A. Details for Section 4

Proof of Claim 4.4. Let τ_e be the first time $t \geq 1$ when element e does not belong to E_t. Then, $\tau = \max_{e \in E} \tau_e$. Hence,

$$\sum_{t=0}^{\tau-1} p_t = \max_{e \in E} \sum_{t=0}^{\tau_e - 1} p_t.$$

By the union bound, for all $\lambda \geq 0$, we have

$$\mathbb{P}\left[\sum_{t=0}^{\tau-1} p_t \geq \lambda\right] \leq \sum_{e \in E} \mathbb{P}\left[\sum_{t=0}^{\tau_e-1} p_t \geq \lambda\right]. \qquad (12)$$

Define a new stochastic process $Z_t(e)$ as follows: $Z_0(e) = 1$ and for $t \geq 1$,

$$Z_t(e) = \begin{cases} e^{\sum_{t'=0}^{t-1} p_{t'}}, & \text{if } e \in E_t; \\ 0, & \text{otherwise.} \end{cases}$$

Note that if $\sum_{t=0}^{\tau_e - 1} p_t \geq \lambda$, then $\max_{t \geq 0} Z_t(e) \geq e^{\lambda - 1}$. Thus, we will bound $\Pr[\max_{t \geq 0} Z_t(e) \geq e^{\lambda-1}]$. Observe that Z_t is a supermartingale, since

$$\begin{aligned} \mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] &= \Pr[e \in E_{t+1} \mid \mathcal{F}_t] \cdot e^{p_t} \cdot Z_t \\ &\leq (1 - p_t) \cdot e^{p_t} \cdot Z_t \leq Z_t. \end{aligned}$$

By Doob's maximal martingale inequality, we have

$$\Pr[\max_{t \geq 0} Z_t(e) \geq e^{\lambda - 1}] \leq Z_0(e)/e^{\lambda-1} = e^{-(\lambda-1)}.$$

Using (12), we get

$$\mathbb{P}\left[\sum_{t=0}^{\tau-1} p_t \geq \lambda\right] \leq |E| \cdot e^{-(\lambda-1)}.$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\sum_{t=0}^{\tau-1} p_t\right] &= \int_0^\infty \mathbb{P}\left[\sum_{t=0}^{\tau-1} p_t \geq \lambda\right] d\lambda \\ &\leq \ln|E| + \int_{\ln|E|}^\infty |E| \cdot e^{-(\lambda-1)} d\lambda \\ &= \ln|E| + |E| e^{-\ln|E|+1} = \ln|E| + e. \end{aligned}$$

□

B. Terminal Embedding

Lemma B.1. *For every finite set of real numbers K, the function ψ_K defined in Lemma 6.1 is a cut preserving embedding that satisfies for every $x \in \mathbb{R}$ and $y \in K$*

$$\begin{aligned} |\psi_K(x) - \psi_K(y)| &\leq |x - y|^2 \\ &\leq 8|K| \cdot |\psi_K(x) - \psi_K(y)|. \end{aligned}$$

Proof. We first show that this function ψ_K is continuous and differentiable in $\mathbb{R}$. Consider $2k$ open intervals on the real line divided by points in K and points $(y_i + y_{i+1})/2$ for $i \in \{1, 2, \cdots, k-1\}$. In every such open interval, the function ψ_K is a quadratic function, which is continuous and differentiable. Since ψ_K is also continuous and differentiable at the endpoints of these intervals, the function ψ_K is continuous and differentiable in $\mathbb{R}$. For any $x \in \mathbb{R}$, we have $\psi'_K(x) = 2\,|x - y^*| \geq 0$ where y^* is the closest point in K to x. Thus, the function ψ_K is increasing in $\mathbb{R}$, which implies ψ_K is cut preserving.

We now prove that ψ_K satisfies two inequalities. We first show that for every $x \in \mathbb{R}$ and $y \in K$, $|\psi_K(x) - \psi_K(y)| \leq |x - y|^2$. Suppose that $x \geq y$ (The case $x \leq y$ is handled similarly.) If $x = y$, then this inequality clearly holds. Thus, to prove $|\psi_K(x) - \psi_K(y)| \leq |x - y|^2$, it is sufficient to prove the following inequality on derivatives

$$(\psi_K(x) - \psi_K(y))'_x \leq \left((x-y)^2\right)'_x.$$

Let y^* be the closest point in K to x. Then,

$$\begin{aligned} &(\psi_K(x) - \psi_K(y))'_x = (\psi_K(x))'_x = \\ &\qquad = (\psi_K(y^*) + \varepsilon_x (x - y^*)^2)'_x = 2|x - y^*|. \end{aligned}$$

Since y^* is the closest point in K to x, we have $|x - y^*| \leq |x - y| = \left((x-y)^2\right)'_x/2$. This finishes the proof of the first inequality.

We now verify the second inequality. First, consider two points y_i and y_j ($y_i < y_j$). Write,

$$\psi_K(y_j) - \psi_K(y_i) = z_j - z_i = \frac{1}{2} \sum_{m=i}^{j-1} (y_{m+1} - y_m)^2.$$

By the arithmetic mean–quadratic mean inequality, we have

$$(j-i) \cdot \sum_{m=i}^{j-1} (y_{m+1} - y_m)^2 \geq \Big(\sum_{m=i}^{j-1} y_{m+1} - y_m\Big)^2 = (y_j - y_i)^2.$$

Thus,

$$\psi_K(y_j) - \psi_K(y_i) \geq \frac{(y_j - y_i)^2}{2(j-i)} \geq \frac{(y_j - y_i)^2}{2(k-1)}.$$

Now we consider the case when x is an arbitrary real number in $\mathbb{R}$ and $y \in K$. Let y^* be the closest point in K to x. Then,

$$|x - y|^2 \leq 2|x - y^*|^2 + 2|y^* - y|^2.$$

The first term on the right hand side equals $4|\psi_K(x) - \psi_K(y^*)|$; the second term is upper bounded by $4(k-1)|\psi_K(y) - \psi_K(y^*)|$. Thus,

$$|x - y|^2 \leq 4|\psi_K(x) - \psi_K(y^*)| + 4(k-1)|\psi_K(y^*) - \psi_K(y)|.$$

Note that $|\psi_K(x) - \psi_K(y^*)| \le |\psi_K(x) - \psi_K(y)|$ since y^* is the closest point in K to x. Also, we have

$$\begin{aligned}|\psi_K(y^*) - \psi_K(y)| &\le \\ &\le |\psi_K(x) - \psi_K(y^*)| + |\psi_K(x) - \psi_K(y)| \\ &\le 2|\psi_K(x) - \psi_K(y)|.\end{aligned}$$

Hence,

$$|x - y|^2 \le 8k|\psi_K(x) - \psi_K(y)|.$$

This completes the proof. □

Lemma B.2. *The terminal embedding ψ is coordinate cut preserving. For every threshold cut (i, θ), there exists a threshold cut (i, θ') such that*

$$\{x \in \mathbb{R}^d : x_i \le \theta'\} = \{x \in \mathbb{R}^d : \psi(x)_i \le \theta\}.$$

Proof. By the construction of φ, we have for any threshold cut (i, θ)

$$\{x \in \mathbb{R}^d : \psi(x)_i \le \theta\} = \{x \in \mathbb{R}^d : \psi_i(x_i) \le \theta\}.$$

Since ψ_i is a cut preserving terminal embedding by Lemma 6.2, there exists a threshold $\theta' \in \mathbb{R}$ such that

$$\{x \in \mathbb{R}^d : x_i \le \theta'\} = \{x \in \mathbb{R}^d : \psi_i(x_i) \le \theta\}.$$

□

C. k-medians in ℓ_2

In this section, we present an algorithm for the k-medians in ℓ_2 and show that it provides an explainable clustering with cost at most $O(\log^{3/2} k)$ times the original cost.

C.1. Algorithm for k-medians in ℓ_2

Our algorithm builds a binary threshold tree T using a top-down approach, as shown in Algorithm 2. It starts with a tree containing only the root node r. The root r is assigned the set of points X_r that contains all points in the data set X and all reference centers c^i. Then, the algorithm calls function BUILD_TREE(r). Function BUILD_TREE(u) partitions centers in u in several groups X_v using function PARTITION_LEAF(u) and then recursively calls itself (BUILD_TREE(v)) for every new group X_v that contains more than one reference center c^i.

Most work is done in the function PARTITION_LEAF(u). The argument of the function is a leaf node u of the tree. We denote the set of data points and centers assigned to u by X_u. Function PARTITION_LEAF(u) partitions the set of centers assigned to node u into several groups. Each group contains at most half of all centers c^i from the set X_u. When PARTITION_LEAF(u) is called, the algorithm finds the ℓ_1-median of all reference centers in node u. Denote this point by m^u. We remind the reader that the i-th coordinate of the median m^u (which we denote by m^u_i) is a median for i-th coordinates of centers in X_u. That is, for each coordinate i, both sets $\{c \in X_u \cap C : c_i < m^u_i\}$ and $\{c \in X_u \cap C : c_i > m^u_i\}$ contain at most half of all centers in X_u. Then, function PARTITION_LEAF(u) iteratively partitions X_u into pieces until each piece contains at most half of all centers from X_u. We call the piece that contains the median m^u the main part (note that we find the median m^u when PARTITION_LEAF(u) is called and do not update m^u afterwards).

At every iteration t, the algorithm finds the maximum distance R^u_t from centers in the main part to the point m^u. The algorithm picks a random coordinate $i^u_t \in \{1, 2, \cdots, d\}$, random number $\theta^u_t \in [0, (R^u_t)^2]$, and random sign $\sigma^u_t \in \{\pm 1\}$ uniformly. Then, it splits the main part using the threshold cut $(i^u_t, m^u_i + \sigma^u_t\sqrt{\theta^u_t})$ if this cut separates at least two centers in the main part. Function PARTITION_LEAF(u) stops, when the main part contain at most half of all centers in X_u. Note that all pieces separated from m^u during the execution of PARTITION_LEAF(u) contain at most half of all centers in X_u because m^u is the median of all centers in X_u.

Theorem C.1. *Given a set of points X in $\mathbb{R}^d$ and a set of centers $C = \{c^1, \dots, c^k\} \subset \mathbb{R}^d$, Algorithm 2 finds a threshold tree T with expected k-medians in ℓ_2 cost at most*

$$\mathbb{E}[\text{cost}_{\ell_2}(X, T)] \le O(\log^{3/2} k) \cdot \text{cost}_{\ell_2}(X, C).$$

Proof. Let $T_t(u)$ be the threshold tree at the beginning of iteration t in function PARTITION_LEAF(u). For every point $x \in X_u$, define its cost at step t of function PARTITION_LEAF(u) to be the distance from x to the closest center in the same leaf of $T_t(u)$ as x. That is, if x belongs to a leaf node v in the threshold tree $T_t(u)$, then

$$\text{cost}_{\ell_2}(x, T_t(u)) = \min\{\|x - c\|_2 : c \in X_v \cap C\}.$$

If the point x is separated from its original center in C by the cut generated at time step t, then x will be eventually assigned to some other center in the main part of $T_t(u)$. By the triangle inequality, the new cost of x at the end of the algorithm will be at most $\text{cost}_{\ell_2}(x, C) + 2R^u_t$, where R^u_t is the maximum radius of the main part in $T_t(u)$ i.e., R^u_t is the distance from the median m^u to the farthest center c^i in the main part. Define a penalty function $\phi^u_t(x)$ as follows: $\phi^u_t(x) = 2R^u_t$ if x is separated from its original center c at time t; $\phi^u_t(x) = 0$, otherwise. Let U_x be the set of all nodes u for which the algorithm calls BUILD_TREE(u) and $x \in X_u$. Note that some nodes v of the threshold tree with $x \in X_v$ do not belong to U_x. Such nodes v are created and split into two groups in the same call of PARTITION_LEAF(u). Observe that $\phi^u_t(x) \ne 0$ for at most one step t in the call of PARTITION_LEAF(u) for some node

Algorithm 2 Threshold tree construction for k-medians in L_2

```
Input: a data set $X \subset \mathbb{R}^d$, centers $C = \{c_1, c_2, \ldots, c_k\} \subset \mathbb{R}^d$
Output: a threshold tree $T$

function MAIN$(X, C)$
    Create a root $r$ of the threshold tree $T$ containing $X_r = X \cup C$.
    BUILD_TREE$(r)$.
end function

function PARTITION_LEAF$(u)$
    Compute the $\ell_1$ median $m^u$ of all centers in $X_u$.
    Set the main part $u_0 = u$ and set $t = 0$.
    while node $u_0$ contains more than $1/2$ of centers in $X_u$ do
        Update $t = t + 1$.
        Let $R_t^u = \max_{c \in X_{u_0}} \|c\|_2$.
        Sample $i_t^u \in \{1, 2, \cdots, d\}$, $\theta_t^u \in [0, (R_t^u)^2]$, and $\sigma_t^u \in \{\pm 1\}$ uniformly at random.
        if two centers in $X_{u_0}$ are separated by $(i_t^u, m_i^u + \sigma_t^u \sqrt{\theta_t^u})$ then
            Assign to $u_0$ two children $u_\leq = \{x \in X_{u_0} : x_i \leq \vartheta\}$ and $u_> = \{x \in X_{u_0} : x_i > \vartheta\}$ where $i = i_t^u, \vartheta = m_i^u + \sigma_t^u \theta_t^u$.
            Update the main part $u_0$ be $u_\leq$ if $\sigma_t^u = 1$, and be $u_>$ otherwise (thus, the main part always contains $m^u$).
        end if
    end while
end function

function BUILD_TREE$(u)$
    Call PARTITION_LEAF$(u)$.
    Call BUILD_TREE$(v)$ for each leaf $v$ in the subtree of $u$ containing more than one center.
end function
```

$u \in U_x$, and

$$\mathrm{cost}_{\ell_2}(x, T) \leq \mathrm{cost}_{\ell_2}(x, C) + \sum_{u \in U_x} \sum_t \phi_t^u(x). \qquad (13)$$

The sum in the right hand side is over all iterations t in all calls of function PARTITION_LEAF(u) with $u \in U_x$. Since each piece in the partition returned by function PARTITION_LEAF(u) contains at most half of all centers from X_u, the depth of the recursion tree is at most $O(\log k)$ (note that the depth of the threshold tree can be as larger as $k - 1$). This means that the size of U_x is at most $O(\log k)$. In Lemma C.3, we show that the expected total penalty in the call of PARTITION_LEAF(u) for every $u \in U_x$ is at most $O(\sqrt{\log k})$ times the original cost. Before that, we upper bound the expected penalty $\phi_t^u(x)$ for each step t in the call of PARTITION_LEAF(u) for every node $u \in U_x$.

Lemma C.2. *The expected penalty $\phi_t^u(x)$ is upper bounded as follows:*

$$\mathbb{E}[\phi_t^u(x)] \leq \mathbb{E}\left[2\|x - c\|_2 \cdot \frac{\|c - m^u\|_2 + \|x - m^u\|_2}{d \cdot R_t^u}\right],$$

where c is the closest center to the point x in C.

Proof. We first bound the probability that point x is separated from its original center c at iteration t. For any coordinate $i \in \{1, 2, \cdots, d\}$, let x_i and c_i be the i-th coordinates of point x and center c respectively. For any point $x \in \mathbb{R}^d$, we define the indicator function $\delta_x(i, \theta) = 0$ if $x_i \leq \theta$, and $\delta_x(i, \theta) = 1$ otherwise. To determine whether the threshold cut sampled at iteration t separates x and c, we consider the following two cases: (1) x and c are on the same side of the median m^u in coordinate i (i.e. $(x_i - m_i^u)(c_i - m_i^u) \geq 0$), and (2) x and c are on the opposite sides of the median m^u in coordinate i (i.e. $(x_i - m_i^u)(c_i - m_i^u) < 0$).

If x and c are on the same side of the median m^u in coordinate i, then the threshold cut $(i, m_i^u + \sigma_t^u \sqrt{\theta_t^u})$ separates x and c if and only if σ_t^u has the same sign as $x_i - m_i^u$ and θ_t^u is between $(x_i - m_i^u)^2$ and $(c_i - m_i^u)^2$. Thus,

$$\begin{aligned} &\mathbb{P}\left[\delta_x(i, \vartheta_t^u) \neq \delta_c(i, \vartheta_t^u) \mid T_t(u)\right] = \\ &\quad = \frac{\left|(c_i - m_i^u)^2 - (x_i - m_i^u)^2\right|}{2(R_t^u)^2} \leq \\ &\quad\quad \leq \frac{|c_i - x_i|\left(|c_i - m_i^u| + |x_i - m_i^u|\right)}{2(R_t^u)^2}, \end{aligned}$$

where $\vartheta_t^u = m_i^u + \sigma_t^u \sqrt{\theta_t^u}$.

Now, suppose x and c are on the opposite sides of the median m^u in coordinate i, i.e. $(x_i - m_i^u)(c_i - m_i^u) < 0$. The threshold cut $(i, m_i^u + \sigma_t^u\sqrt{\theta_t^u})$ separates x and c if and only if $\sigma_t^u(x_i - m_i^u) \geq 0$, $\theta_t^u \leq (x_i - m_i^u)^2$ or $\sigma_t^u(c_i - m_i^u) \geq 0$, $\theta_t^u \leq (c_i - m_i^u)^2$. Thus, we have for every coordinate i with $(x_i - m_i^u)(c_i - m_i^u) < 0$,

$$\begin{aligned}\mathbb{P}\left[\delta_x(i,\vartheta_t^u) \neq \delta_c(i,\vartheta_t^u) \mid T_t(u)\right] = \\ = \frac{(c_i - m_i^u)^2 + (x_i - m_i^u)^2}{2(R_t^u)^2} \leq \\ \leq \frac{|c_i - x_i|\left(|c_i - m_i^u| + |x_i - m_i^u|\right)}{2(R_t^u)^2},\end{aligned}$$

where the last inequality follows from $|c_i - x_i| \geq \max\{|c_i - m_i^u|, |x_i - m_i^u|\}$, since c_i, x_i are on the different sides of m_i^u.

Since the coordinate i_t^u is chosen randomly and uniformly from $\{1, \cdots d\}$, the probability that x and c are separated at iteration t is

$$\begin{aligned}\mathbb{P}[\delta_x(i_t^u,\vartheta_t^u) \neq \delta_c(i_t^u,\vartheta_t^u) \mid T_t(u)] \leq \\ \leq \sum_{i=1}^{d} \frac{|c_i - x_i|\left(|c_i - m_i^u| + |x_i - m_i^u|\right)}{2d \cdot (R_t^u)^2} \\ \leq \frac{\|c - x\|_2(\|x - m^u\|_2 + \|c - m^u\|_2)}{d \cdot (R_t^u)^2},\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and $(|c_i| + |x_i|)^2 \leq 2c_i^2 + 2x_i^2$.

Then, the expected penalty is

$$\begin{aligned}\mathbb{E}[\phi_t^u(x)] &\leq \mathbb{E}\left[\mathbb{P}\left[\delta_x(i_t^u,\vartheta_t^u) \neq \delta_c(i_t^u,\vartheta_t^u) \mid T_t(u)\right] \cdot 2R_t^u\right] \\ &\leq \mathbb{E}\left[2\|c - x\|_2 \cdot \frac{\|c - m^u\|_2 + \|x - m^u\|_2}{d \cdot R_t^u}\right].\end{aligned}$$

□

To bound the expected penalty for point x, we consider two types of cuts based on three parameters: the maximum radius R_t^u and distances $\|x - m^u\|_2$, $\|c - m^u\|_2$ between x, c and the median m^u . If x is separated from its original center c at iteration t with

$$R_t^u \leq \sqrt{\log_2 k} \cdot \max\{\|x - m^u\|_2, \|c - m^u\|_2\},$$

then we call this cut a light cut. Otherwise, we called it a heavy cut.

Lemma C.3. *In every call of* PARTITION_LEAF(u) *(see Algorithm 2), the expected penalty for a point $x \in X$ is upper bounded as follows:*

$$\mathbb{E}\left[\sum_t \phi_t^u(x)\right] \leq O(\sqrt{\log k}) \cdot \mathrm{cost}_{\ell_2}(x, C).$$

Proof. If point x is not separated from its original center c in PARTITION_LEAF(u), then the total penalty is 0. If x is separated from its center c in this call, then there are two cases: (1) the point x is separated by a light cut; (2) the point x is separated by a heavy cut. We first show that the expected penalty due to a heavy cut is at most $O(\sqrt{\log k})\mathrm{cost}_{\ell_2}(x, C)$.

Denote the set of all heavy cuts at iteration t in PARTITION_LEAF(u) by H_t^u:

$$H_t^u = \{x : \max\{\|x - m^u\|_2, \|c - m^u\|_2\} < R_t^u / \sqrt{\log_2 k}\}.$$

Then, by Lemma C.2, the expected penalty x incurs due to a heavy cut is at most

$$\begin{aligned}&\mathbb{E}\left[\sum_{t: x \in H_t^u} \phi_t^u(x)\right] \leq \\ &\leq 2\|x - c\|_2 \cdot \mathbb{E}\left[\sum_{t: x \in H_t^u} \frac{\|x - m^u\|_2 + \|c - m^u\|_2}{d \cdot R_t^u}\right].\end{aligned}$$

Since the maximum radius R_t^u is a non-increasing function of t, we split all steps of this call of PARTITION_LEAF into phases with exponentially decreasing values of R_t^u. At phase s, the maximum radius R_t^u is in the range $(R_1^u/2^{s+1}, R_1^u/2^s]$, where R_1^u is the maximum radius at the beginning of PARTITION_LEAF(u).

Consider an arbitrary phase s and step t in that phase. Let $R = R_1^u/2^s$. For every center c' with $\|c' - m^u\|_2 \in (R/2, R]$, the probability that this center c' is separated from the main part at step t in phase s is at least

$$\begin{aligned}&\mathbb{P}\left[\delta_{c'}(i_t^u,\vartheta_t^u) \neq \delta_{m^u}(i_t^u,\vartheta_t^u) \mid T_t(u)\right] = \\ &= \sum_{j=1}^{d} \frac{1}{d} \cdot \frac{(c'_j - m_j^u)^2}{2(R_t^u)^2} = \frac{\|c' - m^u\|_2^2}{2d \cdot (R_t^u)^2} \geq \frac{1}{4d},\end{aligned}$$

where the last inequality is due to $\|c' - m^u\|_2 > R/2 \geq R_t^u/2$ for step t in the phase s. Since there are at most k centers, all centers with norm in $(R/2, R]$ are separated from the main part in at most $4d \ln k$ steps in expectation. Thus, the expected length of each phase is $O(d \log k)$ steps, and hence, the expected penalty x incurred during phase s is at most

$$\begin{aligned}&2\|x - c\|_2 \cdot \mathbb{E}\Bigg[\sum_{\substack{t: x \in H_t^u \\ R_t^u \in (R/2, R]}} \frac{\|x - m^u\|_2 + \|c - m^u\|_2}{d \cdot R_t^u}\Bigg] \leq \\ &\leq 2\|x - c\|_2 \cdot \mathbb{E}\Bigg[\sum_{\substack{t: x \in H_t^u \\ R_t^u \in (R/2, R]}} \frac{\|x - m^u\|_2 + \|c - m^u\|_2}{d \cdot R/2}\Bigg] \leq \\ &\leq O(\log k) \cdot \|x - c\|_2 \cdot \frac{\|x - m^u\|_2 + \|c - m^u\|_2}{R}.\end{aligned}$$

Let s' be the last phase for which

$$R_1^u/2^{s'} \geq \sqrt{\log_2 k}\cdot \max\{\|x-m^u\|_2, \|c-m^u\|_2\}. \quad (14)$$

Then, in every phase $s > s'$, all cuts separating x from its original center c are light. Hence, the total expected penalty due to a heavy cut is upper bounded by

$$O(\log k)\cdot\|x-c\|_2\cdot(\|x-m^u\|_2+\|c-m^u\|_2)\cdot\sum_{s=0}^{s'}\frac{2^s}{R_1^u} =$$
$$= O(\log k)\cdot\|x-c\|_2\cdot(\|x-m^u\|_2+\|c-m^u\|_2)\cdot\frac{2^{s'+1}}{R_1^u}.$$

Using the definition (14) of s', we write

$$(\|x-m^u\|_2+\|c-m^u\|_2)\cdot\frac{2^{s'+1}}{R_1^u} \leq$$
$$\leq 2\frac{\|x-m^u\|_2+\|c-m^u\|_2}{R_1^u/2^{s'}} \leq \frac{4}{\sqrt{\log_2 k}}.$$

Thus, the expected penalty due to a heavy cut is at most $O(\sqrt{\log k})\mathrm{cost}_{\ell_2}(x, C)$.

We now analyze the expected penalty due to a light cut. Consider an iteration t in PARTITION_LEAF(u) with $x \notin H_t^u$. By the analysis in Lemma C.2, the probability that x and c are separated at iteration t is at most

$$\frac{\|c-x\|_2(\|x-m^u\|_2+\|c-m^u\|_2)}{d\cdot(R_t^u)^2}.$$

The probability that x or c is separated from the main part at iteration t is at least

$$\frac{\max\{\|x-m^u\|_2^2, \|c-m^u\|_2^2\}}{d(R_t^u)^2}.$$

If x or c is separated from the main part, then the point x will not incur penalty at any step after t. Thus, the probability that x and c are separated by a light cut in the end of PARTITION_LEAF(u) is at most

$$\frac{\|c-x\|_2(\|x-m^u\|_2+\|c-m^u\|_2)}{\max\{\|x-m^u\|_2^2, \|c-m^u\|_2^2\}} \leq$$
$$\leq \frac{2\|c-x\|_2}{\max\{\|x-m^u\|_2, \|c-m^u\|_2\}}.$$

Since the penalty of a light cut is at most $R_t^u \leq \sqrt{\log_2 k}\cdot\max\{\|x-m^u\|_2, \|c-m^u\|_2\}$, the expected penalty due to a light cut is at most $O(\sqrt{\log k})\cdot\mathrm{cost}_{\ell_2}(x, C)$.

This concludes the proof of Lemma C.3. □

For every node u, the main part contains the median m^u, which is also the ℓ_1-median of all centers in X_u. Thus, each cut sampled in the call PARTITION_LEAF(u) separates at most half of all centers in X_u from the origin. The main part contains at most half of centers in X_u at the end of the call PARTITION_LEAF(u). Therefore, each leaf node generated in the end of PARTITION_LEAF(u) contains at most half of centers in X_u. Thus, the depth of the recursion tree is at most $O(\log k)$. By Lemma C.3 and Equation (13), we get the conclusion. □

D. Lower Bound for Threshold Tree

D.1. Lower bound for k-means

In this section, we show a lower bound on the price of explainability for k-means.

Theorem D.1. *For any k, there exists an instance X with k clusters such that the cost of explainable k-means clustering for every tree T is at least*

$$\mathrm{cost}_{\ell_2^2}(X, T) \geq \Omega\left(\frac{k}{\log k}\right)\mathrm{OPT}_{\ell_2^2}(X).$$

To prove this lower bound, we construct an instance as follows. We uniformly sample k centers $C = \{c^1, c^2, \cdots, c^k\}$ from the d-dimensional unit cube $[0, 1]^d$ where the dimension $d = 300\ln k$. For each center c^i, we add two points $c^i \pm (\varepsilon, \varepsilon, \cdots, \varepsilon)$ with $\varepsilon = 300\ln k/k$. We also add many points at each center such that the optimal centers for any threshold tree remain almost the same. Specially, we can add k^2 points co-located with each center c^i. Then, if one center c^i is shifted by a distance of ε in the threshold tree clustering, the cost of the co-located points at c^i is at least $k^2\varepsilon^2$. Since the optimal regular cost for this instance is $kd\varepsilon^2$, the total cost of the threshold tree is lower bounded by $\Omega(k/\log k)\mathrm{OPT}_{\ell_2^2}(X)$. Consequently, we consider the threshold tree with optimal centers shifted by at most ε.

First, we show that any two centers defined above are far apart with high probability.

Lemma D.2. *With probability at least $1 - 1/k^2$ the following holds: The squared distance between every two distinct centers c and c' in C is at least $d/12$.*

Proof. Consider any fixed two centers $c, c' \in C$. Since c, c' are uniformly sampled from $[0, 1]^d$, each coordinate of c, c' is sampled from $[0, 1]$; and centers c, c' are sampled independently. Thus, we have

$$\mathbb{E}_{c,c'}[\|c-c'\|^2] = \sum_{i=1}^{d}\mathbb{E}_{c_i,c_i'}[(c_i-c_i')^2] = \frac{d}{6}.$$

We use a random variable X_i to denote $(c_i - c_i')^2$ for each coordinate $i \in \{1, \ldots, d\}$. Since random variables $\{X_i\}_{i=1}^d$ are independent, by Hoeffding's inequality, we have

$$\mathbb{P}\left[\sum_{i=1}^{d} X_i - \mathbb{E}\left[\sum_{i=1}^{d} X_i\right] \le -\sqrt{2d\ln k}\right] \le$$
$$\le e^{-4\ln k} = \frac{1}{k^4},$$

where we used that $d = 300\ln k$. This implies that the squared distance between c and c' is less than $d/12$ with probability at most $1/k^4$. Using the union bound over all pairs of centers in C, we conclude that the squared distance between all pairs in C is at least $d/12$ with probability at least $1 - 1/k^2$. □

If any two centers are far apart, then a point x separated from its original center will incur a large penalty. Thus, we can get a lower bound if there exists an instance which satisfies: (1) any two centers are separated by a large distance; (2) every threshold tree separates a relatively large portion of points from their original centers. In particular, we prove that with probability $1 - o(1)$, every threshold cut separates a relatively large portion of points from their original centers in the random instance we constructed.

Lemma D.3. *With probability at least $1 - 1/k^2$, the following holds: every threshold cut (i, θ) with $i \in \{1, 2, \cdots, d\}$ and $\theta \in [0, 1)$ separates at least $\varepsilon k/4$ points from their original centers.*

Proof. Consider a fixed coordinate $i \in \{1, \ldots, d\}$. We project each center and its rectangular neighborhood onto this coordinate. For each center $c^j \in C$, we define an interval I_i^j as the intersection of $[0, 1]$ and the ε-neighborhood of its projection c_i^j, i.e. $I_i^j = (c_i^j - \varepsilon, c_i^j + \varepsilon) \cap [0, 1]$. Each interval I_i^j has length at least ε. If we pick a threshold cut inside any interval I_i^j, then we separate at least one points from center c^j. In this case, the interval I_i^j is called covered by this threshold cut. Then, we give the lower bound on the minimum number of intervals covered by a threshold cut.

For a fixed set of centers C, we consider at most $2k$ special positions for the threshold cut at coordinate i as follows. Let E_i be the set containing two end points of intervals I_i^j for all centers c^j. For any threshold cut at coordinate i, the closest position in set E_i covers exactly the same set of intervals as this threshold cut. Thus, we only need to consider threshold cuts at positions in E_i.

For centers chosen uniformly from $[0, 1]^d$, the set E_i contains $2k$ random variables. Suppose we pick a threshold cut at a position θ in E_i related to interval I_i^j. Conditioned on the position θ, the other $k - 1$ centers c^j for $j \ne j^*$ are uniformly distributed in $[0, 1]^d$ since all centers are chosen independently. For $j \in \{1, 2, \cdots, k\} \setminus \{j^*\}$, let Y_i^j be the indicator random variable that the interval I_i^j contains this position θ. For each variable Y_i^j, we have $\varepsilon \le \mathbb{P}\left[Y_i^j = 1\right] \le 2\varepsilon$. Since random variables Y_i^j are independent, by the Chernoff bound for Bernoulli random variables, we have

$$\mathbb{P}\left[\sum_j Y_i^j - \mathbb{E}\left[\sum_j Y_i^j\right] \le -\sqrt{18\varepsilon k\ln k} \mid \theta\right] \le$$
$$\le e^{-4\ln k} = \frac{1}{k^4}.$$

Thus, we have the number of intervals containing this position θ is at least $\varepsilon k/4$ with probability at least $1 - 1/k^4$.

Since we have $2k$ positions E_i for each coordinate $i \in \{1, 2, \cdots, d\}$, there are total $2dk$ positions for threshold cuts. Using the union bound over all positions, we have the minimum number of intervals covered by a threshold cut is at least $\varepsilon k/4$ with probability at least $1 - 1/k^2$. Since the threshold cut separates one point from its original center for each covered interval, we have every threshold cut separates at least $\varepsilon k/4$ points from their original centers in this case. □

Proof of Theorem D.1. By Lemma D.2, we can only consider the instance where any two centers are separated with the squared distance at least $d/12$. Note that the optimal centers for any threshold tree remain almost the same as centers C. Thus, we analyze the k-means cost given by any threshold tree with respect to center C. If a point in X is separated from its original center, this point will finally be assigned to another center in C. By the triangle inequality, the k-means cost of this point is at least $d/20$. By Lemma D.3, there exists an instance such that any threshold cut separates at least $\varepsilon k/4$ points from their original centers. Thus, there exists an instance X such that any threshold tree T has the k-means cost at least

$$\mathrm{cost}_{\ell_2^2}(X, T) \ge \frac{\varepsilon k}{4} \cdot \frac{d}{20} = \frac{\varepsilon k d}{80}.$$

Note that the optimal regular k-means cost for this instance X is

$$\mathrm{OPT}_{\ell_2^2}(X) = 2k \cdot \varepsilon^2 d.$$

Therefore, the k-means cost for this instance X given by any threshold tree T is at least

$$\begin{aligned}\mathrm{cost}_{\ell_2^2}(X, T) &\ge \frac{1}{160\varepsilon} \cdot \mathrm{OPT}_{\ell_2^2}(X) \\ &= \Omega\left(\frac{k}{\log k}\right) \cdot \mathrm{OPT}_{\ell_2^2}(X).\end{aligned}$$

□

D.2. Lower bound for k-medians in ℓ_2

In this section, we show a lower bound on the price of explainability for k-medians in ℓ_2.

Theorem D.4. *For every $k \geq 1$, there exists an instance X with k clusters such that the k-medians with ℓ_2 objective cost of every threshold tree T is at least*

$$\mathrm{cost}_{\ell_2}(X, T) \geq \Omega(\log k)\mathrm{OPT}_{\ell_2}(X).$$

To prove this lower bound, we use the construction similar to that used in Theorem D.1. We discretize the d-dimensional unit cube $[0,1]^d$ into grid with length $\varepsilon = 1/\lceil \ln k \rceil$, where the dimension $d = 300 \ln k$. We uniformly sample k centers $C = \{c^1, c^2, \cdots, c^k\}$ from the above grid $\{0, \varepsilon, 2\varepsilon, \cdots, 1\}^d$. For each center c^i, we add 2 points $c^i \pm (\varepsilon, \varepsilon, \cdots, \varepsilon)$ to this center. Similar to Theorem D.1, we also add many points at each center such that the optimal centers for any threshold tree remain almost the same.

Similar to Lemma D.2, we show that any two centers defined above are far apart with high probability.

Lemma D.5. *With probability at least $1 - 1/k^2$ the following holds: The distance between every two distinct centers c and c' in C is at least $\sqrt{d}/4$.*

Proof. To sample a center from the grid uniformly, we can first sample a candidate center uniformly from the cube $[-\varepsilon/2, 1 + \varepsilon/2]^d$ and then move it to the closest grid point. Note that the ℓ_2-distance from every point in this cube to its closest grid point is at most $\varepsilon\sqrt{d} = o(1)$. By Lemma D.2, the ℓ_2 distance between every pairs of candidate centers is at least $\sqrt{d/12}$ with probability at least $1 - 1/k^2$. Thus, the distance between every two distinct centers is at least $\sqrt{d}/4$ with probability at least $1 - 1/k^2$. □

For every node in the threshold tree, we can specify it by threshold cuts in the path from the root to this node. Thus, we define a path π as an ordered set of tuples $(i_j, \theta_j, \sigma_j)$, where (i_j, θ_j) denotes the j-th threshold cut in this path and $\sigma_j \in \{\pm 1\}$ denotes the direction with respect to this cut. We use $u(\pi)$ be the node specified by the path π. We define a center is damaged if one of its two points are separated by this cut, otherwise a center is undamaged. Let F_u be the set of undamaged centers in node u.

Lemma D.6. *With probability at least $1 - 1/k$, the following holds: For every path π with length less than $\log_2 k/4$, we have (a) the node $u(\pi)$ contains at most $\sqrt{k}$ undamaged centers; or (b) every cut in node $u(\pi)$ damages at least $\varepsilon|F_{u(\pi)}|/2$ centers in $F_{u(\pi)}$.*

Proof. Consider any fixed path π with length less than $\log_2 k/4$. We upper bound the probability that both events (a) and (b) do not happen conditioned on $F_{u(\pi)}$. If $|F_{u(\pi)}| \leq \sqrt{k}$, then the event (a) happens. For the case $F_{u(\pi)}$ contains more than $\sqrt{k}$ centers, we pick an arbitrary threshold cut (i, θ) in the node $u(\pi)$. For every center c in $F_{u(\pi)}$, the probability we damage this center c is at least ε. Let X_j be the indicator random variable that the j-th center in $F_{u(\pi)}$ is damaged by the threshold cut (i, θ). Then, we have the expected number of centers in $F_{u(\pi)}$ damaged by this cut (i, θ) is

$$\mathbb{E}\left[\sum_j X_j\right] \geq \varepsilon \left|F_{u(\pi)}\right|.$$

Let $\mu = \mathbb{E}[\sum_j X_j]$. By the Chernoff bound for Bernoulli random variables, we have

$$\mathbb{P}\left[\sum_j X_j \leq \varepsilon\left|F_{u(\pi)}\right|/2\right] \leq \mathbb{P}\left[\sum_j X_j \leq \mu/2\right] \leq e^{-\mu/8} \leq e^{-\varepsilon\sqrt{k}/8}.$$

Using the union bound over all threshold cuts in $u(\pi)$, the failure probability that both event (a) and (b) do not happen is at most $e^{-\varepsilon\sqrt{k}/16}$. The number of paths with length less than $\log_2 k/4$ is at most $m(2d/\varepsilon)^m \leq e^{-\log^2 k}$. Thus, by the union bound over all paths with length less than $\log_2 k/4$, we get the conclusion. □

Proof of Theorem D.4. By Lemma D.5 and Lemma D.6, we can find an instance X such that both two properties hold. We first show that the threshold tree must separate all centers. Suppose there is a leaf contains more than one center. Since the distance between every two centers is at least $\sqrt{d}/4$ and there are many points at each center, the cost for this leaf can be arbitrary large. To separate all centers, the depth of the threshold tree is at least $\lceil \log_2 k \rceil$.

We now lower bound the cost for every threshold tree that separates all centers. Consider any threshold tree T that separates all centers. We consider the following two cases. If the number of damaged centers at level $\lfloor \log_2 k \rfloor/4$ of threshold tree T is more than $k/2$, then the cost given by T is at least

$$\mathrm{cost}_{\ell_2}(X, T) \geq \frac{k}{2} \cdot \frac{\sqrt{d}}{8} = \frac{k\sqrt{d}}{16}.$$

If the number of damaged centers at level $\lfloor \log_2 k \rfloor/4$ of threshold tree T is less than $k/2$, then the number of undamaged centers at every level $i = 1, 2, \ldots, \lfloor \log_2 k \rfloor/4$ is at least $k/2$. We call a node u a small node if it contains at most $\sqrt{k}$ undamaged centers, otherwise we call it a large node. Then, we lower bound the number of damaged centers generated at any fixed level $i \in \{1, 2, \cdots, \lfloor \log_2 k \rfloor/4\}$. Since the number of nodes at level i is at most $k^{1/4}$, the number of undamaged centers in small nodes at level i is at most $k^{3/4}$. Thus, the number of undamaged centers in large nodes at level i is at least $k/4$. By Lemma D.6, the number of damaged centers generated at level i is at least $\varepsilon k/8$. Therefore, the cost given by this threshold tree T is at least

$$\mathrm{cost}_{\ell_2}(X,T) \geq \frac{\lfloor \log_2 k \rfloor}{4} \frac{\varepsilon k}{8} \frac{\sqrt{d}}{8} = \Omega(k\sqrt{d}\varepsilon \log k).$$

Note that the optimal cost for this instance is at most $k\varepsilon\sqrt{d}$ and $\varepsilon = 1/\lceil \log k \rceil$. Combining the two cases above, we have the cost given by threshold tree T is at least

$$\mathrm{cost}_{\ell_2}(X,T) = \Omega(k\sqrt{d}\varepsilon \log k) = \Omega(\log k)\mathrm{OPT}_{\ell_2}(X).$$

□

E. Fast Algorithm

In this section, we provide a fast variant of Algorithm 1 with running time $O(kd \log^2 k)$. The input of this algorithm is the set of reference centers $c^1, \dots, c^k$ and the output is a threshold tree that splits all centers. The algorithm does not consider the data points (hence, it does not explicitly assign them to clusters). It takes an extra $O(nk)$ time to assign every point in the data set to one of the leaves of the threshold tree.

This fast variant of Algorithm 1 picks a separate threshold cut ω^u for each leaf u. This cut is chosen uniformly at random from R^u, where

$$R^u = \bigcup_{c^i, c^j \in X_u} S_{ij}.$$

That is, R^u is the set of all cuts ω that separate at least two centers in X_u. The algorithm then splits leaf u into two parts using ω^u.

A straightforward implementation of the algorithm partitions each leaf by computing $\delta_c(\omega^u)$ for all centers c in X_u. It takes $O(d \cdot |X_u \cap C|)$ time to find R^u and sample ω^u for each u. It takes time $O(|X_u \cap C|)$ to split X_u into two groups. Thus, the total running time of this implementation of the algorithm is $O(k^2 d)$. We now discuss how to implement this algorithm with running time $O(kd \log^2 k)$ using red-black trees.

The improved algorithm stores centers for each leaf of the threshold tree in d red-black trees. Centers in the i-th red-black tree are sorted by the i-th coordinate. Using red-black trees, we can find the minimum and maximum values of c_i for $c \in C \cap X_u$ in time $O(d \log k)$. Denote these values by a_i and b_i, then

$$R^u = \bigcup_i \{i\} \times [a_i, b_i].$$

Hence, we can find R^u and sample a random cut ω^u in time $O(d \log k)$ for each u.

To partition set X^u into two groups with respect to $\omega^u = (i, \theta)$, we consider the i-th red-black tree for leaf u and find the sizes of the new parts, $Left = \{c \in X_u \cap C : c_i \leq \theta\}$ and $Right = \{c \in X_u \cap C : c_i > \theta\}$. We choose the set that contains fewer centers. Let us assume that the second set ($Right$) is smaller the first one ($Left$). Then, we find all centers in $Right$ and delete them from this red-black tree and all other red-black trees for node u. We assign the updated red-black trees (with deleted $Right$) to the left child of u. For the right child, we build d new red-black trees, which store centers for $Right$. Since we delete at most half of all centers in the red-black tree, each center is deleted at most $O(\log k)$ times. Each time it is deleted from d trees and inserted into d trees. Each deletion and insertion operation takes time $O(\log k)$. Thus, the total time of all deletion and insertion operations is $O(kd \log^2 k)$.

We note that though this algorithm slightly differs from the algorithm presented in Section 4, its approximation guarantees are the same.