How to Find a Good Explanation for Clustering?

Sayan Bandyapadhyay[1], Fedor Fomin[1], Petr A Golovach[1], William Lochet[1], Nidhi Purohit[1], Kirill Simonov [2]

[1] Department of Informatics, University of Bergen, Norway
[2] Algorithms and Complexity Group, TU Wien, Vienna, Austria
{sayan.bandyapadhyay, fedor.fomin, petr.golovach, william.lochet, nidhi.purohit}@uib.no, kirillsimonov@gmail.com

Abstract

k-means and k-median clustering are powerful unsupervised machine learning techniques. However, due to complicated dependences on all the features, it is challenging to interpret the resulting cluster assignments. Moshkovitz, Dasgupta, Rashtchian, and Frost proposed an elegant model of explainable k-means and k-median clustering in ICML 2020. In this model, a decision tree with k leaves provides a straightforward characterization of the data set into clusters.

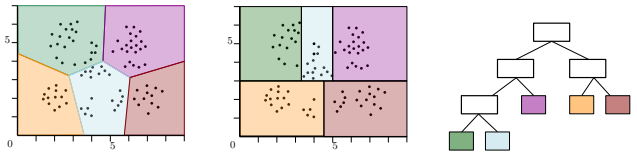
We study two natural algorithmic questions about explainable clustering. (1) For a given clustering, how to find the "best explanation" by using a decision tree with k leaves? (2) For a given set of points, how to find a decision tree with k leaves minimizing the k-means/median objective of the resulting explainable clustering? To address the first question, we introduce a new model of explainable clustering. Our model, inspired by the notion of outliers in robust statistics, is the following. We are seeking a small number of points (outliers) whose removal makes the existing clustering well-explainable. For addressing the second question, we initiate the study of the model of Moshkovitz et al. from the perspective of multivariate complexity. Our rigorous algorithmic analysis sheds some light on the influence of parameters like the input size, dimension of the data, the number of outliers, the number of clusters, and the approximation ratio, on the computational complexity of explainable clustering.

Introduction

Interpretation or explanation of decisions produced by learning models, including clustering, is a significant direction in machine learning (ML) and artificial intelligence (AI), and has given rise to the subfield of Explainable AI. Explainable AI has attracted a lot of attention from the researchers in recent years (see the surveys by Carvalho et al. (2019) and Marcinkevičs and Vogt (2020)). All these works can be divided into two main categories: *pre-modelling* (Wang and Rudin 2015; Ustun and Rudin 2016; Hastie and Tibshirani 1986; Feng and Simon 2017; Lu et al. 2018) and *post-modelling* (Ribeiro, Singh, and Guestrin 2016; Shrikumar, Greenside, and Kundaje 2017; Breiman 2001; Sundararajan, Taly, and Yan 2017; Lundberg and Lee 2017) explainability. While post-modeling explainability focuses on giving reasoning behind decisions made by black box models, pre-modeling explainability deals with ML systems that are inherently understandable or perceivable by humans. One of the canonical approaches to pre-modelling explainability builds on decision trees (Molnar 2020; Murdoch et al. 2019). In fact, a significant amount of work on explainable clustering is based on unsupervised decision trees (Bertsimas, Orfanoudaki, and Wiberg 2021; Fraiman, Ghattas, and Svarc 2013; Geurts et al. 2007; Ghattas, Michel, and Boyer 2017; Lipton 2018; Moshkovitz et al. 2020). In each node of the decision tree, the data is partitioned according to some features' threshold value. While such a *threshold tree* provides a clear interpretation of the resulting clustering, its cost measured by the standard k-means/median objective can be significantly worse than the cost of the optimal clustering. Thus, on the one hand, the efficient algorithms developed for k-means/median clustering (Aggarwal and Reddy 2013) are often challenging to explain. On the other hand, the easily explainable models could output very costly clusterings. Subsequently, Moshkovitz et al. (2020), in a fundamental work, posed the natural algorithmic question of whether it is possible to kill two birds with one stone? To be precise, is it possible to design an efficient procedure for clustering that

- Is explainable by a small decision tree; and
- Does not cost significantly more than the cost of an optimal k-means/median clustering?

To address this question, Moshkovitz et al. (2020) introduced explainable k-means/median clustering. In this scheme, a clustering is represented by a binary (*threshold*) tree whose leaves correspond to clusters, and each internal node corresponds to partitioning a collection of points by a threshold on a fixed coordinate. Thus, the number of leaves in such a tree is k, the number of clusters sought. Also, any cluster assignment can be explained by the thresholds along the corresponding root-leaf path. For example, consider Fig. 1: Fig. 1a shows an optimal 5-means clustering of a 2D data set; Fig. 1b shows an explainable 5-means clustering of the same data set; The threshold tree inducing the explainable clustering is shown in Fig. 1c. The tree has five leaves, corresponding to 5 clusters. Note that in this model of explainability, any clustering has a clear geometric interpretation, where each cluster is formed by a set of axis-aligned cuts defined by the tree. As Moshkovitz et al. argue, the classical k-means clustering algorithm leads to more com-



(a) (b) (c)

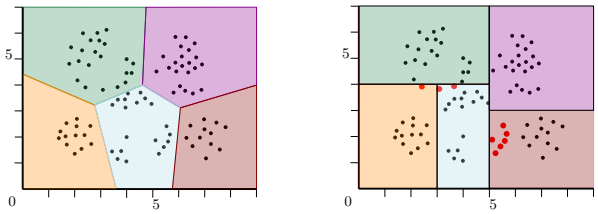
Figure 1: (a) An example of an optimal solution to 5-means. (b) An explainable 5-means clustering and (c) the corresponding threshold tree.

plicated clusters while the threshold tree leads to an easy explanation. The advantage of the explainable approach becomes even more evident in higher dimensions when many feature values in k-means contribute to the formation of the clusters.

Moshkovitz et al. (2020) define the quality of any explainable clustering as the "cost of explainability", that is the ratio of the cost of the explainable clustering to the cost of an optimal clustering. Subsequently, they obtain efficient algorithms for computing explainable clusterings whose "cost of explainability" is $\mathcal{O}(k)$ for k-median and $\mathcal{O}(k^2)$ for k-means. Recently, these bounds have been improved significantly (Charikar and Hu 2021; Esfandiari, Mirrokni, and Narayanan 2021; Gamlath et al. 2021; Laber and Murtinho 2021; Makarychev and Shan 2021).

Our contributions. In this work, we propose a new model for explaining a clustering, called CLUSTERING EXPLANATION. Our approach to explainability is inspired by the research on robustness in statistics and machine learning, especially the vast field of outlier detection and removal in the context of clustering (Chen 2008; Friggstad et al. 2019; Feng et al. 2019; Charikar et al. 2001; Chakrabarty, Goyal, and Krishnaswamy 2016; Harris et al. 2019; Krishnaswamy, Li, and Sandeep 2018). In this model, we are given a k-means/median clustering and we would like to explain the clustering by a threshold tree *after removing a subset of points*. To be precise, we are interested in finding a subset of points S (which are to be removed) and a threshold tree T such that the explainable clustering induced by the leaves of T is exactly the same as the given clustering after removing the points in S. For the given clustering, we define an optimal (or best) explainable clustering to be the one that minimizes the size of S, i.e. for which the given clustering can be explained by removing the minimum number of points. Thus in CLUSTERING EXPLANATION we measure the "explainability" as the number of outlying points whose removal turns the given clustering into an explainable clustering. The reasoning behind the new measure of cluster explainability is the following. In certain situations, we would be satisfied with a small decision tree explaining clustering of all but a few outlying data points. We note that for a given clustering that is already an explainable clustering, i.e. can be explained by a threshold tree, the size of S is 0.

In Fig. 2, we provide an example of an optimal 5-means clustering of exactly the same data set as in Fig. 1. However, the new explainable clustering is obtained in a different way.



(a) (b)

Figure 2: (a) An optimal 5-clustering and (b) an explainable clustering that fits this clustering after removing the larger (red) points.

If we remove a small number of points (in Fig. 2b these are the 9 red larger points), then the explainable clustering is same as the optimal clustering after removing those 9 points.

We note that CLUSTERING EXPLANATION corresponds to the classical machine learning setting of interpreting a black-box model, i.e. it lies within the scope of post-modeling explainability. Surprisingly, this area is widely unexplored when it comes to rigorous algorithmic analysis of clustering explanation. Consequently, we study CLUSTERING EXPLANATION from the perspective of computational complexity. Our new model naturally raises the following algorithmic questions: (i) *Given a clustering, how efficiently can one decide whether the clustering can be explained by a threshold tree (without removing any points)?* and (ii) *Given a clustering and an integer s, how efficiently can one decide whether the clustering can be explained by removing s points?*

In our work, we design a polynomial time algorithm that resolves the first question. Regarding the second question, we give an algorithm that in time $2^{2\min\{s,k\}} \cdot n^{2d} \cdot (dn)^{\mathcal{O}(1)}$ decides whether a given clustering of n points in $\mathbb{R}^d$ could be explained by removing s points. We also give an $n^{\mathcal{O}(1)}$ time $(k-1)$-approximation algorithm for CLUSTERING EXPLANATION. That is, we give a polynomial time algorithm that returns a solution set of at most $s(k-1)$ points that are to be removed, whereas any best explainable clustering removes s points. Moreover, we provide an efficient data reduction procedure that reduces an instance of CLUSTERING EXPLANATION to an equivalent instance with at most $r = 2(s+1)dk$ points in $\mathbb{R}^d$ with integer coordinates within the range $\{1, \ldots, r\}$. The procedure can be used to speed up *any* algorithm for CLUSTERING EXPLANATION, as long as $n > 2(s+1)dk$. We complement our algorithms by showing a hardness lower bound. In particular, we show that CLUSTERING EXPLANATION cannot be approximated within a factor of $F(s)$ in time $f(s)(nd)^{o(s)}$, for any functions F and f, unless Exponential Time Hypothesis (ETH) (Impagliazzo, Paturi, and Zane 2001) fails. All these results appear in Section .

We also provide new insight into the computational complexity of the model of Moshkovitz et al. (2020). While the vanilla k-median and k-means problems are NP-hard for $k = 2$ (Aloise et al. 2009; Drineas et al. 2004; Dasgupta 2008) or $d = 2$ (Mahajan, Nimbhorkar, and Varadarajan

Model	Algorithms	Lower bounds
Clustering Explanation	$2^{2\min\{s,k\}}n^{2d}n^{\mathcal{O}(1)}$ $(k-1)$-approximation Red. to $\mathcal{O}(sdk)$ points	No $F(s)$-appr. in $f(s)(nd)^{o(s)}$
Explainable Clustering	$(4nd)^{k+\mathcal{O}(1)}$ $n^{2d}\cdot n^{\mathcal{O}(1)}$	$f(k)\cdot n^{o(k)}$
Approx. Explainable Clustering	$(\frac{8dk}{\epsilon})^k\cdot n^{\mathcal{O}(1)}$	

Table 1: A summary of our results.

2012), this is not the case for explainable clustering! We design two simple algorithms computing optimal (best) explainable clustering with k-means/median objective that run in time $(4nd)^{k+\mathcal{O}(1)}$ and $n^{2d}\cdot n^{\mathcal{O}(1)}$, respectively. Hence for constant k or constant d, an optimal explainable clustering can be computed in polynomial time. The research on approximation algorithms on the "cost of explainability" in (Moshkovitz et al. 2020; Charikar and Hu 2021; Esfandiari, Mirrokni, and Narayanan 2021; Gamlath et al. 2021; Laber and Murtinho 2021; Makarychev and Shan 2021) implicitly assumes that solving the problem exactly is NP-hard. However, we did not find a proof of this fact in the literature. To fill this gap, we obtain the following hardness lower bound: An optimal explainable clustering cannot be found in $f(k)\cdot n^{o(k)}$ time for any computable function $f(\cdot)$, unless Exponential Time Hypothesis (ETH) fails. This lower bound demonstrates that asymptotically the running times of our simple algorithms are unlikely to be improved. Our reduction also yields that the problem is NP-hard. These results are described in Section .

Finally, we combine the above two explainability models to obtain the Approximate Explainable Clustering model: For a collection of n points in $\mathbb{R}^d$ and a positive real constant $\varepsilon < 1$, we seek whether we can identify at most εn outliers, such that the cost of explainable k-means/median of the remaining points does not exceed the optimal cost of an explainable k-means/median clustering of the original data set. Thus, if we are allowed to remove a small number of points, can we do as good as any original optimal solution? While our hardness result of Section holds for explaining the whole dataset, by "sacrificing" a small fraction of points it might be possible to solve the problem more efficiently. And indeed, for this model, we obtain an algorithm whose running time $(\frac{8dk}{\epsilon})^k\cdot n^{\mathcal{O}(1)}$ has a significantly better dependence on d and k. For example, compare this with the above time bounds of $(4nd)^{k+\mathcal{O}(1)}$ and $n^{2d}\cdot(dn)^{\mathcal{O}(1)}$. This algorithm appears in Section . See Table 1 for a summary of all our results. Some of the proofs are deferred to the full version of the paper (Bandyapadhyay et al. 2021).

Preliminaries

k-means/median. Given a collection $\mathbf{X}=\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ of n points in $\mathbb{R}^d$ and a positive integer k, the task of k-*clustering* is to partition $\mathbf{X}$ into k parts $\mathbf{C}_1,\dots,\mathbf{C}_k$, called *clusters*, such that the *cost* of clustering is minimized. We follow the convention in the previous work (Moshkovitz et al. 2020) for defining the cost. In particular, for k-means, we consider the Euclidean distance and for k-median, the Manhattan distance. For a collection of points $\mathbf{X}'$ of $\mathbb{R}^d$, we define

$$\mathsf{cost}_2(\mathbf{X}')=\min_{\mathbf{c}\in\mathbb{R}^d}\sum_{\mathbf{x}\in\mathbf{X}'}\|\mathbf{c}-\mathbf{x}\|_2^2, \tag{1}$$

and call the point $\mathbf{c}^*\in\mathbb{R}^d$ minimizing the sum in (1) the *mean* of $\mathbf{X}'$. For a clustering $\{\mathbf{C}_1,\dots,\mathbf{C}_k\}$ of $\mathbf{X}\subseteq\mathbb{R}^d$, its *$k$-means* (or simply means) cost is $\mathsf{cost}_2(\mathbf{C}_1,\dots,\mathbf{C}_k)=\sum_{i=1}^k\mathsf{cost}_2(\mathbf{C}_i)$. With respect to the Manhattan distance, we define analogously $\mathsf{cost}_1(\mathbf{X}')=\min_{\mathbf{c}\in\mathbb{R}^d}\sum_{\mathbf{x}\in\mathbf{X}'}\|\mathbf{c}-\mathbf{x}\|_1$, which is minimized at the *median* of $\mathbf{X}'$, and $\mathsf{cost}_1(\mathbf{C}_1,\dots,\mathbf{C}_k)=\sum_{i=1}^k\mathsf{cost}_1(\mathbf{C}_i)$, which we call the *$k$-median* (or simply median) cost of the clustering.

Explainable clustering. For a vector $\mathbf{x}\in\mathbb{R}^d$, we use $\mathbf{x}[i]$ to denote the i-th element (coordinate) of the vector for $i\in\{1,\dots,d\}$. Let $\mathbf{X}$ be a collection of points of $\mathbb{R}^d$. For $i\in\{1,\dots,d\}$ and $\theta\in\mathbb{R}$, we define $\mathsf{Cut}_{i,\theta}(\mathbf{X})=(\mathbf{X}_1,\mathbf{X}_2)$, where $\{\mathbf{X}_1,\mathbf{X}_2\}$ is a partition of $\mathbf{X}$ with

$$\mathbf{X}_1=\{\mathbf{x}\in\mathbf{X}\mid\mathbf{x}[i]\le\theta\}\text{ and }\mathbf{X}_2=\{\mathbf{x}\in\mathbf{X}\mid\mathbf{x}[i]>\theta\}.$$

Then, given a collection $\mathbf{X}\subseteq\mathbb{R}^d$ and a positive integer k, we cluster $\mathbf{X}$ as follows. If $k=1$, then $\mathbf{X}$ is the unique cluster. If $k=2$, then we choose $i\in\{1,\dots,d\}$ and $\theta\in\mathbb{R}$ and construct two clusters $\mathbf{C}_1$ and $\mathbf{C}_2$, where $(\mathbf{C}_1,\mathbf{C}_2)=\mathsf{Cut}_{i,\theta}(\mathbf{X})$. For $k>2$, we select $i\in\{1,\dots,d\}$ and $\theta\in\mathbb{R}$, and construct a partition $(\mathbf{X}_1,\mathbf{X}_2)=\mathsf{Cut}_{i,\theta}(\mathbf{X})$ of $\mathbf{X}$. Then clustering of $\mathbf{X}$ is defined recursively as the union of a k_1-clustering of $\mathbf{X}_1$ and a k_2-clustering of $\mathbf{X}_2$ for some integers k_1 and k_2 such that $k_1+k_2=k$. We say that a clustering $\{\mathbf{C}_1,\dots,\mathbf{C}_k\}$ is an *explainable k-clustering* of a collection of points $\mathbf{X}\subseteq\mathbb{R}^d$ if $\mathbf{C}_1,\dots,\mathbf{C}_k$ can be constructed by the described procedure.

Threshold tree. It is useful to represent an explainable k-clustering as a triple (T,k,φ), called a *threshold tree*, where T is a rooted binary tree with k leaves, where each nonleaf node has two children called *left* and *right*, respectively, and $\varphi\colon U\to\{1,\dots,d\}\times\mathbb{R}$, where U is the set of nonleaf nodes of T. For each node v of T, we compute a collection of points $\mathbf{X}_v\subseteq\mathbf{X}$. For the root r, $\mathbf{X}_r=\mathbf{X}$. Let v be a nonleaf node of T and let u and w be its left and right children, respectively, and assume that $\mathbf{X}_v$ is constructed. We compute $(\mathbf{X}_u,\mathbf{X}_w)=\mathsf{Cut}_{\varphi(v)}(\mathbf{X})$. If v is a leaf, then $\mathbf{X}_v$ is a cluster. A clustering $\{\mathbf{C}_1,\dots,\mathbf{C}_k\}$ is an explainable k-clustering of a collection of points $\mathbf{X}\subseteq\mathbb{R}^d$ if there is a threshold tree (T,k,φ) such that $\mathbf{C}_1,\dots,\mathbf{C}_k$ are the clusters corresponding to the leaves of T. Note that T is a full binary tree with k leaves and the total number of such trees is the $(k-1)$-th Catalan number, which is upper bounded by 4^k.

For a collection $\mathbf{X}=\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ of n points and $i\in\{1,\dots,d\}$, we denote by $\mathsf{coord}_i(\mathbf{X})$ the set of distinct values of i-th coordinates $\mathbf{x}_j[i]$ for $j\in\{1,\dots,n\}$. It is easy to observe that in the construction of a threshold tree for a set of points $\mathbf{X}\subseteq\mathbb{R}^d$, it is sufficient to consider cuts $\mathsf{Cut}_{i,\theta}$ with

$\theta \in \mathsf{coord}_i(\mathbf{X})$; we call such values of θ and cuts *canonical*. We say that a threshold tree (T, k, φ) for a collection of points $\mathbf{X} \subseteq \mathbb{R}^d$ is *canonical*, if for every nonleaf node $u \in V(T)$, $\varphi(u) = (i, \theta)$ where $\theta \in \mathsf{coord}_i(\mathbf{X})$. Throughout the paper we consider only canonical threshold trees.

Clustering Explanation

Clustering explanation. In the CLUSTERING EXPLANATION problem, the input contains a k-clustering $\{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbf{X} \subseteq \mathbb{R}^d$ and a nonnegative integer s, and the task is to decide whether there is a collection of points $W \subseteq \mathbf{X}$ with $|W| \leq s$ such that $\{\mathbf{C}_1 \setminus W, \ldots, \mathbf{C}_k \setminus W\}$ is an explainable k-clustering. Note that some $\mathbf{C}_i \setminus W$ may be empty here.

A Polynomial-time $(k-1)$-Approximation

In the optimization version of CLUSTERING EXPLANATION, we are given a k-clustering $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbf{X}$ in $\mathbb{R}^d$, and the goal is to find a minimum-sized subset $W \subseteq \mathbf{X}$ such that $\{\mathbf{C}_1 \setminus W, \ldots, \mathbf{C}_k \setminus W\}$ is an explainable clustering. In the following, we design an approximation algorithm for this problem based on a greedy scheme.

For any subset $W \subseteq \mathbf{X}$, let $\mathcal{C} - W = \{\mathbf{C}_1 \setminus W, \ldots, \mathbf{C}_k \setminus W\}$. Also, for any subset $Y \subseteq \mathbf{X}$, define the clustering induced by Y as $\mathcal{C}(Y) = \{\mathbf{C}_1 \cap Y, \ldots, \mathbf{C}_k \cap Y\}$. Denote by $\mathrm{OPT}(Y)$ the size of the minimum-sized subset W such that the clustering $\mathcal{C}(Y) - W$ is explainable. First, we have the following simple observation which follows trivially from the definition of $\mathrm{OPT}(.)$.

Observation 1. *For any subset $Y \subseteq \mathbf{X}$, $\mathrm{OPT}(Y) \leq \mathrm{OPT}(\mathbf{X})$.*

For any cut (i, θ) where $i \in \{1, \ldots, d\}$ and $\theta \in \mathsf{coord}_i(\mathbf{X})$, let $L(i, \theta) = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}[i] \leq \theta\}$ and $R(i, \theta) = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}[i] > \theta\}$.

Lemma 1. *Consider any subset $Y \subseteq \mathbf{X}$ such that $\mathcal{C}(Y)$ contains at least two non-empty clusters. It is possible to select a cut (i, θ) for $i \in \{1, \ldots, d\}$ and $\theta \in \mathsf{coord}_i(Y)$, and a subset $W \subseteq Y$, in polynomial time, such that (i) each cluster in $\mathcal{C}(Y) - W$ is fully contained in either $L(i, \theta)$ or in $R(i, \theta)$, (ii) at least one cluster in $\mathcal{C}(Y) - W$ is in $L(i, \theta)$, (iii) at least one cluster in $\mathcal{C}(Y) - W$ is in $R(i, \theta)$ and (iv) size of W is at most $\mathrm{OPT}(Y)$.*

Before we prove this lemma, we show how to use it to design the desired approximation algorithm.

The Algorithm. We start with the set of all points $\mathbf{X}$. We apply the algorithm in Lemma 1 with $Y = \mathbf{X}$ to find a cut (i, θ) and a subset $W_1 \subseteq \mathbf{X}$ such that each cluster in $\mathcal{C}(\mathbf{X}) - W_1$ is fully contained in either $L(i, \theta)$ or in $R(i, \theta)$. Let $\mathbf{X}_1 = (\mathbf{X} \setminus W_1) \cap L(i, \theta)$ and $\mathbf{X}_2 = (\mathbf{X} \setminus W_1) \cap R(i, \theta)$. We recursively apply the above step on both $\mathbf{X}_1$ and $\mathbf{X}_2$ separately. If at some level the point set is a subset of a single cluster, we simply return.

The correctness of the above algorithm trivially follows from Lemma 1. In particular, the recursion tree of the algorithm gives rise to the desired threshold tree. Also, the algorithm runs in polynomial time, as each successful cut (i, θ) can be found in polynomial time and the algorithm finds only $k-1$ such cuts that separate the clusters. The last claim follows due to the properties (ii) and (iii) in Lemma 1.

Consider the threshold tree generated by the algorithm. For each internal node u, let X_u be the corresponding points and W_u be the points removed from X_u for finding an explainable clustering of the points in $X_u \setminus W_u$. Note that we have at most $k-1$ such nodes. The total number of points removed from $\mathbf{X}$ for finding the explainable clustering is $\sum_u |W_u|$. By Lemma 1,

$$|W_u| \leq \mathrm{OPT}(X_u).$$

Now, as $X_u \subseteq \mathbf{X}$, by Observation 1, $\mathrm{OPT}(X_u) \leq \mathrm{OPT}(\mathbf{X})$. It follows that

$$\sum_u |W_u| \leq (k-1) \cdot \mathrm{OPT}(\mathbf{X}).$$

Theorem 1. *There is a polynomial-time $(k-1)$-approximation algorithm for the optimization version of* CLUSTERING EXPLANATION.

By noting that $\mathrm{OPT}(\mathbf{X}) = 0$ if $\mathcal{C}$ is an explainable clustering, we obtain the following corollary.

Corollary 1. *Explainability of any given k-clustering in $\mathbb{R}^d$ can be tested in polynomial time.*

Proof of Lemma 1. We probe all possible choices for cuts (i, θ) with $i \in \{1, \ldots, d\}$ and $\theta \in \mathsf{coord}_i(Y)$, and select one which incurs the minimum cost. We also select a subset W of points to be removed w.r.t. each cut. The cost of such a cut is exactly the size of W.

Fix a cut (i, θ). We have the following three cases. In the first case, for all clusters in $\mathcal{C}(Y)$, strictly more than half of the points are contained in $L(i, \theta)$. In this case select a cluster $\mathbf{C}$ which has the minimum intersection with $L(i, \theta)$. Put all the points in $\mathbf{C} \cap L(i, \theta)$ into W. Also, for any other cluster $\mathbf{C}' \in \mathcal{C}(Y)$, put the points in $\mathbf{C}' \cap R(i, \theta)$ into W. The second case is symmetric to the first one – for all clusters in $\mathcal{C}(Y)$, strictly more than half of the points are contained in $R(i, \theta)$. In this case we again select a cluster $\mathbf{C}$ which has the minimum intersection with $R(i, \theta)$. Put all the points in $\mathbf{C} \cap R(i, \theta)$ into W. Also, for any other cluster $\mathbf{C}' \in \mathcal{C}(Y)$, put the points in $\mathbf{C}' \cap L(i, \theta)$ into W. In both of the above cases, the first three desired properties are satisfied for $\mathcal{C}(Y) - W$. In the third case, for each cluster $\mathbf{C} \in \mathcal{C}(Y)$, add the smaller part among $\mathbf{C} \cap L(i, \theta)$ and $\mathbf{C} \cap R(i, \theta)$ to W. In case $|\mathbf{C} \cap L(i, \theta)| = |\mathbf{C} \cap R(i, \theta)|$, we break the tie in a way so that properties (ii) and (iii) are satisfied. As $\mathcal{C}(Y)$ contains at least two clusters this can always be done. Moreover, property (i) is trivially satisfied.

In the above we showed that for all the choices of the cuts, it is possible to select W so that the first three properties are satisfied. Let w_m be the minimum size of the set W over all cuts. As we select a cut for which the size of W is minimized, it is sufficient to show that $w_m \leq \mathrm{OPT}(Y)$.

Let k' be the number of clusters in $\mathcal{C}(Y)$. Consider any optimal set W^* for Y such that $\mathcal{C}(Y) - W^*$ is explainable. Let (i^*, θ^*) be the canonical cut corresponding to the root of the threshold tree corresponding to the explainable clustering $\mathcal{C}(Y) - W^*$. Such a cut exists, as $\mathcal{C}(Y)$ contains at

least two clusters. Let $\widehat{W}$ be the set selected in our algorithm corresponding to the cut (i^*, θ^*). In the first of the above mentioned three cases, suppose W^* does not contain the part $\mathbf{C} \cap L(i^*, \theta^*)$ fully for any of the k' clusters $\mathbf{C} \in \mathcal{C}(Y)$. In other words, $\mathcal{C}(Y) - W^*$ contains points from each such part $\mathbf{C} \cap L(i^*, \theta^*)$. But, then even after choosing the root cut (i^*, θ^*) we still need k' more cuts to separate the points in $(Y \setminus W^*) \cap L(i^*, \theta^*)$, which contains points from all the k' clusters. However, by definition, the threshold tree must use only k' cuts and hence we reach to a contradiction. Hence, $\mathbf{C}^* \cap L(i^*, \theta^*)$ must be fully contained in W^* for some $\mathbf{C}^* \in \mathcal{C}(Y)$. In this case, our algorithm adds the points in $\mathbf{C} \cap L(i^*, \theta^*)$ to $\widehat{W}$ such that the size $|\mathbf{C} \cap L(i^*, \theta^*)|$ is minimized over all $\mathbf{C} \in \mathcal{C}(Y)$ and for any other cluster $\mathbf{C}' \in \mathcal{C}(Y)$, we put the points in $\mathbf{C}' \cap R(i^*, \theta^*)$ into $\widehat{W}$. Thus, $|\widehat{W}| \leq |W^*| = \mathrm{OPT}(Y)$. The proof for the second case is the same as the one for the first case. We discuss the proof for the third case. Consider the clusters $\mathbf{C} \in \mathcal{C}(Y)$ such that both $\mathbf{C} \cap L(i^*, \theta^*)$ and $\mathbf{C} \cap R(i^*, \theta^*)$ are non-empty. Note that these are the only clusters whose points are put into $\widehat{W}$. But, then W^* must contain all the points from at least one of the parts $\mathbf{C} \cap L(i^*, \theta^*)$ and $\mathbf{C} \cap R(i^*, \theta^*)$. For each such cluster $\mathbf{C}$, we add the smaller part among $\mathbf{C} \cap L(i, \theta)$ and $\mathbf{C} \cap R(i, \theta)$ to $\widehat{W}$. Hence, in this case also $|\widehat{W}| \leq |W^*| = \mathrm{OPT}(Y)$. The lemma follows by noting that $w_m \leq |\widehat{W}|$. □

Exact Algorithm

Our $2^{2\min\{s,k\}} \cdot n^{2d} \cdot (dn)^{\mathcal{O}(1)}$ time algorithm is based on a novel dynamic programming scheme. Here, we briefly describe the algorithm. Our first observation is that each subproblem can be defined w.r.t. a bounding box in $\mathbb{R}^d$, as each cut used to split a point set in any threshold tree is an axis-parallel hyperplane. The number of such distinct bounding boxes is at most n^{2d}, as in each dimension a box is specified by two bounding values. This explains the n^{2d} factor in the running time. Now, consider a fixed bounding box corresponding to a subproblem containing a number of given clusters, may be partially. If a new canonical cut splits a cluster, then one of the two resulting parts has to be removed, and this choice has to be passed on along the dynamic programming. As we remove at most s points and the number of clusters is at most k, the number of such distinct choices can be bounded by $2^{2\min\{s,k\}}$. This roughly gives us the following theorem. The detailed proof is quite technical and is deferred to the full version of the paper.

Theorem 2. CLUSTERING EXPLANATION *can be solved in* $2^{2\min\{s,k\}} \cdot n^{2d} \cdot (dn)^{\mathcal{O}(1)}$ *time.*

Data Reduction

Theorem 3. *Let $r = 2(s+1)dk$. There is a polynomial-time algorithm that, given an instance of* CLUSTERING EXPLANATION*, produces an equivalent one with at most r points in $\{1, \ldots, r\}^d$.*

Proof. Let $(\mathcal{C}, s)$ be an instance of CLUSTERING EXPLANATION, where $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ for disjoint collections of points $\mathbf{C}_i$ of $\mathbf{R}^d$. Let $\mathbf{X} = \bigcup_{i=1}^{k} \mathbf{C}_i$.

Our first aim is to reduce the number of points. For this, we use a procedure that marks essential points.

For every $i \in \{1, \ldots, k\}$ and every $j \in \{1, \ldots, d\}$, do the following:

- Order the points of $\mathbf{C}_i$ by the increase of their j-th coordinate; the ties are broken arbitrarily.
- *Mark* the first $\min\{s+1, |\mathbf{C}_i|\}$ points and the last $\min\{s+1, |\mathbf{C}_i|\}$ points in the ordering.

The procedure marks at most $2(s+1)dk$ points. Then we delete the remaining unmarked points. Formally, we denote by $\mathbf{Y}$ the collection of marked points and set $\mathbf{S}_i = \mathbf{C}_i \cap \mathbf{Y}$ for all $i \in \{1, \ldots, k\}$. Then we consider the instance $(\mathcal{S}, s)$ of CLUSTERING EXPLANATION, where $\mathcal{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_k\}$. We show the following claim.

Claim 0.1. *$(\mathcal{C}, s)$ is a yes-instance of* CLUSTERING EXPLANATION *if and only if $(\mathcal{S}, s)$ is a yes-instance.*

Proof of Claim 0.1. Trivially, if $(\mathcal{C}, s)$ is a yes-instance, then $(\mathcal{S}, s)$ is a yes-instance, because we just deleted some point to construct $(\mathcal{S}, s)$. We show that if $(\mathcal{S}, s)$ is a yes-instance, then $(\mathcal{C}, s)$ is a yes-instance.

Because $(\mathcal{S}, s)$ is a yes-instance, there is a collection of at most s points $\mathbf{W} \subseteq \mathbf{Y}$ such that $\{\mathbf{S}_1 \setminus \mathbf{W}, \ldots, \mathbf{S}_k \setminus \mathbf{W}\}$ is an explainable k-clustering. In other words, there is an explainable clustering of $\mathbf{Y} \setminus \mathbf{W}$ with a canonical threshold tree (T, k, φ) such that the clusters $\mathbf{S}_1 \setminus \mathbf{W}, \ldots, \mathbf{S}_k \setminus \mathbf{W}$ correspond to the leaves of the threshold tree. We claim that if we use the same threshold tree for $\mathbf{X} \setminus \mathbf{W}$, then $\mathbf{C}_1 \setminus \mathbf{W}, \ldots, \mathbf{C}_k \setminus \mathbf{W}$ correspond to the leaves.

The proof is by contradiction. Assume that at least one collections of points corresponding to a leaf is distinct from every $\mathbf{C}_1 \setminus \mathbf{W}, \ldots, \mathbf{C}_k \setminus \mathbf{W}$. Then there is a node $v \in V(T)$ such that for some $j \in \{1, \ldots, k\}$, $\mathbf{C}_j \setminus \mathbf{W}$ is split by the cut $\mathsf{Cut}_{i,\theta}$ for $(i, \theta) = \varphi(v)$, that is, for $(\mathbf{A}, \mathbf{B}) = \mathsf{Cut}_{i,\theta}(\mathbf{X})$, $\mathbf{A} \cap (\mathbf{C}_j \setminus \mathbf{W}) \neq \emptyset$ and $\mathbf{B} \cap (\mathbf{C}_j \setminus \mathbf{W}) \neq \emptyset$. Observe that either $\mathbf{A} \cap (\mathbf{S}_j \setminus \mathbf{W}) = \emptyset$ or $\mathbf{B} \cap (\mathbf{S}_j \setminus \mathbf{W}) = \emptyset$. We assume without loss of generality that $\mathbf{A} \cap (\mathbf{S}_j \setminus \mathbf{W}) = \emptyset$ (the other case is symmetric). This means that there is an unmarked point $\mathbf{x} \in \mathbf{C}_j \setminus \mathbf{W}$ in $\mathbf{A}$ and all the marked points of $\mathbf{C}_j \setminus \mathbf{W}$ are in $\mathbf{B}$. Because $\mathbf{C}_j$ has an unmarked point, $|\mathbf{C}_j| \geq 2(s+1)+1$. Following the marking procedure, we order the points of $\mathbf{C}_j$ by the increase of the i-th coordinate breaking ties exactly as in the marking procedure. Let L be the collection of the first $s+1$ points that are marked. Since $|\mathbf{W}| \leq s$, there is $\mathbf{y} \in L \setminus \mathbf{W}$. Because $L \setminus \mathbf{W} \subseteq \mathbf{S}_j \setminus \mathbf{W} \subseteq \mathbf{B}$, we have that $\mathbf{y}[i] > \theta$. Then $\mathbf{x}[i] \geq \mathbf{y}[i] > \theta$ and $\mathbf{x} \in \mathbf{B}$; a contradiction.

We conclude that if we use (T, k, φ) to cluster $\mathbf{X} \setminus \mathbf{W}$, then $\mathbf{C}_1 \setminus \mathbf{W}, \ldots, \mathbf{C}_k \setminus \mathbf{W}$ correspond to the leaves. This proves that $(\mathcal{C}, s)$ is a yes-instance of CLUSTERING EXPLANATION. □

We obtained the instance $(\mathcal{S}, s)$, where $\mathbf{Y} = \bigcup_{i=1}^{k} \mathbf{S}_i$ has $\ell \leq 2(s+1)dk$ points, that is equivalent to the original instance. Now we modify the points to ensure that they are in $\{1, \ldots, \ell\}^d$. For this, we observe that for each $i \in \{1, \ldots, d\}$, the values of the i-th coordinates can be changed if we maintain their order. Formally, we do the following.

For every $i \in \{1,\dots,d\}$, let $\mathsf{coord}_i(\mathbf{Y}) = \{\theta^i_1,\dots,\theta^i_{r_i}\}$, where $\theta^i_1 < \cdots < \theta^i_{r_i}$. For every $\mathbf{y} \in \mathbf{Y}$, we construct a point $\mathbf{z}$, by setting $\mathbf{z}[i] = j$, where $\theta^i_j = \mathbf{y}[i]$, for each $i \in \{1,\dots,d\}$. Then for $\mathbf{S}_i$ containing $\mathbf{y}$, we replace $\mathbf{y}$ by $\mathbf{z}$. Denote by $\mathbf{Z}$ the constructed collection of points, and let $\mathcal{R} = \{\mathbf{R}_1,\dots,\mathbf{R}_k\}$ be the family of the collections of points constructed from $\mathbf{S}_1,\dots,\mathbf{S}_k$.

We have that $(\mathcal{R}, s)$ is a yes-instance of CLUSTERING EXPLANATION if and only if $(\mathcal{S}, s)$ is a yes-instance, and $\mathbf{Z} \subseteq \{1,\dots,\ell\}^d$. Then the data reduction algorithm returns $(\mathcal{R}, s)$. To complete the proof, it remains to observe that the marking procedure is polynomial, and the coordinates replacement also can be done in polynomial time. □

Hardness of Approximation

We show that the CLUSTERING EXPLANATION problem remains hard when the number of points to delete s is small. Specifically, we provide a parameter-preserving reduction from HITTING SET to CLUSTERING EXPLANATION that transfers known results about hardness of approximation for the HITTING SET problem to CLUSTERING EXPLANATION.

Theorem 4. *For any functions f and F, there is no algorithm that approximates* CLUSTERING EXPLANATION *within a factor of $F(s)$ in time $f(s)(nd)^{o(s)}$, unless ETH fails.*

Intuitively, given an instance $(U, \mathcal{A}, \ell)$ of HITTING SET, we construct clusters $\mathbf{C}_0, \dots, \mathbf{C}_m$ in $\mathbb{R}^{\sum_{j\in[m]}|A_j|}$. The clusters $\mathbf{C}_1, \dots, \mathbf{C}_m$ represent the sets in the family $\mathcal{A} = \{A_1,\dots,A_m\}$, and $\mathbf{C}_0$ is a special cluster that needs to be separated from each of $\mathbf{C}_1, \dots, \mathbf{C}_m$ so that the clustering is explainable. The separation can only be performed by removing special points from $\mathbf{C}_0$ each of which corresponds to an element of the universe U. Removing such a point allows for separation between $\mathbf{C}_0$ and each $\mathbf{C}_j$ such that the corresponding set A_j contains the corresponding universe element. The two clusters can be separated along a special coordinate where only that special point "blocks" the separation. This is the crux of the reduction. The full proof appears in the full version.

Explainable Clustering

Explainable k-means/median clustering. We consider the EXPLAINABLE k-MEANS (resp. EXPLAINABLE k-MEDIAN) problem where given a collection $\mathbf{X} \subseteq \mathbb{R}^d$ of $n \geq k$ points, the task is to find an explainable k-clustering $\{\mathbf{C}_1,\dots,\mathbf{C}_k\}$ of $\mathbf{X}$ of minimum k-means (resp. k-median) cost.

Exact Algorithms

Our $(nd)^{k+\mathcal{O}(1)}$ time algorithm is indeed very simple and based on branching technique. At each non-leaf node of threshold tree, we would like to find an optimal cut. As we focus on canonical threshold trees, the number of distinct choices for branching is at most nd. Also as the number of non-leaf nodes in the threshold binary tree is $k-1$, we have the following theorem.

Theorem 5. EXPLAINABLE k-MEANS *and* EXPLAINABLE k-MEDIAN *can be solved in* $(nd)^{k+\mathcal{O}(1)}$ *time.*

Our $n^{2d}\cdot(dn)^{\mathcal{O}(1)}$ time algorithm is based on dynamic programming, which we describe in the following. For two vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$, we write $\mathbf{x} \leq \mathbf{y}$ ($\mathbf{x} < \mathbf{y}$, respectively) to denote that $\mathbf{x}[i] \leq \mathbf{y}[i]$ ($\mathbf{x}[i] < \mathbf{y}[i]$, respectively) for every $i \in \{1,\dots,d\}$. We highlight that when we write $\mathbf{x} < \mathbf{y}$, we require the strict inequality for *every* coordinate.

Theorem 6. EXPLAINABLE k-MEANS *and* EXPLAINABLE k-MEDIAN *can be solved in* $n^{2d}\cdot(dn)^{\mathcal{O}(1)}$ *time.*

Proof. The algorithms for both problems are almost the same. Hence, we demonstrate it for EXPLAINABLE k-MEANS. For simplicity, we only show how to find the minimum cost of clustering but the algorithm can be easily modified to produce an optimal clustering as well by standard arguments.

Let $(\mathbf{X}, k)$ be an instance of the problem with $\mathbf{X} = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ and $\mathbf{X} \subseteq \mathbb{R}^d$. We say that a vector $\mathbf{z} \in (\mathbb{R}\cup\{\pm\infty\})^d$ is *canonical* if $\mathbf{z}[i] \in \mathsf{coord}_i(\mathbf{X}) \cup \{\pm\infty\}$ for every $i \in \{1,\dots,d\}$. For every pair of canonical vectors $(\mathbf{a},\mathbf{b})$ such that $\mathbf{a} \leq \mathbf{b}$ and every positive integer $s \leq k$, we compute the minimum means cost of an explainable s-clustering of $\mathbf{X}_{\mathbf{a},\mathbf{b}} = \{\mathbf{x}_i \in \mathbf{X} \mid \mathbf{a} < \mathbf{x}_i \leq \mathbf{b}\}$ and denote this value $\omega(\mathbf{a},\mathbf{b},s)$. We assume that $\omega(\mathbf{a},\mathbf{b},s) = +\infty$ if $\mathbf{X}_{\mathbf{a},\mathbf{b}}$ does not admit an explainable s-clustering. It is also convenient to assume that $\omega(\mathbf{a},\mathbf{b},s) = +\infty$ if $\mathbf{X}_{\mathbf{a},\mathbf{b}} = \emptyset$, because we are not interested in empty clusters. Notice that the minimum means cost of an explainable k-clustering of $\mathbf{X}$ is $\omega(\mathbf{a}^*,\mathbf{b}^*,k)$, where $\mathbf{a}^*[i] = -\infty$ and $\mathbf{b}^*[i] = +\infty$ for $i \in \{1,\dots,d\}$. We compute the table of values of $\omega(\mathbf{a},\mathbf{b},s)$ consecutively for $s = 1,2,\dots,k$.

If $s = 1$, then by definition,

$$\omega(\mathbf{a},\mathbf{b},s) = \begin{cases} \mathsf{cost}_2(\mathbf{X}_{\mathbf{a},\mathbf{b}}) & \text{if } \mathbf{X}_{\mathbf{a},\mathbf{b}} \neq \emptyset, \\ +\infty & \text{if } \mathbf{X}_{\mathbf{a},\mathbf{b}} = \emptyset, \end{cases}$$

and this value can be computed in polynomial time. Let $s \geq 2$ and assume that the tables are already constructed for the lesser values of s. Consider a pair $(\mathbf{a},\mathbf{b})$ of canonical vectors of $(\mathbb{R}\cup\{\pm\infty\})^d$ such that $\mathbf{a} \leq \mathbf{b}$. For $i \in \{1,\dots,d\}$ and $\theta \in \mathsf{coord}_i(\mathbf{X})$ such that $\mathbf{a}[i] < \theta < \mathbf{b}[i]$, we define the vectors $\mathbf{a}^{i,\theta}$ and $\mathbf{b}^{i,\theta}$ by setting

$$\mathbf{a}^{i,\theta}[j] = \begin{cases} \theta & \text{if } j = i, \\ \mathbf{a}[j] & \text{if } j \neq i, \end{cases} \text{ and } \mathbf{b}^{i,\theta}[j] = \begin{cases} \theta & \text{if } j = i, \\ \mathbf{b}[j] & \text{if } j \neq i. \end{cases}$$

Then we compute $\omega(\mathbf{a},\mathbf{b},s)$ using the following recurrence

$$\begin{aligned} \omega(\mathbf{a},\mathbf{b},s) = \min\{&\omega(\mathbf{a},\mathbf{b}^{i,\theta},s_1) + \omega(\mathbf{a}^{i,\theta},\mathbf{b},s_2) \\ &\text{for } 1 \leq i \leq d, \theta \in \mathsf{coord}_i(\mathbf{X}), \\ &\mathbf{a}[i] < \theta < \mathbf{b}[i], \\ &s_1, s_2 \geq 1, \text{ and } s_1 + s_2 = s\}. \quad (2) \end{aligned}$$

The correctness of (2) follows form the definition of the explainable clustering. It is sufficient to observe that to compute the optimum means cost of an explainable s-clustering of $\mathbf{X}_{\mathbf{a},\mathbf{b}}$, we have to take minimum over the sums of optimum costs of explainable s_1-clusterings and s_2-clusterings of X_1 and X_2, respectively, where $(X_1, X_2) =$

$\mathsf{Cut}_{i,\theta}(\mathbf{X}_{\mathbf{a},\mathbf{b}})$ for some $i \in \{1,\ldots,d\}$, $\theta \in \mathsf{coord}_i(\mathbf{X}_{\mathbf{a},\mathbf{b}})$ and $s_1 + s_2 = s$, and this is exactly what is done in (2).

To evaluate the running time, observe that to compute $\omega(\mathbf{a},\mathbf{b},s)$ using (2), we consider d values of i, at most n values of θ and at most $k \leq n$ values of s_1 and s_2, that is, we go over at most dn^2 choices. Thus, computing $\omega(\mathbf{a},\mathbf{b},s)$ for $s \geq 2$ and fixed $\mathbf{a}$ and $\mathbf{b}$ can be done in $\mathcal{O}(dn^2)$ time. Since there are at most $(n+2)^{2d}$ pairs of canonical vectors $\mathbf{a}$ and $\mathbf{b}$, we obtain that the time to compute the table of values of $\mathbf{X}_{\mathbf{a},\mathbf{b}}$ for all pairs of vectors is $n^{2d} \cdot (dn)^{\mathcal{O}(1)}$. Since the table for $s = 1$ can be constructed in $n^{2d} \cdot (dn)^{\mathcal{O}(1)}$ and we iterate using (2) $k - 1 \leq n$ times, the total running time is $n^{2d} \cdot (dn)^{\mathcal{O}(1)}$. □

Hardness

Theorem 7. EXPLAINABLE k-MEANS *and* EXPLAINABLE k-MEDIAN *are* NP*-complete and cannot be solved in* $f(k) \cdot n^{o(k)}$ *time for a computable function* $f(\cdot)$ *unless ETH fails.*

To prove this theorem, we again reduce from HITTING SET, but the construction is different. Here, we construct a point for each set and also for each element. Then, there is a hitting set of size k iff there is an explainable $(k+1)$-clustering of a suitable cost. The details appear in the full version of the paper.

Approximate Explainable Clustering

Approximate explainable k-means/median clustering. In APPROXIMATE EXPLAINABLE k-MEANS, we are given a collection of n points $\mathbf{X} \subseteq \mathbb{R}^d$, a positive integer $k \leq n$, and a positive real constant $\varepsilon < 1$. Then the task is to find a collection of points $\mathbf{Y} \subseteq \mathbf{X}$ with $|\mathbf{Y}| \geq (1-\varepsilon)|\mathbf{X}|$ and an explainable k-clustering of $\mathbf{Y}$ whose k-median cost does not exceed the optimum k-median cost of an explainable k-clustering for the original collection of points $\mathbf{X}$. Note that we ask about the construction of $\mathbf{Y}$ and the corresponding clustering as the decision variant is trivial. Observe also that the optimum cost is unknown a priori. APPROXIMATE EXPLAINABLE k-MEDIAN differs only by the clustering measure.

Theorem 8. APPROXIMATE EXPLAINABLE k-MEANS *and* APPROXIMATE EXPLAINABLE k-MEDIAN *are solvable in* $(\frac{8dk}{\epsilon})^k \cdot n^{\mathcal{O}(1)}$ *time.*

As the proofs for both problems are identical, we describe only the algorithm for APPROXIMATE EXPLAINABLE k-MEANS.

Proof. Let $\mathbf{X} \subseteq \mathbb{R}^d$ be an instance of APPROXIMATE EXPLAINABLE k-MEANS, (T,k,φ) be the optimal (canonical) threshold tree for explainable k-means clustering and $(C_1,\ldots,C_k)$ the clustering induced by (T,k,φ). The goal of the algorithm will be to guess an approximation of (T,k,φ). Since T is a binary tree with k leaves, guessing T only requires 4^k tries. Guessing φ is more complicated however, as there is $d \cdot n$ choice at each node of T, which gives potentially $(dn)^k$ possibilities, where $n = |\mathbf{X}|$. The idea here will be to guess for every nonleaf node u of T the second element of $\varphi(u)$ up to a precision of $\mathcal{O}(\frac{\epsilon n}{k})$, which gives only $\mathcal{O}(d \cdot \frac{k}{\epsilon})$ choices at each nonleaf node.

More formally, let $n' = \lfloor \frac{\epsilon n}{k} \rfloor$ and note first that if $n' = 0$, then $\frac{\epsilon n}{k} < 1$ and thus $n \leq \frac{k}{\epsilon}$. This means that if $n' = 0$, then the algorithm trying all the possible values of T and φ, and computing the value of the obtained clustering, runs in time $4^k(\frac{dk}{\epsilon}) \cdot n^{\mathcal{O}(1)}$, which ends the proof. From now on, let us assume that $n' \neq 0$.

Let U denote the set of nonleaf nodes of T. Let $\varphi' \colon U \to \{1,\ldots,d\} \times \mathbb{R}$ be the function obtained from φ by rounding, for every $u \in U$, the value of the first element of $\varphi(u)$ to the closest multiple of n'. In other words, if $\varphi(u) = (j,r)$, then $\varphi'(u) = (j, i \cdot n')$ where i is the largest integer such that $i \cdot n' \leq r$.

Consider now the clustering obtained from the threshold tree (T,k,φ'). At each node $v \in T$ such that $\varphi(v) = (j,r)$ and $\varphi'(v) = (j, i \cdot n')$, the points x of $\mathbf{X}$ that can be misplaced by the $\mathsf{Cut}_{\varphi'(u)}(\mathbf{X})$ are exactly the points such that $i \cdot n' < \mathbf{x}[i] \leq r$. This means that, if $\mathbf{Z}_u$ denotes the set of all the points x such that $i \cdot n' \leq \mathbf{x}[j] \leq (i+1)n'$, then $|\mathbf{Z}_u| \leq n'$ and the partitions $\mathsf{Cut}_{\varphi'(u)}(\mathbf{X} \setminus \mathbf{Z}_u)$ and $\mathsf{Cut}_{\varphi(u)}(\mathbf{X} \setminus \mathbf{Z}_u)$ are identical. Therefore, if $\mathbf{Z} = \bigcup_{u \in U} \mathbf{Z}_u$, then (T,k,φ') and (T,k,φ) induce the exact same clustering on $\mathbf{X} \setminus \mathbf{Z}$. Note that $|\mathbf{Z}| \leq kn' \leq \epsilon n$.

Therefore the algorithm will try all possible choices for T, and for every nonleaf node u, it tries all possible values for $\varphi'(u)$ of the form $(j, i \cdot n')$, where $j \in [d]$ and $i \in [\frac{2k}{\epsilon}]$. For each such try, the algorithm also removes the set $\mathbf{Z}$ consisting of all the points x such that $i \cdot n' \leq \mathbf{x}[j] \leq (i+1) \cdot n'$ whenever there exists $u \in U$ such that $\varphi'(u) = (j, i \cdot n')$ and computes the value of the clustering induces by (T,k,φ') on $\mathbf{X} \setminus \mathbf{Z}$. Finally it outputs the set $\mathbf{X} \setminus \mathbf{Z}$ as well as the threshold tree (T,k,φ') which minimises the value of the clustering.

Note that for every set of choices of T and $\varphi'(u)$ of the form $(j, i \cdot n')$, the set $\mathbf{Z}$ has size at most $k \cdot n' \leq \epsilon n$, which implies that the algorithm indeed outputs the desired set and threshold tree. Moreover, since there are at most 4^k possible trees T and $d \cdot \frac{2k}{\epsilon}$ possible choices of $\varphi'(u)$ for every node of the tree, we conclude that the running time of the algorithm is $(\frac{8dk}{\epsilon})^k n^{\mathcal{O}(1)}$. □

Conclusion

In this paper, we initiated the study of computational complexity of several variants of explainable clustering. Concluding, we would like to outline some further directions of research and state a number of open problems.

We showed that CLUSTERING EXPLANATION admits a polynomial-time approximation with a factor of $(k-1)$. Can this factor be improved in polynomial-time, say, to $\log k$? We proved that EXPLAINABLE k-MEANS and EXPLAINABLE k-MEDIAN can be solved in $n^{2d} \cdot (dn)^{\mathcal{O}(1)}$ time. Is this result tight? Or is it possible to obtain an $f(d) \cdot (dn)^{\mathcal{O}(1)}$ time algorithm for some function f? Also, is it possible to obtain approximation schemes parameterized by k, i.e., $(1+\varepsilon)$-approximation in $g(k,\varepsilon)(nd)^{\mathcal{O}(1)}$ time for some function g. Regular k-means/median admits such approximation schemes (Kumar, Sabharwal, and Sen 2010).

Acknowledgments

This research was supported by the Research Council of Norway via the project MULTIVAL (grant no. 263317) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819416). Kirill Simonov acknowledges support by the Austrian Science Fund (FWF) via project P31336 (New Frontiers for Parameterized Complexity). The research leading to these results has received funding from the Research Council of Norway via the project BWCA (grant no. 314528).

References

Aggarwal, C. C.; and Reddy, C. K., eds. 2013. *Data Clustering: Algorithms and Applications*. CRC Press.

Aloise, D.; Deshpande, A.; Hansen, P.; and Popat, P. 2009. NP-hardness of Euclidean sum-of-squares clustering. *Mach. Learn.*, 75(2): 245–248.

Bandyapadhyay, S.; Fomin, F.; Golovach, P.; Lochet, W.; Purohit, N.; and Simonov, K. 2021. How to Find a Good Explanation for Clustering? arXiv:2112.06580.

Bertsimas, D.; Orfanoudaki, A.; and Wiberg, H. M. 2021. Interpretable clustering: an optimization approach. *Mach. Learn.*, 110(1): 89–138.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8): 832.

Chakrabarty, D.; Goyal, P.; and Krishnaswamy, R. 2016. The Non-Uniform k-Center Problem. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, 67:1–67:15.

Charikar, M.; and Hu, L. 2021. Near-Optimal Explainable k-Means for All Dimensions. *CoRR*, abs/2106.15566.

Charikar, M.; Khuller, S.; Mount, D. M.; and Narasimhan, G. 2001. Algorithms for facility location problems with outliers. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, 642–651. Society for Industrial and Applied Mathematics.

Chen, K. 2008. A constant factor approximation algorithm for k-median clustering with outliers. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, 826–835.

Dasgupta, S. 2008. *The hardness of k-means clustering*. Department of Computer Science and Engineering, University of California.

Drineas, P.; Frieze, A. M.; Kannan, R.; Vempala, S. S.; and Vinay, V. 2004. Clustering Large Graphs via the Singular Value Decomposition. *Mach. Learn.*, 56(1-3): 9–33.

Esfandiari, H.; Mirrokni, V. S.; and Narayanan, S. 2021. Almost Tight Approximation Algorithms for Explainable Clustering. *CoRR*, abs/2107.00774.

Feng, J.; and Simon, N. 2017. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*.

Feng, Q.; Zhang, Z.; Huang, Z.; Xu, J.; and Wang, J. 2019. Improved Algorithms for Clustering with Outliers. In *30th International Symposium on Algorithms and Computation (ISAAC 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Fraiman, R.; Ghattas, B.; and Svarc, M. 2013. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7(2): 125–145.

Friggstad, Z.; Khodamoradi, K.; Rezapour, M.; and Salavatipour, M. R. 2019. Approximation Schemes for Clustering with Outliers. *ACM Trans. Algorithms*, 15(2).

Gamlath, B.; Jia, X.; Polak, A.; and Svensson, O. 2021. Nearly-Tight and Oblivious Algorithms for Explainable Clustering. *CoRR*, abs/2106.16147.

Geurts, P.; Touleimat, N.; Dutreix, M.; and d'Alché Buc, F. 2007. Inferring biological networks with output kernel trees. *BMC bioinformatics*, 8(2): 1–12.

Ghattas, B.; Michel, P.; and Boyer, L. 2017. Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. *Pattern Recognition*, 67: 177–185.

Harris, D. G.; Pensyl, T.; Srinivasan, A.; and Trinh, K. 2019. A lottery model for center-type problems with outliers. *ACM Transactions on Algorithms (TALG)*, 15(3): 1–25.

Hastie, T.; and Tibshirani, R. 1986. Generalized Additive Models. *Statistical Science*, 1(3): 297–318.

Impagliazzo, R.; Paturi, R.; and Zane, F. 2001. Which problems have strongly exponential complexity. *J. Computer and System Sciences*, 63(4): 512–530.

Krishnaswamy, R.; Li, S.; and Sandeep, S. 2018. Constant approximation for k-median and k-means with outliers via iterative rounding. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 646–659.

Kumar, A.; Sabharwal, Y.; and Sen, S. 2010. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2): 5:1–5:32.

Laber, E. S.; and Murtinho, L. 2021. On the price of explainability for some clustering problems. In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, 5915–5925.

Lipton, Z. C. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Lu, Y. Y.; Fan, Y.; Lv, J.; and Noble, W. S. 2018. DeepPINK: reproducible feature selection in deep neural networks. In *NeurIPS*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30: 4765–4774.

Mahajan, M.; Nimbhorkar, P.; and Varadarajan, K. R. 2012. The planar k-means problem is NP-hard. *Theor. Comput. Sci.*, 442: 13–21.

Makarychev, K.; and Shan, L. 2021. Near-Optimal Algorithms for Explainable k-Medians and k-Means. In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, volume 139, 7358–7367. PMLR.

Marcinkevičs, R.; and Vogt, J. E. 2020. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805*.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Moshkovitz, M.; Dasgupta, S.; Rashtchian, C.; and Frost, N. 2020. Explainable k-Means and k-Medians Clustering. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, 7055–7065. PMLR.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " 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*, 1135–1144.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 3145–3153. PMLR.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.

Ustun, B.; and Rudin, C. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3): 349–391.

Wang, F.; and Rudin, C. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*, 1013–1022. PMLR.