Explainable Clustering Beyond Worst Case Guarantees

Maximilian Fleissner MAXIMILIAN.FLEISSNER@TUM.DE
Maedeh Zarvandi MAEDEH.ZARVANDI@TUM.DE
Debarghya Ghoshdastidar GHOSHDAS@IN.TUM.DE
Technical University of Munich

Keywords: Explainability, Clustering

Abstract

In this paper we study the explainable clustering problem first posed by Dasgupta, Frost, Moshkovitz, and Rashtchian in their ICML 2020 work. The goal of explainable clustering is to fit a binary, axis-aligned decision tree with K leaves and minimal clustering cost (where every leaf is a cluster). The fundamental theoretical question in this line of work is the *price of explainability*, defined as the ratio between the cost of the best tree-based clustering and the optimal cost. Numerous papers have provided worst-case guarantees on this quantity. For the K-medians cost, it has recently been shown that the worst-case price of explainability is precisely $\Theta(\log K)$. While this settles the matter from a data-agnostic point of view, two important questions have remained unanswered: (i) Are tighter guarantees possible for well-clustered data? (ii) When can we trust decision trees to recover underlying cluster structures with high probability? In this paper, we place ourselves in the probabilistic setting of a mixture model ν and provide a statistical analysis for both of these questions. We prove that better guarantees on the price of explainability are indeed feasible for sufficiently well-clustered data, and also derive upper and lower bounds on how well decision trees can recover mixture models. Our algorithm takes as input a mixture model and constructs a suitable tree in data-independent time. We then extend our probabilistic analysis to explainable kernel clustering, deriving data-dependent guarantees that significantly improve over known worst case bounds for well-clustered data.

1. Introduction

Over the past decade, explainable machine learning has emerged as an active area of research, promising to cast light into black box algorithms. While a plethora of methods has been developed for the supervised setting, the explainability of unsupervised learning has attracted significantly less attention, despite models deployed in practice often leveraging large amounts of unlabeled data. Recently however, several researchers have begun to address the problem of explainable K-means or K-medians clustering. In their ICML 2020 work, Moshkovitz et al. (2020) suggested to cluster the data using binary axis-aligned decision trees with K leaves[1]. At every node of the tree, a one sided axis-aligned threshold cut $x_i \leq \theta$ partitions the data into two child nodes (here x_i denotes the i-th coordinate of x). The clustering cost of this tree is then compared to the optimal K-medians or K-means solution. Of course, unless all clusters are incidentally already separable by axis-aligned cuts, the tree has a strictly higher cost. The quality of the approximation is measured by the *price of explainability*, which is the ratio between the clustering cost induced by an optimal axis-aligned tree (where every leaf is a cluster) and the optimal clustering cost. Notably, Moshkovitz et al. (2020) were able to prove an upper bound on the price of explainability that depends only on the number

[1] Naturally, other notions of interpretable clustering models exist. We do not study these in this paper.

of clusters K, with $\mathcal{O}(K)$ guarantees for K-medians and $\mathcal{O}(K^2)$ guarantees for the K-means cost. These bounds hold for any dataset, regardless of how well it is clustered. Analyzing and gradually improving worst-case guarantees on the price of explainability has been of keen interest, and numerous groups have tackled the problem of finding tighter upper and lower bounds (see related works). Most recently, Makarychev and Shan (2023) were able to prove that the asymptotic price of explainability is precisely $\Theta(\log K)$ for K-medians. A similar analysis was given by Gupta et al. (2023). While this result settles the matter from a worst-case point of view, it leaves at least two important open questions that we tackle in this work.

First of all, all previously derived bounds are entirely *distribution-agnostic*. This is overly pessimistic: In practice, it is unlikely that we would ever need to explain a clustering model that doesn't cluster the data well (instead, the data scientist would most likely be asked to fit a better model). Indeed, Frost et al. (2020) empirically show that the price of explainability is actually close to 1 for several real-world clustering datasets. Interestingly, already Moshkovitz et al. (2020) ponder whether better guarantees are feasible for well-clustered data, such as Gaussian mixture models. On the other hand, as Gamlath et al. (2021) point out, the currently known examples which provide lower bounds on the price of explainability already consist of rather well-clustered instances, suggesting that care must be taken in the analysis.

Secondly, prior works do not give insight into when a decision tree accurately recovers the underlying distribution, in the sense that a sample is assigned to the ground truth cluster with high probability. We emphasize that this is a related, but not equivalent question to the one of the price of explainability: Even though a tree may recover the underlying clusters almost perfectly, the price of explainability can be lower bounded by a constant > 1 for mixture models (we will come back to this later). Understanding when trees are capable of recovering cluster structures is crucial. Otherwise, can we truly trust the interpretable model to provide faithful explanations of the data?

Contributions. We introduce an explainability-to-noise ratio $ENR(\nu)$ for mixture models ν and show that it is $ENR(\nu)$ that influences the price of explainability of K-medians, as well as the ability of a tree to recover a given mixture model. We prove upper and lower bounds that depend on the number of clusters K and $ENR(\nu)$, and are tight in the latter. Our upper bounds are attained by the Mixture Model Decision Tree (MMDT) algorithm that we propose. Given a mixture model ν, it constructs a decision tree such that every leaf represents exactly one of the mixture components. Importantly, it runs in data-independent time and is much faster than classification and regression trees (CART) (Breiman, 2017). Furthermore, we use our statistical analysis to derive tighter guarantees on the price of explainability for kernel clustering. Our bounds significantly improve over existing worst case guarantees, explaining why the price of explainability is often quite low on real world clustering datasets.

Related work. Starting with Moshkovitz et al. (2020), a number of works have tackled the explainable clustering problem. Most works focus on a tighter analysis of the worst case price of explainability (Makarychev and Shan, 2022; Gamlath et al., 2021; Esfandiari et al., 2022; Makarychev and Shan, 2023; Gupta et al., 2023). Some results also incorporate the role of the dimension d into the analysis (Charikar and Hu, 2022; Laber and Murtinho, 2021). Other works on interpretable clustering loosen the requirement of exactly K leaves. Frost et al. (2020) empirically demonstrate that this improves the practical performance, and Makarychev and Shan (2022) formally prove it. Furthermore, Fleissner et al. (2024) theoretically analyze the worst case price of explainability for kernel clustering, obtaining essentially $\mathcal{O}(dK^2)$ bounds for the Gaussian kernel.

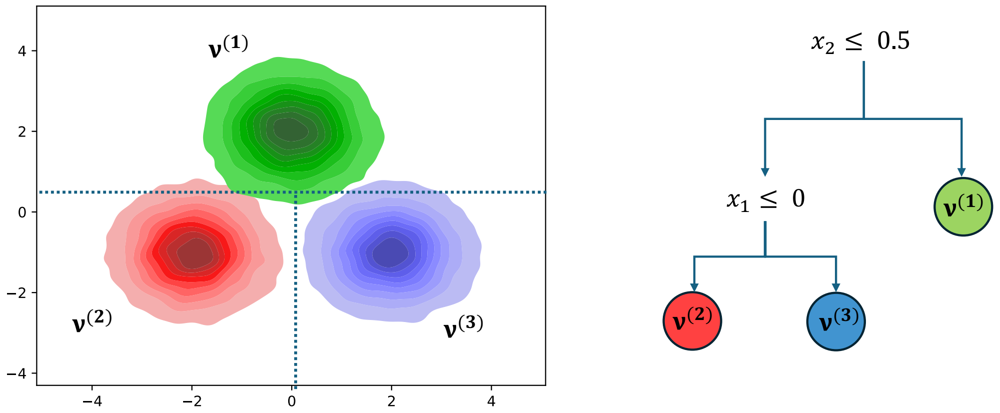

Figure 1: Illustration of our objective: Given three distributions $\nu^{(1)}, \nu^{(2)}, \nu^{(3)}$, we investigate how well a decision tree can cluster these components.

2. Problem Setup and Definitions

We begin by formalizing the model, discussing the problem statement, and introducing key definitions and notation.

Data model. We assume that we are given a mixture model $\nu = \sum_{k=1}^{K} p^{(k)} \nu^{(k)}$, where each $\nu^{(k)}$ is a discrete or absolutely continuous probability distribution on $\mathbb{R}^d$, and the mixing weights satisfy $\sum_{k=1}^{K} p^{(k)} = 1$. We denote by $\mu^{(k)}$ the mean of mixture component $\nu^{(k)}$, and assume that the probability density or probability mass function of each $\nu^{(k)}$ is symmetric around its mean, so that means and medians coincide for all k. We assume further that every mixture component has finite variances given by some positive real numbers $\sigma_1^2, \dots, \sigma_d^2$. We use the superscript for enumerating properties of the mixture components, and the subscript to refer to axes of the ambient space $\mathbb{R}^d$. We assume that the the mixing weights $p^{(k)}$ are safely bounded away from 1, satisfying $p^{(k)} \leq \alpha / K$ for some fixed $\alpha \geq 1$ independent of K.

Problem statement. Given a mixture model ν on $\mathbb{R}^d$, the goal is to explain each of its components in terms of individual features. Since unsupervised learning by its very *raison d'être* is concerned with extracting global patterns in data, and not with prediction on instances, it is important that these explanations provide a global perspective on the underlying data. In the explainable clustering problem, one chooses decision trees that sequentially partition the data into K leaves, selecting at most K features. Formally, given a mixture model ν, our goal is therefore to construct an axis-aligned decision tree with K leaves such that each leaf approximates exactly one of the K mixture components $\nu^{(1)}, \dots, \nu^{(K)}$. At every internal node of the tree, a one-sided threshold cut $x_i \leq \theta$ partitions the data into two child nodes. Here, we denote x_i for the i-th coordinate of the point x. We ensure correspondence between leafs and mixture components by constraining the decision trees to satisfy that each leaf contains exactly one mean $\mu^{(k)}$. While this need not be enforced, all existing algorithms also demand every leaf to contain exactly one of the K baseline cluster centers. Figure 2 illustrates our setting.

Price of explainability for mixtures. In this paper, we analyze the ratio between the cost of the tree and the cost of the underlying mixture model, defined as follows.

Definition 1 ***(Price of explainability for mixtures)*** *Given a mixture model ν as specified above, and a decision tree T that partitions $\mathbb{R}^d$ into K leaves each containing one of the means $\mu^{(k)}$, we define*

$$Price(\nu, T) = \frac{\mathbb{E}_{x\sim\nu}\left[\|x - \tilde{\mu}(x)\|_1\right]}{\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_1\right]} \tag{1}$$

as the price of explainability. Here, $\mu(x) = \mu^{(k)}$ if x is sampled from $\nu^{(k)}$, and $\tilde{\mu}(x)$ is the median of the leaf that x ends up in. We denote further $\hat{\mu}(x)$ for the mean $\mu^{(k)}$ that ends up in the same leaf as x. This need not be $\tilde{\mu}(x)$.

This is a direct adaption of the price of explainability to mixture models. However, in contrast to the clustering setting, the reference partition (i.e. the denominator) that we compare the decision tree against is the K-medians cost of assigning to each $x \sim \nu$ the median of the mixture component from which x is drawn (i.e. one of the $\mu^{(k)}$). This is not necessarily the optimal K-medians partition associated with ν. The reason is that the optimal centers of the K-medians cost

$$cost_{opt}(\nu) := \min_{c^{(1)},\ldots,c^{(K)}} \mathbb{E}_{x\sim\nu}\left[\min_{k\in[K]} \|x - c^{(K)}\|_1\right] \tag{2}$$

are typically **not** given by the set $\{\mu^{(k)}\}_{k=1}^K$. As an example, think of ν being a mixture of two Gaussians in $\mathbb{R}$, with means $\mu^{(1)} = 1 = -\mu^{(2)}$ and unit variance. The minimizers of $cost_{opt}(\nu)$ are in general not ± 1, because some $x \sim \nu$ will be sampled closer to the "other" mean, making it beneficial to push $c^{(1)}, c^{(2)}$ further apart. While this effect could potentially be incorporated into the analysis, we choose not to do so. In a mixture model setting, we are typically interested in the mixture components $\{\nu^{(k)}\}_{k=1}^K$ and not the optimal population clusters, and most works on mixture modeling aim to recover the true means and covariances, not the clustering (Dasgupta, 1999; Ashtiani et al., 2020). In other words, our ground truth to compare an explainable clustering method against is the set of means $\{\mu^{(k)}\}_{k=1}^K$ of the mixture components, and we therefore stick to Definition 1 in this paper.

Error rate. Next to the price of explainability, we also consider the error rate of T in this paper, that is $P_{x\sim\nu}\left(\hat{\mu}(x) \neq \mu(x)\right)$. It quantifies how well the decision tree recovers the underlying mixture components.

Explainability-to-noise ratio. Prior works that study the explainable clustering problem are distribution-independent. In this paper, we move beyond worst-case guarantees by incorporating information on ν into our analysis. In particular, we introduce a novel quantity that governs both the approximation ratio $Price(\nu, T)$ as well as the error rate $P_{x\sim\nu}\left(\hat{\mu}(x) \neq \mu(x)\right)$, which we call the explainability-to-noise ratio of the mixture model ν. It can be viewed as an axis-aligned version of the classic signal-to-noise ratio.

Definition 2 ***(Explainability-to-noise ratio)*** *For a mixture model ν, we define its explainability-to-noise ratio as*

$$ENR(\nu) := \min_{k\neq l} \max_{j\in[d]} \left(\frac{|\mu_j^{(k)} - \mu_j^{(l)}|^2}{\sigma_j^2}\right) \tag{3}$$

Intuitively, a large explainability-to-noise ratio ensures that at any node of a decision tree that is not a leaf, we can find a pair of remaining means that can be separated along some axis $i \in [d]$ without increasing the probability $P_{x\sim\nu}(\mu(x) \neq \hat{\mu}(x))$ too much. Note that if the signal-to-noise ratio of ν is given by $SNR(\nu) = \min_{k\neq l} \sum_{j=1}^{d} \sigma_j^{-2} \left(\mu_j^{(k)} - \mu_j^{(l)}\right)^2$ then certainly $ENR(\nu) \geq SNR(\nu)/d$.

3. Proposed Algorithm

Algorithm 1 Mixture Model Decision Tree (**MMDT**)

```
Input: Mixture components $\{\nu^{(k)}\}_{k=1}^{K}$ with means $\{\mu^{(k)}\}_{k=1}^{K} \in \mathbb{R}^d$ and coordinate-wise variances $\{\sigma_i^2\}_{i=1}^{d}$.
Output: Decision tree with $K$ leaves.
Initialize list of nodes with a single array $\mathcal{L} \leftarrow [K]$.
Initialize empty list of threshold cuts $\mathcal{T}$.
$\boldsymbol{Z} \leftarrow$ matrix with columns $\mu^{(k)}$ for $k \in [K]$.
while $|\mathcal{L}| < K$ do
  for $\boldsymbol{N} \in \mathcal{L}$ do
    $i(\boldsymbol{N}) = \text{argmax}_{i\in[d]} \left\{\sigma_i^{-1} \max_{k,l\in\boldsymbol{N}} \left|\mu_i^{(k)} - \mu_i^{(l)}\right|\right\}$
    Pick $\theta(\boldsymbol{N})$ as in Equation (4) or, when exact densities are unknown, (5).
    Append to $\mathcal{T}$ a new binary cut described through the tuple $(\boldsymbol{N}, i(\boldsymbol{N}), \theta(\boldsymbol{N}))$.
    Update $\mathcal{L}$ with two new child nodes $\boldsymbol{N}_1, \boldsymbol{N}_2$ in place of $\boldsymbol{N}$.
  end for
end while
```

We propose the Mixture Model Decision Tree (MMDT) algorithm to obtain an explainable approximation from any given mixture model. MMDT is described in Algorithm 1, and only requires the mixture model ν with its means $\mu^{(1)}, \ldots, \mu^{(k)}$ and coordinate-wise variances $\sigma_1^2, \ldots, \sigma_d^2$ as input. It iteratively partitions the means into K leaves, each of which eventually contains exactly one mean, representing one mixture component. At every node t with a set of remaining mixture components $N(t) \subset [K]$, this is achieved by identifying the axis $i \in [d]$ along which the separation between a pair of remaining means (projected to that axis) is highest, after normalizing by the coordinate-wise standard deviation σ_i. The binary cut is then chosen at $\theta \in \mathbb{R}$ such as to minimize the probability that any $x \sim \sum_{k\in N(t)} \bar{p}^{(k)}\nu^{(k)} =: \nu(N)$ is lowest, where $\bar{p}^{(k)}$ simply rescales $p^{(k)}$ to ensure the weights of $\nu(N)$ still sum to one. Formally,

$$\theta = \text{argmin} \left\{ P_{x\sim\nu(N)} \left(x \text{ is separated from } \mu(x) \text{ through } \theta\right) : \min_{k\in N(t)} \mu_i^{(k)} < \theta < \max_{k\in N(t)} \mu_i^{(k)} \right\} \quad (4)$$

If multiple possible thresholds exist, one is chosen randomly. If the exact probability density of each $\nu^{(k)}$ is not known, then θ is computed by instead minimizing an upper bound on the probability. Such can always be obtained using Chebyshev's inequality, which guarantees that

$$P_{x\sim\nu^{(k)}} \left(x \text{ is separated from } \mu(x) \text{ through } \theta\right) \leq \sigma_i^2 \cdot \left|\mu_i^{(k)} - \theta\right|^{-2} \quad (5)$$

Under stronger concentration such as a Gaussian mixture model, we could instead plug in Gaussian concentration bounds into Equation (5). Note that in both cases, all we really need to know are the means and variances of the mixture components. Therefore, θ can be found in data-independent time.

4. Theoretical Analysis of MMDT

In this section, we theoretically analyze Algorithm 1 regarding the price of explainability for K-medians (4.1) and the error rate (4.2).

4.1. Price of explainability

We begin by analyzing $Price(\nu, T)$, giving both upper and lower bounds. We need an assumption on the first absolute centralized moment.

Assumption 3 *We assume that the mixture model ν satisfies the following property: There exists $\beta \in \mathbb{R}$ such that for all $i \in [d]$*

$$\beta \cdot \mathbb{E}_{x\sim\nu}\left[|x_i - \mu_i(x)|\right] \geq \sqrt{\mathbb{E}_{x\sim\nu}\left[|x_i - \mu_i(x)|^2\right]} \tag{6}$$

In particular, this implies that the K-medians cost satisfies

$$\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_1\right] \geq \beta^{-1} \cdot \sum_{j=1}^{d} \sigma_j \tag{7}$$

Note that certainly $\beta \geq 1$ by Jensen's inequality.

Assumption 3 essentially asks the baseline cost to be lower bounded in terms of the standard deviations $\sigma_1, \ldots, \sigma_d$. For Gaussian mixture models, we can always pick $\beta = 1$. We now state our first result on the price of explainability for mixture models.

Theorem 4 (Upper bounds on price) *Consider a mixture model ν as specified before. Then, there exists a decision tree T such that*

$$Price(\nu, T) \leq 1 + \frac{(4 + 2\pi^2/3)\alpha\beta K(K-1)}{\sqrt{q}} \tag{8}$$

where q is the explainability-to-noise ratio of ν.

Proof We give a proof sketch here and include the full result in Appendix A.1. The main idea is to bound the probability that $x \sim \nu$ is separated from $\mu(x)$ for all nodes t of the tree. Because $ENR(\nu) = q$, we can always find an axis $i \in [d]$ for which at least one pair of means is separated by $R_i(t) \geq \sigma_i\sqrt{q}$, where we define $R_j(t) = \max_{k,l\in N(t)} \mu_j^{(k)} - \mu_j^{(l)}$ for all $j \in [D]$ and $N(t)$ is the set of means that arrive at node t. Following Equation (5), for any threshold value θ chosen along axis i, Chebyshev's inequality ensures

$$P_{x\sim\nu}\left(x \text{ is separated from } \mu(x) \text{ through } \theta\right) \leq \sigma_i^2 \sum_{k=1}^{K'} p^{(k)} \cdot \left|\mu_i^{(k)} - \theta\right|^{-2} \tag{9}$$

where we without loss of generality assume that $N(t) = [K']$. It turns out that an upper bound on this probability is given by the configuration of means where $K' = K$ and all $\mu^{(k)}$ are equidistant along axis i, with $\left|\mu_i^{(k)} - \mu_i^{(l)}\right| = \frac{R_i(t)}{K}$ for all $k \neq l$. We prove this in the appendix. In that case,

$$P_{x\sim\nu}\left(x \text{ is separated from } \mu(x) \text{ through } \theta\right) = \mathcal{O}(\alpha K \sigma_i^2 / R_i(t)^2) \tag{10}$$

where we exploited $p^{(k)} \leq \alpha/K$ and the fact that the series of reciprocal squares converges. As previous works on explainable clustering, we upper bound the cost of the decision tree as

$$\mathbb{E}_{x\sim\nu}\left[\|x - \tilde{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\|x - \hat{\mu}(x)\|_1\right] \tag{11}$$

where we denote $\hat{\mu}(x)$ for the mean that ends up in the same leaf as x. This is well-defined because Algorithm 1 ensures every leaf contains exactly one of the original K means. Simplifying further,

$$\mathbb{E}_{x\sim\nu}\left[\|x - \hat{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_1\right] + \mathbb{E}_{x\sim\nu}\left[\|\mu(x) - \hat{\mu}(x)\|_1\right] \tag{12}$$

Now, we bound the second term by noting that if x is separated from $\mu(x)$ at node t, then $\|\mu(x) - \hat{\mu}(x)\|_1 \leq \sum_{j=1}^d R_j(t)$. Therefore,

$$\mathbb{E}_{x\sim\nu}\left[\|\mu(x) - \hat{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\sum_{t\in T} \mathbf{1}(x \text{ separated from } \mu(x) \text{ at } t) \cdot \sum_{j=1}^d R_j(t)\right] \tag{13}$$

$$= \sum_{t\in T} P_{x\sim\nu}\left(x \text{ is separated from } \mu(x) \text{ at } t\right) \cdot \sum_{j=1}^d R_j(t) \tag{14}$$

$$\leq \sum_{t\in T} \mathcal{O}\left(\alpha K \sigma_i^2 R_i(t)^{-2}\right) \cdot \sum_{j=1}^d R_j(t) \tag{15}$$

where we plugged in the upper bound on the probability from before. Since $\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_1\right] \geq \beta^{-1}\sum_{j=1}^d \sigma_j$ by Assumption 3 and exploiting the fact that i is chosen such that

$$\frac{\sum_{j=1}^d R_j(t)}{\sum_{j=1}^d \sigma_j} \leq \max_{j\in[d]} \frac{R_j(t)}{\sigma_j} =: \frac{R_i(t)}{\sigma_i} \tag{16}$$

for all nodes t, we obtain

$$Price(\nu, T) = \sum_{t\in T} \mathcal{O}\left(\alpha\beta K \sigma_i / R_i(t)\right) \tag{17}$$

Plugging in $R_i(t)/\sigma_i \geq \sqrt{q}$ at all nodes t and using the fact that there are no more than $K-1$ nodes, we obtain

$$Price(\nu, T) = \mathcal{O}\left(\alpha\beta K(K-1)/\sqrt{q}\right) \tag{18}$$

as desired. ■

In particular, Theorem 4 explains why for Gaussian mixture models, the price of explainability approaches 1 as the variance of the Gaussians shrinks. The next result proves a lower bound that is logarithmic in K and tight in $ENR(\nu)$. It relies on existing $\Omega(\log K)$ bounds for K-medians while ensuring that $ENR(\nu) = q$ and $\beta = \sqrt{q}/2$.

Theorem 5 (Lower bounds on price) *Let $K \geq 2$ and $q \geq 2K^3$ be arbitrary. Then, there exists a mixture model ν with explainability-to-noise ratio given by $ENR(\nu) = q$ such that*

$$Price(\nu, T) = \Omega(\alpha\beta \log(K)/\sqrt{q}) \tag{19}$$

for any decision tree with K leaves, each of which contains exactly one the mean of some $\nu^{(k)}$. The mixture components have symmetric densities, $\alpha = 1$ and $\beta = \sqrt{q}/2$.

The proof is included in Appendix A.2. We conclude this subsection with two remarks.

Remark 6 *The best known worst case guarantees for K-medians rely on a random cut algorithm. When all means $\mu^{(k)}$ are contained in a box $[-M, M]^d \subset \mathbb{R}^d$, it sequentially chooses the threshold cut (i, θ) with uniform probability from the set $[d] \times [-M, M]^d$, and discards cuts that do not partition any means. We do not adopt this approach for our probabilistic analysis: Under Chebyshev's inequality, $Price(\nu, T)$ no longer retains information on $ENR(\nu)$, as simple computations reveal. Then again, Chebyshev's inequality itself is an upper bound, and there may very well exist classes of distributions for which the random cut attains rates that decay with $ENR(\nu)$.*

Remark 7 *While our proof technique extends to the explainable K-means clustering problem, the guarantees we obtain are no longer dependent on $ENR(\nu)$. This can be resolved by using concentration inequalities of higher moments, and adjusting Assumption 3. See Appendix A.3.*

4.2. Error rates

Our probabilistic analysis also sheds light on the error rate. We obtain bounds that have a linear gap in the number of mixture components K, and are tight in $ENR(\nu)$. This further justifies our definition of the explainability-to-noise ratio introduced earlier.

Theorem 8 (Upper bounds on error rate) *Consider a mixture model ν as specified before. Then, Algorithm 1 returns a decision tree with*

$$P_{x\sim\nu}\left(\hat{\mu}(x) \neq \mu(x)\right) \leq \frac{(4 + 2\pi^2/3)\alpha K(K-1)}{q} \tag{20}$$

where the explainability-to-noise ratio of the mixture ν is given by $q > 0$.

The proof follows exactly the same procedure as Theorem 4, except we do not need to take care of the cost incurred by the event that $\mu(x) \neq \hat{\mu}(x)$. For details, see Appendix B.1. We also give a lower bound.

Theorem 9 (Lower bounds on error rate) *Let $q \geq K$ be arbitrary. Then, there exists a mixture model ν with K components, equal mixing weights and an explainability-to-noise ratio $ENR(\nu) = q$ such that*

$$P_{x\sim\nu}\left(\hat{\mu}(x) \neq \mu(x)\right) \geq \frac{K-1}{4q} \tag{21}$$

for any tree with K leaves that contains at least one mean in every leaf.

The proof is in Appendix B.2. With these results in hand, it is natural to ask what the relationship between $Price(\nu, T)$ and $P_{x\sim\nu}\left(\hat{\mu}(x) \neq \mu(x)\right)$ is. As mentioned earlier, it can happen that the price stays bounded away from 1 despite the error rate of the tree approaching zero. We formally prove this in Appendix B.3.

5. Extension to Kernel Clustering

Kernel clustering is a nonparametric clustering technique based on the theory of reproducing kernel Hilbert spaces (Smola and Schölkopf, 1998). Intuitively, it relies on a suitable kernel function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that measures the similarity between points $x, x' \in \mathbb{R}^d$. The potential nonlinearity of κ allows discovering nonlinear cluster structures, which is impossible with ordinary K-means or K-medians. For kernel K-means with the Gaussian kernel $\kappa(x, x') = \exp(-\gamma\|x - x'\|^2)$ it has been shown that the price of explainability is essentially $\mathcal{O}(dK^2)$ in the worst case (Fleissner et al., 2024). *In this section, we extend our probabilistic analysis of explainable clustering to kernels, providing us with tighter, distribution-dependent guarantees*. The first step is to reformulate our problem in a nonparametric setting using kernel mean embeddings. We then introduce additional assumptions and definitions specific to this section.

Kernel mean embeddings. When a kernel κ is positive semi-definite, there exists a Hilbert space $\mathcal{H}$ of functions and a feature map $\phi : \mathbb{R}^d \to \mathcal{H}$ such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \mathbb{R}^d$. As before, let ν be a distribution on $\mathbb{R}^d$ with $\mathbb{E}_{x\sim\nu}\left[\sqrt{\kappa(x, x)}\right] < \infty$. Then, the Bochner integral $\phi_\nu := \mathbb{E}_{x\sim\nu}[\phi(x)] \in \mathcal{H}$ is referred to as the kernel mean embedding of the distribution ν. For any $f \in \mathcal{H}$ it holds that $\mathbb{E}_{x\sim\nu} f(x) = \langle f, \phi_\nu \rangle$. Given two distributions ν, ν', we define the maximum mean discrepancy (MMD) as $d_\kappa(\nu, \nu') := \|\phi_\nu - \phi_{\nu'}\|_{\mathcal{H}}$. The MMD has numerous applications, such as two-sample testing (Gretton et al., 2012) and can be used to prove that kernel clustering recovers nonparametric mixture models (Vankadara et al., 2021). For more details on kernel mean embeddings, see Muandet et al. (2017).

Assumptions and definitions. We make three assumptions on the mixture ν and the kernel κ. Firstly, we assume that the kernel κ is a bounded, distance-based product kernel

$$\kappa(x, x') = \prod_{i=1}^{d} \kappa_i(x, x') = \prod_{i=1}^{d} g_i(|x_i - x'_i|) \tag{22}$$

where each $g_i : [0, \infty) \to \mathbb{R}$ is a positive, monotone decreasing function satisfying $g_i(0) = 1$. Examples include the Gaussian kernel with $g_i(t) = \exp(-t^2)$ for all i and the Laplace kernel with $g_i(t) = \exp(-t)$ for all i. Secondly, we assume that the components $\nu^{(k)}$ of the mixture model ν all satisfy

$$\|\phi_{\nu^{(k)}}\|_{\mathcal{H}}^2 = \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}}\left[\kappa(x, x')\right] = \sigma^2 \tag{23}$$

for some $\sigma^2 \in (0, 1)$, and that $\nu^{(k)} \neq \nu^{(l)}$ for all $k \neq l$. Thirdly, we assume that for all $i \in [d]$ and all pairs $k, l \in [K]$, the variance of κ_i is bounded by

$$\mathbb{V}_{x\sim\nu^{(k)}, x'\sim\nu^{(l)}}\left[\kappa_i(x, x'))\right] \leq \epsilon^2 \tag{24}$$

for some $\epsilon > 0$. Since $\|\kappa_i\|_\infty = 1$, we can certainly choose $\epsilon^2 = 1$, although this will in general not provide useful bounds. From here, we define the price of explainability of a mixture model with respect to a kernel function κ as

$$Price_\kappa(T, \nu) = \frac{E_{x\sim\nu}\left[\|\phi(x) - \tilde{\mu}(x)\|_{\mathcal{H}}^2\right]}{E_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} \tag{25}$$

Here, $\tilde{\mu}^{(l)}$ denotes the kernel mean embedding of $x \sim \nu$ conditioned on the event that x arrives in the l-th leaf of the decision tree, and $\tilde{\mu}(x) = \tilde{\mu}^{(l)}$ if and only if x ends up in the l-th leaf. Since the the map $h \mapsto \mathbb{E}\left[\|\phi(x) - h\|_{\mathcal{H}}^2\right]$ has its minimum at the mean $h = \mathbb{E}[\phi(x)]$, we can certainly upper bound the price via

$$Price_\kappa(T, \nu) \leq \frac{E_{x\sim\nu}\left[\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2\right]}{E_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} \tag{26}$$

where $\hat{\mu}(x) = \phi_{\nu^{(l)}}$ if x ends up in a leaf that is assigned to the l-th mixture component $\nu^{(l)}$.

Explainable kernel clustering. Note that it is not possible to *directly* extend MMDT (1) or even the definition of the explainability-to-noise ratio $ENR(\nu)$ to the kernel setting, since the kernel mean embeddings $\phi_{\nu^{(k)}}$ of the mixture components are now elements of the Hilbert space $\mathcal{H}$, not $\mathbb{R}^d$. But for our decision tree to be interpretable, its axis-aligned cuts naturally need to be in the input space, not $\mathcal{H}$. Therefore, we instead give our guarantees on $Price_\kappa(\nu, T)$ in terms of the following axis-aligned quantity.

Definition 10 ***(Axis-aligned κ-similarity of the mixture components)*** *For a mixture model ν and a kernel κ as introduced before, we define*

$$\tau = \max_{k \neq l} \min_{i \in [d]} \mathbb{E}_{x\sim\nu^{(k)}, y\sim\nu^{(l)}}\left[\kappa_i(x, y)\right] \tag{27}$$

as the axis-aligned κ-similarity of the mixture components.

The axis-aligned κ-similarity plays the role of the $ENR(\nu)$ in a nonparametric setting, except that explainability becomes easier as it decreases. It can be upper bounded by the MMD between the mixture components, capturing the idea that well-clustered data is easier to explain even in kernel clustering. Indeed, the following result is proved in Appendix C.1.

Lemma 11 *Consider a mixture model ν as specified before. Suppose that for any two distinct $\nu^{(k)}, \nu^{(l)}$ it holds that $d_\kappa(\nu^{(k)}, \nu^{(l)}) \geq \gamma$. Then, ν has a κ-similarity at most $\tau \leq \sigma^2 - \frac{\gamma^2}{2d}$. In particular, this implies that $\tau < \sigma^2$ since all $\nu^{(k)}$ are distinct.*

The main hurdle facing us in adapting the MMDT-algorithm to the kernel setting is that, as discussed before, the kernel means are no longer elements of the space our tree is aiming to partition. Of course, one could in principle still operate on the means $\mu^{(k)} = \mathbb{E}_{x\sim\nu^{(k)}}[x] \in \mathbb{R}^d$. But these are in general not informative, since the clustering cost $\mathbb{E}_{x\sim\nu^{(k)}}[\|\phi(x) - \phi_{\nu^{(k)}}\|_{\mathcal{H}}^2$ is measured in the Hilbert space. We resolve this by instead choosing prototype points $x^{(k)} \sim \nu^{(k)}$ for each mixture component at every node, and deciding based on these $x^{(k)}$ whether a mixture component goes to the left or the right child node. The intuition is that $\kappa(x^{(k)}, x')$ is large for x' from the same cluster as $x^{(k)}$, and small otherwise. One thing changes compared to the previous section: The axis-aligned threshold cuts will no longer be one sided $x_i \leq \theta$ but rather two-sided intervals $|x_i - x_i^{(k)}| \leq \theta$. This does not harm the interpretability of the resulting tree, as we are still only checking one axis at every node. As before, the tree T is fitted sequentially. Suppose we have a set $N(t) \subset [K]$ of remaining mixture components at a node t. We choose an axis $i \in [d]$ and $k, l \in N(t)$ that minimize

$$\xi(i, k, l) := \mathbb{E}_{x\sim\nu^{(k)}, y\sim\nu^{(l)}}[\kappa_i(x, y)] \tag{28}$$

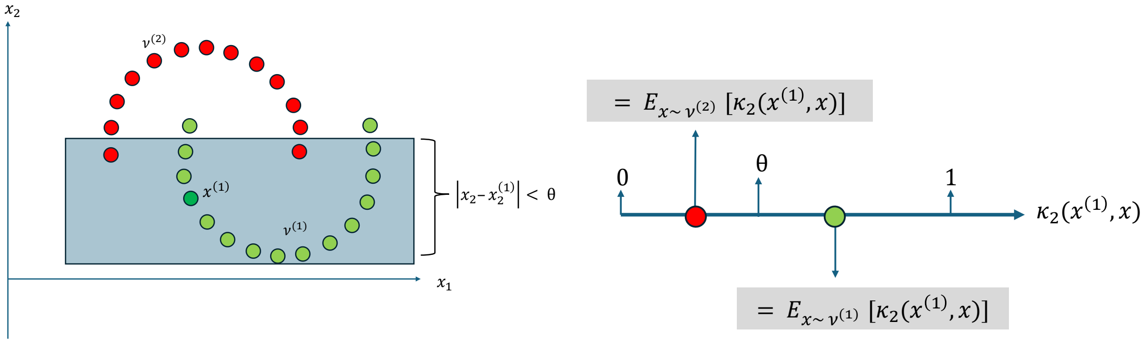
x_2
$\nu^{(2)}$
$x^{(1)}$
$\nu^{(1)}$
$|x_2 - x_2^{(1)}| < \theta$
x_1
$= E_{x\sim\nu^{(2)}}[\kappa_2(x^{(1)},x)]$
0
θ
1
$\kappa_2(x^{(1)},x)$
$= E_{x\sim\nu^{(1)}}[\kappa_2(x^{(1)},x)]$

Figure 2: Example for Kernel MMDT with two mixture components. When projected to axis $i = 2$, the components can be separated well by thresholding with respect to the distance to a randomly sampled point $x^{(1)} \sim \nu^{(1)}$.

Note that certainly $\xi(i,k,l) \leq \tau$. A threshold value $\xi(i,k,l) < \theta < \sigma^2$ is decided as follows: Sort the expectations $\{\xi(i,k,l)\}_{l \in N(t)\setminus\{k\}}$ in non-decreasing order. Without loss of generality, $k = 1$ and $\xi(i,k,l) \leq \xi(i,k,l+1) \leq \sigma^2$ for all $2 \leq l \leq |N(t)| - 1$. Then, θ is chosen halfway between the two expectations that are separated furthest. Finally, we sample $x^{(k)} \sim \nu^{(k)}$ as a reference point for the node. Now, for any new $x \sim \nu$, we send it left if $\kappa_i(x, x^{(k)}) < \theta$ and right otherwise. To keep track of which mixture components go to which child node of t, we simply evaluate on which side of θ they are located. Since $\sigma^2 \leq \xi(i,k,k)$ (this follows from the proof of Lemma 11, see Appendix C.1) the k-th mixture component certainly goes to the right leaf. Formally, we partition $N(t)$ into child nodes t_{left}, t_{right} by defining

$$N(t_{left}) = \{l \in N(t) : \xi(i,k,l) < \theta\} \tag{29}$$

$$N(t_{right}) = \{l \in N(t) : \xi(i,k,l) > \theta\} \tag{30}$$

From there, the procedure continues at both child nodes. Eventually, each node only contains a single index $k \in [K]$. All points in this leaf are then assigned to the k-th mixture, that is $\hat{\mu}(x) = \phi_{\nu^{(k)}}$. The algorithm is illustrated in Figure 5 for $K = 2$. Using similar ideas as in Section 4, we can relate the price of explainability to the K, σ^2, ϵ^2 and the axis-aligned κ-alignment τ. The proof is in Appendix C.2.

Theorem 12 ***(Price of explainability for kernels)*** *Consider a mixture model ν that satisfies the conditions discussed in this section. Denote by $\tau \in [0, \sigma^2)$ its axis-aligned κ-alignment. Then, it holds that*

$$Price_\kappa(\nu, T) \leq 1 + \frac{1+\sigma^2}{1-\sigma^2} \cdot \max\left(1, \frac{(4 + 2\pi^2/3)\alpha\epsilon^2 K(K-1)}{\sigma^2 - \tau}\right) \tag{31}$$

where T is the tree fitted using the above algorithm, and the expectation is over randomly sampled prototypes $x^{(k)} \sim \nu^{(k)}$ at every node.

We remark that if $\sigma^2 \to 1$, then also $\epsilon^2 \to 0$ because all points will concentrate around K unique centers. Also note that as τ decreases, the upper bound improves, and explainability becomes easier.

6. Discussion and Conclusion

This paper provides the first statistical analysis of the explainable clustering problem. While a significant number of previous works have focused on worst case bounds, we are the first to show that it is possible to obtain tighter guarantees on the price of explainability by incorporating information on the data. We do so by introducing a new quantity that we refer to as the explainability-to-noise ratio of a mixture model. Our work explains the empirical phenomenon that the price of explainability is very close to 1 for well-clustered data, and we provide a new algorithm with provable guarantees and data-independent runtime.

It is noteworthy that our results even extend to a kernel setting, and both the Gaussian as well as the Laplace kernel fall under the umbrella of our analysis. Given that both are characteristic kernels capable of distinguishing between arbitrary distributions (Sriperumbudur et al., 2008), one might intuitively not expect interpretable approximations to exist. Our results provide a more nuanced look on this matter, demonstrating that axis-aligned cuts can work even for nonparametric mixture models provided the average similarity between samples from distinct mixture components is small. On the technical side, we believe that our approach to provably explainable kernel models extends well beyond clustering, and it would be quite interesting to explore this further. In addition, we see the following opportunities for future work.

Explicitly dealing with finite sample sizes. In this paper, we assume exact knowledge of the means and variances of the mixture model ν. This is the probabilistic equivalent of assuming knowledge of the true K-means or K-medians clusters, as is done in prior works. However, an immediate extension of our work would be to incorporate finite sample complexity bounds on estimation of mixture models, e.g. when every component is Gaussian (Ashtiani et al., 2020). Since our proof relies on the explainability-to-noise ratio $ENR(\nu)$ that depends only on means and variances, we do not believe that our results change for reasonably large sample sizes. Then again, it may not always be possible to exactly estimate ν. This touches on the important question of identifiability in mixture models (Aragam et al., 2020).

Tighter bounds. An obvious challenge for future work is to tighten our upper and lower bounds from Section 4. Additionally, one could incorporate the role of the dimension d into the analysis. In particular, when $d < K$, better guarantees are most likely feasible. Similar questions pose themselves in the kernel setting.

Acknowledgments

We are supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, and the German Research Foundation (Research Grant GH257/4-1).

References

Bryon Aragam, Chen Dan, Eric P Xing, and Pradeep Ravikumar. Identifiability of nonparametric mixture models and bayes optimal clustering. 2020.

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

Heinz Bauer. Minimalstellen von funktionen und extremalpunkte. *Archiv der Mathematik*, 9(4): 389–393, 1958.

Leo Breiman. *Classification and regression trees*. Routledge, 2017.

Moses Charikar and Lunjia Hu. Near-optimal explainable k-means for all dimensions. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2580–2606. SIAM, 2022.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. Almost tight approximation algorithms for explainable clustering. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2641–2663. SIAM, 2022.

Maximilian Fleissner, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Explaining kernel clustering via decision trees. *arXiv preprint arXiv:2402.09881*, 2024.

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

Buddhima Gamlath, Xinrui Jia, Adam Polak, and Ola Svensson. Nearly-tight and oblivious algorithms for explainable clustering. *Advances in Neural Information Processing Systems*, 34: 28929–28939, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Anupam Gupta, Madhusudhan Reddy Pittu, Ola Svensson, and Rachel Yuan. The price of explainability for clustering. *arXiv preprint arXiv:2304.09743*, 2023.

Eduardo S Laber and Lucas Murtinho. On the price of explainability for some clustering problems. In *International Conference on Machine Learning*, pages 5915–5925. PMLR, 2021.

Konstantin Makarychev and Liren Shan. Explainable k-means: don't be greedy, plant bigger trees! In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1629–1642, 2022.

Konstantin Makarychev and Liren Shan. Random cuts are optimal for explainable k-medians. *arXiv preprint arXiv:2304.09113*, 2023.

Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In *International conference on machine learning*, pages 7055–7065. PMLR, 2020.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective hilbert space embeddings of probability measures. In *21st annual conference on learning theory (COLT 2008)*, pages 111–122. Omnipress, 2008.

Leena C Vankadara, Sebastian Bordt, Ulrike von Luxburg, and Debarghya Ghoshdastidar. Recovery guarantees for kernel-based clustering under non-parametric mixture models. In *International Conference on Artificial Intelligence and Statistics*, pages 3817–3825. PMLR, 2021.

Appendix A. Proofs from Section 4.1

A.1. Proof of Theorem 4

Proof We begin by observing that

$$\begin{aligned}\mathbb{E}_{x\sim\nu}\left[\|x-\mu(x)\|_1\right] &= \sum_{i=1}^{d}\mathbb{E}_{x\sim\nu}\left[|x_i-\mu_i(x)|\right] &(32)\\ &\geq \beta^{-1}\sum_{i=1}^{d}\sqrt{\mathbb{E}_{x\sim\nu}\left[|x_i-\mu_i(x)|^2\right]} &(33)\\ &= \beta^{-1}\left(\sigma_1+\cdots+\sigma_d\right) &(34)\end{aligned}$$

from Assumption 3. Moreover,

$$\mathbb{E}_{x\sim\nu}\left[\|x-\tilde{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\|x-\hat{\mu}(x)\|_1\right] \tag{35}$$

because $\tilde{\mu}(x)$ is the median of the points that end up in the same leaf as x. Using the triangle inequality, we obtain

$$\mathbb{E}_{x\sim\nu}\left[\|x-\hat{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\|x-\mu(x)\|_1\right] + \mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right] \tag{36}$$

Therefore, the price of explainability is given by

$$Price(\nu,T) \leq 1 + \frac{\mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right]}{\sigma_1+\cdots+\sigma_d} \tag{37}$$

The random variable $\|\mu(x)-\hat{\mu}(x)\|_1$ can be bounded uniformly from above.

$$\|\mu(x)-\hat{\mu}(x)\|_1 \leq \sum_{t\in T}\mathbf{1}(x \text{ is separated from } \mu(x) \text{ at node } t)\cdot \max_{k,l\in N(t)}\|\mu^{(k)}-\mu^{(l)}\|_1 \tag{38}$$

where each $t\in T$ denotes a node of the tree T and we bound $\|\mu(x)-\hat{\mu}(x)\|_1$ in terms of the worst-case distance between any pair from the set of remaining centers $N(t)$ at this very node. Note that

$$\|\mu(x)-\hat{\mu}(x)\|_1 \leq \max_{k,l\in N(t)}\|\mu^{(k)}-\mu^{(l)}\|_1 \tag{39}$$

conditioned on the event that a cut at node t indeed separates $\mu(x)$ from $\hat{\mu}(x)$. Taking expectations over $x\sim\nu$, we obtain

$$\mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right] \leq \sum_{t\in T}P(x \text{ is separated from } \mu(x) \text{ at node } t)\cdot \max_{k,l\in N(t)}\|\mu^{(k)}-\mu^{(l)}\|_1 \tag{40}$$

Observe that $x\sim\nu$ can only be separated from $\mu(x)$ at node t if x lies on the wrong side of the threshold cut θ that is performed at node t, and also satisfies $\mu(x)\in N(t)$. The former can only happen if $|x_i - \mu_i(x)| > |\theta - \mu_i(x)|$. Using Chebyshev's inequality, the probability is therefore bounded as

$$P(x \text{ is separated from } \mu(x) \text{ at node } t) = \sum_{k=1}^{K} p^{(k)} \cdot P(x \text{ is separated from } \mu^{(k)} \text{ at node } t) \tag{41}$$

$$\leq \frac{\alpha}{K} \sum_{k \in N(t)} \frac{\sigma_i^2}{\left|\theta - \mu_i^{(k)}\right|^2} \tag{42}$$

plugging in $p^{(k)} \leq \alpha/K$. Recall that the explainability-to-noise ratio $ENR(\nu) = q$ is defined as

$$q := \min_{k \neq l} \max_{j \in [d]} \left(\frac{|\mu_j^{(k)} - \mu_j^{(l)}|^2}{\sigma_j^2} \right) \tag{43}$$

We now claim the following.

Claim 1: At any internal node t of T with at least two remaining centers in $N(t)$, there exists a pair $\mu^{(k)}, \mu^{(l)} \in N(t)$ and a coordinate $i \in [d]$ such that

$$\max_{k \in N(t)} \mu_i^{(k)} - \min_{l \in N(t)} \mu_i^{(l)} \geq \sigma_i \sqrt{q} \tag{44}$$

Proof of Claim 1: Suppose this was not true at some node t. Then, for all $j \in [d]$, all remaining centers in $N(t)$ are at a distance strictly smaller than $\sigma_j \sqrt{q}$ when projected to the j-th coordinate. This is a contradiction to the explainability-to-noise ratio being $ENR(\nu) = q$. □

For all $j \in [d]$, denote

$$R_j(t) = \max_{k \in N(t)} \mu_j^{(k)} - \min_{l \in N(t)} \mu_j^{(l)} \tag{45}$$

for the length of the interval that contains all remaining centers in $N(t)$, projected to the j-th axis. Recall that the algorithm selects the axis $i \in [d]$ with

$$i = \text{argmax}_{j \in [d]} \frac{R_j(t)}{\sigma_j} \tag{46}$$

We assume w.l.o.g. that $N(t) = [K']$ for some $K' \leq K$. We also assume that the projections are sorted in non-decreasing order, that is

$$0 = \mu_i^{(1)} \leq \mu_i^{(2)} \leq \cdots \leq \mu_i^{(K')} = R_i(t) \tag{47}$$

Let us define

$$f(\theta) = \sum_{k=1}^{K'} \frac{\sigma_i^2}{\left|\theta - \mu_i^{(k)}\right|^2} \tag{48}$$

and denote

$$\theta^* := \text{argmin}_{0<\theta<R_i(t)} f(\theta) \tag{49}$$

The minimum is sure to exist since $f(\theta)$ is continuous everywhere except at any $\mu^{(k)}$, where it is unbounded. Note that $f(\theta^*)$ is a upper bound on the probability of separating any $x \sim \nu$ from its center at node t through the threshold cut chosen by the algorithm.

Claim 2: Define $\delta = \frac{R_i(t)}{2(K-1)}$. We claim that for any $0 = \mu_i^{(1)} \leq \cdots \leq \mu_i^{(K')} = R_i(t)$,

$$f(\theta^*) \leq 2 \sum_{k=1}^{\lceil K/2 \rceil} \frac{\sigma_i^2}{(2k-1)^2 \delta^2} \tag{50}$$

The right hand side corresponds to the case where $K' = K$, all centers are equidistant, and the threshold is halfway between the two neighboring centers in the middle.

Proof of Claim 2: We begin by noting that

$$\max_{\{\mu^{(k)}\}_{k=1}^{K'}} f(\theta^*) \leq \max_{\{\mu^{(k)}\}_{k=1}^{K}} f(\theta^*) \tag{51}$$

because adding more centers to the same interval $[0, R_i(t)]$ can only ever increase $f(\theta^*)$. By the same logic, we may without loss of generality assume hat $K = K'$ is even. For $K = 2$, Inequality 50 surely holds, as simply $\mu_i^{(1)} = 0 < R_i(t) = \mu_i^{(2)}$. So we may assume $K \geq 4$. Recall that Algorithm 1 chooses θ^* such as to minimize $f(\theta)$. We can certainly give an upper bound on $f(\theta^*)$ by instead choosing a suboptimal threshold $\hat{\theta}$ that is located halfway inside the longest subinterval. Here, we define subintervals $\Delta_1, \ldots \Delta_{K-1}$ as

$$\Delta_k = \left[\mu_i^{(k)}, \mu_i^{(k+1)}\right] \tag{52}$$

The surrogate $\hat{\theta}$ is chosen as

$$\hat{\theta} := \frac{\mu_i^{(\hat{k}+1)} + \mu_i^{(\hat{k})}}{2} \text{, where} \tag{53}$$

$$\hat{k} := \text{argmax}_{k \in [K-1]} \Delta_k \tag{54}$$

When multiple possible $\hat{k}$ exist any one of them is selected. Now, fix some $1 \leq K_1 \leq K$ and $0 < R_1 \leq R$. Consider the optimization problem of finding the worst configuration of K_1 means $\mu_i^{(1)} \leq \cdots \leq \mu_i^{(K_1)}$ **on the right** of a fixed threshold cut, on an interval of fixed length R_1. The location of means is described through auxiliary positive variables δ_k which serve the purpose of letting $\mu_i^{(k)} = \delta_1 + \cdots + \delta_k$.

$$\text{maximize} \quad h(\delta) = \sum_{k=1}^{K_1} \frac{1}{(\delta_1 + \cdots + \delta_k)^2} \tag{55}$$

$$\text{s.t.} \quad \delta_1 \geq 2\delta_k \text{ for all } k \geq 2 \tag{56}$$

$$\delta_1 \geq R_1/2K_1 \tag{57}$$

$$\delta_1 + \cdots + \delta_{K_1} = R_1 \tag{58}$$

The function $g(x) = x^{-2}$ satisfies $g''(x) = 6x^{-4} > 0$ and is hence convex on the positive real line. Since the ℓ_1 norm is convex (as is any ℓ_p norm for $p \geq 1$), each

$$h_k(\delta_1, \dots, \delta_k) = \frac{1}{(\delta_1 + \dots + \delta_k)^2} \tag{59}$$

is convex, being the composition of two convex functions. Their sum $h(\delta)$ is also convex, as can be seen from looking at its Hessian. For $k < K_1$, the Hessian $\boldsymbol{H}_k$ of h_k simply gets an additional zero block if we treat it as a function of $\delta = (\delta_1, \dots, \delta_{K_1})$, and thus stays positive semi-definite. Moreover, $\boldsymbol{H}_{K_1}$ is positive definite. The sum of a positive definite matrix with $K_1 - 1$ positive semi-definite matrices is positive definite, and convexity of $h(\delta)$ follows. Overall, we are maximizing a continuous, convex function over a closed (all constraints are inequalities or equalities), bounded, convex domain. By Bauer's maximum principle, the maximum of h is attained on its boundary (Bauer, 1958). It is easy to see that this implies $\delta_1^* = R_1/2K_1$ and $\delta_k^* = R_1/K_1$ for all $k \geq 2$. The value of h in this scenario is

$$h(\delta^*) = \sum_{k=1}^{K_1} \frac{1}{\left(\frac{R_1}{2K_1} + (k-1)\frac{R}{K_1}\right)^2} = \frac{4K_1^2}{R_1^2} \sum_{k=1}^{K_1} \frac{1}{(2k-1)^2} \tag{60}$$

Thus, when $\hat{\theta}$ is selected halfway in the longest subinterval, we certainly have

$$\max_{\{\mu^{(k)}\}_{k=1}^K} f(\hat{\theta}) \leq \max_{K_1, K_2, R_1, R_2} \left\{ \frac{4K_1^2}{R_1^2} \sum_{k=1}^{K_1} \frac{1}{(2k-1)^2} + \frac{4K_2^2}{R_2^2} \sum_{k=1}^{K_2} \frac{1}{(2k-1)^2} \right\} \tag{61}$$

where the maximum on the RHS is constrained over $K_1, K_2 \geq 1$ and $R_1 + R_2 = R$ and $R_1/K_1 = R_2/K_2$ (only then will $\hat{\theta}$ lie halfway between two neighboring centers). Clearly, this expression can be upper bounded by the case where $K_1 = K_2 = K/2$ and $R_1 = R_2 = R/2$. This concludes the statement. □

Combining this result with the previous steps, we see that

$$\begin{aligned} P(x \text{ is separated from } \mu(x) \text{ at node } t) &\leq \frac{\alpha}{K} \sum_{k=1}^{K} \frac{\sigma_i^2}{\left|\theta - \mu_i^{(k)}\right|^2} && (62) \\ &\leq \frac{2\alpha\sigma_i^2}{K} \sum_{k=1}^{K} \frac{1}{(2k-1)^2 \delta^2} && (63) \\ &\leq \frac{2\alpha\sigma_i^2}{K} \sum_{k=1}^{K} \frac{4K^2}{(2k-1)^2 R_i(t)^2} && (64) \\ &\leq \frac{8\alpha\sigma_i^2 K}{R_i(t)^2} \left(0.5 + \frac{\pi^2}{12}\right) && (65) \\ &= \frac{(4 + 2\pi^2/3)\alpha\sigma_i^2 K}{R_i(t)^2} && (66) \end{aligned}$$

where we bound the sum of odd reciprocal squares

$$\sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \leq 1 + 0.5 \sum_{k=2}^{\infty} \frac{1}{k^2} = 0.5 + \frac{\pi^2}{12} \tag{67}$$

Returning to the price of explainability, and exploiting the fact that

$$\max_{k,l \in N(t)} \|\mu^{(k)} - \mu^{(l)}\|_1 \leq \sum_{j=1}^{d} R_j(t) \tag{68}$$

at every node t, we obtain

$$\begin{aligned}
\frac{\mathbb{E}_{x\sim\nu}\left[\|\mu(x) - \hat{\mu}(x)\|_1\right]}{\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_1\right]} &\leq \beta \sum_{t\in T} P(x \text{ is separated from } \mu(x) \text{ at node } t) \cdot \max_{k,l\in N(t)} \frac{\sum_j R_j(t)}{\sum_j \sigma_j} &(69)\\
&\leq \sum_{t\in T} \frac{(4+2\pi^2/3)\alpha\beta\sigma_i^2 K}{R_i(t)^2} \cdot \frac{\sum_j R_j(t)}{\sum_j \sigma_j} &(70)\\
&\leq \sum_{t\in T} \frac{(4+2\pi^2/3)\alpha\beta\sigma_i^2 K}{R_i(t)^2} \cdot \max_{i\in[d]} \frac{R_i(t)}{\sigma_i} &(71)\\
&= (4+2\pi^2/3)\alpha\beta K \cdot \sum_{t\in T} \frac{\sigma_i}{R_i(t)} &(72)\\
&\leq \frac{(4+2\pi^2/3)\alpha\beta K(K-1)}{\sqrt{q}} &(73)
\end{aligned}$$

where we used the fact that i is chosen as the maximizer of $\frac{R_j(t)}{\sigma_j}$ at every node t, and plugged in the explainability-to-noise ratio as a lower bound on this quantity. In the final step, we use that there are no more than $K-1$ nodes in the tree. ■

A.2. Proof of Theorem 5

Proof The proof builds on a construction by Moshkovitz et al. (2020), mildly adjusted to our probabilistic setting. We refer the reader to Appendix C from their paper for details. First, K centers $\{\mu^{(k)}\}_{k=1}^{K}$ are constructed on $\{\pm 1\}^d$. Using Hoeffding's inequality and by randomly sampling points from $\{\pm 1\}^d$, it can be shown that if $d = K^3$, the following two properties hold for any $\epsilon \geq \log(K)/\sqrt{K}$.

1. All centers are distinct along at least $d/4$ axes.
2. For any selection of $l \leq \log(K)/50$ axes of $[d]$, the number of centers that agree on all these axes is at least $K(2^{-l} - \epsilon) > 1$.

From there, we define K clusters $C^{(k)}$, each containing $M + 2d$ points where M will be chosen later. We let

$$C^{(k)} = \left\{ \underbrace{\mu^{(k)}, \ldots, \mu^{(k)}}_{M \text{ times}}, \mu^{(k)} + e_i, \mu^{(k)} - e_i \right\}_{i=1}^{d} \tag{74}$$

so that $|C^{(k)}| = M + 2d$ and the mean and median in each cluster is at $\mu^{(k)}$. We then let

$$\mathcal{X} = \bigcup_{k=1}^{K} C^{(k)} \tag{75}$$

denote the set of all points in the support of the mixture model. We give $1/(M + 2d)K$ probability mass to each point in the union of all $\mathcal{X}$, and denote by $\nu^{(k)}$ the distribution that uniformly randomly samples from $C^{(k)}$. We also choose equal mixing weights over all $\nu^{(k)}$. Note that by adaptively choosing M, the explainability-to-noise ratio changes. For all $i \in [d]$ we have

$$\mathbb{E}_{x\sim\nu^{(k)}} \left[|x_i - \mu(x)_i|\right] = \mathbb{E}_{x\sim\nu^{(k)}} \left[|x_i - \mu(x)_i|^2\right] = \frac{2}{2d + M} \tag{76}$$

Therefore,

$$ENR(\nu) = \frac{2}{\frac{2}{2d+M}} = 2d + M \tag{77}$$

which can certainly be made larger than q for any $q \geq 2d = 2K^3$. Also note that $\beta = \sqrt{2d + M}$. Moreover, the optimal ℓ_1 clustering cost is

$$\sum_{x\in\mathcal{X}} \|x - \mu(x)\|_1 = 2dK \tag{78}$$

From here, we follow the proof of Moshkovitz et al. (2020). Their proof shows that when $M = 0$, any decision tree constructed on this dataset satisfies

$$\sum_{x\in\mathcal{X}} \|x - \tilde{\mu}(x)\|_1 \geq \Omega(dK \log K) \tag{79}$$

where $\tilde{\mu}(x)$ is the median of the leaf constructed by a tree. To provide some intuition, this is true because the means are at a distance of $\Omega(d)$, and any tree needs to select $\Omega(\log K)$ axes on all its root-to-leaf paths, at each of which new errors happen. In our case with $M > 0$, adding more points can only increase the total value of (79) but keeps (78) the same (the added points lie exactly at the means). This implies that for any M,

$$\frac{\sum_{x\in\mathcal{X}} \|x - \tilde{\mu}(x)\|_1}{\sum_{x\in\mathcal{X}} \|x - \mu(x)\|_1} = \Omega(\log K) \tag{80}$$

Using the fact that we give each point equal probability mass in ν, we obtain

$$Price(\nu, T) = \Omega(\log K) = \Omega(\alpha\beta \log K/\sqrt{q}) \tag{81}$$

as desired. ■

A.3. Extending to K-means

An extension of the proof technique from Theorem 4 to the explainable K-means problem is possible. In that case,

$$Price(\nu, T) = \frac{\mathbb{E}_{x\sim\nu}\left[\|x - \hat{\mu}(x)\|_2^2\right]}{\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_2^2\right]} \tag{82}$$

and we do not need Assumption 3. The probability of separating a point $x \sim \nu$ from its mean at a given node t remains of order $\mathcal{O}(\alpha K \sigma_i^2 / R_i^2)$, where as before we denote

$$R_j(t) = \max_{k,l \in N(t)} \mu_j^{(k)} - \mu_j^{(l)} \tag{83}$$

for the side length of the interval that contains all remaining centers along the j-th coordinate. The maximum cost incurred by separating x from $\mu(x)$ at node t is now $R_1^2 + \cdots + R_d^2$ because we are clustering in the squared Euclidean norm. The baseline cost is simply

$$\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_2^2\right] = \sigma_1^2 + \cdots + \sigma_d^2 \tag{84}$$

Now, using the fact that

$$\frac{\sum_{j=1}^d R_j^2}{\sum_{j=1}^d \sigma_j^2} \leq \max_{j\in[d]} \frac{R_j^2}{\sigma_j^2} = \frac{R_i^2}{\sigma_i^2} \tag{85}$$

by design of the algorithm, we see that the dependency on R_i and σ_i cancels out. Thus, we obtain $\mathcal{O}(K^2)$ bounds just as Moshkovitz et al. (2020) did, and $ENR(\nu)$ is removed from the analysis. This result of course does **not** imply that information on the mixture ν cannot be incorporated into our analysis — it is just that, loosely speaking, the quadratic dependency introduced by Chebyshev's inequality cancels out with the quadratic cost function of K-means. Concentration inequalities of higher moments (provided they exist) allow reintroducing information on the cluster-friendliness of ν into the upper bounds. One can redefine $ENR(\nu)$ in terms of these higher moments and reformulate Assumption 3 accordingly.

Appendix B. Proofs from Section 4.2

B.1. Proof of Theorem 8

Proof The proof follows from Appendix A.1, which shows that the probability of separating a point $x \sim \nu$ from its correct underlying mean $\mu(x)$ is bounded as

$$P(x \text{ is separated from } \mu(x) \text{ at node } t) \leq \frac{(4 + 2\pi^2/3)\alpha\sigma_i^2 K}{R_i(t)^2} \tag{86}$$

at any node t. Recall that i is chosen as the maximizer of $R_i(t)/\sigma_i$, which is certainly lower bounded by $\sqrt{q}$. As a consequence, $\sigma_i^2/R_i(t)^2$ is upper bounded by $1/q$. Summing over all nodes t of the tree (of which there are no more than $K-1$), we get the desired bound

$$P_{x\sim\nu}\left(\mu(x) \neq \hat{\mu}(x)\right) \leq \frac{(4 + 2\pi^2/3)\alpha K(K-1)}{q} \tag{87}$$

This concludes the proof. ■

B.2. Proof of Theorem 9

Proof We begin by constructing the mixture components $\nu^{(k)}$. To this end, let $d = K$ and define the means as the standard basis vectors

$$\mu_j^{(k)} = \begin{cases} 0 \text{ , if } k \neq j \\ 1 \text{ , if } k = j \end{cases} \tag{88}$$

For all $k \in [K]$ and all coordinates $j \in [d]$, let the random variable $X \sim \nu^{(k)}$ be given via

$$P(X_j = s) = \begin{cases} \epsilon \text{ , if } s = \mu^{(k)} + q \\ \epsilon \text{ , if } s = \mu^{(k)} - q \\ 1 - 2\epsilon \text{ , if } s = \mu^{(k)} \end{cases} \tag{89}$$

where $\epsilon = \frac{1}{2q}$. This ensures that $\nu^{(k)}$ has a variance equal to $\sigma_i^2 = 2\epsilon = \frac{1}{q}$ along any axis $i \in [d]$. Therefore, the explainability-to-noise ratio is

$$\max_{i\in[d]} \frac{1}{\sigma_i^2} = q \tag{90}$$

as desired. So far, we only specified the marginals of each distribution $\nu^{(k)}$. The joint distribution of each $\nu^{(k)}$ is chosen such that for all $k \in [K]$, and all $i \neq j \in [d]$, we have

$$P_{X\sim\nu^{(k)}}\left(X \neq \mu_j^{(k)} \text{ and } X_k^{(i)} \neq \mu_i^{(k)}\right) = 0 \tag{91}$$

This is certainly possible because $q \geq K \implies 2\epsilon K \leq 1$. It ensures that for all $\nu^{(k)}$, the probability that there exists an axis-aligned cut of the tree that separates $x \sim \nu^{(k)}$ from $\mu^{(k)}$ is precisely the sum over all the probabilities that this occurs at a given node of the tree, provided that different dimensions are chosen at each cut. Note that any decision tree must separate exactly one mean $\mu^{(k)}$

from all other remaining means at each iteration. Without loss of generality, we may assume that first $\mu^{(1)}$ is separated, then $\mu^{(2)}$ and so forth. Then, we obtain

$$\begin{aligned} P(\hat{\mu}(x) \neq \mu(x)) &= \frac{1}{K} \sum_{k=1}^{K} P_{x \sim \nu^{(k)}}(\hat{\mu}(x) \neq \mu^{(k)}) && (92) \\ &= \frac{1}{K} (\epsilon + 2\epsilon + \cdots + (K-1)\epsilon) && (93) \\ &= \frac{\epsilon(K-1)}{2} && (94) \\ &= \frac{K-1}{4q} && (95) \end{aligned}$$

This gives the desired lower bound on the error rate. ■

B.3. Constant Price Despite Perfect Recovery

Consider a mixture of two discrete measures on $\mathbb{R}^d$. For all $i \in [d]$ define a random variable X_i via

$$X_i = \begin{cases} +1, & \text{with probability } \epsilon \\ -1, & \text{with probability } \epsilon \\ 0, & \text{else} \end{cases} \tag{96}$$

where $\epsilon = \frac{1}{2d}$. Then, define $X = (X_1, \dots, X_d)$ such that $P(X_i \neq 0 \wedge X_j \neq 0) = 0$ for all $i \neq j$. This ensures that the mean (and median) of X is 0 along every axis, and that X always deviates from its median along exactly one axis. Moreover, let $X^{(1)} = X + 0.5$ and $X^{(2)} = X - 0.5$ and let $\nu^{(1)}, \nu^{(2)}$ be the measures associated with $X^{(1)}, X^{(2)}$. Choose equal mixing weights. Then, the baseline cost is

$$\mathbb{E}_{x \sim \nu}\left[\|x - \mu(x)\|_1\right] = 2d\epsilon = 1 \tag{97}$$

Now any tree that separates $\mu^{(1)} = 1$ from $\mu^{(2)} = -1$ necessarily makes an error with probability ϵ. Without loss of generality, we may assume that the tree chooses axis $i = 1$, and thresholds at $x_1 \leq 0$. Let us now look at the distribution of points in the "negative" leaf, i.e. we consider $x \sim \nu$ conditioned on $x_1 \leq 0$. Note that

$$\begin{aligned} & P(x_1 = -0.5 | x_1 < 0) && (98) \\ &= P(\mu(x) = \hat{\mu}(x) | x_1 < 0) \cdot P(x_1 = -0.5 | x_1 < 0, \mu(x) = \hat{\mu}(x)) + && (99) \\ &\quad P(\mu(x) \neq \hat{\mu}(x) | x_1 < 0) \cdot P(x_1 = -0.5 | x_1 < 0, \mu(x) \neq \hat{\mu}(x)) && (100) \\ &= \frac{P(\mu(x) = \hat{\mu}(x) \wedge x_1 < 0)}{P(x_1 < 0)} \cdot P(x_1 = -0.5 | x_1 < 0, \mu(x) = \hat{\mu}(x)) + && (101) \\ &\quad \frac{P(\mu(x) \neq \hat{\mu}(x) \wedge x_1 < 0)}{P(x_1 < 0)} \cdot P(x_1 = -0.5 | x_1 < 0, \mu(x) \neq \hat{\mu}(x)) && (102) \\ &= (1-\epsilon) \cdot \frac{1-2\epsilon}{1-\epsilon} + \epsilon \cdot 1 && (103) \\ &= 1 - \epsilon && (104) \end{aligned}$$

This immediately implies that the median in the negative leaf is at -0.5. Similar computations for the other axes reveal that the median of the negative leaf is also located at $x_i = -0.5$. Intuitively, this is true because the distribution of points in the negative leaf and the distribution $\nu^{(2)}$ are almost the same, and the median is robust to the small change. By symmetry, the median of the positive leaf $x_1 > 0$ is at 0.5 for all axes. Thus,

$$\hat{\mu}(x) = \tilde{\mu}(x) \tag{105}$$

This result implies that if a point $x \sim \nu$ is assigned to the wrong leaf not containing $\mu(x)$, then $\tilde{\mu}(x)$ is at a distance of 1 along every axis except the one that the tree chooses (where the distance is 0). Hence

$$\mathbb{E}_{x\sim\nu}\left[\|x - \tilde{\mu}(x)\|_1 | \tilde{\mu}(x) \neq \mu(x)\right] = (d-1) \tag{106}$$

If $x \sim \nu$ is not separated from $\mu(x)$ by the tree (this happens with probability $1 - \epsilon$), then

$$\mathbb{E}_{x\sim\nu}\left[\|x - \hat{\mu}(x)\|_1 | \tilde{\mu}(x) = \mu(x)\right] = 1 \tag{107}$$

which implies

$$\begin{aligned} \mathbb{E}_{x\sim\nu}\left[\|x - \hat{\mu}(x)\|_1\right] &= \epsilon(d-1) + (1-\epsilon) && (108) \\ &= 1.5 - \frac{1}{d} && (109) \end{aligned}$$

Clearly, as $d \to \infty \iff \epsilon \to 0$, the price approaches $Price(\nu, T) = 1.5$ despite the error rate decaying to zero.

Appendix C. Proofs from Section 5

C.1. Proof of Lemma 11

Proof Recall that for all $k \in [K]$, we have

$$\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} \left[\kappa(x,x')\right] = \|\phi_{\nu^{(k)}}\|_{\mathcal{H}}^2 = \sigma^2 \tag{110}$$

Since $g_i(|x_i - x_i'|) \in (0,1]$ for all $i \in [d]$ this implies that for all axes $i \in [d]$, we also have

$$\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} \left[g_i(|x_i - x_i'|)\right] \geq \sigma^2 \tag{111}$$

Now take any pair of distinct mixture components $\nu^{(k)}, \nu^{(l)}$. First of all, we have

$$\begin{aligned} \gamma &\leq d_\kappa(\nu^{(k)}, \nu^{(l)}) & (112)\\ &= \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} \left[\kappa(x,x')\right] + \mathbb{E}_{y,y' \sim \nu^{(l)}} \left[\kappa(y,y')\right] - 2\mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} \left[\kappa(x,y)\right] & (113)\\ &= 2\sigma^2 - 2\mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} \left[\kappa(x,y)\right] & (114) \end{aligned}$$

Therefore, we can write

$$\begin{aligned} \gamma/2 &\leq \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} \left[\kappa(x,x')\right] - \mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} \left[\kappa(x,y)\right] & (115)\\ &= \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}, y \sim \nu^{(l)}} \left[\prod_{i=1}^d g_i(|x_i - x_i'|) - \prod_{i=1}^d g_i(|x_i - y_i|)\right] & (116)\\ &\leq \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}, y \sim \nu^{(l)}} \left[\sum_{i=1}^d g_i(|x_i - x_i'|) - g_i(|x_i - y_i|)\right] & (117) \end{aligned}$$

The inequality step above exploits the fact that for any set of positive real numbers $a_i, b_i \in (0,1]$ we can write

$$\begin{aligned} \prod_{i=1}^d a_i - \prod_{i=1}^d b_i &= (a_1 - b_1)\prod_{i=2}^d a_i + b_1(a_2 - b_2)\prod_{i=3}^d a_i + b_1 b_2 (a_3 - b_3)\prod_{i=4}^d a_i + \dots & (118)\\ &= \sum_{i=1}^d (a_i - b_i) \underbrace{\prod_{j<i} b_j}_{\in(0,1]} \cdot \underbrace{\prod_{j>i} a_j}_{\in(0,1]} & (119) \end{aligned}$$

and therefore, there must exist $i \in [d]$ with

$$a_i - b_i \geq \frac{\prod_{i=1}^d a_i - \prod_{i=1}^d b_i}{d} \tag{120}$$

It follows from Equation (115) that there certainly exists $i \in [d]$ with

$$\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}, y \sim \nu^{(l)}} \left[g_i(|x_i - x_i'|) - g_i(|x_i - y_i|)\right] \geq \frac{\gamma}{2d} \tag{121}$$

Since for all $i \in [d]$ we have $\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} \left[g_i(|x_i - x_i'|)\right] \geq \sigma^2$ this implies that

$$\mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} \left[g_i(|x_i - x_i'|)\right] \leq \sigma^2 - \gamma/2d = \tau \tag{122}$$

Since the pair $k \neq l$ was arbitrary, this proves the statement.

■

C.2. Proof of Theorem 12

Proof We begin by noting that for any $x \sim \nu$, if $\hat{\mu}(x) \neq \mu(x)$, then

$$\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2 = 1 + \|\hat{\mu}(x)\|_{\mathcal{H}}^2 - 2\langle \phi(x), \hat{\mu}(x) \rangle \leq 1 + \sigma^2 \tag{123}$$

Therefore, the random variable $Z = \|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2$ is almost surely bounded from above by another random variable $W = \|\phi(x) - \mu(x)\|_{\mathcal{H}}^2 + \mathbf{1}(\hat{\mu}(x) \neq \mu(x)) \cdot (1 + \sigma^2)$. Thus,

$$\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2\right] \leq \mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right] + P(\hat{\mu}(x) \neq \mu(x)) \cdot (1 + \sigma^2) \tag{124}$$

Therefore,

$$\frac{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2\right]}{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} \leq \frac{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right] + P(\hat{\mu}(x) \neq \mu(x)) \cdot (1 + \sigma^2)}{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} \tag{125}$$

$$= 1 + \frac{1 + \sigma^2}{1 - \sigma^2} \cdot P(\hat{\mu}(x) \neq \mu(x)) \tag{126}$$

where we used the fact that $\|\phi_{\nu^{(k)}}\|_{\mathcal{H}}^2 = \sigma^2$ for all k, which implies

$$\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right] = 1 - \sigma^2 \tag{127}$$

by virtue of $\kappa(x, x) = 1$. Thus, we see that it is sufficient to bound the error rate of the tree. Fix a node t with remaining components $N(t)$. Let i be the axis chosen by the algorithm, and let $k \neq l \in N(t)$ be the pair that minimizes

$$\xi(i, k, l) = \mathbb{E}_{x\sim\nu^{(k)}, y\sim\nu^{(l)}}[\kappa_i(x, y)] \tag{128}$$

Without loss of generality, we may assume that $N(t) = [K']$ for some $K' \leq K$ and that $l = 1, k = K'$. We assume further that $\xi(i, k, m) < \xi(i, k, m')$ for all $m < m' \in [K' - 1]$. If this does not hold, simply sort and relabel. Because the axis-aligned κ-alignment of the mixture model ν is given by τ, we know that $\xi(i, K', 1) \leq \tau$. Additionally, recall that $\|g_i\|_\infty \leq 1$ and hence $\kappa_i(x, x') \leq 1$ for all x, x'. Consequently,

$$\mathbb{E}_{x,x'\sim_{i.i.d.}\nu^{(K')}}[\kappa_i(x, x')] \geq \mathbb{E}_{x,x'\sim_{i.i.d.}\nu^{(K')}}[\kappa(x, x')] = \sigma^2 \tag{129}$$

This implies that $\xi(i, 1, K') \geq \sigma^2$. Therefore, the scalars $\{\xi(i, K', m)\}_{m\in N(t)}$ are spread out on an interval of length at least $\Delta \geq \sigma^2 - \tau$. Recall that the threshold θ is chosen exactly halfway between the two neighboring values of $\xi(i, K', m)$ that are contained inside (τ, σ^2) and separated furthest therein. Formally,

$$\theta = \frac{\xi(i, K', m' + 1) + \xi(i, K', m')}{2}, \text{ where} \tag{130}$$

$$m = \operatorname*{argmax}_{m' : \xi(i, K', m'+1) \leq \sigma^2} \xi(i, K', m' + 1) - \xi(i, K', m') \tag{131}$$

For a randomly sampled reference point $x^{(K')} \sim \nu^{(K')}$, a mixture component $m \in N(t)$ and some new $x \sim \nu^{(m)}$, the threshold cut at θ can only make a mistake (i.e. assign x to the child node that erroneously does not contain index m) if either $\kappa_i(x^{(K')}, x) < \theta < \xi(i, K', m)$ or vice versa. This can only happen if

$$\left|\kappa_i(x^{(K')}, x) - \xi(i, K', m)\right| = \left|\kappa_i(x^{(K')}, x) - \mathbb{E}_{x^{(K')} \sim \nu^{(K')}, x \sim \nu^{(m)}}[\kappa_i(x^{(K')}, x)]\right| \tag{132}$$

$$> |\theta - \xi(i, K', m)| \tag{133}$$

which brings us back to the setting where we can invoke Chebyshev's inequality to bound the probability of making mistakes. Using $p^{(m)} \leq \alpha/K$ for all $m \in [K]$, we obtain

$$P_{x^{(k)} \sim \nu^{(k)}, x \sim \nu}(x \text{ goes to wrong child}) \leq \sum_{m \in N(t)} p^{(m)} \cdot \frac{\epsilon^2}{|\theta - \xi(i, K', m)|^2} \tag{134}$$

$$\leq \frac{\alpha \epsilon^2}{K} \sum_{m \in N(t)} \frac{1}{|\theta - \xi(i, K', m)|^2} \tag{135}$$

Now we argue just as we did in the proof of Theorem 4 (see Claim 2 from Appendix A.1). An upper bound on this probability occurs when all $\xi(i, K', m)$ are equidistant and $N(t) = [K]$. In that case,

$$\sum_{m \in N(t)} \frac{1}{|\theta - \xi(i, K, m)|^2} \leq 2 \sum_{k=1}^{\lceil K/2 \rceil} \frac{1}{(2k-1)^2 (\Delta/2K)^2} \tag{136}$$

$$\leq \frac{8K^2}{\Delta^2} \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \tag{137}$$

$$\leq \frac{8K^2(0.5 + \pi^2/12)}{\Delta^2} \tag{138}$$

This implies

$$P_{x^{(k)} \sim \nu^{(k)}, x \sim \nu}(x \text{ goes to wrong child}) \leq \frac{(4 + 2\pi^2/3)\alpha \epsilon^2 K}{\sigma^2 - \tau} \tag{139}$$

Since there are no more than $K - 1$ nodes, we obtain

$$P_{x \sim \nu}(\hat{\mu}(x) \neq \mu(x)) \leq \frac{(4 + 2\pi^2/3)\alpha \epsilon^2 K(K-1)}{\sigma^2 - \tau} \tag{140}$$

Obviously, the probability can never exceed 1, so we take the maximum. Plugging this back into the price of explainability yields

$$Price_\kappa(\nu, T) \leq 1 + \frac{1 + \sigma^2}{1 - \sigma^2} \cdot \max\left(1, \frac{(4 + 2\pi^2/3)\alpha \epsilon^2 K(K-1)}{\sigma^2 - \tau}\right) \tag{141}$$

■