Near-Optimal Explainable k-Means for All Dimensions

Moses Charikar* Lunjia Hu†

Abstract

Many clustering algorithms are guided by certain cost functions such as the widely-used k-means cost. These algorithms divide data points into clusters with often complicated boundaries, creating difficulties in explaining the clustering decision. In a recent work, Dasgupta, Frost, Moshkovitz, and Rashtchian (ICML'20) introduced explainable clustering, where the cluster boundaries are axis-parallel hyperplanes and the clustering is obtained by applying a decision tree to the data. The central question here is: how much does the explainability constraint increase the value of the cost function?

Given d-dimensional data points, we show an efficient algorithm that finds an explainable clustering whose k-means cost is at most $k^{1-2/d}\,\mathrm{poly}(d\log k)$ times the minimum cost achievable by a clustering without the explainability constraint, assuming $k, d \geq 2$. Combining this with an independent work by Makarychev and Shan (ICML'21), we get an improved bound of $k^{1-2/d}\,\mathrm{polylog}(k)$, which we show is optimal for every choice of $k, d \geq 2$ up to a poly-logarithmic factor in k. For $d = 2$ in particular, we show an $O(\log k \log\log k)$ bound, improving exponentially over the previous best bound of $\widetilde{O}(k)$.

1 Introduction

As a result of the rapid growth of data analysis and machine learning technologies, many important decisions that impact our lives are made based on rules learned from data by algorithms, rather than rules designed explicitly by people. Clustering algorithms play an important role here when data is not supervised by labels, and it is important to make sure that the clusterings they produce can be understood easily by people. Many clustering algorithms make their decisions by trying to optimize a cost function such as the widely-used k-means cost. While this simple decision rule is helpful when designing and analyzing clustering algorithms, it raises interpretability issues when applied in practice [SGZ20, BOW21].

We study a notion of explainable clustering introduced by Dasgupta, Frost, Moshkovitz, and Rashtchian [MDRF20], where a clustering is considered explainable if a decision tree with k leaves can explain the clustering result for all data points. Specifically, every non-leaf node of the decision tree corresponds to an axis-parallel hyperplane that divides the current set of data points into two subsets, which are passed to the two children of the node respectively. Every leaf of the decision tree thus corresponds to all the data points contained in a (possibly unbounded) rectangular box with axis-parallel faces, and these data points are required to be placed in the same cluster.

To understand the increase in the clustering cost caused by the explainability constraint, [MDRF20] studied the *competitive ratio* of an explainable clustering, which is defined to be the

*Computer Science Department, Stanford University. Email: moses@cs.stanford.edu. Supported by a Simons Investigator Award.

†Computer Science Department, Stanford University. Email: lunjia@stanford.edu. Supported by NSF Award IIS-1908774 and a VMware fellowship.

ratio between its cost and the minimum cost achievable by a clustering using k clusters without the explainability constraint. For the k-medians and k-means cost, [MDRF20] showed efficient algorithms computing explainable clusterings with competitive ratios being at most $O(k)$ and $O(k^2)$ respectively. (In their work, distances are measured using the L_1-norm for k-medians, whereas the L_2-norm is used for k-means.)

In a recent work independent of ours, Makarychev and Shan [MS21] significantly improved the competitive ratio bound of [MDRF20]. For k-means, they showed an efficient algorithm computing an explainable clustering with competitive ratio at most $k\,\mathrm{polylog}(k)$. They also showed that the bound is near-optimal when the dimension d is *worst-case* for the given k: there exists a *worst-case* dimension d depending on k for which it is impossible to achieve a competitive ratio better than $k/\mathrm{polylog}(k)$. For k-medians, they achieved a competitive ratio bound of $\widetilde{O}(\log k)$ for the L_1 norm and $O((\log k)^{3/2})$ for the L_2-norm.

The work [MS21] leaves open the question of achieving better competitive ratios when the dimension is not worst-case. Before our work, Laber and Murtinho [LM21] gave algorithms with better competitive ratios in lower dimensions, but their bounds for k-medians and k-means were sub-optimal and (almost) subsumed by [MS21]. In this work, we show that significantly better competitive ratios than that in [MS21] can be achieved for k-means when the dimension d is not worst-case. We give a competitive ratio bound that depends on both d and k, which we show is near-optimal for all choices of $d, k \geq 2$.

Our results. Our main result is an efficient algorithm that takes a set of d dimensional points and computes an explainable clustering with competitive ratio at most $k^{1-2/d}\mathrm{poly}(d \log k)$ for the k-means cost (Theorem 11). Similarly to previous work [MDRF20, LM21], we design a post-processing algorithm that takes an arbitrary clustering $\mathcal{C}$ of the input points with k clusters, and computes an explainable clustering with cost at most $k^{1-2/d}\mathrm{poly}(d \log k)$ times the cost of $\mathcal{C}$ using the same cluster centroids in $\mathcal{C}$. To get the desired competitive ratio, one only needs to run the post-processing algorithm on a clustering $\mathcal{C}$ computed by a constant-factor approximation algorithm for k-means [KMN+04, ANFSW17].

Our bound has a better dependence on k for *every* fixed dimension d than [MS21], but our dependence on d is not as good. However, the dependence on d can be improved by combining our algorithm with [MS21]. Specifically, if we run the algorithm in [MS21] instead when $d = \Omega(\log k)$, we get an improved bound of $k^{1-2/d}\,\mathrm{polylog}(k)$ for all $k, d \geq 2$ (Corollary 12). We show this is nearly optimal for all $k, d \geq 2$ by constructing a set of d dimensional points for which the competitive ratio is at least $k^{1-2/d}/\mathrm{polylog}(k)$ for the k-means cost (Theorem 24).

In the special case of $d = 2$, we show an efficient algorithm computing explainable clusterings with competitive ratio $O(\log k \log\log k)$ (Theorem 1). This is an exponential improvement compared to the previous best bound of $\widetilde{O}(k)$ [LM21, MS21].

Technical overview. We build the decision tree recursively starting from the root. That is, the first step of the algorithm is to find the hyperplane corresponding to the root of the decision tree, and then solve the two induced subproblems recursively. In a similar spirit to [LM21], we make sure that the direct excess cost of every hyperplane is small, and that the two subproblems have similar "sizes" in order to minimize the depth of the decision tree. When the two goals are in conflict, it is important to make a balanced tradeoff: [LM21] applies binary search with non-uniform probing cost due to [CFG+02], whereas we take a more flexible approach originated from an argument of Seymour [Sey95].

In previous analysis [MDRF20, LM21], whenever a point is separated from its current centroid,

the diameter of the subproblem is used to upper bound the distance from the point to its new centroid. While this can be a good estimate for the worst-case dimension, we need a more careful bound to get a near optimal competitive ratio for non-worst-case dimensions. To this end, we compute a collection of *forbidden regions* when selecting every hyperplane to prevent separating a point from its current centroid if it would suffer a cost too large compared to its current cost. We use a volume argument to bound the size of the forbidden region, so that we have enough non-forbidden space to apply Seymour's argument.

The diameter upper bound used in previous works allows them to completely ignore a point once it is separated, because the distance cannot grow larger than the current diameter even if the point is separated further. In our analysis, however, we need to deal with situations where a point is separated multiple times. We keep track of the points that are separated from their assigned centroids, and we show that "essentially" no point can be separated too many times. Roughly speaking, for $d = 2$, we show that a point can be separated at most twice before either a) the point has a large distance to its current centroid so that we can afford to use the diameter upper bound, or b) both the point and its current centroid are close to a corner of the current rectangle so that we can avoid separating them. For the more general case $d > 2$, it is possible that none of the above cases happen for some "bad" points, but we show that the "bad" points are assigned to only a small number of centroids, in which case we can also avoid separating them from their assigned centroids.

Related work. Decision trees are a classic method for classifying labeled data [HMS66]. Due to its intrinsic interpretability, people also applied decision trees and related algorithms to clustering unlabeled data: [DRB97, CJ02, BK05, LXY05, YM10, FGS13, CCH+16, GMB17, BOW18, BOW21]. We use the framework of [MDRF20], who gave the first competitive ratio analysis for explainable clustering using decision trees. [FMR20] relaxed the framework of [MDRF20] by allowing more leaves in the decision tree than the number of clusters, so that a cluster can correspond to multiple leaves.

In clustering tasks, the dimension of the input points plays an important role. Many influential results were obtained while studying clustering in different dimensions. While it is NP-hard in general to approximately optimize the k-medians and the k-means cost within a small constant factor [JMS02, ACKS15], polynomial-time approximation schemes (PTAS) have been found for both k-medians and k-means in fixed dimensions [ARR99, KR99, FRS16, CAKM16]. Randomly projecting any k clusters to $O(\log k)$ dimensions approximately preserves the the k-medians and the k-means cost with high probability [BBCA+19, MMR19].

Interpretability and explainability are important aspects of making machine learning reliable, and they have received growing research attention. Compared to clustering and unsupervised learning in general, more work on interpretability considered supervised learning [RSG16, LL17, AB18, RSG18, Lip18, Rud19, MSK+19, AMDIVW19, DF19, Mol20, SF20, GL20]. Besides decision trees, neural nets have been used to improve clustering explainability [KEM+19], whereas an interpretability score was formulated by [SGZ20], who studied the tradeoff between the interpretability score and the clustering cost. Fairness is another important consideration towards making clustering more trustworthy. There is a large body of recent work on fair clustering [BIO+19, BCFN19, HJV19, KAM19, SSS19, MV20, JKL20, DC21].

Paper organization. We formally define explainable clusterings and various notations in Section 2. We prove our $O(\log k \log\log k)$ competitive ratio upper bound for $d = 2$ in Section 3, and our $k^{1-2/d}\,\mathrm{poly}(d \log k)$ upper bound for $d > 2$ in Section 4. The lower bound $k^{1-2/d}/\mathrm{polylog}(k)$ is

shown in Section 5. Some helper claims and lemmas used in our analysis are stated and proved in Appendix A.

2 Preliminaries

We use $x(j)$ to denote the j-th coordinate of a point x in the d-dimensional space $\mathbb{R}^d$, where j is chosen from $[d]$, namely, $\{1, \ldots, d\}$. We use $\|x\|_2 = (\sum_{j=1}^{d} x(j)^2)^{1/2}$ and $\|x\|_\infty = \max_{j\in[d]} |x(j)|$ to denote the L_2-norm and the L_∞ norm of a point $x \in \mathbb{R}^d$, respectively.

Every pair $(j, \theta) \in [d] \times \mathbb{R}$ defines an axis-parallel hyperplane that partitions $\mathbb{R}^d$ into two subsets: $B_{\le}(j, \theta) := \{x \in \mathbb{R}^d : x(j) \le \theta\}$ and $B_{>}(j, \theta) := \{x \in \mathbb{R}^d : x(j) > \theta\}$. We say two points $x, x' \in \mathbb{R}^d$ lie on the same side of the hyperplane (j, θ) if $x, x' \in B_{\le}(j, \theta)$ or $x, x' \in B_{>}(j, \theta)$; otherwise we say the two points lie on different sides of the hyperplane, or equivalently, they are separated by the hyperplane.

We consider decision trees as rooted directed trees. Nodes in the tree with no child are called *leaves*, and we require that every non-leaf node has exactly 2 children—a left child and a right child. Every non-leaf node corresponds to an axis-parallel hyperplane $(j, \theta) \in [d] \times \mathbb{R}$. This naturally make every node in the tree define a subset of $\mathbb{R}^d$ with axis-parallel boundaries: the root defines the entire space $\mathbb{R}^d$; if a non-leaf node corresponding to hyperplane (j, θ) defines the region B, its left child defines the region $B \cap B_{\le}(j, \theta)$, and its right child defines the region $B \cap B_{>}(j, \theta)$. Clearly, the regions defined by the leaves of a decision tree form a partition of $\mathbb{R}^d$.

For positive integers d and k, a k-clustering $\mathcal{C}$ for a set of points $x_1, \ldots, x_n \in \mathbb{R}^d$ consists of k centroids $y_1, \ldots, y_k \in \mathbb{R}^d$ and an assignment mapping $\xi : [n] \to [k]$. The k-means cost of the clustering $\mathcal{C}$ is given by $\mathrm{cost}(\mathcal{C}) = \sum_{i=1}^{n} \|x_i - y_{\xi(i)}\|_2^2$. We say the clustering $\mathcal{C}$ is k-explainable with respect to a decision tree T if T has at most k leaves and $\xi(i) = \xi(i')$ holds for all points $x_i, x_{i'}$ in the region defined by the same leaf of T.

For a subset $W \subseteq \mathbb{R}^d$, we use $|W|$ to denote its Lebesgue measure. We only care about the Lebesgue measure of bounded subsets that can be represented as a union of finitely many rectangles (or intervals when $d = 1$). For those subsets W, the Lebesgue measure $|W|$ always exists.

We use log to denote the base-e logarithm, and $\log_2$ to denote the base-2 logarithm.

3 Explainable k-means in the plane

We focus on the simpler case $d = 2$ in this section and give an efficient algorithm for finding a k-explainable clustering with competitive ratio $O(\log k \log\log k)$. Before we describe our algorithm, we remark that there exists a poly-time algorithm that computes a k-explainable clustering with *minimum* k-means cost given a set of n input points in $d = 2$ dimensions. In fact, for general $d \ge 2$ and $n \ge 2$, there exists such an algorithm with running time $n^{O(d)}$ via dynamic programming: if $k \ge n$, it is trivial to achieve zero cost; if $k < n$, the algorithm solves all subproblems each consisting of a box $(\alpha(1), \beta(1)] \times (\alpha(2), \beta(2)] \times \cdots \times (\alpha(d), \beta(d)]$ and a positive integer $k' \le k$, where the goal is to find a k'-explainable clustering with minimum cost for the input points inside the box. Although there are infinitely many such boxes, at most $n^{O(d)}$ among them define distinct subsets of input points, so essentially there are at most $kn^{O(d)} = n^{O(d)}$ different subproblems. Also, every subproblem can be solved in $(nd)^{O(1)}$ time using solutions to smaller subproblems.

While the above algorithm guarantees to find a k-explainable clustering with minimum cost and thus minimum competitive ratio, it does not give us a concrete bound on the competitive ratio. We develop a different algorithm that post-processes an arbitrary k-clustering $\mathcal{C}$ into a k-explainable clustering, and we show that the k-means cost of the explainable clustering is at most

$O(\log k \log\log k)$ times the cost of $\mathcal{C}$ assuming $d = 2$. Choosing $\mathcal{C}$ as the output of a constant-factor approximation algorithm for k-means ensures that the explainable clustering has competitive ratio $O(\log k \log\log k)$.

Theorem 1. *Assume $k \geq 2$. There exists a poly-time algorithm* `post-process_2d` *that takes a k-clustering $\mathcal{C}$ of n points in 2 dimensions, and outputs a clustering $\mathcal{C}'$ of the n points and a decision tree T with at most k leaves such that*

1. *$\mathcal{C}'$ is k-explainable with respect to T;*
2. *$\text{cost}(\mathcal{C}') \leq O(\log k \log\log_2(2k)) \cdot \text{cost}(\mathcal{C})$;*
3. *$\mathcal{C}'$ uses the same k centroids as $\mathcal{C}$ does.*

Consequently, there exists a poly-time algorithm that takes n points in 2 dimensions and outputs a k-explainable clustering with competitive ratio $O(\log k \log\log_2(2k))$.

In the rest of the section, we assume $d = 2$ and $k \geq 2$.

3.1 Subproblem

Our algorithm `post-process_2d` works in a recursive manner, constructing the tree from root to leaf. Thus, in each stage of the algorithm, we focus on a subset of the points and the centroids. Moreover, our algorithm keeps track of some helper information for every point. This leads us to the definition of a *subproblem*.

Definition 1 (Subproblem for $d = 2$)**.** *Given points $x_1, \ldots, x_n \in \mathbb{R}^d$ and centroids $y_1, \ldots, y_k \in \mathbb{R}^d$, a subproblem $\mathcal{P}$ consists of the following:*

1. *A subset $X \subseteq \{x_1, \ldots, x_n\}$. We focus on points $x \in X$.*
2. *A subset $Y \subseteq \{y_1, \ldots, y_k\}$. We focus on centroids $y \in Y$.*
3. *An assigned centroid $\sigma_x \in Y$ for every $x \in X$.*
4. *A length $\ell_x \geq 0$ for every point $x \in X$. We always enforce ℓ_x to be an upper bound on $\|x - \sigma_x\|_\infty$ (see Definition 4 Item 1). While we define the k-means cost using the L_2 norm, in our analysis we find it more convenient to keep track of the L_∞ norm instead.*
5. *A type t_x for every point $x \in X$. The type t_x is either a function $t_x : [d] \to \{0, 1, 2\}$ or the* irrelevant type *$t_x = \perp$. This gives a partition of X into two subsets: a subset*

$$R = \{x \in X : t_x \neq \perp\}$$

consisting of all relevant points*, and a subset $X \backslash R$ consisting of all* irrelevant points*. The set R is further partitioned into $R_0, R_1, \ldots, R_d$ defined as follows:*

$$R_i = \{x \in R : \|t_x\|_0 = i\}, \quad \text{for all } i \in [d],$$

where $\|t_x\|_0$ denotes the number of $j \in [d]$ with $t_x(j) \neq 0$. If $x \in R$ has $t_x(j) \neq 0$ for some $j \in [d]$, we ensure that x is close to one of the boundaries *in the j-th dimension, which we formalize in Definition 3 and Definition 4 Item 4.*

We fix a positive real number m as the *centroid mass* which we determine later. We can now define various quantities for a subproblem with respect to the centroid mass m. Later in our analysis, we relate these quantities to the cost of the explainable clustering we find for the subproblem.

Definition 2. *Given a subproblem $\mathcal{P}$, we define the following quantities:*

$$M(\mathcal{P}) := m|Y| + \sum_{x \in R_0} \ell_x^2,$$
$$A(\mathcal{P}) := f(M(\mathcal{P})) + 2^{32} \sum_{x \in R_1} \ell_x^2 + 2^9 \sum_{x \in R_2} \ell_x^2 + \sum_{x \in X \setminus R} \ell_x^2,$$

where

$$f(M) := 2^{57} M(1 + \log(M/m)) \log \log_2(2k).$$

Our algorithm crucially uses the boundaries and the diameter of a subproblem defined as follows.

Definition 3 (Subproblem boundary)**.** *Given a subproblem $\mathcal{P}$, for every $j \in [d]$, we define*

$$b_1(j) = \min_{y \in Y} y(j), \text{ and}$$
$$b_2(j) = \max_{y \in Y} y(j)$$

as the lower and upper boundaries *in the j-th dimension. Define $L := \max_{j \in [d]}(b_2(j) - b_1(j))$ as the diameter of the subproblem $\mathcal{P}$.*

To impose necessary constraints on the subproblems we deal with, we focus on *valid* subproblems defined as follows.

Definition 4 (Valid subproblem)**.** *Given points $x_1, \ldots, x_n \in \mathbb{R}^d$ and centroids $y_1, \ldots, y_k \in \mathbb{R}^d$, a subproblem $\mathcal{P} = (X, Y, (\sigma_x)_{x \in X}, (\ell_x)_{x \in X}, (t_x)_{x \in X})$ is valid if all of the following hold:*

1. *$\ell_x \geq \|x - \sigma_x\|_\infty$ for all $x \in X$.*
2. *$M(\mathcal{P})/m \leq 2k$.*
3. *For every point $x \in X \setminus R$ and every $y \in Y$, $\ell_x \geq \|x - y\|_\infty$.*
4. *If point $x \in R$ has $t_x(j) = 1$ for some $j \in [d]$, then $|x(j) - b_1(j)| \leq \ell_x$. Similarly, if $x \in R$ has $t_x(j) = 2$, then $|x(j) - b_2(j)| \leq \ell_x$.*

3.2 Making a single cut

We describe an efficient algorithm `single_cut_2d` that takes a valid subproblem $\mathcal{P}$, and produces two smaller valid subproblems $\mathcal{P}_1$ and $\mathcal{P}_2$ together with an axis-parallel hyperplane (j^*, θ) that separates them. Later in Section 3.3, we invoke this algorithm recursively to construct the algorithm `post-process_2d` required by Theorem 1.

Specifically, given an input subproblem $\mathcal{P} = (X, Y, (\sigma_x)_{x \in X}, (\ell_x)_{x \in X}, (t_x)_{x \in X})$, the algorithm `single_cut_2d` computes a partition X_1, X_2 of X, a partition Y_1, Y_2 of Y, new assignments $(\sigma'_x)_{x \in X}$, new lengths $(\ell'_x)_{x \in X}$, new types $(t'_x)_{x \in X}$, and outputs two smaller subproblems

$$\mathcal{P}_1 = (X_1, Y_1, (\sigma'_x)_{x \in X_1}, (\ell'_x)_{x \in X_1}, (t'_x)_{x \in X_1}),$$

$$\mathcal{P}_2 = (X_2, Y_2, (\sigma'_x)_{x\in X_2}, (\ell'_x)_{x\in X_2}, (t'_x)_{x\in X_2}). \tag{1}$$

The partitions (X_1, X_2) and (Y_1, Y_2) are determined by an axis-parallel hyperplane $(j^*, \theta) \in [d] \times (b_1(j^*), b_2(j^*))$:

$$\begin{aligned} X_1 &= X \cap B_{\le}(j^*, \theta) = \{x \in X : x(j^*) \le \theta\}, \\ X_2 &= X \cap B_{>}(j^*, \theta) = \{x \in X : x(j^*) > \theta\}, \\ Y_1 &= Y \cap B_{\le}(j^*, \theta) = \{y \in Y : y(j^*) \le \theta\}, \\ Y_2 &= Y \cap B_{>}(j^*, \theta) = \{y \in Y : y(j^*) > \theta\}. \end{aligned} \tag{2}$$

We always choose $j^* \in [d]$ so that $b_2(j^*) - b_1(j^*)$ is maximized, i.e., $b_2(j^*) - b_1(j^*) = L$. Note that by choosing $\theta \in (b_1(j^*), b_2(j^*))$, we are implicitly requiring $b_1(j^*) < b_2(j^*)$, or equivalently, $L > 0$, which we assume to be the case. Moreover, the choice $\theta \in (b_1(j^*), b_2(j^*))$ guarantees that Y_1 and Y_2 are both non-empty, and thus they both have sizes smaller than $|Y|$, which means that the two subproblems are indeed "smaller".

We say a point $x \in X$ is *σ-separated* if x and σ_x are separated by the hyperplane (j^*, θ). In other words, σ-separated points form the subset $X_+ \subseteq X$ defined as follows:

$$X_+ := \{x \in X_1 : \sigma_x \in Y_2\} \cup \{x \in X_2 : \sigma_x \in Y_1\}. \tag{3}$$

Consequently, non-σ-separated points belong to one of the following two sets

$$\begin{aligned} X_{11} &:= \{x \in X_1 : \sigma_x \in Y_1\}, \quad \text{and} \\ X_{22} &:= \{x \in X_2 : \sigma_x \in Y_2\}. \end{aligned} \tag{4}$$

To make sure that the two subproblems $\mathcal{P}_1, \mathcal{P}_2$ are well-defined, we require that $\sigma'_x \in Y_1$ whenever $x \in X_1$ and $\sigma'_x \in Y_2$ whenever $x \in X_2$. In other words, we require that no point is σ'-separated. This implies that for every σ-separated point $x \in X$, we must ensure $\sigma'_x \neq \sigma_x$. On the other hand, our algorithm `single_cut_2d` guarantees $\sigma'_x = \sigma_x$ whenever x is not σ-separated.

Our goal is to show that the two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ created by the `single_cut_2d` algorithm satisfy the following lemmas, which are crucial in our analysis to obtain Theorem 1. We always assume that the input subproblem $\mathcal{P}$ is valid even when we do not explicitly state so.

Lemma 2. *The two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ output by* `single_cut_2d` *are both valid.*

Lemma 3. *The two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ output by* `single_cut_2d` *satisfy $A(\mathcal{P}_1) + A(\mathcal{P}_2) \le A(\mathcal{P})$.*

We prove the above lemmas after describing the `single_cut_2d` algorithm step by step in the following subsections.

3.2.1 Preprocessing

For every $x \in R$ with $\ell_x \ge L/16$, we have

$$\|x - y\|_\infty \le \|x - \sigma_x\|_\infty + \|\sigma_x - y\|_\infty \le \ell_x + L \le 17\ell_x, \quad \text{for all } y \in Y. \tag{5}$$

For every such point x, we replace the current value of ℓ_x by $17\ell_x$, and set $t_x = \perp$ (thus removing x from R). The new subproblem is still valid (Definition 4 Item 3 follows from (5)), and all points $x \in R$ in the new subproblem satisfies $\ell_x \le L/16$. Moreover, it is clear that the value of $A(\mathcal{P})$ does not increase (note that $17^2 < 2^9$). From the rest of Section 3.2, we use $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X})$ to denote the subproblem *after* the preprocessing step.

3.2.2 Forbidding

We specify a subset F of the interval $(b_1(j^*), b_2(j^*))$ as the forbidden region. By making the algorithm `single_cut_2d` choose θ outside of the forbidden region, we can guarantee some desired properties for σ-separated points (see Lemma 5).

For every point $x \in X$, we define an interval W_x as follows:

$$W_x = \begin{cases} [\sigma_x(j^*), x(j^*)) \cap (b_1(j^*), b_2(j^*)), & \text{if } x(j^*) \geq \sigma_x(j^*); \\ [x(j^*), \sigma_x(j^*)) \cap (b_1(j^*), b_2(j^*)), & \text{if } x(j^*) < \sigma_x(j^*). \end{cases} \tag{6}$$

For every point $x \in X$, define η_x as its target new centroid in the event that x is σ-separated. Specifically, define

$$\begin{aligned} \eta_x &= \operatorname*{argmin}_{y \in Y(x)} \|\sigma_x - y\|_\infty, \\ q_x &= \|\sigma_x - \eta_x\|_\infty, \end{aligned} \tag{7}$$

where

$$Y(x) := \begin{cases} \{y \in Y : y(j^*) \geq \min\{x(j^*), b_2(j^*)\}\}, & \text{if } x(j^*) > \sigma_x(j^*); \\ \{y \in Y : y(j^*) \leq \max\{x(j^*), b_1(j^*)\}\}, & \text{if } x(j^*) < \sigma_x(j^*). \end{cases} \tag{8}$$

If x is σ-separated, it is clear that x and η_x must lie on the same side of the hyperplane (j^*, θ), in which case defining $\sigma'_x = \eta_x$ prevents x from being σ'-separated.

Now we focus on points x in the subset $T \subseteq S \subseteq R_0 \cup R_1$ defined as follows:

$$\begin{aligned} S &= \{x \in R_0 \cup R_1 : t_x(j^*) = 0, x(j^*) \neq \sigma_x(j^*)\}. \\ T &= \left\{ x \in S : \frac{q_x}{L} > 2^{11} \left(\frac{\ell_x}{L} \right)^{1/(d - \|t_x\|_0)} \right\}. \end{aligned} \tag{9}$$

In other words, a point $x \in S \cap R_0$ belongs to T if $q_x^2 > 2^{22} \ell_x L$; a point $x \in S \cap R_1$ belongs to T if $q_x > 2^{11} \ell_x$.

We define the forbidden region F as

$$F = (b_1(j^*), b_1(j^*) + L/8] \cup [b_2(j^*) - L/8, b_2(j^*)) \cup \bigcup_{x \in T} W_x.$$

It is clear that F can be represented as a union of finitely many *disjoint* intervals, and the representation can be computed in poly-time. We define F as above because choosing θ outside F guarantees that all relevant points that can possibly be σ-separated must have good properties summarized in Lemma 4 and Lemma 5 below, including having a relatively small value of q_x. Since the algorithm needs to choose $\theta \in (b_1(j^*), b_2(j^*))$ outside the forbidden region, it is necessary to show that the forbidden region does not cover the entire interval $(b_1(j^*), b_2(j^*))$. Lemma 6 below makes a stronger guarantee.

Lemma 4. *If we choose $\theta \in (b_1(j^*), b_2(j^*)) \backslash F$, then every $x \in R$ with $t_x(j^*) = 1$ belongs to X_{11}, and similarly every $x \in R$ with $t_x(j^*) = 2$ belongs to X_{22}. Consequently, every $x \in R \cap X_+$ satisfies $t_x(j^*) = 0$.*

Proof. Consider a point $x \in R$ with $t_x(j^*) = 1$. By Definition 4 Item 4, we know $|x(j^*) - b_1(j^*)| \leq \ell_x \leq L/16$, where the last inequality is guaranteed by the preprocessing step. By Definition 4

Item 1 and the triangle inequality, $|\sigma_x(j^*) - b_1(j^*)| \le \|x - \sigma_x\|_\infty + |x - b_1(j^*)| \le \ell_x + \ell_x \le L/8$. Therefore,

$$\max\{x(j^*), \sigma_x(j^*)\} \le b_1(j^*) + L/8 < \theta,$$

where the last inequality is because $(b_1(j^*), b_1(j^*) + L/8] \subseteq F$ and $\theta \in (b_1(j^*), b_2(j^*))\backslash F$. This implies $x \in X_{11}$. Similarly, every $x \in R$ with $t_x(j^*) = 2$ belongs to X_{22}. Since X_+ is disjoint from $X_{11} \cup X_{22}$, points $x \in R \cap X_+$ cannot have $t_x(j^*) \in \{1, 2\}$, and thus $t_x(j^*) = 0$. □

Lemma 5. *If we choose $\theta \in (b_1(j^*), b_2(j^*))\backslash F$, then every σ-separated relevant point $x \in R \cap X_+$ satisfies all of the following:*

1. *$x(j^*) \ne \sigma_x(j^*)$;*
2. *$t_x(j^*) = 0$ (and thus $\|t_x\|_0 \le 1$ and $x \in R_0 \cup R_1$);*
3. *$\frac{q_x}{L} \le 2^{11}(\frac{\ell_x}{L})^{1/(d-\|t_x\|_0)}$.*

Proof. Item 1 is obvious, since a point x cannot be σ-separated unless $x(j^*) \ne \sigma_x(j^*)$. Item 2 follows directly from Lemma 4.

Assume for the sake of contradiction that Item 3 is not satisfied by $x \in R \cap X_+$. We already know that Item 1 and Item 2 are both satisfied, so $x \in S$, and therefore $x \in T$. This implies $W_x \subseteq F$ by the definition of F. However, the fact that $x \in X_+$ implies $\theta \in W_x$, and thus $\theta \in F$, a contradiction. □

Lemma 6. *The forbidden region F has length (i.e. Lebesgue measure) at most $L/2$.*

Lemma 6 is a direct consequence of the following lemma:

Lemma 7. $|\bigcup_{x\in T} W_x| \le L/4$.

Proof. While we are dealing with $d = 2$ specifically, we prove the lemma using a more general language so that the proof can be reused in Section 4 where we deal with $d > 2$.

By Claim 27, we can find $U \subseteq T$ such that the intervals $(W_x)_{x\in U}$ are disjoint, and $|\bigcup_{x\in T} W_x| \le 3|\bigcup_{x\in U} W_x|$.

It remains to prove that $|\bigcup_{x\in U} W_x| \le L/12$. Define $U_> = \{x \in U : x(j^*) > \sigma_x(j^*)\}$.

We prove $|\bigcup_{x\in U_>} W_x| \le L/24$ via a volume argument. For every point $x \in U_>$, define a rectangular box $B_x \subseteq \mathbb{R}^d$ as follows:

$$B_x = \{z \in \mathbb{R}^d : \|z - \sigma_x\|_\infty \le q_x/3, z(j^*) - \sigma_x(j^*) > q_x/6\}.$$

We show that the boxes are pair-wise disjoint. Assume for the sake of contradiction that a point z lies in both boxes B_x and $B_{x'}$ where x, x' are distinct points in $U_>$. Assume w.l.o.g. $x(j^*) \le x'(j^*)$. Since W_x and $W_{x'}$ are disjoint, we have $\sigma_x(j^*) < x(j^*) \le \sigma_{x'}(j^*) < x'(j^*)$. Therefore, $\sigma_{x'} \in Y(x)$, and thus

$$q_x \le \|\sigma_x - \sigma_{x'}\|_\infty \le \|\sigma_x - z\|_\infty + \|\sigma_{x'} - z\|_\infty \le q_x/3 + q_{x'}/3.$$

This implies that $q_x \le q_{x'}/2$, and thus $z(j^*) \le \sigma_x(j^*) + q_x/3 \le \sigma_{x'}(j^*) + q_{x'}/6 < z(j^*)$, a contradiction.

It is clear by definition that $q_x \le L$ for all $x \in U_>$. Therefore, the boxes B_x are all contained in the large box

$$B := \{z \in \mathbb{R}^d : \forall j \in [d], b_1(j) - L/3 \le z(j) \le b_2(j) + L/3\}.$$

We define a projection $\pi : B \to B$ in the following way. For all $z \in B$, we define $\pi(z)(j^*) = z(j^*)$. For $j \neq j^*$, we define

$$\pi(z)(j) = \begin{cases} z(j), & \text{if } b_1(j) \leq z(j) \leq b_2(j), \\ b_1(j), & \text{if } z(j) < b_1(j), \\ b_2(j), & \text{if } z(j) > b_2(j). \end{cases}$$

This allows us to define another family of disjoint boxes. Specifically, define $\widetilde{B}_x$ as the set of points $z \in \mathbb{R}^d$ satisfying all of the following:

1. $z(j^*) \in (\sigma_x(j^*) + q_x/6, \sigma_x(j^*) + q_x/3)$;
2. for all $j \neq j^*$ with $t_x(j) = 0$, $z(j) \in (\sigma_x(j) - q_x/3, \sigma_x(j) + q_x/3)$;
3. for all j with $t_x(j) = 1$, $z(j) \in (b_1(j) - L/3, b_1(j))$;
4. for all j with $t_x(j) = 2$, $z(j) \in (b_2(j), b_2(j) + L/3)$.

It is clear that $\widetilde{B}_x \subseteq B$. Moreover, whenever $t_x(j) = 1$, we have $|b_1(j) - \sigma_x(j)| \leq |b_1(j) - x(j)| + \|x - \sigma_x\|_\infty \leq 2\ell_x \leq q_x/3$. Similarly, whenever $t_x(j) = 2$, we have $|b_2(j) - \sigma_x(j)| \leq q_x/3$. This implies that any $z \in \widetilde{B}_x$ has $\pi(z) \in B_x$. It is then easy to show that $\widetilde{B}_x$ are disjoint: if $z \in \widetilde{B}_x \cap \widetilde{B}_{x'}$, then $\pi(z) \in B_x \cap B_{x'}$, a contradiction.

The volume of $\widetilde{B}_x$ can be lower bounded as follows:

$$|\widetilde{B}_x| = (1/4)(2q_x/3)^{d-\|t_x\|_0}(L/3)^{\|t_x\|_0} \geq (5L/3)^d(24\ell_x/L),$$

where the last inequality is by the fact that $\frac{q_x}{L} > 2^{11}(\frac{\ell_x}{L})^{1/(d-\|t_x\|_0)}$, $d \geq 2$, and $\|t_x\|_0 \leq 1$. Summing up, we have

$$\sum_{x \in U_>} (5L/3)^d(24\ell_x/L) \leq \sum_{x \in U_>} |\widetilde{B}_x| \leq |B| \leq (5L/3)^d.$$

Therefore,

$$\sum_{x \in U_>} |W_x| \leq \sum_{x \in U_>} \ell_x \leq L/24.$$

A similar argument proves $|\bigcup_{x \in U \setminus U_>} W_x| \leq L/24$, which implies $|\bigcup_{x \in U} W_x| \leq L/12$ and completes the proof of the lemma. □

3.2.3 Cutting

Our algorithm `single_cut_2d` chooses $\theta \in (b_1(j^*), b_2(j^*)) \setminus F$ using a method by Seymour [Sey95] based on Lemma 8 below. Recall that any choice of θ defines a partition X_1, X_2 of X and a partition Y_1, Y_2 of Y as specified in (2). It also defines X_+, X_{11}, X_{22} as specified in (3) and (4). We further define

$$\begin{aligned} M_1^* &= m|Y_1| + \sum_{x \in R_0 \cap X_{11}} \ell_x^2, \\ M_2^* &= m|Y_2| + \sum_{x \in R_0 \cap X_{22}} \ell_x^2, \\ M^* &= \min\{M(\mathcal{P})/2, M(\mathcal{P}) - M_1^*, M(\mathcal{P}) - M_2^*\}. \end{aligned}$$

It is clear that $M_1^* + M_2^* \leq M(\mathcal{P})$.

Lemma 8. *There exists $\theta \in (b_1(j^*), b_2(j^*))\backslash F$ satisfying*

$$\sum_{x \in R_0 \cap X_+} \ell_x L \leq 8M^* \log(M(\mathcal{P})/M^*) \log\log_2(M(\mathcal{P})/m). \tag{10}$$

Moreover, θ can be computed in poly-time.

Proof. The fact that θ can be computed in poly-time follows immediately from its existence, because there are at most $|X|+|Y|$ choices of $\theta \in (b_1(j^*), b_2(j^*))\backslash F$ that lead to distinct partitions (X_1, X_2) and (Y_1, Y_2). It only takes poly-time to check (10) for each of the choices using the representation of F as a union of disjoint intervals. Below we prove the existence of θ.

Define $M = M(\mathcal{P})$. For every point $x \in R_0$, define a function $g_x : (b_1(j^*), b_2(j^*)) \to [0, +\infty)$ such that $g_x(\theta) = \ell_x^2/|W_x|$ if $\theta \in W_x$, and $g_x(\theta) = 0$ otherwise. Define $h(\theta)$ as the number of centroids $y \in Y$ with $y(j^*) < \theta$. Define

$$G(\theta) = mh(\theta) + \sum_{x \in R_0} \int_{b_1(j^*)}^{\theta} g_x(\theta')\mathrm{d}\theta'.$$

It is clear that G is non-decreasing, and bounded between m and $M - m$ for all $\theta \in (b_1(j^*), b_2(j^*))$. Moreover, for every choice of θ, we have $M_1^* \leq G(\theta)$ and $M_2^* \leq M - G(\theta)$. Define $I_1 = (\{x(j^*) : x \in X\} \cup \{y(j^*) : y \in Y\}) \cap (b_1(\ell^*), b_2(\ell^*))$. G is differentiable on $(b_1(\ell^*), b_2(\ell^*))\backslash I_1$, where $G'(\theta) = \sum_{x \in R_0} g(\theta)$.

By Lemma 6, the total length of the non-forbidden region is at least $L/2$. Therefore, we can find real numbers $\alpha_1, \ldots, \alpha_u$ and $\beta_1, \ldots, \beta_u$ such that

1. $b_1(j^*) \leq \alpha_1 < \beta_1 \leq \alpha_2 < \beta_2 \leq \cdots \leq \alpha_u < \beta_u \leq b_2(j^*)$;
2. every (α_i, β_i) is disjoint from the forbidden region F;
3. $\sum_{i=1}^{u} (\beta_i - \alpha_i) = L/2$.

Define $z_i := \sum_{i'=1}^{i} (\beta_{i'} - \alpha_{i'})$. We define a bijection γ from $\bigcup_{i=1}^{u} (z_{i-1}, z_i)$ to $\bigcup_{i=1}^{u} (\alpha_i, \beta_i)$ as follows: for all $z \in (z_{i-1}, z_i)$, define $\gamma(z) = \alpha_i + (z - z_{i-1})$. It is clear that γ is non-decreasing and has derivative $\gamma'(z) = 1$ for all $z \in \bigcup_{i=1}^{u} (z_{i-1}, z_i)$.

Define $I = \{z_1, \ldots, z_{m-1}\} \cup \{\gamma^{-1}(\theta) : \theta \in I_1 \cap \bigcup_{i=1}^{u} (\alpha_i, \beta_i)\}$. I is a finite subset of $(0, L/2)$. Define $V : (0, L/2)\backslash I \to [m, M - m]$ by $V(z) = G(\gamma(z))$. Then V is a non-decreasing function on $(0, L/2)\backslash I$ with derivative $V'(z) = G'(\gamma(z))$. By Lemma 29, we can find $z \in (0, L/2)\backslash I$ such that

$$V'(z) \leq (4/L) M' \log(M/M') \log\log_2(M/m),$$

where $M' := \min\{V(z), M - V(z)\}$. Choose $\theta = \gamma(z)$. We have $M' \leq V(z) = G(\theta) \leq M - M_2^*$ and $M' \leq M - V(z) = M - G(\theta) \leq M - M_1^*$. Therefore, $M' \leq M^*$. By Claim 26, we have

$$G'(\theta) = V'(z) \leq (8/L) M^* \log(M/M^*) \log\log_2(M/m).$$

The lemma is proved by noting that

$$G'(\theta) = \sum_{x \in R_0} g_x(\theta) \geq \sum_{x \in R_0 \cap X_+} g_x(\theta) \geq \sum_{x \in R_0 \cap X_+} \ell_x,$$

where the last inequality is by the easy fact that $\theta \in W_x$ whenever $x \in X_+$ and that $\ell_x^2/|W_x| \geq \ell_x$. □

3.2.4 Updating

We now specify the new assignments σ'_x, new lengths ℓ'_x, new types t'_x. The two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ can then be formed by (1).

For every non-σ-separated point $x \in X \backslash X_+$ we define $\sigma'_x = \sigma_x$, $\ell'_x = \ell_x$, and $t'_x = t_x$. For every σ-separated irrelevant point $x \in X_+ \backslash R$, we define $\ell'_x = \ell_x$, $t'_x = t_x (= \perp)$, and define σ'_x to be an arbitrary centroid in Y that lies on the same side of the hyperplane (j^*, θ) with x. Such a centroid exists because Y_1, Y_2 are both non-empty since we choose θ from $(b_1(j^*), b_2(j^*))$.

It remains to consider relevant points that are σ-separated, i.e. points $x \in R \cap X_+$. These points satisfy the properties in Lemma 5. For these points, we define

$$\ell'_x = \ell_x + 2^{11} L (\ell_x / L)^{1/(d - \|t_x\|_0)} \tag{11}$$

and $\sigma'_x = \eta_x$. We define t'_x to be equal to t_x, except that we change $t'_x(j^*)$ to either 1 or 2 from the original value $t_x(j^*) = 0$. Specifically, $t'_x(j^*) = 1$ if $x \in X_2$, and $t'_x(j^*) = 2$ if $x \in X_1$.

This completes our definition of σ'_x, ℓ'_x and t'_x. The algorithm `single_cut_2d` returns the two subproblem $\mathcal{P}_1, \mathcal{P}_2$ formed by (1) together with the hyperplane (j^*, θ). Before we prove Lemma 2 and Lemma 3, we first prove Lemma 9 below. Define $R' = \{x \in X : t'_x \neq \perp\}$. For $i = \{0, 1, 2\}$, define $R'_i = \{x \in R' : \|t'_x\|_0 = i\}$. It is clear from our update rules that $t'_x = \perp$ if and only if $t_x = \perp$, so $R' = R$.

Lemma 9. *$M(\mathcal{P}_1) = M_1^*$, and $M(\mathcal{P}_2) = M_2^*$.*

Proof. According to our updating rule, no point $x \in X_+$ has $\|t'_x\|_0 = 0$ because either $x \in X_+ \backslash R$ and $t'_x = \perp$, or $x \in R \cap X_+$ and $\|t'_x\|_0 \geq 1$. Therefore, a point $x \in X_1 \cap R'_0$ if and only if $x \in X_1 \backslash X_+ = X_{11}$, $t_x \neq \perp$, and $\|t_x\|_0 = 0$, or equivalently, $x \in R_0 \cap X_{11}$. This implies

$$M(\mathcal{P}_1) = m|Y_1| + \sum_{x \in X_1 \cap R'_0} (\ell'_x)^2 = m|Y_1| + \sum_{x \in R_0 \cap X_{11}} (\ell'_x)^2 = m|Y_1| + \sum_{x \in R_0 \cap X_{11}} \ell_x^2 = M_1^*.$$

Similarly, $M(\mathcal{P}_2) = M_2^*$. □

We concluse Section 3.2 by proving Lemma 2 and Lemma 3.

Proof of Lemma 2. Let $\mathcal{P} = (X, Y, (\sigma_x)_{x \in X}, (\ell_x)_{x \in X}, (t_x)_{x \in X})$ denote the valid subproblem *after* the preprocessing step.

We check every item in Definition 4. Item 3 follows immediately from the validity of $\mathcal{P}$ and the fact that $\sigma'_x = \sigma_x$ whenever $x \in X \backslash R' = X \backslash R$.

Now we prove Item 1. All non-σ-separated points $x \in X \backslash X_+$ have $\sigma'_x = \sigma_x$, and $\ell'_x = \ell_x$, so they satisfy $\ell'_x = \ell_x \geq \|x - \sigma_x\|_\infty = \|x - \sigma'_x\|_\infty$. By Item 3, all points in $X \backslash R = X \backslash R'$ also satisfy $\ell'_x \geq \|x - \sigma'_x\|_\infty$. It remains to check Item 1 for σ-separated relevant points $x \in R \cap X_+$. By Lemma 5, these points satisfy

$$\ell'_x = \ell_x + 2^{11} L (\ell_x / L)^{1/(d - \|t_x\|_0)} \geq \ell_x + q_x \geq \|x - \sigma_x\|_\infty + \|\sigma_x - \eta_x\|_\infty \geq \|x - \eta_x\|_\infty = \|x - \sigma'_x\|_\infty.$$

We now move on to Item 2. It suffices to prove that $\max\{M(\mathcal{P}_1), M(\mathcal{P}_2)\} \leq M(\mathcal{P})$, which follows directly from Lemma 9 and the fact that $M_1^* + M_2^* \leq M(\mathcal{P})$.

Now we prove Item 4. We prove it for $\mathcal{P}_1$, and omit the similar proof for $\mathcal{P}_2$. Define $b'_1(j), b'_2(j)$ similarly as $b_1(j), b_2(j)$ are defined in Definition 3 except that we replace Y by Y_1.

Suppose $x \in X_1 \cap R'$ has $t'_x(j) \neq 0$. If $j = j^*$ and $t'_x(j) = 2$, then by Lemma 4 it must be the case that $x \in R \cap X_+$. We have $\sigma'_x(j) \leq b'_2(j) \leq \sigma_x(j)$, so

$$|x(j) - b'_2(j)| \leq \max\{|x(j) - \sigma'_x(j)|, |x(j) - \sigma_x(j)|\} \leq \max\{\ell'_x, \ell_x\} = \ell'_x.$$

If $j \neq j^*$ or $t'_x(j) \neq 2$, we have $t'_x(j) = t_x(j)$. Define $i = t'_x(j) = t_x(j)$. If $i = 1$, we have $b_i(j) \leq b'_i(j) \leq \sigma'_x(j)$; if $i = 2$, we have $b_i(j) \geq b'_i(j) \geq \sigma'_x(j)$. In both cases,

$$|x(j) - b'_i(j)| \leq \max\{|x(j) - \sigma'_x(j)|, |x(j) - b_i(j)|\} \leq \max\{\ell'_x, \ell_x\} = \ell'_x. \qquad \square$$

Proof of Lemma 3. Since the preprocessing step preserves the validity of $\mathcal{P}$ and does not increase $A(\mathcal{P})$, we assume w.l.o.g. that $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X})$ is the subproblem *after* the preprocessing step. Define $M = M(\mathcal{P})$. We have

$$A(\mathcal{P}_1) + A(\mathcal{P}_2) = f(M(\mathcal{P}_1)) + f(M(\mathcal{P}_2)) + 2^{32} \sum_{x\in R'_1} (\ell'_x)^2 + 2^9 \sum_{x\in R'_2} (\ell'_x)^2 + \sum_{x\in X\setminus R} (\ell'_x)^2. \tag{12}$$

Moreover,

$$\begin{aligned}
\sum_{x\in R'_1} (\ell'_x)^2 &= \sum_{x\in R_1\cap R'_1} (\ell'_x)^2 + \sum_{x\in R_0\cap R'_1} (\ell'_x)^2 \\
&= \sum_{x\in R_1\cap R'_1} (\ell'_x)^2 + \sum_{x\in R_0\cap X_+} (\ell'_x)^2 \\
&= \sum_{x\in R_1\cap R'_1} \ell_x^2 + \sum_{x\in R_0\cap X_+} (\ell_x + 2^{11} L(\ell_x/L)^{1/2})^2 && \text{(by (11))} \\
&\leq \sum_{x\in R_1\cap R'_1} \ell_x^2 + 2^{23} \sum_{x\in R_0\cap X_+} \ell_x L \\
&\leq \sum_{i\in R_1\cap R'_1} \ell_x^2 + 2^{25} M^* \log(M/M^*) \log\log_2(2k), && (13)
\end{aligned}$$

where the last inequality is by Lemma 8. Similarly,

$$\begin{aligned}
\sum_{x\in R'_2} (\ell'_x)^2 &= \sum_{x\in R_2} (\ell'_x)^2 + \sum_{x\in R_1\cap R'_2} (\ell'_x)^2 \\
&= \sum_{x\in R_2} \ell_x^2 + \sum_{x\in R_1\cap R'_2} (\ell_x + 2^{11}\ell_x)^2 && \text{(by (11))} \\
&\leq \sum_{x\in R_2} \ell_x^2 + 2^{23} \sum_{x\in R_1\cap R'_2} \ell_x^2, && (14)
\end{aligned}$$

Applying Lemma 9,

$$\begin{aligned}
&f(M(\mathcal{P}_1)) + f(M(\mathcal{P}_2)) + 2^{57} M^* \log(M/M^*) \log\log_2(2k) \\
&= f(M_1^*) + f(M_2^*) + 2^{57} M^* \log(M/M^*) \log\log_2(2k) \\
&\leq f(M^*) + f(M - M^*) + 2^{57} M^* \log(M/M^*) \log\log_2(2k) \\
&\leq 2^{57}\Big(M^*(1 + \log(M^*/m)) \log\log_2(2k) + (M - M^*)(1 + \log(M/m)) \log\log_2(2k) \\
&\quad + M^* \log(M/M^*) \log\log_2(2k)\Big) \\
&= f(M). && (15)
\end{aligned}$$

Combining the inequalities as (13) $\times\, 2^{32}$ + (14) $\times\, 2^9$ + (15) and simplifying using (12), we get $A(\mathcal{P}_1) + A(\mathcal{P}_2) \leq A(\mathcal{P})$, as desired. $\square$

3.3 Building a decision tree

We prove Theorem 1 by describing the algorithm `post-process_2d` that takes an arbitrary k-clustering $\mathcal{C}$ and turns it into a k-explainable clustering with respect to a decision tree T.

Given the algorithm `single_cut_2d`, a natural algorithm `decision_tree_2d` that takes a valid subproblem $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X})$ and produces a decision tree T and an assigned centroid $\delta_x \in Y$ for every $x \in X$ works recursively as follows: if the diameter L of $\mathcal{P}$ is zero, which means that all centroids $y \in Y$ are at the same location, return $((\delta_x)_{x\in X}, T)$, where δ_x is identical for all x and equals to an arbitrary centroid $y \in Y$, and T is the tree with a single node; else, the algorithm `decision_tree_2d` calls `single_cut_2d` on $\mathcal{P}$, obtains a hyperplane (j^*, θ) and two subproblems $\mathcal{P}_1, \mathcal{P}_2$. The algorithm `decision_tree_2d` recursively calls itself on $\mathcal{P}_1$ and $\mathcal{P}_2$, and obtains $((\delta_x)_{x\in X_1}, T_1)$ and $((\delta_x)_{x\in X_2}, T_2)$. The algorithm constructs a tree T with root corresponding to (j^*, θ) and its left and right sub-trees being T_1 and T_2. The algorithm returns $((\delta_x)_{x\in X}, T)$.

Lemma 10. *Assuming the input $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X})$ to `decision_tree_2d` is valid. The output $((\delta_x)_{x\in X}, T)$ of `decision_tree_2d` satisfies the following properties. The decision tree T has at most $|Y|$ leaves. For every leaf v of the decision tree T and every pair of points $x, x' \in X$ in the region defined by v, we have $\delta_x = \delta'_x$. Moreover, $\sum_{x\in X} \|x - \delta_x\|_\infty^2 \le A(\mathcal{P})$.*

Proof. We prove the lemma by induction on $|Y|$. When $|Y| = 1$, we have $L = 0$ and thus `decision_tree_2d` returns without calling `single_cut_2d`. The lemma is trivial in this case. Now suppose the lemma is true when $|Y| < u$ for an integer $u > 1$, and we prove the lemma for $|Y| = u$. If $L = 0$, then again the lemma is trivial; otherwise, the algorithm `decision_tree_2d` calls `single_cut_2d`, and the lemma follows from the induction hypothesis together with Lemma 2 and Lemma 3. □

Given a k-clustering $\mathcal{C}$ of points $x_1, \ldots, x_n$ consisting of centroids $y_1, \ldots, y_k$, and an assignment mapping $\xi : [n] \to [k]$, the algorithm `post-process_2d` computes an *initial subproblem* $\widetilde{\mathcal{P}}$ by setting $X = \{x_1, \ldots, x_n\}, Y = \{y_1, \ldots, y_k\}, \sigma_{x_i} = y_{\xi(i)}, \ell_x = \|x - \sigma_x\|_\infty, t_x(j) = 0, \forall j \in [d]$. Setting the centroid mass $m = \frac{1}{k}\sum_{x\in X} \ell_x^2$, we have $\widetilde{\mathcal{P}}$ is valid and $A(\widetilde{\mathcal{P}}) = O(\log k \log\log_2(2k)) \cdot \mathrm{cost}(\mathcal{C})$. Algorithm `post-process` then calls `decision_tree_2d` on $\widetilde{\mathcal{P}}$ and obtains $(\delta_x)_{x\in X}$ and T. Algorithm `post-process` returns the decision tree T and a clustering $\mathcal{C}'$ with centroids $y_1, \ldots, y_k$ and assignment mapping $\xi' : [n] \to [k]$ such that $y_{\xi'(i)} = \delta_{x_i}$.

Proof of Theorem 1. Since `single_cut_2d` computes θ in polynomial time by Lemma 8, the entire algorithm `single_cut_2d` can be implemented in polynomial time. Consequently, `decision_tree_2d` and `post-process_2d` both run in polynomial time. Let $\mathcal{C}'$ and T be the output of algorithm `post-process_2d`. By Lemma 10, we know $\mathcal{C}'$ is k-explainable w.r.t. T, and

$$\begin{aligned}\mathrm{cost}(\mathcal{C}') = \sum_{i=1}^n \|x_i - y_{\xi'(i)}\|_2^2 \le 2\sum_{i=1}^n \|x_i - y_{\xi'(i)}\|_\infty^2 &= 2\sum_{i=1}^n \|x_i - \delta_{x_i}\|_\infty^2 \\ &\le 2A(\widetilde{\mathcal{P}}) \le O(\log k \log\log_2(2k))\mathrm{cost}(\mathcal{C}).\end{aligned}$$

It is by definition that $\mathcal{C}'$ uses the same centroids as $\mathcal{C}$ does. Choosing $\mathcal{C}$ as the output of a poly-time constant factor approximation algorithm for k-means, e.g. [KMN+04, ANFSW17], or since $d = 2$, a PTAS for k-means [FRS16, CAKM16], gives the competitive ratio bound $O(\log k \log\log_2(2k))$. □

4 Explainable k-means in $d > 2$ dimensions

We now describe our algorithm for higher dimensions, i.e., $d > 2$. The algorithm also works for $d = 2$ despite giving a worse bound than Theorem 1. We follow the same structure as the previous section, and emphasize the differences from it. Our goal is to prove the following high-dimensional analogue of Theorem 1.

Theorem 11. *Assume $k, d \geq 2$. There exists a poly-time algorithm* `post-process` *that takes a k-clustering $\mathcal{C}$ of n points in d dimensions, and outputs a clustering $\mathcal{C}'$ of the n points and a decision tree T with at most k leaves such that*

1. *$\mathcal{C}'$ is k-explainable with respect to T;*
2. *$\mathrm{cost}(\mathcal{C}') \leq O(k^{1-2/d}(\log k)^8(\log\log_2(2k))^3 d^4) \cdot \mathrm{cost}(\mathcal{C})$;*
3. *$\mathcal{C}'$ uses the same k centroids as $\mathcal{C}$ does.*

Consequently, there exists a poly-time algorithm that takes n points in d dimensions and outputs a k-explainable clustering with competitive ratio $O(k^{1-2/d}(\log k)^8(\log\log_2(2k))^3 d^4)$.

The bound in Theorem 11 can be improved when combined with [MS21].

Corollary 12 (In light of [MS21])**.** *Assume $k, d \geq 2$. There exists a poly-time algorithm that takes n points in d dimensions and outputs a k-explainable clustering with competitive ratio $k^{1-2/d}\,\mathrm{polylog}(k)$.*

Proof. Use the following algorithm: when $d \leq \log k$, invoke the algorithm in Theorem 11 to achieve competitive ratio

$$O(k^{1-2/d}(\log k)^8(\log\log_2(2k))^3 d^3) = k^{1-2/d}\,\mathrm{polylog}(k);$$

when $d > \log k$, invoke the algorithm in [MS21] to achieve competitive ratio

$$k\,\mathrm{polylog}(k) = k^{2/d}k^{1-2/d}\,\mathrm{polylog}(k) = O(1) \cdot k^{1-2/d}\,\mathrm{polylog}(k) = k^{1-2/d}\,\mathrm{polylog}(k). \qquad \square$$

In the rest of the section, we assume $k, d \geq 2$.

4.1 Subproblem

Compared to subproblems for $d = 2$ defined in Section 3.1, subproblems for $d > 2$ contain more information:

Definition 5 (Subproblem for $d > 2$)**.** *Given points $x_1, \ldots, x_n \in \mathbb{R}^d$ and centroids $y_1, \ldots, y_k \in \mathbb{R}^d$, besides what is included in Definition 1, a subproblem $\mathcal{P}$ consists of the following in addition:*

1. *A color $c_x \in \{-1, 0, \ldots, \lfloor \log_2 k \rfloor - 1\}$ for every point $x \in X$.*
2. *A scale $s_x \geq 1$ for every point $x \in X$;*
3. *A potential $p_x \in (0, +\infty)$ for every point $x \in X$.*

We fix a positive real number m as the *centroid mass* and define various quantities for a subproblem with respect to the centroid mass m similarly to Definition 2 but taking the potentials p_x into account.

Definition 6. *Given a subproblem $\mathcal{P}$, we define the following quantities:*

$$M(\mathcal{P}) := m|Y| + \sum_{x \in R} p_x \ell_x^2 (16(\log(2k))^2 \log\log_2(2k))^{2-\|t_x\|_0},$$
$$A(\mathcal{P}) := f(M(\mathcal{P})) + \sum_{x \in X \setminus R} p_x \ell_x^2,$$

where

$$f(M) := 16M(M/m)^{1/\log(2k)}(1 + \log(M/m)) \log\log_2(2k).$$

We define subproblem boundaries in the same way as Definition 3. Our definition for valid subproblems includes more requirements than Definition 7:

Definition 7 (Valid subproblem)**.** *Given points $x_1, \ldots, x_n \in \mathbb{R}^d$ and centroids $y_1, \ldots, y_k \in \mathbb{R}^d$, a subproblem $\mathcal{P} = (X, Y, (\sigma_x)_{x \in X}, (\ell_x)_{x \in X}, (t_x)_{x \in X}, (c_x)_{x \in X}, (s_x)_{x \in X}, (p_x)_{x \in X})$ is valid if in addition to the requirements in Definition 4, it satisfies all of the following:*

1. *For all $x \in R$, $\|t_x\|_0 \le 2$. (Thus, $R = R_0 \cup R_1 \cup R_2$.)*
2. *If $x \in R_0$, then ℓ_x is either zero or an integer power of 2, i.e., $\ell_x = 0$ or $\ell_x = 2^a$ for an integer a.*
3. *If $x \in R_0$, then $s_x = k$.*
4. *For every relevant point $x \in R$, its color $c_x = -1$ if and only if $x \in R_0$.*
5. *For every color $c \in \{0, \ldots, \lfloor \log_2 k \rfloor - 1\}$ and every relevant type $t : [d] \to \{0, 1, 2\}$, define $R_{c,t} = \{x \in R : c_x = c, t_x = t\}$ and $Y_{c,t} = \{\sigma_x : x \in R_{c,t}\}$. If $R_{c,t} \neq \emptyset$, there exists scale $s_{c,t} \ge 1$ and length $\ell_{c,t} \ge 0$ such that all points $x \in R_{c,t}$ have $s_x = s_{c,t}$ and $\ell_x = \ell_{c,t}$. Moreover, $|Y_{c,t}| \le s_{c,t}$.*

4.2 Making a single cut

As in Section 3.2, we describe an efficient algorithm `single_cut` that takes a valid subproblem $\mathcal{P} = (X, Y, (\sigma_x)_{x \in X}, (\ell_x)_{x \in X}, (t_x)_{x \in X}, (c_x)_{x \in X}, (s_x)_{x \in X}, (p_x)_{x \in X})$, and produces two smaller valid subproblems $\mathcal{P}_1$ and $\mathcal{P}_2$ together with an axis-parallel hyperplane (j^*, θ) that separates them. The two subproblems are defined by new assignments $(\sigma'_x)_{x \in X}$, new lengths $(\ell'_x)_{x \in X}$, new types $(t'_x)_{x \in X}$, new colors $(c'_x)_{x \in X}$, new scales $(s'_x)_{x \in X}$, and new potentials $(p'_x)_{x \in X}$ as follows

$$\mathcal{P}_1 = (X_1, Y_1, (\sigma'_x)_{x \in X_1}, (\ell'_x)_{x \in X_1}, (t'_x)_{x \in X_1}, (c'_x)_{x \in X_1}, (s'_x)_{x \in X_1}, (p'_x)_{x \in X_1}),$$
$$\mathcal{P}_2 = (X_2, Y_2, (\sigma'_x)_{x \in X_2}, (\ell'_x)_{x \in X_2}, (t'_x)_{x \in X_2}, (c'_x)_{x \in X_2}, (s'_x)_{x \in X_2}, (p'_x)_{x \in X_2}). \tag{16}$$

Again, we choose $j^* = \operatorname{argmax}_{j \in [d]}(b_2(j) - b_1(j))$, and define $X_1, X_2, Y_1, Y_2, X_+, X_{11}, X_{22}$ as in (2), (3) and (4). We assume that the input subproblem $\mathcal{P}$ is valid even when we do not explicitly state so, and that the diameter of $\mathcal{P}$ satisfies $L > 0$. At the end of the section, we prove the following lemmas for the algorithm `single_cut`.

Lemma 13. *The two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ output by* `single_cut` *are both valid.*

Lemma 14. *The two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ output by* `single_cut` *satisfy $A(\mathcal{P}_1) + A(\mathcal{P}_2) \le A(\mathcal{P})$.*

4.2.1 Preprocessing

If $x \in R_0 \cup R_1$ has $\ell_x \geq L/64$, we replace ℓ_x by $65\ell_x$, replace p_x by $p_x/65^2$, and set $t_x = \perp$ (thus removing x from R). If $x \in R_2$ has $\ell_x \geq L$, we replace ℓ_x by $2\ell_x$, replace p_x by $p_x/4$, and set $t_x = \perp$ (again removing x from R).

Similarly to Section 3.2.1, it is clear that the new subproblem is still valid, and in the new subproblem every point $x \in R_0 \cup R_1$ satisfies $\ell_x \leq L/64$, and every point $x \in R_2$ satisfies $\ell_x \leq L$. Moreover, the value of $A(\mathcal{P})$ does not increase. For the rest of Section 4.2, we use $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X}, (c_x)_{x\in X}, (s_x)_{x\in X}, (p_x)_{x\in X})$ to denote the subproblem *after* the preprocessing step.

4.2.2 Forbidding

Similarly to Section 3.2.2, We specify a subset F of the interval $(b_1(j^*), b_2(j^*))$ as the forbidden region.

For all $x \in X$, we define $W_x, \eta_x, q_x, Y(x)$ as in (6), (7), and (8). We also define S, T as in (9).

For a point $x \in R_0$ with $\ell_x \geq L/32k$, we change its color from $c_x = -1$ (guaranteed by Item 4 in Definition 7) to $c_x = \lfloor \log_2(\ell_x/(L/32k)) \rfloor$. Our preprocessing guarantees that $\ell_x \leq L/64$, so the new c_x is an integer between 0 and $\lfloor \log_2 k \rfloor - 1$. Moreover, points $x \in R_0$ have ℓ_x being an integer power of 2 by Item 2 in Definition 7. This means that if $x, x' \in R_0$ have $c_x = c_{x'} \geq 0$ after the color change, then $\ell_x = \ell_{x'}$. Since all points $x \in R_0$ have $s_x = k$ by Item 3 in Definition 7, Item 5 in Definition 7 still holds with the new colors and the induced new definitions of $R_{c,t}, Y_{c,t}, s_{c,t}, \ell_{c,t}$.

For every color $c \geq 0$, every type $t \neq \perp$, if $R_{c,t} = \emptyset$, define $H_{c,t} = \emptyset$. If $R_{c,t} \neq \emptyset$, we know $Y_{c,t} \neq \emptyset$, and we define $H_{c,t}$ as follows, where $E_{c,t}(y)$ is the interval $[y(j^*) - \ell_{c,t}, y(j^*) + \ell_{c,t}]$ for every $y \in Y_{c,t}$, and $\mathbb{1}_E(\cdot)$ denotes the indicator function of $E \subseteq (b_1(j^*), b_2(j^*))$:

$$H_{c,t} = \left\{ \theta \in (b_1(j^*), b_2(j^*)) : \sum_{y\in Y_{c,t}} \mathbb{1}_{E_{c,t}(y)}(\theta) > 48 \cdot 2^{\|t\|_0} \binom{d}{\|t\|_0} s_{c,t}\ell_{c,t}(\log_2 k)/L \right\}.$$

The entire forbidden region is

$$\begin{aligned} F = {} & (b_1(j^*), b_1(j^*) + L/32] \cup [b_2(j^*) - L/32, b_2(j^*)) \\ & \cup \bigcup_{y\in Y} [y(j^*) - L/32k, y(j^*) + L/32k] \cap (b_1(j^*), b_2(j^*)) \\ & \cup \bigcup_{x\in T} W_x \\ & \cup \bigcup_{c\geq 0, t\neq\perp} H_{c,t}. \end{aligned}$$

Similarly to the $d = 2$ case, F can be represented as a union of finitely many *disjoint* intervals, and the representation can be computed in poly-time. The following lemma has essentially the same proof as Lemma 4:

Lemma 15. *If we choose $\theta \in (b_1(j^*), b_2(j^*))\backslash F$, then every $x \in R_0 \cup R_1$ with $t_x(j^*) = 1$ belongs to X_{11}, and similarly every $x \in R_0 \cup R_1$ with $t_x(j^*) = 2$ belongs to X_{22}. Consequently, every $x \in (R_0 \cup R_1) \cap X_+$ satisfies $t_x(j^*) = 0$.*

Similarly to Lemma 5 and Lemma 6, we prove Lemma 16 and Lemma 17 below.

Lemma 16. *If we choose $\theta \in (b_1(j^*), b_2(j^*))\backslash F$, then every σ-separated relevant point $x \in R \cap X_+$ satisfies all of the following:*

1. *$x(j^*) \neq \sigma_x(j^*)$;*
2. *$\ell_x \geq L/64k$ (and thus $c_x \geq 0$);*
3. *if $x \in R_0 \cup R_1$, then $t_x(j^*) = 0$;*
4. *if $x \in R_0 \cup R_1$, then $\frac{q_x}{L} \leq 2^{11}(\frac{\ell_x}{L})^{1/(d-\|t_x\|_0)}$.*

Moreover, for every color $c \geq 0$ and every type $t \neq \perp$,

$$|\{\sigma_x : x \in R_{c,t} \cap X_+\}| \leq 48 \cdot 2^{\|t\|_0} \binom{d}{\|t\|_0} s_{c,t}\ell_{c,t}(\log_2 k)/L. \tag{17}$$

Proof. Item 3 follows directly from Lemma 15. Item 1 and Item 4 follow from the same argument in the proof of Lemma 5.

Assume for the sake of contradiction that Item 2 does not hold for $x \in R \cap X_+$. Then by Definition 4 Item 1, $|x(j^*) - \sigma_x(j^*)| \leq \|x - \sigma_x\|_\infty \leq \ell_x < L/64k$. Therefore,

$$W_x \subseteq [\sigma_x(j^*) - L/64k, \sigma_x(j^*) + L/64k] \cap (b_1(j^*), b_2(j^*)) \subseteq F.$$

However, the fact that $x \in X_+$ implies $\theta \in W_x$, and thus $\theta \in F$, a contradiction.

Assume for the sake of contradiction that (17) does not hold for color $c \geq 0$ and type $t \neq \perp$. For every $y \in \{\{\sigma_x : x \in R_{c,t} \cap X_+\}\} \subseteq Y_{c,t}$, we have $\theta \in E_{c,t}(y)$. Since (17) is violated, this means $\theta \in H_{c,t} \subseteq F$, a contradiction. □

Lemma 17. *The forbidden region F has length at most $L/2$.*

Proof. Using a similar argument to the proof of Lemma 7, we have $|\bigcup_{x\in T} W_x| \leq L/4$. It suffices to prove $|\bigcup_{c\geq 0, t\neq\perp} H_{c,t}| \leq L/8$.

For a uniform random $\theta \in (b_1(j^*), b_2(j^*))$, we have $\mathbb{E}[\mathbb{1}_{E_{c,t}(y)}(\theta)] \leq |E_{c,t}(y)|/L = 2\ell_{c,t}/L$. Therefore,

$$\mathbb{E}[\sum_{y\in Y_{c,t}} \mathbb{1}_{E_{c,t}(y)}(\theta)] \leq |Y_{c,t}| \cdot 2\ell_{c,t}/L \leq 2s_{c,t}\ell_{c,t}/L.$$

By Markov's inequality, $|H_{c,t}| \leq L/\left(24 \cdot 2^{\|t\|_0}\binom{d}{\|t\|_0}\log_2 k\right)$. Summing up over c, t, we have

$$\begin{aligned}
\sum_{c\geq 0, t\neq\perp} |H_{c,t}| &\leq \sum_{i=0}^{2} \sum_{\|t\|_0=i} \sum_{c=0}^{\lfloor \log_2 k\rfloor - 1} L/\left(24 \cdot 2^i \binom{d}{i} \log_2 k\right) \\
&\leq \sum_{i=0}^{2} \sum_{\|t\|_0=i} L/\left(24 \cdot 2^i \binom{d}{i}\right) \\
&= \sum_{i=0}^{2} L/24 \\
&\leq L/8.
\end{aligned}$$

□

4.2.3 Cutting

We choose $\theta \in (b_1(j^*), b_2(j^*))$ in a similar way as in Section 3.2.3 based on Lemma 18 below. Define

$$M_1^* = m|Y_1| + \sum_{x \in R \cap X_{11}} p_x \ell_x^2 (16(\log(2k))^2 \log\log_2(2k))^{2-\|t_x\|_0},$$
$$M_2^* = m|Y_2| + \sum_{x \in R \cap X_{22}} p_x \ell_x^2 (16(\log(2k))^2 \log\log_2(2k))^{2-\|t_x\|_0},$$
$$M^* = \min\{M(\mathcal{P})/2, M(\mathcal{P}) - M_1^*, M(\mathcal{P}) - M_2^*\}.$$

It is clear that $M_1^* + M_2^* \le M(\mathcal{P})$.

Lemma 18. *There exists a cut position $\theta \in (b_1(j^*), b_2(j^*))\backslash F$ with the following property.*

$$\sum_{x \in R \cap X_+} p_x \ell_x L (16(\log(2k))^2 \log\log_2(2k))^{2-\|t_x\|_0} \le 8M^* \log(M(\mathcal{P})/M^*) \log\log_2(M(\mathcal{P})/m).$$

Moreover, θ can be computed in poly-time.

We omit the proof as it is essentially the same as the proof of Lemma 8.

4.2.4 Updating

We now specify the new assignment σ'_x, new lengths ℓ'_x, new types t'_x, new colors c'_x, new scales s'_x, and new potentials p'_x. The two new subproblems $\mathcal{P}_1, \mathcal{P}_2$ can then be formed by (16).

For every non-σ-separated point $x \in X\backslash X_+$ we define $\sigma'_x = \sigma_x$, $\ell'_x = \ell_x$, $t'_x = t_x$, $s'_x = s_x$, $p'_x = p_x$, and define $c'_x = -1$ if $x \in R_0$ and $c'_x = c_x$ otherwise. For every σ-separated irrelevant point $x \in X_+\backslash R$, we define $\ell'_x = \ell_x$, $t'_x = t_x (= \perp)$, $s'_x = s_x$, $p'_x = p_x$, $c'_x = c_x$, and define σ'_x to be an arbitrary centroid in Y that lies on the same side of the hyperplane (j^*, θ) with x.

It remains to consider relevant points that are σ-separated, i.e. points $x \in R \cap X_+$. We define

$$s'_x = 48 \cdot 2^{\|t_x\|_0} \binom{d}{\|t_x\|_0} s_x \ell_x (\log_2 k)/L. \tag{18}$$

It is clear that $s'_x \ge 1$ because otherwise no point in R_{c_x,t_x} should be σ-separated by (17). Define σ'_x to be the centroid $y \in Y$ with the minimum $\|y - \sigma_x\|_\infty$ that lies on the same side of the hyperplane (j^*, θ) with x. Since every centroid in $Y(x)$ is a candidate for σ'_x, we have $\|\sigma_x - \sigma'_x\|_\infty \le q_x$. We define $c'_x = c_x$. The definition for ℓ'_x, p'_x, t'_x depends on whether $x \in R_0 \cup R_1$ or $x \in R_2$ as follows.

If $x \in (R_0 \cup R_1) \cap X_+$, define

$$\ell'_x = 2^{12} L (\ell_x/L)^{1/(d-\|t_x\|_0)}. \tag{19}$$

Define p'_x so that $p'_x(\ell'_x)^2 = p_x \ell_x L$, or equivalently,

$$p'_x = p_x (\ell_x/L)^{1-2/(d-\|t_x\|_0)}/2^{24}. \tag{20}$$

We define t'_x to be equal to t_x, except that we change $t'_x(j^*)$ to either 1 or 2 from the original value $t_x(j^*) = 0$. Specifically, $t'_x(j^*) = 1$ if $x \in X_2$, and $t'_x(j^*) = 2$ if $x \in X_1$.

If $x \in R_2 \cap X_+$, define $\ell'_x = 2L$. Again, define p'_x so that $p'_x(\ell'_x)^2 = p_x \ell_x L$, or equivalently,

$$p'_x = p_x (\ell_x/L)/4. \tag{21}$$

Also, define $t'_x = \perp$.

This completes our definition of $\sigma'_x, \ell'_x, t'_x, c'_x, s'_x$ and p'_x. The algorithm `single_cut` outputs the two new subproblems $\mathcal{P}_1$ and $\mathcal{P}_2$ formed according to (16) together with the hyperplane (j^*, θ). Before we prove Lemma 13 and Lemma 14, we need some inequalities about $M(\mathcal{P}_1), M(\mathcal{P}_2), M_1^*, M_2^*$, and M^*.

Lemma 19. *We have the following inequalities:*

$$\begin{aligned} M(\mathcal{P}_1) \geq M_1^*, \quad \text{and} \quad M(\mathcal{P}_2) \geq M_2^*, \\ \min\{M_1^*, M_2^*\} \leq M^*, \quad \text{and} \quad \max\{M_1^*, M_2^*\} \leq M - M^*. \end{aligned}$$

Moreover,

$$M(\mathcal{P}_1) + M(\mathcal{P}_2) - M_1^* - M_2^* \leq M^*/2\log(2k).$$

Proof. By our updating rules,

$$\begin{aligned} M(\mathcal{P}_1) - M_1^* &= \sum_{x \in (R_0 \cup R_1) \cap X_+ \cap X_1} p'_x (\ell'_x)^2 (16\log(2k)\log\log_2(2k))^{2-\|t'_x\|_0} \geq 0, \\ M(\mathcal{P}_2) - M_1^* &= \sum_{x \in (R_0 \cup R_1) \cap X_+ \cap X_2} p'_x (\ell'_x)^2 (16\log(2k)\log\log_2(2k))^{2-\|t'_x\|_0} \geq 0. \end{aligned}$$

Summing up,

$$\begin{aligned} & M(\mathcal{P}_1) + M(\mathcal{P}_2) - M_1^* - M_2^* \\ = & \sum_{x \in (R_0 \cup R_1) \cap X_+} p'_x (\ell'_x)^2 (16\log(2k)\log\log_2(2k))^{2-\|t'_x\|_0} \\ = & \sum_{x \in (R_0 \cup R_1) \cap X_+} p_x \ell_x L (16\log(2k)\log\log_2(2k))^{2-\|t'_x\|_0} \\ = & \sum_{x \in (R_0 \cup R_1) \cap X_+} p_x \ell_x L (16\log(2k)\log\log_2(2k))^{2-(\|t_x\|_0+1)} \\ \leq & (16(\log(2k))^2 \log\log_2(2k))^{-1} \sum_{x \in R \cap X_+} p_x \ell_x L (16\log(2k)\log\log_2(2k))^{2-\|t_x\|_0} \\ \leq & M^*/2\log(2k). \qquad \text{(by Lemma 18)} \end{aligned}$$

It remains to prove $\min\{M_1^*, M_2^*\} \leq M^*$ and $\max\{M_1^*, M_2^*\} \leq M - M^*$. Assume w.l.o.g. that $M_1^* \leq M_2^*$. Since $M_1^* + M_2^* \leq M$, we have $M_1^* \leq \min\{M/2, M - M_2^*\}$. Combining this with the definition of M^*, we have $M_1^* \leq M^*$. Also, the definition of M^* directly implies $M_2^* \leq M - M^*$. □

We conclude Section 4.2 by proving Lemma 13 and Lemma 14.

Proof of Lemma 13. Let $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X}, (c_x)_{x\in X}, (s_x)_{x\in X}, (p_x)_{x\in X})$ denote the valid subproblem *after* the preprocessing step.

We first check every item in Definition 4. Item 3 follows from a similar argument to the proof of Lemma 2. The only difference is that we need to consider points $x \in R_2 \cap X_+$ separately. For these points, we have $t'_x = \perp$. By our preprocessing step, we have $\ell_x \leq L$, and thus for all $y \in Y$,

$$\ell'_x = 2L \geq L + \ell_x \geq \|\sigma_x - y\|_\infty + \|x - \sigma_x\|_\infty \geq \|x - y\|_\infty.$$

Item 1 also follows from a similar argument to the proof of Lemma 2, noting that for $x \in (R_0 \cup R_1) \cap X_+$, update rule (19) implies

$$\ell'_x \geq \ell_x + 2^{11} L(\ell_x/L)^{1/(d-\|t_x\|_0)} \geq \ell_x + q_x \geq \ell_x + \|\sigma_x - \sigma'_x\|_\infty \geq \|x - \sigma'_x\|_\infty.$$

Defining $M = M(\mathcal{P})$, Item 2 holds because of Lemma 19:

$$\max\{M(\mathcal{P}_1), M(\mathcal{P}_2)\} \leq \max\{M_1^*, M_2^*\} + M^*/2\log(2k) \leq M - M^* + M^*/2\log(2k) \leq M \leq km.$$

Definition 4 Item 4 follows from the same argument as the proof of Lemma 2.

We now check every item in Definition 7. Item 1 is clear from our update rules. Item 2, Item 3 and Item 4 follow from the fact that $\|t'_x\|_0 = 0$ if and only if $x \in R_0 \backslash X_+$.

We prove Definition 7 Item 5 for $\mathcal{P}_1$, and omit the similar proof for $\mathcal{P}_2$. For color $c \geq 0$ and type $t \neq \perp$, define $R'_{c,t} = \{x \in X_1 : c'_x = c, t'_x = t\}$.

If $t(j^*) \neq 2$, we know $R'_{c,t} \cap X_+ = \emptyset$, so $R'_{c,t} \subseteq R_{c,t} \backslash X_+$. Therefore, for $x \in R'_{c,t}$, we have $s'_x = s_x = s_{c,t}, \ell'_x = \ell_x = s_{c,t}, \sigma'_x = \sigma_x = \sigma_{c,t}$. Definition 7 Item 5 is trivial in this case.

If $t(j^*) = 2$, by Lemma 15, we know $R'_{c,t} \subseteq (R_0 \cup R_1) \cap X_+$, and $R'_{c,t} \subseteq R_{c,\widetilde{t}}$, where $\widetilde{t}$ is equal to t except that $\widetilde{t}(j^*) = 0$. By the update rules (18) and (19), all points $x \in R'_{c,t}$ have

$$\ell'_x = 2^{12} L(\ell_{c,\widetilde{t}}/L)^{1/(d-\|\widetilde{t}\|_0)}, \tag{22}$$

$$s'_x = 48 \cdot 2^{\|\widetilde{t}\|_0} \binom{d}{\|\widetilde{t}\|_0} s_{c,\widetilde{t}} \ell_{c,\widetilde{t}} (\log_2 k)/L. \tag{23}$$

Denote the right-hand-sides of (22) and (23) as $\ell'_{c,t}$ and $s'_{c,t}$, respectively. We have

$$|\{\sigma'_x : x \in R'_{c,t}\}| \leq |\{\sigma'_x : x \in R_{c,\widetilde{t}} \cap X_+ \cap X_1\}| \leq |\{\sigma_x : x \in R_{c,\widetilde{t}} \cap X_+ \cap X_1\}| \leq s'_{c,t},$$

where the second inequality is because points in $R \cap X_+ \cap X_1$ with the same σ_x have the same σ'_x, and the last inequality is by (17) and (23). □

Proof of Lemma 14. Since the preprocessing step preserves the validity of $\mathcal{P}$ and does not increase $A(\mathcal{P})$, we assume w.l.o.g. that $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X}, (c_x)_{x\in X}, (s_x)_{x\in X}, (p_x)_{x\in X})$ is the subproblem *after* the preprocessing step. Assume w.l.o.g. $M_1^* \leq M_2^*$. Define $M = M(\mathcal{P}), M_1 = M^* + M(\mathcal{P}_1) - M_1^*$, and $M_2 = (M - M^*) + M(\mathcal{P}_2) - M_2^*$. By Lemma 19,

$$M^* \leq M_1 \leq M^* + M^*/2\log(2k) \leq 3M^*/2 \leq 3M/4, \tag{24}$$

$$M^* \leq M - M^* \leq M_2, \tag{25}$$

$$M(\mathcal{P}_1) \leq M_1, \quad \text{and} \tag{26}$$

$$M(\mathcal{P}_2) \leq M_2. \tag{27}$$

Lemma 19, (24) and (25) give us

$$M_1 + M_2 \leq M + M^*/2\log(2k) \leq M + \min\{M_1, M_2\}/2\log(2k). \tag{28}$$

Define $R' = \{x \in X : t'_x \neq \perp\}$. We have

$$\begin{aligned}\sum_{x\in X\backslash R'} p'_x(\ell'_x)^2 - \sum_{x\in X\backslash R} p_x \ell_x^2 &= \sum_{x\in R_2\cap X_+} p'_x(\ell'_x)^2 \\ &= \sum_{x\in R_2\cap X_+} p_x \ell_x L\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{x\in R\cap X_+} p_x\ell_x L((\log k)^2\log\log k)^{2-\|t_x\|_0}\\
&\le 8M^*\log(M/M^*)\log\log_2(M/m) && \text{(by Lemma 18)}\\
&\le 16M_1\log(M/M_1)\log\log_2(M/m). && \text{(by (24) and Claim 26)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&f(M(\mathcal{P}_1))+f(M(\mathcal{P}_2))+\sum_{x\in X\setminus R'} p'_x(\ell'_x)^2-\sum_{x\in X\setminus R} p_x\ell_x^2\\
&\le f(M_1)+f(M_2)+16M_1\log(M/M_1)\log\log_2(M/m) && \text{(by (26), (27))}\\
&\le 16(M_1(M_1/m)^{1/\log(2k)}+M_2(M_2/m)^{1/\log(2k)})(1+\log(M/m))\log\log_2(2k)\\
&\le 16M(M/m)^{1/\log(2k)}(1+\log(M/m))\log\log_2(2k) && \text{(by (28) and Claim 25)}\\
&= f(M).
\end{aligned}
$$

Rearranging the inequality above,

$$
A(\mathcal{P}_1)+A(\mathcal{P}_2)=f(M(\mathcal{P}_1))+f(M(\mathcal{P}_2))+\sum_{x\in X\setminus R'} p'_x(\ell'_x)^2\le f(M)+\sum_{x\in X\setminus R} p_x(\ell_x)^2=A(\mathcal{P}). \quad \square
$$

4.3 Building a decision tree

The algorithm `post-process` we use to prove Theorem 11 is similar to `post-process_2d` in Section 3.3. We first construct an algorithm `decision_tree` similar to `decision_tree_2d`, except that it takes a subproblem $\mathcal{P}$ in $d\ge 2$ dimensions and invokes the algorithm `single_cut` we developed in Section 4.2 instead of `single_cut_2d`. The algorithm `post-process` calls `decision_tree` on an initial subproblem $\widetilde{\mathcal{P}}$, which we defined as follows.

Given a k-clustering $\mathcal{C}$ of points $x_1,\dots,x_n$ consisting of centroids $y_1,\dots,y_k$, and an assignment mapping $\xi:[n]\to[k]$, the algorithm `post-process` computes the initial subproblem $\widetilde{\mathcal{P}}$ by setting $X=\{x_1,\dots,x_n\}, Y=\{y_1,\dots,y_k\}, \sigma_{x_i}=y_{\xi(i)}, t_x(j)=0, \forall j\in[d], s_x=k, c_x=-1$, and

$$
\begin{aligned}
\ell_x &= \begin{cases} 2^{\lceil\log_2\|x-\sigma_x\|_\infty\rceil}, & \text{if } \|x-\sigma_x\|_\infty>0;\\ 0, & \text{if } \|x-\sigma_x\|_\infty=0;\end{cases}\\
p_x &= 2^{52}k^{1-2/d}d^3(48\log_2 k)^3.
\end{aligned} \tag{29}
$$

Set $m=\frac{1}{k}\sum_{x\in X}p_x\ell_x^2(16(\log(k))^2\log\log_2(2k))^2$ so that $M(\widetilde{\mathcal{P}})/m=2k$. It is clear that $\widetilde{\mathcal{P}}$ is valid and

$$
A(\widetilde{\mathcal{P}})=O(k^{1-2/d}(\log k)^8(\log\log_2(2k))^3d^3)\cdot\mathrm{cost}(\mathcal{C}).
$$

After obtaining the output $((\delta_x)_{x\in X},T)$ from `decision_tree`, `post-process` returns the decision tree T and a clustering $\mathcal{C}'$ with centroids $y_1,\dots,y_k$ and assignment mapping $\xi':[n]\to[k]$ such that $y_{\xi'(i)}=\delta_{x_i}$.

Before we prove Theorem 11, we need the following lemma showing that the potential of a point never drops below 1:

Lemma 20. *In the process of running algorithm* `post-process`*, whenever the algorithm* `single_cut` *is called, the input subproblem* $\mathcal{P}=(X,Y,(\sigma_x)_{x\in X},(\ell_x)_{x\in X},(t_x)_{x\in X},(c_x)_{x\in X},(s_x)_{x\in X},(p_x)_{x\in X})$ *to* `single_cut` *is valid and satisfies* $\forall x\in X, p_x\ge 1$.

Proof. The validity of $\mathcal{P}$ follows from an induction using the validity of $\widetilde{\mathcal{P}}$ and Lemma 13. Below we prove $\forall x \in X, p_x \geq 1$.

Fix a point x_u for $u \in [n]$. For $i \in \{0,1,2\}$ consider the moments when `single_cut` is called with a subproblem $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X}, (c_x)_{x\in X}, (s_x)_{x\in X}, (p_x)_{x\in X})$ that satisfy $x_u \in X$ and $\|t_{x_u}\|_0 = i$. These moments may not exist, or there may be multiple such moments. However, as long as there is at least one such moment, the tuple $(s_{x_u}, p_{x_u}, \ell_{x_u}, t_{x_u})$ must be identical for all such moments, because algorithm `single_cut` always keeps $(s'_x, p'_x, \ell'_x, t'_x)$ equal (s_x, p_x, ℓ_x, t_x) except when $t'_x = \perp \neq t_x$ or $\|t'_x\|_0 = \|t_x\|_0 + 1$. We use $(s(i), p(i), \ell(i))$ to denote the identical tuple $(s_{x_i}, p_{x_i}, \ell_{x_i})$ over all such moments given $i \in \{0,1,2\}$. The diameter L of $\mathcal{P}$ may not be identical over all such moments, so we use $L(i)$ to denote the value of L for the *last* such moment. We define $u(i) = L(i)/\ell(i)$. Similarly, we use $(s(\perp), p(\perp))$ to denote the identical value of (s_{x_u}, p_{x_u}) whenever $t_{x_u} = \perp$.

Our goal is to show each of $p(0), p(1), p(2), p(\perp)$, whenever exists, is at least 1.

For $i \in \{0,1,2\}$, whenever $p(i)$ exists, by (20) we have

$$p(i) = p(0) / \prod_{i'=0}^{i-1} \left(2^{24} u(i')^{1-2/(d-i')}\right) \geq p(0) 2^{-24i} / \left(\prod_{i'=0}^{i-1} u(i')\right)^{1-2/d}. \tag{30}$$

By (18) we have

$$1 \leq s(i) \leq k \prod_{i'=0}^{i-1} \left(48 \cdot 2^{i'} \binom{d}{i'} (\log_2 k)/u(i')\right).$$

Therefore,

$$\prod_{i'=0}^{i-1} u(i') \leq k \prod_{i'=0}^{i-1} \left(48 \cdot 2^{i'} \binom{d}{i'} \log_2 k\right).$$

Combining this with (29) and (30),

$$\begin{aligned} p(0) &= 2^{52} k^{1-2/d} d^3 (48 \log_2 k)^3 \geq 65^2 > 1, \\ p(1) &\geq 2^{28} d^3 (48 \log_2 k)^2 \geq 65^2 > 1, \\ p(2) &\geq 2^3 d^2 (48 \log_2 k) \geq 4 > 1. \end{aligned} \tag{31}$$

Finally, we show $p(\perp) \geq 1$ whenever $p(\perp)$ exists. If the transition of t_{x_u} to $\perp$ happens at preprocessing, it is clear that either $p(\perp) \geq p(0)/65^2 \geq 1$ or $p(\perp) \geq p(1)/65^2 \geq 1$ or $p(\perp) \geq p(2)/4 \geq 1$. Otherwise, by (20) and (21) we have

$$p(\perp) = p(0) / \left(4u(2) \prod_{i=0}^{1} (2^{24} u(i)^{1-2/(d-i)})\right) = p(0) 2^{-50} / (u(0)^{1-2/d} u(1)^{1-2/(d-1)} u(2)). \tag{32}$$

By (18), we have

$$1 \leq s(\perp) \leq k \prod_{i=0}^{2} \left(48 \cdot 2^{i} \binom{d}{i} (\log_2 k)/u(i)\right). \tag{33}$$

Therefore,

$$u(0)u(1)u(2) \leq 4kd^3 (48 \log_2 k)^3.$$

From the update rule (19), we know

$$u(2) = \frac{L(2)}{\ell(2)} \leq \frac{L(1)}{\ell(2)} = u(1)^{1/(d-1)} 2^{-12} \leq u(1)^{1/(d-1)},$$

which implies

$$u(2)^{2/d} \le u(1)^{2/(d(d-1))} = u(1)^{2/(d-1)-2/d}.$$

Therefore,

$$u(0)^{1-2/d}u(1)^{1-2/(d-1)}u(2) \le u(0)^{1-2/d}u(1)^{1-2/d}u(2)^{1-2/d} \le 4k^{1-2/d}d^3(48\log_2 k)^3, \tag{34}$$

where the last inequality is by (33). Plugging (31) and (34) into (32), we get $p(\perp) \ge 1$, as desired. □

Lemma 21. *In the process of running algorithm* `post-process`*, whenever* `decision_tree` *is called with input being subproblem* $\mathcal{P} = (X, Y, (\sigma_x)_{x\in X}, (\ell_x)_{x\in X}, (t_x)_{x\in X}, (c_x)_{x\in X}, (s_x)_{x\in X}, (p_x)_{x\in X})$*, the output* $((\delta_x)_{x\in X}, T)$ *of* `decision_tree` *satisfies the following properties. The decision tree* T *has at most* $|Y|$ *leaves. For every leaf* v *of the decision tree* T *and every pair of points* $x, x' \in X$ *in the region defined by* v*, we have* $\delta_x = \delta'_x$*. Moreover,* $\sum_{x\in X} \|x - \delta_x\|_\infty^2 \le A(\mathcal{P})$*.*

Proof. The proof is essentially the same as the proof of Lemma 10. The only difference is that when $L = 0$, to prove $\sum_{x\in X} \|x - \delta_x\|_\infty^2 \le A(\mathcal{P})$, we need $p_x \ge 1$ from Lemma 20. □

Proof of Theorem 11. The proof is essentially the same as the proof of Theorem 1 except that we use Lemma 18 instead of Lemma 8, and Lemma 21 instead of Lemma 10. □

5 Lower Bound

We prove a lower bound of $k^{1-2/d}/\text{polylog}(k)$ on the competitive ratio for d-dimensional explainable k-means for all $k, d \ge 2$ (Theorem 24). Our proof is based on a construction by [LM21] summarized in Lemma 22 below. For brevity, we do not repeat its proof here. The construction allows us to show a competitive ratio lower bound depending on two other parameters p and b in Lemma 23. We then prove Theorem 24 by specifying the values of p and b.

Lemma 22 ([LM21])**.** *Let* b, p *be positive integers satisfying* $b \ge 3$*. There exists* b^p *points in* p *dimensions with the following properties:*

1. *For every* $j \in [p]$*, the* j*-th coordinate of the* b^p *points form a permutation of* $\{0, \dots, b^p - 1\}$*.*
2. *The* L_2 *distance between any two of the points is at least* $b^{p-1}/2$*.*

Lemma 23. *Given positive integers* k, d, p, b *satisfying* $p \le d, b \ge 3, b^p \le k$*, there exists a set of points in* d *dimensions for which any* k*-explainable clustering has competitive ratio* $\Omega(b^{p-2}/p)$ *for the* k*-means cost.*

Proof. When $p = 1$, it is trivial to show a competitive ratio lower bound of $1 = \Omega(b^{p-2}/p)$, so we assume $p \ge 2$.

Let Z denote the set of b^p points $\mathbb{R}^p$ from Lemma 22. Every point $z \in Z$ creates a point $u(z) \in \mathbb{R}^d$, where the first p coordinates of $u(z)$ are equal to the p coordinates of z, and the remaining $d - p$ coordinates of $u(z)$ are zeros. Let $Y_1 = \{u(z) : z \in Z\}$ be the resulting set of b^p points in $\mathbb{R}^d$. We construct a set $Y_2 \subseteq \mathbb{R}^d$ consisting of $k - b^p$ points that are sufficiently far away from each other and from the points in Y_1.

For every $y \in Y_1$, we create a set X_y consisting of $2p$ points in $\mathbb{R}^d$ as follows: for every $j \in [p]$, X_y contains 2 points $y_j^+, y_j^- \in \mathbb{R}^d$ by adding $2/3$ and $-2/3$ to the j-th coordinate of y. For every $y \in Y_2$, we define X_y to be the set containing only y itself.

Given the $2pb^p + (k - b^p)$ points in $\bigcup_{y \in Y_1 \cup Y_2} X_y$, there exists a k-clustering $\mathcal{C}$ with $\text{cost}(\mathcal{C}) = O(pb^p)$ by choosing $Y_1 \cup Y_2$ to be the centroids. On the other hand, for any k-explainable clustering $\mathcal{C}'$, by Lemma 22 Item 2 there exist two points x, x' assigned to the same centroid with $x \in X_y, x' \in X_{y'}$ for distinct $y, y' \in Y_1 \cup Y_2$. Therefore, $\text{cost}(\mathcal{C}') \geq 1/2\|x - x'\|_2^2 \geq \Omega(b^{2p-2})$. The competitive ratio is thus lower bounded by $\Omega(b^{p-2}/p)$. □

Theorem 24. *For every $k, d \geq 2$, there exists a set of points in d dimensions for which any k-explainable clustering has competitive ratio $\Omega(k^{1-2/d}(\log k)^{-5/3} \log\log(2k))$ for the k-means cost.*

Proof. We can assume $k \geq 3$ w.l.o.g. because it is trivial to show a competitive ratio lower bound of 1.

We specify the integers b and p in Lemma 23 to get concrete lower bounds for the competitive ratio.

When $k^{1/d} \geq 3d$, we choose $p = d$ and $b = \lfloor k^{1/p} \rfloor$. It is clear that $b \geq 3p$, so

$$b^p = (b+1)^p/(1+1/b)^p \geq k/\exp(p/b) \geq \Omega(k).$$

This implies that $b^{p-2} = b^p/b^2 \geq \Omega(k^{1-2/d})$. Lemma 23 gives us a competitive ratio lower bound of

$$\Omega(b^{p-2}/p) \geq \Omega(k^{1-2/d}/d) \geq \Omega(k^{1-2/d}(\log k)^{-1} \log\log k),$$

where the last inequality is by Claim 30.

When $(\log k)^{2/3}/\log\log k \leq k^{1/d} < 3d$, we choose p to be the maximum integer such that $k^{1/p} \geq 3p$, and choose $b = \lfloor k^{1/p} \rfloor$. Again we have $b \geq 3p$, so $b^p \geq \Omega(k)$. We also have $b \leq k^{1/p} \leq (k^{1/(p+1)})^{1+1/p} \leq (p+1)^{1+1/p} \leq O(p) = O(\log k/\log\log k)$, where the last inequality is by Claim 30. This implies $b^{p-2}/p = b^p/(b^2 p) \geq \Omega(k(\log k)^{-3}(\log\log k)^3)$. The competitive ratio lower bound from Lemma 23 is

$$\Omega(b^{p-2}/p) \geq \Omega(k(\log k)^{-3}(\log\log k)^3) \geq \Omega(k^{1-2/d}(\log k)^{-5/3} \log\log k).$$

When $3 \leq k^{1/d} < (\log k)^{2/3}/\log\log k$, we choose $b = \lceil k^{1/d} \rceil$ and $p = \lfloor \log_b k \rfloor$. We have $b^p = b^{p+1}/b \geq k/b$, and $b \leq 2k^{1/d}$. Therefore, $b^{p-2} = b^{p+1}/b^3 \geq k/b^3 \geq k^{1-3/d}/8$. The competitive ratio lower bound from Lemma 23 is

$$\Omega(b^{p-2}/p) \geq \Omega(k^{1-3/d}/p) \geq \Omega(k^{1-3/d}/d) \geq \Omega(k^{1-3/d}/\log k) \geq \Omega(k^{1-2/d}(\log k)^{-5/3} \log\log k).$$

When $k^{1/d} < 3$, we choose $b = 3$ and $p = \lfloor \log_3 k \rfloor$. We have $b^{p-2} = b^{p+1}/b^3 \geq k/27$. The competitive ratio lower bound from Lemma 23 is

$$\Omega(b^{p-2}/p) \geq \Omega(k/p) \geq \Omega(k/\log k) \geq \Omega(k^{1-2/d}(\log k)^{-1}).$$ □

References

[AB18] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[ACKS15] Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximations of Euclidean k-means. In *31st International Symposium on Computational Geometry*, volume 34 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages 754–767. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2015.

[AMDIVW19] David Alvarez-Melis, Hal Daumé III, Jennifer Wortman Vaughan, and Hanna Wallach. Weight of evidence as a basis for human-oriented explanations. *arXiv preprint arXiv:1910.13503*, 2019.

[ANFSW17] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and Euclidean k-median by primal-dual algorithms. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 61–72. IEEE Computer Soc., Los Alamitos, CA, 2017.

[ARR99] Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for Euclidean k-medians and related problems. In *STOC '98 (Dallas, TX)*, pages 106–113. ACM, New York, 1999.

[BBCA+19] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss lemma. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1039–1050. ACM, New York, 2019.

[BCFN19] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[BIO+19] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019.

[BK05] Jayanta Basak and Raghu Krishnapuram. Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE transactions on knowledge and data engineering*, 17(1):121–132, 2005.

[BOW18] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering via optimal trees. *arXiv preprint arXiv:1812.00539*, 2018.

[BOW21] Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering: an optimization approach. *Mach. Learn.*, 110(1):89–138, 2021.

[CAKM16] Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in Euclidean and minor-free metrics. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 353–364. IEEE Computer Soc., Los Alamitos, CA, 2016.

[CCH+16] Junxiang Chen, Yale Chang, Brian Hobbs, Peter Castaldi, Michael Cho, Edwin Silverman, and Jennifer Dy. Interpretable clustering via discriminative rectangle mixture model. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 823–828. IEEE, 2016.

[CFG+02] Moses Charikar, Ronald Fagin, Venkatesan Guruswami, Jon Kleinberg, Prabhakar Raghavan, and Amit Sahai. Query strategies for priced information. volume 64, pages 785–819. 2002. Special issue on STOC 2000 (Portland, OR).

[CJ02] Jae-Woo Chang and Du-Seok Jin. A new cell-based clustering method for large, high-dimensional data in data mining applications. In *Proceedings of the 2002 ACM symposium on Applied computing*, pages 503–507, 2002.

[DC21] Maryam Negahbani Deeparnab Chakrabarty. Better algorithms for individually fair k-clustering. *arXiv preprint arXiv:2106.12150*, 2021.

[DF19] Daniel Deutch and Nave Frost. Constraints-based explanations of classifications. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 530–541. IEEE, 2019.

[DRB97] Luc De Raedt and Hendrik Blockeel. Using logical decision trees for clustering. In *International Conference on Inductive Logic Programming*, pages 133–140. Springer, 1997.

[FGS13] Ricardo Fraiman, Badih Ghattas, and Marcela Svarc. Interpretable clustering using unsupervised binary trees. *Adv. Data Anal. Classif.*, 7(2):125–145, 2013.

[FMR20] Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. ExKMC: Expanding explainable k-means clustering. *arXiv preprint arXiv:2006.02399*, 2020.

[FRS16] Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 365–374. IEEE Computer Soc., Los Alamitos, CA, 2016.

[GL20] Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1296. PMLR, 2020.

[GMB17] Badih Ghattas, Pierre Michel, and Laurent Boyer. Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. *Pattern Recognition*, 67:177–185, 2017.

[HJV19] Lingxiao Huang, Shaofeng Jiang, and Nisheeth Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[HMS66] Earl B Hunt, Janet Marin, and Philip J Stone. Experiments in induction. 1966.

[JKL20] Christopher Jung, Sampath Kannan, and Neil Lutz. Service in your neighborhood: Fairness in center location. *Foundations of Responsible Computing (FORC)*, 2020.

[JMS02] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pages 731–740. ACM, New York, 2002.

[KAM19] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pages 3448–3457. PMLR, 2019.

[KEM+19] Jacob Kauffmann, Malte Esders, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. From clustering to cluster explanations via neural networks. *arXiv preprint arXiv:1906.07633*, 2019.

[KMN+04] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

[KR99] Stavros G. Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the Euclidean k-median problem. In *Algorithms—ESA '99 (Prague)*, volume 1643 of *Lecture Notes in Comput. Sci.*, pages 378–389. Springer, Berlin, 1999.

[Lip18] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

[LM21] Eduardo Laber and Lucas Murtinho. On the price of explainability for some clustering problems. *arXiv preprint arXiv:2101.01576*, 2021.

[LXY05] Bing Liu, Yiyuan Xia, and Philip S Yu. Clustering via decision tree construction. In *Foundations and advances in data mining*, pages 97–124. Springer, 2005.

[MDRF20] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7055–7065. PMLR, 13–18 Jul 2020.

[MMR19] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1027–1038. ACM, New York, 2019.

[Mol20] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[MS21] Konstantin Makarychev and Liren Shan. Near-optimal algorithms for explainable k-medians and k-means. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[MSK+19] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[MV20] Sepideh Mahabadi and Ali Vakilian. Individual fairness for k-clustering. In *International Conference on Machine Learning*, pages 6586–6596. PMLR, 2020.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[Sey95] P. D. Seymour. Packing directed circuits fractionally. *Combinatorica*, 15(2):281–288, 1995.

[SF20] Kacper Sokol and Peter Flach. Limetree: Interactively customisable explanations based on local surrogate multi-output regression trees. *arXiv preprint arXiv:2005.01427*, 2020.

[SGZ20] Sandhya Saisubramanian, Sainyam Galhotra, and Shlomo Zilberstein. Balancing the tradeoff between clustering value and interpretability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 351–357, 2020.

[SSS19] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.

[YM10] Yasser Yasami and Saadat Pour Mozaffari. A novel unsupervised classification approach for network anomaly detection by k-means clustering and id3 decision tree learning methods. *The Journal of Supercomputing*, 53(1):231–245, 2010.

A Helper lemmas and claims

Claim 25. *Define $\alpha(z) = z^{1+2u}$ for $u \in (0, 1/2)$. Suppose non-negative real numbers z_1, z_2, z, Δ satisfy $z_1 + z_2 \le z + u\Delta$, and $\Delta \le \min\{z_1, z_2\}$. Then $\alpha(z_1) + \alpha(z_2) \le \alpha(z)$.*

Proof. Assume w.l.o.g. $z_1 \le z_2$. By the monotonicity and convexity of α, we have

$$\alpha(z_1) + \alpha(z_2) \le \alpha(z_1) + \alpha(z + u\Delta - z_1) \le \alpha(\Delta) + \alpha(z + u\Delta - \Delta).$$

We only need to prove that

$$\Delta^{1+2u} + (z - (1-u)\Delta)^{1+2u} \le z^{1+2u}. \tag{35}$$

Note that $0 \le \Delta \le z/(2-u)$ because $2\Delta \le z_1 + z_2 \le z + u\Delta$. Since the left-hand-side is convex in Δ, we just need to check (35) when $\Delta = 0$ and $\Delta = z/(2-u)$. It reduces to checking $1 + 2u \ge \frac{\log 2}{\log(2-u)}$. This holds by a standard calculation using our assumption $u \in (0, 1/2)$. □

Claim 26. *If $0 \le u \le v \le 3M/4$, then $u \log(M/u) \le 2v \log(M/v)$.*

Proof. The function $\alpha(z) = z \log(M/z)$ is non-decreasing when $z \le M/e$, and non-increasing when $z \ge M/e$. We just need to check the claim for $u = M/e$ and $v = 3M/4$, which follows from a standard calculation. □

Claim 27. *Let $I_1, \ldots, I_n \in \mathbb{R}$ be intervals. There exists $S \subseteq [n]$ such that $I_u \cap I_v = \emptyset$ for all distinct $u, v \in S$ and*

$$|\bigcup_{u \in [n]} I_u| \le 3|\bigcup_{v \in S} I_v|.$$

Proof. Assume w.l.o.g. that $|I_1| \ge \cdots \ge |I_n|$. For $u = 1, \ldots, n$, construct $S_u \subseteq [u]$ inductively as follows. We set $S_1 = \{1\}$, and for every $u > 1$, we set $S_u = S_{u-1}$ if there exists $v \in S_{u-1}$ such that $I_u \cap I_v \ne \emptyset$, and set $S_u = S_{u-1} \cup \{u\}$ otherwise. Define $S = S_n$. It is clear that $I_u \cap I_v = \emptyset$ for all distinct $u, v \in S$.

For every $u \in [n] \backslash S$, there exists $\alpha(u) \in S$ such that $\alpha(u) < u$ and $I_{\alpha(u)} \cap I_u \ne \emptyset$. We define $\alpha(u) = u$ when $u \in S$. For every $v \in S$ and $u \in \alpha^{-1}(v) \backslash \{v\}$, we have $I_v \cap I_u \ne \emptyset$ and $|I_u| \le |I_v|$. Therefore, $|\bigcup_{u \in \alpha^{-1}(v)} I_u| \le 3|I_v|$. We have

$$|\bigcup_{u \in [n]} I_u| = |\bigcup_{v \in S} \bigcup_{u \in \alpha^{-1}(v)} I_u| \le \sum_{v \in S} |\bigcup_{u \in \alpha^{-1}(v)} I_u| \le \sum_{v \in S} 3|I_v| = 3|\bigcup_{v \in S} I_v|.$$ □

Lemma 28 ([Sey95])**.** *Let m, M be positive real numbers with $M \ge 2m$. Suppose $U : (0, L) \to [m, M/2]$ is a non-decreasing function, and there is a finite set $I \subseteq (0, L)$ such that U is differentiable over $(0, L) \backslash I$. Then, there exists $z \in (0, L) \backslash I$ such that $U'(z) \le (1/L) U(z) \log \frac{M}{U(z)} \log \log_2 \frac{M}{m}$.*

Proof. Consider the function $\alpha(z) := -\log \log_2 \frac{M}{U(z)}$. It is clear that $\alpha(z)$ is non-decreasing and differentiable over $(0, L) \backslash I$. The function has derivative

$$\alpha'(z) = \frac{U'(z)}{U(Z) \log(M/U(z))}.$$

Suppose there does not exist such $z \in (0, L) \backslash I$ that makes the desired inequality hold. This implies that the derivative of $\alpha(z)$ exceeds $(1/L) \log \log_2 \frac{M}{m}$ whenever $z \in (0, L) \backslash I$. This contradicts with the fact that $\alpha(z)$ is bounded between $-\log \log_2(M/m)$ and 0. □

Lemma 29 ([Sey95])**.** *Let m, M be positive real numbers with $M \geq 2m$. Suppose $V : (0, L) \to [m, M - m]$ is a non-decreasing function, and there is a finite set $I \subseteq (0, L)$ such that V is differentiable over $(0, L)\backslash I$. Then, there exists $z \in (0, L)\backslash I$ such that $V'(z) \leq (2/L)M' \log \frac{M}{M'} \log \log_2 \frac{M}{m}$, where $M' = \min\{V(z), M - V(z)\}$.*

Proof. When $V(L/2) \leq M/2$, the lemma follows from Lemma 28 by setting $U(z) = V(z/2)$. When $V(L/2) \geq M/2$, the lemma follows from Lemma 28 by setting $U(z) = M - V(L - z/2)$. □

Claim 30. *For real numbers $k \geq 3$ and $p > 0$, if $k^{1/p} \geq p$, then $p \leq 2 \log k / \log \log k$.*

Proof. Assume for the sake of contradiction that $p > 2 \log k / \log \log k$, then $k^{1/p} < (\log k)^{1/2}$. Therefore,

$$p/k^{1/p} > 2(\log k)^{1/2} / \log \log k = (\log k)^{1/2} / \log((\log k)^{1/2}) \geq 1,$$

a contradiction. □