Explainable Clustering of Mixture Models

Maximilian Fleissner[1] **Maedeh Zarvandi**[1,2] **Debarghya Ghoshdastidar**[1,2]

[1]Technical University of Munich
[2] Munich Center for Machine Learning

Abstract

The explainable clustering problem was first posed by Moshkovitz et al. (ICML 2020) and studies how well an axis-aligned decision tree with K leaves can approximate a given clustering. The performance of the tree is measured via the *price of explainability*, defined as the ratio between the clustering cost of the tree (where every leaf is a cluster) and the optimal cost. Several recent works have given worst-case characterizations of the price of explainability for different cost functions. However, these guarantees are data-agnostic and therefore notoriously pessimistic in practical clustering settings. In this paper, we study explainable clustering from the point of view of mixture models, which allows us to give the first data-dependent bounds on the price of explainability. First, we focus on K-medians clustering of mixture models with subexponential tails. We propose an algorithm that leverages information about the distribution of the data to find better cuts, and prove new upper and lower bounds. Second, we extend our algorithm and the theoretical guarantees it provides to kernel clustering, thereby refining the existing worst-case analysis.

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 several methods have been developed for the supervised setting, the explainability of unsupervised learning has attracted 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] suggest to cluster the data using binary axis-aligned decision trees with K leaves. At every node of the tree, a one sided axis-aligned threshold cut $x_i \leq \theta$ partitions the data into two child nodes (where 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. Moshkovitz et al. [2020] prove an upper bound on the price of explainability that depends only on the number of clusters K, with $\mathcal{O}(K)$ guarantees for K-medians. This bound holds for *any* dataset, regardless of how well it is clustered.

Numerous recent works have given improved upper and lower bounds on the price of explainability for various cost functions (for an overview, see related works). Notably, Makarychev and Shan [2023] show that for K-medians the price of explainability is $\Theta(\log K)$, and that the upper bound is attained by a randomized procedure known as Random Coordinate Cut. From a worst-case point of view, this settles the matter. However, all existing theory is entirely distribution-agnostic: It treats well-separated, low-noise mixture models and adversarial worst-case examples on equal footing.

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 in their discussion 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

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1585–1603.

the price of explainability already consist of rather well-clustered instances, suggesting that care must be taken in the analysis.

Contributions. In this work, we study explainable clustering under assumptions on the data distribution, which allows us to derive tighter, more realistic guarantees. The first part of the paper tackles the classic explainable K-medians problem under mixture models with subexponential tails. We introduce an explainability-to-noise ratio $\mathrm{ENR}(\nu)$ for mixtures ν and show that $\mathrm{ENR}(\nu)$ influences how well a decision tree can approximate ν. We prove upper and lower bounds on the price of explainability that depend only on the number of clusters K and $\mathrm{ENR}(\nu)$, and are almost tight in the latter. Our upper bounds are attained by the Mixture Model Decision Tree (MMDT) algorithm that we propose. MMDT takes specific information about the distribution of the mixture model into account and runs in data-independent time. The second part of this paper applies our new probabilistic analysis to kernel clustering. We derive data-dependent guarantees on the price of explainability that significantly improve over existing worst-case guarantees, thereby explaining the empirically observed phenomenon that the price of explainability is often quite low on real world clustering datasets.

Related Work. Starting with Moshkovitz et al. [2020], a significant number of works have tackled the explainable clustering problem. Most works focus on a tighter analysis of the worst-case price of explainability for K-medians and K-means [Makarychev and Shan, 2021, 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], or focus on other cost functions [Laber and Murtinho, 2021, 2023]. 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. Laber et al. [2023] discuss explainable clustering with shallow decision trees. Fleissner et al. [2024] theoretically analyze the worst-case price of explainability for kernel clustering, focusing on Gaussian and Laplace kernel. Several other works also propose explainable clustering algorithms that can be used for partitions induced by kernel clustering, but do not bound the price of explainability [Bertsimas et al., 2021, Ohl et al., 2024, Argov and Wagner, 2026].

2 PROBLEM SETUP AND DEFINITIONS

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

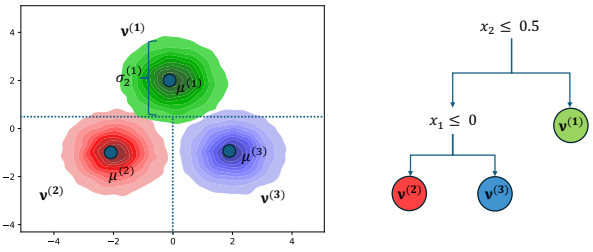

Figure 1: Illustration of our setting: Given three distributions $\nu^{(1)}, \nu^{(2)}, \nu^{(3)}$, we investigate how well a decision tree with axis-aligned binary cuts and three leaves can cluster these components. The objective is the ratio between the clustering cost of the tree, and the cost of the baseline partition into $\nu^{(1)}, \nu^{(2)}, \nu^{(3)}$.

2.1 DATA MODEL

Let $K \geq 2$ be an integer. 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 $\sum_{k=1}^{K} p^{(k)} = 1$. We assume that the the mixing weights $p^{(k)}$ are safely bounded away from 1, satisfying $p^{(k)} \leq \frac{\alpha}{K}$ for some $\alpha \geq 1$. 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 the mean absolute deviations of all mixture components are finite.

$$\forall k, i : \mathbb{E}_{x \sim \nu^{(k)}}[|x_i - \mu_i^{(k)}|] = \sigma_i^{(k)} < \infty. \tag{1}$$

For all $i \in [d]$, we define the average mean absolute deviation as $\sigma_i = \sum_{k=1}^{K} p^{(k)} \sigma_i^{(k)}$. We use the superscript for enumerating the mixture components, and the subscript to refer to axes of the ambient space $\mathbb{R}^d$. The support of ν is denoted $\mathrm{supp}(\nu)$.

2.2 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. 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 $\nu^{(1)}, \ldots, \nu^{(K)}$. 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 $\mu^{(k)}$. This is natural in a mixture

model setting, but technically need not be enforced (there could exist trees with low clustering cost despite multiple means in the same leaf). However, all existing algorithms also demand every leaf to contain exactly one of the K baseline cluster centers. Figure 1 illustrates our setting.

2.3 PRICE OF EXPLAINABILITY

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*

$$\mathrm{Price}(T, \nu) = \frac{\mathbb{E}_{x\sim\nu}\left[\|x - \tilde{\mu}(x)\|_1\right]}{\mathbb{E}_{x\sim\nu}\left[\|x - \mu(x)\|_1\right]} \tag{2}$$

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, 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

$$cost_{opt}(\nu) := \min_{c^{(1)},\dots,c^{(K)}} \mathbb{E}_{x\sim\nu}\left[\min_{k\in[K]} \|x - c^{(k)}\|_1\right] \tag{3}$$

are 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. However, in a mixture model setting, we are typically interested in the 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.

In this paper, we incorporate information on ν into the analysis of $\mathrm{Price}(T, \nu)$.

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

$$\mathrm{ENR}(\nu) := \min_{k\neq l} \max_{j\in[d]} \left(\frac{|\mu_j^{(k)} - \mu_j^{(l)}|}{\sigma_j}\right). \tag{4}$$

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

Input: Components $\{\nu^{(k)}\}_{k=1}^K$ with means $\{\mu^{(k)}\}_{k=1}^K \subset \mathbb{R}^d$ and coordinate-wise mean absolute deviations $\{\sigma_i^{(k)}\}_{i,k}$.
Output: Binary tree with K leaves (one per component).
Create root t_0 with $N(t_0) \leftarrow [K]$, active set $\mathcal{A} \leftarrow \{t_0\}$.
while $\exists\, t \in \mathcal{A}$ with $|N(t)| > 1$ **do**
  Pick any such t, set $N \leftarrow N(t)$.
  Choose (i, θ) by minimizing (5) over $i \in [d]$ and $\theta \in I_t$. If densities are unknown, by minimizing the bound in (7) over i, θ.
  Split index set into $N_L \leftarrow \{k \in N : \mu_i^{(k)} \leq \theta\}$ and $N_R \leftarrow \{k \in N : \mu_i^{(k)} > \theta\}$.
  Store at t the cut (i, θ), create children t_L, t_R with $N(t_L) = N_L$, $N(t_R) = N_R$.
  Update $\mathcal{A} \leftarrow (\mathcal{A} \setminus \{t\}) \cup \{t_L, t_R\}$.
end while
Route x at node (i, θ) to left if $x_i \leq \theta$ and right otherwise.
If leaf t has $N(t) = \{k\}$, assign leaf to component k.

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 $\mathbb{P}_{x\sim\nu}(\mu(x) \neq \hat{\mu}(x))$ too much. The explainability-to-noise ratio can be viewed as an axis-aligned version of the classic signal-to-noise ratio.

At this point, one may be tempted to think that if there exists a tree that makes few mistakes on a mixture model ν, then the tree also automatically achieves a low price of explainability. However, this is not necessarily correct: The cost of the tree is not only influenced by the error rate of the tree, but also by the severity of the mistakes it makes. In some cases, both effects exactly "cancel out". For example, known $\Omega(\log K)$ lower bounds on the price of explainability for K-medians can be adapted such that the resulting tree almost perfectly recovers the mixture components $\nu^{(k)}$, in the sense that $\mathbb{P}(\hat{\mu}(x) = \mu(x)) = 1 - o(1)$.

However, for mixture models with light tails, the error rate of the tree is indeed the deciding quantity, as it generally decays much faster than the cost of mistakes. This observation forms the starting point of our analysis and suggested algorithm.

3 PROPOSED ALGORITHM

We propose the Mixture Model Decision Tree (MMDT) algorithm to obtain an explainable approximation for mixture models. MMDT is described in Algorithm 1. 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 $i \in [d]$ and binary cut $\theta \in \mathbb{R}$ such that the probability that any $x \sim \sum_{k\in N(t)} \bar{p}^{(k)}\nu^{(k)} =: \nu(N)$ is separated

from its mean is minimal. Here, $\bar{p}^{(k)}$ simply rescales $p^{(k)}$ to ensure the weights of $\nu(N)$ still sum to one. Formally,

$$i, \theta = \underset{i \in [d], \theta \in I_t}{\arg\min}\, \mathbb{P}_{x \sim \nu(N)}(x \text{ separated from } \mu(x) \text{ by } \theta)$$
$$I_t = \left(\min_{k \in N(t)} \mu_i^{(k)}, \max_{k \in N(t)} \mu_i^{(k)} \right) \tag{5}$$

If multiple possible thresholds exist, one is selected randomly. Typically, the exact probability density of each $\nu^{(k)}$ is not known (or difficult to minimize), so we choose θ by instead minimizing an upper bound on the probability in (5). If nothing is known about the mixture components except its means and mean absolute deviations, we can use Markov's inequality: For any threshold θ chosen on the i-th axis, we can certainly bound the probability that $x \sim \nu^{(k)}$ gets separated from its mean by $\frac{|\mu_i^{(k)} - \theta|}{\sigma_i^{(k)}}$. We assume lighter, subexponential tails.

Assumption 3 (Subexponential tails). *We assume there exist constants $c_1, c_2 > 0$ independent of K such that for all $k \in [K], i \in [d], s > 0$,*

$$\mathbb{P}_{x \sim \nu^{(k)}}(|x_i - \mu_i^{(k)}| > s) \leq c_1 \cdot \exp\left(-\frac{s c_2}{\sigma_i^{(k)}} \right). \tag{6}$$

Under Assumption 3, for all $k \in [K]$, and any threshold cut θ along some axis $i \in [d]$, we can bound

$$\mathbb{P}_{x \sim \nu^{(k)}}\left(x \text{ is separated from } \mu(x) \text{ by } \theta\right) \leq c_1 \exp\left(-\frac{c_2 |\theta - \mu_i^{(k)}|}{\sigma_i^{(k)}} \right). \tag{7}$$

Plugging this bound into (5), θ can be found in data-independent time (i.e. without iterating over all datapoints).

Remark 4 (Connection to IMM). *Algorithm 1 can be viewed as a "population version" of IMM, the initial explainable clustering algorithm proposed by Moshkovitz et al. [2020]. IMM minimizes the number of points that go to the wrong cluster center at a given node. It admits data-independent $\mathcal{O}(K)$ bounds for K-medians, and its runtime scales linearly with the number of samples (unlike MMDT, which only relies on means and mean absolute deviations).*

Remark 5 (Improved runtime for high dimensions). *For high dimensions, finding the minimizer of (7) over all $i \in [d]$ may be inefficient. For these cases, we suggest the following acceleration: At any node t of the tree, and for all $i \in [d]$, compute*

$$r_i(t) = \sigma_i^{-1}\left(\max_{l \in N(t)} \mu_i^{(l)} - \min_{k \in N(t)} \mu_i^{(k)} \right). \tag{8}$$

Then, for some small integer $d' \ll d$, only minimize (7) over the axes $i_1, \ldots, i_{d'}$ that achieve highest $r_i(t)$. Our theoretical guarantees are preserved even for $d' = 1$.

Remark 6 (Relaxing assumptions). *If the assumption of subexponential tails is considered too restrictive, one should replace Assumption 3 with a heavier-tailed decay rate. This in turn changes Equation (7) and also the theoretical guarantees that can be attained. For the sake of clarity, we stick to Assumption 3 in the following section.*

4 THEORETICAL ANALYSIS OF MMDT

In this section, we theoretically analyze Algorithm 1 in terms of the price of explainability for K-medians. To simplify the analysis, we assume in this section that $\sigma_i = \sigma_i^{(k)}$ for all $i \in [d], k \in [K]$.

Theorem 7 (Upper bounds on price). *Consider a mixture model ν satisfying Assumption 3. Denote $q = \mathrm{ENR}(\nu)$. Then, Algorithm 1 yields a decision tree T such that*

$$\mathrm{Price}(T, \nu) \leq 1 + \frac{c_1 \alpha \cdot q}{\exp\left(\frac{c_2 \cdot q}{2(K-1)}\right) - 1}.$$

The proof is in Appendix A. Theorem 7 quantifies the phenomenon that we alluded to in the introduction: *Mixture models with light tails can usually be explained very well*, with price of explainability close to 1 despite large K. We now provide a lower bound that is almost tight in q for those trees that recover the individual mixture components sufficiently well.

Theorem 8 (Lower bounds on price). *For any $K \geq 3$ and $q \geq (K-1)\log(1.5)$ there exists a mixture model ν with K components, $\mathrm{ENR}(\nu) = q$, equal mixing weights, and constants $c_1 = c_2 = 1$ such that the following holds: For any decision tree T satisfying $\mathbb{P}_{x \sim \nu^{(k)}}(\hat{\mu}(x) = \mu^{(k)}) \geq 0.5$ for all k, and denoting $C_K = \frac{K! - K + 1}{K!}$,*

$$\mathrm{Price}(T, \nu) \geq \frac{C_K(2K-1)}{2K} + \frac{q}{96(K-1) \cdot (e^{\frac{q}{K-1}} - 1)}$$

The proof is included in Appendix B. Since we need the constants c_1, c_2 to be independent of K, our construction differs from previous lower bounds on explainable K-medians.

Remark 9 (Comparison to Random Cuts). *The Random Coordinate Cut algorithm analyzed by Makarychev and Shan [2023] achieves $\mathcal{O}(\log K)$ worst case bounds for K-medians. In contrast, the upper bound for MMDT takes $\mathrm{ENR}(\nu)$ into account. If $\mathrm{ENR}(\nu) = \Theta(K)$, MMDT is loose by a factor of order $\Theta(\frac{K}{\log K})$ compared to Random Coordinate Cut. However, if the data is well-clustered and $\mathrm{ENR}(\nu) = \Omega(K \log K)$, the upper bound for MMDT is tighter. Being greedy at each node pays off if the data is easy to explain.*

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. *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.

5.1 BACKGROUND

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', the maximum mean discrepancy (MMD) is defined 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].

5.2 ASSUMPTIONS

The kernel setting is a bit different from the setup considered in the first part of this paper. Therefore, instead of the assumptions introduced in Sections 2 and 3, we impose three conditions on the mixture ν and the kernel κ. Firstly, we assume that the kernel κ is a bounded, distance-based product kernel

$$\kappa(x, x') = \prod_{i=1}^{d} \kappa_i(x, x') = \prod_{i=1}^{d} g_i(|x_i - x_i'|) \tag{9}$$

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

$$\|\phi_{\nu^{(k)}}\|_{\mathcal{H}}^2 = \mathbb{E}_{x,x'\sim_{i.i.d.}\nu^{(k)}}[\kappa(x, x')] = \sigma^2 \tag{10}$$

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 $i \in [d], k \in [K]$ and $\bar{x} \in \mathrm{supp}(\nu)$, the variance of $\kappa_i(\bar{x}, \cdot)$ is bounded by

$$\mathbb{V}_{x\sim\nu^{(k)}}[\kappa_i(\bar{x}, x)] \leq \epsilon^2 \tag{11}$$

for some (small) $\epsilon > 0$. This is a concentration assumption on the coordinate-wise similarity between any fixed $\bar{x}$ and random $x \sim \nu^{(k)}$. Due to boundedness of the kernel, unbounded support of ν does not pose a problem, though the decay of the g_i impacts how small we can hope ϵ to be.

Next, we extend Definition 1 to kernels.

Definition 10 (Price of explainability in kernel clustering). *Consider a kernel function κ satisfying* (9)*, and a mixture model ν satisyfing* (10) *and* (11)*. Then, we define the price of explainability of ν with respect to κ as*

$$\mathrm{Price}_\kappa(T, \nu) = \frac{E_{x\sim\nu}\left[\|\phi(x) - \tilde{\mu}(x)\|_{\mathcal{H}}^2\right]}{E_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]}. \tag{12}$$

Here, $\tilde{\mu}(x)$ denotes the kernel mean embedding of the distribution ν conditioned on the event that it arrives in the l-th leaf of the decision tree.

Remark 11 (Upper bound on price for kernels). *We can certainly upper bound*

$$\mathrm{Price}_\kappa(T, \nu) \leq \frac{E_{x\sim\nu}\left[\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2\right]}{E_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} \tag{13}$$

where $\hat{\mu}(x) = \phi_{\nu^{(l)}}$ if x ends up in a leaf that is assigned to the l-th mixture component $\nu^{(l)}$. This is true because the second moment functional $\mathcal{M} : \mathcal{H} \to \mathbb{R}$ defined via

$$\mathcal{M}(h) := \mathbb{E}\left[\|\phi(x) - h\|_{\mathcal{H}}^2\right] \tag{14}$$

has its minimum at the mean $h = \mathbb{E}[\phi(x)]$.

The explainability-to-noise ratio (Definition 2) does not translate directly to the kernel setting, where clustering happens in $\mathcal{H}$ instead of $\mathbb{R}^d$. In the following, we instead give our guarantees on $\mathrm{Price}_\kappa(T, \nu)$ in terms of the following axis-aligned quantity.

Definition 12 (Axis-aligned kernel distance). *For a mixture model ν and a kernel κ as introduced before, we define*

$$\tau = \min_{k\neq l} \max_{i\in[d]} \mathbb{E}_{x,x'\sim\nu^{(k)}, y\sim\nu^{(l)}}[\kappa_i(x, x') - \kappa_i(x, y)] \tag{15}$$

as the axis-aligned kernel distance of the mixture components.

Algorithm 2 Kernel Mixture Model Decision Tree (**Kernel MMDT**)

Input: Mixture components $\{\nu^{(k)}\}_{k=1}^K$ on $\mathbb{R}^d$, product kernel $\kappa(x,x') = \prod_{i=1}^d \kappa_i(x,x')$ where $\kappa_i(x,x') = g_i(|x_i - x_i'|)$ and g is monotone decreasing.
Output: Decision tree T with K leaves (one per component).
Root t_0 with $N(t_0) \leftarrow [K]$, active set $\mathcal{A} \leftarrow \{t_0\}$.
while $\exists\, t \in \mathcal{A}$ with $|N(t)| > 1$ **do**
  Pick any such t, set $N \leftarrow N(t)$.
  Find maximizers $(\bar{x}, i^*, k^*, l^*)$ of

$$\xi(x,i,k,\ell) := \mathbb{E}_{x' \sim \nu^{(k)}}[\kappa_i(x,x')] - \mathbb{E}_{y \sim \nu^{(\ell)}}[\kappa_i(x,y)].$$

  For each $m \in N$, compute $\zeta(m) := \mathbb{E}_{y \sim \nu^{(m)}}[\kappa_{i^*}(\bar{x}, y)]$.
  If densities are known, choose $\rho \in (\min \zeta(m), \max \zeta(m))$ such as to minimize (19). If densities are unknown, select ρ as midpoint of largest consecutive gap among sorted values $\{\zeta(m) : m \in N\}$.
  Split index set into $N_L := \{m \in N : \zeta(m) \le \rho\}$ and $N_R := \{m \in N : \zeta(m) > \rho\}$.
  Define $\theta = g_{i^*}^{-1}(\rho)$.
  Store at t the cut $(i^*, \bar{x}_{i^*}, \theta)$, create children t_L, t_R with $N(t_L) = N_L$, $N(t_R) = N_R$.
  Update $\mathcal{A} \leftarrow (\mathcal{A} \setminus \{t\}) \cup \{t_L, t_R\}$.
end while
At node $(i, \bar{x}_i, \theta)$ send x right iff $|\bar{x}_i - x_i| < \theta$ (else left).
If $N(t) = \{k\}$, assign leaf to component k and predict $\hat{\mu}(x) = \phi_{\nu^{(k)}}$.

The axis-aligned kernel distance plays the role of the $\mathrm{ENR}(\nu)$ in a nonparametric setting, except that is defined via inner products instead of distances (which is more natural when working with kernels). 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. The following result is proved in Appendix C.

Lemma 13 (MMD implies bounds on τ). *Consider a mixture model ν such that for any two distinct $\nu^{(k)}, \nu^{(l)}$ it holds that $d_\kappa(\nu^{(k)}, \nu^{(l)}) \ge \gamma$. Then, $\tau \ge \frac{\gamma}{2d}$.*

5.3 ALGORITHM

The main hurdle facing us in adapting the MMDT-algorithm to the kernel setting is that the kernel mean embeddings are no longer elements of $\mathbb{R}^d$. But for our decision tree to be interpretable, its axis-aligned cuts naturally need to be in the input space, not $\mathcal{H}$ (in general, the "coordinates" of $\mathcal{H}$ are functions, not coordinates of $\mathbb{R}^d$). 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, because the clustering cost is measured in the Hilbert space. In fact, the input space means may even coincide despite the mixture components being well-clustered in $\mathcal{H}$.

Example 1. *For real numbers $r_1, \dots, r_K > 0$ denote by $\nu^{(k)}$ the uniform distribution on the sphere of radius r_k in $\mathbb{R}^d$. Let κ be the Gaussian kernel. Then $\kappa(x,x') = \exp(-\|x - x'\|_2^2) \le \exp(r_l - r_k)$ for all $x \in \mathrm{supp}(\nu^{(k)}), x' \in \mathrm{supp}(\nu^{(l)})$. However, $\mathbb{E}_{x \sim \nu^{(k)}}[x] = 0$ for all $k \in [K]$.*

We resolve this by instead choosing *prototype points* $\bar{x}$ at every node, and deciding based on these $\bar{x}$ whether a mixture component goes to the left or the right child node. The intuition is that, for at least one axis $i \in [d]$, $\kappa_i(\bar{x}, x')$ should be larger than some $\rho > 0$ for those x' sampled from the same cluster as $\bar{x}$, and smaller otherwise. One thing changes compared to the previous section: Since

$$\kappa_i(\bar{x}, x) \le \rho \iff |\bar{x}_i - x_i| \ge g_i^{-1}(\rho) =: \theta \qquad (16)$$

the axis-aligned threshold cuts are no longer one sided cuts $x_i \le \theta$ but rather two-sided intervals $|x_i - \bar{x}_i| \le \theta$. This does not harm the interpretability of the resulting tree, as we are still only checking one axis at every node. This way, we can construct a decision tree in the input space, while at the same time relating its clustering cost to properties of the mixture components in $\mathcal{H}$. The very same relaxation was done by Fleissner et al. [2024].

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 $i^* \in [d]$, a pair $k^*, l^* \in N(t)$, and a reference point $\bar{x} \in \mathrm{supp}(\nu^{(k^*)})$ such that

$$\xi(x,i,k,l) := \mathbb{E}_{x' \sim \nu^{(k)}, y \sim \nu^{(l)}}[\kappa_i(x,x') - \kappa_i(x,y)] \quad (17)$$

is maximal. Note that certainly $\xi(\bar{x}, i^*, k^*, l^*) \ge \tau$ at each node if the axis-aligned kernel distance is $\ge \tau$. A threshold value ρ is then decided as follows: For $m \in N(t)$, sort the terms

$$\zeta(m) = \mathbb{E}_{y \sim \nu^{(m)}}[\kappa_{i^*}(\bar{x}, y)] \qquad (18)$$

in non-decreasing order. If densities are known for each mixture component, then choose $\rho \in (\min \zeta(m), \max \zeta(m))$ such that

$$\sum_{k \in N(t)} p^{(k)} \mathbb{P}_{x \sim \nu^{(k)}}\left(\kappa_i(\bar{x}, x), \zeta(k) \text{ lie on other sides of } \rho\right) \qquad (19)$$

is minimized. If densities are unknown, ρ is chosen halfway between the two neighboring terms that are separated furthest. Now, for any $x \sim \nu$ that arrives at this node, we send it left if $\kappa_{i^*}(\bar{x}, x) < \rho$ and right otherwise. To keep track of which child node of t receives the m-th mixture component, we simply evaluate on which side of ρ the scalar $\zeta(m)$ is located.

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 $\hat{\mu}(x) = \phi_{\nu^{(k)}}$. We refer to this algorithm as Kernel MMDT, see Algorithm 2.

Remark 14 (**Choice of threshold in Kernel MMDT**). *Similar as in MMDT, the threshold ρ in Kernel MMDT is chosen with the aim of minimizing the probability that a point x is separated from its corresponding mixture component. For large datasets, choosing the cut halfway in between two neighboring $\zeta(m)$ is a fast proxy.*

We can now give data-dependent guarantees on the price of explainability for kernels that depend on the axis-aligned kernel similarity τ and ϵ, which measures the concentration within each cluster. The proof is in Appendix D.

Theorem 15 (**Price of explainability for kernels**). *Consider a mixture model ν as introduced above. Denote by $\tau > 0$ its axis-aligned kernel distance. Then, it holds that*

$$\mathrm{Price}_\kappa(T, \nu) \leq 1 + \frac{1+\sigma^2}{1-\sigma^2} \cdot \frac{(4 + \frac{2\pi^2}{3})\alpha\epsilon^2(K-1)^2}{\tau^2}$$

where T is the tree fitted using the above algorithm (under known densities).

We conclude this section with two final remarks.

Remark 16 (**Empirical Algorithm**). *For a finite dataset $\mathcal{X} = \{x^{(1)}, \ldots, x^{(n)}\}$ of size n, the algorithm can be run on the empirical estimate $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n \phi(x^{(i)})$ of ν. Note that $\mathrm{supp}(\hat{\nu})$ is finite, and every point $x^{(i)} \in \mathcal{X}$ is a potential prototype point to be selected by Kernel MMDT. If n is large, we propose speeding the algorithm up by sampling a set of $S \in \mathbb{N}$ candidate prototypes $\mathcal{P} = \{x^{(s,k)}\}_{s\in[S],k\in[K]}$.*

Remark 17 (**Generalized threshold cuts**). *As Lemma 13 indicates, the axis-aligned kernel distance τ decreases with the input dimension d, suggesting that finding a "good" axis $i \in [d]$ is difficult in high dimensions, even when the mixture components are well-separated. For Gaussian and Laplace kernel, this can be resolved by allowing generalized threshold cuts of the form*

$$\kappa(\bar{x}, x) \leq \rho \iff \|\bar{x} - x\|_p^p \geq -\log\rho. \qquad (20)$$

where $p = 1$ for Laplace and $p = 2$ for Gaussian kernel. Of course, these cuts are no longer axis-aligned, but instead define ℓ_p balls around prototype points $\bar{x}$ in $\mathbb{R}^d$.

6 EXPERIMENTS

To evaluate how robust our algorithms is, we compare it to IMM, CART, and Random Coordinate Cut on a number of synthetic and real world clustering datasets. First we look at two Gaussian mixture models (GMM) where the components have different covariance structures and non-uniform weights. The first GMM has $K = 5$ clusters in $d = 2$ dimensions, the second one has $K = 10$ clusters in $d = 10$ dimensions. Then, we run both algorithms on wine and the Wisconsin breast cancer dataset. K is chosen based on the true number of classes in the dataset.

The algorithms are **not** provided with the true mixture model but instead estimates weights, means and mean absolute deviations from the data using the GMM algorithm provided in scikit-learn [Pedregosa et al., 2011]. The output of this algorithm is also used to compute the baseline K-medians cost. Note: Since the GMM algorithm may not find the optimal K-medians partition, it is possible that the price of explainability is < 1 if the tree actually clusters slightly better than the estimated mixture model. We do not estimate the entire distribution, but use Equation 7 to compute the optimal threshold cut (i, θ) at each node. For Random Coordinate Cut, we report the average price of explainability over 500 runs. *MMDT outperforms the worst-case optimal RCC on all four datasets*. The difference is particularly pronounced when the assumption of a mixture model with light tails is satisfied. The full result is summarized in Table 1. Figure 2 shows an example with $K = 5$ Gaussians in $\mathbb{R}^2$.

We also include empirical evaluations for Kernel MMDT on three datasets: Iris, MNIST digits 0-3, 20newsgroups. For the latter, the features are constructed via tfidf. We use the Gaussian kernel with the bandwidth chosen proportional to the median Euclidean distance between points from the dataset, and compare with CART and SpeX [Argov and Wagner, 2026]. As Table 2 shows, Kernel MMDT performs comparably well, sometimes slightly weaker. Note that we do not claim to have derived the optimal explainable kernel clustering algorithm — neither of the two competitors analyze price of explainability. Our contribution lies in showing that refined guarantees are possible by taking the data distribution into account.

7 DISCUSSION AND CONCLUSION

This paper provides the first analysis of the explainable clustering problem for mixture models. 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 more realistic guarantees on the price of explainability by incorporating information on the data distribution. It is noteworthy that our ideas 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 Gaussian and Laplace 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 non-parametric mixture models provided the average similarity between samples from distinct mixture components is small, and dimensionality is not too high (see Lemma 13).

There are a few open questions for future works to explore.

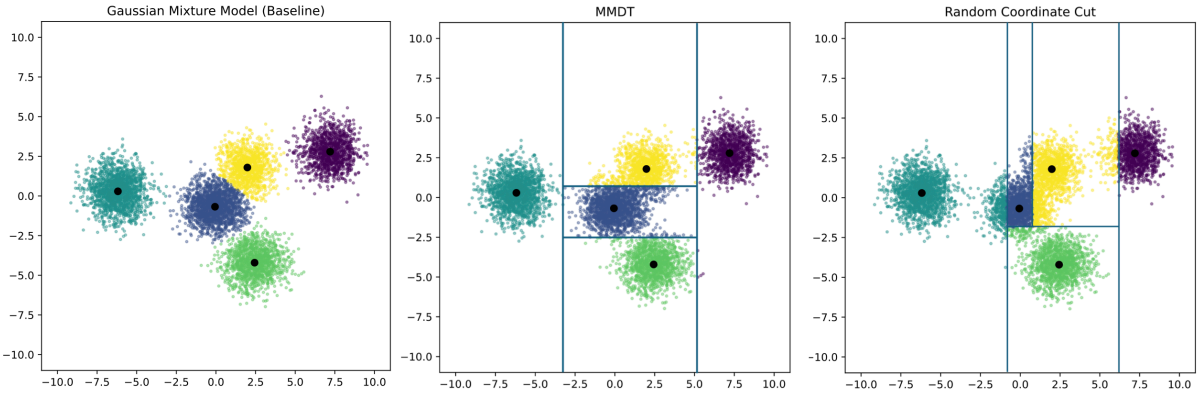

Figure 2: On a Gaussian mixture model with $K = 5$ components (left plot), MMDT exploits the concentration within each cluster to find an axis-aligned tree with low clustering cost (center plot). The distribution-agnostic Random Coordinate Cut randomly samples threshold cuts, leading to higher cost (right plot).

Dataset	**n**	**d**	**K**	**Price MMDT**	**Price CART**	**Price IMM**	**Average Price RCC**
Gaussians I	7081	2	5	1.022	1.021	1.022	1.425
Gaussians II	14970	10	10	1.000	1.010	1.000	1.967
Wine	178	13	3	1.003	1.016	1.003	1.042
Breast Cancer	569	30	2	0.985	0.986	0.986	1.062

Table 1: Price of explainability for four clustering datasets. On average over 500 runs, proposed MMDT outperforms RCC on all four datasets. The improvements are particularly noticeable for Gaussian mixture models. Note: The algorithm is not provided with the true clusters on all datasets, but instead estimates a mixture model.

Dataset	**n**	**d**	**K**	**Price KMMDT**	**Price CART**	**Price SpeX**
Iris	150	4	3	1.0742	1.0636	1.0636
Digits	720	64	4	1.1240	1.0474	1.0263
20newsgroups	2756	1000	3	1.0045	1.0064	1.0040

Table 2: We report the price of explainability for kernel clustering on three popular clustering datasets. We compare with CART and SpeX. The performance is comparably good for all three, indicating that the kernel clustering cost of the decision trees is close to the baseline partition.

For one, this paper assumes exact knowledge of the means and mean absolute deviations of the mixture model ν, and studies *approximation* guarantees for decision trees. 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 move towards *generalization on new data* by incorporating finite sample complexity bounds on estimation of mixture models, e.g. when every component is Gaussian [Ashtiani et al., 2020]. Since our proof mainly relies on the explainability-to-noise ratio $\mathrm{ENR}(\nu)$ that depends only on means and mean absolute deviations, 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].

Acknowledgements

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

References

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

Tal Argov and Tal Wagner. Spex: A spectral approach to explainable clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*,

2026. URL https://openreview.net/forum?id=heoHY4vcRl.

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.

Dimitris Bertsimas, Agni Orfanoudaki, and Holly Wiberg. Interpretable clustering: an optimization approach. *Machine Learning*, 110(1):89–138, 2021.

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.

Bálint Farkas, Béla Nagy, and Szilárd Gy Révész. Fenton type minimax problems for sum of translates functions. *Journal of Mathematical Analysis and Applications*, 543 (2):128931, 2025.

Maximilian Fleissner, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Explaining kernel clustering via decision trees. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FAGtjl7HOw.

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. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1131–1148. IEEE, 2023.

Eduardo Laber, Lucas Murtinho, and Felipe Oliveira. Shallow decision trees for explainable k-means clustering. *Pattern Recognition*, 137:109239, 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.

Eduardo Sany Laber and Lucas Saadi Murtinho. Nearly tight bounds on the price of explainability for the k-center and the maximum-spacing clustering problems. *Theoretical Computer Science*, 949:113744, 2023.

Konstantin Makarychev and Liren Shan. Near-optimal algorithms for explainable k-medians and k-means. In *International Conference on Machine Learning*, pages 7358–7367. 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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=MFWgLCWgUB.

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.

Louis Ohl, Pierre-Alexandre Mattei, Mickaël Leclercq, Arnaud Droit, and Frédéric Precioso. Kernel kmeans clustering splits for end-to-end unsupervised decision trees. *arXiv preprint arXiv:2402.12232*, 2024.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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.

Supplementary Material

Maximilian Fleissner[1] **Maedeh Zarvandi**[1,2] **Debarghya Ghoshdastidar**[1,2]

[1]Technical University of Munich
[2] Munich Center for Machine Learning

A PROOF OF THEOREM 7

Proof. Throughout this proof we assume $c_1 = c_2 = 1$ to avoid carrying all constants with us. Note that $\mathbb{E}_{x\sim\nu}\left[\|x-\mu(x)\|_1\right] = \sigma_1 + \cdots + \sigma_d$. Moreover,

$$\mathbb{E}_{x\sim\nu}\left[\|x-\tilde{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\|x-\hat{\mu}(x)\|_1\right] \tag{21}$$

because $\tilde{\mu}(x)$ is the median of the points that end up in the same leaf as x, and thus achieves the minimum possible $\|\cdot\|_1$ cost. Using the triangle inequality, we obtain

$$\mathbb{E}_{x\sim\nu}\left[\|x-\hat{\mu}(x)\|_1\right] \leq \mathbb{E}_{x\sim\nu}\left[\|x-\mu(x)\|_1\right] + \mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right]. \tag{22}$$

Therefore, the price of explainability is given by

$$\mathrm{Price}(T,\nu) \leq 1 + \frac{\mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right]}{\sigma_1+\cdots+\sigma_d}. \tag{23}$$

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

$$\|\mu(x)-\hat{\mu}(x)\|_1 \leq \sum_{t\in T} \mathbf{1}(x \text{ is separated from } \mu(x) \text{ at node } t) \cdot \max_{k,l\in N(t)} \|\mu^{(k)}-\mu^{(l)}\|_1 \tag{24}$$

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. Taking expectations over $x \sim \nu$, we obtain

$$\mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right] \leq \sum_{t\in T} \mathbb{P}(x \text{ is separated from } \mu(x) \text{ at node } t) \cdot \max_{k,l\in N(t)} \|\mu^{(k)}-\mu^{(l)}\|_1. \tag{25}$$

Let t be an arbitrary node of the tree. For all $j \in [d]$, define $R_j(t) = \max_{k\in N(t)} \mu_j^{(k)} - \min_{l\in N(t)} \mu_j^{(l)}$ for the length of the interval that contains all remaining centers in $N(t)$, projected to the j-th axis. Note that there certainly exists $i \in [d]$ such that $R_i(t) \geq \sigma_i q$. This is true because $\mathrm{ENR}(\nu) = q$.

Recall that Algorithm 1 chooses the threshold cut that *minimizes* the probability of making a mistake at the node t. We can certainly upper bound this probability by analyzing only cuts along the particular axis $i \in [d]$.

Observe that $x \sim \nu$ can only be separated from $\mu(x)$ at node t along axis i if it lies on the wrong side of the threshold cut (i,θ). This can only happen if $|x_i - \mu_i(x)| > |\theta - \mu_i(x)|$. Using Assumption 3, the probability is therefore bounded as

$$\mathbb{P}(x \text{ is separated from } \mu(x) \text{ at node } t) = \sum_{k=1}^{K} p^{(k)} \cdot \mathbb{P}(x \text{ is separated from } \mu^{(k)} \text{ at node } t) \tag{26}$$

$$\leq \frac{\alpha}{K} \sum_{k\in N(t)} \exp\left(-\frac{|\theta-\mu_i^{(k)}|}{\sigma_i}\right) \tag{27}$$

plugging in $p^{(k)} \leq \frac{\alpha}{K}$. 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

$$0 = \mu_i^{(1)} \leq \mu_i^{(2)} \leq \cdots \leq \mu_i^{(K')} = R_i(t). \tag{28}$$

Define the function

$$f_i(\theta) = \sum_{k=1}^{K'} \exp\left(-\frac{|\theta - \mu_i^{(k)}|}{\sigma_i}\right). \tag{29}$$

This is an 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 1: Let $K' \geq 2$, $R_i(t) > 0$, $\sigma_i > 0$, and let $0 = \mu_i^{(1)} \leq \cdots \leq \mu_i^{(K')} = R_i(t)$. Define the function

$$f_i(\theta) := \sum_{k=1}^{K'} \exp\left(-\frac{|\theta - \mu_{j^*}^{(k)}|}{\sigma_{j^*}}\right), \tag{30}$$

and let $\theta^* \in \arg\min_{\theta \in [0, R_i(t)]} f_i(\theta)$. Then, the following upper bound holds.

$$f_i(\theta^*) \leq 2 \sum_{k=1}^{\lceil K'/2 \rceil} \exp\left(-\frac{(2k-1)R_i(t)}{2(K'-1)\sigma_i}\right). \tag{31}$$

Proof of Claim 1: The idea of the proof is to connect to Theorem 1.3 in [Farkas et al., 2025], which gives a very general result on minimax sums of translation-invariant functions. Set $m := K' - 1$, $R := R_i(t)$, $\sigma := \sigma_i$, and $\rho := \frac{R}{\sigma}$. After the change of variables $s = \frac{\theta}{R}$ and with

$$y_k := \frac{\mu_i^{(k+1)}}{R} \tag{32}$$

for all $k = 0, \ldots, m$ we obtain a normalized configuration

$$0 = y_0 \leq \cdots \leq y_m = 1 \tag{33}$$

on the unit interval. We define

$$\Phi_y(s) := \sum_{k=0}^{m} e^{-\rho|s - y_k|}, \tag{34}$$

$$f_i(\theta^*) = \inf_{s \in [0,1]} \Phi_y(s). \tag{35}$$

Then, taking the supremum over the points y_k,

$$f_i(\theta^*) \leq \sup_{0 \leq y_1 \leq \cdots \leq y_{m-1} \leq 1} \inf_{s \in [0,1]} \Phi_y(s). \tag{36}$$

We now apply Theorem 1.3 of Farkas et al. [2025] to what is referred to as a non-singular kernel $K(u) := -e^{-\rho|u|}$ and a field function $J(s) := -e^{-\rho s} - e^{-\rho(1-s)}$. Note that $K(u)$ is piecewise concave and monotone on $(-1, 0) \cup (0, 1)$, and that $J(s)$ is finite everywhere (thus, for any m, it is a m-field function). Let

$$F(y, s) := J(s) + \sum_{k=1}^{m-1} K(y_k - s). \tag{37}$$

Their result asserts that, in our notation,

$$\inf_{0 \leq y_1 \leq \cdots \leq y_{m-1} \leq 1} \sup_{s \in [0,1]} F(y, s) = \sup_{0 \leq y_1 \leq \cdots \leq y_{m-1} \leq 1} \min_{0 \leq j \leq m-1} \sup_{s \in [y_j, y_{j+1}]} F(y, s). \tag{38}$$

Note that $F(y,s) = -\Phi_y(s)$. By switching signs, this implies that

$$\sup_{0\le y_1\le\cdots\le y_{m-1}\le 1}\ \inf_{s\in[0,1]} \Phi_y(s) = \inf_{0\le y_1\le\cdots\le y_{m-1}\le 1}\ \max_{0\le j\le m-1}\ \inf_{s\in[y_j,y_{j+1}]} \Phi_y(s). \tag{39}$$

The left hand side is what we're after. The right hand side can be upper bounded by evaluating the respective $\max\inf$ objective for the equidistant grid $y_k = \frac{k}{m}$ where $k = 0,\dots,m$. This grid places all points at a distance of $\frac{1}{m}$. For each interval $I_j(y) := \left[\frac{j}{m}, \frac{j+1}{m}\right]$ where $j = 0,\dots,m-1$, define the midpoint $s_j := \frac{2j+1}{2m}$. Surely then, for all $j = 0,\dots,m-1$, we have

$$\inf_{s\in I_j(z)} \Phi_z(s) \le \Phi_z(s_j). \tag{40}$$

At such a midpoint, the two nearest grid points are at a distance of $\frac{1}{2m}$, the next two possible grid points are at distance of $\frac{3}{2m}$, and so on. Since there are only $m = K' - 1$ grid points, we obtain

$$\Phi_z(s_j) \le 2\sum_{k=1}^{\lceil m/2\rceil} \exp\left(-\frac{\rho\cdot(2k-1)}{2m}\right). \tag{41}$$

This bound is independent of j. Consequently,

$$\sup_y \inf_{s\in[0,1]} \Phi_y(s) \le 2\sum_{k=1}^{\lceil m/2\rceil} \exp\left(-\frac{\rho\cdot(2k-1)}{2m}\right). \tag{42}$$

Substituting back $m = K' - 1$, and $\rho = \frac{R_i(t)}{\sigma_i}$ gives the claimed result. □

Combining this result with the previous steps and using $K' \le K$, we see that at any node t, the probability of a mistake is bounded as follows.

$$\begin{aligned}
\mathbb{P}(x \text{ separated from } \mu(x) \text{ at } t) &\le \frac{2\alpha}{K}\sum_{k=1}^{\lceil K'/2\rceil} \exp\left(-\frac{(2k-1)R_i(t)}{2(K'-1)\sigma_i}\right) \\
&\le \frac{2\alpha}{K}\sum_{k=1}^{\infty} \exp\left(-\frac{R_i(t)}{2(K'-1)\sigma_i}\right)^{(2k-1)} \\
&\le \frac{2\alpha}{K}\cdot\frac{1}{2}\cdot\sum_{k=1}^{\infty} \exp\left(-\frac{R_i(t)}{2(K'-1)\sigma_i}\right)^{k} \\
&= \frac{\alpha}{K}\cdot\frac{\exp\left(-\frac{R_i(t)}{2(K'-1)\sigma_i}\right)}{1-\exp\left(-\frac{R_i(t)}{2(K'-1)\sigma_i}\right)} \\
&= \frac{\alpha}{K}\cdot\frac{1}{\exp\left(\frac{R_i(t)}{2(K-1)\sigma_i}\right)-1}
\end{aligned} \tag{43}$$

Returning to the price of explainability, we first note that

$$\max_{k,l\in N(t)} \|\mu^{(k)} - \mu^{(l)}\|_1 \le \sum_{j=1}^{d} R_j(t) \tag{44}$$

Therefore,

$$\begin{aligned}\frac{\mathbb{E}_{x\sim\nu}\left[\|\mu(x)-\hat{\mu}(x)\|_1\right]}{\mathbb{E}_{x\sim\nu}\left[\|x-\mu(x)\|_1\right]} &\leq \sum_{t\in T}\frac{\sum_{j=1}^{d}R_j(t)}{\sum_{j=1}^{d}\sigma_j}\cdot\frac{\alpha}{K}\cdot\frac{1}{\exp\left(\frac{R_i(t)}{2(K-1)\sigma_i}\right)-1}\\ &\leq \sum_{t\in T}\frac{R_i(t)}{\sigma_i}\cdot\frac{\alpha}{K}\cdot\frac{1}{\exp\left(\frac{R_i(t)}{2(K-1)\sigma_i}\right)-1}\\ &\leq \sum_{t\in T}\frac{\alpha q}{K}\cdot\frac{1}{\exp\left(\frac{q}{2(K-1)}\right)-1}\\ &\leq \frac{\alpha q}{\exp\left(\frac{q}{2(K-1)}\right)-1}.\end{aligned} \tag{45}$$

In the third step, we used the fact that the function $x\mapsto\frac{x}{\exp(x)-1}$ is decreasing for all $x\geq 0$, and that $\frac{R_i(t)}{\sigma_i}\geq q$ by definition of the explainability-to-noise ratio. In the final step, we use that there are no more than $K-1$ nodes in the tree. □

B PROOF OF THEOREM 8

Proof. Fix any $K\in\mathbb{N}$ and $q\geq 1$. Define $\sigma=\frac{K-1}{q}>0$. We first construct means $\mu^{(k)}$. Let $d=K!$ and for all $j\in[d]$, let Π_j be a permutation of the set $\{0,\dots,K-1\}$. Then, define $\mu_j^{(k)}=\Pi_j(k)$ for all $j\in[d],k\in[K]$. Thus, for all $l\neq k$, we have $\|\mu^{(l)}-\mu^{(k)}\|_1\geq\frac{dK}{3}$. Let $\sigma=\frac{K-1}{q}$. For all $j\in[d],k\in[K]$, pick each $\nu^{(k)}$ to be a Laplace distribution (with independent marginals of scale σ) and mean $\mu^{(k)}$. This construction ensures that the mean absolute deviation is σ. Moreover, the distribution is subexponential with constants $c_1=1,c_2=1$. For all $j\in[d],k\in[K],s>0$, standard results on the CDF for Laplace random variables give

$$\mathbb{P}_{x\sim\nu^{(k)}}(x_j\geq\mu_j^{(k)}+s)=\frac{1}{2}\exp\left(-\frac{s}{\sigma}\right). \tag{46}$$

We have uniform mixing weights, so $\nu=\frac{1}{K}\sum_{k=1}^{K}\nu^{(k)}$. Thus, $\mathrm{ENR}(\nu)=\frac{K-1}{\sigma}=q$ and $\mathbb{E}[\|x-\mu(x)\|_1]=\sum_{j=1}^{d}\mathbb{E}[|x_j-\mu_j(x)|]=d\sigma$. Since the distribution is symmetric, means and medians coincide, so this is also the baseline K-medians cost. Now consider as tree T that puts every $\mu^{(k)}$ in its own leaf. At its root, the probability of putting a sample $x\sim\nu$ to a leaf that dos not contain $\mu(x)$ is at least

$$\begin{aligned}\mathbb{P}(x\text{ separated from }\mu(x)) &\geq \frac{1}{2K}\sum_{k=1}^{K-1}\exp\left(-\frac{k}{\sigma}\right)\\ &= \frac{1}{2K}\frac{e^{-\frac{1}{\sigma}}-e^{-\frac{K}{\sigma}}}{1-e^{-\frac{1}{\sigma}}}\\ &\geq \frac{1}{4K}\frac{1}{e^{\frac{1}{\sigma}}-1}.\end{aligned} \tag{47}$$

The last line uses $q\geq(K-1)\log 1.5\geq\log 2$. Denote by $\tilde{\mu}(x)$ the median of the leaf that a sample x ends up. Denote by $\mathcal{I}(T)\subset[d]$ the coordinates used in a tree T. To lower bound the price of explainability, we will rely on the lower bound

$$\mathbb{E}_{x\sim\nu}[\|x-\tilde{\mu}(x)\|_1\geq\sum_{j\notin\mathcal{I}(T)}\mathbb{E}_{x\sim\nu}[|x_j-\tilde{\mu}_j(x)|]. \tag{48}$$

Note: For $j\notin\mathcal{I}(T)$, the distribution of x_j is independent of any events that depend on the tree (such as whether the point was split in other axes). Therefore, we can later condition on these events without changing the expectation.

For all $l\neq k$, the distance between the means restricted to the coordinates $j\notin\mathcal{I}(T)$ is still of order dK. This is true because the worst that can happen for any pair $l\neq k$ is that the coordinates in $\mathcal{I}(T)$ (of which there are no more than $K-1$ since the depth of the tree is at most K) remove exactly $K-1$ of those $2(K-2)!$ coordinates where the distance between $\mu^{(k)}$ and $\mu^{(l)}$ is $K-1$. Thus,

$$\sum_{j\notin\mathcal{I}(T)}|\mu_j^{(k)}-\mu_j^{(l)}|\geq\frac{dK}{3}-(K-1)^2\geq dK\cdot\left(\frac{1}{3}-\frac{K-1}{K^2}\right)\geq\frac{dK}{12}. \tag{49}$$

This already shows a weaker result, namely that

$$\begin{aligned}\frac{\mathbb{E}_{x\sim\nu}[\|x-\hat{\mu}(x)\|_1]}{\mathbb{E}_{x\sim\nu}[\|x-\mu(x)\|_1]} &\geq \frac{1}{d\sigma}\cdot\Big(\sum_{j\notin\mathcal{I}(T)} P(\hat{\mu}(x)=\mu(x))\cdot\mathbb{E}[|x_j-\mu(x)_j|]+P(\hat{\mu}(x)\neq\mu(x))\cdot\frac{dK}{12}\Big)\\ &\geq \frac{d-K+1}{d}+\frac{K}{12\sigma}\cdot\frac{1}{4K}\frac{1}{e^{\frac{1}{\sigma}}-1}\\ &= \underbrace{\frac{d-K+1}{d}}_{\approx 1 \text{ for large } K}+\frac{q}{48(K-1)}\cdot\frac{1}{e^{\frac{q}{K-1}}-1}.\end{aligned} \tag{50}$$

To proceed further, we need to account for the fact that $\hat{\mu}(x)\neq\tilde{\mu}(x)$. We start with a lower bound on the median of a sum of Laplace random variables with "distant" means.

Claim 2: Let $K\in\mathbb{N}, p\in\mathbb{N}$ and $\alpha_1,\dots,\alpha_K\in\mathbb{R}^p$, and $\sigma>0$. For all $k\in[K]$ let $L_k\sim\mathrm{Lap}(\alpha_k,\sigma)$ denote the p-dimensional Laplace distribution with mean $\alpha_k\in\mathbb{R}^p$ and independent marginals of scale σ. Suppose that there exists $C>0$ such that $\|\alpha_k-\alpha_l\|_1>C$ for all $l\neq k$. Let $p_1\geq p_2\geq\cdots\geq p_K\geq 0$ denote mixing weights that sum to one, and let $m\in\mathbb{R}^p$ be the median of the mixture. Then,

$$\sum_{k=1}^{K} p_k\mathbb{E}[\|L_k-m\|_1]\geq p_1\cdot p\sigma+(1-p_1)\left(0.5C+p\sigma e^{-\frac{C}{2p\sigma}}\right). \tag{51}$$

Proof of Claim 2: For $k\in[K]$, define $r_k:=\|\alpha_k-m\|_1$. Since the coordinates of L_k are independent Laplace random variables with scale σ, the one-dimensional identity $\mathbb{E}_{X\sim\mathrm{Lap}(\mu,\sigma)}[|X-a|]=|\mu-a|+\sigma\exp\left(-\frac{|\mu-a|}{\sigma}\right)$ yields

$$\begin{aligned}\mathbb{E}[\|L_k-m\|_1] &= \|\alpha_k-m\|_1+\sigma\sum_{j=1}^{p}\exp\left(-\frac{|e_j^\top\alpha_k-m_j|}{\sigma}\right)\\ &\geq \|\alpha_k-m\|_1+p\sigma\cdot\exp\left(-\frac{\|\alpha_k-m\|_1}{p\sigma}\right).\end{aligned} \tag{52}$$

In the second step we used the AM-GM inequality, i.e. $\sum_{i=1}^n u_i\geq n\prod_{i=1}^n u_i^{1/n}$ for any $u_1,\dots,u_n\geq 0$. Therefore,

$$\mathbb{E}[\|L_k-m\|_1]\geq r_k+p\sigma\exp\left(-\frac{r_k}{p\sigma}\right)=\psi(r_k), \tag{53}$$

where we define the function $\psi(t):=t+p\sigma\exp(-\frac{t}{p\sigma})$ for all $t\geq 0$. Observe that $\psi(0)=p\sigma$ and $\psi'(t)=1-\exp(-\frac{t}{p\sigma})\geq 0$. Thus, ψ is non-decreasing. Let $i\in[K]$ be the index of the mean α_i that is closest to the median. We write

$$r_i=\min_{k\in[K]} r_k=:c. \tag{54}$$

We distinguish two cases.

1. If $c\geq 0.5C$, then $r_k\geq 0.5C$ for all k, and thus

$$\sum_{k=1}^{K} p_k\psi(r_k)\geq\psi(0.5C)=0.5C+p\sigma\cdot\exp\left(-\frac{0.5C}{p\sigma}\right)\geq p_1\cdot p\sigma+(1-p_1)\left(0.5C+p\sigma e^{-\frac{C}{2p\sigma}}\right). \tag{55}$$

2. If $c<0.5C$, then for all **other** $k\neq i$, the triangle inequality gives

$$r_k=\|\alpha_k-m\|_1\geq\|\alpha_k-\alpha_i\|_1-\|\alpha_i-m\|_1>C-c\geq 0.5C. \tag{56}$$

Therefore, using that ψ is non-decreasing and that p_1 is the largest weight,

$$\begin{aligned}\sum_{k=1}^{K} p_k\psi(r_k) &\geq p_i\psi(r_i) + \sum_{k\neq i} p_k\psi(r_k)\\ &\geq p_i\psi(0) + \sum_{k\neq i} p_k\psi\left(0.5C\right)\\ &\geq p_1\psi(0) + \sum_{k\neq 1} p_k\psi\left(0.5C\right)\\ &= p_1 p\sigma + (1-p_1)\cdot\left(0.5C + p\sigma\exp\left(-\frac{0.5C}{p\sigma}\right)\right).\end{aligned} \tag{57}$$

This proves the claim. □

We are ready to complete the proof. Let $p_{k\to l} = \mathbb{P}(\mu(x) = \mu^{(k)}, \hat{\mu}(x) = \mu^{(l)})$ denote the probability mass of component k assigned to leaf l. Summing the cost over all leaves and using the law of total expectation,

$$\begin{aligned}\mathbb{E}_{x\sim\nu}[\|x-\tilde{\mu}(x)\|_1] &= \sum_{l,k=1}^{K} \mathbb{P}_{x\sim\nu}(\mu(x)=\mu^{(k)}, \hat{\mu}(x)=\mu^{(l)})\cdot\mathbb{E}_{x\sim\nu}\left[\|x-\tilde{\mu}^{(l)}\|_1\middle|\mu(x)=\mu^{(k)}, \hat{\mu}(x)=\mu^{(l)}\right]\\ &\geq \sum_{j\notin\mathcal{I}(T)}\sum_{l,k=1}^{K} \mathbb{P}_{x\sim\nu}(\mu(x)=\mu^{(k)}, \hat{\mu}(x)=\mu^{(l)})\cdot\mathbb{E}_{x\sim\nu^{(k)}}\left[|x_j-\tilde{\mu}_j^{(l)}|\middle|\hat{\mu}(x)=\mu^{(l)}\right]\\ &= \sum_{l,k=1}^{K} p_{k\to l}\sum_{j\notin\mathcal{I}(T)}\mathbb{E}_{x\sim\nu^{(k)}}[|x_j-\tilde{\mu}_j^{(l)}|]\end{aligned} \tag{58}$$

We used the fact that for $x\sim\nu^{(k)}$, the random variable x_j is independent of the event $\hat{\mu}(x)=\mu^{(l)}$ for $j\notin\mathcal{I}(T)$. As stated before, for all $k\neq k'$, we have $\sum_{j\notin\mathcal{I}(T)}|\mu_j^{(k)}-\mu_j^{(k')}| \geq \frac{dK}{12} =: C$. Now, we may invoke Claim 2 for all leaves $l\in[K]$. Define $\bar{p}_l := \sum_{k=1}^{K} p_{k\to l}$ for the probability mass of the l-th leaf. Recall that we restrict ourselves to trees T satisfying

$$\forall k\in[K]: \mathbb{P}_{x\sim\nu^{(k)}}(\hat{\mu}(x)=\mu(x)) \geq \frac{1}{2}. \tag{59}$$

Due to equal mixing weights, this implies $p_{l\to l}\geq\frac{1}{2K}$ for all l. Furthermore, for all $l\neq k$, $p_{k\to l}\leq\frac{1}{2K}$ and hence $\max_k p_{k\to l} = p_{l\to l}$. Define $p = d - |\mathcal{I}(T)|$. Then,

$$\begin{aligned}\sum_{l,k=1}^{K} p_{k\to l}\sum_{j\notin\mathcal{I}(T)}\mathbb{E}_{x\sim\nu^{(k)}}[|x_j-\tilde{\mu}_j^{(l)}|] &= \sum_{l=1}^{K}\bar{p}_l\cdot\sum_{k=1}^{K}\frac{p_{k\to l}}{\bar{p}_l}\sum_{j\notin\mathcal{I}(T)}\mathbb{E}_{x\sim\nu^{(k)}}[|x_j-\tilde{\mu}_j^{(l)}|]\\ &\geq \sum_{l=1}^{K}\bar{p}_l\cdot\left(\frac{p_{l\to l}}{\bar{p}_l}\cdot p\sigma + \left(1-\frac{p_{l\to l}}{\bar{p}_l}\right)\cdot\left(0.5C + p\sigma e^{-\frac{0.5C}{p\sigma}}\right)\right)\\ &= \sum_{l=1}^{K} p_{l\to l}\cdot p\sigma + (\bar{p}_l - p_{l\to l})\cdot\left(0.5C + p\sigma e^{-\frac{0.5C}{p\sigma}}\right)\\ &= p\sigma + \sum_{l=1}^{K}(\bar{p}_l - p_{l\to l})\cdot\left(0.5C + p\sigma e^{-\frac{0.5C}{p\sigma}} - p\sigma\right)\end{aligned} \tag{60}$$

where we used $\sum_{l=1}^{K}\bar{p}_l = 1$ in the last line. Furthermore,

$$\sum_{l=1}^{K}\bar{p}_l - p_{l\to l} = \sum_{l=1}^{K}\sum_{k\neq l} p_{k\to l} = \mathbb{P}(\hat{\mu}(x)\neq\mu(x)) \geq \frac{1}{4K}\frac{1}{e^{\frac{1}{\sigma}}-1}. \tag{61}$$

Plugging this into the bound,

$$\mathbb{E}_{x\sim\nu}[\|x-\tilde{\mu}(x)\|_1] \geq p\sigma + \frac{1}{4K}\cdot\frac{0.5C + p\sigma e^{-\frac{0.5C}{p\sigma}} - p\sigma}{e^{\frac{1}{\sigma}}-1}. \tag{62}$$

Recall $C = \frac{dK}{12}$ and $q = \frac{K-1}{\sigma}$, so $\frac{1}{\sigma} = \frac{q}{K-1}$ and $\frac{0.5C}{4d\sigma K} = \frac{q}{96(K-1)}$. Let $C_K := \frac{K!-K+1}{K!}$ as defined in the theorem. Note: $\frac{p}{d} \geq C_K$ since $p \geq d - K + 1$. Dividing the cost of the tree by the baseline cost $d\sigma$, some algebra gives

$$\begin{aligned}
\text{Price}(T, \nu) &\geq C_K + \frac{1}{e^{\frac{q}{K-1}} - 1} \left(\frac{q}{96(K-1)} + \frac{p}{4Kd} \left(e^{-\frac{C}{2p\sigma}} - 1 \right) \right) \\
&\geq C_K + \frac{1}{e^{\frac{q}{K-1}} - 1} \left(\frac{q}{96(K-1)} - \frac{p}{4Kd} \right) \\
&\geq C_K - \frac{p}{2Kd} + \frac{q}{96(K-1)\left(e^{\frac{q}{K-1}} - 1 \right)} \\
&\geq \frac{C_K(2K-1)}{2K} + \frac{q}{96(K-1)\left(e^{\frac{q}{K-1}} - 1 \right)}.
\end{aligned} \tag{63}$$

where $p \leq d$ and $q \geq (K-1)\log 1.5$ are used. This completes the proof. □

C PROOF OF LEMMA 13

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

$$\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} [\kappa(x, x')] = \|\phi_{\nu^{(k)}}\|^2_{\mathcal{H}} = \sigma^2. \tag{64}$$

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

$$\gamma \leq d_\kappa(\nu^{(k)}, \nu^{(l)}) \tag{65}$$
$$= \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} [\kappa(x, x')] + \mathbb{E}_{y,y' \sim \nu^{(l)}} [\kappa(y, y')] - 2\mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} [\kappa(x, y)] \tag{66}$$
$$= 2\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} [\kappa(x, x')] - 2\mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} [\kappa(x, y)]. \tag{67}$$

Therefore, we can write

$$\begin{aligned}
\gamma/2 &\leq \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}} [\kappa(x, x')] - \mathbb{E}_{x \sim \nu^{(k)}, y \sim \nu^{(l)}} [\kappa(x, y)] \\
&= \mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}, y \sim \nu^{(l)}} \left[\prod_{i=1}^{d} g_i(|x_i - x'_i|) - \prod_{i=1}^{d} g_i(|x_i - y_i|) \right].
\end{aligned} \tag{68}$$

Next, we use the fact that for any set of positive real numbers $a_i, b_i \in (0, 1]$ we can write

$$\begin{aligned}
\prod_{i=1}^{d} a_i - \prod_{i=1}^{d} b_i &= (a_1 - b_1) \prod_{i=2}^{d} a_i + b_1(a_2 - b_2) \prod_{i=3}^{d} a_i + b_1 b_2 (a_3 - b_3) \prod_{i=4}^{d} a_i + \dots \\
&= \sum_{i=1}^{d} (a_i - b_i) \underbrace{\prod_{j<i} b_j}_{\in (0,1]} \cdot \underbrace{\prod_{j>i} a_j}_{\in (0,1]} \\
&\leq \sum_{i: a_i \geq b_i} (a_i - b_i).
\end{aligned} \tag{69}$$

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

$$a_i - b_i \geq \frac{\prod_{i=1}^{d} a_i - \prod_{i=1}^{d} b_i}{d}. \tag{70}$$

It follows from Equation (68) that there certainly exists $i \in [d]$ with

$$\mathbb{E}_{x,x' \sim_{i.i.d.} \nu^{(k)}, y \sim \nu^{(l)}} [g_i(|x_i - x'_i|) - g_i(|x_i - y_i|)] \geq \frac{\gamma}{2d}. \tag{71}$$

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

□

D PROOF OF THEOREM 15

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

$$\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2 = \|\phi(x)\|_{\mathcal{H}}^2 + \|\hat{\mu}(x)\|_{\mathcal{H}}^2 - 2\langle \phi(x), \hat{\mu}(x)\rangle \leq 1 + \sigma^2. \tag{72}$$

Here, we plugged in $\|\phi(x)\|_{\mathcal{H}}^2 = \kappa(x,x) = 1$ and then used the fact that $\hat{\mu}(x)$ is a kernel mean embedding to get

$$\langle \phi(x), \hat{\mu}(x)\rangle = \mathbb{E}_{y\sim\nu^{\hat{\mu}(x)}}[\kappa(x,y)] \geq 0. \tag{73}$$

Thus,

$$\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2 \leq \|\phi(x) - \mu(x)\|_{\mathcal{H}}^2 + \mathbf{1}(\hat{\mu}(x) \neq \mu(x)) \cdot (1 + \sigma^2). \tag{74}$$

Therefore,

$$\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2\right] \leq \mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right] + \mathbb{P}(\hat{\mu}(x) \neq \mu(x)) \cdot (1 + \sigma^2). \tag{75}$$

This implies that

$$\begin{aligned}\frac{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \hat{\mu}(x)\|_{\mathcal{H}}^2\right]}{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} &\leq \frac{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right] + \mathbb{P}(\hat{\mu}(x) \neq \mu(x)) \cdot (1 + \sigma^2)}{\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right]} \\ &= 1 + \frac{1+\sigma^2}{1-\sigma^2} \cdot \mathbb{P}(\hat{\mu}(x) \neq \mu(x))\end{aligned} \tag{76}$$

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

$$\mathbb{E}_{x\sim\nu}\left[\|\phi(x) - \mu(x)\|_{\mathcal{H}}^2\right] = 1 - \sigma^2 \tag{77}$$

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 $N(t)$. The algorithm chooses $\bar{x}, i^*, k^*, l^* \in N(t)$ that maximize

$$\xi(\bar{x}, i, k, l) := \mathbb{E}_{x'\sim_{i.i.d.}\nu^{(k)}, y\sim\nu^{(l)}}[\kappa_i(\bar{x}, x') - \kappa_i(\bar{x}, y)]. \tag{78}$$

Note that $\mathbb{E}_{\bar{x}\sim\nu^{(k^*)}}[\xi(\bar{x}, i^*, k^*, l^*)] \geq \tau$ by definition of the axis-aligned kernel similarity. Therefore, $\xi(\bar{x}, i, k, l) \geq \tau$. Without loss of generality, we may assume that $N(t) = [K']$ for some $K' \leq K$, and that $k^* = 1, l^* = 2$. For any $m \in [K']$ define

$$\zeta(m) = \mathbb{E}_{y\sim\nu^{(m)}}[\kappa_{i^*}(\bar{x}, y)]. \tag{79}$$

We may assume that $\zeta(m) \leq \zeta(m+1)$ for all $2 \leq m \leq K'$. If this does not hold, simply sort and relabel. We know that $\zeta(2) \leq \zeta(1) - \tau$. Therefore, the scalars $\{\zeta(m)\}_{m=1}^{K'}$ are spread out on an interval of length at least $\Delta \geq \tau$. Recall that the threshold θ is chosen exactly halfway where the jump between $\zeta(m)$ and $\zeta(m+1)$ is largest. For any mixture component $m \in N(t)$ and any $x \sim \nu^{(m)}$, 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 $\kappa_{i^*}(\bar{x}, x) < \theta < \zeta(m)$ or vice versa. This can only happen if

$$|\kappa_{i^*}(\bar{x}, x) - \zeta(m)| > |\theta - \zeta(m)|. \tag{80}$$

We can now use Chebyshev's inequality to bound the probability of making mistakes.

$$\begin{aligned}\mathbb{P}_{x\sim\nu}(x \text{ goes to wrong child at } t) &\leq \sum_{m\in N(t)} p^{(m)} \cdot \mathbb{P}_{x\sim\nu^{(m)}}(|\kappa_{i^*}(\bar{x}, x) - \zeta(m)| > |\theta - \zeta(m)|) \\ &\leq \sum_{m\in N(t)} p^{(m)} \cdot \frac{\epsilon^2}{|\theta - \zeta(m)|^2} \\ &\leq \frac{\alpha\epsilon^2}{K} \sum_{m\in N(t)} \frac{1}{|\theta - \zeta(m)|^2}.\end{aligned} \tag{81}$$

Now we argue just as we did in the proof of Theorem 7 (see Claim 1 from Appendix A), except that we have quadratic concentration and not exponential concentration. This is not a problem for Theorem 1.3 in [Farkas et al., 2025], as the (singular) kernel $K(u) = -u^{-2}$ is still concave and monotone on $(-1, 0)$ and $(0, 1)$. The corresponding field function is $J(s) = -s^{-2} - (1-s)^{-2}$, which is bounded from above on $[0, 1]$ and finite everywhere except at the interval endpoints.

Thus, an upper bound occurs when all $\zeta(m)$ are equidistant on the interval of length Δ and $N(t) = [K]$. In that case, $|\zeta(m+1) - \zeta(m)| = \frac{\Delta}{K-1}$ for all $m \leq K-1$. Then,

$$\begin{aligned}\sum_{m \in N(t)} \frac{1}{|\theta - \zeta(m)|^2} &\leq 2 \sum_{k=1}^{\lceil K/2 \rceil} \frac{1}{(2k-1)^2 (\frac{\Delta}{2(K-1)})^2} \\ &\leq \frac{8(K-1)^2}{\Delta^2} \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \\ &\leq \frac{8(K-1)^2(0.5 + \pi^2/12)}{\Delta^2}.\end{aligned} \tag{82}$$

This implies (at every node)

$$\mathbb{P}_{x \sim \nu} \left(x \text{ goes to wrong child} \right) \leq \frac{(4 + \frac{2\pi^2}{3}) \alpha \epsilon^2 (K-1)^2}{K \tau^2}. \tag{83}$$

Since there are no more than K nodes, we obtain

$$\mathbb{P}_{x \sim \nu} (\hat{\mu}(x) \neq \mu(x)) \leq \frac{(4 + \frac{2\pi^2}{3}) \alpha \epsilon^2 (K-1)^2}{\tau^2}. \tag{84}$$

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