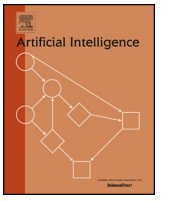
Artificial Intelligence 322 (2023) 103948

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



How to find a good explanation for clustering? ☆

Sayan Bandyapadhyay [a], Fedor V. Fomin [b], Petr A. Golovach [b,*], William Lochet [c], Nidhi Purohit [b], Kirill Simonov [d]

[a] Department of Computer Science, Portland State University, United States of America
[b] Department of Informatics, University of Bergen, Norway
[c] LIRMM, Université Montpellier, CNRS, Montpellier, France
[d] Hasso Plattner Institute, University of Potsdam, Germany

ARTICLE INFO

Article history:
Received 2 November 2022
Received in revised form 3 March 2023
Accepted 22 May 2023
Available online 1 June 2023

Keywords:
Explainable clustering
Clustering with outliers
Multivariate analysis

ABSTRACT

k-means and k-median clustering are powerful unsupervised machine learning techniques. However, due to complicated dependencies 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.

1. 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 researchers in recent years (see the surveys by Carvalho et al. [6] and Marcinkevičs and Vogt [39]). All these works can be divided into two main categories: *pre-modeling* [48,47,26,16,34] and *post-modeling*

☆ A preliminary version of this paper appeared as an extended abstract in the proceedings of AAAI 2022. The research leading to these results has been supported by the Research Council of Norway via the project BWCA (grant no. 314528), European Research Council (ERC) via grant LOPPRE, reference 819416, and DFG Research Group ADYN via grant DFG 411362735.

* Corresponding author at: University of Bergen, PB 7803, N-5020, Bergen, Norway.
E-mail addresses: sayanb@pdx.edu (S. Bandyapadhyay), fedor.fomin@uib.no (F.V. Fomin), petr.golovach@uib.no (P.A. Golovach), william.lochet@gmail.com (W. Lochet), kirillsimonov@gmail.com (K. Simonov).

https://doi.org/10.1016/j.artint.2023.103948

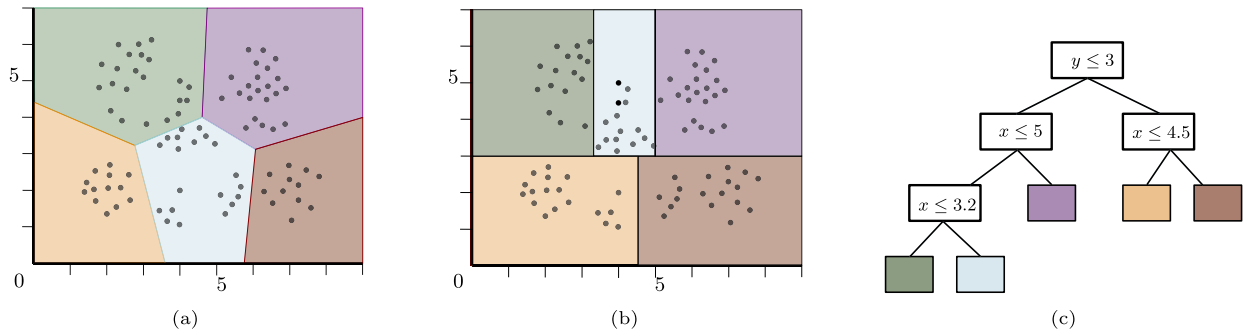


Fig. 1. (a) An example of an optimal solution to 5-means. (b) An explainable 5-means clustering and (c) the corresponding threshold tree. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

[43,45,4,46,35] explainability. While post-modeling explainability focuses on giving the 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-modeling explainability builds on decision trees [40,42]. In fact, a significant amount of work on explainable clustering is based on unsupervised decision trees [3,19,23,24,33,41]. In each node of the decision tree, the data is partitioned according to some feature's 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 [1] are often challenging to explain. On the other hand, the easily explainable models could output very costly clusterings. Subsequently, Moshkovitz et al. [41], 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. [41] 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 complicated 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. [41] 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.

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 [10,20,17,9,7,25,30]. 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 a 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. If we remove a small number of points (in Fig. 2b these are the 9 red larger points), then the explainable clustering is the same as the optimal clustering after removing those 9 points.

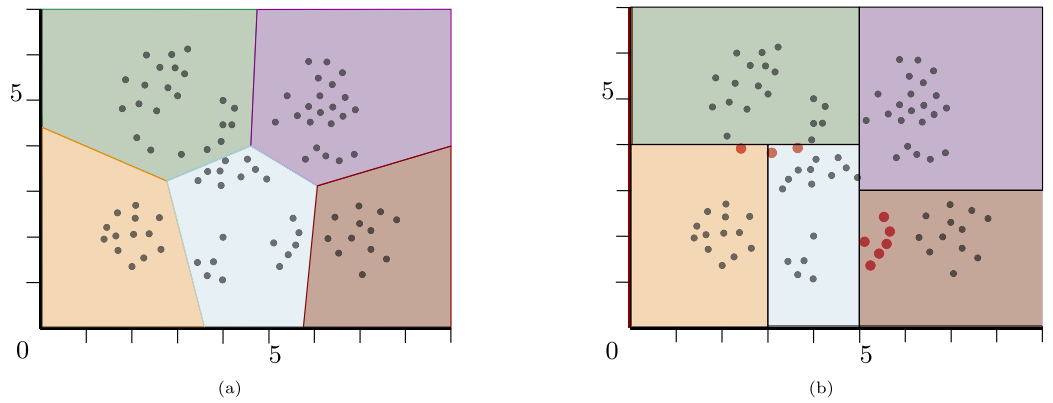


Fig. 2. (a) An optimal 5-clustering and (b) an explainable clustering that fits this clustering after removing the larger (red) points.

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 the 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 computable functions F and f, unless the Exponential Time Hypothesis (ETH) [28] fails. All these results appear in Section 3.

We also provide new insight into the computational complexity of the model of Moshkovitz et al. [41]. While the vanilla k-median and k-means problems are NP-hard for $k = 2$ [2,14,12] or $d = 2$ [36], 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 [41,8,15,22,32,37] 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 the 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 4.

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 4 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 $(\frac{4d(k+\epsilon)}{\epsilon})^k \cdot n^{\mathcal{O}(1)}$ 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 5. See Table 1 for a summary of all our results.

Related work A recent series of work [8,15,22,32,37,38] have further investigated the "cost of explainability", that is the ratio of the cost of the explainable clustering to the cost of an optimal clustering. The work of these papers has significantly improved the bound of $\mathcal{O}(k)$ for k-median and $\mathcal{O}(k^2)$ for k-means, which were shown by Moshkovitz et al. [41]. The state-of-the-art bound for k-median is $O(poly \log k)$ and for k-means $O(k \cdot poly \log k)$. We note that all of these works are based on the random sampling of orthogonal cuts and are different from the techniques we use. In another related work, Izza et al. [29] studied the redundancy in explanations provided by decision trees and gave efficient algorithms to eliminate such

Table 1
A summary of our results.

Model	Algorithms/Upper bounds	Hardness/Lower bounds
Clustering Explanation	$2^{2\min\{s,k\}}n^{2d}n^{\mathcal{O}(1)}$	No $F(s)$-approx. in $f(s)(nd)^{o(s)}$
	$(k-1)$-approximation	
	Reduction to $\mathcal{O}(sdk)$ points	
Explainable Clustering	$(4nd)^{k+\mathcal{O}(1)}\ n^{2d}\cdot n^{\mathcal{O}(1)}$	$f(k)\cdot n^{o(k)}$
Approximate Explainable Clustering	$(\frac{4d(k+\epsilon)}{\epsilon})^k\cdot n^{\mathcal{O}(1)}$	

redundancy. We note that investigation of such redundancy is beyond the scope of our work. Finally, we mention the recent survey of Cañete-Sifuentes, Monroy, and Medina-Pérez [5] about experimental results for multivariate decision trees.

2. Preliminaries

k-means/median Given a collection $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ 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 $\mathbf{C}_1,\ldots,\mathbf{C}_k$, called *clusters*, such that the *cost* of clustering is minimized. We follow the convention in the previous work [41] 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 $\{\mathbf{C}_1,\ldots,\mathbf{C}_k\}$ of $\mathbf{X}\subseteq\mathbb{R}^d$, its *k-means* (or simply means) cost is $\mathsf{cost}_2(\mathbf{C}_1,\ldots,\mathbf{C}_k)=\sum_{i=1}^k\mathsf{cost}_2(\mathbf{C}_i)$. 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 $\mathsf{cost}_1(\mathbf{C}_1,\ldots,\mathbf{C}_k)=\sum_{i=1}^k\mathsf{cost}_1(\mathbf{C}_i)$, 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 $i\in\{1,\ldots,d\}$. Let $\mathbf{X}$ be a collection of points of $\mathbb{R}^d$. For $i\in\{1,\ldots,d\}$ 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

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

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 $i\in\{1,\ldots,d\}$ 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 $i\in\{1,\ldots,d\}$ 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 $\{\mathbf{C}_1,\ldots,\mathbf{C}_k\}$ is an *explainable k-clustering* of a collection of points $\mathbf{X}\subseteq\mathbb{R}^d$ if $\mathbf{C}_1,\ldots,\mathbf{C}_k$ 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 $\varphi\colon U\to\{1,\ldots,d\}\times\mathbb{R}$, where U is the set of nonleaf nodes of T. For each node v of T, we define 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 set $(\mathbf{X}_u,\mathbf{X}_w)=\mathsf{Cut}_{\varphi(v)}(\mathbf{X})$. If v is a leaf, then $\mathbf{X}_v$ is a cluster. A clustering $\{\mathbf{C}_1,\ldots,\mathbf{C}_k\}$ 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 $\mathbf{C}_1,\ldots,\mathbf{C}_k$ 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. Note also that a full binary tree with k leaves has exactly $k-1$ nonleaf nodes.

For a collection $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ of n points and $i\in\{1,\ldots,d\}$, we denote by $\mathsf{coord}_i(\mathbf{X})$ the set of distinct values of i-th coordinates $\mathbf{x}_j[i]$ for $j\in\{1,\ldots,n\}$. 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.

Parameterized complexity and ETH A *parameterized problem* Π is a subset of $\Sigma^*\times\mathbb{N}$, where Σ is a finite alphabet. Thus, an instance of Π is a pair (I,k), where $I\subseteq\Sigma^*$ and k is a nonnegative integer called a *parameter*. A parameterized problem Π is *fixed-parameter tractable* (FPT) if it can be solved in $f(k)\cdot|I|^{\mathcal{O}(1)}$ time for some computable function $f(\cdot)$. Parameterized complexity theory also provides tools to refute the existence of an FPT algorithm for a parameterized problem. The standard way is to show that the considered problem is hard in the parameterized complexity classes W[1] or W[2]. We refer to the book [11] for the formal definitions of the parameterized complexity classes. The basic complexity assumption of the theory

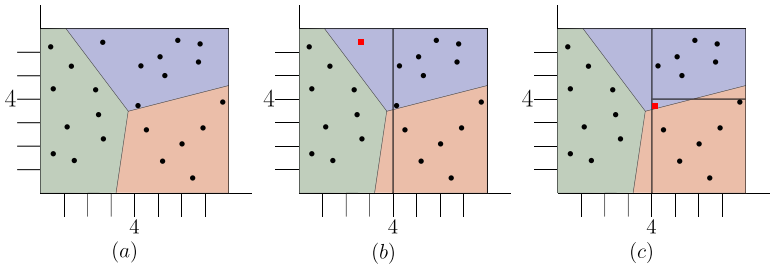


Fig. 3. (a) An example 3-means clustering. (b) The first (vertical) cut. (c) The second (horizontal) cut.

is that for the class FPT, formed by all parameterized fixed-parameter tractable problems, FPT $\subset$ W[1] $\subset$ W[2]. The hardness is proved by demonstrating a parameterized reduction from a problem known to be hard in the considered complexity class. A *parameterized reduction* is a many-one reduction that takes an input (I, k) of the first problem, and in $f(k)|I|^{\mathcal{O}(1)}$ time outputs an equivalent instance (I', k') of the second problem with $k' \leq g(k)$, where $f(\cdot)$ and $g(\cdot)$ are computable functions. Another way to obtain lower bounds is to use the *Exponential Time Hypothesis (ETH)* formulated by Impagliazzo, Paturi and Zane [27,28]. For an integer $k \geq 3$, let δ_k be the infimum of the real numbers c such that the k-Satisfiability problem can be solved in time $\mathcal{O}(2^{cn})$, where n is the number of variables. The Exponential Time Hypothesis states that $\delta_3 > 0$. In particular, ETH implies that k-Satisfiability cannot be solved in time $2^{o(n)}n^{\mathcal{O}(1)}$.

A *kernelization* (or *kernel*) for a parametrized problem Π is a polynomial time algorithm that, given an instance (I, k) of Π, outputs an instance (I', k') of Π such that (i) $(I, k) \in \Pi$ if and only if $(I', k') \in \Pi$ and (ii) $|I'| + k' \leq f(k)$ for a computable function $f(\cdot)$. The function $f(\cdot)$ is called the kernel *size*; a kernel is *polynomial* if $f(\cdot)$ is a polynomial. It can be shown that every decidable FPT problem admits a kernel. However, it is unlikely that all FPT problems have polynomial kernels. We refer to [11] and the recent book on kernelization of Fomin et al. [18] for details.

3. 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.

3.1. 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.

An illustration of the algorithm on an example with three clusters is shown in Fig. 3. Here any horizontal or vertical cut separates at least one point of a cluster. We select the first cut to be the vertical one with the coordinate value 4. W_1 contains only the square (red) point in Fig. 3(b). After this cut, $\mathbf{X}_1$ contains only one cluster, so we just return in that recursive branch. $\mathbf{X}_2$ still contains two clusters. So, we further divide the point set by choosing a horizontal cut with

coordinate value 4. This cut separates a point from the top cluster as shown in Fig. 3(c). Thus, in this case, W_1 is this square (red) point. After this cut, each part contains a single cluster, and thus we return on both branches. Hence, we obtain an explainable clustering by removing only 2 points.

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 arrive at 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 $w_m \leq |\widehat{W}|$. □

3.2. 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

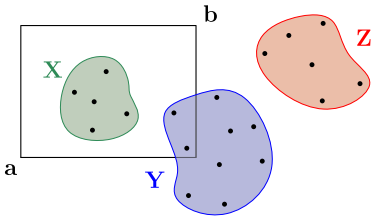


Fig. 4. The collection $\mathbf{X}$ is in the interval $(\mathbf{a}, \mathbf{b}]$, $\mathbf{Z}$ is outside $(\mathbf{a}, \mathbf{b}]$, and $(\mathbf{a}, \mathbf{b}]$ splits $\mathbf{Y}$. Subfamilies $\{\mathbf{X}, \mathbf{Y}\}$ and $\{\mathbf{X}\}$ are all $(\mathbf{a}, \mathbf{b}]$-proper subfamilies of $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$. $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ is 1-feasible with respect to $(\mathbf{a}, \mathbf{b}]$ since only $\mathbf{Y}$ is split by the interval.

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, maybe 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.

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

Before we move to the formal proof of the theorem, let us introduce some specific notations (illustrated by Fig. 4). Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ be such that $\mathbf{a} < \mathbf{b}$. We denote $(\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{a} < \mathbf{x} \leq \mathbf{b}\}$ and call $(\mathbf{a}, \mathbf{b}]$ an *interval*. For a collection of points $X \subseteq \mathbb{R}^d$, we say that X is *in* $(\mathbf{a}, \mathbf{b}]$ if $X \subseteq (\mathbf{a}, \mathbf{b}]$, X is *outside* $(\mathbf{a}, \mathbf{b}]$ if $X \cap (\mathbf{a}, \mathbf{b}] = \emptyset$, and we say that $(\mathbf{a}, \mathbf{b}]$ *splits* X if $X \cap (\mathbf{a}, \mathbf{b}] \neq \emptyset$ and $X \setminus (\mathbf{a}, \mathbf{b}] \neq \emptyset$.

Let $\mathcal{X}$ be a family of disjoint collections of points of $\mathbb{R}^d$. A subfamily $\mathcal{Y} \subseteq \mathcal{X}$ is said to be $(\mathbf{a}, \mathbf{b}]$-*proper* if (i) every $X \in \mathcal{X}$ that is in $(\mathbf{a}, \mathbf{b}]$ is in $\mathcal{Y}$, and (ii) every $X \in \mathcal{X}$ that is outside $(\mathbf{a}, \mathbf{b}]$ is not included in $\mathcal{Y}$. Note that $X \subseteq \mathcal{X}$ that are split by $(\mathbf{a}, \mathbf{b}]$ may be either in $\mathcal{Y}$ or not in $\mathcal{Y}$. The *truncation* of $\mathcal{X}$ with respect to $(\mathbf{a}, \mathbf{b}]$, is the family

$$\mathrm{tr}_{(\mathbf{a},\mathbf{b}]}(\mathcal{X}) = \{X \cap (\mathbf{a}, \mathbf{b}] \mid X \in \mathcal{X} \text{ s.t. } X \cap (\mathbf{a}, \mathbf{b}] \neq \emptyset\}.$$

For an integer $s \geq 0$, $\mathcal{X}$ is *s-feasible* with respect to $(\mathbf{a}, \mathbf{b}]$ if $(\mathbf{a}, \mathbf{b}]$ splits at most s collections in $\mathcal{X}$.

Proof of Theorem 2. 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 $\mathbf{X} = \bigcup_{i=1}^{k} \mathbf{C}_i$. 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 $i \in \{1, \ldots, d\}$. Recall that it suffices to consider only canonical cuts for the construction of a decision tree; canonical vectors represent sequences of canonical cuts along various dimensions.

For every pair of canonical vectors $(\mathbf{a}, \mathbf{b})$ such that $\mathbf{a} < \mathbf{b}$ and $\mathcal{C}$ is s-feasible with respect to $(\mathbf{a}, \mathbf{b}]$, and every $(\mathbf{a}, \mathbf{b}]$-proper $\mathcal{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_\ell\} \subseteq \mathcal{C}$, we denote by $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S})$ the minimum size of a collection of points $\mathbf{W} \subseteq \mathbf{X} \cap (\mathbf{a}, \mathbf{b}]$ such that $\{\mathbf{S}'_1 \setminus \mathbf{W}, \ldots, \mathbf{S}'_\ell \setminus \mathbf{W}\}$, where $\{\mathbf{S}'_1, \ldots, \mathbf{S}'_\ell\} = \mathrm{tr}_{(\mathbf{a},\mathbf{b}]}(\mathcal{S})$, is an explainable ℓ-clustering. We assume that $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) = 0$ if $\mathcal{S}$ is empty. Intuitively, the tuple $(\mathbf{a}, \mathbf{b}, \mathcal{S})$ defines a subproblem in the dynamic programming, bounded by the box $(\mathbf{a}, \mathbf{b}]$ in which the clusters of $\mathcal{S}$ have to be explained. That is why $\mathcal{S}$ is $(\mathbf{a}, \mathbf{b}]$-proper, as it necessarily contains all clusters in $(\mathbf{a}, \mathbf{b}]$, and some clusters that are split by $(\mathbf{a}, \mathbf{b}]$; those split clusters that are part of $\mathcal{S}$ are assumed to be contained in $(\mathbf{a}, \mathbf{b}]$ after removing the extra points. Clearly, $\mathcal{S}$ should be s-feasible with respect to $(\mathbf{a}, \mathbf{b}]$, as otherwise simply accounting for clusters split by $(\mathbf{a}, \mathbf{b}]$ already exceeds the budget of s.

Now, instead of $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S})$, we compute the following slightly different value

$$w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = \omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) + \sum_{\mathbf{C}_i \in \mathcal{S}} |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_i \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}]|. \tag{2}$$

Observe that $w(\mathbf{a}, \mathbf{b}, \mathcal{S})$ additionally accounts for the cost of removing extra points from clusters that are split by $(\mathbf{a}, \mathbf{b}]$, while $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S})$ only computes the cost of explanation for the truncation of $\mathcal{S}$ in $(\mathbf{a}, \mathbf{b}]$. Since we are interested only in clustering that can be obtained by deleting at most s points, we assume that $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = +\infty$ if this value is bigger than s. This slightly informal definition simplifies arguments. In particular, observe the two sums in (2) give a value that is bigger than s if $\mathcal{S}$ is not s-feasible with respect to $(\mathbf{a}, \mathbf{b}]$. In fact, this is the reason why these sums are included in (2).

Notice that $(\mathcal{C}, s)$ is a yes-instance of Clustering Explanation if and only if $w(\mathbf{a}^*, \mathbf{b}^*, \mathcal{C}) \leq s$, where $\mathbf{a}^*[i] = -\infty$ and $\mathbf{b}^*[i] = +\infty$ for $i \in \{1, \ldots, d\}$.

We now proceed to define the computation. The values $w(\mathbf{a}, \mathbf{b}, \mathcal{S})$ are computed in different ways depending on $\ell = |\mathcal{S}|$. If $\ell = 0$, that is, $\mathcal{S} = \emptyset$, then $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) = 0$ and $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = \sum_{\mathbf{C}_j \in \mathcal{C}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]|$. If $\ell = 1$, then $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) = 0$ by definition. Then $\mathcal{S} = \{\mathbf{C}_i\}$ for some $i \in \{1, \ldots, k\}$ such that $\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}] \neq \emptyset$ and

$$w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{C} \setminus \{\mathbf{C}_i\}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]|.$$

Assume that $\ell \geq 2$, and the values of $\omega(\mathbf{a}', \mathbf{b}', \mathcal{S}')$ are computed for $|\mathcal{S}'| < \ell$.

For $i \in \{1, \ldots, d\}$ 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

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

That is, $(\mathbf{a}, \mathbf{b}^{i,\theta}]$ and $(\mathbf{a}^{i,\theta}, \mathbf{b}]$ are the two intervals formed by splitting $(\mathbf{a}, \mathbf{b}]$ along the cit (i, θ). We also say that (i, θ) is *s-feasible* if $\mathcal{C}$ is s-feasible with respect to $(\mathbf{a}, \mathbf{b}^{i,\theta}]$ and $(\mathbf{a}^{i,\theta}, \mathbf{b}]$. For an s-feasible (i, θ), a partition $(\mathcal{S}_1, \mathcal{S}_2)$ of $\mathcal{S}$ is (i, θ)*-proper* if $\mathcal{S}_1$ and $\mathcal{S}_2$ are $(\mathbf{a}, \mathbf{b}^{i,\theta}]$ and $(\mathbf{a}^{i,\theta}, \mathbf{b}]$-proper, respectively. We define

$$\delta_{i,\theta}(\mathcal{S}) = \sum_{\substack{\mathbf{C}_i \in \mathcal{S} \colon \mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}^{i,\theta}] \neq \emptyset \\ \text{and } \mathbf{C}_i \cap (\mathbf{a}^{i,\theta}, \mathbf{b}] \neq \emptyset}} |\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}]|,$$

i.e., the total size of the clusters of $\mathcal{C}$ split by the cut (i, θ) in $(\mathbf{a}, \mathbf{b}]$.

We compute $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S})$ by the following recurrence.

$$w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = \min\{(**), (***)\}, \tag{3}$$

where the right part is denoted by $(*)$, and

$$(**) = \min\{|\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{S} \setminus \{C_i\}} |\mathbf{C}_j| + \sum_{\mathbf{C}_j \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]| \mid \mathbf{C}_i \in \mathcal{S}\},$$

$$(***) = \min\{w(\mathbf{a}, \mathbf{b}^{i,\theta}, \mathcal{S}_1) + w(\mathbf{a}^{i,\theta}, \mathbf{b}, \mathcal{S}_2) - \delta_{i,\theta}(\mathcal{S}) \text{ for } 1 \leq i \leq d,\ \theta \in \mathsf{coord}_i(\mathbf{X}),$$
$$(\mathcal{S}_1, \mathcal{S}_2) \text{ is partition of } \mathcal{S} \text{ s.t. } \mathbf{a}[i] < \theta < \mathbf{b}[i], (i, \theta) \text{ is } s\text{-feasible}, (\mathcal{S}_1, \mathcal{S}_2) \text{ is } (i, \theta)\text{-proper}\}.$$

We assume that $(***) = +\infty$ if there is no triple $(i, \theta, (\mathcal{S}_1, \mathcal{S}_2))$ satisfying the conditions in the definition of the set. We also assume that $(*) = +\infty$ if its value proves to be bigger than s. Intuitively, the term $(***)$ corresponds to trying all feasible cuts at the top level of the subproblem and reducing to the two respective smaller subproblems. The term $(**)$ then additionally covers the case where all except one of the remaining clusters are empty in the solution, which is not covered by $(***)$ since the following cuts do not play a role in this case.

We now prove the correctness of (3) by showing the inequalities between the left and the rights parts in both directions.

First, we show that $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) \geq (*)$. This is trivial if $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = +\infty$. Assume that this is not the case. Then by our assumption, $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) \leq s$. Recall that $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = \omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) + \sum_{\mathbf{C}_i \in \mathcal{S}} |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_i \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}]|$. Let $r = \omega(\mathbf{a}, \mathbf{b}, \mathcal{S})$ and let $\mathbf{W} \subseteq \mathbf{X} \cap (\mathbf{a}, \mathbf{b}]$ be a collection of r points such that $\mathcal{S}' = \{\mathbf{S}'_1 \setminus \mathbf{W}, \ldots, \mathbf{S}'_\ell \setminus \mathbf{W}\}$, where $\{\mathbf{S}'_1, \ldots, \mathbf{S}'_\ell\} = \mathsf{tr}_{(\mathbf{a}, \mathbf{b}]}(\mathcal{S})$, is an explainable ℓ-clustering. Assume that $\mathcal{S} = \{\mathbf{C}_{i_1}, \ldots, \mathbf{C}_{i_\ell}\}$ and $\mathbf{S}'_j = \mathbf{C}_{i_j} \cap (\mathbf{a}, \mathbf{b}]$. Let $\mathbf{W}_i = \mathbf{W} \cap \mathbf{S}'_i$ for $i \in \{1, \ldots, \ell\}$.

Notice that it may happen that $\mathbf{W}_j = \mathbf{S}'_j$ for some $j \in \{1, \ldots, \ell\}$. Then $\mathbf{C}_{i_j} \cap (\mathbf{a}, \mathbf{b}] = \mathbf{W}_j$. Observe, however, that $\mathbf{C}_{i_j} \cap (\mathbf{a}, \mathbf{b}] = \mathbf{W}_j$ for at most $\ell - 1$ values of j. Suppose that there is $h \in \{1, \ldots, \ell\}$ such that $\mathbf{C}_{i_h} \cap (\mathbf{a}, \mathbf{b}] \neq \mathbf{W}_h$ and $\mathbf{C}_{i_j} \cap (\mathbf{a}, \mathbf{b}] = \mathbf{W}_j$ for every $h \in \{1, \ldots, \ell\}$ such that $j \neq h$. In this case, we obtain that

$$\omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) = \sum_{\mathbf{C}_j \in \mathcal{S} \setminus \{C_{i_h}\}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]|$$

and

$$w(\mathbf{a}, \mathbf{b}, \mathcal{S}) = |\mathbf{C}_{i_h} \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{S} \setminus \mathbf{C}_{i_h}} |\mathbf{C}_j| + \sum_{\mathbf{C}_j \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]|.$$

Then $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) \geq (**) \geq (*)$. Assume from now that this is not the case and $\mathbf{S}_j \setminus \mathbf{W}_j \neq \emptyset$ for at least two distinct indices $j \in \{1, \ldots, \ell\}$. Then we show that $w(\mathbf{a}, \mathbf{b}, \mathcal{S}) \geq (***)$.

Because we separate at least two nonempty collections of points, the definition of explainable clustering implies that there are $i \in \{1, \ldots, d\}$ and $\theta \in \mathsf{coord}_i(\mathbf{X})$, such that $\mathbf{a}[i] < \theta < \mathbf{b}[i]$ and there is a partition (I_1, I_2) of $\{1, \ldots, \ell\}$ with the property that (i) $\hat{\mathcal{S}}_1 = \{\mathbf{S}'_i \setminus \mathbf{W}_j \mid j \in I_1\}$ is an explainable $\ell_1 = |I_1|$-clustering with the clusters in $(\mathbf{a}, \mathbf{b}^{i,\theta}]$, and (ii) $\hat{\mathcal{S}}_2 = \{\mathbf{S}'_j \setminus \mathbf{W}_j \mid j \in I_2\}$ is an explainable $\ell_2 = |I_2|$-clustering with the clusters in $(\mathbf{a}^{i,\theta}, \mathbf{b}]$, where both $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ contain nonempty collections of points. Moreover, we assume that if $\mathbf{S}_j \setminus \mathbf{W}_j = \emptyset$ for some $j \in \{1, \ldots, \ell\}$, then $\mathbf{S}_j \setminus \mathbf{W}_j$ is placed in $\hat{\mathcal{S}}_1$ if $\mathbf{S}_j$ has a point in $(\mathbf{a}, \mathbf{b}^{i,\theta}]$ and, otherwise, i.e. if $\mathbf{S}_j$ has only points in $(\mathbf{a}^{i,\theta}, \mathbf{b}]$, it is placed in $\hat{\mathcal{S}}_2$.

We define $\mathcal{S}_1 = \{\mathbf{C}_{i_j} \mid j \in I_1\}$ and $\mathcal{S}_2 = \{\mathbf{C}_{i_j} \mid j \in I_2\}$. For $j \in I_1$, let $\mathbf{W}^1_j = (\mathbf{a}, \mathbf{b}^{i,\theta}] \cap \mathbf{W}_j$, and let $\mathbf{W}^2_j = (\mathbf{a}^{i,\theta}, \mathbf{b}] \cap \mathbf{W}_j$. We set $\mathbf{W}^1 = \bigcup_{j \in I_1} \mathbf{W}^1_j$ and $\mathbf{W}^2 = \bigcup_{j \in I_2} \mathbf{W}^2_j$. Let also $\mathbf{R}_1 = (\mathbf{a}^{i,\theta}, \mathbf{b}] \cap \left(\bigcup_{j \in I_1} \mathbf{W}_j\right)$ and $\mathbf{R}_2 = (\mathbf{a}, \mathbf{b}^{i,\theta}] \cap \left(\bigcup_{j \in I_2} \mathbf{W}_j\right)$. Observe that $(\mathbf{W}^1, \mathbf{R}_1, \mathbf{W}^2, \mathbf{R}_2)$ is a partition of $\mathbf{W}$ where some sets may be empty. Denote by $w_1 = |\mathbf{W}^1|$ and $w_2 = |\mathbf{W}^2|$, and let $r_1 = |\mathbf{R}_1|$ and $r_2 = |\mathbf{R}_2|$. Clearly, $w_1 + w_2 + r_1 + r_2 = r$.

Notice that $\mathcal{S}'_1 = \{\mathbf{S}'_j \setminus \mathbf{R}_1 \mid j \in I_1\} = \mathrm{tr}_{(\mathbf{a},\mathbf{b}^{i,\theta}]}(\mathcal{S}_1)$ and $\mathcal{S}'_2 = \{\mathbf{S}'_j \setminus \mathbf{R}_2 \mid j \in I_2\} = \mathrm{tr}_{(\mathbf{a}^{i,\theta},\mathbf{b}]}(\mathcal{S}_2)$. Also, $\mathcal{S}_1$ and $\mathcal{S}_2$ are $(\mathbf{a},\mathbf{b}^{i,\theta}]$ and $(\mathbf{a}^{i,\theta},\mathbf{b}]$-proper, respectively. Furthermore, for $h \in \{1,2\}$, $\hat{\mathcal{S}}_h$ is obtained from $\mathcal{S}'_h$ by deleting the points of $\mathbf{W}^1$ from the clusters. Also we have that

$$\sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}^{i,\theta}]| = \sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}]| + |\mathbf{R}_1|, \tag{4}$$

$$\sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a}^{i,\theta},\mathbf{b}]| = \sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}]| + |\mathbf{R}_2|, \tag{5}$$

and

$$\sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}^{i,\theta}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_2} |\mathbf{C}_j \cap (\mathbf{a}^{i,\theta},\mathbf{b}]| = \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}]| + |\mathbf{R}_1| + |\mathbf{R}_2|. \tag{6}$$

Note also that

$$\delta_{i,\theta}(\mathcal{S}) = |\mathbf{R}_1| + |\mathbf{R}_2|. \tag{7}$$

Then by (4)–(7),

$$\begin{aligned}
w(\mathbf{a},\mathbf{b},\mathcal{S}) &= (w_1 + w_2 + r_1 + r_2) + \sum_{\mathbf{C}_j \in \mathcal{S}} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}]| \\
&= (w_1 + w_2 + r_1 + r_2) + \sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}]| \\
&\quad + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}^{i,\theta}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_2} |\mathbf{C}_j \cap (\mathbf{a}^{i,\theta},\mathbf{b}]| - 2r_1 - 2r_2 \\
&= w_1 + \sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}^{i,\theta}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}^{i,\theta}]| \\
&\quad + w_2 + \sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a}^{i,\theta},\mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_2} |\mathbf{C}_j \cap (\mathbf{a}^{i,\theta},\mathbf{b}]| - \delta_{i,\theta}(\mathcal{S}).
\end{aligned} \tag{8}$$

Recall that $w(\mathbf{a},\mathbf{b},\mathcal{S}) \le s$. Note that

$$\sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}^{i,\theta}]| \le \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}]| + |\mathbf{R}_2|.$$

Using (4), we obtain that

$$\sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}^{i,\theta}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}^{i,\theta}]| \le w(\mathbf{a},\mathbf{b},\mathcal{S}),$$

which is at most s. This means that $\mathcal{S}_1$ is $(\mathbf{a},\mathbf{b}^{i,\theta}]$-proper. Similarly, we have that $\mathcal{S}_2$ is $(\mathbf{a}^{i,\theta},\mathbf{b}]$-proper. Therefore,

$$w_1 + \sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a},\mathbf{b}^{i,\theta}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}^{i,\theta}]| \ge w(\mathbf{a},\mathbf{b}^{i,\theta},\mathcal{S}_1)$$

and

$$w_2 + \sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a}^{i,\theta},\mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}_2} |\mathbf{C}_j \cap (\mathbf{a}^{i,\theta},\mathbf{b}]| \ge w(\mathbf{a}^{i,\theta},\mathbf{b},\mathcal{S}_2).$$

This allows us to extend (8) and conclude that

$$w(\mathbf{a},\mathbf{b},\mathcal{S}) \ge w(\mathbf{a},\mathbf{b}^{i,\theta},\mathcal{S}_1) + w(\mathbf{a}^{i,\theta},\mathbf{b},\mathcal{S}_2) \ge (***).$$

This shows that $w(\mathbf{a},\mathbf{b},\mathcal{S}) \ge (***) \ge (*)$ and concludes the proof of the first inequality.

Now we show that $w(\mathbf{a},\mathbf{b},\mathcal{S}) \le (*)$. The inequality is trivial if $(*) = +\infty$. Suppose that this is not the case. Then, by our assumption about assigning the value to $(*)$, $(*) \le s$.

Suppose that the minimum in $(*)$ is achieved in the first part, that is, $(**) \le (***)$. Then

$$(*) = |\mathbf{C}_i \setminus (\mathbf{a},\mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{S}\setminus\{C_i\}} |\mathbf{C}_j| + \sum_{\mathbf{C}_j \in \mathcal{C}\setminus\mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a},\mathbf{b}]|$$

for some $\mathbf{C}_i \in \mathcal{S}$. We define $\mathbf{W} = \bigcup_{\mathbf{C}_j \in \mathcal{S} \setminus \{\mathbf{C}_i\}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]|$. We have that the family obtained from $tr_{(\mathbf{a},\mathbf{b}]}(\mathcal{S})$ by the deletion of the points of $\mathbf{W}$ is an explainable ℓ-clustering because we deleted the points in $(\mathbf{a}, \mathbf{b}]$ of every $\mathbf{C}_j \in \mathcal{S}$ excepts $\mathbf{C}_i$. This implies that $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) \leq |\mathbf{W}|$ and we obtain that

$$w(\mathbf{a}, \mathbf{b}, \mathcal{S}) \leq |\mathbf{W}| + \sum_{\mathbf{C}_i \in \mathcal{S}} |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_i \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}]| = |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_j \in \mathcal{S} \setminus \{C_i\}} |\mathbf{C}_j| + \sum_{\mathbf{C}_j \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]|,$$

which is exactly $(**)$, so $\omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) \leq (*)$.

Assume from now that the minimum in $(*)$ is achieved for the second part, that is, $(*) = (***) < (**)$. Suppose that $i \in \{1, \ldots, d\}$, $\theta \in \mathsf{coord}_i(\mathbf{X})$ where $\mathbf{a}[i] \leq \theta \leq \mathbf{b}[i]$ and (i, θ) is s-feasible, and a partition $(\mathcal{S}_1, \mathcal{S}_2)$ of $\mathcal{S}$ that is (i, θ)-proper are chosen in such a way that $(**)$ achieves the minimum value for them, that is, $(***) = w(\mathbf{a}, \mathbf{b}^{i,\theta}, \mathcal{S}_1) + w(\mathbf{a}^{i,\theta}, \mathbf{b}, \mathcal{S}_2) - \delta_{i,\theta}(\mathcal{S})$.

Let $\mathbf{R}_1 = (\mathbf{a}^{i,\theta}, \mathbf{b}] \cap \left(\bigcup_{\mathbf{C}_j \in \mathcal{S}_1} \mathbf{C}_j\right)$ and $\mathbf{R}_2 = (\mathbf{a}, \mathbf{b}^{i,\theta}] \cap \left(\bigcup_{\mathbf{C}_j \in \mathcal{S}_2} \mathbf{C}_j\right)$. Then we obtain that

$$\sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a}, \mathbf{b}^{i,\theta}]| = \sum_{\mathbf{C}_j \in \mathcal{S}_1} |\mathbf{C}_j \setminus (\mathbf{a}, \mathbf{b}]| + |\mathbf{R}_1|, \tag{9}$$

$$\sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a}^{i,\theta}, \mathbf{b}]| = \sum_{\mathbf{C}_j \in \mathcal{S}_2} |\mathbf{C}_j \setminus (\mathbf{a}, \mathbf{b}]| + |\mathbf{R}_2|, \tag{10}$$

$$\sum_{\mathbf{C}_j \in \mathcal{C} \setminus \mathcal{S}_1} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}^{i,\theta}]| + \sum_{\mathbf{C}_j \in \mathcal{C} \setminus \mathcal{S}_2} |\mathbf{C}_j \cap (\mathbf{a}^{i,\theta}, \mathbf{b}]| = \sum_{\mathbf{C}_j \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_j \cap (\mathbf{a}, \mathbf{b}]| + |\mathbf{R}_1| + |\mathbf{R}_2|, \tag{11}$$

and

$$\delta_{i,\theta}(\mathcal{S}) = |\mathbf{R}_1| + |\mathbf{R}_2|. \tag{12}$$

Let $w_1 = \omega(\mathbf{a}, \mathbf{b}^{i,\theta}, \mathcal{S}_1)$ and $w_2 = \omega(\mathbf{a}^{i,\theta}, \mathbf{b}, \mathcal{S}_2)$. Then there is a collection $\mathbf{W}_1 \subseteq (\mathbf{a}, \mathbf{b}^{i,\theta}] \cap \mathbf{X}$ such that the collection of sets obtained from the collections of $\mathrm{tr}_{(\mathbf{a},\mathbf{b}^{i,\theta}]}(\mathcal{S}_1)$ by the deletions of the points of $\mathbf{W}_1$ is an explainable $|\mathcal{S}_1|$-clustering. Similarly, there is a collection $\mathbf{W}_2 \subseteq (\mathbf{a}^{i,\theta}, \mathbf{b}] \cap \mathbf{X}$ such that the family of collections obtained from the collections of $\mathrm{tr}_{(\mathbf{a}^{i,\theta},\mathbf{b}]}(\mathcal{S}_2)$ by the deletions of the points of $\mathbf{W}_2$ is an explainable $|\mathcal{S}_2|$-clustering. Consider $\mathbf{W} = \mathbf{W}_1 \cup \mathbf{W}_2 \cup \mathbf{R}_1 \cup \mathbf{R}_2$. The crucial observation is that the family obtained from the collections of $\mathrm{tr}_{(\mathbf{a},\mathbf{b}]}(\mathcal{S})$ by the deletions of the points of $\mathbf{W}$ is an explainable ℓ-clustering, where the first cut is $\mathsf{Cut}_{i,\theta}$. Using (9)–(12), we obtain that

$$\begin{aligned}(***) = w(\mathbf{a}, \mathbf{b}^{i,\theta}, \mathcal{S}_1) + w(\mathbf{a}^{i,\theta}, \mathbf{b}, \mathcal{S}_2) - \delta_{i,\theta}(\mathcal{S}) &= |\mathbf{W}| + \sum_{\mathbf{C}_i \in \mathcal{S}} |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_i \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}]| \\ &\geq \omega(\mathbf{a}, \mathbf{b}, \mathcal{S}) + \sum_{\mathbf{C}_i \in \mathcal{S}} |\mathbf{C}_i \setminus (\mathbf{a}, \mathbf{b}]| + \sum_{\mathbf{C}_i \in \mathcal{C} \setminus \mathcal{S}} |\mathbf{C}_i \cap (\mathbf{a}, \mathbf{b}]| = w(\mathbf{a}, \mathbf{b}, \mathcal{S}).\end{aligned}$$

Since $(*) = (***)$, this completes the correctness proof of the recurrence (3).

In the final stage of the proof, we evaluate the running time. We construct the table of values of $w(\mathbf{a}, \mathbf{b}, \mathcal{S})$ for pairs $(\mathbf{a}, \mathbf{b})$ of canonical vectors such that $\mathbf{a} < \mathbf{b}$. The total number of such pairs is at most $(n+2)^{2d}$ and they can be constructed in $n^{2d} \cdot (dn)^{\mathcal{O}(1)}$ time. We are interested only in $\mathbf{a}$ and $\mathbf{b}$ such that $\mathcal{C}$ is s-feasible with respect to $(\mathbf{a}, \mathbf{b}]$. Clearly, for given $\mathbf{a}$ and $\mathbf{b}$, the s-feasibility can be checked in $(dn)^{\mathcal{O}(1)}$ time. If $\mathcal{C}$ is s-feasible with respect to $(\mathbf{a}, \mathbf{b}]$, then all $(\mathbf{a}, \mathbf{b}]$-proper subfamilies $\mathcal{S}$ of $\mathcal{C}$ can be listed by brute force as follows. Observe that $\mathcal{X} = \{\mathbf{C}_i \in \mathcal{C} \mid \mathbf{C}_i \text{ is in } (\mathbf{a}, \mathbf{b}]\}$ that can be constructed in polynomial time is a subfamily of every $(\mathbf{a}, \mathbf{b}]$-proper $\mathcal{S}$. Let $\mathcal{Y} = \{\mathbf{C}_i \in \mathcal{C} \mid \mathbf{C}_i \text{ is split by } (\mathbf{a}, \mathbf{b}]\}$. Since $\mathcal{C}$ is s-feasible, $|\mathcal{Y}| \leq \min\{s, k\}$. We can construct $\mathcal{Y}$ in polynomial time and then generate at most $2^{\min\{s,k\}}$ subfamilies $\mathcal{Z}$ of $\mathcal{Y}$ in total $2^{\min\{s,k\}} \cdot (dn)^{\mathcal{O}(1)}$ time. Then the $(\mathbf{a}, \mathbf{b}]$-proper subfamilies $\mathcal{S}$ are exactly the families of the form $\mathcal{Z} \cup \mathcal{X}$. We obtain that there are at most $2^{\min\{s,k\}}$ $(\mathbf{a}, \mathbf{b}]$-proper subfamilies that can be generated in time $2^{\min\{s,k\}} \cdot (dn)^{\mathcal{O}(1)}$ time. Then we conclude that the dynamic programming algorithm computes at most $2^{\min\{s,k\}} \cdot n^{2d}$ values of $w(\mathbf{a}, \mathbf{b}, \mathcal{S})$.

The value of $w(\mathbf{a}, \mathbf{b}, \mathcal{S})$ for $|\mathcal{S}|$ is constructed in $(dn)^{\mathcal{O}(1)}$ time if $\ell \leq 1$. If $\ell \geq 2$, we are using the recurrence (3). Computing $(**)$ can be done in polynomial time. To compute $(***)$, we go through $i \in \{1, \ldots, d\}$, $\theta \in \mathsf{coord}_i(\mathbf{X})$, and partitions $(\mathcal{S}_1, \mathcal{S})$ of $\mathcal{S}$, where $(\mathcal{S}_1, \mathcal{S}_2)$ is required to be (i, θ)-proper. This implies that $(***)$ can be computed in $2^{\min\{s,k\}} \cdot (dn)^{\mathcal{O}(1)}$ time. Summarizing, we have that the total running time of the dynamic programming algorithm is $2^{2\min\{s,k\}} \cdot n^{2d} \cdot (dn)^{\mathcal{O}(1)}$. This concludes the proof. □

3.3. Data reduction

In this section, we prove that Clustering Explanation admits a polynomial kernel when parameterized by $k+d+s$. Thus, we show that an instance of Clustering Explanation can be effectively compressed to an equivalent one of size polynomial in k, d, and s, irrespective of the original number of points in the instance.

Theorem 3. Clustering Explanation *parameterized by k, d and s admits a kernel with* $r = 2(s+1)dk$ *points of* $\{1,\dots,r\}^d$.

Proof. Let $(\mathcal{C}, s)$ be an instance of Clustering Explanation, where $\mathcal{C} = \{\mathbf{C}_1,\dots,\mathbf{C}_k\}$ 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 $i \in \{1,\dots,k\}$ and every $j \in \{1,\dots,d\}$, do the following:

- Order the points of $\mathbf{C}_i$ in increasing order 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 $i \in \{1,\dots,k\}$. Then we consider the instance $(\mathcal{S}, s)$ of Clustering Explanation, where $\mathcal{S} = \{\mathbf{S}_1,\dots,\mathbf{S}_k\}$. We show the following claim.

Claim 3.1. $(\mathcal{C}, s)$ *is a yes-instance of* Clustering Explanation *if and only if* $(\mathcal{S}, s)$ *is a yes-instance.*

Proof of Claim 3.1. Trivially, if $(\mathcal{C}, s)$ is a yes-instance, then $(\mathcal{S}, s)$ is a yes-instance because we just deleted some points 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 $\{\mathbf{S}_1 \setminus \mathbf{W},\dots,\mathbf{S}_k \setminus \mathbf{W}\}$ 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 $\mathbf{S}_1 \setminus \mathbf{W},\dots,\mathbf{S}_k \setminus \mathbf{W}$ 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 $\mathbf{C}_1 \setminus \mathbf{W},\dots,\mathbf{C}_k \setminus \mathbf{W}$ correspond to the leaves.

The proof is by contradiction. Assume that at least one collection of points corresponding to a leaf is distinct from every $\mathbf{C}_1 \setminus \mathbf{W},\dots,\mathbf{C}_k \setminus \mathbf{W}$. Then there is a node $v \in V(T)$ such that for some $j \in \{1,\dots,k\}$, $\mathbf{C}_j \setminus \mathbf{W}$ 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 $\mathbf{C}_1 \setminus \mathbf{W},\dots,\mathbf{C}_k \setminus \mathbf{W}$ 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 $\{1,\dots,\ell\}^d$. For this, we observe that for each $i \in \{1,\dots,d\}$, 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 $\mathsf{coord}_i(\mathbf{Y}) = \{\theta_1^i,\dots,\theta_{r_i}^i\}$, where $\theta_1^i < \cdots < \theta_{r_i}^i$. For every $\mathbf{y} \in \mathbf{Y}$, we construct a point $\mathbf{z}$, by setting $\mathbf{z}[i] = j$, where $\theta_j^i = \mathbf{y}[i]$, 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. □

3.4. Hardness of approximation

In Theorem 1, we proved that the optimization Clustering Explanation problem admits a polynomial $(k-1)$-approximation. Here we show that it is hard to approximate the minimum number of deleted points s within a factor that depends on s. Specifically, we provide a parameter-preserving reduction from Hitting Set to Clustering Explanation that transfers known results about the hardness of parameterized approximation for the Hitting Set problem to Clustering Explanation. Recall that in the decision version of the Hitting Set problem, the input is a family of sets $\mathcal{A}$ over a universe U together with a parameter ℓ, and the goal is to decide whether there is a set $H \subseteq U$ of size at most ℓ such that H has non-empty intersection with each set in $\mathcal{A}$. In the parameterized approximation variant, where ℓ is a parameter, the goal is to find a hitting set of size at most $F(\ell)\ell$ whenever G has a hitting set of size at most ℓ.

The starting point of our reduction is the following result for Hitting Set by Karthik, Laekhanukit, and Manurangsi [44].

Theorem 4 *(Theorem 1.4 in [44]). Assuming ETH, no* $f(\ell)(|U||\mathcal{A}|)^{o(\ell)}$*-time algorithm can approximate* Hitting Set *within a factor of* $F(\ell)$*, for any computable functions* f *and* F *of* ℓ.

Note that Theorem 1.4 in [44] is stated for the Dominating Set problem, however by the standard parameter-preserving reduction from Dominating Set to Hitting Set (see e.g. Theorem 13.28 in [11]), the statement above immediately follows.

Now, intuitively, given an instance $(U, \mathcal{A}, \ell)$ of Hitting Set, where $\mathcal{A} = \{S_1, \ldots, S_m\}$, our reduction constructs clusters $\mathbf{C}_0$, $\ldots$, $\mathbf{C}_m$ in $\mathbb{R}^{\sum_{j=1}^{m} |S_j|}$. The clusters $\mathbf{C}_1, \ldots, \mathbf{C}_m$ represent the sets $S_1, \ldots, S_m$ in the family $\mathcal{A}$, and $\mathbf{C}_0$ is a special cluster that needs to be separated from each of $\mathbf{C}_1, \ldots, \mathbf{C}_m$ 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 S_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, which results in the following theorem.

Theorem 5. *For any computable functions f and F, there is no $F(s)$-approximation algorithm for the optimization version of* Clustering Explanation *with running time $f(s)(nd)^{o(s)}$, unless ETH fails.*

Proof. We show a reduction from Hitting Set. Consider an instance of Hitting Set over a universe U with a family of sets $\mathcal{A}$ and the size of the target hitting set ℓ. We construct the following instance of Clustering Explanation, denote $|\mathcal{A}| = m$. The target dimension d is equal to the sum of set sizes in the family $\mathcal{A}$, $d = \sum_{S \in \mathcal{A}} |S|$. The number of clusters in the constructed instance is $m + 1$, and for clarity, we denote them by $\mathbf{C}_0, \ldots, \mathbf{C}_m$. The target parameter s, that is, the number of points to remove from the cluster, is set exactly to ℓ.

Now we describe how the clusters are composed. Intuitively, the clusters $\mathbf{C}_1, \ldots, \mathbf{C}_m$ represent the sets in the family $\mathcal{A}$, and $\mathbf{C}_0$ is a special cluster that needs to be separated from each of $\mathbf{C}_1, \ldots, \mathbf{C}_m$ so that the clustering is explainable. The separation can only be performed by removing special points from $\mathbf{C}_0$ that correspond 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 S_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.

Formally, we define point sets $\mathbf{C}_0, \ldots, \mathbf{C}_m$ in terms of coordinates in $\mathbb{R}^d$. The constructed instance is binary, that is, only values zero and one are used in the vectors. Order arbitrarily the sets in the family $\mathcal{A}$, i.e. $\mathcal{A} = \{S_1, \ldots, S_m\}$, for each $j \in \{1, \ldots, m\}$, the cluster $\mathbf{C}_j$ corresponds to the set S_j. The d coordinates are partitioned between the m sets in the following way. For each $j \in \{1, \ldots, m\}$, denote $I_j = \left[\sum_{j' < j} |S_{j'}| + 1, \sum_{j' \leq j} |S_{j'}|\right] \subseteq \{1, \ldots, d\}$. That is, the first range of coordinates I_1 is the first $|S_1|$ coordinates, I_2 are the following $|S_2|$ coordinates, and so on. Now, for $j \in \{1, \ldots, m\}$, the set $\mathbf{C}_j$ consists of $F(\ell) \cdot \ell + 1$ identical points $\mathbf{w}_j$, where $\mathbf{w}_j[i] = 1$ for $j \in I_j$ and $\mathbf{w}_j[i] = 0$ for $j \notin I_j$. The set $\mathbf{C}_0$ consists of two parts, $\mathbf{C}_0 = \mathbf{O} \cup \mathbf{V}$. First, there are $F(\ell) \cdot \ell + 1$ identical zero vectors in $\mathbf{C}_0$, we denote this set by $\mathbf{O}$. Second, for each element u in the universe U, there is a point $\mathbf{v}_u$ in $\mathbf{V}$. The coordinates of $\mathbf{v}_u$ are set so that for every $j \in \{1, \ldots, m\}$ such that $u \in S_j$, there is exactly one coordinate in I_j where $\mathbf{v}_u$ is set to one, and this coordinate is unique for each $u \in S_j$. If for $j \in \{1, \ldots, m\}$, $u \notin S_j$, then $\mathbf{v}_u[i] = 0$ for all $i \in I_j$. More specifically, for each $S_j \in \mathcal{A}$, order arbitrarily the elements of S_j, $S_j = \{u_1, \ldots, u_{|S_j|}\}$. For $i \in [|S_j|]$, $\mathbf{v}_{u_i}[\sum_{j' < j} |S_j| + i] = 1$. After performing the above for each $j \in \{1, \ldots, m\}$, set the remaining coordinates of each vector $\mathbf{v}_u$ to zero. This concludes the construction.

Now we show that an $F(s)$-approximate solution to the constructed Clustering Explanation instance would imply an $F(\ell)$-approximate solution to the original Hitting Set instance. Specifically, we prove the following.

Claim 3.2. *Whenever there exists a set $\mathbf{W} \subseteq \mathbf{X}$ of size at most $F(s) \cdot s$ such that $\{\mathbf{C}_0 \setminus \mathbf{W}, \ldots, \mathbf{C}_m \setminus \mathbf{W}\}$ is an explainable clustering, there also exists a set $H \subseteq U$ that is a hitting set of $\mathcal{A}$ and $|H| \leq |\mathbf{W}|$. On the other hand, for any hitting set $H \subseteq U$ there exists a solution $\mathbf{W}$ to the* Clustering Explanation *instance such that $|\mathbf{W}| = |H|$.*

Proof. In the forward direction, consider such a set $\mathbf{W} \subseteq \mathbf{X}$. We may assume that $\mathbf{W} \cap \mathbf{C}_j = \emptyset$ for each $j \in \{1, \ldots, m\}$ and $\mathbf{W} \cap \mathbf{O} = \emptyset$, since these sets consist of $F(s) \cdot s + 1$ identical points, and replacing $\mathbf{W}$ by a smaller set not intersecting the sets above is still a solution. Thus, $\mathbf{W} \subseteq \mathbf{V}$. Recall that each of the points in $\mathbf{V}$ corresponds to an element of the universe U in the Hitting Set instance. Denote by H the subset of U corresponding to $\mathbf{W}$, that is, $H = \{u \in U : \mathbf{v}_u \in \mathbf{W}\}$. Clearly $|H| \leq |\mathbf{W}|$, we claim that H is a solution to the Hitting Set instance.

Consider a threshold tree that provides an explanation for $\{\mathbf{C}_0 \setminus \mathbf{W}, \ldots, \mathbf{C}_m \setminus \mathbf{W}\}$. By the above, all these clusters are non-empty. For $j \in \{0, \ldots, m\}$, denote $\mathbf{C}'_j = \mathbf{C}_j \setminus \mathbf{W}$. Recall that we assume $\mathbf{C}_j = \mathbf{C}'_j$ for all $j \in \{1, \ldots, m\}$. We now show that for any $j \in \{1, \ldots, m\}$, the set H has a non-empty intersection with the set S_j. Since the tree represents the clustering, for each $j \in \{1, \ldots, m\}$, there exists a cut in the tree that separates $\mathbf{C}'_j$ and $\mathbf{C}'_0$. That is, there exists a cut $\mathsf{Cut}_{i,\theta}(\mathbf{X}') = (\mathbf{X}'_1, \mathbf{X}'_2)$ in the tree for $\mathbf{X}' \subseteq \mathbf{X}$ such that $\mathbf{C}'_j \cup \mathbf{C}'_0 \subseteq \mathbf{X}'$, and $\mathbf{C}'_j \subseteq \mathbf{X}'_1$, $\mathbf{C}'_0 \subseteq \mathbf{X}'_2$ (w.l.o.g). The dimension of the cut i necessarily belongs to I_j, since in all other coordinates $\mathbf{C}'_j$ is indistinguishable from $\mathbf{O}$. For $i \in I_j$, there exists a point $\mathbf{v}_u \in \mathbf{V}$ such that $\mathbf{v}_u[i] = 1$ and $u \in S_j$. Since the points in $\mathbf{C}'_j$ are set to one in this coordinate and the points in $\mathbf{O}$ to zero, $\mathbf{v}_u$ has to be in $\mathbf{W}$. Thus, by construction, $u \in H$, and the set S_j is hit by H.

In the other direction, assume that there exists a hitting set H for $\mathcal{A}$. We set $\mathbf{W}$ to be the corresponding to H subset of $\mathbf{V}$, $|\mathbf{W}| = |H|$. For every $j \in \{1, \ldots, m\}$, there exists an element $u \in H$ such that $u \in S_j$. Consider the corresponding element $\mathbf{v}_u$ of $\mathbf{W}$, and the coordinate $i \in I_j$ where $\mathbf{v}_u[i] = 1$. By construction, $\mathbf{v}_u$ is the only element of $\mathbf{C}_0$ that has one in the i-th coordinate, and for any $j' \neq j$, the elements of $\mathbf{C}_{j'}$ have zero in these coordinate. Thus, cutting over the dimension i with

the threshold θ set to zero separates $\mathbf{C}_j$ from all the other clusters. Finally, the threshold tree for $\{\mathbf{C}_0 \setminus \mathbf{W}, \ldots, \mathbf{C}_m \setminus \mathbf{W}\}$ is constructed by separating out each $\mathbf{C}_j$ one by one via the corresponding cut. □

The theorem follows easily from Claim 3.2. Namely, assume there exists an $F(s)$-approximate algorithm for Clustering Explanation with running time $f(s)(nd)^{o(s)}$, that is, an algorithm that correctly decides either that the input instance has a solution of size at most $F(s) \cdot s$, or that it has no solution of size at most s. We construct an $F(\ell)$-approximate algorithm for Hitting Set as follows. First, construct a Clustering Explanation instance via the reduction above. Second, run the Clustering Explanation algorithm on that instance and output its answer. If it returns that there is a solution $\mathbf{W}$ of size at most $F(s) \cdot s$, by Claim 3.2 there is a solution H to the original Hitting Set instance of size at most $F(s) \cdot s = F(\ell) \cdot \ell$. If there is no solution to the Clustering Explanation instance of size at most s, by Claim 3.2 there also cannot be a solution to the Hitting Set instance of size at most $s = \ell$. This shows that the constructed algorithm provides indeed an $F(\ell)$-approximation for the Hitting Set problem. Finally, its running time can be bounded as $f(s)(nd)^{o(s)} = f(\ell)(F(\ell)|U|m)^{o(\ell)}$, which contradicts Theorem 4 unless ETH fails. □

4. 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 $\{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbf{X}$ of minimum k-means (resp. k-median) cost.

4.1. Exact algorithms

Our $(nd)^{k+\mathcal{O}(1)}$ time algorithm is indeed very simple and based on a branching technique. At each non-leaf node of a 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 a threshold binary tree is $k-1$, we have the following theorem.

Theorem 6. 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 $i \in \{1, \ldots, d\}$. We highlight that when we write $\mathbf{x} < \mathbf{y}$, we require the strict inequality for *every* coordinate.

Theorem 7. 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 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{X} \subseteq \mathbb{R}^d$. Following the proof of Theorem 2, 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 $i \in \{1, \ldots, d\}$. 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 $i \in \{1, \ldots, d\}$. We compute the table of values of $\omega(\mathbf{a}, \mathbf{b}, s)$ consecutively for $s = 1, 2, \ldots, k$.

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 $i \in \{1, \ldots, d\}$ 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

$$\omega(\mathbf{a},\mathbf{b},s)=\min\{\omega(\mathbf{a},\mathbf{b}^{i,\theta},s_1)+\omega(\mathbf{a}^{i,\theta},\mathbf{b},s_2) \quad \text{for } 1\leq i\leq d,\theta\in \mathsf{coord}_i(\mathbf{X}), \quad \mathbf{a}[i]<\theta<\mathbf{b}[i], \quad s_1,s_2\geq 1, \text{ and } s_1+s_2=s\}. \tag{13}$$

The correctness of (13) follows from 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 the 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 (13).

To evaluate the running time, observe that to compute $\omega(\mathbf{a},\mathbf{b},s)$ using (13), 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 (13) $k-1\leq n$ times, the total running time is $n^{2d}\cdot(dn)^{\mathcal{O}(1)}$. □

4.2. Hardness

In this section, we show our hardness results for Explainable k-means and Explainable k-median: the problems are NP-complete, W[2]-hard when parameterized by k, and cannot be solved in $f(k)\cdot n^{o(k)}$ time for a computable function $f(\cdot)$ unless ETH fails. Moreover, the hardness holds even if the input points are binary. More precisely, we prove the following theorem.

Theorem 8. *Given a collection of n points $\mathbf{X}\subseteq\{0,1\}^d$, a positive integer $k\leq n$, and a nonnengative integer B, it is* W[2]*-hard to decide whether $\mathbf{X}$ admits an explainable k-clustering of mean (median, respectively) cost at most B, when the problem is parameterized by k. Moreover, the problems are* NP*-complete and cannot be solved in $f(k)\cdot n^{o(k)}$ time for a computable function $f(\cdot)$ unless ETH fails.*

Proof. We show the theorem for the means costs and then briefly explain how the proof is modified for medians. The case of median cost is easier due to the fact that for a collection of binary points, its median is also a binary vector.

We reduce from the Hitting Set problem. The task of the problem is, given a family of sets $\mathcal{W}=\{W_1,\ldots,W_m\}$ over universe $U=\{u_1,\ldots,u_n\}$, and a positive integer k, decide whether there is $S\subseteq U$ of size at most k that is a *hitting* set for $\mathcal{W}$, i.e. such that $S\cap W_i\neq\emptyset$ for every $i\in\{1,\ldots,m\}$. This problem is well-known to be W[2]-complete when parameterized by k even if the input is restricted to families of sets of the same size (see [13]).

Let $\mathcal{W}=\{W_1,\ldots,W_m\}$ be a family of sets over $U=\{u_1,\ldots,u_n\}$ with $|W_1|=\cdots=|W_m|=r$, and let k be a positive integer. We construct the following points.

- Let $s=8nm^2$. For each $i\in\{1,\ldots,n\}$, we construct a vector $\mathbf{u}_i\in\{0,1\}^n$ by setting
$$\mathbf{u}_i[j]=\begin{cases}1 & \text{if } j=i,\\ 0 & \text{otherwise.}\end{cases}$$
We also define $\mathbf{U}_i$ to be the collection of s points identical to $\mathbf{u}_i$.
- For each $i\in\{1,\ldots,m\}$, let $\mathbf{w}_i$ be the characteristic vector of W_i. That is,
$$\mathbf{w}_i[j]=\begin{cases}1 & \text{if } u_j\in W_i,\\ 0 & \text{otherwise.}\end{cases}$$
We define $\mathbf{W}=\{\mathbf{w}_1,\ldots,\mathbf{w}_m\}$.
- For $t=16ns^2=1024n^3m^3$ we construct a collection $\mathbf{Z}$ of t zero points.

Finally, we define

- $\mathbf{X}=\left(\bigcup_{i=1}^n\mathbf{U}_i\right)\cup\mathbf{W}\cup\mathbf{Z}$,
- $k'=k+1$, and
- $B=m(r-1)+s(n-k)=m(r-1)+8nm^2(n-k)$.

Clearly, $\mathbf{X}$, k' and B can be constructed from the considered instance of Hitting Set in polynomial time. We claim that $\mathcal{W}$ has a hitting set S of size at most k if and only if $\mathbf{X}$ has an explainable k'-clustering of means cost at most B.

For the forward direction, assume that S is a hitting set of size at most k. Without loss of generality, we assume that $|S|=k$. Let $S=\{u_{i_1},\ldots,u_{i_k}\}$. We define the k'-clustering $\{\mathbf{C}_1,\ldots,\mathbf{C}_{k+1}\}$ as follows. We set $\mathbf{C}_1=\{\mathbf{x}\in\mathbf{X}\mid\mathbf{x}[i_1]=1\}$ and for $j=2,\ldots,k+1$,

$$\mathbf{C}_j = \{\mathbf{x} \in \mathbf{X} \mid \mathbf{x}[i_j] = 1\} \setminus \bigcup_{h=1}^{j-1} \mathbf{C}_h.$$

Observe that the constructed clustering is explainable. To see this, let $X_0 = \mathbf{X}$ and set $X_j = \mathbf{X} \setminus \bigcup_{h=1}^{j-1} \mathbf{C}_h$. Then $(X_j, \mathbf{C}_j) = \mathsf{Cut}_{i_j,0}(X_{j-1})$ for $j \in \{1, \ldots, k+1\}$. Notice that $\mathbf{U}_{i_j} \subseteq \mathbf{C}_j$ for every $j \in \{1, \ldots, k\}$, and we have that $\mathbf{Z} \subseteq \mathbf{C}_{k+1}$ and $\mathbf{U}_h \subseteq \mathbf{C}_{k+1}$ for all $h \in \{1, \ldots, m\} \setminus \{i_1, \ldots, i_k\}$. Moreover, because S is a hitting set, each $\mathbf{w}_j \in \mathbf{W}$ is included in some cluster $\mathbf{C}_h$ for some $h \in \{1, \ldots, k\}$, that is, $\mathbf{C}_{k+1} = \mathbf{Z} \cup \mathbf{R}$, where $\mathbf{R} = \bigcup_{h \in \{1,\ldots,m\} \setminus \{i_1,\ldots,i_k\}} \mathbf{U}_h$.

For every $j \in \{1, \ldots, k\}$, we define $\mathbf{c}_j = \mathbf{u}_{i_j}$, and we set $\mathbf{c}_{k+1}$ be the zero vector. Clearly, for every $j \in \{1, \ldots, k+1\}$, $\mathsf{cost}_2(\mathbf{C}_j) \leq \sum_{\mathbf{x} \in \mathbf{C}_j} \|\mathbf{x} - \mathbf{c}_j\|_2^2$. Let $\mathbf{W}_j = \mathbf{W} \cap \mathbf{C}_j$ for $j \in \{1, \ldots, k\}$. Note that $\mathbf{C}_j = \mathbf{W}_j \cup \mathbf{U}_{i_j}$ for $j \in \{1, \ldots, k\}$. Because $\mathbf{x}[i_j] = 1$ for every $\mathbf{x} \in \mathbf{C}_j$ for $j \in \{1, \ldots, k\}$ and $\mathbf{x}$ has exactly r nonzero elements, we have that $\|\mathbf{x} - \mathbf{c}_j\|_2^2 = r - 1$ for every $\mathbf{x} \in \mathbf{W}_j$ for $j \in \{1, \ldots, k\}$. If $\mathbf{x} \in \mathbf{R}$, then $\|\mathbf{x} - \mathbf{c}_{k+1}\|_2^2 = 1$. We obtain that

$$\mathsf{cost}_2(\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}) \leq \sum_{j=1}^{k+1} \sum_{\mathbf{x} \in \mathbf{C}_j} \|\mathbf{x} - \mathbf{c}_j\|_2^2 = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathbf{W}_j} \|\mathbf{x} - \mathbf{c}_j\|_2^2 + \sum_{\mathbf{x} \in \mathbf{R}} \|\mathbf{x} - \mathbf{c}_{k+1}\|_2^2 = m(r-1) + s(n-k) = B.$$

Thus, $\{\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}\}$ is an explainable k'-clustering of means cost at most B.

For the opposite direction, let $\{\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}\}$ be an explainable k'-clustering with $\mathsf{cost}_2(\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}) \leq B$. Let also (T, k', φ) be a canonical threshold tree for this clustering. Notice that because every point of $\mathbf{X}$ is binary, for every nonleaf $v \in V(T)$, $\varphi(v) = (i, 0)$ for some $i \in \{1, \ldots, d\}$.

We show the following property of (T, k', φ): for every nonleaf node $v \in V(T)$, its right child is a leaf. Suppose that this is not the case. Denote by x the root of T and let $P = x_1 \cdots x_p$ be the root-leaf path, where $x_1 = x$ and x_i is the left child of x_{i-1} for $i \in \{2, \ldots, p\}$, that is, P is constructed starting from the root and following the left children until we achieve the leftmost leaf. Because T has k' leaves and at least one right child is not a leaf, we have that $p \leq k' - 1 = k$. Denote by $\mathbf{C}_q$ the cluster corresponding to the leaf x_p of T. Notice that $\mathbf{Z} \subseteq \mathbf{C}_q$ and $n - (p-1)$ collections $\mathbf{U}_i$ are included in $\mathbf{C}_q$ for some $i \in \{1, \ldots, n\}$. Let $i_1, \ldots, i_{n-p+1} \in \{1, \ldots, n\}$ be the distinct indices such that $\mathbf{U}_{i_j} \subseteq \mathbf{C}_q$. Denote by $\mathbf{c}$ the mean of $\mathbf{C}_q$. Observe that for each $j \in \{1, \ldots, n\}$, the multiset of the j-th coordinates of the points of $\mathbf{C}_q$ contains at most $s + m$ ones and at least t zeros. Then for every $j \in \{1, \ldots, n\}$,

$$\mathbf{c}[j] \leq \frac{m+s}{|\mathbf{C}_q|} \leq \frac{m+s}{t} \leq \frac{2s}{t} = \frac{1}{8ns} \leq \frac{1}{2s},$$

and we obtain that

$$\begin{aligned}
\mathsf{cost}_2(\mathbf{C}_q) \geq & \sum_{j=1}^{n-p+1} \sum_{\mathbf{x} \in \mathbf{U}_{i_j}} \|\mathbf{x} - \mathbf{c}\|_2^2 = s \sum_{j=1}^{n-p+1} \|\mathbf{u}_{j_i} - \mathbf{c}\|_2^2 \\
& \geq s(n-k+1)\Big(1 - \frac{1}{2s}\Big)^2 > s(n-k+1)\Big(1 - \frac{1}{s}\Big) \\
& \geq s(n-k) + s - n \geq s(n-k) + 2nm - n \\
& > m(n-1) + s(n-k) \geq B,
\end{aligned}$$

contradicting that $\mathsf{cost}_2(\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}) \leq B$.

Because for every nonleaf node $v \in V(T)$, its right child is a leaf, we have that the clustering is obtained by consecutive cutting each cluster from the set of points. Then we can assume without loss of generality that there are k-tuple of distinct indices $(i_1, \ldots, i_k)$ from $\{1, \ldots, n\}$ such that $\mathbf{C}_1 = \{\mathbf{x} \in \mathbf{X} \mid \mathbf{x}[i_1] = 1\}$ and

$$\mathbf{C}_j = \{\mathbf{x} \in \mathbf{X} \mid \mathbf{x}[i_j] = 1\} \setminus \bigcup_{h=1}^{j-1} \mathbf{C}_h$$

for $j = 2, \ldots, k+1$. We claim that $S = \{u_{i_1}, \ldots, u_{i_k}\}$ is a hitting set for $\mathcal{W}$.

The proof is by contradiction. Suppose that S is not a hitting set. For every $i \in \{1, \ldots, k+1\}$, let $\mathbf{W}_i = \mathbf{C}_i \cap \mathbf{W}$. Notice that because S is not a hitting set, $\mathbf{W}_{k+1} \neq \emptyset$. We analyze the structure of clusters and upper bound their means costs. For this, denote by $\mathbf{c}_1, \ldots, \mathbf{c}_{k+1}$ the means of $\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}$.

Let $j \in \{1, \ldots, k\}$ and consider $\mathbf{C}_j$ with the mean $\mathbf{c}_j$. We have that $\mathbf{C}_j = \mathbf{U}_{i_j} \cup W_j$. Then $\mathbf{c}_j[i_j] = 1$. Let $h \in \{1, \ldots, n\}$ be distinct from i_j. If $\mathbf{x} \in \mathbf{U}_{i_j}$, then $\mathbf{x}[h] = 0$. Then the multiset of the h-th coordinates of the points of $\mathbf{C}_j$ contains at most $|\mathbf{W}_j| \leq m$ ones and at least s zeros, and we have that

$$\mathbf{c}_j[h] \leq \frac{m}{|\mathbf{C}_j|} \leq \frac{m}{s} \leq \frac{1}{8mn}.$$

Recall that $\mathbf{x} \in \mathbf{W}$ has exactly r elements that are equal to one. This implies, that for every $\mathbf{x} \in \mathbf{W}_j$,

$$\|\mathbf{x} - \mathbf{c}_j\|_2^2 \geq (r-1)\big(1 - \frac{1}{8mn}\big)^2 \geq (r-1)\big(1 - \frac{1}{4mn}\big)$$

and, therefore,

$$\mathsf{cost}_2(\mathbf{C}_j) \geq (r-1)\big(1 - \frac{1}{4mn}\big)|\mathbf{W}_j| \geq (r-1)|\mathbf{W}_j| - \frac{1}{4m}|\mathbf{W}_j|, \tag{14}$$

because $r \leq n$.

Now we consider $\mathbf{C}_{k+1}$ and the corresponding mean $\mathbf{c}_{k+1}$. Notice that $\mathbf{C}_{k+1} = \mathbf{Z} \cup \mathbf{W}_{k+1} \cup \mathbf{R}$, where $\mathbf{R} = \bigcup_{h \in \{1,\ldots,m\} \setminus \{i_1,\ldots,i_k\}} \mathbf{U}_h$. Let $h \in \{1,\ldots,n\}$. We have that $\mathbf{x}[h] = 0$ for $\mathbf{x} \in \mathbf{Z}$. Hence, the multiset of h-th coordinates of the points of $\mathbf{C}_{k+1}$ contains at most $|\mathbf{W}_{k+1}| + s \leq m + s$ ones and at least t zeros. Therefore,

$$\mathbf{c}_{k+1}[h] \leq \frac{m+s}{|\mathbf{C}_{k+1}|} \leq \frac{2s}{t} \leq \frac{1}{8sn}.$$

Since $\mathbf{x} \in \mathbf{W}$ has exactly $r \leq n$ elements that are equal to one, for every $\mathbf{x} \in \mathbf{W}_{k+1}$,

$$\|\mathbf{x} - \mathbf{c}_{k+1}\|_2^2 \geq r\big(1 - \frac{1}{8sn}\big)^2 \geq r\big(1 - \frac{1}{8mn}\big)^2 \geq r\big(1 - \frac{1}{4mn}\big) \geq r - \frac{1}{4m}.$$

For every $\mathbf{x} \in \mathbf{R}$, $\mathbf{x}$ contains a unique nonzero element and, therefore,

$$\|\mathbf{x} - \mathbf{c}_{k+1}\|_2^2 \geq \big(1 - \frac{1}{8sn}\big)^2 \geq 1 - \frac{1}{4sn}.$$

Note that $\mathbf{R}$ contains exactly $s(n-k)$ points. Then

$$\begin{aligned}\mathsf{cost}_2(\mathbf{C}_{k+1}) &\geq \sum_{\mathbf{x} \in \mathbf{W}_{k+1}} \|\mathbf{x} - \mathbf{c}_{k+1}\|_2^2 + \sum_{\mathbf{x} \in \mathbf{R}} \|\mathbf{x} - \mathbf{c}_{k+1}\|_2^2 \\ &\geq \big(r - \frac{1}{4m}\big)|\mathbf{W}_{k+1}| + s(n-k)\big(1 - \frac{1}{4sn}\big) \\ &\geq r|\mathbf{W}_{k+1}| + s(n-k) - \frac{1}{4m}|\mathbf{W}_{k+1}| - \frac{1}{4}. \end{aligned} \tag{15}$$

Recall that $|\mathbf{W}_1| + \cdots + |\mathbf{W}_{k+1}| = m$. Then combining (14) and (15), we obtain that

$$\mathsf{cost}_2(\mathbf{C}_1, \ldots, \mathbf{C}_{k+1}) = \sum_{j=1}^{k+1} \mathsf{cost}_2(\mathbf{C}_j) \geq m(r-1) + s(n-k) + |\mathbf{W}_{k+1}| - \frac{1}{2} \geq B + \frac{1}{2}$$

because $\mathbf{W}_{k+1} \neq \emptyset$. However, this contradicts that the means cost of $\{\mathbf{C}_1, \ldots, \mathbf{C}_{k'}\}$ is at most B. This means that S is a hitting set for $\mathcal{W}$.

This concludes the hardness proof for Explainable k-means. For the median cost, we use exactly the same reduction from Hitting Set. Then we show that $\mathcal{W}$ has a hitting set S of size at most k if and only if $\mathbf{X}$ has an explainable k'-clustering of median cost at most B. The proof for the forward direction is identical to the proof for the means up to the replacement of cost_2 by cost_1 and of $\|\cdot\|_2^2$ by $\|\cdot\|_1$. For the opposite direction, the proof follows the same lines as the above proof for means but gets simplified because we can assume that medians are binary. In particular, if the multiset of h-th coordinates of the points of a cluster $\mathbf{C}_j$ contains more zeros than ones, then $\mathbf{c}_j[h] = 0$ for the median $\mathbf{c}_j$. Notice that the crucial part of the proof for the means is obtaining upper bounds for the values $\mathbf{c}_j[h]$. Now we can immediately assume that $\mathbf{c}_j[h] = 0$ in all considered cases whenever we upper bound $\mathbf{c}_j[h]$, and the lower bounds for the costs in the proof become straightforward.

To see the second part of the theorem, note that our reduction from Hitting Set is polynomial. This immediately implies NP-hardness of the considered problems for both measures. For the lower bound up to ETH, we use the well-known fact (see, e.g. [11, Chapter 14]) that Hitting Set does not admit an algorithm with running time $f(k) \cdot (n+m)^{o(k)}$ for any computable function $f(\cdot)$ assuming ETH. Because our reduction is polynomial and, moreover, the value of the parameter in the constructed instance is $k' = k+1$, we obtain that an algorithm with running time $f(k) \cdot (n+m)^{o(k)}$ for our problems would contradict ETH. □

5. 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.

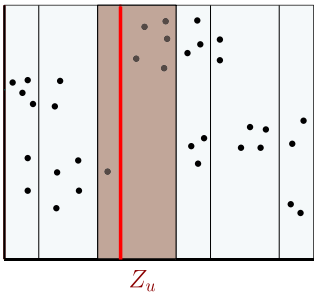


Fig. 5. An example of the definition of a set $\mathbf{Z}_u$ for a node $u \in T$. The points $\mathbf{X}$ are partitioned into stripes containing $n' = \lfloor \frac{\epsilon n}{k} \rfloor$ (6 in the figure) points. The set $\mathbf{Z}_u$ is the stripe containing the optimal cut, represented in red. Any cut inside that stripe gives the same partition of $\mathbf{X} \setminus \mathbf{Z}_u$ as the optimal one. This means that by removing n' points, the algorithm essentially reduces the numbers of different cuts from n to $\mathcal{O}(\frac{k}{\epsilon})$.

Theorem 9. Approximate Explainable k-means *and* Approximate Explainable k-median *are solvable in* $(\frac{4d(k+\epsilon)}{\epsilon})^k \cdot n^{\mathcal{O}(1)}$ *time.*

Proof. As the proofs for both problems are identical, we describe only the algorithm for Approximate Explainable k-means.

Let $\mathbf{X} \subseteq \mathbb{R}^d$ be an instance of Approximate Explainable k-means with $|\mathbf{X}| = n$ and let (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). Recall that the total number of full binary trees with k leaves is the $(k-1)$-th Catalan number $C_{k-1} = \frac{1}{k}\binom{2(k-1)}{k-1} \leq 4^k$. Since T is a full binary tree with k leaves, guessing T only requires at most 4^k tries. Guessing φ is more complicated however, as there is $d \cdot (n-1)$ choices at each node u of T because we have d possibilities to choose a coordinate i and $|\mathsf{coord}_i(\mathbf{X})| - 1$ possibilities to select the threshold value. This gives potentially $(d(n-1))^{k-1}$ possibilities for $k-1$ nonleaf nodes of T. 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.

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 $4^k(\frac{dk}{\epsilon})^k \cdot n^{\mathcal{O}(1)}$, which ends the proof. From now on, let us assume that $n' \geq 1$.

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 the following rounding. Consider $u \in U$ and assume that $\varphi(u) = (j, \theta)$. Denote by $\mathbf{x}_1, \ldots, \mathbf{x}_n$ the points of $\mathbf{X}$ assuming that they are sorted by their j-th coordinate, that is, $\mathbf{x}_1[j] \leq \cdots \leq \mathbf{x}_n[j]$. Recall that because (T, k, φ) is canonical, $\theta \in \mathsf{coord}_j(\mathbf{X}) = \{\mathbf{x}_1[j], \ldots, \mathbf{x}_n[j]\}$. We define $\varphi'(u) = (j, \theta')$, where $\theta' = \mathbf{x}_{in'+1}[j]$ for the largest nonnegative integer i such that $\mathbf{x}_{in'+1}[j] \leq \theta$.

Consider now the clustering obtained from the threshold tree (T, k, φ'). At each node $u \in U$ such that $\varphi(u) = (j, \theta)$ and $\varphi'(u) = (j, \theta')$ for $\theta' = \mathbf{x}_{in'+1}$, the points $\mathbf{x}$ of $\mathbf{X}$ that can be misplaced by $\mathsf{Cut}_{\varphi'(u)}(\mathbf{X})$ are exactly the points such that $\theta' < \mathbf{x}[j] \leq \theta$. By the choice of i, we have that $\theta \leq \mathbf{x}_{(i+1)n'}$. This means that, if $\mathbf{Z}_u$ denotes the set of all the points $\mathbf{x}_h$ such that $i \cdot n' < h \leq (i+1)n'$, then 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 (see Fig. 5). 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 $\mathbf{X} \setminus \mathbf{Z}$. Because $|\mathbf{Z}_u| \leq n'$, we have that $|\mathbf{Z}| \leq kn' \leq \epsilon n$.

Our algorithm is based on these observations. The algorithm tries all possible choices for T. Given T with the set of nonleaf nodes U, it tries all possible choices of $\varphi'(u) = (j, \theta')$ for $u \in U$ as follows. For each u, we first pick a coordinate $j \in \{1, \ldots, d\}$. Then we sort $\mathbf{X}$ by the j-th coordinate of the points and assume that $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{x}_1[j] \leq \cdots \leq \mathbf{x}_n[j]$. We consider all possible values of $\theta' = \mathbf{x}_{in'+1}$ for $0 \leq i \leq \frac{k}{\epsilon}$. Furthermore, the algorithm constructs the set of points $\mathbf{Z}_u = \{\mathbf{x}_h \mid in' < h \leq (i+1)n'\}$.

For each choice T and φ, the algorithm constructs $\mathbf{Z} = \bigcup_{u \in U} \mathbf{Z}_u$, removes $\mathbf{Z}$ and computes the value of the clustering induced by (T, k, φ') on $\mathbf{Y} = \mathbf{X} \setminus \mathbf{Z}$. Finally, it outputs the set $\mathbf{Y}$ as well as the threshold tree (T, k, φ') which minimizes the value of the clustering.

We have that the value of the clustering for (T, k, φ') on $\mathbf{Y}$ does not exceed the value of the clustering for an optimum threshold tree on $\mathbf{Y}$. Therefore, we obtain an explainable k-clustering of $\mathbf{Y}$ whose cost is upper bounded by the minimum cost of an explainable k-clustering of $\mathbf{X}$. Because for every set of choices of T and φ', the set $\mathbf{Z}$ has size at most $k \cdot n' \leq \epsilon n$, the algorithm outputs the set $\mathbf{Y}$ such that $|\mathbf{Y}| \geq (1-\epsilon)|\mathbf{X}|$. This concludes the description of the algorithm and its correctness proof.

Since there are at most 4^k possible choices of T and $d \cdot (\frac{k}{\epsilon} + 1)$ possible choices of $\varphi'(u)$ for every nonleaf node of the tree, we conclude that the total number of possible choices of T and φ' is at most $4^k\big(d(\frac{k}{\epsilon} + 1)\big)^{k-1} \leq \big(\frac{4d(k+\epsilon)}{\epsilon}\big)^k$. This implies that the running time of the algorithm is $(\frac{4d(k+\epsilon)}{\epsilon})^k \cdot n^{\mathcal{O}(1)}$. This concludes the proof. □

6. Conclusion

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

Explainable k-means *and* Explainable k-median Our first question concerns FPT-approximability of Explainable k-means and Explainable k-median. Recall that by Theorem 8, both problems are W[2]-hard being parameterized by k. The reduction also rules out algorithms with running time $f(k) \cdot (nd)^{o(k)}$ for both of these problems up to a widely believed assumption in the complexity theory. Theorem 6 indicates that this lower bound is tight since the naive brute force algorithm solves each of the problems in time $(nd)^{k+\mathcal{O}(1)}$. However, our hardness reduction does not exclude the possibility of the existence of efficient FPT approximation algorithms. Such algorithms exist for vanilla clustering. Particularly, for k-means, an $(1+\varepsilon)$-approximation in $2^{(k/\varepsilon)^{\mathcal{O}(1)}} \cdot nd$ time was demonstrated by Kumar, Sabharwal, and Sen [31]. The same techniques can be used for k-median [31]. Our question is: Is it possible to find a $(1+\varepsilon)$-approximation for Explainable k-means or Explainable k-median in time $2^{f(k/\varepsilon)} \cdot (nd)^{\mathcal{O}(1)}$?

In Theorem 7, we give an algorithm solving Explainable k-means and Explainable k-median in time $n^{2d} \cdot (nd)^{\mathcal{O}(1)}$. The exponential dependence on d makes an algorithm with such a running time unsuitable for instances with large dimension d (number of features). Due to our lower bound, the exponential dependence on d is unavoidable. However, it is plausible that the problem is FPT parameterized by d. In other words, the question is whether it is possible to obtain an $f(d) \cdot (dn)^{\mathcal{O}(1)}$ time algorithm solving Explainable k-means or Explainable k-median for some function $f(\cdot)$? The question about FPT-approximation is also open. In particular, the existence of an $(1+\varepsilon)$-approximation for Explainable k-means or Explainable k-median in time $2^{f(d/\varepsilon)} \cdot (nk)^{\mathcal{O}(1)}$ remains open. Observe that the standard dimension reduction tools seem to be inapplicable for explainable clusterings because a specific dimension could be crucial for separating clusters even when it does not contribute significantly to the distance, and it is very unclear how to distinguish such dimensions.

The same question is valid for approximation under the "stronger" parameterization by $k+d$, that is, $(1+\varepsilon)$-approximation for Explainable k-means or Explainable k-median in time $2^{f((d+k)/\varepsilon)} \cdot n^{\mathcal{O}(1)}$. Our Theorem 9 shows that an approximation can be done in $(\frac{8dk}{\epsilon})^k \cdot n^{\mathcal{O}(1)}$ time, if we are allowed to sacrifice εn of the input points in order to construct an explainable clustering (in fact, we obtain an explainable k-clustering without increasing the cost). Simultaneously, this theorem indicates the main technical difficulty for the approximation which is the fact that a small fraction of points can drastically increase the clustering cost (see also the bounds in the paper Moshkovitz et al. [41] for the cost of the explanation of an optimal k-means (medians) clustering).

In Theorem 1, we show that Clustering Explanation admits a polynomial-time approximation with a factor of $(k-1)$. We do not know whether there exist polynomial time algorithms providing a better ratio. The concrete question we leave open for further research is whether the approximation ratio $\log k$ for Clustering Explanation is achievable by some polynomial time algorithm.

Clustering explanation Further in the paper, we obtained a number of results about the Clustering Explanation problem, where the aim is to explain a given k-clustering by the cost of deletion of a bounded number of points. Theorem 5 shows that it is unlikely that the minimum number s of deleted points may be approximated in $f(s) \cdot n^{o(s)}$ time within $F(s)$ factor for any computable functions $f(\cdot)$ and $F(\cdot)$. However, Theorem 1 shows that Clustering Explanation admits a polynomial-time approximation with a factor of $(k-1)$. The concrete question we leave open for further research is whether the approximation ratio $\log k$ for Clustering Explanation is achievable by some polynomial time algorithm.

The complexity lower bound from Theorem 5 leads to the questions about the parameterized complexity of Clustering Explanation for various parameterizations by k, d, and s and their combinations. In particular, is Clustering Explanation FPT when parameterized by k or d, or for the combined parameterizations by $k+d$, $s+d$, or $s+k$? Note also that in Clustering Explanation, we do not make any assumptions about the input clustering $\{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ except that it is a partition of the input points. However, in a more practical setting, it may be assumed that $\mathbf{C}_1, \ldots, \mathbf{C}_k$ are "real" clusters. For example, $\{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ can be an optimum k-means or k-medians clustering. Would such an assumption make Clustering Explanation more tractable?

Non-unary conditions and other generalizations From a high-level perspective, the idea of Moshkovitz et al. [41], which we adapt in our work, is to design an efficient procedure for clustering that is explainable by a small decision tree and whose quality is close to an optimal k-clustering. In this model, the conditions in the decision tree are unary (in the form $x \leq a$ and $y < b$) and a natural research direction here is to investigate more general linear conditions (like $x + y < a$). An interesting question here is what is the right balance here between the complexity of the conditions of the decision tree and the human ability to grasp such explanations? But in certain situations, it is plausible to assume that a decision tree with non-unary conditions could have a reasonable interpretation for the purposes of clustering. From mathematical and algorithmic perspectives, such trees are also interesting objects. It is easy to construct examples when decision trees with a linear combination of two coordinates provide a much better "price of explainability" than trees with unary conditions. However, providing quantitative bounds as well as efficient algorithms for computing such trees seems to be way more challenging.

Another interesting research direction is to study decision trees that allow more than k leaves. In this case, one cluster may be assigned to several different leaves. How helpful is it to introduce more leaves? Frost, Moshkovitz, and Rashtchian [21] proposed a heuristic algorithm that constructs a decision tree with a given number $k' \geq k$ of leaves and analyzed its performance empirically. Makarychev and Shan [38] took a theoretical approach to this problem and designed for a parameter $\delta \in (0,1)$ a polynomial-time randomized algorithm constructing a threshold decision tree with $(1+\delta)k$ leaves. The constructed tree induces an explainable clustering whose cost is within $1/\delta \cdot polylog(k)$ the cost of an optimal

k-clustering. However, to the best of our knowledge, the parameterized complexity of constructing such types of decision trees remains completely unexplored.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

We thank the anonymous referees for their helpful comments and suggestions.

References

[1] Charu C. Aggarwal, Chandan K. Reddy (Eds.), Data Clustering: Algorithms and Applications, CRC Press, 2013.
[2] Daniel Aloise, Amit Deshpande, Pierre Hansen, Preyas Popat, NP-hardness of Euclidean sum-of-squares clustering, Mach. Learn. 75 (2) (2009) 245–248, https://doi.org/10.1007/s10994-009-5103-0.
[3] Dimitris Bertsimas, Agni Orfanoudaki, Holly M. Wiberg, Interpretable clustering: an optimization approach, Mach. Learn. 110 (1) (2021) 89–138, https://doi.org/10.1007/s10994-020-05896-2.
[4] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[5] Leonardo Cañete-Sifuentes, Raúl Monroy, Miguel Angel Medina-Pérez, A review and experimental comparison of multivariate decision trees, IEEE Access 9 (2021) 110451–110479, https://doi.org/10.1109/ACCESS.2021.3102239.
[6] Diogo V. Carvalho, Eduardo M. Pereira, Jaime S. Cardoso, Machine learning interpretability: a survey on methods and metrics, Electronics 8 (8) (2019) 832.
[7] Deeparnab Chakrabarty, Prachi Goyal, Ravishankar Krishnaswamy, The non-uniform k-center problem, in: 43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy, 2016, 67.
[8] Moses Charikar, Lunjia Hu, Near-optimal explainable k-means for all dimensions, in: Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022, SIAM, 2022, pp. 2580–2606.
[9] Moses Charikar, Samir Khuller, David M. Mount, Giri Narasimhan, Algorithms for facility location problems with outliers, in: Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2001, pp. 642–651.
[10] Chen Ke, A constant factor approximation algorithm for k-median clustering with outliers, in: Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2008, pp. 826–835.
[11] Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, Saket Saurabh, Parameterized Algorithms, Springer, ISBN 978-3-319-21274-6, 2015.
[12] Sanjoy Dasgupta, The Hardness of k-Means Clustering, Department of Computer Science and Engineering, University of California, 2008.
[13] Rodney G. Downey, Michael R. Fellows, Fundamentals of Parameterized Complexity, Texts in Computer Science, Springer, ISBN 978-1-4471-5558-4, 2013.
[14] Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh S. Vempala, V. Vinay, Clustering large graphs via the singular value decomposition, Mach. Learn. 56 (1–3) (2004) 9–33, https://doi.org/10.1023/B:MACH.0000033113.59016.96.
[15] Hossein Esfandiari, Vahab S. Mirrokni, Shyam Narayanan, Almost tight approximation algorithms for explainable clustering, in: Joseph (Seffi) Naor, Niv Buchbinder (Eds.), Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022, SIAM, 2022, pp. 2641–2663.
[16] Jean Feng, Noah Simon, Sparse-input neural networks for high-dimensional nonparametric regression and classification, arXiv preprint, arXiv:1711.07592, 2017.
[17] Qilong Feng, Zhen Zhang, Ziyun Huang, Jinhui Xu, Jianxin Wang, Improved algorithms for clustering with outliers, in: 30th International Symposium on Algorithms and Computation (ISAAC 2019), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.
[18] Fedor V. Fomin, Daniel Lokshtanov, Saket Saurabh, Meirav Zehavi, Kernelization: Theory of Parameterized Preprocessing, Cambridge University Press, 2019.
[19] Ricardo Fraiman, Badih Ghattas, Marcela Svarc, Interpretable clustering using unsupervised binary trees, Adv. Data Anal. Classif. 7 (2) (2013) 125–145.
[20] Zachary Friggstad, Kamyar Khodamoradi, Mohsen Rezapour, Mohammad R. Salavatipour, Approximation schemes for clustering with outliers, ACM Trans. Algorithms 15 (2) (February 2019).
[21] Nave Frost, Michal Moshkovitz, Cyrus Rashtchian, ExKMC: expanding explainable k-means clustering, CoRR, arXiv:2006.02399 [abs], 2020, https://arxiv.org/abs/2006.02399.
[22] Buddhima Gamlath, Xinrui Jia, Adam Polak, Ola Svensson, Nearly-tight and oblivious algorithms for explainable clustering, in: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual, 2021, pp. 28929–28939, https://proceedings.neurips.cc/paper/2021/hash/f24ad6f72d6cc4cb51464f2b29ab69d3-Abstract.html.
[23] Pierre Geurts, Nizar Touleimat, Marie Dutreix, Florence d'AlchéBuc, Inferring biological networks with output kernel trees, BMC Bioinform. 8 (2) (2007) 1–12.
[24] Badih Ghattas, Pierre Michel, Laurent Boyer, Clustering nominal data using unsupervised binary decision trees: comparisons with the state of the art methods, Pattern Recognit. 67 (2017) 177–185.
[25] David G. Harris, Thomas Pensyl, Aravind Srinivasan, Khoa Trinh, A lottery model for center-type problems with outliers, ACM Trans. Algorithms 15 (3) (2019) 1–25.
[26] Trevor Hastie, Robert Tibshirani, Generalized additive models, Stat. Sci. 1 (3) (1986) 297–318.
[27] Russell Impagliazzo, Ramamohan Paturi, Complexity of *k*-SAT, in: Proceedings of the 14th Annual IEEE Conference on Computational Complexity, Atlanta, Georgia, USA, May 4–6, 1999, IEEE Computer Society, 1999, pp. 237–240.
[28] Russell Impagliazzo, Ramamohan Paturi, Francis Zane, Which problems have strongly exponential complexity, J. Comput. Syst. Sci. 63 (4) (2001) 512–530.

[29] Yacine Izza, Alexey Ignatiev, Joao Marques-Silva, On tackling explanation redundancy in decision trees, J. Artif. Intell. Res. 75 (2022).
[30] Ravishankar Krishnaswamy, Shi Li, Sai Sandeep, 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, 2018, pp. 646–659.
[31] Amit Kumar, Yogish Sabharwal, Sandeep Sen, Linear-time approximation schemes for clustering problems in any dimensions, J. ACM 57 (2) (2010) 5, https://doi.org/10.1145/1667053.1667054.
[32] Eduardo Sany Laber, Lucas Murtinho, On the price of explainability for some clustering problems, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 5915–5925, http://proceedings.mlr.press/v139/.
[33] Zachary C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57.
[34] Yang Young Lu, Yingying Fan, Jinchi Lv, William Stafford Noble, DeepPINK: reproducible feature selection in deep neural networks, in: NeurIPS, 2018.
[35] Scott M. Lundberg, Su-In Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) 4765–4774.
[36] Meena Mahajan, Prajakta Nimbhorkar, Kasturi R. Varadarajan, The planar k-means problem is NP-hard, Theor. Comput. Sci. 442 (2012) 13–21, https://doi.org/10.1016/j.tcs.2010.05.034.
[37] Konstantin Makarychev, Liren Shan, Near-optimal algorithms for explainable k-medians and k-means, in: Proceedings of the 38th International Conference on Machine Learning (ICML), vol. 139, PMLR, 2021, pp. 7358–7367, http://proceedings.mlr.press/v139/.
[38] Konstantin Makarychev, Liren Shan, Explainable k-means: don't be greedy, plant bigger trees!, in: Stefano Leonardi, Anupam Gupta (Eds.), STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022, ACM, 2022, pp. 1629–1642.
[39] Ričards Marcinkevičs, Julia E. Vogt, Interpretability and explainability: a machine learning zoo mini-tour, arXiv preprint, arXiv:2012.01805, 2020.
[40] Christoph Molnar, Interpretable machine learning, Lulu.com, 2020.
[41] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, Nave Frost, Explainable k-means and k-medians clustering, in: Proceedings of the 37th International Conference on Machine Learning (ICML), vol. 119, PMLR, 2020, pp. 7055–7065, http://proceedings.mlr.press/v119/moshkovitz20a.html.
[42] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yu, Interpretable machine learning: definitions, methods, and applications, arXiv preprint, arXiv:1901.04592, 2019.
[43] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[44] C.S. Karthik, Bundit Laekhanukit, Pasin Manurangsi, On the parameterized complexity of approximating dominating set, J. ACM (ISSN 0004-5411) 66 (5) (August 2019), https://doi.org/10.1145/3325116.
[45] Avanti Shrikumar, Peyton Greenside, Anshul Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.
[46] Mukund Sundararajan, Ankur Taly, Qiqi Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.
[47] Berk Ustun, Cynthia Rudin, Supersparse linear integer models for optimized medical scoring systems, Mach. Learn. 102 (3) (2016) 349–391.
[48] Fulton Wang, Cynthia Rudin, Falling rule lists, in: Artificial Intelligence and Statistics, PMLR, 2015, pp. 1013–1022.