# Connecting Interpretability and Robustness in Decision Trees through Separation

**Michal Moshkovitz** [1] **Yao-Yuan Yang** [1] **Kamalika Chaudhuri** [1]

## Abstract

Recent research has recognized interpretability and robustness as essential properties of trustworthy classification. Curiously, a connection between robustness and interpretability was empirically observed, but the theoretical reasoning behind it remained elusive. In this paper, we rigorously investigate this connection. Specifically, we focus on interpretation using decision trees and robustness to $l_\infty$-perturbation. Previous works defined the notion of $r$-separation as a sufficient condition for robustness. We prove upper and lower bounds on the tree size in case the data is $r$-separated. We then show that a tighter bound on the size is possible when the data is linearly separated. We provide the first algorithm with provable guarantees both on robustness, interpretability, and accuracy in the context of decision trees. Experiments confirm that our algorithm yields classifiers that are both interpretable and robust and have high accuracy. The code for the experiments is available at https://github.com/yangarbiter/interpretable-robust-trees.

## 1. Introduction

Deploying machine learning (ML) models in high-stakes fields like healthcare, transportation, and law, requires the ML models to be trustworthy. Essential ingredients of trustworthy models are explainability and robustness: if we do not understand the reasons for the model's prediction, we cannot trust the model; if small changes in the input modifies the model's prediction, we cannot trust the model. Previous works hypothesized that there is a strong connection between robustness and explainability. They empirically observed that robust models lead to better explanations (Chen et al., 2019; Ross & Doshi-Velez, 2017). In this work, we take a rigorous approach towards understanding the connection between robustness and interpretability.

We focus on binary predictions, where each example has $d$ features and the label of each example is in $\{-1,+1\}$, so an ML model is a hypothesis $f:\mathbb{R}^d \to \{-1,1\}$. We want our model to be (i) robust to adversarial $\ell_\infty$ perturbations, i.e., for a small distortion, $\|\delta\|_\infty$, the model's response is similar, $f(x)=f(x+\delta)$, for most examples $x$, (ii) interpretable, i.e., the model itself is simple and so self-explanatory, and (iii) have high-accuracy. A common type of interpretable models are decision trees (Molnar, 2019), which we call *tree-based explanation* and focus on in this paper.

Prior literature (Yang et al., 2020b) showed that data *separation* is a sufficient condition for a robust and accurate classifier. A dataset is $r$-separated if the distance between the two closest examples with different labels is at least $2r$. Intuitively, if $r$ is large, then the data is well-separated. A separated data guarantees that points with opposite labels are far from each other, which is essential to construct a robust model.

In this paper, we examine whether separation implies tree-based explanation. We first show that for a decision tree to have accuracy strictly above $1/2$ (i.e., better than random), the data must be bounded. Henceforth, we assume that the data is in $[-1,1]^d$. We start with a trivial algorithm that constructs a tree-based explanation with complexity (i.e., tree size) $2^{O(d/r)}$. For constant $r$, we show a matching lower bound of $2^{\Omega(d)}$. Thus, we have a matching lower and upper bound on the explanation size of $2^{\Theta(d)}$. Thus, separation implies robustness and interpretability. Unfortunately, for a large number of features, $d$, the explanation size is too high to be useful in practice.

In this paper, we show that designing a simpler explanation is possible with a stronger separability assumption — linear separability with a $\gamma$-margin. This assumption was recently used to gain a better understanding of neural networks (Soudry et al., 2018; Nacson et al., 2019; Shamir, 2020). More formally, this assumption means that there is a vector $w$ with $\|w\|=1$ such that $yw\cdot x \geq \gamma$ for each labeled example $(x,y)\in\mathbb{R}^d\times\{-1,1\}$ in the data (Shalev-Shwartz

[1] University of California, San Diego. Correspondence to: Michal Moshkovitz <mmoshkovitz@eng.ucsd.edu>, Yao-Yuan Yang <yay005@eng.ucsd.edu>.

*Proceedings of the 38<sup>th</sup> International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

& Ben-David, 2014).

One can hope that standard methods for learning linear models will suffice, but this may not be the case. Standard linear models such as $\ell_2$-regularized logistic regressions or support vector machines (Shalev-Shwartz & Ben-David, 2014) may produce models that use too many features (in other words, the weights are not sparse), and this can make the model not interpretable. Many other approaches (Rudin, 2019; Bertsimas et al., 2019) try to solve this issue by enforcing sparsity on the weight vector. However, these models may not be adversarially robust. In this paper, our goal is to find a model that is interpretable, robust, and have a high-accuracy.

Utilizing ideas from (Shalev-Shwartz & Singer, 2008), we show that under the linearity assumption, there is always at least one feature that provides non-trivial information for the prediction. To formalize this, we use the known notion of *weak learners* (Kearns, 1988), which guarantees the existence of hypothesis with accuracy bounded below by more than $1/2$.

The weak-learnability theorem, together with Kearns & Mansour (1999), implies that a popular CART-type algorithm (Breiman et al., 1984) provides a decision tree with size $1/\epsilon^{O(1/\gamma^2)}$ and accuracy $1-\epsilon$. Therefore, under the linearity assumption, we can design a tree with complexity independent of the number of features. Thus, even if the number of features, $d$, is large, the interpretation complexity is not affected. This achieves our first goal of constructing an interpretable model with provable guarantees.

Recently, several research papers give a theoretical justification for CART's empirical success (Brutzkus et al., 2020; 2019; Blanc et al., 2019; 2020; Fiat & Pechyony, 2004). Those papers assume that the underlying distribution is uniform or features chosen independently. For many cases, this assumption does not hold. For example, in medical data, there is a strong correlation between age and different diseases. On the other hand, we give a theoretical justification for CART without resorting to the feature-independence assumption. We use, instead, the linear separability assumption. We believe that this method will allow, in the future, proofs with less restrictive assumptions.

So far, we have shown how to construct an interpretable model, but we want a model that is not just interpretable but also robust. Decision trees are not robust by-default (Chen et al., 2019). For example, a slight change in the feature at the root of the decision tree leads to an entirely different model (and thus to entirely different predictions): the model defined by the left subtree and the model defined by the right subtree. We are left with the question, are we able to constrct a tree that is both robust and interpretable. To design such model, we focus on a specific kind of decision tree — risk score (Ustun & Rudin, 2017). A risk score is composed of several conditions (e.g., $age \geq 75$) and each matched with a weight, i.e., a small integer. A score $s(x)$ of an example $x$ is the weighted sum of all the satisfied conditions. The label is then a function of the score $s(x)$. A risk score is a specific case of decision trees, wherein at each level in the tree, the same feature is queried. The number of parameters required to represent a risk score is much smaller than their corresponding decision trees, hence they might be considered more interpretable than decision trees (Ustun & Rudin, 2017).

We design a new learning algorithm, BBM-RS, for learning risk scores that rely on the Boost-by-Majority (BBM) algorithm (Freund, 1995) and our weak learner theorem. It yields a risk score of size $O(\gamma^{-2}\log(1/\epsilon))$ and accuracy $1-\epsilon$. Thus, we found an algorithm that creates a risk score with provable guarantees on size and accuracy. As a side effect, note that BBM allows to control the interpretation complexity easily. Importantly, we show that risk scores are also guaranteed to be robust to $\ell_\infty$ perturbations, by deliberately adding a small noise to dataset (but not too much noise to make sure that the noisy dataset is still linearly separable). Therefore, we design a model that is guaranteed to have high accuracy and be both interpretable and robust, achieving our final goal.

Finally, in Section 6, we test the validity of the separability assumption and the quality of the new algorithm on real-world datasets that were used previously in tree-based explanation research. On most of the datasets, less than $12\%$ points were removed to achieve an $r$-separation with $r \geq 0.05$. For comparison, for binary feature-values $\{-1, 1\}$, and $\ell_\infty$ distance, the best value for $r$ is $r = 1$. The added percentage of points required to be removed for the dataset to be linearly separable is less than $7\%$ on average. Thus, we observe that real datasets are close to being separable and even linearly separable. Then, we explored the quality of our new algorithm, BBM-RS. Even though it has provable guarantees only if the data is linearly separable, we run it on real datasets that do not satisfy this property. We compare BBM-RS to different algorithms learning: decision trees (Quinlan, 1986), small risk scores (Ustun & Rudin, 2017), and robust decision trees (Chen et al., 2019). All algorithms try to maximize accuracy, but different algorithms try to, additionally, minimize interpretation complexity or maximize robustness. None of the algorithms aimed to optimize both interpretability and robustness. We compared the (i) interpretation complexity, (ii) robustness, and (iii) accuracy of all four algorithms. We find that our algorithm provides a model with better interpretation complexity and robustness while having comparable accuracy.

To summarize, our main contributions are:

**Interpretability under separability: optimal bounds.** We show lower and upper bounds on decision tree size for

$r$-separable data with $r = \Theta(1)$, of $2^{\Theta(d)}$. Namely, our upper bound proves that for any separable data, there is a tree of size $2^{O(d)}$, and the lower bound proves that separability cannot guarantee an explanation smaller than $2^{\Omega(d)}$.

**Algorithm with provable guarantees on interpretability and robustness.** Designing algorithms that have provable guarantees both on interpretability, robustness, and accuracy in the context of decision trees is highly sought-after, yet there was no such algorithm before our work. We design the first learning algorithm that has provable guarantees both on interpretability, robustness, and accuracy of the returned model, under the assumption that the data is linearly separable with a margin.

While the CART algorithm is empirically highly effective, its theoretical analysis has been elusive for a long time. As a side effect, we provide an analysis of CART under the assumption of linear separability. To the best of our knowledge, this is the first proof with a distributional assumption that does not include feature independence.

**Experiments.** We verify the validity of our assumptions empirically and show that for real datasets, if a small percentage of points is removed then we get a linear separable dataset. We also compare our new algorithm to other algorithms that return interpretable models (Quinlan, 1986; Ustun & Rudin, 2017; Chen et al., 2019) and show that if the goal is to design a model that is both interpretable and robust, then our method is preferable.

## 2. Related Work

**Post-hoc explanations.** There are two main types of explanations: post hoc explanations (Ribeiro et al., 2016a) and intrinsic explanations (Rudin, 2019). Algorithms for post hoc explanation take as an input a black-box model and return some form of explanation. Intrinsic explanations are simple models, so the models are self-explanatory. The main advantage of algorithms for post hoc explanations (Lundberg & Lee, 2017; Lundberg et al., 2018; Ribeiro et al., 2016b; Koh & Liang, 2017; Ribeiro et al., 2018; Deutch & Frost, 2019; Li et al., 2020; Boer et al., 2020) is that they can be used on any model. However, they host a variety of problems: they introduce a new source of error stemming from the explanation method (Rudin, 2019); they can be fooled (Lakkaraju & Bastani, 2020; Slack et al., 2020); some explanations methods are not robust to common pre-processing steps (Kindermans et al., 2019), and can be independent both of the model and the data generating process (Adebayo et al., 2018). Because of the critics against post hoc explanations, in this paper, we focus on intrinsic explanations.

**Explainability and robustness.** Prior studies research the intersection of explanation and robustness of black-box models (Lakkaraju et al., 2020), decision trees (Chen et al., 2019; Andriushchenko & Hein, 2019), and deep neural networks (Szegedy et al., 2013; Goodfellow et al., 2014; Madry et al., 2017; Ross & Doshi-Velez, 2017). Unfortunately, the quality of these methods are only verified empirically. On the theoretical side, most works analyzed explainability and robustness separately. Explainability was researched for supervised learning (Garreau & von Luxburg, 2020b;a; Mardaoui & Garreau, 2020; Hu et al., 2019) and unsupervised learning (Moshkovitz et al., 2020; Frost et al., 2020; Laber & Murtinho, 2021). For robustness, Cohen et al. (2019) showed that the technique of randomized smoothing has robustness guarantees. Ignatiev et al. (2019) connected adversarial examples and a different type of explainability from the point of view of formal logic.

**Risk scores.** Ustun and Rudin (Ustun & Rudin, 2017) designed a new algorithm for learning risk scores by solving an appropriate optimization problem. They focused on constructing an interpretable model with high accuracy and did not consider robustness, as we do in this work.

## 3. Preliminaries

We investigate models that are (i) with high-accuracy, (ii) robust, and (iii) interpretable, as formalized next.

**High accuracy.** We consider the task of binary classification over a domain $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\mu$ be an underlying probability distribution[1] over labeled examples $\mathcal{X} \times \{-1, +1\}$. The input to a learning algorithm $\mathcal{A}$ consists of a labeled sample $S \sim \mu^m$, and its output is a hypothesis $h : \mathcal{X} \to \{-1, +1\}$. The accuracy of $h$ is equal to $\Pr_{(x,y)\sim\mu}(h(x) = y)$. The sample complexity of $\mathcal{A}$ under the distribution $\mu$, denoted $m(\epsilon, \delta) : (0, 1)^2 \to \mathbb{N}$, is a function mapping desired accuracy $\epsilon$ and confidence $\delta$ to the minimal positive integer $m(\epsilon, \delta)$ such that for any $m \geq m(\epsilon, \delta)$, with probability at least $1 - \delta$ over the drawn of an i.i.d. sample $S \sim \mu^m$, the output $A(S)$ has accuracy of at least $1 - \epsilon$.

**Robustness.** We focus on the $\ell_\infty$ ball, $\mathbb{B}$, and denote the $r$-radius ball around a point $x \in \mathcal{X}$ as $\mathbb{B}(x, r)$. A hypothesis $h : \mathcal{X} \to \{-1, +1\}$ is *robust* at $x$ with radius $r$ if for all $x' \in \mathbb{B}(x, r)$ we have that $h(x) = h(x')$. In (Wang et al., 2018), the notion of *astuteness* was introduced to measure the robustness of a hypothesis $h$. The astuteness of $h$ at radius $r > 0$ under a distribution $\mu$ is

$$\Pr_{(x,y)\sim\mu}[\forall x' \in \mathbb{B}(x, r).\ h(x') = y].$$

For a hypothesis to have high astuteness the positive and negative examples need to be separated. A binary labeled data is *$r$-separated* if for every two labeled examples $(x^1, +1)$,$(x^2, -1)$, it holds that $\|x^1 - x^2\|_\infty \geq 2r$.

[1] In the paper, we will assume that $\mu$ has additional properties, like separation or linear separation.

**Interpretability.** We focus on intrinsic explanations, also called interpretable models (Rudin, 2019), where the model itself is the explanation. There are several types of interpretable models, e.g., logistic regression, decision rules, and anchors (Molnar, 2019). One of the most fundamental interpretable models, which we focus on in this paper, is *decision trees* (Quinlan, 1986). In a decision tree, each leaf corresponds to a label, and each inner node corresponds to a threshold and a feature. The label of an example is the leaf's label of the corresponding path.

In this paper we focus on a specific type of decision trees, *risk score* (Ustun & Rudin, 2019); see Table 1. Risk score is defined by a series of $m$ conditions and a weight for each condition. Each condition compares one feature to a threshold, and the weights should be small integers. A score, $s(x)$, of an example $x$ is the number of satisfied conditions out of the $m$ conditions, each multiplied by the corresponding weight. The prediction of the risk model $f$ is the sign of the score, $f(x) = sign(s(x))$. A risk score can be viewed as a decision tree where at each level there is the same feature-threshold pair. Since the risk-score model has fewer parameters than the corresponding decision tree, it may be considered more interpretable.

| feature | weights | | |
|---|---|---|---|
| | LCPA | BBM-RS | |
| Bias term | -6 | -7 | + ... |
| Age $\geq 75$ | - | 2 | + ... |
| Called in Q1 | 1 | 2 | + ... |
| Called in Q2 | -1 | - | + ... |
| Called before | 1 | 4 | + ... |
| Previous call was Successful | 1 | 2 | + ... |
| Employment variation rate $< -1$ | 5 | 4 | + ... |
| Consumer price index $\geq 93.5$ | 1 | - | + ... |
| 3 month euribor rate $\geq 200$ | -2 | - | + ... |
| 3 month euribor rate $\geq 400$ | 5 | - | + ... |
| 3 month euribor rate $\geq 500$ | 2 | - | + ... |
| | | total score = | |

*Table 1.* Two risk score models: LCPA (Ustun & Rudin, 2019) and our new BBM-RS algorithm on the bank dataset (Moro et al., 2014). Each satisfied condition is multiplied by its weight and summed. Bias term is always satisfied. If the total score $> 0$, the risk model predicts "1" (i.e., the client will open a bank account after a marketing call). All features are binary (either 0 or 1).

# 4. Separation and Interpretability

We want to understand whether separation implies the existence of a small tree-based explanation. Our first observation is that the data has to be bounded for a tree-based explanation to exist. If the data is unbounded, then to achieve a training error slightly better than random, the tree size must depend on the size of the training data, see Section 4.1, Theorem 1.

In Section 4.2 we investigate lower and upper bounds for decision tree's size, assuming separation. Specifically, in Theorem 2, we show that if the data is bounded, in $[-1, 1]^d$, then $r$-separability implies a tree based-explanation with tree depth $O(^d/_r)$. Importantly, the depth of the tree is independent of the training size, so a tree-based explanation exists. Nevertheless, even for a constant $r$, the size of the tree is exponential in $d$. In Theorem 3, we show that this bound is tight as there is a 1-separable dataset that requires an exponential size to achieve accuracy even negligibly better than random. To conclude, if all we know is that the data is $r$-separability for constant $r$, the interpretation complexity is $2^{\Theta(d)}$. Unfortunately, this explanation has size exponential in $d$. In Section 5, we improve the interpretation complexity by assuming a stronger separability assumption. We will assume linear separability with a margin. All proofs are in Section A.1.

## 4.1. Bounded

In Theorem 1, we show that the data has to be bounded for a small decision tree to exist. In fact, boundedness is necessary, even if the data is constrained to be linearly separable. For any tree size $s$ and a given accuracy, we can construct a linearly-separable dataset such that any tree of size $s$ cannot have the desired accuracy.

**Theorem 1.** *For any tree size $s$ and $\gamma > 0$, there is a dataset in $\mathbb{R}^2$ that is linearly separable, and any decision tree with size $s$ has accuracy less than $\frac{1}{2} + \gamma$.*

## 4.2. Upper and lower bounds

Assuming the data in $[-1, 1]^d$ is $r$-separated, Theorem 2 tells us that one can construct a decision tree with depth $6d/r$ and training error 0 (and from standard VC-arguments also accuracy $1 - \epsilon$, with enough examples). Importantly, the depth of the tree is independent of the training size $n$. Nevertheless, it means the size of the trees is exponential in $d$. The idea of the proof is to fine-grain the data to bins of size about $r$, in each coordinate. From this construction, it is clear that the returned model is robust at any training data.

**Theorem 2.** *For any labeled data in $[-1, 1]^d \times \{-1, 1\}$ that is $r$-separated, there is a decision tree of depth at most $\frac{6d}{r}$ which has a training error 0.*

Theorem 3 proves a matching lower bound by constructing a dataset such that any tree that achieves error better than random, the tree size must be exponential in $d$. The dataset proving this lower bound is parity. More specifically, it contains the points $\{-1, +1\}^d$ and the label of each point $x$ is the xor of all of its coordinates.

**Theorem 3.** *There is a labeled dataset in $[-1, 1]^d$ which is 1-separated and has the following property. For any $\gamma > 0$*

*and any decision tree $T$ that achieves accuracy $0.5 + \gamma$, the size of $T$ is at least $\gamma 2^d$.*

## 5. Linear Separability

In the previous section, we showed that $\Theta(1)$-separability implies a decision tree with size exponential in $d$, and we showed a matching lower bound. This section explores a stronger assumption than separability that will guarantee a smaller tree, i.e., a simpler explanation. This assumption is that the data is linearly separable with a margin. More formally, data is $\gamma$-linearly separable if there is $w \in \mathbb{R}^d$, $\|w\|_1 = 1$, such that for each positive example $x$ it holds that $w \cdot x \geq \gamma$ and for each negative example $x$ it holds that $w \cdot x \leq -\gamma$. Note that without loss of generality $w_i \geq 0$ (if the inequality does not hold, multiply the $i$-th feature in each example by $-1$). Thus, we can interpret $w$ as a distribution over all the features. Linear separability might seem at first like a strong assumption, but besides being a widespread assumption (Soudry et al., 2018; Nacson et al., 2019; Shamir, 2020), in Section 6 we show that this assumption is reasonable for real datasets.

As a first attempt, one might hope that $w$ is a good explanation, but this explanation might use all the features, and the corresponding tree-based explanation might be of exponential size. As a second attempt, one might take the highest $w_i$'s, since one might interpret the highest $w_i$ as the most important feature. However, this can be misleading. For example, if all data has the same value at the $i$-th feature, this feature is meaningless. In this section, we explore a different approach for constructing an interpretable model.

One of our key ideas is to use boosting method (Schapire & Freund, 2013) to construct a model which is both interpretable, robust, and accurate. This will allow us to gradually add features to the model until we achieve a high-accuracy model. To implement this idea, we show that one feature can provide a nontrivial prediction. In particular, in Section 5.1 we show that the hypotheses class, $\mathcal{H}_t = \{h_{i,\theta}\}$, is a weak learner, where

$$h_{i,\theta}(x) = \begin{cases} +1 & \text{if } x_i \geq \theta \\ -1 & \text{o.w.} \end{cases}$$

This class is similar to the known decision stumps class, but it does not contain hypotheses of the form "if $x_i \leq \theta$ then $+1$ else $-1$". The reason will become apparent in Section 5.3, but for now, we will hint that it helps achieve robustness.

In Section 5.2, we observe that weak learnability immediately implies that the known CART algorithm constructs a tree of size independent of $d$ (Kearns & Mansour, 1999). Unfortunately, decision trees are not necessarily robust. To overcome this difficulty, we focus on one type of decision trees, risk scores, which are interpretable models on their own. In Section 5.3 we show how to use (Freund, 1995) together with our weak learnability theorem to construct a risk score model. We also show that this model is robust. This concludes our quest of finding a model that is *guaranteed* to be robust, interpretable, and have high-accuracy under the linearity separable assumption. In Section 6 we will evaluate the model on several real datasets.

### 5.1. Weak learner

This section shows that under the linearity assumption, we can always find a feature that gives nontrivial information, which is formally defined using the concept of a *weak learner* class. We say that a class $\mathcal{H}$ is a weak learner if for every distribution $\mu$ over the examples and a function $f$ that are $\gamma$-linearly separable, there is hypothesis $h \in \mathcal{H}$ such that $\Pr_{x \sim \mu}(h(x) = f(x))$ is strictly larger than $1/2$, preferably at least $1/2 + \Omega(\gamma)$. Finding the best hypothesis in $\mathcal{H}_t$ can be done efficiently using dynamic programming (Shalev-Shwartz & Ben-David, 2014). The question is how to prove that there must be a weak learner in $\mathcal{H}_t$.

One might suspect that if the data is linearly separable by the vector $w$ (i.e., for each labeled example $(x, y)$ it holds that $ywx \geq \gamma$), then $h_i$ which corresponds to the highest $w_i$ is a weak learner. Conversely, if $w_i$ is small, then the corresponding hypotheses $h_i$ will have a low accuracy. These claims are not true. To illustrate this, think about the extreme example where $w_1 = 0$ but $x_1$ completely predicts the output of any example $x$. From the viewpoint of $w$, the first feature is irrelevant, as it does not contribute to the term $w \cdot x$, but the first feature is a perfect predictor.

One can prove that there is always a hypothesis in $\mathcal{H}_t$ with accuracy $0.5 + \Omega(\gamma)$ by binarizing the input and applying (Shalev-Shwartz & Singer, 2008). More specifically, they formed a different connection between linear separability and weak learning. They view each example in the hypotheses basis, and on this basis, the famous minimax theorem implies that linearity is equivalent to weak learnability. In this paper, we focus on the case that the data, *in its original form*, is linearly separable. Nonetheless, when the features are binary, the two views, the original and hypotheses bases, coincide.

For completeness, in the appendix, Section A.2, we provide a different proof of Theorem 4, by viewing $\mathcal{H}_t$ as a graph. Namely, define a bipartite graph where the vertices are the examples and the hypotheses and there is an edge between a hypothesis $h$ and example $x$ if $h$ correctly predicts $x$. The edges of the graph are defined so that (i) the degree of the hypotheses vertices corresponds to its accuracy and (ii) the linearity assumption ensures that the degree of the example vertices is high. These two properties of the graph proves the theorem.

**Theorem 4.** *Fix $\alpha > 0$. For any data in $[-1,1]^d \times \{-1,1\}$ that is labeled by a $\gamma$-linearly separable hypothesis $f$ and for any distribution $\mu$ on the examples, there is a hypothesis $h \in \mathcal{H}_t$ such that*

$$\Pr_{x\sim\mu}(h(x) = f(x)) \geq \frac{1}{2} + \frac{\gamma}{2} - \alpha.$$

So far, we showed the existence of hypothesis in $\mathcal{H}_t$ with accuracy $0.5 + \Omega(\gamma)$. Standard arguments in learning theory imply that the hypothesis that maximizes the accuracy on a sample also has accuracy $0.5 + \Omega(\gamma)$. Specifically, for any sample $S$, denote by $h_S$ the best hypothesis in $\mathcal{H}_t$ on the sample $S$. Basic arguments in learning theory shows that for a sample of size $m = O\left(d+\log\frac{1}{\delta}/\gamma^2\right)$, the hypothesis $h_S$ has a good accuracy, as the following theorem proves.

**Theorem 5** (weak-learner)**.** *Fix $\alpha > 0$. For any distribution $\mu$ over $[-1,+1]^d \times \{-1,+1\}$ that satisfies linear separability with a $\gamma$-margin, and for any $\delta \in (0,1)$ there is $m = O\left(\frac{d+\log\frac{1}{\delta}}{\gamma^2}\right)$, such that with probability at least $1-\delta$ over the sample $S$ of size $m$, it holds that*

$$\Pr_{(x,y)\sim\mu}(h_S(x) = y) \geq \frac{1}{2} + \frac{\gamma}{4} - \alpha.$$

### 5.2. Decision tree using CART

CART is a popular algorithm for learning decision trees. In (Kearns & Mansour, 1999) it was shown that if the internal nodes define a $\gamma$-weak learner and number of samples is some polynomial of $t\log(1/\delta)d$, then a CART-type algorithm returns a tree with size $t = 1/\epsilon^{O(1/\gamma^2)}$ and accuracy at least $1-\epsilon$, with probability at least $1-\delta$. Under the linearity assumption, we know that the internal nodes indeed define a $\gamma$-weak learner by Theorem 5. Thus, we get a model with a tree size independent of the training size and the dimension. But the model is not necessarily robust.

The above results can be interrupted as a proof for the CART's algorithm success. This proof does not use the strong assumption of feature independence, which is assumed in recent works (Brutzkus et al., 2020; 2019; Blanc et al., 2019; 2020; Fiat & Pechyony, 2004).

Designing robust decision trees is inherently a difficult task. The reason is that, generally, the model defined by the right and left subtrees can be completely different. The feature $i$ in the root determines if the model uses the right or left subtrees. Thus, a small change in the $i$-th feature completely changes the model. To overcome this difficulty we focus on a specific type of decision tree, risk scores (Ustun & Rudin, 2019), see Table 1 for an example. In the decision tree that corresponds to the risk score, the right and left subtrees are the same. In the next section, we design risk scores that have guarantees on the robustness and the accuracy.

### 5.3. Risk score

This section designs an algorithm that returns a risk score model with provable guarantees on its accuracy and robustness, assuming that the data is linearly separable. In the previous section, we used (Kearns & Mansour, 1999) that viewed CART as a boosting method. This section uses a more traditional boosting method — the Boost-by-Majority algorithm (BBM) (Freund, 1995). This boosting algorithm gets as an input training data and an integer $T$, and at each step $t \leq T$ it reweigh the examples and apply a $\gamma$-weak learner that returns a hypothesis $h_t : \mathbb{R}^d \to \{-1,+1\}$. At the end, after $T$ steps, BBM returns $sign\left(\sum_{t=1}^{T} h_t\right)$. In (Freund, 1995; Schapire & Freund, 2013) it was shown that BBM returns hypothesis with accuracy at least $1-\epsilon$ after at most $T = O(\gamma^{-2}\log(1/\epsilon))$ rounds.

The translation from BBM, which uses $\mathcal{H}_t$ as a weak learner, to a risk score model, is straightforward. The hypotheses in $\mathcal{H}_t$ exactly correspond to the conditions in the risk score. Each condition has weight of 1. If the number of conditions that hold is at least $T/2$ then our risk model returns $+1$, else it returns $-1$. Together with Theorem 4 and (Freund, 1995) we get that BBM returns a risk score with accuracy at least $1-\epsilon$ and with $T = O(\gamma^{-2}\log(1/\epsilon))$ conditions.

We remark that other boosting methods, e.g., (Freund & Schapire, 1997; Kanade & Kalai, 2009), cannot replace BBM in the suggested scheme, since the final combination has to be a simple sum of the weak learners and *not* arbitrary linear combination. The letter corresponds to a risk score where the weights are in $\mathbb{R}$ and not a small integer, as desired.

Our next and final goal is to prove that our risk score model is also robust. For that, we use the concept of *monotonicity*. For $x, y \in \mathbb{R}^d$, we say that $x \leq y$ if and only if for all $i \in [d]$ it holds that $x_i \leq y_i$. A model $f : \mathbb{R}^d \to \{0,1\}$ is monotone if for all $x \leq y$ it holds that $f(x) \leq f(y)$. We will show that BBM with weak learners from $\mathcal{H}_t$ yields a monotone model. The reasons being (i) all conditions are of the form "$x_i \geq \theta$", (ii) all weights are non-negative, except the bias term, and (iii) classification of a risk score is detriment by the score's sign. All proofs appear in Section A.3.

**Claim 6.** *If every condition in a risk-score model $R$ is of the form "$x_i \geq \theta$" and all weights are positive, except the bias term, then $R$ is a monotone model.*

In Claim 7 we show that, by carefully adding a small noise to each feature, we can transform any algorithm that returns a monotone model to one that returns a robust model.

**Claim 7.** *Assume a learning algorithm $A$ gets as an input a sample from a $\gamma$-linearly separable data and returns a* monotone *model with accuracy $1-\epsilon(\gamma)$. Then, there is an algorithm that returns a model with astuteness at least*

$1 - \epsilon\left(\frac{\gamma}{2}\right)$ *at radius* $\gamma/2$.

To summarize, in Algorithm 1 we show the pseudocode of our new algorithm, BBM-RS. In the first step we add noise to each example by replacing each example $(x, y)$ by $(x - \tau y \mathbf{1}, y)$, where $\tau \in (0, 1)$ is a parameter that defines the noise level and $\mathbf{1}$ is the all-one vector. In other words, we add noise $y\tau$ to each feature. In the second step, the algorithm iteratively adds conditions to the risk score. At each iteration, we first find the distribution $\mu$ defined by BBM (Freund, 1995). Then, we find the best hypothesis $h_{i,\theta}$ in $\mathcal{H}_t$, according to $\mu$. We add to the risk score a condition "$x_i \geq \theta$". Finally, we add a bias term of $-T/2$, to check if at least half of the conditions are satisfied.

**Algorithm 1** BBM-RS (BBM-Risk Score)

```
input: $D$: linearly separable training data by $w$; WLOG $\forall i.w_i \geq 0$
        $T$: bound on interpretation complexity
        $\tau$: noise level
output: risk score
# Add noise:
for $(x, y) \in D$ do
  replace $(x, y)$ with $(x - \tau y \mathbf{1}, y)$
end for
for $i = 1 \dots T$ do
  $\mu \leftarrow$ BBM distrbution on $D$
  $i, \theta \leftarrow \arg\max_{i,\theta} \sum_{(x,y) \in D} \mu(x) I_{(x_i - \theta)y > 0}$
  Add condition "$x_i \geq \theta$" to $RS$
end for
Add a bias term of $-T/2$ to $RS$
return $RS$
```

## 6. Experiments

In previous sections, we designed new algorithms and gave provable guarantees for separated data. We next investigate these results on real datasets. Concretely, we ask the following questions:

- How separated are real datasets?
- How well does BBM-RS perform compared with other interpretable methods?
- How do interpretability, robustness, and accuracy tradeoff with one another in BBM-RS?

**Datasets.** To maintain compatibility with prior work on interpretable and robust decision trees (Ustun & Rudin, 2019; Lin et al., 2020), we use the following pre-processed datasets from their repositories – adult, bank, breastcancer, mammo, mushroom, spambase, careval, ficobin, and campasbin. We also use some datasets from other sources such as LIBSVM (Chang & Lin, 2011) datasets and Moro et al. (2014). These include diabetes, heart, ionosphere, and bank2. All features are normalized to $[0, 1]$. More details can be found in Appendix B. The dataset statistics are shown in Table 2.

| | dataset statistics | | | | $r$-separation | | $\gamma$-linear separation | |
|---|---|---|---|---|---|---|---|---|
| | # samples | # features | # binary features | portion of positive label | sep. | $2r$ | sep. | $\gamma$ |
| adult | 32561 | 36 | 36 | 0.24 | 0.88 | 1.00 | 0.84 | 0.001 |
| bank | 41188 | 57 | 57 | 0.11 | 0.97 | 1.00 | 0.90 | 0.33 |
| bank2 | 41188 | 63 | 53 | 0.11 | 1.00 | 0.0004 | 0.91 | 0.00002 |
| breastcancer | 683 | 9 | 0 | 0.35 | 1.00 | 0.11 | 0.97 | 0.0003 |
| careval | 1728 | 15 | 15 | 0.30 | 1.00 | 1.00 | 0.96 | 0.003 |
| compasbin | 6907 | 12 | 12 | 0.46 | 0.68 | 1.00 | 0.65 | 0.20 |
| diabetes | 768 | 8 | 0 | 0.65 | 1.00 | 0.11 | 0.77 | 0.0008 |
| ficobin | 10459 | 17 | 17 | 0.48 | 0.79 | 1.00 | 0.70 | 0.33 |
| heart | 270 | 20 | 13 | 0.44 | 1.00 | 0.13 | 0.89 | 0.0003 |
| ionosphere | 351 | 34 | 1 | 0.64 | 1.00 | 0.80 | 0.95 | 0.0007 |
| mammo | 961 | 14 | 13 | 0.46 | 0.83 | 0.33 | 0.79 | 0.14 |
| mushroom | 8124 | 113 | 113 | 0.48 | 1.00 | 1.00 | 1.00 | 0.02 |
| spambase | 4601 | 57 | 0 | 0.39 | 1.00 | 0.000063 | 0.94 | 0.000002 |

*Table 2.* Dataset statistics. Columns "sep." records the separateness of each dataset. Columns "$2r$" and "$\gamma$" are calculated after dataset is separated by removing $1 -$ sep points.

### 6.1. Separation of real datasets

To understand how separated they are, we measure the closeness of each dataset to being $r$- or linearly separated. The *separateness* of a dataset is one minus the fraction of examples needed to be removed for it to be $r$- or linearly separated.

For $r$-separation, we use the algorithm designed by Yang et al. (2020a) that calculates the minimum number of examples needed to be removed for a dataset to be $r$-separated with $r \geq 10^{-5}$. This ensures that after removal, there will be no pair of examples that are very similar but with different labels. Finding the optimal separateness for linear separation is NP-hard (Ben-David et al., 2003), thus we run a $\ell_1$ regularized linear SVM with regularization terms $C = \{10^{-10}, 10^{-8}, \dots, 10^{10}\}$ and record the lowest training error as an approximation to one minus the optimal separateness.

The separation results are shown in Table 2. Eight datasets are already $r$-separated (separateness $= 100\%$). In the five datasets with separateness $< 100\%$, there are examples with very similar features but different labels. This occurs mostly in binarized datasets; see Appendix C for an example. Three datasets are almost separated with separateness equal to $97\%$, $88\%$, and $83\%$, and two have separateness $68\%$ and $79\%$. To summarize, $84\%$ of the datasets are $r$-separated with $r \geq 10^{-5}$, after removing at most $17\%$ of the points.

Linear separation is a stricter property than $r$-separation, so the separateness for linear separation is smaller or equal to the separateness for $r$-separation. Seven datasets have separateness $\geq 90\%$, three separateness between $79\%$ and $89\%$, and the remaining three have separateness $< 79\%$. After removing the points, all datasets are $\gamma$-linearly separable and nine datasets have $\gamma \geq 0.001$. To summarize, (i) $77\%$ of the datasets are close to being linearly separated (ii) requiring linear-separability reduces the separateness of the $r$-separated dataset by only an average of $6.77\%$. From this we conclude that for these datasets at least, the assumption of $r-$ or linear-separability is approximately correct.

## 6.2. Performance of BBM-RS

Next, we want to understand how our proposed BBM-RS performs on real datasets. We compare the performance of BBM-RS with three different baselines on three evaluation criteria: interpretability, accuracy, and robustness.

**Baselines.** We compare BBM-RS with three baselines: (i) LCPA (Ustun & Rudin, 2019), an algorithm for learning risk scores, (ii) DT (Breiman et al., 1984), standard algorithm for learning decision trees, and (iii) Robust decision tree (RobDT) (Chen et al., 2019), an algorithm for learning robust decision trees.

We use a 5-fold cross-validation based on accuracy for hyperparameters selection. For DT and RobDT, we search through $5, 10, \ldots 30$ for the maximum depth of the tree. For BBM-RS, we search through $5, 10, \ldots 30$ for the maximum number of weak learners ($T$). The algorithm stops when it reaches $T$ iterations or if no weak learner can produce a weighted accuracy $> 0.51$. For LCPA, we search through $5, 10, \ldots 30$ for the maximum $\ell_0$ norm of the weight vector. We set the robust radius for RobDT and the noise level $\tau$ for BBM-RS to $0.05$. More details about the setup of the algorithms can be found in Appendix B.

### 6.2.1. Evaluation

We evaluate interpretability, accuracy, and robustness of each baseline. The data is randomly split into training and testing sets by 2:1. The experiment is repeated 10 times with different training and testing splits. The mean and standard error of the evaluation criteria are recorded.

**Interpretability.** We measure a model's interpretability by evaluating its *Interpretation Complexity (IC)*, which is the number of feature-thresholds pairs in the model (one can think of this as the number of tests the model performs). For decision trees (DT and RobDT), the IC is the number of internal nodes in the tree, and for risk scores (LCPA and BBM-RS), the number of non-zero terms in the weight vector. The lower the IC is, the more interpretable the model is. This is a global measure of the models' complexity as we constructed a model which is self-explainable. One can also measure the local complexity of the model, measured by depth.

**Robustness.** We measure model's robustness by evaluating its *Empirical robustness (ER)* (Yang et al., 2020a). ER on a classifier $f$ at an input $x$ is $ER(f, x) := \min_{f(x') \neq f(x)} \|x' - x\|_\infty$. We evaluate ER on 100 randomly chosen correctly predicted examples in the test set. The larger ER is, the more robust the classifier is.

### 6.2.2. Results

The results are shown in Table 3 (only the means are shown, the standard errors can be found in Appendix C). We see that BBM-RS performs well in terms of interpretability and robustness. BBM-RS performs the best on nine and eleven out of thirteen datasets in terms of interpretation complexity and robustness, respectively. In terms of accuracy, in nine out of the thirteen datasets, BBM-RS is the best or within $3\%$ to the best. These results show that on most datasets, BBM-RS is better than other algorithms in IC and ER while being comparable in accuracy.

In addition to using the number of feature-thresholds pairs as a (global) measure for IC, we also present in Table 4 the results in terms of the local measure for IC, i.e., the depth. This local measure considerably favors decision trees (DT and RobDT), since in the same depth, DT and RobDT can use *exponentially* more feature-threshold pairs than LCPA and BBM-RS, which can be much less interpretable. From the table, we see that even in this case, BBM-RS can still have a comparable results with DT and RobDT. In Appendix C.4, the standard error of Table 4 is recorded.

## 6.3. Tradeoffs in BBM-RS

The parameter $\tau$ gives us the opportunity to explore the tradeoff between interpretability, robustness, and accuracy within BBM-RS. Figure 1 shows that for small $\tau$, BBM-RS's IC is high, and its ER is low, and when $\tau$ is high, IC is low, and ER is high. This empirical observation strengthens the claim that interpretability and robustness are correlated. See Appendix C for experiments on other datasets and experiments on the tradeoffs between IC and accuracy.

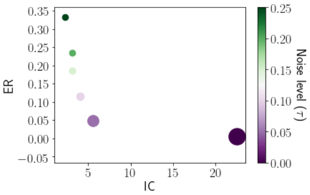


*Figure 1.* Interaction of interpretability, accuracy, and robustness with different noise level $\tau$ on the spambase dataset. The size of each ball represents the accuracy. For $\tau = 0$: $IC = 22.5, ER = 0.006$ and for higher noise $\tau = 0.25$: $IC = 2.3, ER = 0.33$

| | IC (lower=better) | | | | test accuracy (higher=better) | | | | ER (higher=better) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DT | RobDT | LCPA | BBM-RS | DT | RobDT | LCPA | BBM-RS | DT | RobDT | LCPA | BBM-RS |
| adult | 414.20 | 287.90 | 14.90 | **6.00** | **0.83** | **0.83** | 0.82 | 0.81 | **0.50** | **0.50** | 0.12 | **0.50** |
| bank | 30.70 | 26.80 | 8.90 | **8.00** | **0.90** | **0.90** | **0.90** | **0.90** | **0.50** | **0.50** | 0.20 | **0.50** |
| bank2 | 30.00 | 30.70 | 13.80 | **4.50** | **0.91** | 0.90 | 0.90 | 0.90 | 0.12 | 0.18 | 0.10 | **0.50** |
| breastcancer | 15.20 | 7.40 | **6.00** | 11.00 | 0.94 | 0.94 | **0.96** | **0.96** | 0.23 | **0.29** | 0.28 | 0.27 |
| careval | 59.30 | 28.20 | 10.10 | **8.70** | **0.97** | 0.96 | 0.91 | 0.77 | **0.50** | **0.50** | 0.19 | **0.50** |
| compasbin | 67.80 | 33.70 | **5.40** | 7.60 | **0.67** | **0.67** | 0.65 | 0.66 | **0.50** | **0.50** | 0.15 | 0.33 |
| diabetes | 31.20 | 27.90 | 6.00 | **2.10** | 0.74 | 0.73 | **0.76** | 0.65 | 0.08 | 0.08 | 0.09 | **0.15** |
| ficobin | 30.60 | 59.60 | **6.40** | 11.80 | 0.71 | 0.71 | 0.71 | **0.72** | **0.50** | **0.50** | 0.22 | **0.50** |
| heart | 20.30 | 13.60 | 11.90 | **9.50** | 0.76 | 0.79 | **0.82** | **0.82** | 0.23 | 0.31 | 0.14 | **0.32** |
| ionosphere | 11.30 | 8.60 | 17.90 | **6.80** | 0.89 | **0.92** | 0.88 | 0.86 | 0.15 | 0.25 | 0.07 | **0.28** |
| mammo | 27.40 | 12.40 | 7.20 | **1.90** | **0.79** | **0.79** | **0.79** | 0.77 | 0.47 | **0.50** | 0.21 | **0.50** |
| mushroom | 10.80 | **9.10** | 23.80 | 9.90 | **1.00** | **1.00** | **1.00** | 0.97 | **0.50** | **0.50** | 0.10 | **0.50** |
| spambase | 153.90 | 72.30 | 29.50 | **5.60** | **0.92** | 0.87 | 0.88 | 0.79 | 0.00 | 0.04 | 0.02 | **0.05** |

*Table 3.* Comparison of BBM-RS with other interpretable models. In bold: the best algorithm for each dataset and criterion. Note that several datasets (adult, bank, careval, compasbin, ficobin, and mushroom) have ER = 0.5 for tree-based models (DT, RobDT, and BBM-RS), because these datasets have all binary features and tree-based models set the threshold in the middle of 0 and 1.

## 7. Conclusion

We found that linear separability is a hidden property of the data that guarantees both interpretability and robustness. We designed an efficient algorithm, BBM-RS, that returns a model, risk-score, which we prove is interpretable, robust, and have high-accuracy. An interesting open question is whether a weaker notion than linear separability can give similar guarantees.

## Acknowledgements

Kamalika Chaudhuri and Yao-Yuan Yang thank NSF under CIF 1719133 and CNS 1804829 for research support.

| | DT | RobDT | LCPA | BBM-RS |
|---|---|---|---|---|
| adult | 10.00 | 12.50 | 14.90 | **6.00** |
| bank | **5.00** | 6.00 | 8.90 | 8.00 |
| bank2 | 5.00 | 6.00 | 13.80 | **4.50** |
| breastcancer | 6.00 | **5.20** | 6.00 | 11.00 |
| careval | 12.30 | 11.40 | 10.10 | **8.70** |
| compasbin | 7.40 | 7.90 | **5.40** | 7.60 |
| diabetes | 6.00 | 7.50 | 6.00 | **2.10** |
| ficobin | **5.00** | 7.00 | 6.40 | 11.80 |
| heart | **6.00** | 6.10 | 11.90 | 9.50 |
| ionosphere | **6.00** | 7.90 | 17.90 | 6.80 |
| mammo | 5.60 | 6.20 | 7.20 | **1.90** |
| mushroom | **5.80** | 6.00 | 23.80 | 9.90 |
| spambase | 17.40 | 17.60 | 29.50 | **5.60** |

*Table 4.* The IC of four different methods across all datasets. Here, we use the depth of the tree as the interpretable complexity measure for DT and RobDT.

## References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31: 9505–9515, 2018.

Alon, N. and Spencer, J. H. *The probabilistic method.* John Wiley & Sons, 2004.

Andriushchenko, M. and Hein, M. Provably robust boosted decision stumps and trees against adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 13017–13028, 2019.

Ben-David, S., Eiron, N., and Long, P. M. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

Bertsimas, D., Delarue, A., Jaillet, P., and Martin, S. The price of interpretability. *arXiv preprint arXiv:1907.03419*, 2019.

Blanc, G., Lange, J., and Tan, L.-Y. Top-down induction of decision trees: rigorous guarantees and inherent limitations. *arXiv preprint arXiv:1911.07375*, 2019.

Blanc, G., Lange, J., and Tan, L.-Y. Provable guarantees for decision tree induction: the agnostic setting. *arXiv preprint arXiv:2006.00743*, 2020.

Boer, N., Deutch, D., Frost, N., and Milo, T. Personal insights for altering decisions of tree-based ensembles over time. *Proceedings of the VLDB Endowment*, 13(6): 798–811, 2020.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees.* CRC press, 1984.

Brutzkus, A., Daniely, A., and Malach, E. On the optimality of trees generated by id3. *arXiv preprint arXiv:1907.05444*, 2019.

Brutzkus, A., Daniely, A., and Malach, E. Id3 learns juntas for smoothed product distributions. In *Conference on Learning Theory*, pp. 902–915. PMLR, 2020.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Chen, H., Zhang, H., Boning, D., and Hsieh, C.-J. Robust decision trees against adversarial examples. *arXiv preprint arXiv:1902.10660*, 2019.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Deutch, D. and Frost, N. Constraints-based explanations of classifications. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 530–541. IEEE, 2019.

Fiat, A. and Pechyony, D. Decision trees: More theoretical justification for practical algorithms. In *International Conference on Algorithmic Learning Theory*, pp. 156–170. Springer, 2004.

Freund, Y. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Frost, N., Moshkovitz, M., and Rashtchian, C. Exkmc: Expanding explainable $k$-means clustering. *arXiv preprint arXiv:2006.02399*, 2020.

Garreau, D. and von Luxburg, U. Explaining the explainer: A first theoretical analysis of lime. *arXiv preprint arXiv:2001.03447*, 2020a.

Garreau, D. and von Luxburg, U. Looking deeper into lime. *arXiv preprint arXiv:2008.11092*, 2020b.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hu, X., Rudin, C., and Seltzer, M. Optimal sparse decision trees. In *Advances in Neural Information Processing Systems*, pp. 7267–7275, 2019.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. On relating explanations and adversarial examples. In *Advances in Neural Information Processing Systems*, pp. 15883–15893, 2019.

Kanade, V. and Kalai, A. Potential-based agnostic boosting. *Advances in neural information processing systems*, 22: 880–888, 2009.

Kantchelian, A., Tygar, J. D., and Joseph, A. Evasion and hardening of tree ensemble classifiers. In *International Conference on Machine Learning*, pp. 2387–2396, 2016.

Kearns, M. Learning boolean formulae or finite automata is as hard as factoring. *Technical Report TR-14-88 Harvard University Aikem Computation Laboratory*, 1988.

Kearns, M. and Mansour, Y. On the boosting ability of top–down decision tree learning algorithms. *Journal of Computer and System Sciences*, 58(1):109–128, 1999.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

Laber, E. and Murtinho, L. On the price of explainability for some clustering problems. *arXiv preprint arXiv:2101.01576*, 2021.

Lakkaraju, H. and Bastani, O. ”how do i fool you?” manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020.

Lakkaraju, H., Arsov, N., and Bastani, O. Robust and stable black box explanations. In *International Conference on Machine Learning*, pp. 5628–5638. PMLR, 2020.

Li, X.-H., Shi, Y., Li, H., Bai, W., Song, Y., Cao, C. C., and Chen, L. Quantitative evaluations on saliency methods: An experimental study. *arXiv preprint arXiv:2012.15616*, 2020.

Lin, J., Zhong, C., Hu, D., Rudin, C., and Seltzer, M. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pp. 6150–6160. PMLR, 2020.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774, 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mardaoui, D. and Garreau, D. An analysis of lime for text data. *arXiv preprint arXiv:2010.12487*, 2020.

Molnar, C. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Moshkovitz, M., Dasgupta, S., Rashtchian, C., and Frost, N. Explainable k-means and k-medians clustering. In *International Conference on Machine Learning*, pp. 7055–7065. PMLR, 2020.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019.

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

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Ribeiro, M. T., Singh, S., and Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016b.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ross, A. S. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404*, 2017.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Schapire, R. E. and Freund, Y. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shalev-Shwartz, S. and Singer, Y. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *21st Annual Conference on Learning Theory, COLT 2008*, 2008.

Shamir, O. Gradient methods never overfit on separable data. *arXiv preprint arXiv:2007.00028*, 2020.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ustun, B. and Rudin, C. Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1125–1134, 2017.

Ustun, B. and Rudin, C. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pp. 5133–5142. PMLR, 2018.

Yang, Y.-Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Robustness for non-parametric classification: A generic attack and defense. In *International Conference on Artificial Intelligence and Statistics*, pp. 941–951. PMLR, 2020a.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. Adversarial robustness through local lipschitzness. *arXiv preprint arXiv:2003.02460*, 2020b.

# A. Proofs

## A.1. Separation and interpretability

**Theorem 1.** *For any tree size $s$ and $\gamma > 0$, there is a dataset in $\mathbb{R}^2$ that is linearly separable, and any decision tree with size $s$ has accuracy less than $\frac{1}{2} + \gamma$.*

*Proof.* Fix an integer $s$ and $\gamma > 0$. We will start by describing the dataset and prove it is linearly separable. We will then show that any decision tree of size $s$ must have accuracy smaller than $1/2 + \gamma$.

**Dataset.** The dataset size is $n$, to be fixed later. The dataset includes, for any $i = 1 \dots n/4$, a group, $G_i$, of four points: two points $(i, -i + \epsilon), (i + \epsilon, -i)$ that are labeled positive and two points $(i, -i - \epsilon), (i - \epsilon, -i)$ that are labeled negative for some small $\epsilon > 0$, see Figure 2.

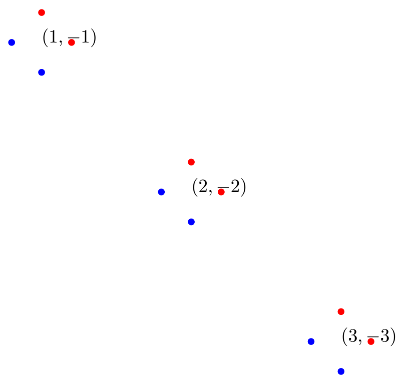


*Figure 2.* Proof of Theorem 1. Linearly separable dataset that is not interpretable by a small decision tree. Around point $(i, -i)$ there are 4 close points: two points $(i, -i + \epsilon), (i + \epsilon, -i)$ that are labeled positive and two points $(i, -i - \epsilon), (i - \epsilon, -i)$ that are labeled negative for some small $\epsilon > 0$.

To prove that the dataset is linearly separable focus on the vector $w = (0.5, 0.5)$ with $|w|_1 = 1$. For each labeled examples $(x, y)$ in the dataset, the inner product is equal to $yw \cdot x = \epsilon/2 > 0$.

**Accuracy.** We will prove by induction that the points arriving in each node are a series of consecutive groups, perhaps except a few points that are in the "tails". In each node, the number of positive and negative examples is about the same, so one cannot predict well. See intuition in Figure 3.

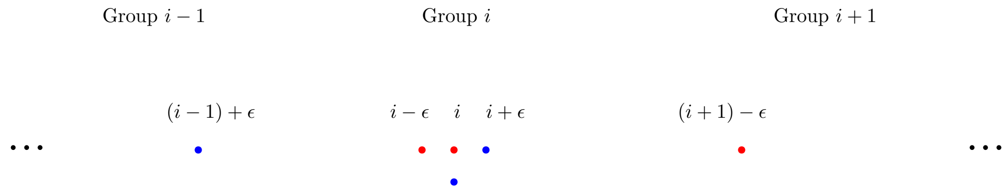


*Figure 3.* Projection of the points in the dataset to the first feature. We see that the groups appear one after another, each cut defined by an inner node in the tree, will leave the order between the groups as is. In a series of consecutive groups, the number of positive and negative examples is equal. Thus, the accuracy is close to $1/2$. Similar figure when projecting to the second feature.

More formally, a tail $G'_i$ is a subset of $G_i$, i.e., $G'_i \in \wp(G_i)$, where $\wp$ is the power set. We prove the following claim.

*Claim.* For each inner node $v$, there are integers $j_0 \leq j_1$ such that the points arriving to this node, $P_v$, satisfy

$$P_v = \cup_{i=j_0}^{j_1} G_i \cup G'_{j_0-1} \cup G'_{j_1+1},$$

where $G'_{j_0-1} \in \wp(G_{j_0-1})$ and $G'_{j_1+1} \in \wp(G_{j_1+1})$. In other words, $P_v$ contains a series of consecutive groups and a few points from the group after and before the series.

*Proof.* We will prove the claim by induction on the level of the tree. The claim is correct for the root with $j_0 = 1$ and $j_1 = n/4$. To prove the claim by induction, suppose an inner node uses a threshold $\theta$ and the first feature. Denote the closest integer to $\theta$ by $j^* \leq \theta < j^* + 1$. The points reaching the left son are

$$\cup_{i=j_0}^{j^*} G_i \cup G'_{j_0-1} \cup G'_{j^*+1},$$

and the nodes reaching the right son are

$$\cup_{i=j^*}^{j_1} G_i \cup G'_{j^*-1} \cup G'_{j_1+1}.$$

A similar argument also hold if $v$ uses the second feature, which proves our claim. □

To finish the proof of Theorem 1, first note that the number of positive and negative examples in each $G_i$ is exactly equal. Together with the claim that we just proved, for each leaf $v$, the number of points that are correctly classified out of the points, $P_v$, reaching $v$, is at most $|P_v|/2 + 4$. Thus the total number of points correctly classified by the entire tree is at most $n/2 + 4s$. Thus, the accuracy of the tree is at most $1/2 + 4s/n$. We take $n > 4s/\gamma$ and get that the accuracy is smaller than $1/2 + \gamma$, which is what we wanted to prove.

□

**Theorem 2.** *For any labeled data in $[-1, 1]^d \times \{-1, 1\}$ that is $r$-separated, there is a decision tree of depth at most $\frac{6d}{r}$ which has a training error $0$.*

*Proof.* Each feature is in $[-1, 1]$, so $r \leq 1$. Fix $r \leq \Delta < 2r$ and $L = \lceil \frac{1}{\Delta} \rceil$. We can bound $L$ by

$$L \leq \frac{1}{\Delta} + 1 \leq \frac{1}{r} + 1 \leq \frac{2}{r}.$$

Take two data points, one labeled positive, $x^+$, and one negative $x^-$. By the $r$-separation, we know that there is a feature $i'$ such that

$$|x^+_{i'} - x^-_{i'}| \geq 2r > \Delta.$$

This means that we can find a threshold $\theta$ among the $2L + 1$ thresholds $-L \cdot \Delta, \dots, 0 \cdot \Delta, 1 \cdot \Delta, \dots, L \cdot \Delta$ that distinguishes the examples $x^+$ and $x^-$, i.e., there is $j' \in \{-L, \dots, L\}$, such that

$$sign(x^+_{i'} - \Delta \cdot j') \neq sign(x^-_{i'} - \Delta \cdot j').$$

We focus on the decision tree with all possible features and the $2L + 1$ thresholds. All examples reaching the same leaf, has the same label. In other words, the training error is 0. Since there are $d \cdot (2L + 1)$ pairs of feature and thresholds, the depth of the tree is at most $d(2L + 1) \leq 3dL \leq \frac{6d}{r}$.

□

**Theorem 3.** *There is a labeled dataset in $[-1, 1]^d$ which is $1$-separated and has the following property. For any $\gamma > 0$ and any decision tree $T$ that achieves accuracy $0.5 + \gamma$, the size of $T$ is at least $\gamma 2^d$.*

*Proof.* We start with describing the dataset, then we will show that any decision tree must be large if we want to achieve $0.5 + \gamma$ accuracy.

**The dataset.** The inputs are all strings in $\{-1, 1\}^d$. The hypothesis is the parity function, i.e., and

$$f(x_1, \dots, x_d) = \begin{cases} 1 & \text{if } \sum_{i=1}^{d} \frac{x_i+1}{2} \equiv 1 \pmod 2 \\ -1 & \text{o.w.} \end{cases}$$

**Each node has equal number of positive and negative examples.** The main idea is that as long as the depth of a node is not $d$, it has exactly the same number of positive and negative examples reaching it. This is true since as long as at most $d - 1$ features are fixed, there exactly the same number of positive and negative examples that agree on these features.

**Large size.** Denote by $N_1$ the set of all nodes that exactly one example reach them. Denote this set size by $|N_1| = n_1$. We want to prove that $n_1$ is large. So far, we proved that for each node that contains more than one example, exactly half of the examples are labeled positive and half negative. Number of examples correctly classified is half of all the examples not in $N_1$ plus $n_1$. There are $2^d$ examples in total. So the accuracy is equal to $\frac{(2^d-n_1)/2+n_1}{2^d} = \frac{1}{2} + \frac{n_1}{2^{d+1}}$. The latter should be at least $1/2 + \gamma$. Therefore, the size of the tree is at least $\gamma 2^d$.

□

### A.2. Linear separability: weak learner

We start with the more restricted case where the features are binary. This will give the necessary foundations for the general case, where features are in $[-1,1]$. In this case $\mathcal{H}_t$ is simplified to the set $\{x \mapsto x_i : i \in [d]\}$.

**Theorem 8.** *For any data in $\{-1,1\}^d \times \{-1,1\}$ that is labeled by a $\gamma$-linearly separable hypothesis $f$ and for any distribution $\mu$ on the examples, there is hypothesis $h \in \mathcal{H}_t$ such that*

$$\Pr_{x \sim \mu}(h(x) = f(x)) \geq \frac{1}{2} + \frac{\gamma}{2}.$$

*Proof.* The proof's high-level idea is to represent the class as a bipartite graph and lower bound the weighted density of this graph. The high lower-bound leads to a high degree vertex, which will correspond to our desired weak learner.

Recall that by the fact that the data is $\gamma$-linearly separable, we know that there is a vector $w$ with $|w|_1 = 1$ such that for eac labeled example $(x, y)$ it holds that

$$yw \cdot x \geq \gamma. \tag{1}$$

**Bipartite graph description.** Consider the following bipartite graph. The vertices are the $m$ examples in the training data and the $d$ hypotheses in the binary version of $\mathcal{H}_t$. There is an edge $(x, h_i) \in E$ between an example $x$ and hypothesis $h_i$, if it correctly classify $x$, i.e., if $f(x) = h_i(x)$. Each vertex is given a weight: example $x^j$ gets weight $\mu_j$ and a hypothesis $h_i$ gets weight $w_i$, where $w$ is in Equation (1).

**Assumption: $\sum_{i=1}^d w_i = 1$.** We assume without loss of generality that $w_i \geq 0$, otherwise if there is $w_i < 0$, we can multiply the $i$-th feature in all the examples by $-1$. After this multiplication, the linearity assumption still holds. We know that $\sum_{i=1}^d |w_i| = 1$, and since $w_i \geq 0$, we get that $\sum_{i=1}^d w_i = 1$.

**Lower bound weighted-density.** The main idea is to lower bound the weighted density $\rho$ of the bipartite graph, which is the sum over all edges $(h_i, x^j)$ in the graph, each with weight $w_i, \mu_j$:

$$\rho = \sum_{(x^j, h_i) \in E} w_i \mu_j.$$

To prove a lower bound on $\rho$, we focus on one labeled example $(x^j, y^j)$. From the linearity assumption, Equation (1), we know that $\sum_{i=1}^d y^j w_i x_i^j \geq \gamma$. Recall that $y^j x_i^j$ is equal to $+1$ or $-1$, since we are in the binary case. We can separate the sum depending on whether $y^j x_i^j$ is equal to $+1$ or $-1$ and get that

$$\sum_{i: y^j x_i^j = 1} w_i - \sum_{i: y^j x_i^j = -1} w_i \geq \gamma.$$

We know that $\sum_{i=1}^d w_i = 1$, thus $\sum_{i: y^j x_i^j = -1} w_i = 1 - \sum_{i: y^j x_i^j = 1} w_i$, and the inequality can be rewritten as

$$2 \sum_{i: y^j x_i^j = 1} w_i - 1 \geq \gamma.$$

Notice that $(x^j, h_i) \in E \Leftrightarrow y^j x_i^j = 1$. Thus, the inequality can be further rewritten as

$$\sum_{i: (x^j, h_i) \in E} w_i \geq \frac{1+\gamma}{2}.$$

This inequality holds for any labeled example, so we can sum all these inequalities, each with weight $\mu_j$ and get that

$$\rho \geq (1+\gamma)/2.$$

**Finding a weak learner.** Since $w$ is a probability distribution, we can rewrite $\rho$ and get

$$\mathbb{E}_{i \sim w}\left[\sum_{j:(x^j,h_i)\in E} \mu_j\right] \geq (1+\gamma)/2.$$

From the probabilistic method (Alon & Spencer, 2004), there is a hypothesis $h_i$ such that

$$\sum_{j:(x^j,h_i)\in E} \mu_j \geq (1+\gamma)/2.$$

By the definition of the graph, $(x^j, h_i) \in E \Leftrightarrow h_i(x^j) = f(x^j)$. Thus, we get that

$$\Pr_{x\sim\mu}(h_i(x) = f(x)) \geq \frac{1}{2} + \frac{\gamma}{2},$$

which is exactly what we wanted to prove.

□

**Theorem 9** (binary weak-learner)**.** *For any distribution $\mu$ over labeled examples $\{-1,+1\}^d \times \{-1,+1\}$ that satisfies linear separability with a $\gamma$-margin, and for any $\delta \in (0,1)$ there is $m = O\left(\frac{d+\log\frac{1}{\delta}}{\gamma^2}\right)$, such that with probability at least $1-\delta$ over the sample $S$ of size $m$, it holds that*

$$\Pr_{(x,y)\sim\mu}(h_S(x) = y) \geq \frac{1}{2} + \frac{\gamma}{4}.$$

*Proof.* We start with a known fact[2] that

$$VC(\mathcal{H}_t) \leq d.$$

Denote the best hypothesis in the binary version in $\mathcal{H}_t$ as $h^*$. From Theorem 8, we know that $\Pr_x(h^*(x) = f(x)) \geq \frac{1}{2} + \frac{\gamma}{2}$. For every sample $S$, denote by $h_S$ the hypothesis in the binary version of $\mathcal{H}_t$ that optimizes the accuracy on the sample $S$. From the fundamental theorem of statistical learning (Shalev-Shwartz & Ben-David, 2014), we know that for $m = O\left(\frac{d+\log\frac{1}{\delta}}{(\gamma/4)^2}\right)$, with probability at least $1-\delta$ over the sample $S$ of size $m$, it holds that

$$\Pr_{(x,y)\sim\mu}(h_S(x) = y) \geq \Pr_{(x,y)\sim\mu}(h^*(x) = y) - \frac{\gamma}{4} \geq \frac{1}{2} + \frac{\gamma}{4}.$$

□

**Theorem 4.** *Fix $\alpha > 0$. For any data in $[-1,1]^d \times \{-1,1\}$ that is labeled by a $\gamma$-linearly separable hypothesis $f$ and for any distribution $\mu$ on the examples, there is a hypothesis $h \in \mathcal{H}_t$ such that*

$$\Pr_{x\sim\mu}(h(x) = f(x)) \geq \frac{1}{2} + \frac{\gamma}{2} - \alpha.$$

*Proof.* The proof is similar in spirit to the proof of Theorem 8. The main difference is the technique to lower bound $\rho$, the weighted density. In the current proof we use the fact that if for some positive example $x$, its $i$-th feature, $x_i$, is high, then it contains many edges in the graph. For a negative example $x$, if its $i$-th feature, $x_i$, is low, then it contains many edges in the graph.

[2] To bound the $VC(\mathcal{H}_t)$ note that for any $d+1$ points in $\mathbb{R}^d$ there is at least one point $v$, where in each coordinate it is not the largest one among the $d+1$ points. Thus, it's impossible that $v$ is labeled $+1$ while all the rest of the points are labeled $-1$.

**Bipartite graph description.** We first discretize the segment $[-1,1]$ to $\ell \geq 2/\alpha+1$ values with $\Delta = 2/(\ell-1)$:

$$Z = \{-1, -1+\Delta, \ldots, 1-\Delta, 1\}.$$

The value of $\Delta$ was chosen such that $|Z| = \ell$. We focus on the following subclass of $\mathcal{H}'_t$ which is a discretization of $\mathcal{H}_t$

$$\mathcal{H}'_t = \{h_{i,\theta}\} \quad \text{for } i \in [d], \theta \in Z.$$

We use a similar bipartite graph as in the proof of Theorem 8 for the subclass $\mathcal{H}'_t$: the vertices are the $m$ examples and the hypotheses in $\mathcal{H}'_t$; and there is an edge $(x, h_{i,\theta}) \in E$ between an example $x$ with label $y$ and hypothesis $h_{i,\theta}$ whenever $yx_i \geq \theta$, i.e., when $h_{i,\theta}$ correctly classify $x$.

Recall that by the fact that the data is $\gamma$-linearly separable, we know that there is a vector $w$ with $|w|_1 = 1$ and for any labeled example $(x,y)$ it holds that

$$yw \cdot x \geq \gamma. \tag{2}$$

From the same argument as in Theorem 8, we assume that $\sum_{i=1}^{d} w_i = 1$.

We are now ready to give each vertex in the bipartite graph a weight: example $x^j$ gets weight $\mu_j$ and a hypothesis $h_{i,\theta}$ gets weight $w_{i,\theta} = w_i/\ell$, where $w$ is as in Equation (2). The weights on the hypotheses were chosen such that they will sum up to 1:

$$\sum_{i,\theta} w_{i,\theta} = \sum_{i,\theta} \frac{w_i}{\ell} = \sum_i w_i = 1. \tag{3}$$

**Lower bound weighted-density.** We will lower bound the weighted density $\rho$ of the bipartite graph, which is the sum over all edges $(x^j, h_{i,\theta})$ in the graph, each with weight $w_{i,\theta}\mu_j$:

$$\rho = \sum_{(x^j, h_{i,\theta}) \in E} w_{i,\theta}\mu_j.$$

To show a lower bound on $\rho$, we focus on one labeled example $(x,y)$ and one feature $i$ and we will lower bound the following sum

$$\sum_{\theta:(x,h_{i,\theta})\in E} w_{i,\theta} = \frac{w_i}{\ell} \sum_{\theta:(x,h_{i,\theta})\in E} 1. \tag{4}$$

The sum in the RHS is equal to the number of $\theta$'s such that $yx_i^j \geq \theta$:

$$d_{x,i} = |\{\theta \in Z : yx_i \geq \theta\}|. \tag{5}$$

To lower bound $d_{x,i}$ we separate the analysis depending on the label $y$. If $x$ is a positive example, i.e., $y = +1$, we have that

$$d_{x,i} \geq (1 + x_i - \Delta)/\Delta$$

For a negative example $x$, we can lower bound $d_{x,i}$ by

$$d_{x,i} \geq (1 - x_i - \Delta)/\Delta$$

To summarize these two equations we get that

$$d_{x,i} \geq (1 + yx_i - \Delta)/\Delta$$

Going back to Equation (4),

$$\sum_{\theta:(x,h_{i,\theta})\in E} w_{i,\theta} \geq \frac{w_i}{\ell} \cdot \frac{1+yx_i-\Delta}{\Delta} = w_i(1+yx_i-\Delta)\cdot\frac{\ell-1}{2\ell}.$$

Summing over all features

$$\begin{aligned} \sum_{i,\theta:(x,h_{i,\theta})\in E} w_{i,\theta} &= \frac{\ell-1}{2\ell}\left(\sum_i w_i(1-\Delta) + \sum_i yx_i w_i\right) && (6)\\ &\geq \frac{\ell-1}{2\ell}(1+\gamma-\Delta) = \left(\frac{1}{2}+\frac{\gamma}{2}-\frac{\Delta}{2}\right)\left(1-\frac{1}{\ell}\right) && (7) \end{aligned}$$

Weighted sum over all examples yields the desired lower bound on $\rho$

$$\rho \geq \left(\frac{1}{2} + \frac{\gamma}{2} - \frac{\Delta}{2}\right)\left(1 - \frac{1}{\ell}\right) \geq \frac{1}{2} + \frac{\gamma}{2} - \alpha.$$

**Finding a weak learner.** The proof now continues similarly to the proof of Theorem 8. Since $w_{i,\theta}$ is a probability distribution over all hypotheses in $\mathcal{H}'_t$ (see Equation 3), we can rewrite $\rho$ and get

$$\mathbb{E}_{h_{i,\theta}\sim w}\left[\sum_{j:(x^j,h_{i,\theta})\in E} \mu_j\right] \geq \frac{1}{2} + \frac{\gamma}{2} - \alpha.$$

From the probabilistic method (Alon & Spencer, 2004), there is hypothesis $h_i$ such that

$$\sum_{j:(x^j,h_i)\in E} \mu_j \geq \frac{1}{2} + \frac{\gamma}{2} - \alpha.$$

By the definition of the graph, $(x^j, h_i) \in E \Leftrightarrow h_i(x^j) = f(x^j)$. Thus, we get that

$$\Pr_{x\sim\mu}\left(h_i(x) = f(x)\right) \geq \frac{1}{2} + \frac{\gamma}{2} - \alpha,$$

which is exactly what we wanted to prove.

□

**Theorem 5** (weak-learner)**.** *Fix $\alpha > 0$. For any distribution $\mu$ over $[-1,+1]^d \times \{-1,+1\}$ that satisfies linear separability with a $\gamma$-margin, and for any $\delta \in (0,1)$ there is $m = O\left(\frac{d+\log\frac{1}{\delta}}{\gamma^2}\right)$, such that with probability at least $1-\delta$ over the sample $S$ of size $m$, it holds that*

$$\Pr_{(x,y)\sim\mu}(h_S(x) = y) \geq \frac{1}{2} + \frac{\gamma}{4} - \alpha.$$

*Proof.* Proof is similar to Theorem 9. □

### A.3. Linear separability: risk scores

**Claim 6.** *If every condition in a risk-score model $R$ is of the form "$x_i \geq \theta$" and all weights are positive, except the bias term, then $R$ is a monotone model.*

*Proof.* Fix a risk score model $f$ which is defined by a series of $m$ conditions "$x_{i_1} \geq \theta_1$", …, "$x_{i_m} \geq \theta_m$" and weights $w_0, w_1, \ldots, w_m$. The score, $s(z)$, of an example $z$ is the weighted-sum of all satisfied conditions,

$$s(z) = w_0 + \sum_{j=1}^{m} w_j I_{z_j \geq \theta_j},$$

where $I_A = 1$ if $A$ is true, and $I_A = 0$ otherwise. The prediction of the model $f$ is equal to

$$f(z) = sign(s(z)).$$

Fix examples $x, y \in \mathbb{R}^d$ with $x \leq y$. Our goal is to show that $f(x) \leq f(y)$. The key observation is that any condition $x_{i_j} \geq \theta_j$ satisfied by $x$ is also satisfied by $y$ because $y_{i_j} \geq x_{i_j} \geq \theta_j$, by our assumption that $x \leq y$. In different words we have that

$$I_{y_j \geq \theta_j} \geq I_{x_j \geq \theta_j}. \tag{8}$$

This implies that the score of $y$ is at least the score of $x$, since it holds that

$$s(y) = w_0 \sum_{j=1}^{m} w_j I_{y_j \geq \theta_j} \geq w_0 + \sum_{j=1}^{m} w_j I_{x_j \geq \theta_j} = s(x).$$

Thus, we get exactly what we wanted to prove $f(y) = sign(s(y)) \geq sign(s(x)) = f(x)$.

As an aside, at this point, it should become apparent why we restricted our conditions to be of the form "$x_{i_j} \geq \theta_j$" and did not allow natural conditions of the form "$x_{i_j} \leq \theta_j$", such conditions will not be monotone and Inequality (8) will not hold. □

**Claim 7.** *Assume a learning algorithm $A$ gets as an input a sample from a $\gamma$-linearly separable data and returns a* monotone *model with accuracy $1 - \epsilon(\gamma)$. Then, there is an algorithm that returns a model with astuteness at least $1 - \epsilon\left(\frac{\gamma}{2}\right)$ at radius $\gamma/2$.*

*Proof.* The high-level idea is to focus on a noisier distribution than the original one. The noise is small enough so that the new data will remain linearly separable with a (slightly worse) margin. Therefore, the learning algorithm can be applied. We call the learning algorithm with a sample from the noisy distribution and return its result. We will prove that a point correctly classified in the noisy dataset implies that the noiseless point is robust to adversarial examples.

**Noisy data.** Fix a data $D \subseteq \mathcal{X} \times \{-1, +1\}$ and a distribution $\mu$ on $D$. We will create a noisy distribution $\mu'$ on the labeled examples by mapping each example $(x, y) \in D$ to a new labeled example $(x', y)$ where

$$x' = x - y\frac{\gamma}{2}\mathbf{1},$$

where $\mathbf{1}$ is the vector of all 1. Thus, if $x$ is a positive labeled example ($y = +1$) then $x' = x - \frac{\gamma}{2}\mathbf{1}$. This means that from each coordinate we decrease $\gamma/2$. Intuitively, we make $x$ looks more negative. Similarly, if $x$ is a negative labeled example ($y = -1$) we create a new example $x' = x + \frac{\gamma}{2}\mathbf{1}$, intuitively, making $x$ looks more positive by adding $\gamma/2$ to each coordinate.

If we have a samples from $\mu$, then we can easily sample from $\mu'$ by subtracting $y\frac{\gamma}{2}\mathbf{1}$ from each labeled example $(x, y)$ in the sample.

**Noisy data is $\gamma/2$-linearly separable.** Suppose that the original data $D$ is $\gamma$-separable. This means that there is a vector $w$ with $|w|_1 = 1$ such that for every labeled example $(x, y) \in D$ in it holds that $yw \cdot x \geq \gamma$. We know that the corresponding example in the noisy data is equal to $x' = x - y\frac{\gamma}{2}\mathbf{1}$. We will prove a lower bound on $yw \cdot x'$ and this will prove that the noisy input is also linearly separable with a margin. We will use the fact that $y^2 = 1$ for any label $y$ and get that

$$yw \cdot x' = yw\left(x - y\frac{\gamma}{2}\mathbf{1}\right) = ywx - y^2 w \cdot \frac{\gamma}{2}\mathbf{1} = ywx - w \cdot \frac{\gamma}{2}\mathbf{1}$$

Recall that $|w|_1 = 1$, which means that $w \cdot \mathbf{1} = \sum_i w_i \leq \sum_i |w_i| = |w|_1 = 1$. This means that $-w \cdot \frac{\gamma}{2}\mathbf{1} \geq -\gamma/2$. Together with the fact that $yw \cdot x \geq \gamma$, we can now give the desired lower bounds on $yw \cdot x'$

$$yw \cdot x' \geq ywx - \gamma/2 \geq \gamma - \gamma/2 = \gamma/2.$$

In different words, the new data is $\gamma/2$-linearly separable.

**Model is robust.** In order to construct a robust model, we take our sample $S$ from $\mu$. Then we transform it to a sample from $\mu'$ by subtracting noise of $y\frac{\gamma}{2}\mathbf{1}$ to each labeled example $(x, y)$, and then we call algorithm $A$ with the noisy training data. From our assumption, the resulting model has accuracy $1 - \epsilon(\frac{\gamma}{2})$. We will show that the returned model has astuteness of $1 - \epsilon(\frac{\gamma}{2})$ at radius $\gamma/2$ with respect to $\mu$. We do this by proving that if $(x', y)$ is correctly classified than the model is robust at $x$ with radius $\gamma/2$.

Fix a noisy labeled example $(x', y)$ and an example $z$ that is $\gamma/2$ close to $x$, in $\ell_\infty$, i.e., $\|x - z\|_\infty \leq \gamma/2$. If $x$ is positively labeled then $x'$ is also positively labeled, and from the construction of the noisy dataset it holds that

$$x' \leq z.$$

Hence, from the monotone property of the model, if $x'$ is labeled correctly then so is $z$. A similar argument holds if $x$ is negatively labeled.

□

## B. Additional Experiment Details

### B.1. Setups

The experiments are performed on a Intel Core i9 9940X machine with 128GB of RAM. The code for the experiments is available at `https://github.com/yangarbiter/interpretable-robust-trees`.

**Additional dataset details.** The adult, bank, breastcancer, mammo, mushroom, and spambase datasets are retrieved from a publicly available repository [3], and these datasets are used by Ustun & Rudin (2019). The careval, ficobin, and campasbin datasets are also retrieved from a publicly available source[4] and used by Lin et al. (2020). We also added the diabetes, heart, and ionosphere dataset from [5] and bank2 dataset from Moro et al. (2014). All features are scaled to $[0, 1]$ by the following formula $(x–min)/(max–min)$, where $x$ represents the feature value, and $min$ and $max$ represents the minimum and maximum value of the given feature across the entire data.

In the adult dataset, the target is to predict whether the person's income is greater then $50,000$. For bank and bank2 datasets, we want to predict whether the client opens a bank account after a marketing call. In the breastcancer dataset, we want to predict whether the given sample is benign. In the heart dataset, we want to detect the presence of heart disease in a patient. In the mammo dataset, we want to predict whether the sample from the mammography is malignant. In the mushroom dataset, we want to predict whether the mushroom is poisonous. In the spambase dataset, we want to predict whether an email is a spam. In the careval dataset, the goal is to evaluate cars. In the ficobin dataset, we want to predict a person credit risk. In the campasbin dataset, we want to predict whether a convicted criminal will re-offend again. In the diabetes dataset, we want to predict whether or not the patients in the dataset have diabetes. In the ionosphere dataset, we want to predict whether the radar data are showing some evidence of some type of structure in the ionosphere.

**Baseline implementations** For DT, we use the implementation from `scikit-learn` (Pedregosa et al., 2011) and set the splitting criteria to entropy. For LCPA and RobDT, we use the implementation from their original authors[6].

**Measure empirical robustness.** The IR for DT and RobDT can be measured using the method in (Kantchelian et al., 2016; Yang et al., 2020a). The IR for BBM-RS is measured using the method in (Andriushchenko & Hein, 2019). The IR for LCPA can be measured by solving a linear program.

## C. Additional Results

### C.1. Examples that are similar but labeled differently

The compasbin dataset has the lowest $r$-separateness. Its binary features are:

- sex:Female
- age:$< 21$
- age:$< 23$
- age:$< 26$
- age:$< 46$
- juvenile-felonies:=0
- juvenile-misdemeanors:=0
- juvenile-crimes:=0
- priors:=0
- priors:=1
- priors:2-3
- priors:>3

[3] `https://github.com/ustunb/risk-slim/tree/master/examples/data`

[4] `https://github.com/Jimmy-Lin/GeneralizedOptimalSparseDecisionTrees/tree/master/experiments/datasets`

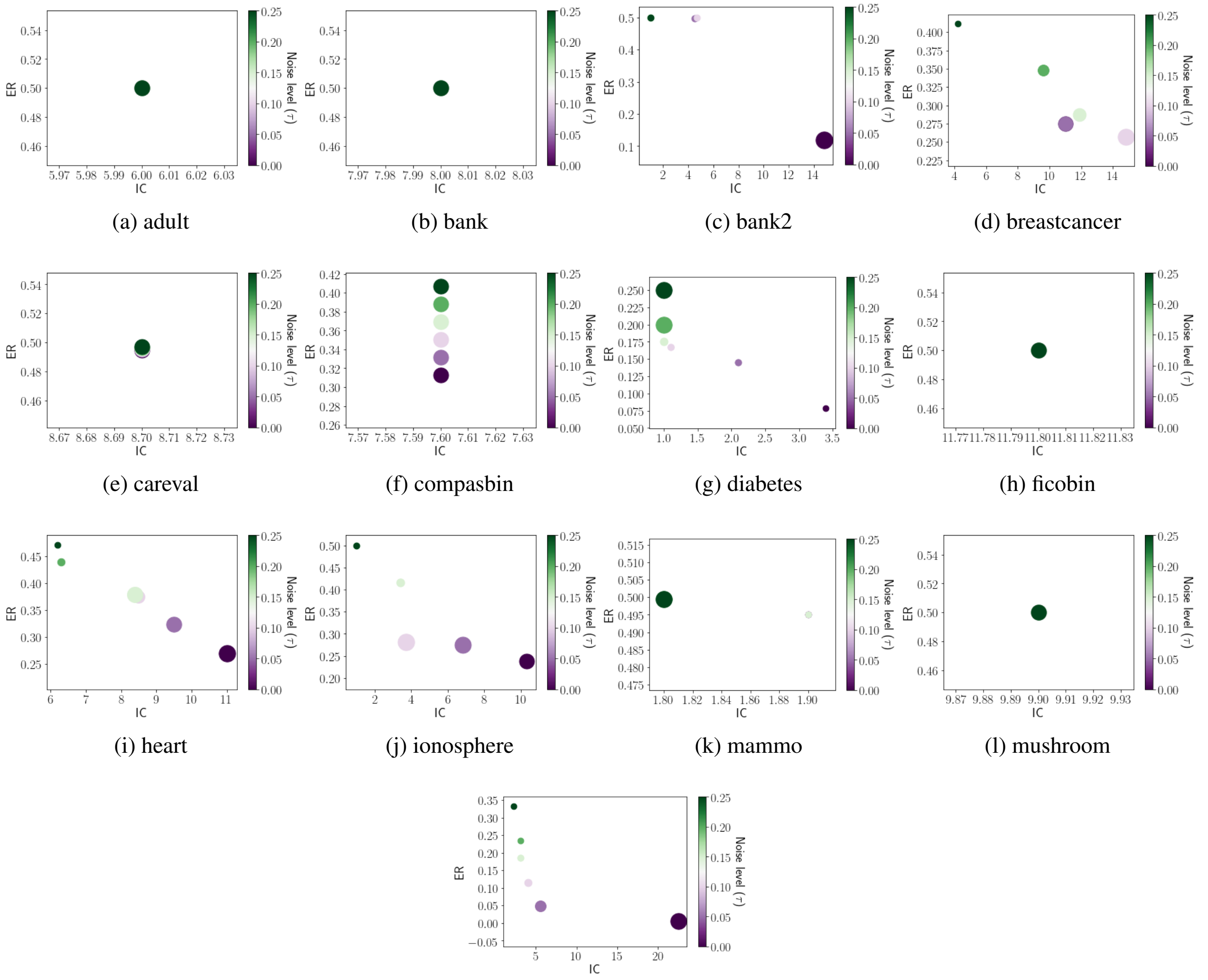
[5] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

[6] `https://github.com/ustunb/risk-slim` and `https://github.com/chenhongge/RobustTrees`

There are 852 people who are: male, age between 26 to 46, did not commit any juvenile felonies, misdemeanors, and crimes, and have more than 3 previous criminal conviction. These people will have the same feature vector while for their labels, 542 recidivate within two years while 310 people did not.

### C.2. Relationship between explainability, accuracy, and robustness in BBM-RS

To understand the interaction between explainability, accuracy, and robustness, we measure these criteria of BBM-RS with different noise levels $\tau$. The results are shown in Figure 4. We observed that, by changing the noise level, that robustness and the explanation complexity go hand in hand. For higher noise levels, we have a higher robustness and lower explanation complexity and accuracy. The shows that by making the model simpler, we can have better robustness and explainability while loosing some accuracy.



(a) adult (b) bank (c) bank2 (d) breastcancer

(e) careval (f) compasbin (g) diabetes (h) ficobin

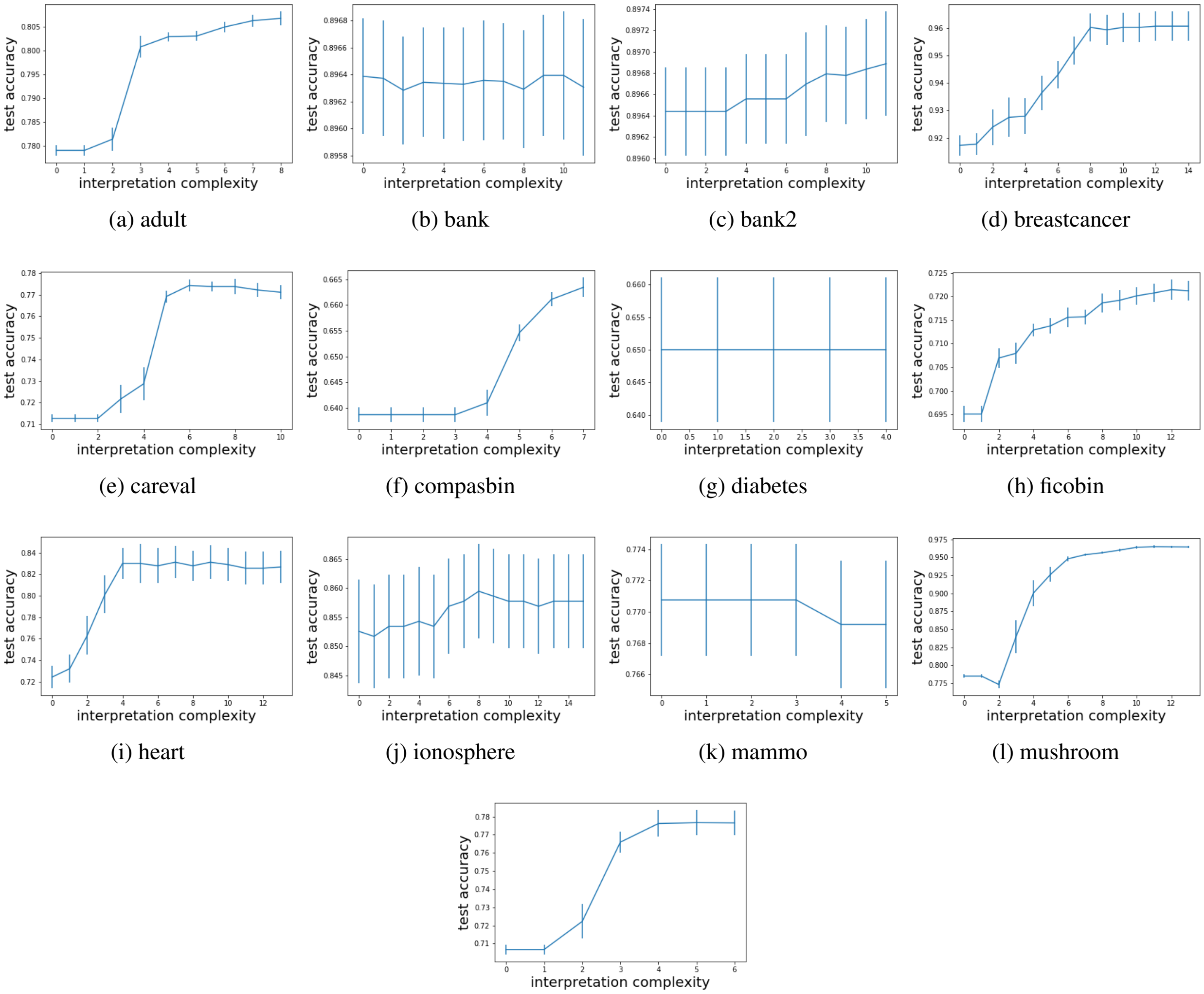
(i) heart (j) ionosphere (k) mammo (l) mushroom

(m) spambase

*Figure 4.* Trade-off between explainability and accuracy for BBM-RS. The size of the ball represents the accuracy.

### C.3. Trade-off between explanation complexity and accuracy for BBM-RS

To understand how explanation complexity effects accuracy, we first train a BBM-RS classifier. The learned BBM-RS classifier consists of $T$ weak learners. We then measure the test accuracy of using only $i$ weak learners for prediction, where $i = 1 \dots T$. Finally, we plot out the figure of accuracy versus explanation complexity (number of unique weak learners) in Table 5. Note that for the same explanation complexity, there may be more then one test accuracy. In this case we show the highest test accuracy. In Table 5, we see that with the increase of explanation complexity, generally the test accuracy also increases.



(a) adult

(b) bank

(c) bank2

(d) breastcancer

(e) careval

(f) compasbin

(g) diabetes

(h) ficobin

(i) heart

(j) ionosphere

(k) mammo

(l) mushroom

(m) spambase

*Figure 5.* Trade-off between explanation complexity and test accuracy for BBM-RS.

### C.4. Local interpretable complexity

| | EC | | | |
|---|---|---|---|---|
| | DT | RobDT | Rudin's | BBM |
| adult | 414.20 ± 5.66 | 287.90 ± 35.66 | 14.90 ± 1.46 | **6.00 ± .60** |
| bank | 30.70 ± .15 | 26.80 ± .20 | 8.90 ± .66 | **8.00 ± 1.41** |
| bank2 | 30.00 ± .30 | 30.70 ± .15 | 13.80 ± 1.54 | **4.50 ± 1.34** |
| breastcancer | 15.20 ± 1.25 | 7.40 ± .60 | **6.00 ± .00** | 11.00 ± .89 |
| careval | 59.30 ± 2.22 | 28.20 ± .65 | 10.10 ± .97 | **8.70 ± .47** |
| compasbin | 67.80 ± 13.01 | 33.70 ± 3.05 | **5.40 ± .22** | 7.60 ± .16 |
| diabetes | 31.20 ± 6.96 | 27.90 ± 2.95 | 6.00 ± .00 | **2.10 ± .53** |
| ficobin | 30.60 ± .22 | 59.60 ± 29.82 | **6.40 ± .16** | 11.80 ± .65 |
| heart | 20.30 ± 1.60 | 13.60 ± .88 | 11.90 ± 1.46 | **9.50 ± .82** |
| ionosphere | 11.30 ± .98 | 8.60 ± .76 | 17.90 ± 3.14 | **6.80 ± 1.96** |
| mammo | 27.40 ± 5.09 | 12.40 ± .65 | 7.20 ± .65 | **1.90 ± .60** |
| mushroom | 10.80 ± .25 | **9.10 ± .10** | 23.80 ± 1.50 | 9.90 ± .89 |
| spambase | 153.90 ± 8.51 | 72.30 ± 2.89 | 29.50 ± .76 | **5.60 ± .48** |
| | test accuracy | | | |
| adult | **0.83 ± .00** | **0.83 ± .00** | 0.82 ± .00 | 0.81 ± .00 |
| bank | **0.90 ± .00** | **0.90 ± .00** | **0.90 ± .00** | **0.90 ± .00** |
| bank2 | **0.91 ± .00** | 0.90 ± .00 | 0.90 ± .00 | 0.90 ± .00 |
| breastcancer | 0.94 ± .00 | 0.94 ± .01 | **0.96 ± .00** | **0.96 ± .01** |
| careval | **0.97 ± .00** | 0.96 ± .00 | 0.91 ± .01 | 0.77 ± .00 |
| compasbin | **0.67 ± .00** | **0.67 ± .00** | 0.65 ± .00 | 0.66 ± .00 |
| diabetes | 0.74 ± .01 | 0.73 ± .01 | **0.76 ± .01** | 0.65 ± .01 |
| ficobin | 0.71 ± .00 | 0.71 ± .00 | 0.71 ± .00 | **0.72 ± .00** |
| heart | 0.76 ± .01 | 0.79 ± .01 | **0.82 ± .01** | **0.82 ± .01** |
| ionosphere | 0.89 ± .01 | **0.92 ± .01** | 0.88 ± .01 | 0.86 ± .01 |
| mammo | **0.79 ± .00** | **0.79 ± .00** | **0.79 ± .00** | 0.77 ± .00 |
| mushroom | **1.00 ± .00** | **1.00 ± .00** | **1.00 ± .00** | 0.97 ± .00 |
| spambase | **0.92 ± .00** | 0.87 ± .00 | 0.88 ± .00 | 0.79 ± .01 |
| | ER | | | |
| adult | **0.50 ± .00** | **0.50 ± .00** | 0.12 ± .02 | **0.50 ± .00** |
| bank | **0.50 ± .00** | **0.50 ± .00** | 0.20 ± .03 | **0.50 ± .00** |
| bank2 | 0.12 ± .01 | 0.18 ± .02 | 0.10 ± .01 | **0.50 ± .00** |
| breastcancer | 0.23 ± .01 | **0.29 ± .01** | 0.28 ± .00 | 0.27 ± .01 |
| careval | **0.50 ± .00** | **0.50 ± .00** | 0.19 ± .02 | **0.50 ± .00** |
| compasbin | **0.50 ± .00** | **0.50 ± .00** | 0.15 ± .01 | 0.33 ± .01 |
| diabetes | 0.08 ± .01 | 0.08 ± .00 | 0.09 ± .00 | **0.15 ± .05** |
| ficobin | **0.50 ± .00** | **0.50 ± .00** | 0.22 ± .01 | **0.50 ± .00** |
| heart | 0.23 ± .02 | 0.31 ± .02 | 0.14 ± .01 | **0.32 ± .02** |
| ionosphere | 0.15 ± .01 | 0.25 ± .01 | 0.07 ± .01 | **0.28 ± .01** |
| mammo | 0.47 ± .01 | **0.50 ± .00** | 0.21 ± .02 | **0.50 ± .00** |
| mushroom | **0.50 ± .00** | **0.50 ± .00** | 0.10 ± .01 | **0.50 ± .00** |
| spambase | 0.00 ± .00 | 0.04 ± .00 | 0.02 ± .00 | **0.05 ± .00** |

*Table 5.* The comparison of BBM-RS with other interpretable models (with standard error).

| | DT | RobDT | Rudin's | BBM |
|---|---|---|---|---|
| adult | 10.00 ± .00 | 12.50 ± 1.07 | 14.90 ± 1.46 | **6.00 ± .60** |
| bank | **5.00 ± .00** | 6.00 ± .00 | 8.90 ± .66 | 8.00 ± 1.41 |
| bank2 | 5.00 ± .00 | 6.00 ± .00 | 13.80 ± 1.54 | **4.50 ± 1.34** |
| breastcancer | 6.00 ± .54 | **5.20 ± .20** | 6.00 ± .00 | 11.00 ± .89 |
| careval | 12.30 ± .54 | 11.40 ± .16 | 10.10 ± .97 | **8.70 ± .47** |
| compasbin | 7.40 ± .81 | 7.90 ± .53 | **5.40 ± .22** | 7.60 ± .16 |
| diabetes | 6.00 ± .67 | 7.50 ± .76 | 6.00 ± .00 | **2.10 ± .53** |
| ficobin | **5.00 ± .00** | 7.00 ± 1.00 | 6.40 ± .16 | 11.80 ± .65 |
| heart | **6.00 ± .52** | 6.10 ± .10 | 11.90 ± 1.46 | 9.50 ± .82 |
| ionosphere | **6.00 ± .42** | 7.90 ± .59 | 17.90 ± 3.14 | 6.80 ± 1.96 |
| mammo | 5.60 ± .60 | 6.20 ± .13 | 7.20 ± .65 | **1.90 ± .60** |
| mushroom | **5.80 ± .25** | 6.00 ± .00 | 23.80 ± 1.50 | 9.90 ± .89 |
| spambase | 17.40 ± 1.36 | 17.60 ± .67 | 29.50 ± .76 | **5.60 ± .48** |

*Table 6.* Interpretable complexity. DT measured by depth