Explainable k-Means: Don't Be Greedy, Plant Bigger Trees![*]

Konstantin Makarychev
Northwestern University
Evanston, IL, USA

Liren Shan
Northwestern University
Evanston, IL, USA

ABSTRACT

We provide a new *bi-criteria* $\tilde{O}(\log^2 k)$ competitive algorithm for explainable k-means clustering. Explainable k-means was recently introduced by Dasgupta, Frost, Moshkovitz, and Rashtchian (ICML 2020). It is described by an easy to interpret and understand (threshold) decision tree or diagram. The cost of the *explainable k-means* clustering equals to the sum of costs of its clusters; and the cost of each cluster equals the sum of squared distances from the points in the cluster to the center of that cluster. The best non bi-criteria algorithm for explainable clustering $\tilde{O}(k)$ competitive, and this bound is tight.

Our randomized bi-criteria algorithm constructs a threshold decision tree that partitions the data set into $(1+\delta)k$ clusters (where $\delta \in (0,1)$ is a parameter of the algorithm). The cost of this clustering is at most $\tilde{O}(1/\delta \cdot \log^2 k)$ times the cost of the optimal unconstrained k-means clustering. We show that this bound is almost optimal.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**.

KEYWORDS

Clustering; k-means; Decision Tree

ACM Reference Format:
Konstantin Makarychev and Liren Shan. 2022. Explainable k-Means: Don't Be Greedy, Plant Bigger Trees!. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC '22), June 20–24, 2022, Rome, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3519935.3520056

1 INTRODUCTION

In this paper, we study explainable k-means clustering. k-means is one of the most popular ways to cluster data. It is widely used in data science and machine learning. A k-means clustering of data set X in $\mathbb{R}^d$ is determined by its k centers $c^1, c^2, \ldots, c^k$. Specifically, k-means clustering is a d-dimensional Voronoi diagram for centers $c^1, c^2, \ldots, c^k$, in which, the i-th cluster contains those points in X that are closer to c^i than to any other center c^j. The cost of the clustering equals

$$\text{cost}(X; c^1, c^2, \ldots, c^k) \equiv \sum_{i=1}^{d} \sum_{x \in P_i} \|x - c^i\|_2^2, \tag{1}$$

where P_i is the i-th cluster.

In a recent paper, Dasgupta, Frost, Moshkovitz, and Rashtchian [14] observed that it can be hard for a human to understand k-means clustering. Clusters in k-means are determined by all features (coordinates) of the data. Thus, usually there is no a concise explanation of why a particular point belongs to one cluster or another. To make k-means more understandable for humans, Dasgupta et al. [14] proposed an alternative way to cluster data, which they called *explainable k-means*. In *explainable k-means*, the data set is partitioned into clusters using a threshold decision tree with k leaves (a variant of a binary space partitioning tree). Every internal node u of the tree splits the data into two disjoint groups based on a single feature (coordinate). A point x is assigned to the left subtree of u, if $x_i \le \theta$; it is assigned to the right subtree of u, if $x_i > \theta$. Points assigned to each of the k leaves form a cluster. The cost of explainable k-means clustering is defined in the same way as for k-means. It is equal to the sum of cluster costs:

$$\text{cost}(X, \mathcal{T}) = \sum_{i=1}^{d} \sum_{x \in P_i} \|x - c^i\|_2^2,$$

where $P_1, \ldots, P_k$ are clusters; $c^1, \ldots, c^k$ are centers of $P_1, \ldots, P_k$; and $\mathcal{T}$ is the decision tree that defines the clustering.

Note that explainable k-means clustering can be represented by a simple decision diagram as in Figure 1. This diagram is easy to understand, and humans can easily determine to which cluster a given data point x belongs to.

The cost of explainability or the competitive ratio of an explainable k-means clustering is the ratio between the cost of that clustering and the cost of the optimal unconstrained k-means clustering for the same data set. Dasgupta et al. [14] showed how to obtain a k-means clustering with a competitive ratio of $O(k^2)$. This competitive ratio was improved to a near-optimal[1] bound of $\tilde{O}(k)$ by Makarychev and Shan [28]; Gamlath, Jia, Polak, and Svensson [19]; and Esfandiari, Mirrokni, and Narayanan [16]. This guarantee does not depend on the size and dimension of the data set. However, it is large for large data sets. For comparison, the best competitive ratio for explainable *k-medians* is exponentially better than $\tilde{O}(k)$. It equals $\tilde{O}(\log k)$ (see Esfandiari et al. [16], Makarychev and Shan [28]). Nevertheless, Dasgupta et al. [14] and then Frost et al. [18] empirically demonstrated that, in practice, the price of explainability for k-means clustering is fairly small. In this work, we provide a

[*]This work was supported by NSF Awards CCF-1955351 and CCF-1934931.

STOC '22, June 20–24, 2022, Rome, Italy

ACM ISBN 978-1-4503-9264-8/22/06…$15.00
https://doi.org/10.1145/3519935.3520056

[1]It is possible to get a better competitive ratio for low dimensional data. For details, see Section 1.2

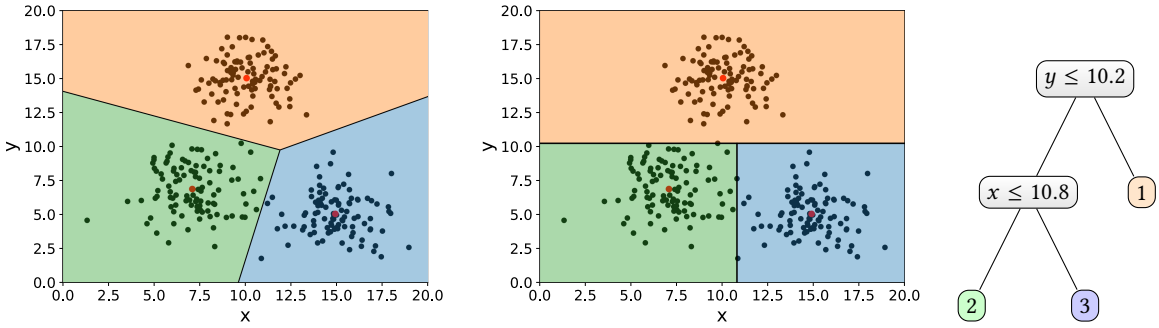

Figure 1: Explainable and non-explainable k-means. The left diagram shows the optimal Voronoi partition of the plane. The middle diagram shows an explainable partition. The right diagram shows the corresponding decision tree for explainable clustering.

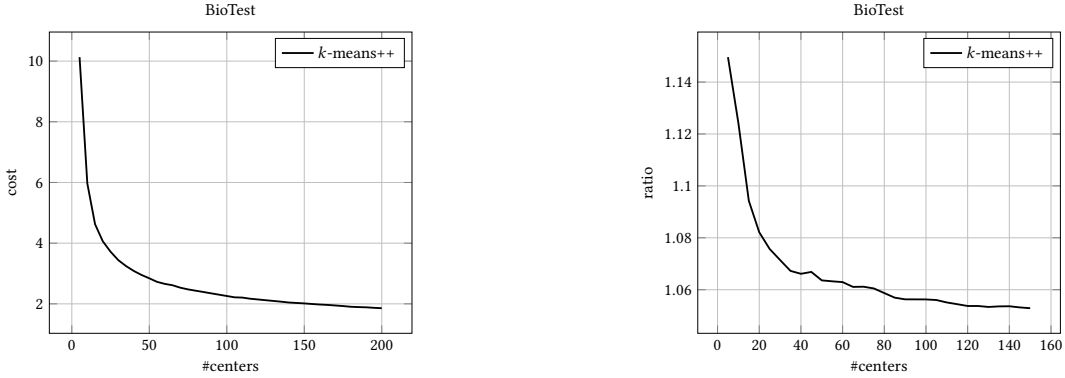

Figure 2: Performance of k-means++ on BioTest data set. The left diagram shows the cost of k-means++ for $k = 5, 10, 15, \ldots, 200$. The clustering cost is divided by the cost of k-means with 1000 clusters. The right diagram shows the ratio between the clustering cost with k centers and the cost with $(1+\delta)k$ centers for $k = 5, 10, \ldots, 150$ and $\delta = 0.2$.

theoretical justification for this observation. Specifically, we show a bi-criteria approximation algorithm which finds a decision tree with $(1+\delta)k$ leaves and has a competitive ratio of $O(1/\delta \log^2 k \log\log k)$, where δ is a parameter between 0 and 1.

We note that in practice the cost of the optimal k-means clustering is approximately the same for k and $(1+\delta)k$ clusters (here $\delta \in (0,1)$ is a small constant). In other words, for many data sets X, we have $\mathrm{OPT}_k(X) \approx \mathrm{OPT}_{(1+\delta)k}(X)$, where $\mathrm{OPT}_k(X)$ is the cost of the optimal unconstrained k-means clustering of X with k clusters[2]. The plot in Figure 2 shows that the cost of k-means++ clustering for BioTest data set from KDD Cup [15] is about the same for k and $(1+\delta)k$ centers when k is between 50 and 200. If $\mathrm{OPT}_k(X) \approx \mathrm{OPT}_{(1+\delta)k}(X)$, then our algorithm gives a *true* $\tilde{O}(\log^2 k)$ approximation, because

$$\mathrm{cost}(X, \mathcal{T}) \le \tilde{O}(\log^2 k)\, \mathrm{OPT}_k(X) \approx \tilde{O}(\log^2 k)\, \mathrm{OPT}_{(1+\delta)k}(X).$$

[2]In the worst case, we may have $\mathrm{OPT}_{(1+\delta)k}(X) \ll \mathrm{OPT}_k(X)$. For example, if X contains exactly $(1+\delta)k$ points, then $\mathrm{OPT}_{(1+\delta)k}(X) = 0$ but $\mathrm{OPT}_k(X) > 0$.

1.1 Our Results

We now formally state our results. We provide a randomized algorithm for finding bi-criteria explainable k-means. Similarly to the algorithm by Frost et al. [18], our algorithm takes k centers $\{c^1, c^2, \ldots, c^k\}$ and a parameter $\delta > 0$ and returns a threshold decision tree with $(1+\delta)k$ leaves. Each leaf of the tree is labeled with one of the centers $c^1, c^2, \ldots, c^k$. Let us denote the center returned by the decision tree $\mathcal{T}$ for point x by $\mathcal{T}(x)$. Then, the cost of explainable clustering defined by $\mathcal{T}$ equals

$$\mathrm{cost}(X, \mathcal{T}) \equiv \sum_{x \in X} \|x - \mathcal{T}(x)\|_2^2. \tag{2}$$

Theorem 1.1. *There exists a polynomial-time randomized algorithm that given a data set X, a set of k centers $C = \{c^1, c^2, \ldots, c^k\}$, and parameter $\delta \in (0,1)$, creates a threshold decision tree $\mathcal{T}$ whose leaves are labeled with centers from C. The expected number of leaves in $\mathcal{T}$ is $(1+\delta)k$, and the expected cost of explainable clustering defined by $\mathcal{T}$ is*

$$\mathbf{E}[\mathrm{cost}(X, \mathcal{T})] \le O(1/\delta \cdot \log^2 k \log\log k) \cdot \mathrm{cost}(X, C).$$

Observe that our algorithm constructs a tree with $(1+\delta)k$ leaves and only k centers. Thus, we can use this algorithm to partition X into k clusters. In this case, one cluster may be assigned to several different leaves. Alternatively, we can assign its own cluster to every leaf. Then, we will have a proper threshold decision tree with $(1+\delta)k$ clusters. In either case, we can further improve the clustering by replacing the original center c^i assigned to each leaf u with the optimal center for the cluster assigned to u (the optimal center is the centroid of that cluster).

If C is the optimal set of centers for k means, then the explainable clustering provided by our algorithm has an expected cost of at most $O(1/\delta \cdot \log^2 k \log\log k)\,\mathrm{OPT}_k(X)$. Furthermore, if C is obtained by a constant factor bi-criteria approximation algorithm such as k-means++ (in which case, $|C| = (1+\delta)k$ and $\mathrm{cost}(X, C) \le O(1) \cdot \mathrm{OPT}_k(X)$), then the expected cost of the explainable clustering is also at most $O(1/\delta \cdot \log^2 k \log\log k)\,\mathrm{OPT}_k(X)$ and the number of leaves in the threshold decision tree is at most $(1+3\delta)k$ in expectation.

As we note above, our work is influenced by the paper of Frost, Moshkovitz, and Rashtchian [18], who showed a bi-criteria algorithm for explainable k-means. However, our algorithm for this problem is very different from theirs. It uses the approach from our previous paper (Makarychev and Shan [28]). In that paper, we gave an algorithm for finding explainable *k-medians with ℓ_2 norm*. Our new algorithm has an additional crucial step: It duplicates some centers when the algorithm splits nodes. This step gives an exponential improvement to the competitive ratio for k-means. The analysis of our algorithm is considerably more involved than the analysis of the previous algorithm.

We complement our algorithmic results with an almost matching lower bound of $\Omega(1/\delta \cdot \log^2 k)$ for all threshold trees with at most $(1+\delta)k$ leaves.

Theorem 1.2. *For every $k > 500$ and $\ln^3 k/\sqrt{k} < \delta < 1/100$, there exists an instance X with k clusters such that the k-means cost for every threshold tree $\mathcal{T}$ with $(1+\delta)k$ leaves is at least*

$$\mathrm{cost}(X, \mathcal{T}) \ge \Omega\left(\frac{\log^2 k}{\delta}\right)\mathrm{OPT}_k(X).$$

In the full version, we provide a family of k-means instances for which a greedy bi-criteria algorithm finds a solution of cost $\mathrm{cost}(X, \mathcal{T}) \ge \widetilde{\Omega}(k^2)\,\mathrm{OPT}_k(X)$ for $k \to \infty$.

1.2 Related Work

Decision trees have been widely used for classification and clustering due to their simplicity. Examples of decision tree algorithms for supervised classification include CART by Breiman et al. [10], ID3 by Quinlan [29], and C4.5 by Quinlan [30]. Examples of decision tree algorithms for unsupervised clustering include algorithms by Liu et al. [23], Fraiman et al. [17], Silhouette Metric (Bertsimas et al. [7]), Saisubramanian et al. [31].

Dasgupta et al. [14] proposed the problems of explainable k-means and k-medians clustering in ℓ_1. They defined these problems and offered algorithms for explainable k-means and k-medians with the competitive ratios of $O(k^2)$ and $O(k)$, respectively. Later, Frost et al. [18] designed a new bi-criteria algorithm for these problems and evaluated its performance in practice. Laber and Murtinho [21], Makarychev and Shan [28], Charikar and Hu [11], Esfandiari, Mirrokni, and Narayanan [16], and Gamlath, Jia, Polak, and Svensson [19] provided improved upper and lower bounds for explainable k-means and k-medians. The best competitive ratios for explainable k-means and k-medians are $\tilde{O}(k)$ and $\tilde{O}(\log k)$, respectively. Makarychev and Shan [28], Esfandiari et al. [16], and Gamlath et al. [19] gave a $\tilde{O}(k)$ competitive ratio for explainable k-means; and Makarychev and Shan [28] and Esfandiari et al. [16] gave a $\tilde{O}(\log k)$ bound for k-medians. Charikar and Hu [11] provided $k^{1-2/d} \cdot \mathrm{poly}(d \log k)$ algorithm for k-means (this algorithm gives stronger approximation guarantees when the dimension of the space, d, is small). Additionally, Makarychev and Shan [28] gave an $\tilde{O}(\log^{3/2} n)$ competitive algorithm for explainable k-medians in ℓ_2.

Boutsidis et al. [8], Boutsidis et al. [9], Makarychev et al. [25], Cohen et al. [12], and Becchetti et al. [6] showed how to reduce the dimensionality of a data set for k-means clustering. Particularly, Makarychev et al. [25] proved that we can use the Johnson Lindenstrauss transform to reduce the dimensionality of k-means to $d' = O(\log k)$. Note, however, that the Johnson Lindenstrauss transform cannot be used for the explainable k-means, because this transform does not preserve the set of features. Instead, one can use a *feature selection* algorithm by Boutsidis et al. [9] or Cohen et al. [12] to reduce the dimensionality to $d' = \tilde{O}(k)$.

The classic k-means clustering has been extensively studied by researchers in machine learning and theoretical computer science. Lloyd's algorithm (Lloyd [24]) is the most popular heuristic for k-means clustering. Arthur and Vassilvitskii [4] proposed a randomized seeding algorithm called k-means++, which achieves an expected $O(\log k)$ approximation. Ahmadian, Norouzi-Fard, Svensson, and Ward [2] designed a primal-dual algorithm with an approximation factor of 6.357. It was recently improved to 6.12903 by Grandoni, Ostrovsky, Rabani, Schulman, and Venkat [20]. Dasgupta [13] and Aloise, Deshpande, Hansen, and Popat [3] showed that k-means problem is NP-hard. Awasthi et al. [5] showed that it is also NP-hard to approximate the k-means objective within a factor of $(1+\varepsilon)$ for some positive constant ε (see also Lee, Schmidt, and Wright [22]). The bi-criteria approximation for k-means has also been studied before. Aggarwal, Deshpande, and Kannan [1] proved that k-means++ that picks $(1+\delta)k$ centers gives a constant factor bi-criteria approximation for some constant $\delta > 0$. Later, Wei [32] and Makarychev, Reddy, and Shan [27] gave improved bi-criteria approximation guarantees for k-means++. Makarychev, Makarychev, Sviridenko, and Ward [26] designed local search and LP-based algorithms with better bi-criteria approximation guarantees.

2 PRELIMINARIES

Consider a set of points $X \subseteq \mathbb{R}^d$ and an integer $k > 1$. A k-means clustering consists of k clusters $P_1, \dots, P_k$. Each cluster P_i is assigned a center c^i, which is the centroid (geometric center) of P_i.

The cost of the clustering equals

$$\text{cost}(X; c^1, \dots, c^k) \equiv \sum_{i=1}^{d} \sum_{x \in P_i} \|x - c^i\|_2^2.$$

The optimal k-means clustering is the clustering of the minimum cost. We denote the cost of the optimal k-means clustering with k clusters by $\text{OPT}_k(X)$.

A threshold decision tree is a tree that recursively partitions $\mathbb{R}^d$ into cells using hyperplane cuts. Every node in the tree corresponds to a cell (polytope) of the space. The root corresponds to the entire space $\mathbb{R}^d$. In this paper, we will identify nodes of the tree with the cells they correspond to. Thus, a threshold decision tree defines a hierarchical partitioning of $\mathbb{R}^d$ into k cells or clusters.

Each internal node (cell) u in the threshold tree is split into two nodes u_{left} and u_{right} using a threshold cut (i, ξ) as follows:

$$u_{left} = \{x \in u : x_i \le \xi\} \quad \text{and} \quad u_{right} = \{x \in u : x_i > \xi\}.$$

We assign a center c to every leaf of the threshold decision tree. Let $\mathcal{T}(x)$ (where $x \in \mathbb{R}^d$) be the center assigned to the unique leaf u of $\mathcal{T}$ that contains x. In this paper, we will also assign centers to internal nodes of the tree. We will denote the set of centers assigned to node u by C_u. For leaf nodes, we have $|C_u| = 1$.

Consider a data set X and threshold decision tree $\mathcal{T}$. The k-means cost of $\mathcal{T}$ equals

$$\text{cost}(X, \mathcal{T}) \equiv \sum_{x \in X} \|x - \mathcal{T}(x)\|_2^2.$$

The competitive ratio of explainable clustering defined by $\mathcal{T}$ is $\text{cost}(X, \mathcal{T})/\text{OPT}_k(X)$. We say that a randomized algorithm is α-competitive if the expected cost of the explainable clustering returned by the algorithm is at most $\alpha \,\text{cost}(X, C)$, where C is the set of centers provided to the algorithm.

A bi-criteria solution to explainable k-means clustering with parameter δ is a threshold decision tree with at most $(1+\delta)k$ leaves. In this paper, we describe an algorithm that finds a tree with at most $(1+\delta)k$ leaves and k distinct centers assigned to them.

3 ALGORITHM

In this section, we present an algorithm for explainable k-means clustering. The input of the algorithm is a set of centers $c^1, \dots, c^k$ and parameter δ. The output is a threshold decision tree in which every leaf node is labeled with one of the centers c^i. In Sections 5 and 6, we will show that the expected number of leaves in the decision tree is $(1+\delta)k$ and the approximation factor of the obtained clustering is $O(1/\delta \cdot \log^2 k \cdot \log\log k)$.

Algorithm. Our algorithm builds a binary threshold tree using a top-down approach. The algorithm assigns every node u in the tree a subset of centers $c^1, \dots, c^k$. We denote this subset C_u. First, the algorithm creates a tree $\mathcal{T}_1$ with a root vertex r and assigns all centers $c^1, c^2, \dots, c^k$ to it. Then, the algorithm recursively splits leaf nodes in the threshold tree until each leaf is assigned exactly one center. At each step t, the algorithm chooses a coordinate $i_t \in \{1, 2, \dots, d\}$, a positive threshold $\theta_t \in (0, 1)$, and number σ_t in $\{\pm 1\}$ uniformly at random. For each leaf u with more than one center, it calls function *Divide-and-Share* to split node u into two parts.

Threshold Tree Construction

Input: a data set X and set of centers $C = \{c^1, c^2, \dots, c^k\}$, a parameter $\delta \in (0, 1)$

Output: a threshold tree $\mathcal{T}$

- Create a tree $\mathcal{T}_1$ containing a root r. Let $C_r = C$.
- **while** $\mathcal{T}_t$ contains a leaf with at least two distint centers **do**:
  - Sample $i_t \in \{1, 2, \dots, d\}$, $\theta_t \in (0, 1)$, and $\sigma_t \in \{\pm 1\}$ uniformly at random.
  - For each leaf u in the tree $\mathcal{T}_t$ containing more than one center, split node u using *Divide-and-Share* with parameters u, i_t, θ_t, σ_t, and $\varepsilon = \min\{\delta/15 \ln k, 1/384\}$.
  - Update $t = t + 1$.

Figure 3: Threshold Tree Construction

Function Divide-and-Share

Input: a node u, a coordinate $i \in \{1, \dots, d\}$, a positive threshold θ, a number $\sigma \in \{\pm 1\}$, and a parameter ε

Output: if successful, the function splits u into two parts

- Find the median of all centers assigned to node u. Denote it by m^u.
- Let $R_u = \max\{\|c - m^u\|_2 : c \in C_u\}$ be the maximum distance from m^u to one of the centers in C_u.
- Let

$$Left = \{c \in C_u : c_i \le m_i^u + \sigma\sqrt{\theta} R_u + \varepsilon\sqrt{\theta} R_u\};$$

$$Right = \{c \in C_u : c_i \ge m_i^u + \sigma\sqrt{\theta} R_u - \varepsilon\sqrt{\theta} R_u\}.$$

- If both sets – $Left$ and $Right$ – are nonempty, then
  - Split u into two parts using cut $(i, m^u + \sigma\sqrt{\theta} R_u)$.
  - Assign the set of centers $Left$ to the left child u_{left} and the set of centers $Right$ to the right child, u_{right}.
- Otherwise, return the unmodified tree (in this case, we say that Divide-and-Share *fails*).

Figure 4: Function Divide-and-Share

Function *Divide-and-Share* first finds a median[3] of all centers assigned to u, which we denote by m^u. Let R_u be the maximum distance from centers in node u to the median m^u. The algorithm creates two child nodes for u using cut $\omega_t = (i_t, \xi_t)$ with $\xi_t = m_i^u + \sigma_t\sqrt{\theta_t} R_u$. Then, *Divide-and-Share* assigns two sets of centers, $Left$ and $Right$, defined in Figure 4 to the left and right children of u, respectively. Note that these sets share centers in the strip of width $2\varepsilon\sqrt{\theta_t} R_u$:

$$Left \cap Right = \{c \in C_u : \xi_t - \varepsilon\sqrt{\theta_t} R_u \le c_i \le \xi_t + \varepsilon\sqrt{\theta_t} R_u\}.$$

If one of the sets, $Left$ or $Right$, is empty, then *Divide-and-Share* discards both newly created children of u.

We show that the bi-criteria approximation factor of the algorithm is $O(1/\delta \log^2 k \log\log k)$ and the expected number of leaves is $(1+\delta)k$. In the next section, we give a proof overview. Then,

[3] Median m^u satisfies the following property: For ever coordinate i, each of the sets $\{c \in C_u : c_i < m_i^u\}$ and $\{c \in C_u : c_i > m_i^u\}$ contains at most half of all points from C_u.

we prove the upper bounds on the expected number of leaves and approximation factor of the algorithm in Sections 5 and 6, respectively.

4 PROOF OVERVIEW

In this section, we provide an overview of the analysis of our algorithm, give definitions, and discuss the motivation for the proofs. In Sections 5 and 6, we present detailed proofs.

4.1 Cost of Clustering

We first analyze approximation guarantees for our algorithm. We show that the expected approximation factor is $O(1/\delta \log^2 k \log\log k)$ $= O(1/\varepsilon \log k \log\log k)$, particularly for constant δ (e.g., $\delta = 0.05$), the expected approximation factor is $O(\log^2 k \log\log k)$. We denote the final tree returned by the algorithm by $\mathcal{T}$. Let $\mathcal{T}(x)$ be the center assigned by the threshold tree $\mathcal{T}$ to point x.

Theorem 4.1. *For every set of centers $c^1, \dots, c^k$ in $\mathbb{R}^d$, every $\delta \in (0,1)$, and every $x \in \mathbb{R}^d$, we have*

$$\mathbf{E}\left[\|x - \mathcal{T}(x)\|_2^2\right] \le O(1/\delta \ \log^2 k \log\log k) \min_{c \in \{c^1,\dots,c^k\}} \|x-c\|_2^2. \quad (3)$$

This theorem guarantees that the expected approximation factor for every point x is at most $O(1/\delta \ \log^2 k \log\log k)$. Consequently, the expected approximation factor for any data set X is also bounded by $O(1/\delta \ \log^2 k \log\log k)$.

Fix an arbitrary point x for the entire proof of Theorem 4.1. If x equals one of the centers c^i, then $\mathcal{T}(x)$ also always equals c^i. Hence, $\|x - \mathcal{T}(x)\|_2^2 = 0$ and bound (3) trivially holds. So, from now on, we will assume that x is not one of the centers.

Denote by $\mathcal{T}_t$ the tree built by the algorithm in the first $(t-1)$ steps. Tree $\mathcal{T}_1$ contains only one node – the root. The root corresponds to the entire space $\mathbb{R}^d$ and all centers $c^1, \dots, c^k$ are assigned to it. Since point x is fixed, we will only consider nodes u in $\mathcal{T}$ that contain x. Let u_t be the leaf node of the tree $\mathcal{T}_t$ that contains x. That is, u_t is the leaf node that contains x at the beginning of iteration t. Nodes $u_1, u_2, \dots$ form a path in the tree $\mathcal{T}$ from the root to the unique leaf of $\mathcal{T}$ that contains x. To simplify notation, we denote

$$C_t = C_{u_t}, \quad R_t = R_{u_t}, \quad m^t = m^{u_t}.$$

Also, let D_t be the diameter of set C_t:

$$D_t = \max\{\|c' - c''\|_2 : c', c'' \in C_t\}.$$

Finally, let $\mathcal{T}_t(x)$ be the closest center from the set C_t to point x. We call this center the tentative center for point x at step t. The tentative cost of x at step t is $\|x - \mathcal{T}_t(x)\|_2^2$.

Initially, at step 1, the tentative center for point x is the closest center $c \in \{c^1, \dots, c^k\}$ to x. If the tentative center for x does not change, then the eventual cost of x, $\|x - \mathcal{T}(x)\|_2^2$ exactly equals the optimal cost $\|x - c\|_2^2$. However, at some step t, point x may be separated from its tentative center c (see below for a formal definition), in which case another tentative center $\mathcal{T}_{t+1}(x)$ is assigned to x. At this step, the tentative cost of x may significantly increase. Moreover, the tentative cost of x may further increase if x is separated from the new tentative center. Our goal is to give an upper bound on the expected total cost increase.

Definition 4.2. *We say that x is separated from its tentative center $c = \mathcal{T}_t(x)$ at step t, if $c \notin C_{t+1}$.*

Note that x is separated from its tentative center $c = \mathcal{T}_t(x)$ at step t if and only if c is no longer the tentative center for x at step $t+1$ ($\mathcal{T}_{t+1}(x) \neq \mathcal{T}_t(x)$). We now define A_k. Loosely, speaking A_k is the approximation factor of the algorithm for the given set of centers $c^1, \dots, c^k$ and point x. For technical reasons, the formal definition is more involved.

Definition 4.3. *Let A_k be the smallest number such that the following inequality holds with probability 1 for every partially built tree $\mathcal{T}_t$:*

$$\mathbf{E}\left[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_t\right] \le A_k \ \|x - \mathcal{T}_t(x)\|_2^2. \quad (4)$$

In this definition, $\mathbf{E}\left[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_t\right]$ is the conditional expectation of the eventual cost of x given that at step t the partially built tree is $\mathcal{T}_t$. Thus, if at some step t, the tentative center for x is c, then the expected final cost $\mathbf{E}[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_t]$ is upper bounded by $A_k \ \|x - c\|_2^2$. Observe, that A_k is well defined and finite, because $\mathcal{T}(x)$ and $\mathcal{T}_t(x)$ take at most k different values (namely, values in $\{c^1, \dots, c^k\}$).

We show an upper bound of $O(1/\varepsilon \log k \log\log k)$ on A_k (note: $\varepsilon = \min\{\delta/15\ln k, 1/384\}$). To illustrate the proof, we make a number of simplifying assumptions in this section. The actual proof is considerably more involved. We give it in Section 6.

Informal Proof of the Upper Bound on A_k. Suppose c^* is the tentative center for x at step t^*. If at some step $t \ge t^*$, center c^* is separated from x, then we assign a new tentative center to x. We call this center a *fallback* center for x. This fallback center depends on the tree $\mathcal{T}_t$ and cut (i, ξ) that separates x and c^*. However, to illustrate the idea behind the proof, let us assume that the distance from the *fallback* center to x does not depend on the cut (i, ξ). Specifically, we suppose that the distance from x to the fallback center is M_t at step t *for every cut* (i, ξ).

We consider four possibilities:

A. Point x and c^* are never separated.
B. Point x is separated from c^* at step t and $D_t^2 \le \|x - c^*\|_2^2$.
C. Point x is separated from c^* at step t and $\|x - c^*\|_2^2 < D_t^2 \le A_k M_t^2/2$.
D. Point x is separated from c^* at step t and $D_t^2 > A_k M_t^2/2$.

In case (A), the cost of x in the resulting tree $\mathcal{T}$ equals $\|x - c^*\|_2^2$. In cases (B) and (C), the eventual cost of x is upper bounded by $(D_t + \|x - c^*\|_2)^2 \le 2D_t^2 + 2\|x - c^*\|_2^2$ because no matter which center c^{**} in C_t is assigned to x in $\mathcal{T}$, the distance from c^{**} to x is at most $\|x - c^*\|_2 + \|c^* - c^{**}\|_2 \le \|x - c^*\|_2 + D_t$ (note: D_t is the maximum distance between centers in C_t). Furthermore, in case (B), $2D_t^2 + 2\|x - c^*\|^2 \le 4\|x - c^*\|^2$. In case (D), after step t, the distance from x to the new tentative center is M_t. Hence, by the definition of A_k (see Definition 4.3), the expected cost of x in $\mathcal{T}$ is bounded by $A_k M_t^2$. To summarize, in case (A) or (B), the final cost of x is at most $4\|x - c^*\|_2^2$. In case (C) and (D), the final cost is upper bounded by $2\|x - c^*\|_2^2 + \min\left\{2D_t^2, A_k M_t^2\right\}$, where t is the step when x and c^* are separated.

Let t^{**} be the first step t of the algorithm, when $D_t \le \|x - c^*\|_2$ or c^* is no longer the tentative center for x. Note that for some step t, C_t contains only one center and $D_t = 0$. Hence, the stopping time t^{**} is well defined. Let

$$p_{x,c^*,t} = \mathbb{P}\{x \ \&\ c^* \text{ are separated at step } t \mid \mathcal{T}_t\}$$

be the probability that point x is separated from c^* at step t conditioned on $\mathcal{T}_t$. Then, we have

$$\mathbf{E}[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_{t^*}] \le \\ \le 4\|x - c^*\|_2^2 + \mathbf{E}\left[\sum_{t=t^*}^{t^{**}-1} p_{x,c^*,t} \min\left\{2D_t^2, A_k M_t^2\right\} \mid \mathcal{T}_{t^*}\right].$$

We need to estimate the probability that x and c^* are separated at step t. Observe that if x and c^* are separated, then $x_i - m_i^t \le \sigma_t\sqrt{\theta_t}R_t$ and $c_i^* - m_i^t \ge (\sigma_t + \varepsilon)\sqrt{\theta}R_t$ or $x_i - m_i^t \ge \sigma_t\sqrt{\theta}R_t$ and $c_i^* - m_i^t \le (\sigma_t - \varepsilon)\sqrt{\theta}R_t$, where $i = i_t$ is the coordinate chosen by the algorithm. We consider the case when x_i and c_i^* are on the same side of m_i^t, i.e. $(x_i - m_i^t)(c_i^* - m_i^t) \ge 0$. The case when x_i and c_i^* are on the opposite sides of m_i^t is handled similarly. Since θ_t is uniformly distributed in $[0,1]$ and coordinate i_t is chosen randomly from $\{1,\dots,d\}$, we have

$$p_{x,c^*,t} = \mathbb{P}\{x \ \&\ c^* \text{ are separated at step } t \mid \mathcal{T}_t\} \le \\ \le \frac{1}{d\,R_t^2}\sum_{i=1}^{d} \max\left\{\frac{|c_i^* - m_i^t|^2}{(1+\varepsilon)^2} - |x_i - m_i|^2, |x_i - m_i|^2 - \frac{|c_i^* - m_i^t|^2}{(1-\varepsilon)^2}, 0\right\}.$$

Remark: In the formula above, we divide $|c_i^* - m_i^t|^2$ by $(1+\varepsilon)^2$ and $|c_i^* - m_i^t|^2$ by $(1-\varepsilon)^2$. These factors – $1/(1+\varepsilon)^2$ and $1/(1-\varepsilon)^2$ – are essential for the analysis. If we did not have them, we would get $\tilde{\Theta}(k)$ instead of $O(1/\varepsilon \log k \log\log k)$ approximation!

We now use the following inequality: For all positive numbers a, b and $\varepsilon \in (0,1)$, we have

$$\max\left\{\frac{b^2}{(1+\varepsilon)^2} - a^2, b^2 - \frac{a^2}{(1-\varepsilon)^2}\right\} \le \frac{(b-a)^2}{2\varepsilon - \varepsilon^2} \le \frac{(b-a)^2}{\varepsilon}. \quad (5)$$

This inequality can be verified by dividing the left and right hand sides by a^2 and solving the obtained quadratic equation for $\lambda = b/a$. We have

$$p_{x,c^*,t} \le \frac{1}{d\,R_t^2}\sum_{i=1}^{d}\frac{(x_i - c_i^*)^2}{\varepsilon} = \frac{\|x - c^*\|_2^2}{\varepsilon d\,R_t^2}.$$

Note that the separation probability is proportional to the squared distance between x and its tentative center c^* (i.e., $\|x - c^*\|_2^2$) rather than the distance $\|x - c^*\|_2$ itself.

In Section 6, we are going to use a slightly different version of inequality (5) to bound the probability that x and c^* are separated using a particular cut (i, ξ) (see Claim 6.9).

We use the upper bound on the separation probability to obtain a convenient bound on the expected final cost of x:

$$\mathbf{E}[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_{t^*}] \le \\ \le 4\|x - c^*\|_2^2 + \mathbf{E}\left[\sum_{t=t^*}^{t^{**}-1}\frac{\|x - c^*\|_2^2}{\varepsilon d R_t^2}\cdot\min\left\{2D_t^2, A_kM_t^2\right\} \mid \mathcal{T}_{t^*}\right] \\ = \|x - c^*\|_2^2 \cdot \left(4 + \mathbf{E}\left[\frac{1}{\varepsilon d}\sum_{t=t^*}^{t^{**}-1}\frac{\min\left\{2D_t^2, A_kM_t^2\right\}}{R_t^2} \mid \mathcal{T}_{t^*}\right]\right).$$

Thus,

$$\mathbf{E}\left[\frac{\|x - \mathcal{T}(x)\|_2^2}{\|x - c^*\|_2^2} \mid \mathcal{T}_{t^*}\right] \le \\ \le 4 + \mathbf{E}\left[\frac{1}{\varepsilon d}\sum_{t=t^*}^{t^{**}-1}\frac{\min\left\{2D_t^2, A_kM_t^2\right\}}{R_t^2} \mid \mathcal{T}_{t^*}\right]. \quad (6)$$

Our goal is to bound the right hand side of this inequality by $O(1/\varepsilon \log k \log\log k)$.

In Lemma 6.1, we show that $R_t \approx D_t$. Specifically, $1/\sqrt{2}R_t \le D_t \le 2R_t$. This inequality would be trivial if m^t was one of the centers c^j. However, generally speaking, this is not the case. In fact, m^t does not have to belong to the convex hull of centers in C_t. Nevertheless, $D_t \in [1/\sqrt{2}R_t, 2R_t]$ because m^t is the median of C_t (see Lemma 6.1).

It is easy to see that the diameter D_t is a non-increasing function of t (since $C_{t+1} \subset C_t$) and M_t is a non-decreasing function of t. In Lemma 6.2, we show that, in fact, D_t decreases by a factor of 2 every $L = \Theta(d \ln k)$ steps with high probability. That is, $D_{t+L} \le D_t/2$. This happens because for every step t, each pair of centers c' and c'' with $\|c' - c''\|_2 \ge D_t/2$ assigned to u_t is separated with probability at least $\Omega(1/d)$ (see Corollary 6.4). So, in $L = \Theta(d \ln k)$ steps all pairs of centers in C_t at distance at least $D_t/2$ are separated with high probability.

We upper bound the right hand side of (6). Write

$$\frac{1}{\varepsilon d}\sum_{t=t^*}^{t^{**}-1}\frac{\min\left\{2D_t^2, A_kM_t^2\right\}}{R_t^2} \le \\ \le \underbrace{\sum_{\substack{t\in\{t^*,\cdots,t^{**}-1\}\\ A_kM_t^2 \le 2D_t^2}}\frac{A_kM_t^2}{\varepsilon d R_t^2}}_{\Sigma_I} + \underbrace{\sum_{\substack{t\in\{t^*,\cdots,t^{**}-1\}\\ 2D_t^2 < A_kM_t^2}}\frac{2D_t^2}{\varepsilon d R_t^2}}_{\Sigma_{II}} \\ \le \underbrace{\sum_{\substack{t\in\{t^*,\cdots,t^{**}-1\}\\ 2D_t^2 \ge A_kM_t^2}}\frac{4A_kM_t^2}{\varepsilon d D_t^2}}_{\Sigma_I} + \underbrace{\sum_{\substack{t\in\{t^*,\cdots,t^{**}-1\}\\ 2D_t^2 < A_kM_t^2}}\frac{8}{\varepsilon d}}_{\Sigma_{II}}. \quad (7)$$

Consider the first sum, Σ_I on the right hand side of (7). It is upper bounded by $2L$ times the maximum term in that sum, because D_t halves every L steps and therefore $(M_t/D_t)^2$ increases by 4 times every L steps. The maximum term in Σ_I is, in turn, upper bounded by $8/(\varepsilon d)$ (because $2D_t^2 \ge A_kM_t^2$ for all terms in Σ_I).

Now consider the second sum, Σ_{II} on the right hand side of (7). Let t' be the first step t for which $2D_t^2 < A_kM_t^2$. Using that $D_{t+L} \le D_t/2$, we obtain the following upper bound on the number

of steps $t < t^{**}$ in Σ_{II}:

$$t^{**} - t' \le L + L \cdot \log_2 \frac{D_{t'}}{D_{t^{**}-1}} \le$$
$$\le L + L \cdot \log_2 \frac{\sqrt{A_k/2}\, M_{t'}}{D_{t^{**}-1}} \le L + L \cdot \log_2 \frac{\sqrt{A_k/2}\, M_{t^{**}-1}}{D_{t^{**}-1}}.$$

The last inequality holds because M_t is a non-decreasing function of t. Recall, that the distance to the fallback center, M_t is upper bounded by $\|x-c^*\|_2 + D_t$ for every step $t \in \{t^*, \cdots, t^{**}-1\}$. Also, by the definition of stopping time t^{**}, for every $t < t^{**}$, we have $D_t > \|x - c^*\|_2$. Thus,

$$\frac{M_{t^{**}-1}}{D_{t^{**}-1}} \le \frac{\|x-c^*\|_2 + D_{t^{**}-1}}{D_{t^{**}-1}} \le 2.$$

Therefore, $t^{**} - t' \le L \cdot (1 + \log_2 \sqrt{2A_k})$. Consequently, the second sum, Σ_I as well as $\Sigma_I + \Sigma_{II}$ are upper bounded by $O((L \log A_k)/(\varepsilon d)) = O(1/\varepsilon \log k \log A_k)$. We obtained the following bound:

$$\mathbf{E}\left[\frac{\|x - \mathcal{T}(x)\|_2^2}{\|x-c^*\|_2^2} \mid \mathcal{T}_{t^*}\right] \le O(1/\varepsilon \log k \log A_k).$$

Therefore, $A_k \le O(1/\varepsilon \log k \log A_k)$. This recurrence relation gives us an upper bound of $O(1/\varepsilon \log k \log\log k)$ on A_k. This concludes the proof overview of Theorem 3.

4.2 Expected Number of Leaves

We show that the expected number of leaves in the threshold tree given by our algorithm is at most $e^{\delta/2}k$. Particularly, for $\delta \in (0,1)$, the expected number of leaves is at most $(1+\delta)k$. We now give an overview of the analysis. We provide a complete proof in Section 5.

In this section, we consider the case when the space is one dimensional. That is, all centers and data points lie on the real line. Consider a fixed center c. Let $N_c(\mathcal{T})$ be the number of leaves in tree $\mathcal{T}$ containing c. We show that $\mathbf{E}[N_c(\mathcal{T})]$ is at most $e^{\delta/2}$.

Suppose c is assigned to node u at step t (note that c may be assigned to several nodes). Denote the total number of centers assigned to u by $k' = |C_u|$. We prove by induction on k' that the expected number of leaves to which u is assigned in the subtree rooted at u is at most $(1+5\varepsilon)^{\log_2 k'}$. If $k' = 1$, then the claim trivially holds, since u is a leaf. Assume $k' > 1$.

Our algorithm divides u into two parts u_{left} and u_{right}. One of them contains the median m^u. We call that part the main child and denote it by u'. In turn, the main child u' is also divided into two parts, one of them – denoted by u'' – is the main child of u'. We call the sequence of nodes $u, u', u'', \ldots$ the main branch rooted at u. Note that the main child always contains at least half of all centers assigned to its parent. This is the case, because m^u is the median of all centers assigned to u. Thus, the part containing m^u contains at least half of all centers in C_u, and the other (secondary) child contains at most half of all centers in C_u.

Suppose that center c is assigned to a node v in the main branch $u, u', u'', \ldots$. When v is divided into two parts, one of the following three events may occur: (1) c is assigned only to the main child of v; (2) c is assigned to both the main and secondary children of v; (3) c is assigned only to the secondary child of v. Denote these events by $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, respectively. We estimate the number of nodes w such that c is assigned to w, and w is a secondary child of a node in the main branch. This number equals to the number of events $\mathcal{E}_2$ that occur in the main branch before the first event $\mathcal{E}_3$ occurs plus 1. If the probabilities of events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ were the same for all nodes in the main branch containing c, the expected number above would be equal to $1/\mathbb{P}(\mathcal{E}_3 \mid \mathcal{E}_2 \cup \mathcal{E}_3)$. Without loss of generality assume that $m^u = 0$, then for $\varepsilon \le 1/10$, we have

$$\frac{1}{\mathbb{P}(\mathcal{E}_3 \mid \mathcal{E}_2 \cup \mathcal{E}_3)} = \frac{\mathbb{P}(\mathcal{E}_2 \cup \mathcal{E}_3)}{\mathbb{P}(\mathcal{E}_3)} = \frac{c^2}{(1-\varepsilon)^2 R_t^2} \Big/ \frac{c^2}{(1+\varepsilon)^2 R_t^2}$$
$$= \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} \le 1 + 5\varepsilon.$$

Every secondary child w contains at most $k'/2$ centers. So, by the inductive hypothesis, the expected number of leaves containing c in the subtree rooted at w is at most $(1+5\varepsilon)^{\lfloor \log_2 k'/2 \rfloor}$. Therefore, the expected number of leaves containing c in the subree rooted at u is at most

$$(1+5\varepsilon) \cdot (1+5\varepsilon)^{\lfloor \log_2 k'/2 \rfloor} \le (1+5\varepsilon)^{\lfloor \log_2 k' \rfloor}.$$

This concludes the proof of the inductive claim. We now observe that

$$\mathbf{E}[N_c(\mathcal{T})] \le (1+5\varepsilon)^{\lfloor \log_2 k \rfloor} \le e^{\delta/2}$$

for $\varepsilon \le \frac{\delta}{15 \ln k}$.

5 EXPECTED NUMBER OF LEAVES

In this section, we prove a bound the expected number of leaves in the threshold tree constructed by our algorithm. Our algorithm assigns all centers $c^1, \ldots, c^k$ to the root r of the threshold tree $\mathcal{T}$. Then, it recursively divides centers assigned to every node u between its children. However, centers in a narrow strip $Left \cap Right$ are shared by the both children of node u. Thus, the total number of leaves in the threshold tree $\mathcal{T}$ may be larger than k. Let $N(\mathcal{T})$ be the number of leaves in $\mathcal{T}$. We show an upper bound of $e^{\delta/2}k$ on the expected number of leaves $\mathbf{E}[N(\mathcal{T})]$, where the expectation is over the randomness of our algorithm.

Theorem 5.1. *For every set of centers $c^1, c^2, \ldots, c^k$ in $\mathbb{R}^d$ and every $\delta \in (0, \ln k/32)$, the expected number of leaves in the threshold tree $\mathcal{T}$ given by our algorithm is at most*

$$\mathbf{E}_{\mathcal{T}}[N(\mathcal{T})] \le e^{\delta/2}k.$$

In particular, for $\delta \in (0,1)$,

$$\mathbf{E}_{\mathcal{T}}[N(\mathcal{T})] \le (1+\delta)k.$$

Proof. For every center c, we bound the expected number of leaves containing c by $e^{\delta/2}$. Consider a fixed center c. For a node u in the threshold tree $\mathcal{T}$, let $N_c^u(\mathcal{T})$ denote the number of leaves in the subtree of $\mathcal{T}$ rooted at node u to which center c is assigned to.

Definition 5.2. *For every integer $k' \in \{1, 2, \ldots, k\}$, let $B_{k'}$ be the minimum number such that the following inequality holds for every partially built tree $\mathcal{T}_t$ and every leaf u with $|C_u| \le k'$ in $\mathcal{T}_t$ to which center c is assigned,*

$$\mathbf{E}[N_c^u(\mathcal{T}) \mid \mathcal{T}_t] \le B_{k'}.$$

That is, $B_{k'}$ is an upper bound on the expected number of leaves in the subtree rooted at u that contain c if at most k' centers are assigned to u. To prove Theorem 5.1, it is sufficient to show that B_k is at most $1+\delta$. We derive the following recurrence relation on $B_{k'}$.

Lemma 5.3. *The upper bound on the expected number of leaves $B_{k'}$ satisfies the following recurrence relation:*

$$B_1 = 1, \tag{8}$$

$$B_{k'} \le (1+5\varepsilon)B_{\lfloor k'/2 \rfloor}, \tag{9}$$

where $\varepsilon = \min\{\delta/15\ln k, 1/384\}$.

Proof. It is easy to see that $B_1 = 1$, because if c is the only center assigned to node u, then u is a leaf and $N_c^u(\mathcal{T}) = 1$. We now prove (9). Consider a partially built tree $\mathcal{T}_t$, node u in $\mathcal{T}_t$, and center c in X_u for which inequality (5.2) is tight i.e., $B_{k'} = \mathbf{E}[N_c^u(\mathcal{T}) \mid \mathcal{T}_t]$.

Examine the call of function *Divide-and-Share* that splits node u. Let i_t be the coordinate randomly chosen for this call of function *Divide-and-Share*. Without loss of generality, we assume that $c_i \ge m_i^u$. If σ_t is negative, then center c is assigned only to the right child of u. In this case, the expected number of leaves containing c in the subtree rooted at u is at most $B_{k'}$.

We now consider the case when $\sigma_t = 1$. Define three disjoint events: (1) center c is assigned only to the left child of u and $\sigma_t = 1$; (2) center c is assigned to both children of u and $\sigma_t = 1$; (3) center c is assigned only to the right child of u and $\sigma_t = 1$. Denote these events by $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, respectively.

The number of centers assigned to node u is k'. Thus, the number of centers assigned to each child of u is at most k'. Moreover, if $\sigma_t = 1$, the number of centers assigned to the *right* child u_{right} of u is at most $\lfloor k'/2 \rfloor$, because m^u is the median of all centers in C_u and for all centers c' assigned to u_{right}, $c'_i > m_i^u$. Hence, if event $\mathcal{E}_1$ occurs, then the expected number of leaves containing c in the subtree rooted at u is bounded by $B_{k'}$. If event $\mathcal{E}_2$ occurs, then the expected number of leaves containing c in the subtree rooted at u is bounded by $B_{k'} + B_{\lfloor k'/2 \rfloor}$. Finally, if event $\mathcal{E}_3$ occurs, then the expected number of leaves containing c in the subtree rooted at u is bounded by $B_{\lfloor k'/2 \rfloor}$.Let $p_{t,1} = \mathbb{P}(\mathcal{E}_1 \mid \mathcal{T}_t)$, $p_{t,2} = \mathbb{P}(\mathcal{E}_2 \mid \mathcal{T}_t)$, and $p_{t,3} = \mathbb{P}(\mathcal{E}_3 \mid \mathcal{T}_t)$. Thus,

$$\begin{aligned}
&\mathbf{E}[N_c^u(\mathcal{T}) \mid \mathcal{T}_t] \le \\
&\le 1/2 \cdot B_{k'} + B_{k'} \cdot p_{t,1} + \big(B_{k'} + B_{\lfloor k'/2 \rfloor}\big)p_{t,2} + B_{\lfloor k'/2 \rfloor} \cdot p_{t,3} \\
&= \Big(1/2 + p_{t,1} + p_{t,2}\Big)B_{k'} + \Big(p_{t,2} + p_{t,3}\Big)B_{\lfloor k'/2 \rfloor}.
\end{aligned}$$

Since $1/2 + p_{t,1} + p_{t,2} + p_{t,3} = 1$, we have

$$B_{k'} = \mathbf{E}[N_c^u(\mathcal{T}) \mid \mathcal{T}_t] \le \Big(1 - p_{t,3}\Big)B_{k'} + \Big(p_{t,2} + p_{t,3}\Big)B_{\lfloor k'/2 \rfloor}.$$

Thus,

$$B_{k'} \le \frac{p_{t,2} + p_{t,3}}{p_{t,3}} B_{\lfloor k'/2 \rfloor} = \frac{\mathbb{P}(\mathcal{E}_2 \cup \mathcal{E}_3 \mid \mathcal{T}_t)}{\mathbb{P}(\mathcal{E}_3 \mid \mathcal{T}_t)} B_{\lfloor k'/2 \rfloor}.$$

Compute $\mathbb{P}(\mathcal{E}_2 \cup \mathcal{E}_3 \mid \mathcal{T}_t)$ and $\mathbb{P}(\mathcal{E}_3 \mid \mathcal{T}_t)$:

$$\begin{aligned}
&\mathbb{P}(\mathcal{E}_2 \cup \mathcal{E}_3 \mid \mathcal{T}_t) = \\
&= \frac{1}{2d}\sum_{i=1}^{d} \mathbb{P}\left\{\left|c_i - m_i^t\right| \ge (1-\varepsilon)\sqrt{\theta_t}R_t\right\} = \frac{1}{2d}\sum_{i=1}^{d} \frac{(c_i - m_i^t)^2}{(1-\varepsilon)^2 R_t^2}; \\
&\mathbb{P}(\mathcal{E}_3 \mid \mathcal{T}_t) = \\
&= \frac{1}{2d}\sum_{i=1}^{d} \mathbb{P}\left\{\left|c_i - m_i^t\right| \ge (1+\varepsilon)\sqrt{\theta_t}R_t\right\} = \frac{1}{2d}\sum_{i=1}^{d} \frac{(c_i - m_i^t)^2}{(1+\varepsilon)^2 R_t^2}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
B_{k'} &\le B_{\lfloor k'/2 \rfloor} \cdot \sum_{i=1}^{d} \frac{(c_i - m_i^t)^2}{(1-\varepsilon)^2 R_t^2} \Bigg/ \sum_{i=1}^{d} \frac{(c_i - m_i^t)^2}{(1+\varepsilon)^2 R_t^2} = \\
&= \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} B_{\lfloor k'/2 \rfloor} \le (1+5\varepsilon)B_{\lfloor k'/2 \rfloor}.
\end{aligned}$$

where the last inequality holds because $\varepsilon \le 1/10$. □

We now bound the expected number of leaves in the threshold tree $\mathcal{T}$. By Lemma 5.3, the expected number of leaves containing center c in the threshold tree $\mathcal{T}$ is at most

$$\begin{aligned}
\mathbf{E}[N_c^r(\mathcal{T})] \le B_k &\le (1+5\varepsilon)^{\lfloor \log_2 k \rfloor} \cdot B_1 \le \\
&\le \left(1 + \frac{\delta}{3\ln k}\right)^{\log_2 k} \le \left(e^{\frac{\delta}{3\ln k}}\right)^{\log_2 k} < e^{\delta/2}.
\end{aligned}$$

Since $e^{\delta/2} < 1 + \delta$ for $\delta \in (0,1)$, we have for $\delta \in (0,1)$

$$\mathbf{E}[N_c^r(\mathcal{T})] \le e^{\delta/2} \le 1 + \delta.$$

□

6 APPROXIMATION FACTOR

We now prove Theorem 4.1. Our proof follows the outline given in Section 4. We fix a point x, step t^*, and estimate $\mathbf{E}[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_{t^*}]$. Let $c^* = \mathcal{T}_{t^*}(x)$ be the tentative center assigned to x at step t^*. As in Section 4, let u_t be the leaf node of $\mathcal{T}_t$ that contains x, $C_t = C_{u_t}$, $R_t = R_{u_t}$, and $m^t = m^{u_t}$. We denote the diameter of C_t by D_t.

6.1 Bounds on the Diameter

We prove several facts about the diameter D_t. First, we show that $D_t \approx R_t$.

Lemma 6.1. *For every leaf node u in a partially built tree $\mathcal{T}_t$, we have*

$$1/\sqrt{2}R_u \le D_u \le 2R_u.$$

Proof. The second bound easily follows from the triangle inequality: for every c' and c'' in C_u,

$$\|c' - c''\|_2 \le \|c' - m^u\|_2 + \|m^u - c''\|_2 \le 2R_u.$$

We now show the first bound. Let c be the farthest center in C_u from m^u. Then, $R_u = \|c - m^u\|_2$. Consider a center c' in C_u. The distance between c and c' is upper bounded by D_u because D_u is the diameter of C_u. Hence, for each c' in C_u, we have $\|c - c'\|_2^2 \le D_u^2$. Thus,

$$\begin{aligned}
D_u^2 &\ge \mathrm{Avg}_{c' \in C_u} \|c - c'\|_2^2 = \\
&= \mathrm{Avg}_{c' \in C_u} \sum_{i=1}^{d} |c_i - c'_i|^2 = \sum_{i=1}^{d} \mathrm{Avg}_{c' \in C_u} |c_i - c'_i|^2,
\end{aligned}$$

where $\mathrm{Avg}_{c' \in C_u} f(c')$ denotes the average of f over c' in C_u. Observe that

$$\mathrm{Avg}_{c' \in C_u} |c_i - c'_i|^2 \ge 1/2 |c_i - m_i^u|^2.$$

This is because m^u is the median point in C_u, consequently, at least a half of all points $c' \in C_u$ are on the other side of the hyperplane

$\{x : x_i = m_i^u\}$ from c (including centers c' on the hyperplane). For these centers c', we have $|c_i - c_i'| \geq |c_i - m_i^u|$. Therefore,

$$D_u^2 \geq \sum_{i=1}^{d} \mathrm{Avg}_{c' \in C_u} |c_i - c_i'|^2 \geq 1/2 \sum_{i=1}^{d} |c_i - m_i^u|^2 = 1/2 R_u^2.$$

□

We prove that the diameter D_t is exponentially decaying with t. To this end, we estimate the probability that two centers c' and c'' with $\|c' - c''\|_2 \geq D_t/2$ are separated at step t. We say that two centers $c', c'' \in C_t$ are separated at step t if $c' \notin C_t$ or $c'' \notin C_t$.

Lemma 6.2. *For every two centers $c', c'' \in C_t$ at distance at least $D_t/2$,*

$$\Pr\left\{c' \notin C_{t+1} \text{ or } c'' \notin C_{t+1} \mid \mathcal{T}_t\right\} \geq 1/128d.$$

Proof. Suppose, at step t, the algorithm picks coordinate $i_t = i$. For every two centers $c', c'' \in C_t$, we consider the following two cases: (1) c' and c'' are on the same side of the median m^t in coordinate i (i.e. $\mathrm{sign}(c_i' - m_i^t) = \mathrm{sign}(c_i'' - m_i^t)$), and (2) c' and c'' are on the opposite sides of the median m^t in coordinate i (i.e. $\mathrm{sign}(c_i' - m_i^t) = -\mathrm{sign}(c_i'' - m_i^t)$).

Consider the first case, when c' and c'' are on the same side of the median m^t in coordinate i. Without loss of generality, assume that $c_i'' \geq c_i' \geq m_i^t$. Observe that if $\sigma_t = 1$, $c_i'' - m_i^t > (1+\varepsilon)R_t\sqrt{\theta_t}$, and $c_i' - m_i^t \leq (1-\varepsilon)R_t\sqrt{\theta_t}$, then centers c' and c'' are separated at step t. Let $\mathcal{E}_{t,i,c'} = \{i_t = i, \sigma_t = 1\}$ be the event that the threshold cut at step t is in coordinate i and $\sigma_t = 1$. Then, the conditional probability that c' and c'' are separated given $\mathcal{E}_{t,i,c'}$ is

$$\begin{aligned}\mathbb{P}\Big[c_i' - m_i^t \leq (1-\varepsilon)R_t\sqrt{\theta_t} \;\&\; c_i'' - m_i^t > (1+\varepsilon)R_t\sqrt{\theta_t} \mid \mathcal{T}_t,\ \mathcal{E}_{t,i,c'}\Big] \\ = \mathbb{P}\Big\{\theta_t \in \Big[\frac{(c_i' - m_i^t)^2}{(1-\varepsilon)^2 R_t^2}, \frac{(c_i'' - m_i^t)^2}{(1+\varepsilon)^2 R_t^2}\Big]\Big\} \\ = \left(\frac{(c_i'' - m_i^t)^2}{(1+\varepsilon)^2 R_t^2} - \frac{(c_i' - m_i^t)^2}{(1-\varepsilon)^2 R_t^2}\right)^+,\end{aligned}$$

where $(x)^+$ denotes $\max\{x, 0\}$.

Now, consider the second case, when c' and c'' are on the opposite sides of the median m^u in coordinate i. Assume without loss of generality that $c_i'' \geq m_i^t \geq c_i'$ and $|c_i'' - m_i^t| \geq |c_i' - m_i^t|$. If $c_i'' - m_i^u \geq (1+\varepsilon)R_t\sqrt{\theta_t}$ and $\sigma_t = 1$, then c' and c'' are separated at this step. Thus, the conditional probability that c' and c'' are separated given $i_t = i$ and parameter $\sigma_t = 1$ is at least

$$\mathbb{P}\left\{c_i'' - m_i^t \geq (1+\varepsilon)R_t\sqrt{\theta_t} \mid \mathcal{T}_t, i_t = i, \sigma_t = 1\right\} = \frac{(c_i'' - m_i^t)^2}{(1+\varepsilon)^2 R_t^2}.$$

Define

$$a_i = \min\left\{|c_i' - m_i^t|, |c_i'' - m_i^t|\right\} \quad \text{and} \quad b_i = \max\left\{|c_i' - m_i^t|, |c_i'' - m_i^t|\right\}.$$

Let $I_1, I_2 \subset \{1, 2, \ldots, d\}$ be the set of indices i for which c_i' and c_i'' lie on the same side and opposite sides of m^t, respectively. Then,

$$\begin{aligned}&\Pr\left\{c' \notin C_{t+1} \text{ or } c'' \notin C_{t+1} \mid \mathcal{T}_t\right\} \geq \\ &\geq \frac{1}{2d}\sum_{i \in I_1}\Big(\frac{b_i^2}{(1+\varepsilon)^2 R_t^2} - \frac{a_i^2}{(1-\varepsilon)^2 R_t^2}\Big)^+ + \frac{1}{2d}\sum_{i \in I_2}\frac{b_i^2}{(1+\varepsilon)^2 R_t^2}.\end{aligned}$$

Now observe that

$$\begin{aligned}\frac{1}{2d}\sum_{i \in I_1}\Big(\frac{b_i^2}{(1+\varepsilon)^2 R_t^2} - \frac{a_i^2}{(1-\varepsilon)^2 R_t^2}\Big)^+ &\geq \\ &\geq \frac{1}{2dR_t^2}\sum_{i \in I_1}\frac{b_i^2}{(1+\varepsilon)^2} - \frac{a_i^2}{(1-\varepsilon)^2} \\ &\geq \frac{1}{2dR_t^2}\sum_{i \in I_1} b_i^2 - a_i^2 - (2\varepsilon b_i^2 + 3\varepsilon a_i^2).\end{aligned}$$

Similarly, we have

$$\frac{1}{2d}\sum_{i \in I_2}\frac{b_i^2}{(1+\varepsilon)^2 R_t^2} \geq \frac{\sum_{i \in I_2} b_i^2 - 2\varepsilon b_i^2}{2R_t^2 d}.$$

When c' and c'' are on the same side of m^t in coordinate i, we have

$$b_i^2 - a_i^2 = (b_i - a_i)(b_i + a_i) \geq (b_i - a_i)^2 = (c_i' - c_i'')^2.$$

When c' and c'' are on the opposite side of m^t in coordinate i, we have

$$4b_i^2 \geq (b_i + a_i)^2 = (c_i' - c_i'')^2.$$

Note that $\sum_{i=1}^d b_i^2 + a_i^2 = \sum_{i=1}^d (c_i' - m_i^t)^2 + (c_i'' - m_i^t)^2 = \|c' - m^t\|_2^2 + \|c'' - m^t\|_2^2$. Therefore, the probability that c' and c'' are separated at step t is at least

$$\begin{aligned}\mathbb{P}\left\{c' \notin C_{t+1} \text{ or } c'' \notin C_{t+1} \mid \mathcal{T}_t\right\} &\geq \sum_{i=1}^{d}\frac{(c_i' - c_i'')^2}{8dR_t^2} - \frac{2\varepsilon b_i^2 + 3\varepsilon a_i^2}{2dR_t^2} \\ &\geq \frac{\|c' - c''\|_2^2}{8R_t^2 d} - \frac{6\varepsilon}{2d}.\end{aligned}$$

where the second inequality is due to $\sum_{i=1}^d 2\varepsilon b_i^2 + 3\varepsilon a_i^2 \leq \sum_{i=1}^d 3\varepsilon b_i^2 + 3\varepsilon a_i^2 = 3\varepsilon\|c' - m^t\|_2^2 + 3\varepsilon\|c'' - m^t\|_2^2 \leq 6\varepsilon R_t^2$. We conclude that for centers c' and c'' with $\|c' - c''\|_2^2 \geq D_t^2/4$, we have

$$\begin{aligned}\mathbb{P}\left\{c' \notin C_{t+1} \text{ or } c'' \notin C_{t+1} \mid \mathcal{T}_t\right\} &\geq \frac{1}{2d}\cdot\Big(\frac{D_t^2}{16R_t^2} - 6\varepsilon\Big) \\ &\geq \frac{1}{2d}\cdot\Big(\frac{1}{32} - 6\varepsilon\Big) \geq \frac{1}{128d}.\end{aligned}$$

Here, we used that $D_t \geq 1/\sqrt{2}R_t$ and $\varepsilon \leq 1/384$. □

We obtain the following corollary from Lemma 6.1.

Lemma 6.3. *Let $L = \lceil 640d \ln k \rceil$. Then, for every t, we have*

$$\mathbb{P}\left\{D_{t+L} \geq D_t/2 \mid \mathcal{T}_t\right\} \leq \frac{1}{k^3}.$$

Proof. Consider a fixed time step t. Suppose the distance between centers c' and c'' is at least $D_t/2$. Since the diameter D_t is non-increasing as t increases, the distance between c' and c'' is greater than $D_{t'}/2$ for any step $t' \geq t$. By Lemma 6.2, the probability that these centers c' and c'' are separated at step t' is at least $1/128d$.

Thus, these two centers c' and c'' are not separated in $\lceil 640d \ln k \rceil$ steps from step t with probability at most

$$\left(1 - \frac{1}{128d}\right)^{640d \ln k} \leq e^{-5 \ln k}.$$

Since there are at most $\binom{k}{2}$ pairs of centers with distance greater than $D_t/2$, by the union bound over all such pairs, we have for $L = \lceil 640d \ln k \rceil$

$$\mathbb{P}\{D_{t+L} \geq D_t/2 \mid \mathcal{T}_t\} \leq \binom{k}{2} \cdot e^{-5\ln k} \leq \frac{1}{k^3}.$$

□

To simplify the exposition, we define a stopping time t^{**}. Let t^{**} be the first step $t > t^*$ of the algorithm when one of the following happens: (A) $D_t \leq \|x - c^*\|_2$ (note: if c^* is the only center remaining in C_t, then $D_t = 0$); (B) x and c^* are separated before step t (i.e., $c^* \notin C_t$); or (C) $D_t > D_{t-L'}/2$ and $t \geq t^* + L'$ for $L' = \lceil 1280d \ln k \rceil$. For some step t, C_t contains only one center and $D_t = 0$. Thus, the stopping time t^{**} is well-defined. We show that it is very unlikely that the case (C) happens, i.e. $D_{t^{**}} > D_{t^{**}-L'}/2$ and $t^{**} \geq t^* + L'$.

Corollary 6.4. *Let $L' = \lceil 1280d \ln k \rceil$ be twice as large as L in Lemma 6.3. Then,*

$$\mathbb{P}\left\{D_{t^{**}} > D_{t^{**}-L'}/2 \;\&\; t^{**} \geq t^* + L' \mid \mathcal{T}_{t^*}\right\} \leq \frac{1}{k}.$$

Proof. Let $L = \lceil 640d \ln k \rceil$ be as in Lemma 6.3. We consider the set of steps

$$S_L = \{t \leq t^{**} : t = t^* + Lz, z \geq 1\}.$$

By Lemma 6.3, we have for each step $t = t^* + Lz$ in this set S_L

$$\mathbb{P}\{D_t > D_{t-L}/2 \mid \mathcal{T}_{t-L}\} \leq \frac{1}{k^3}.$$

We consider every step $t = t^* + L'z$ for $z \geq 1$. If $D_t > D_{t-L'}/2$, then we have $t^{**} \leq t$. If $D_t \leq D_{t-L'}/2$, then we must separate at least one center from $C_{t-L'}$ in L' steps, which means $|C_t| < |C_{t-L'}|$. Since there are at most k centers in C_{t^*}, we have at most k such steps t with $D_t \leq D_{t-L'}/2$. Thus, we have $t^{**} \leq t^* + L'k = t^* + 2kL$. Then, the set of steps S_L contains at most $2k$ steps. By the union bound over all steps $t \in S_L$, we have $D_t \leq D_{t-L}/2$ for all steps $t \in S_L$ with probability at least $1 - 1/k$. Suppose that $D_t \leq D_{t-L}/2$ holds for all steps $t \in S_L$. For every $t^* + L' \leq t \leq t^{**}$, there exists a $t' \in S_L$ such that $t - L' \leq t' - L < t' \leq t$. Since D_t is a non-increasing sequence, we have for every $t^* + L' \leq t \leq t^{**}$

$$D_t \leq D_{t'} \leq D_{t'-L}/2 \leq D_{t-L'}/2.$$

Therefore, we have $D_{t^{**}} > D_{t^{**}-L'}/2$ and $t^{**} \geq t^* + L'$ with probability at most $1/k$. □

6.2 Cost of Separation

In this section, we complete the proof of Theorem 4.1. The proof is similar to the overview we gave in Section 4. The key difference is that we no longer assume that the distance from x to the nearest fallback center does not depend on the cut that separates x and c^*.

To simplify the exposition, from now on, we shall assume that $c^*_i \geq x_i$ for all i. We make this assumption without loss of generality, because if $c^*_i < x_i$ for some i, we can mirror all centers c in C and point x across the hyperplane $\{y_i = 0\}$, or, in other words, we can change the sign of the i-th coordinate for all centers c in C and point x. This transformation does not affect the algorithm but makes $c^*_i \geq x_i$.

For every (i, η) with $x_i \leq \eta < c_i$, define $M_t(i, \eta)$ as follows: $M_t(i, \eta)$ equals the distance from x to the closest center c' in C_t with $c'_i \leq \eta$. If there are no centers c' in C_t with $c'_i \leq \eta$, then we let $M_t(i, \eta) = \infty$. Observe that if x and c^* are separated at step t, then

$$x_i \leq m^t_i + \sigma_t\sqrt{\theta_t}R_t < \underbrace{m^t_i + \sigma_t\sqrt{\theta_t}R_t + \varepsilon\sqrt{\theta_t}R_t}_{\eta_t} < c^*_i,$$

where i is the coordinate chosen at step t. Thus, if x and c^* are separated at step t, the distance from x to the fallback center is $M_t(i, \eta_t)$, where $\eta_t = m^t_i + \sigma_t\sqrt{\theta_t}R_t + \varepsilon\sqrt{\theta_t}R_t$.

At each step t, our algorithm calls function *Divide-and-Share* with parameters $(i_t, \sigma_t, \theta_t)$ to split node u_t. Let $\omega_t = (i_t, \xi_t)$ be the cut chosen by the algorithm for node u_t where $\xi_t = m^t_i + \sigma_t\sqrt{\theta_t}R_t$; ω_t is undefined ($\omega_t = \perp$), if the algorithm does not make any cut at step t. Note that the cut ω_t is determined by the tuple $(i_t, \sigma_t, \theta_t)$. Then, x and c^* are separated at step t by the tuple (i, σ, θ) if $c^* \in C_t$, $\omega_t = (i, m^t_i + \sigma\sqrt{\theta}R_t)$ and $x_i \leq \xi_t < \eta_t < c^*_i$. Let $f_{x,t}(i, \sigma, \theta)$ be the indicator of the event x and c^* are separated at step t by the tuple (i, σ, θ). If $x_i \leq m^t_i + \sigma\sqrt{\theta}R_t < m^t_i + (\sigma + \epsilon)\sqrt{\theta}R_t < c^*_i$, then $f_{x,t}(i, \sigma, \theta) = 1$; otherwise, $f_{x,t}(i, \sigma, \theta) = 0$.

We define a penalty function $Z_t(i, \sigma, \theta)$ for every tuple (i, σ, θ) with $i \in \{1, 2 \ldots, d\}$, $\sigma \in \{\pm 1\}$, $\theta \in (0, 1)$ as follows:

$$Z_t(i, \sigma, \theta) = \mathbf{E}\left[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_t,\ (i_t, \sigma_t, \theta_t) = (i, \sigma, \theta)\right] f_{x,t}(i, \sigma, \theta).$$

In other words, $Z_t(i, \sigma, \theta)$ equals 0 if the tuple (i, σ, θ) does not separate x and c^* at step t. Otherwise, it is equal to the expected cost of x in the final tree $\mathcal{T}$ assuming that the algorithm chooses the tuple (i, σ, θ) at step t. Note that if x and c^* are already separated at step t, then $Z_t(i, \sigma, \theta) = 0$.

Claim 6.5. *For every step t and every tuple (i, σ, θ), we have*

$$Z_t(i, \sigma, \theta) \leq \min\left\{2\|x - c^*\|_2^2 + 2D_t^2, A_k\ M_t^2(i, \eta)\right\},$$

where $\eta = m^t_i + (\sigma + \varepsilon)\sqrt{\theta}R_t$.

Proof. If x and c^* are not separated by the tuple (i, σ, θ) at step t or x and c^* are already separated at step t, then we have $Z_t(i, \sigma, \theta) = 0$. Thus, we only need to consider the case when x and c^* are separated by the tuple (i, σ, θ) at step t, i.e. $f_{x,t}(i, \sigma, \theta) = 1$. By the triangle inequality, we have

$$\|x - \mathcal{T}(x)\|_2^2 \leq (\|x - c^*\|_2 + \|c^* - \mathcal{T}(x)\|_2)^2 \leq$$
$$\leq (\|x - c^*\|_2 + D_t)^2 \leq 2\|x - c^*\|_2 + 2D_t^2.$$

By Definition 4.3 of the approximation factor A_k, we have

$$Z_t(i, \sigma, \theta) = \mathbf{E}\left[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_t,\ (i_t, \sigma_t, \theta_t) = (i, \sigma, \theta)\right] \leq$$
$$\leq A_k\|x - \mathcal{T}_{t+1}(x)\|_2^2 = A_k M_t^2(i, \eta).$$

Combining these two bounds, we get the conclusion. □

Our goal is to show that $A_k \leq O(1/\varepsilon \log k \log\log k)$. We prove Lemma 6.6, which provides the following recurrence relation on A_k: $A_k \leq \max\{4, A_k/k\} + \alpha/\varepsilon \log k \log A_k$. Using this recurrence relation, we get the desired bound on A_k.

Lemma 6.6. *For some absolute constant α, we have*

$$\mathbf{E}\left[\frac{\|x - \mathcal{T}(x)\|_2^2}{\|x - c^*\|_2^2} \mid \mathcal{T}_{t^*}\right] \leq \max\{4, A_k/k\} + \alpha/\varepsilon \log k \log A_k. \quad (10)$$

Proof. Let t^{**} be the stopping time from Corollary 6.4: t^{**} is the first step t when (A) $D_t \le \|x - c^*\|_2$ (note: if c^* is the only center remaining in C_t, then $D_t = 0$); (B) x and c^* are separated before step t (i.e., $c^* \notin C_t$); or (C) $D_t > D_{t-L'}/2$ (where $L' = O(d \ln k)$ as in Corollary 6.4; $t \ge t^* + L'$). Let $\mathcal{E}_A$, $\mathcal{E}_B$, and $\mathcal{E}_C$ be events corresponding to the the stopping rules (A), (B), and (C):

$$\mathcal{E}_A = \left\{D_{t^{**}} \le \|x - c^*\|_2 \ \&\ c^* \in C_{t^{**}}\right\};$$
$$\mathcal{E}_B = \left\{x \ \&\ c^* \text{ are separated at step } t^{**} - 1\right\};$$
$$\mathcal{E}_C = \left\{D_{t^{**}} > D_{t^{**}-L'}/2 \ \&\ t^{**} \ge t^* + L'\right\} \setminus (\mathcal{E}_A \cup \mathcal{E}_B).$$

Note that $\mathcal{E}_A$, $\mathcal{E}_B$, and $\mathcal{E}_C$ are disjoint collectively exhaustive events (one of them must always occur) and by Corollary 6.4, $\Pr(\mathcal{E}_C \mid \mathcal{T}_{t^*}) \le 1/k$. We further partition $\mathcal{E}_B$ into disjoint events

$$\mathcal{E}_{B,t} = \{x \ \&\ c^* \text{ are separated at step } t\}.$$

If event $\mathcal{E}_A$ occurs, then the eventual cost of x is at most $(\|x - c^*\|_2 + D_{t^{**}})^2 \le 4\|x - c^*\|_2^2$ because every center in $C_{t^{**}}$ is at distance at most $\|x - c^*\|_2 + D_{t^{**}}$ from x. If event $\mathcal{E}_{B,t}$ occurs, then the expected cost of x is upper bounded by $Z(i_t, \sigma_t, \theta_t)$. Finally, if event $\mathcal{E}_C$ occurs, then the expected cost of x in $\mathcal{T}$ is upper bounded by $A_k\|x - c^*\|_2^2$ (because c^* is the tentative center for x at step t^{**}). We have

$$\begin{aligned}
&\mathbf{E}\left[\|x - \mathcal{T}(x)\|_2^2 \mid \mathcal{T}_{t^*}\right] \le \\
&\le 4\|x - c^*\|_2^2 \cdot \mathbb{P}(\mathcal{E}_A \mid \mathcal{T}_{t^*}) + A_k\|x - c^*\|_2^2 \cdot \mathbb{P}(\mathcal{E}_C \mid \mathcal{T}_{t^*}) \\
&\quad + \sum_{t=t^*}^{\infty} \mathbf{E}\left[Z_t(i_t, \sigma_t, \theta_t) \mid \mathcal{E}_{B,t},\ \mathcal{T}_{t^*}\right] \Pr(\mathcal{E}_{B,t} \mid \mathcal{T}_{t^*}) \\
&\le \max\{4, A_k/k\} \cdot \|x - c^*\|_2^2 \\
&\quad + \sum_{t=t^*}^{\infty} \mathbf{E}\left[(Z_t(i_t, \sigma_t, \theta_t) - 4\|x - c^*\|_2^2) \cdot \mathbf{1}(\mathcal{E}_{B,t}) \mid \mathcal{T}_{t^*}\right].
\end{aligned}$$

Let $\widetilde{Z}_t(i_t, \sigma_t, \theta_t) = \max\{Z_t(i_t, \sigma_t, \theta_t) - 4\|x - c^*\|_2^2, 0\}$. Then,

$$\begin{aligned}
&\mathbf{E}\left[\frac{\|x - \mathcal{T}(x)\|_2^2}{\|x - c^*\|_2^2} \mid \mathcal{T}_{t^*}\right] \le \\
&\qquad \le \max\{4, A_k/k\} + \sum_{t=t^*}^{\infty} \mathbf{E}\left[\frac{\widetilde{Z}_t(i_t, \sigma_t, \theta_t)}{\|x - c^*\|_2^2} \cdot \mathbf{1}(\mathcal{E}_{B,t}) \mid \mathcal{T}_{t^*}\right].
\end{aligned}$$

Our goal is to upper bound the second term by $\alpha/\varepsilon \log k \log A_k$. Write,

$$\begin{aligned}
&\mathbf{E}\left[\widetilde{Z}_t(i_t, \sigma_t, \theta_t) \cdot \mathbf{1}(\mathcal{E}_{B,t}) \mid \mathcal{T}_{t^*}\right] \\
&= \sum_{i=1}^{d} \mathbf{E}\left[\int_0^1 \frac{\widetilde{Z}_t(i, -1, \theta) + \widetilde{Z}_t(i, 1, \theta)}{2d}\, d\theta \cdot \mathbf{1}\{t < t^{**}\} \mid \mathcal{T}_{t^*}\right]. \quad (11)
\end{aligned}$$

Here, we used that parameters i_t, σ_t, and θ_t are randomly chosen from $\{1, \dots, d\}$, $\{\pm 1\}$, and $[0, 1]$, respectively. We need the following lemma, which we prove in Section 6.3.

Lemma 6.7. *For every i, we have*

$$\int_0^1 \frac{\widetilde{Z}_t(i, -1, \theta) + \widetilde{Z}_t(i, 1, \theta)}{2}\, d\theta \le \le \frac{c_i^* - x_i}{\varepsilon(1 - \varepsilon)} \int_{x_i}^{c_i^*} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2}\, d\eta.$$

Using Lemma 6.7, we can upper bound (11) as follows

$$\begin{aligned}
&\mathbf{E}\left[\widetilde{Z}_t(i_t, \sigma_t, \theta_t) \cdot \mathbf{1}(\mathcal{E}_{B,t}) \mid \mathcal{T}_{t^*}\right] \\
&\le \frac{1}{d}\sum_{i=1}^{d} \frac{c_i^* - x_i}{\varepsilon(1 - \varepsilon)}\, \mathbf{E}\left[\sum_{t=t^*}^{t^{**}-1} \int_{x_i}^{c_i^*} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2}\, d\eta \mid \mathcal{T}_{t^*}\right] \\
&= \frac{1}{d}\sum_{i=1}^{d} \frac{c_i^* - x_i}{\varepsilon(1 - \varepsilon)} \int_{x_i}^{c_i^*} \mathbf{E}\left[\sum_{t=t^*}^{t^{**}-1} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2} \mid \mathcal{T}_{t^*}\right] d\eta \\
&\le \sum_{i=1}^{d} \frac{2(c_i^* - x_i)^2}{d\varepsilon} \max_{\eta \in [x_i, c_i^*]} \mathbf{E}\left[\sum_{t=t^*}^{t^{**}-1} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2} \mid \mathcal{T}_{t^*}\right].
\end{aligned}$$

We now show that for every $\eta \in [x_i, c_i^*]$ the following bound holds with probability 1:

$$\sum_{t=t^*}^{t^{**}-1} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2} \le O(d \log k \log A_k). \quad (12)$$

This will conclude the proof of Lemma 6.6 because (12) implies that

$$\begin{aligned}
&\mathbf{E}\left[\widetilde{Z}_t(i_t, \sigma_t, \theta_t) \cdot \mathbf{1}(\mathcal{E}_{B,t}) \mid \mathcal{T}_{t^*}\right] \\
&\le \frac{1}{d}\sum_{i=1}^{d} \frac{2(c_i^* - x_i)^2}{\varepsilon} \cdot O(d \log k \log A_k) \\
&= \frac{2\|c^* - x\|_2^2}{\varepsilon} \cdot O(\log k \log A_k).
\end{aligned}$$

□

Lemma 6.8. *Inequality (12) holds with probability* 1.

Proof. By Lemma 6.1, $R_t \ge D_t/2$. Thus,

$$\begin{aligned}
&\sum_{t=t^*}^{t^{**}-1} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2} \le \\
&\le 8\sum_{t=t^*}^{t^{**}-1} \frac{\min\{D_t^2, A_k M_t^2(i, \eta)\}}{D_t^2} = 8\sum_{t=t^*}^{t^{**}-1} \min\left\{1, \frac{A_k\, M_t^2(i, \eta)}{D_t^2}\right\}.
\end{aligned}$$

Let

$$f_t(i, \eta) = \frac{A_k\, M_t^2(i, \eta)}{D_t^2}.$$

Observe that $M_t(i, \eta)$ is a non-decreasing sequence and D_t is a non-increasing sequence for fixed i, η and $t \in \{t^*, \dots, t^{**} - 1\}$. Moreover, by the definition of stopping time t^{**}, $D_t \le D_{t-L'}/2$ for $t \in \{t^* + L', \dots, t^{**} - 1\}$, where $L' = O(d \log k)$ (see stopping rule (C)). Hence, $f_t(i, \eta)$ is a non-decreasing sequence, and $f_t(i, \eta) \ge 4 f_{t-L'}(i, \eta)$ for $t \in \{t^* + L', \cdots, t^{**} - 1\}$. Let t' be the first step t in $[t^*, t^{**} - 1]$ when $f_{t'}(i, \eta) \ge 1$. If $f_t(i, \eta) < 1$ for all $t \in \{t^*, \cdots, t^{**} - 1\}$, then $t' = t^{**}$. We have

$$\begin{aligned}
&\frac{1}{8}\sum_{t=t^*}^{t^{**}-1} \frac{\min\{2D_t^2, A_k M_t^2(i, \eta)\}}{R_t^2} \le \\
&\qquad \le \sum_{t=t^*}^{t^{**}-1} \min\{1, f_t(i, \eta)\} = \underbrace{\sum_{t=t^*}^{t'-1} f_t(i, \eta)}_{\Sigma_I} + \underbrace{\sum_{t=t'}^{t^{**}-1} 1}_{\Sigma_{II}}.
\end{aligned}$$

The first sum (Σ_I) on the right hand side is upper bounded by $2L' \cdot f_{t'}(i,\eta)$, because $f_t(i,\eta) \geq 4 f_{t-L'}(i,\eta)$ for $t < t^{**}$. In turn, $2L' \cdot f_{t'}(i,\eta) \leq 2L' = O(d \log k)$, because $f_t(i,\eta) \leq 1$ for $t < t'$. The second sum (Σ_{II}) equals $t^{**} - t'$. Since $f_t(i,\eta) \geq 4 f_{t-L'}(i,\eta)$ for every $t \in [t^* + L', t^{**} - 1]$, we have

$$\left\lfloor \frac{(t^{**}-1)-t'}{L'} \right\rfloor \leq \log_4 \frac{f_{t^{**}-1}(i,\eta)}{f_{t'}(i,\eta)} \leq$$
$$\leq \log_4 f_{t^{**}-1}(i,\eta) = \log_4 \left(\frac{A_k M^2_{t^{**}-1}(i,\eta)}{D^2_{t^{**}-1}} \right).$$

It remains to show that $M_{t^{**}-1}(i,\eta) = O(D_{t^{**}-1})$ and thus

$$t^{**} - t' = O(L' \log A_k) = O(d \log k \log A_k).$$

We have, $M_{t^{**}-1}(i,\eta) \leq \|x - c^*\|_2 + D_t \leq 2D_t$, where we used that for every $t < t^{**}$, $D_t > \|x - c^*\|_2$ (see stopping rule (C)). This finishes the proof of Lemma 6.8. □

6.3 Proof of Lemma 6.7

We first make the following simple but crucial observation.

Claim 6.9. *If $\widetilde{Z}_t(i,\sigma,\theta) > 0$, then for $\eta = m_i^t + (\sigma+\varepsilon)\sqrt{\theta} R_t$, we have*

$$|\eta - m_i^t| \equiv |(\sigma+\varepsilon)\sqrt{\theta} R_t| \leq \frac{c_i^* - x_i}{\varepsilon}.$$

Proof of Claim 6.9. If $\widetilde{Z}_t(i,\sigma,\theta) > 0$, then the cut with parameters i, σ, θ separates x and c^* (otherwise, $Z_t(i,\sigma,\theta)$ and $\widetilde{Z}_t(i,\sigma,\theta)$ would be equal to 0). That is, $x_i \leq m_i^t + \sigma\sqrt{\theta} R_t$ and $c_i^* > m_i^t + (\sigma + \varepsilon)\sqrt{\theta} R_t$. Write,

$$c_i^* - x_i = (c_i^* - m_i^t) - (x_i - m_i^t) > (\sigma+\varepsilon)\sqrt{\theta} R_t - \sigma\sqrt{\theta} R_t = \varepsilon\sqrt{\theta} R_t.$$

Hence,

$$|(\sigma+\varepsilon)\sqrt{\theta} R_t| = \frac{|\sigma+\varepsilon|}{\varepsilon} \cdot \varepsilon\sqrt{\theta} R_t < \frac{|\sigma+\varepsilon|}{\varepsilon}(c_i^* - x_i).$$

□

Proof of Lemma 6.7. We have

$$\int_0^1 \frac{\widetilde{Z}_t(i,-1,\theta) + \widetilde{Z}_t(i,1,\theta)}{2}\, d\theta = \frac{1}{2} \sum_{\sigma \in \{\pm 1\}} \int_0^1 \widetilde{Z}_t(i,\sigma,\theta)\, d\theta.$$

Make the substitutions $\eta_\sigma = m_i^t + (\sigma+\varepsilon) R_t \sqrt{\theta}$. Then, we have $d\theta = \frac{2(\eta_\sigma - m_i^t)}{(\sigma+\varepsilon)^2 R_t^2}\, d\eta_\sigma$ and

$$\int_0^1 \frac{\widetilde{Z}_t(i,-1,\theta) + \widetilde{Z}_t(i,1,\theta)}{2}\, d\theta =$$
$$= \sum_{\sigma \in \{\pm 1\}} \int_{m_i^t}^{m_i^t + (\sigma+\varepsilon)R_t} \frac{\widetilde{Z}_t(i,\sigma,\theta)}{(\sigma+\varepsilon)^2 R_t^2} \cdot (\eta_\sigma - m_i^t)\, d\eta_\sigma.$$

By Claim 6.7, $|\eta_\sigma - m_i^t| \leq |\sigma+\varepsilon|/\varepsilon \cdot (c_i^* - x_i)$. Since $Z(i,\sigma,\theta) \geq 0$, we have $\widetilde{Z}(i,\sigma,\theta) = \max\{Z_t(i_t,\sigma_t,\theta_t) - 4\|x - c^*\|_2^2, 0\} \leq Z(i,\sigma,\theta)$. As we discuss in the previous section, $\widetilde{Z}(i,\sigma,\theta) \leq Z(i,\sigma,\theta) \leq \min\{2D_t^2, A_k M_t^2(i,\eta_\sigma)\}$ (see Claim 6.5). Also, if $\eta_\sigma \notin [x_i, c_i^*]$, then x and c^* are not separated by the tuple (i,σ,θ), which implies $\widetilde{Z}(i,\sigma,\theta) = 0$. Thus,

$$\int_0^1 \frac{\widetilde{Z}_t(i,-1,\theta) + \widetilde{Z}_t(i,1,\theta)}{2}\, d\theta \leq$$
$$\leq \frac{c_i^* - x_i}{\varepsilon(1-\varepsilon)} \int_{x_i}^{c_i^*} \frac{\min\{2D_t^2, A_k M_t^2(i,\eta)\}}{R_t^2}\, d\eta.$$

This concludes the proof of Lemma 6.7. □

7 LOWER BOUND ON THE BI-CRITERIA APPROXIMATION

In this section, we prove Theorem 1.2. We show a lower bound on the price of explainability for k-means in the bi-criteria setting. Our proof follows the general approach by Makarychev and Shan [28].

Theorem 1.2. *For every $k > 500$ and $\ln^3 k/\sqrt{k} < \delta < 1/100$, there exists an instance X with k clusters such that the k-means cost for every threshold tree $\mathcal{T}$ with $(1+\delta)k$ leaves is at least*

$$\mathrm{cost}(X,\mathcal{T}) \geq \Omega\left(\frac{\log^2 k}{\delta}\right) \mathrm{OPT}_k(X).$$

Proof of Theorem 1.2. We construct a hard instance for explainable clustering as follows. Let $d = 300\lceil \ln k \rceil$. Consider the grid $\{0, \varepsilon, 2\varepsilon, \dots, 1\}^d$ with step size $\varepsilon = 50\delta/\lceil \ln k \rceil$ in the d-dimensional unit cube $[0,1]^d$. We uniformly sample k centers $C = \{c^1, c^2, \dots, c^k\}$ from the nodes of the grid. Then, we create a data set X. For every center c^i in C, data set X contains many (namely, $k^2\lceil \ln^3 k \rceil$) points co-located with c^i and two special points $c^i \pm (\varepsilon, \varepsilon, \dots, \varepsilon)$. Hence, the total number of points in X is $k^3\lceil \ln^3 k \rceil + 2k$. Note that all centers and all points in X lie in the nodes of the grid.

The cost of the k-means clustering with centers $C = \{c^1, \dots, c^k\}$ equals $2kd\varepsilon^2$, since the distance from the special points $c^i \pm (\varepsilon, \dots, \varepsilon)$ to c^i is $\varepsilon\sqrt{d}$. Hence, the cost of the optimal k-means clustering is at most $2kd\varepsilon^2$. We now show that there exists an instance such that the cost of every *explainable* k-means clustering with $(1+\delta)k$ centers is at least $2kd\varepsilon^2 \cdot \Omega(1/\delta \log^2 k)$. In this instance, every explainable k-means clustering with $(1+\delta)k$ centers separates at least $\delta k = \Omega(\varepsilon k \ln k)$ special points $c^i \pm (\varepsilon, \varepsilon, \dots, \varepsilon)$ from c^i. The cost of each special point separated from its original center is at least $\Omega(d)$. Thus, the total cost of every explainable k-means clustering is at least $\Omega(d\varepsilon k \ln k) = 2kd\varepsilon^2 \cdot \Omega(1/\delta \log^2 k)$. First, with high probability every two centers in C are far apart. We use the following lemma from Makarychev and Shan [28].

Lemma 7.1 (Makarychev and Shan [28]). *With probability at least $1 - 1/k^2$ the following statement holds: The distance between every two distinct centers c' and c'' in C is at least $\sqrt{d}/5$.*

All data points in X are in the grid $\{-\varepsilon, 0, \varepsilon, 2\varepsilon, \dots, 1, 1+\varepsilon\}^d$. Every internal node u in the threshold tree should contain a threshold cut that separates at least two data points in that node u. Otherwise, we can ignore this threshold cut since one side of this cut contains no data points. If two threshold cuts have the same coordinate and thresholds within the same grid interval $(j\varepsilon, j\varepsilon + \varepsilon)$, then these two threshold cuts create the same partition of data points contained in the internal node. Since there are at most $1/\varepsilon + 2$ different grid intervals for each coordinate, the number of distinct threshold cuts

for each internal node is at most $d(1/\varepsilon+2) \le 2d/\varepsilon$. Every node in the threshold tree corresponds to a cell in $\mathbb{R}^d$. This cell is determined by the threshold cuts on the path from the root to that node. Let π be an ordered set of tuples (i_j, ξ_j, λ_j), where (i_j, ξ_j) is the j-th threshold cut on the path from the root to the node, and $\lambda_j \in \{\pm 1\}$ specifies one of the sides of the cut. Then, every ordered set π corresponds to a path in the threshold tree starting in the root.

Let $u(\pi)$ be the intersection of the cuts in π. We say that a center c^i in $u(\pi)$ is damaged if one of the special points $c^i \pm (\varepsilon, \dots, \varepsilon)$ is separated from c^i by one of the threshold cuts in π. In other words, c^i is damaged if $c^i \in u(\pi)$, but $c^i - (\varepsilon, \dots, \varepsilon) \notin u(\pi)$ or $c^i + (\varepsilon, \dots, \varepsilon) \notin u(\pi)$. Otherwise, we say that c^i is not damaged. Similarly, we say that a node of the grid $x \in u(\pi)$ is not damaged if $x \pm (\varepsilon, \dots, \varepsilon) \in u(\pi)$. Let $F_{u(\pi)}$ be the set of all centers that are not damaged in node $u(\pi)$. We show that with high probability, if a node $u(\pi)$ contains more than $\sqrt{k}$ centers, every threshold cut that splits node $u(\pi)$ damages at least $\varepsilon|F_{u(\pi)}|/2$ centers in $F_{u(\pi)}$.

Lemma 7.2. *With probability at least $1 - 1/k$, the following holds: For every path (ordered set of cuts) π of length at most $\log_2 k/4$, we have (a) $|F_{u(\pi)}| \le \sqrt{k}$; or (b) every threshold cut that separates at least two data points in $u(\pi)$ damages at least $\varepsilon|F_{u(\pi)}|/2$ centers in $F_{u(\pi)}$.*

Proof. Consider a fixed ordered set of cuts π of size at most $\log_2 k/4$. We upper bound the probability that both events (a) and (b) do not occur for this fixed path π on the random instance X. If $|F_{u(\pi)}| \le \sqrt{k}$, then the event (a) happens. So, we assume that $F_{u(\pi)}$ contains more than $\sqrt{k}$ centers. We then bound the probability that event (b) happens conditioned on the size of $F_{u(\pi)}$. Observe that all centers in $F_{u(\pi)}$ are distributed uniformly and independently among the grid nodes in $u(\pi)$ that are not damaged by the cuts in π conditioned on $|F_{u(\pi)}|$. Pick an arbitrary threshold cut (i, ξ) in $u(\pi)$ that separates at least two nodes of the grid in $u(\pi)$. For every center c in $F_{u(\pi)}$, the probability that the threshold cut (i, ξ) damages this center c is at least ε. Let X_j be the indicator random variable that the j-th center in $F_{u(\pi)}$ is damaged by (i, ξ). The expected number of centers in $F_{u(\pi)}$ damaged by cut (i, ξ) conditioned on $|F_{u(\pi)}| = l$ equals

$$\mathbf{E}\left[\sum_{j=1}^{l} X_j \;\middle|\; |F_{u(\pi)}| = l\right] \ge \varepsilon l.$$

Let $\mu = \mathbf{E}[\sum_j X_j \mid |F_{u(\pi)}| = l]$. By the Chernoff bound for Bernoulli random variables, we have

$$\mathbb{P}\left\{\sum_{j=1}^{l} X_j \le \varepsilon |F_{u(\pi)}|/2 \;\middle|\; |F_{u(\pi)}| = l\right\} \le$$

$$\le \mathbb{P}\left\{\sum_{j=1}^{l} X_j \le \mu/2 \;\middle|\; |F_{u(\pi)}| = l\right\} \le e^{-\mu/8} \le e^{-\varepsilon\sqrt{k}/8}.$$

Combining all conditional probabilities for $|F_{u(\pi)}| > \sqrt{k}$, the probability that the event (b) doesn't happen is at most $e^{-\varepsilon\sqrt{k}/8}$. Since all data points are in the grid $\{-\varepsilon, 0, \varepsilon, 2\varepsilon, \dots, 1, 1+\varepsilon\}^d$, there are at most $2d/\varepsilon$ different threshold cuts that separates at least two data points in node $u(\pi)$. By the union bound, the probability that both events (a) and (b) do not happen is at most $e^{-\varepsilon\sqrt{k}/8} \cdot 2d/\varepsilon \le e^{-2\ln^2 k}$. Since there are at most $4d/\varepsilon$ different choices for each tuple (i_j, ξ_j, λ_j) in π, the number of paths with length less than $m = \log_2 k/4$ is at most $m(4d/\varepsilon)^m \le e^{\ln^2 k}$. Thus, by the union bound over all paths with length less than $\log_2 k/4$, we get that (a) or (b) holds with probability at least

$$1 - m(4d/\varepsilon)^m \cdot e^{-\varepsilon\sqrt{k}/8} \cdot 2d/\varepsilon \ge 1 - e^{\ln^2 k} \cdot e^{-2\ln^2 k} \ge 1 - \frac{1}{k}.$$

since $d/\varepsilon \le 15000\sqrt{k}\ln^3 k$ for $d = 300\lceil \ln k \rceil$ and $\varepsilon = 50\delta/\lceil \ln k \rceil \ge 50\sqrt{k}\ln^2 k$. □

By Lemma 7.1 and Lemma 7.2, we can find an instance X such that the following conditions hold:

- The distance between every two distinct centers c' and c'' in C is at least $\sqrt{d}/5$.
- For every path (ordered set of cuts) π of length at most $\log_2 k/4$, we have (a) $|F_{u(\pi)}| \le \sqrt{k}$; or (b) every threshold cut that separates at least two data points in $u(\pi)$ damages at least $\varepsilon|F_{u(\pi)}|/2$ centers in $F_{u(\pi)}$.

We first show that the threshold tree must separate all centers. Suppose there is a leaf contains more than one center. Since the distance between every two centers is at least $\sqrt{d}/5$, there exists at least one center in this leaf with distance greater than $\sqrt{d}/10$ to the optimal center of this leaf. Since we add $k^2\lceil \ln^3 k \rceil$ points co-located with each center, the cost for the leaf that contains more than one center is greater than $k^2\lceil \ln^3 k \rceil \cdot d/100 = 2kd\varepsilon^2 \cdot \Omega(1/\delta \log^2 k)$. Thus, the lower bound holds for any threshold tree that does not separate all centers. To separate all centers, the depth of the threshold tree must be at least $\lceil \log_2 k \rceil$. We show the following lower bound on the number of damaged centers for every threshold tree that separates all centers.

Lemma 7.3. *Consider any instance X with k centers satisfies two conditions in Lemma 7.1 and Lemma 7.2. For every threshold tree that separates all centers in C, there are at least $2\delta k$ damaged centers.*

Proof. Consider any threshold tree $\mathcal{T}$ that separates all centers. We consider the following two cases. If the number of damaged centers at level $\lfloor \log_2 k \rfloor/4$ of threshold tree $\mathcal{T}$ is more than $k/2$, then the total number of damaged centers generated by this threshold tree is more than $2\delta k$.

If the number of damaged centers at level $\lfloor \log_2 k \rfloor/4$ of threshold tree $\mathcal{T}$ is less than $k/2$, then the number of centers that are not damaged at each level $i = 1, 2, \dots, \lfloor \log_2 k \rfloor/4$ is at least $k/2$. We call a node u a small node if it contains at most $\sqrt{k}$ centers which are not damaged, otherwise we call it a large node. We now lower bound the number of centers damaged at a fixed level $i \in \{1, 2, \cdots, \lfloor \log_2 k \rfloor/4\}$. For every level $i \in \{1, 2, \cdots, \lfloor \log_2 k \rfloor/4\}$, the number of nodes at level i is at most $k^{1/4}$. Since each small node contains at most $\sqrt{k}$ centers that are not damaged, the total number of centers that are not damaged in small nodes at level i is at most $k^{3/4}$. Since the total number of centers that are not damaged at level i is at least $k/2$, the number of centers that are not damaged in large nodes at level i is at least $k/4$. By Lemma 7.2, the number of damaged centers generated at level i is at least $\varepsilon k/8$. Therefore, the total number of damaged centers generated by this threshold

tree $\mathcal{T}$ is at least

$$\frac{\lfloor \log_2 k \rfloor}{4} \cdot \frac{\varepsilon k}{8} \geq \frac{50 \lfloor \log_2 k \rfloor \delta k}{32 \ln k} \geq 2\delta k,$$

which completes the proof. □

We now lower bound the cost for every threshold tree with $(1+\delta)k$ leaves that separates all centers. Consider any threshold tree $\mathcal{T}$ with $(1+\delta)k$ leaves that separates all centers in C. By Lemma 7.3, we have more than $2\delta k$ data points separated from their original centers by $\mathcal{T}$. For each point x separated from its original center c, one and only one of the following may occur: (1) the data point x is assigned to a leaf containing a center $c' \neq c$; (2) the data point x is assigned to a leaf containing no center. Among these $2\delta k$ data points, we show that there are at least δk data points that have distances to their new centers greater than $\sqrt{d}/20$.

For each leaf containing a center c', the optimal center for this leaf is shifted from c' by at most $\varepsilon\sqrt{d}$. Otherwise, the cost of this leaf is at least $k^2\lceil \ln^3 k \rceil \cdot \varepsilon^2 d = 2kd\varepsilon^2 \cdot \Omega(1/\delta \log^2 k)$ since there are $k^2\lceil \ln^3 k \rceil$ data points co-located at each center. Suppose a point x separated from its original center c is assigned to a leaf containing a center $c' \neq c$. By Lemma 7.1 and the triangle inequality, the distance from the point x to the optimal center for this leaf is at least $\sqrt{d}/10$.

For each leaf containing no center, it may contain several points from distinct clusters. Among these points, there is at most one point within $\sqrt{d}/20$ distance of the optimal center for this leaf. Suppose two points x' and x'' from distinct clusters are within $\sqrt{d}/20$ distance of the optimal center for this leaf. Then, the distance between x' and x'' is at most $\sqrt{d}/10$. Let c' and c'' be the original centers for points x' and x'' respectively. The distance between c' and c'' is at most $\sqrt{d}/10 + 2\varepsilon\sqrt{d} \leq \sqrt{d}/5$, which contradicts the distance between every two centers is at least $\sqrt{d}/5$.

Since the threshold tree $\mathcal{T}$ has $(1+\delta)k$ leaves, there are δk leaves that do not contain a center. Thus, among points separated from their original centers, there are at most δk points with distance less than $\sqrt{d}/20$ to their new centers. Since there are more than $2\delta k$ points separated from their original centers, we have at least δk points with cost greater than $d/400$. Therefore, the cost given by this threshold tree $\mathcal{T}$ is at least

$$\mathrm{cost}(X, \mathcal{T}) \geq \delta k \cdot \frac{d}{400} = \Omega(\delta d k).$$

Recall that the optimal k-means cost for this instance is at most $2k\varepsilon^2 d$ and $\varepsilon = 50\delta/\lceil \ln k \rceil$. Thus, the cost given of this explainable clustering is at least

$$\mathrm{cost}(X, \mathcal{T}) = \Omega(\delta d k) \geq \Omega\left(\frac{\log^2 k}{\delta}\right) \mathrm{OPT}_k(X).$$

□

REFERENCES

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

[2] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. 2019. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *SIAM J. Comput.* 49, 4 (2019), FOCS17–97.

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

[4] David Arthur and Sergei Vassilvitskii. 2007. k-means++ the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 1027–1035.

[5] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. 2015. The hardness of approximation of euclidean k-means. *arXiv preprint arXiv:1502.03316* (2015).

[6] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. 2019. Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 1039–1050.

[7] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. 2018. Interpretable clustering via optimal trees. *arXiv preprint arXiv:1812.00539* (2018).

[8] Christos Boutsidis, Michael W Mahoney, and Petros Drineas. 2009. An improved approximation algorithm for the column subset selection problem. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 968–977.

[9] Christos Boutsidis, Anastasios Zouzias, Michael W Mahoney, and Petros Drineas. 2014. Randomized dimensionality reduction for k-means clustering. *IEEE Transactions on Information Theory* 61, 2 (2014), 1045–1062.

[10] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 2017. *Classification and regression trees*. Routledge.

[11] Moses Charikar and Lunjia Hu. 2022. Near-Optimal Explainable k-Means for All Dimensions. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

[12] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. 2015. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 163–172.

[13] Sanjoy Dasgupta. 2008. *The hardness of k-means clustering*. Department of Computer Science and Engineering, University of California, San Diego.

[14] Sanjoy Dasgupta, Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. 2020. Explainable k-Means and k-Medians Clustering. In *International Conference on Machine Learning*. PMLR, 7055–7065.

[15] Ron Elber. 2004. KDD-Cup. http://osmot.cs.cornell.edu/kddcup/

[16] Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. 2022. Almost Tight Approximation Algorithms for Explainable Clustering. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

[17] Ricardo Fraiman, Badih Ghattas, and Marcela Svarc. 2013. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification* 7, 2 (2013), 125–145.

[18] Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. 2020. ExKMC: Expanding Explainable k-Means Clustering. *arXiv preprint arXiv:2006.02399* (2020).

[19] Buddhima Gamlath, Xinrui Jia, Adam Polak, and Ola Svensson. 2021. Nearly-Tight and Oblivious Algorithms for Explainable Clustering. In *Advances in Neural Information Processing Systems*.

[20] Fabrizio Grandoni, Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Rakesh Venkat. 2022. A refined approximation for Euclidean k-means. *Inform. Process. Lett.* 176 (2022), 106251.

[21] Eduardo Laber and Lucas Murtinho. 2021. On the price of explainability for some clustering problems. In *International Conference on Machine Learning*. PMLR.

[22] Euiwoong Lee, Melanie Schmidt, and John Wright. 2017. Improved and simplified inapproximability for k-means. *Inform. Process. Lett.* 120 (2017), 40–43.

[23] Bing Liu, Yiyuan Xia, and Philip S Yu. 2005. Clustering via decision tree construction. In *Foundations and advances in data mining*. Springer, 97–124.

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

[25] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. 2019. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 1027–1038.

[26] Konstantin Makarychev, Yury Makarychev, Maxim Sviridenko, and Justin Ward. 2016. A Bi-Criteria Approximation Algorithm for k-Means. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (2016).

[27] Konstantin Makarychev, Aravind Reddy, and Liren Shan. 2020. Improved Guarantees for k-means++ and k-means++ Parallel. *Advances in Neural Information Processing Systems* 33 (2020).

[28] Konstantin Makarychev and Liren Shan. 2021. Near-optimal Algorithms for Explainable k-Medians and k-Means. In *International Conference on Machine Learning*. PMLR, 7358–7367.

[29] J. Ross Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106.

[30] J. Ross Quinlan. 1993. C4. 5, Programs for Machine Learning. In *International Conference on Machine Learning*. 252–259.

[31] Sandhya Saisubramanian, Sainyam Galhotra, and Shlomo Zilberstein. 2020. Balancing the tradeoff between clustering value and interpretability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 351–357.

[32] Dennis Wei. 2016. A constant-factor bi-criteria approximation guarantee for k-means++. *Advances in Neural Information Processing Systems* 29 (2016), 604–612.