. (2015). Computational feasibility of clustering under clusterability assumptions. *CoRR*, abs/1501.00437. 2

Ben-David, S. and Reyzin, L. (2014). Data stability in clustering: A closer look. *Theor. Comput. Sci.*, 558:51–61. 1, 2

Bhattacharya, A., Jaiswal, R., and Kumar, A. (2018). Faster algorithms for the constrained k-means problem. *Theory of Computing Systems*, 62(1):93–115. 8

Dasgupta, S. and Freund, Y. (2009). Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7):3229–3242. 1

Ding, H. and Xu, J. (2014). Sub-linear time hybrid approximations for least trimmed squares estimator and related problems. In *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*, SOCG'14, pages 110:110–110:119, New York, NY, USA. ACM. 6

Ding, H. and Xu, J. (2015). A unified framework for clustering constrained data without locality property. In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1471–1490, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics. 6

Friggstad, Z., Khodamoradi, K., and Salavatipour, M. R. (2018). Exact algorithms and lower bounds for stable instances of euclidean *k*-means. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2958–2972. 2

Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering: (extended abstract). In *Proceedings of the Tenth Annual Symposium on Computational Geometry*, SCG '94, pages 332–339. ACM. 6

Kumar, A. and Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science (FOCS) 2010, October 23-26, 2010*, pages 299–308. 1

Mahajan, M., Nimbhorkar, P., and Varadarajan, K. (2012). The planar k-means problem is np-hard. *Theoretical Computer Science*, 442:13 – 21. Special Issue on the Workshop on Algorithms and Computation (WALCOM 2009). 1

Telgarsky, M. and Vattani, A. (2010). Hartigan's method: k-means clustering without voronoi. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 820–827. PMLR. 3

Vijayaraghavan, A., Dutta, A., and Wang, A. (2017). Clustering stable instances of euclidean k-means. In *Advances in Neural Information Processing Systems 30*, pages 6503–6512. Curran Associates, Inc. 2, 4