A polyhedral study of multivariate decision trees

Carla Michini$^{1[0000-0002-4717-816X]}$ and Zachary Zhou$^{1[0000-0003-0418-7178]}$

Department of Industrial and Systems Engineering, University of Wisconsin-Madison,
Madison, WI, USA
{michini,zzhou246}@wisc.edu

Abstract. Decision trees are a widely used tool for interpretable machine learning. Multivariate decision trees employ hyperplanes at the branch nodes to route datapoints throughout the tree and yield more compact models than univariate trees. Recently, mixed-integer programming (MIP) has been applied to formulate the optimal decision tree problem. To strengthen MIP formulations, it is crucial to introduce polyhedral characterizations of multivariate decision trees. A key component of most MIP formulations is a specification of how to route datapoints in the tree from the root to the leaves. Our goal is to characterize the set of realizable routings, i.e., routings that can be realized using multivariate branching rules. We first focus on shattering inequalities, a class of valid inequalities that can be used to strengthen almost any MIP formulation and that have been shown to be computationally effective. We prove that if all the feature vectors are in general position, then the shattering inequalities defined for the root of the tree are facet-defining for the convex hull of the realizable routings. We then show that every facet-defining inequality of a depth one tree involving at least two variables is also facet-defining for trees of arbitrary depth. Finally, we show that facet-defining inequalities characterizing realizable routings are also facet-defining for a complete MIP formulation.

Keywords: optimal decision tree · mixed-integer programming · facet-defining inequality.

1 Introduction

Interpretable machine learning (ML)–defined as the extraction of relevant knowledge from a ML model concerning relationships either contained in data or learned by the model [22]–is quickly becoming a requirement in many application areas to promote transparency, ensure fairness, and meet regulatory requirements [13, 14]. Decision trees are among the most popular standalone techniques for interpretable ML [10] and form the foundation for several more sophisticated ML algorithms such as random forest [9, 20]. A decision tree is a binary rooted tree where datapoints are directed from the root to the leaves depending on the outcomes of tests performed at each branch (i.e., internal) node of the tree. Each leaf is assigned a class k such that every datapoint routed to that leaf is classified as belonging to class k. An optimal decision tree maximizes the number of datapoints that are correctly classified, while limiting the tree complexity, e.g., the number of nodes in the tree or the number of tests performed.

The problem of learning optimal decision trees has attracted the attention of several communities. Recent approaches range from mixed-integer programming (MIP) [7, 1, 3, 11, 15, 19, 30, 29], to Boolean satisfiability [23, 6, 17, 25], constraint programming [28, 27], and dynamic programming [24, 4, 5, 16, 21, 12]. Most works focus on *univariate* decision trees, where each branch node tests only a single feature. Less work [7, 31] has been done concerning the problem of learning optimal *multivariate* decision trees, where tests involve a linear combination of the features. Even if multivariate tests are less easily interpretable, they provide more flexibility than univariate tests, which can only resort to axis-aligned hyperplanes. As a consequence, multivariate decision trees can be much more compact than univariate decision trees, i.e., the total number of tests needed to achieve a target accuracy can be dramatically smaller.

Most non-MIP approaches proposed for learning univariate trees–including heuristic methods–fail to carry over when considering multivariate trees. As a result, MIP is the premier technique for learning multivariate decision trees. Moreover, MIP has the advantage of providing a more flexible framework for learning optimal decision trees, since its expressive power can be leveraged to explicitly incorporate objectives of practical interest, such as fairness, robustness, and feature selection. Our goal is to provide polyhedral characterizations of multivariate decision trees that can be leveraged within MIP formulations to compute optimal multivariate decision trees.

Related work. Although a number of MIP formulations for optimal decision trees exist, to the best of our knowledge, there are only few polyhedral studies providing tools to strengthen these formulations. In the univariate setting, Aghaei et al. [3] define a flow-based formulation that can be solved using Benders decomposition. This decomposition exploits the max-flow structure of its subproblems by employing a tailored min-cut procedure, which produces facet-defining cuts. This flow-based formulation serves as the backbone of a number of subsequent works [2, 18]. In the multivariate setting, Zhu et al. [31] propose a formulation based on the 1-norm support vector machine and provide different classes of valid inequalities to further tighten the MIP model. Finally, Boutilier et al. [8] have recently introduced a new class of valid inequalities for multivariate trees, called *shattering inequalities*, that are used as feasibility cuts within a Benders-like decomposition. Unlike general-purpose cutting planes, shattering inequalities exploit the specific structure of the dataset to reveal critical substructures within it. In fact, they have an interesting geometric interpretation, being related to minimal subsets of datapoints that cannot be separated by linear classifiers. Another nice feature of shattering inequalities is that they involve only the decision variables specifying how the datapoints are routed throughout the tree. Such variables appear in almost every MIP formulation for optimal decision trees, thus shattering inequalities can be used to tighten most MIP formulations for multivariate trees. Experimental results using classic benchmark instances demonstrate that shattering inequalities can significantly reduce the computational burden of learning optimal trees [8].

Our contributions. The key contributions of this paper aim at characterizing the realizable routings of multivariate decision trees. Given a full binary tree of depth D with 2^D leaves, a *routing* specifies the path of every datapoint from the root to a leaf. We define a routing to be *realizable* if, at every branch node, there exists a hyperplane separating the datapoints that are routed to the left from those that are routed to the right. Each realizable routing can be represented as a binary vector and our goal is to study the convex hull W_D of the realizable routings.

Our first contribution is proving that certain shattering inequalities, specifically those defined at the root of the decision tree, are facet-defining, provided that the feature vectors of the datapoints are in general position. We remark that datasets with numerical features, for which multivariate trees are particularly well suited, typically satisfy this assumption. Thus, this result justifies why shattering inequalities are computationally very effective, especially when dealing with numerical datasets. However, shattering inequalities are not sufficient to describe W_D. Our second main result is that *every* facet-defining inequality of W_1 involving at least two variables is also facet-defining for W_D. This shows that, polyhedral characterizations of W_1 are crucial to describe W_D. Finally, we prove that every facet-defining inequality of W_D is also facet-defining for the convex hull of the feasible solutions of the MIP formulation proposed in [8].

2 Problem setting

The input to the problem is a dataset consisting of N datapoints. Each datapoint has p numerical features and belongs to one of K classes. We assume w.l.o.g. that each feature is in $[0, 1]$. For $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \ldots, n\}$. Thus, for each $i \in [N]$, the datapoint (x^i, y^i) is in $[0, 1]^p \times [K]$. We also assume that the feature vectors $x^1, \ldots, x^n$ are distinct, a mild assumption when working with numerical features. Our goal is to build a decision tree that predicts the class of a datapoint based on its features.

The nodes of a decision tree are partitioned into branch nodes and leaf nodes. The maximum *depth* D of the decision tree is the length of any path from the root to a leaf. The tree depth is used as an input parameter to control the size of the tree. The set of *branch nodes* is denoted by $\mathcal{B} = \{1, \ldots, 2^D - 1\}$, where node 1 is the root. Each branch node $t \in \mathcal{B}$ has exactly two children, $2t$ and $2t + 1$, and is associated with a linear *branching rule* $a_t^\top x = b_t$. When a datapoint x^i reaches t, x^i is routed to the left child $2t$ if $a_t^\top x^i \leq b_t$ and to the right child $2t + 1$ otherwise. *Univariate* branching rules require that only one component of a_t is nonzero, while *multivariate* branching rules impose no restriction on the support of a_t. The set of *leaf nodes* is denoted by $\mathcal{L} = \{2^D, \ldots, 2^{D+1} - 1\}$. Each leaf $t \in \mathcal{L}$ is a terminal node (i.e., it has no children) and is assigned a class label $k \in [K]$. All datapoints routed to leaf t are classified as belonging to class k. A datapoint x^i is correctly classified if the predicted class for x^i coincides with y^i.

We focus on learning optimal multivariate decision trees: our goal is to construct a decision tree of depth at most D maximizing the number of datapoints that are correctly classified.

Realizable routings. Many optimal decision tree formulations rely on binary variables indicating where in the tree each datapoint is routed. Let binary variable w_{it} be 1 if datapoint $i \in [N]$ is sent to node $t \in \mathcal{B} \cup \mathcal{L}$, and 0 otherwise. We refer to the binary vector w as a *routing*. Clearly, a routing has to satisfy

$$\sum_{t \in \mathcal{L}} w_{it} = 1 \; \forall i \in [N] \quad \text{and} \quad w_{it} = w_{i,2t} + w_{i,2t+1} \; \forall i \in [N], \; t \in \mathcal{B}. \tag{1}$$

We say that a routing w is *realizable* if in addition, for each branch node $t \in \mathcal{B}$, there exists a hyperplane $a_t^\top x = b_t$ that strictly separates the datapoints routed to $2t$ from those routed to $2t + 1$.

Many MIP formulations for decision trees contain the branching parameters $(a_t, b_t)_{t \in \mathcal{B}}$ as variables, and link them with the routing variables w using big-M constraints. This strategy has the disadvantage of producing weak linear programming relaxations. Recently, Boutilier et al. [8] proposed an alternative formulation that relies on a new class of valid inequalities, called shattering inequalities. These inequalities are crucially used to characterize the realizable routing vectors w. Let $\mathcal{I}$ be the set of pairs $(I_L, I_R) \in [N]^2$ such that: *(i)* I_L and I_R are disjoint; *(ii)* The sets $\{x^i\}_{i \in I_L}$ and $\{x^i\}_{i \in I_R}$ are *not* linearly separable; and *(iii)* For all $j \in I_L \cup I_R$, the sets $\{x^i\}_{i \in I_L \setminus \{j\}}$ and $\{x^i\}_{i \in I_R \setminus \{j\}}$ are linearly separable. The *shattering inequalities* at node $t \in \mathcal{B}$ are the packing constraints:

$$\sum_{i \in I_L} w_{i,2t} + \sum_{i \in I_R} w_{i,2t+1} \leq |I_L| + |I_R| - 1, \quad \forall (I_L, I_R) \in \mathcal{I}, \; t \in \mathcal{B}. \tag{2}$$

Constraints (1) and (2) are sufficient to ensure that a routing w is realizable. We denote by S_D the set of all the binary vectors satisfying (1) and (2), and we define $W_D = \operatorname{conv}(S_D)$.

Problem formulation. We adopt the formulation proposed in [8] to model the optimal decision tree problem. Let c_{kt} be a binary variable that is 1 if leaf $t \in \mathcal{L}$ is assigned class label $k \in [K]$, and 0 otherwise. Let z_{it} be a binary variable that is 1 if datapoint $i \in [N]$ is sent to leaf $t \in \mathcal{L}$, and 0 otherwise. The problem formulation is as follows.

$$\begin{aligned}
\underset{w,c,z}{\text{maximize}} \quad & \sum_{i \in [N]} \sum_{t \in \mathcal{L}} z_{it} && && \text{(3a)}\\
\text{subject to} \quad & w \in S_D && && \text{(3b)}\\
& \sum_{k \in [K]} c_{kt} = 1 && \forall t \in \mathcal{L}, && \text{(3c)}\\
& z_{it} \leq \min\{w_{it}, c_{y^i,t}\} && \forall i \in [N], \; t \in \mathcal{L}, && \text{(3d)}\\
& c_{kt} \in \{0, 1\} && \forall k \in [K], \; t \in \mathcal{L}, && \text{(3e)}\\
& z_{it} \in \{0, 1\} && \forall i \in [N], \; t \in \mathcal{L}. && \text{(3f)}
\end{aligned}$$

The objective function (3a) maximizes the number of datapoints correctly classified. Constraints (3b) ensure that w is a realizable routing and constraints (3d) model the condition that $z_{it} = 1$ if and only if datapoint i is sent to leaf node t and is correctly classified as y^i. In particular, as w and c are binary vectors, we have that at optimality $z_{it} = w_{it} c_{y^i,t}$ for all $i \in [N]$, $t \in \mathcal{L}$. This immediately implies that we can relax the integrality constraints on z, i.e., constraints (3f) can be replaced with $z_{it} \in \mathbb{R}$ for all $i \in [N]$ and $t \in \mathcal{L}$. We now argue that we can also relax the integrality constraints on the variables c by replacing constraints (3e) with $c_{kt} \in \mathbb{R}_+$ for all $k \in [K]$, $t \in \mathcal{L}$. To this end, we remark that relaxation of (3) obtained by relaxing integrality on z and c always admits an optimal solution that is integral. In fact, when we solve this relaxation we still have that at optimality $z_{it} = w_{it} c_{y^i,t}$ for all $i \in [N]$, $t \in \mathcal{L}$. Therefore, the objective (3a) is equivalent to maximizing

$$\sum_{i\in[N]} \sum_{t\in\mathcal{L}} z_{it} = \sum_{i\in[N]} \sum_{t\in\mathcal{L}} w_{it} c_{y^i,t} = \sum_{t\in\mathcal{L}} \sum_{k\in[K]} c_{kt} \sum_{\substack{i\in[N]\\ y^i=k}} w_{it}.$$

We observe that for each leaf $t \in \mathcal{L}$ the variables $\{c_{kt}\}_{k=1}^K$ can be viewed as weights that we assign to each class label. For all $t \in \mathcal{L}$ define $K(t) = \{k \in [K] : c_{kt} > 0\}$. Suppose that for some $t \in \mathcal{L}$ we have $|K(t)| > 1$, i.e., each variable c_{kt} with $k \in K(t)$ is fractional. It must hold that for each $k \in K(t)$ the sum $\sum_{i\in[N]:y^i=k} w_{it}$ has the same value. Then we can obtain an alternative optimal solution by choosing an arbitrary $k^* \in K(t)$ and redefining $c_{k^*,t} = 1$ and $c_{kt} = 0$ for all $k \in K(t) \setminus \{k^*\}$.

3 Preliminaries

In this section we introduce some preliminary lemmas that are instrumental for proving our main results. Denote by $\hat{\mathcal{L}} = \mathcal{L} \setminus \{2^D\}$ the set of leaves excluding the leftmost leaf. For $w \in W_D$, we denote by $\hat{w}$ the subvector of w consisting only of the components w_{it} such that $i \in [N]$, $t \in \hat{\mathcal{L}}$. Let $\hat{W}_D$ be the orthogonal projection of W_D onto the variables w_{it} with $i \in [N],\ t \in \hat{\mathcal{L}}$. Since the equality system (1) has full row rank, W_D and $\hat{W}_D$ are isomorphic polyhedra. We denote by $\phi(\cdot)$ the affine function which maps a vector $\hat{w} \in \hat{W}_D$ to the vector $w \in W_D$ that satisfies $w_{it} = \hat{w}_{it}$ for all $i \in [N]$, $t \in \hat{\mathcal{L}}$. From our discussion we obtain the following lemma.

Lemma 1. *Let $w^1, \ldots, w^q \in W_D$. The vectors $w^1, \ldots, w^q$ are affinely independent if and only if their orthogonal projections onto the variables w_{it} with $i \in [N],\ t \in \hat{\mathcal{L}}$ are affinely independent vectors.*

First, we determine the dimension of W_D.

Lemma 2. *For $D \in \mathbb{N}$,* $\dim(W_D) = N(|\mathcal{L}| - 1)$.

Proof. Since the system defined by equality constraints (1) has full row rank, we have $\dim(W_D) \leq N(|\mathcal{L}| - 1)$. To show that $\dim(W_D) \geq N(|\mathcal{L}| - 1)$, we exhibit

$N(|\mathcal{L}|-1)+1$ affinely independent routing vectors $w^0, \dots, w^{N(|\mathcal{L}|-1)}$ in W_D. We denote by $\hat{w}^0, \dots, \hat{w}^{N(|\mathcal{L}|-1)}$ their projections onto $\hat{W}_D$.

Recall that $x^1, \dots, x^N$ are assumed to be distinct. There exists a vector $a \in \mathbb{R}^p$ such that $a \neq \theta(x^i - x^j)$ for any $i, j \in [N]$, $i \neq j$ and $\theta \in \mathbb{R}$, i.e., a is distinct from $x^i - x^j$ even up to scaling. Furthermore, for $b \in \mathbb{R}$, the hyperplane $a^\top x = b$ contains at most one point in $x^1, \dots, x^N$. Therefore, we can assume without loss of generality that $x^1, \dots, x^N$ are ordered such that for all $i \in [N-1]$, the first i points $x^1, \dots, x^i$ can be strictly separated from the last $N-i$ points $x^{i+1}, \dots, x^N$ by a hyperplane, i.e., there exists $b_i \in \mathbb{R}$ such that $a^\top x^j < b_i$ for all $j \leq i$, and $a^\top x^j > b_i$ for all $j > i$.

First, we define the routing w^0 that sends all the datapoints to the leftmost leaf 2^D. Note that this routing is realizable and $\hat{w}^0 = \mathbf{0}$. Moreover, for every $i \in [N]$ and $t \in \hat{\mathcal{L}}$ we construct a routing vector $w^{i,t}$ that sends datapoints $j \leq i$ to leaf t, and datapoints $j > i$ to leftmost leaf 2^D. For each $i \in [N]$ we thus obtain $|\mathcal{L}| - 1$ routings. Note that the routings are realizable since, by construction, $x^1, \dots, x^i$ can be separated from $x^{i+1}, \dots, x^N$ by a hyperplane and so only one branch node of the decision tree applies a non-trivial split using $a^\top x = b_i$ as its branching hyperplane. Precisely, for each $i \in [N]$ and $t \in \hat{\mathcal{L}}$, $\hat{w}^{i,t}$ is such that

$$\hat{w}^{i,t}_{ju} = \begin{cases} 1 & \text{if } j \leq i \text{ and } u = t \\ 0 & \text{otherwise.} \end{cases}$$

We partition $\hat{w} \in \hat{W}_D$ by grouping the entries indexed by each $i \in [N]$ and ordering the variables as follows:

$$\hat{w} = (\tilde{w}_1, \dots, \tilde{w}_N), \text{ where } \tilde{w}_i = (w_{i,2^D+1}, w_{i,2^D+2}, \dots, w_{i,2^{D+1}-1}) \; \forall i \in [N].$$

In particular, for each $j \in [N]$ and $u \in \hat{\mathcal{L}}$, $\hat{w}^{i,t}_{ju}$ is the $((2^D-1)(j-2)+u-1)$-th component of $\hat{w}^{i,t}$. Next, we form a $N(|\mathcal{L}|-1) \times N(|\mathcal{L}|-1)$ square matrix where, for each $i \in [N]$, $t \in \hat{\mathcal{L}}$, the vector $w^{i,t}$ is the $((2^D-1)(i-2)+t-1)$-th row of the matrix. Since the $((2^D-1)(i-2)+t-1)$-th component of $\hat{w}^{i,t}$ is 1, and all subsequent components of $\hat{w}^{i,t}$ are 0, the matrix is lower triangular and has nonzero diagonal. This implies that the vectors $\hat{w}^{i,t}$, $i \in [N]$, $t \in \hat{\mathcal{L}}$, are linearly independent. Thus, these vectors and $\hat{w}^0 = \mathbf{0}$ are affinely independent. The result then follows by Lemma 1. □

Let $(I_L, I_R) \in \mathcal{I}$. It will be crucial to determine whether $I = I_L \cup I_R$ has the following property: for each datapoint $i \notin I$, there exists a hyperplane $a^\top x = b$ that traverses x^i and correctly separates all but one datapoint in (I_L, I_R). This is formally stated in the next definition. For all $k \in I$ we define an indicator variable σ^I_k that is 1 if $k \in I_L$ and is -1 otherwise.

Definition 1. *Let $(I_L, I_R) \in \mathcal{I}$ and $I = I_L \cup I_R$. We say that (I_L, I_R) is a* good partition *if, for every $i \in [N] \setminus I$, there exist $a \in \mathbb{R}^p$, $b \in \mathbb{R}$, $j \in I$ such that $a^\top x^i = b$, $\sigma^I_j a^\top x^j > \sigma^I_j b$, and $\sigma^I_k a^\top x^k < \sigma^I_k b$ for all $k \in I \setminus \{j\}$. A good partition is called* very good *if, in addition, $a^\top x^k \neq b$ for all $k \in [N] \setminus (I \cup \{i\})$.*

The next lemma establishes that every good partition is also a very good partition. Let $a^\top x = b$ be an hyperplane for which $(I_L, I_R) \in \mathcal{I}$ satisfies the definition of good partition and let $i \in [N] \setminus (I_L \cup I_R)$. By slightly tilting the hyperplane, we can ensure that the only datapoint traversed by the hyperplane is x^i.

Lemma 3. *If $(I_L, I_R) \in \mathcal{I}$ is a good partition, then it is a very good partition.*

Proof. Let $I = I_L \cup I_R$. Let $i \in [N] \setminus I$ be arbitrary, and let $a \in \mathbb{R}^p,\ b \in \mathbb{R},\ j \in I$ satisfy the definition of a good partition. Define the set $\mathcal{K} = \{k \in [N] \setminus (I \cup \{i\}) : a^\top x^k = b\}$. If $\mathcal{K} = \emptyset$, then (I_L, I_R) is a very good partition and we are done. Otherwise, we iteratively apply an algorithm that tilts the hyperplane $a^\top x = b$ about the point x^i so that at least one point $x^k,\ k \in \mathcal{K}$ is removed from the hyperplane, while also ensuring that the points in $\{x^h\}_{h \in [N] \setminus (\mathcal{K} \cup \{i\})}$ not originally contained in the hyperplane continue to remain outside and to the same side of it. Since $\mathcal{K}$ is finite, we can construct the desired hyperplane by iterating this procedure. More formally, the algorithm takes as input a, b, and an index $k \in \mathcal{K}$, and returns as output $\tilde{a}$ and $\tilde{b}$ such that the following conditions are met:

1. $\tilde{a}^\top x^i = \tilde{b}$
2. $\tilde{a}^\top x^k \neq \tilde{b}$
3. For all $h \in [N] \setminus (\mathcal{K} \cup \{i\})$, if $a^\top x^h < b$, then $\tilde{a}^\top x^h < \tilde{b}$, otherwise if $a^\top x^h > b$, then $\tilde{a}^\top x^h > \tilde{b}$.

We now describe how to compute $\tilde{a}$ and $\tilde{b}$. Let $r = x^k - x^i$. We define $\tilde{a} = a + \varepsilon r$ and $\tilde{b} = b + \varepsilon r^\top x^i$, where $\varepsilon > 0$ is a scalar that can be set arbitrarily small. We claim that for sufficiently small ε, $\tilde{a}$ and $\tilde{b}$ satisfy the three conditions above. It can be easily checked that condition 1 holds for any $\varepsilon > 0,\ r \in \mathbb{R}^p$. For condition 2, we have

$$\tilde{a}^\top x^k = a^\top x^k + \varepsilon r^\top x^k = b + \varepsilon \left(\|x^k\|_2^2 - (x^i)^\top x^k \right) \tag{4}$$

and

$$\tilde{b} = b + \varepsilon \left((x^k)^\top x^i - \|x^i\|_2^2 \right). \tag{5}$$

We next prove that $\tilde{a}^\top x^k > \tilde{b}$ for any $\varepsilon > 0$. From (4) and (5), this is equivalent to proving

$$\|x^i\|_2^2 + \|x^k\|_2^2 > 2(x^i)^\top x^k. \tag{6}$$

Inequality (6) follows directly from the assumption $x^i \neq x^k$ and the following basic claim.

Claim 1. Let $x, y \in \mathbb{R}^p,\ x \neq y$. Then $2x^\top y < \|x\|_2^2 + \|y\|_2^2$.

Proof of claim. By the Cauchy-Schwarz inequality, we have $x^\top y \leq \|x\|_2 \|y\|_2$, where equality holds if and only if $y = \lambda x$ for some $\lambda \geq 0$. For scalars $\alpha, \beta \in \mathbb{R}$, we have $2\alpha\beta \leq \alpha^2 + \beta^2$, where equality holds if and only if $\alpha = \beta$ (this can be shown by studying the inequality $(\alpha - \beta)^2 \geq 0$). Therefore,

$$2x^\top y \leq 2\|x\|_2 \|y\|_2 \leq \|x\|_2^2 + \|y\|_2^2.$$

We require at least one of the above two inequalities to be strict. Suppose the Cauchy-Schwarz inequality holds at equality. Then $y = \lambda x$ for some $\lambda \geq 0$. As $x \neq y$, it must be the case that $\|x\|_2 \neq \|y\|_2$, implying that the second inequality is strict. $\diamond$

We finally prove that condition 3 holds for some sufficiently small $\varepsilon > 0$. Let $h \in [N] \setminus (\mathcal{K} \cup \{i\})$. We have either $a^\top x^h < b$ or $a^\top x^h > b$. We deal with the case where $a^\top x^h < b$ as the other case follows similarly. We further assume without loss of generality that coefficients are scaled so that $a^\top x^h < b - 1$. We have

$$\begin{aligned}\tilde{a}^\top x^h &= a^\top x^h + \varepsilon r^\top x^h \\ &< b - 1 + \varepsilon r^\top x^h \\ &= \tilde{b} + \varepsilon r^\top (x^h - x^i) - 1.\end{aligned}$$

If $r^\top(x^h - x^i) \leq 0$, we will have $\tilde{a}^\top x^h < \tilde{b}$ for any $\varepsilon > 0$. Otherwise if $r^\top(x^h - x^i) > 0$, we have $\tilde{a}^\top x^h < \tilde{b}$ for $\varepsilon \in \left(0, \frac{1}{r^\top(x^h - x^i)}\right]$. Therefore, it is always possible to pick a sufficiently small $\varepsilon > 0$ such that $\tilde{a}^\top x^h < \tilde{b}$. □

Let $(I_L, I_R) \in \mathcal{I}$ be a very good partition and let $i \in [N] \setminus (I_L \cup I_R)$. In the next lemma, we consider a hyperplane $H = \{x \in \mathbb{R}^p : a^\top x = b\}$ for which (I_L, I_R) satisfies the definition of very good partition. For a sufficiently small value $\varepsilon > 0$, we can translate H and obtain two hyperplanes $H^+ = \{x \in \mathbb{R}^p : a^\top x = b + \varepsilon\}$ and $H^- = \{x \in \mathbb{R}^p : a^\top x = b - \varepsilon\}$ such that no datapoint $x^1, \dots, x^N$ lies in either H^+ and H^-, and, with the exception of the point x^i, the hyperplanes H^+ and H^- partition $x^1, \dots, x^N$ identically.

Lemma 4. *Let $(I_L, I_R) \in \mathcal{I}$ be a very good partition. For all $i \in [N] \setminus (I_L \cup I_R)$, there exists two binary routing vectors w^{i-} and w^{i+} in W_1 that satisfy at equality the shattering inequality* (2) *associated with (I_L, I_R) and $t = 1$ and such that, for all $t \in \{2, 3\}$, $w^{i-}_{it} = 1 - w^{i+}_{it}$ and $w^{i-}_{jt} = w^{i+}_{jt}$ for all $j \neq i$.*

Proof. Let $I = I_L \cup I_R$. Let $i \in [N] \setminus I$ be arbitrary, and let $a \in \mathbb{R}^p,\ b \in \mathbb{R},\ j \in I$ satisfy the definition of a very good partition. We deal with the case where $j \in I_L$, as the other case where $j \in I_R$ follows similarly. Let $0 < \varepsilon < \min_{k \in [N] \setminus \{i\}} |a^\top x^k - b|$. Then

$$\begin{aligned}b - \varepsilon < &a^\top x^i < b + \varepsilon, \\ b + \varepsilon < &a^\top x^k && \forall k \in I_R \cup \{j\}, \\ &a^\top x^k < b - \varepsilon && \forall k \in I_L \setminus \{j\}, \\ &a^\top x^k \notin [b - \varepsilon, b + \varepsilon] && \forall k \in [N] \setminus (I_L \cup I_R).\end{aligned}$$

Using the multivariate split parameters $(a, b-\varepsilon)$ at the root results in the routing w^{i-}, and using the multivariate split parameters $(a, b+\varepsilon)$ at the root results in the routing w^{i+}. By construction, datapoint i is routed left to leaf node 2 in w^{i+} and it is routed right to leaf node 3 in w^{i-}. All the other datapoints are routed identically in w^{i-} and w^{i+}. Note that both w^{i-} and w^{i+} satisfy at equality the shattering inequality (2) associated with (I_L, I_R) and $t = 1$ since, with the exception of datapoint j, all the datapoints in I are partitioned as prescribed by (I_L, I_R). □

We recall that a finite set of points X in $\mathbb{R}^p$ are in general position if no n points in X lie in an $(n-2)$-dimensional affine subspace for $n = 2, \ldots, p+1$. If the points in X are in general position, then every subset of at most $p+1$ datapoints in X is affinely independent. We say that the dataset is in general position if $x^1, \ldots, x^N$ are in general position. The next lemma follows from a classic result in statistical learning theory related to the notion of *VC dimension* [26].

Lemma 5. *If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is such that $|I_L| + |I_R| = p + 2$. Equivalently, each shattering inequality (2) has exactly $p+2$ nonzero coefficients.*

Next, we present two technical lemmas providing geometric properties of points partitions that can be linearly separated. For a finite set of points X in $\mathbb{R}^p$, we denote by $\operatorname{cone}(X)$ and $\operatorname{aff}(X)$ the conic hull and affine hull of X, respectively.

Lemma 6. *Let $\{x^k\}_{k \in J_L}$ and $\{x^k\}_{k \in J_R}$ be two sets of points in $\mathbb{R}^p$, and let $z \in \mathbb{R}^p$. If $\{x^k\}_{k \in J_L}$ and $\{x^k\}_{k \in J_R}$ can be strictly separated by a hyperplane traversing z, then*

$$\operatorname{cone}(\{x^k - z\}_{k \in J_L}) \cap \operatorname{cone}(\{x^k - z\}_{k \in J_R}) = \{\mathbf{0}\}. \tag{7}$$

If the points $\{x^k\}_{k \in J_L \cup J_R} \cup \{z\}$ are in general position and both J_L and J_R are nonempty, then the converse holds.

Proof. The points $\{x^k\}_{k \in J_L}$ and $\{x^k\}_{k \in J_R}$ can be strictly separated by a hyperplane containing z if and only if there exist $a \in \mathbb{R}^p,\ b \in \mathbb{R}$ satisfying

$$\begin{aligned} & a^\top x^k - b \le -1 && \forall k \in J_L, \\ & a^\top x^k - b \ge 1 && \forall k \in J_R, \\ & a^\top z = b. \end{aligned}$$

By exploiting the last equality, we can rewrite the above system as

$$a^\top (x^k - z) \le -1 \qquad \forall k \in J_L, \tag{8}$$

$$a^\top (x^k - z) \ge 1 \qquad \forall k \in J_R. \tag{9}$$

By Farkas' lemma, the system (8)-(9) is feasible if and only if the system

$$\sum_{k \in J_L} q_k (x^k - z) = \sum_{k \in J_R} q_k (x^k - z), \quad \sum_{k \in J_L \cup J_R} q_k > 0, \quad q \in \mathbb{R}_+^{J_L \cup J_R} \tag{10}$$

has no solution.

Suppose there exists $\bar{x} \in \operatorname{cone}(\{x^k - z\}_{k \in J_L}) \cap \operatorname{cone}(\{x^k - z\}_{k \in J_R}) \setminus \{\mathbf{0}\}$. This implies the existence of a solution to (10), equivalently $\{x^k\}_{k \in J_L}$ and $\{x^k\}_{k \in J_R}$ cannot be strictly separated by a hyperplane containing z.

Now assume the points $\{x^k\}_{k \in J_L \cup J_R} \cup \{z\}$ are in general position and both J_L and J_R are non-empty. If (7) is true, we distinguish two cases. In the first case, (10) has no solution, thus $\{x^k\}_{k \in J_L}$ and $\{x^k\}_{k \in J_R}$ can be separated by a hyperplane containing z. In the second case, (10) has a solution in which

$$\sum_{k \in J_L} q_k (x^k - z) = \sum_{k \in J_R} q_k (x^k - z) = \mathbf{0}.$$

Without loss of generality, assume the system

$$\sum_{k \in J_L} q_k (x^k - z) = \mathbf{0}, \quad \sum_{k \in J_L} q_k > 0, \quad q \in \mathbb{R}_+^{J_L}$$

is feasible. By scaling q, it follows that $\mathbf{0} \in \operatorname{conv}(\{x^k - z\}_{k \in J_L})$, equivalently $z \in \operatorname{conv}(\{x^k\}_{k \in J_L})$. Since the points $\{x^k\}_{k \in J_L \cup J_R} \cup \{z\}$ are in general position, it must be the case that $|J_L| \geq p + 1$. The general position assumption also implies that $\operatorname{conv}(\{x^k\}_{k \in J_L})$ is a full-dimensional set with z in its interior. Thus, $\operatorname{cone}(\{x^k - z\}_{k \in J_L}) = \mathbb{R}^p$. Finally, we observe that $\operatorname{cone}(\{x^k - z\}_{k \in J_R})$ has dimension at least one, since J_R is non-empty and, since the points are in general position, each point in J_R is distinct from z. This implies that $\operatorname{cone}(\{x^k - z\}_{k \in J_L}) \cap \operatorname{cone}(\{x^k - z\}_{k \in J_R}) \setminus \{\mathbf{0}\} \neq \emptyset$, a contradiction. □

Note that when the points $\{x^k\}_{k \in J_L \cup J_R} \cup \{z\}$ are not in general position, the converse may not hold. For example, consider the following example in $\mathbb{R}^2$: $J_L = \{1, 2\},\ J_R = \{3\},\ x^1 = (1, 0),\ x^2 = (-1, 0),\ x^3 = (0, 1),\ z = (0, 0)$.

Lemma 7. *Suppose that the dataset is in general position and that $p \geq 2$. Let $J \subseteq [N]$ be the indices of $p - 1$ points, partitioned into J_L and J_R such that $|J_L| \leq |J_R|$. Let $i \in [N] \setminus J$ and $z = x^i$.*

1. *The affine hull of $\{x^k\}_{k \in J} \cup \{z\}$ is a hyperplane, i.e., $\operatorname{aff}(\{x^k\}_{k \in J} \cup \{z\}) = \{x \in \mathbb{R}^p : \pi^\top x = \gamma\}$ for some $\pi \in \mathbb{R}^p$ and $b \in \mathbb{R}$.*
2. *If $S \subseteq [N] \setminus (J \cup \{i\})$ is such that $\pi^\top x^k < \gamma$ for each $k \in S$, then there exists a hyperplane traversing z that strictly separates $\{x^k\}_{k \in J_R}$ and $\{x^k\}_{k \in J_L \cup S}$.*

Proof. To prove 1 we observe that, since the datapoints $\{x^k : k \in J\} \cup \{z\}$ are in general position, then the affine hull of these p points has dimension $p - 1$ and it is a hyperplane in $\mathbb{R}^p$, denoted by $\pi^\top x = \gamma$.

To prove 2, we distinguish two cases: (a) S is empty, and (b) S is nonempty. We first consider case (a). When J_L is empty, the statement follows trivially, so suppose J_L is nonempty. Note that $C_J = \operatorname{cone}(\{x^k - z\}_{k \in J})$ is a simplicial cone, i.e., its generators are linearly independent. Thus, $\operatorname{cone}(\{x^k - z\}_{k \in J_L})$ and $\operatorname{cone}(\{x^k - z\}_{k \in J_R})$ are two faces of C_J whose intersection is $\{\mathbf{0}\}$. Since $\{x^k\}_{k \in J} \cup \{z\}$ are in general position, by Lemma 6 we conclude that there exists a hyperplane traversing z that strictly separates $\{x^k\}_{k \in J_L}$ and $\{x^k\}_{k \in J_R}$.

We now consider case (b). By contradiction, suppose there does not exist a hyperplane traversing z that strictly separates $\{x^k\}_{k\in J_R}$ and $\{x^k\}_{k\in J_L\cup S}$. Since $p \geq 2$, both J_R and $J_L \cup S$ are nonempty, so by Lemma 6, there exists $y \neq z$ such that $y - z \in \operatorname{cone}(\{x^k - z\}_{k\in J_R}) \cap \operatorname{cone}(\{x^k - z\}_{k\in J_L\cup S})$. Since $y - z \in \operatorname{cone}(\{x^k - z\}_{k\in J_L\cup S})$, we have

$$y - z = \sum_{k\in J_L} \lambda_k(x^k - z) + \sum_{k\in S} \lambda_k(x^k - z), \quad \lambda_k \geq 0 \;\forall\, k \in J_L \cup S. \tag{11}$$

It cannot be the case that $\lambda_k = 0$ for all $k \in S$; this would imply $y - z \in \operatorname{cone}(\{x^k - z\}_{k\in J_L})$, and since we already have $y - z \in \operatorname{cone}(\{x^k - z\}_{k\in J_R})$, by Lemma 6 we would have that $\{x^k\}_{k\in J_L}$ and $\{x^k\}_{k\in J_R}$ cannot be separated by a hyperplane traversing z, a contradiction. Thus, $\lambda_k > 0$ for some $k \in S$. Recall that $\{x^k\}_{k\in J} \cup \{z\}$ are contained in the hyperplane $\pi^T x = \gamma$. From (11), we have

$$\begin{aligned}\pi^\top y &= \pi^\top z + \sum_{k\in J_L} \lambda_k \pi^\top(x^k - z) + \sum_{k\in S} \lambda_k \pi^\top(x^k - z)\\ &= \gamma + \sum_{k\in S} \lambda_k(\pi^\top x^k - \gamma) < \gamma,\end{aligned}$$

where the last inequality follows from the assumption that $\pi^\top x^k < \gamma$ for all $k \in S$ and $\lambda_k > 0$ for some $k \in S$. This means $y \notin \operatorname{aff}(\{x^k\}_{k\in J} \cup \{z\})$, which contradicts the fact that $y - z \in \operatorname{cone}(\{x^k - z\}_{k\in J_R})$. □

4 Strength of shattering inequalities for depth 1 trees

In this section we prove two sufficient conditions ensuring that a shattering inequality is facet-defining for W_1.

Theorem 1. *Let $(I_L, I_R) \in \mathcal{I}$ be a very good partition. Then the inequality (2) associated with (I_L, I_R) and $t = 1$ is facet-defining for W_1.*

Proof. Let $I = I_L \cup I_R$. We will show that there are N affinely independent vectors in W_1 that satisfy the inequality (2) corresponding to (I_L, I_R), $t = 1$ at equality. To do this, we define $2N - |I|$ vectors in W_1 satisfying (2) at equality, and we show that among them, N vectors are affinely independent.

Define a binary vector χ^I such that for all $i \in I$ $\chi^I_i = 1$ if and only if $i \in I_L$. Since $I \in \mathcal{I}$, for each $i \in I$ we can partition the datapoints in I by misplacing only i. Precisely, there exists a binary routing $w^i \in W_1$ such that $w^i_{j,2} = \chi^I_j$ if $j \in I \setminus \{i\}$ and $w^i_{i,2} = 1 - \chi^I_i$. We remark that w^i satisfies (2) ((I_L, I_R), $t = 1$) at equality. For each $i \in I$, we denote by $\hat{w}^i$ the subvector of w^i containing the entries $w^i_{j,2}$ indexed by $j \in I$. Note that $\hat{w}^i_j = w^i_{j,2}$ for each $j \in I$.

Claim 2. The subvectors $\hat{w}^i$, $i \in I$, are affinely independent. Moreover, if $|I_L| \geq 2$, they are linearly independent.

Proof of claim. If $\chi_i^I = 0$, then $\hat{w}_i^i = 1$ and $\hat{w}_i^j = 0$ for $j \in I \setminus \{i\}$, hence $\hat{w}^i$ is clearly not an affine combination of the other vectors. If $\chi_i^I = 1$, then $\hat{w}_i^i = 0$ and $\hat{w}_i^j = 1$ for $j \in I \setminus \{i\}$, so for $\hat{w}^i$ to be an affine combination of the other vectors, it should hold $0 = \hat{w}_i^i = \sum_{j \in I \setminus \{i\}} \lambda_j \hat{w}_i^j = \sum_{j \in I \setminus \{i\}} \lambda_j = 1$, a contradiction.

Now assume that $|I_L| \geq 2$, and suppose by contradiction that $\hat{w}^i$, $i \in I$ are linearly dependent. Then there exist scalars λ_i, $i \in I$ not all equal to zero, such that $\sum_{i \in I} \lambda_i \hat{w}^i = 0$. For each $i \in I_R$ we have $\hat{w}_i^i = 1$ and $\hat{w}_i^j = 0$ for all $j \in I \setminus \{i\}$, thus it must be $\lambda_i = 0$. Moreover, for each $i \in I_L$ we have $\hat{w}_i^i = 0$ and $\hat{w}_i^j = 1$ for all $j \in I \setminus \{i\}$. Thus $\sum_{j \in I} \lambda_j \hat{w}_i^j = 0$ implies $\sum_{j \in I_R} \lambda_j \hat{w}_i^j + \sum_{j \in I_L} \lambda_j \hat{w}_i^j = \sum_{j \in I_L \setminus \{i\}} \lambda_j = 0$. Summing up over all $i \in I_L$, we obtain $(|I_L| - 1) \sum_{i \in I_L} \lambda_i = 0$. Thus $\sum_{i \in I_L} \lambda_i = 0$, and since $\lambda_i = 0$ for all $i \in I_R$ we obtain $\sum_{i \in I} \lambda_i = 0$. This contradicts the fact that $\hat{w}^i$, $i \in I$, are affinely independent. $\diamond$

By arranging w^i, $i \in I$ as rows of a matrix, we obtain the block matrix $\begin{bmatrix} A \; B \; C \end{bmatrix}$, where A is a $|I| \times |I|$ matrix containing the subvectors $\hat{w}^i$, i.e., the entries $w_{j,2}^i$ for $j \in I$, B is a $|I| \times (N - |I|)$ matrix containing the entries $w_{j,2}^i$ for $j \in [N] \setminus I$, and C is a $|I| \times N$ matrix containing the entries $w_{j,3}^i$ for $j \in [N]$. By Claim 2, A has full affine rank.

Moreover, for each $i \in [N] \setminus I$, we consider the routings $w^{i-}, w^{i+} \in W_1$ given by Lemma 4. Note that all of these routings satisfy the inequality (2) corresponding to I, $t = 1$ at equality. We construct a $(2N - 2|I|) \times 2N$ matrix D using the vectors w^{i+} as first $N - |I|$ rows, and the vectors w^{i-} as the next $N - |I|$ rows. As before, the first N columns are indexed by $t = 2$, and the last N columns are indexed by $t = 3$. We further partition the first N columns into those indexed by I and $[N] \setminus I$. We now perform N elementary row operations where, for each $i \in [N] \setminus I$, we replace w^{i+} with $w^{i+} - w^{i-}$. By Lemma 4, we have that $w^{i+} - w^{i-} = (e_i, -e_i)$, where e_i denotes the i-th row of the $N \times N$ identity matrix I_N. Let D' be the matrix obtained after the elementary row operations. We define a submatrix D'' of D', obtained by selecting the first $N - |I|$ rows and the first N columns of D'. We construct the $N \times N$ block matrix

$$M = \begin{bmatrix} A \; B \\ D'' \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & I_N \end{bmatrix}.$$

We now consider two different cases. In the first case $|I_L| \geq 2$. By Claim 2, we have that A has nonzero determinant, implying that M had also nonzero determinant. In particular, M has full affine rank. In the second case $|I_L| = 1$. Assume w.l.o.g. that $I_L = \{1\}$. In this case we can choose $w^1 = 0$, i.e., all the datapoints are routed to the right towards leaf 3. As a consequence, the first row of M has all zero entries. We then subtract the first row of M from all the subsequent rows. By Claim 2, we know that A has full affine rank, thus we can conclude that M has full affine rank. $\square$

Theorem 2. *If the dataset is in general position, then every $(I_L, I_R) \in \mathcal{I}$ is a good partition.*

Proof. Let $(I_L, I_R) \in \mathcal{I}$ and $I = I_L \cup I_R$. We prove that for every $i \in [N] \setminus I$, there exist $a \in \mathbb{R}^p$, $b \in \mathbb{R}$, $j \in I$ such that $a^\top x^i = b$, $\sigma^I_j a^\top x^j > \sigma^I_j b$ and $\sigma^I_k a^\top x^k < \sigma^I_k b$ for all $k \in I \setminus \{j\}$. Recall that, since the dataset is in general position, by Lemma 5 we have $|I_L| + |I_R| = p + 2$. We distinguish between the cases $\min\{|I_L|, |I_R|\} = 1$ and $\min\{|I_L|, |I_R|\} \geq 2$.

Case 1 ($\min\{|I_L|, |I_R|\} = 1$*).* We assume $|I_L| = 1$ (equivalently $|I_R| = p+1$); the case where $|I_R| = 1$ (equivalently $|I_L| = p + 1$) follows similarly. Let $I_L = \{\ell\}$. Since the dataset is in general position, the points $\{x^k\}_{k \in I_R}$ are the vertices of the p-simplex $P = \operatorname{conv}(\{x^k\}_{k \in I_R})$, and x^ℓ lies in the interior of P. If $x^i \notin P$, then simply take j to be ℓ. Any hyperplane separating x^i from $\{x^k\}_{k \in I_R}$ can be pushed towards x^i until x^i lies on it. If $x^i \in P$, since $x^1, \ldots, x^N$ are general position, x^i lies in the interior of P. Let the system $\{Cx \leq d\} = \{(c^k)^\top x \leq d_k\}_{k \in I_R}$ be a minimal representation of P, where the inequality $(c^k)^\top x \leq d_k$ corresponds to the facet of P that does *not* contain x^k. Note that for each $k \in I_R$, we have $(c^k)^\top x^k = \min_{x \in P} (c^k)^\top x < d_k$ and $(c^k)^\top x^h = d_k$ for $h \in I_R \setminus \{k\}$, i.e., x^k minimizes $(c^k)^\top x$ over all $x \in P$, whereas the other points $\{x^h\}_{h \in I_R \setminus \{k\}}$ do not. We also have $(c^k)^\top x^i < d_k$ for $k \in I_R$, as x^i lies in the interior of P. Define $\underline{d} = Cx^\ell$.

Claim 3. $\{x \in \mathbb{R}^p : Cx \leq \underline{d}\} = \{x^\ell\}$.

Proof of claim. Suppose there exists $z \in \mathbb{R}^p$, $z \neq x^\ell$ such that $Cz \leq d$. We have $C(z - x^\ell) = Cz - d \leq \mathbf{0}$. For every $x \in P$ and $\lambda \geq 0$, we have $C(x + \lambda(z - x^\ell)) = Cx + \lambda C(z - x^\ell) \leq d$, implying P is unbounded, a contradiction as P is the convex hull of a finite set of points. ◇

Therefore, every $x \in \mathbb{R}^p \setminus \{x^\ell\}$, in particular $x = x^i$, must satisfy $(c^j)^\top x > \underline{d}_j$ for some $j \in I_R$. Let $a = c^j$, $b = (c^j)^\top x^i$. Note that, by definition of a and b, $a^\top x^i = b$. As $(c^j)^\top x^\ell = \underline{d}_j$ and j is chosen so that $\underline{d}_j < (c^j)^\top x^i$, we also have $a^\top x^\ell < b$. Moreover, since x^j is the only vertex of P minimizing $(c^j)^\top x$, we obtain $a^\top x^j < b$ and $a^\top x^k > b$ for all $k \in I_R \setminus \{j\}$.

Case 2 ($\min\{|I_L|, |I_R|\} \geq 2$*).* Let $z = x^i$. Define $C_R = \operatorname{cone}(\{x^k - z\}_{k \in I_R})$ and $C_L = \operatorname{cone}(\{x^k - z\}_{k \in I_L})$. For each $h \in I_R$, define $C_R(h) = \operatorname{cone}(\{x^k - z\}_{k \in I_R \setminus \{h\}})$, similarly for each $h \in I_L$, define $C_L(h) = \operatorname{cone}(\{x^k - z\}_{k \in I_L \setminus \{h\}})$. We first prove the following claim.

Claim 4. There exists $j \in I_L \cup I_R$ *such that if* $j \in I_L$*, then* $C_L(j) \cap C_R = \{\mathbf{0}\}$*, otherwise if* $j \in I_R$*, then* $C_L \cap C_R(j) = \{\mathbf{0}\}$.

Proof of claim. Since $|I_L| + |I_R| = p + 2$, we have $\min\{|I_L|, |I_R|\} \geq 2$ implies that $|I_L| \leq p$ and $|I_R| \leq p$. Furthermore, since the dataset is in general position, $z = x^i \notin \operatorname{aff}(\{x^k : k \in I_R\})$, moreover $\{x^k - z\}_{k \in I_R}$ are linearly independent and C_R is a simplicial cone of dimension $|I_R|$. Since the cone is simplicial, for each $k \in I_R$, $C_R(k)$ is a facet of C_R. Similarly, C_L is a simplicial cone of dimension $|I_L|$, and for each $k \in I_L$, $C_L(k)$ is a facet of C_L. We remark that C_L and C_R cannot have a generator in common. If this was the case, we would have three colinear datapoints (including z), a contradiction. This observation implies that $C_L \neq C_R$. Next, we assume w.l.o.g. $|I_L| \leq |I_R|$.

Case 2(a) ($|I_L| = 2$, equivalently $|I_R| = p$). Assume w.l.o.g. $I_L = \{1, 2\}$. We consider two subcases. In the first subcase, $C_L \subset C_R$. We claim that for each $j \in I_R$ we have $C_R(j) \cap C_L = \{\mathbf{0}\}$. By contradiction, suppose there exists $y \neq z$ such that $(y - z) \in C_R(j) \cap C_L$. Since $(y - z) \in C_L$, we have

$$y - z = r_1(x^1 - z) + r_2(x^2 - z), \tag{12}$$

with $r_1, r_2 \geq 0$ and at least one of r_1, r_2 is strictly positive. Since $C_L \subset C_R$, both $(x^1 - z)$ and $(x^2 - z)$ can be obtained as conic combinations of $\{x^k - z\}_{k \in I_R}$, i.e.,

$$x^1 - z = \sum_{k \in I_R} \lambda_k (x^k - z), \qquad x^2 - z = \sum_{k \in I_R} \mu_i (x^k - z), \tag{13}$$

where $\lambda_k, \mu_k \geq 0$ for each $k \in I_R$. Note that in the above conic combinations, λ_j (resp. μ_j) must be strictly positive, otherwise x^1 (resp. x^2) would lie in the affine hull of the p points $\{x^k\}_{k \in I_R \cup \{i\} \setminus \{j\}}$, contradicting the fact that the dataset is in general position. By combining (12) and (13) we obtain

$$y - z = \sum_{k \in I_R} (r_1 \lambda_i + r_2 \mu_i)(x^k - z). \tag{14}$$

Since $r_1 \lambda_j + r_2 \mu_j > 0$, we have contradicted that $(y - z) \in C_R(j)$.

Now we consider the subcase where $C_L \subset C_R$ does not hold. Then we have that exactly one among $(x^1 - z)$ and $(x^2 - z)$ lies inside C_R; if it is $(x^1 - z)$, then, $C_L(1) \cap C_R = \mathbf{0}$, otherwise $C_L(2) \cap C_R = \mathbf{0}$.

Case 2(b) ($|I_L| \geq 3$, equivalently $|I_R| \leq p-1$). We assume w.l.o.g. $\{1, 2, 3\} \subseteq I_L$. Let $J = I_L \cup I_R \setminus \{1, 2, 3\}$, and note that $|J| = p - 1$. Consider the partition of J defined by $J_R = I_R$ and $J_L = I_L \setminus \{1, 2, 3\}$. Note that $p \geq 2$, since $\min\{|I_L|, |I_R|\} \geq 2$. By Lemma 7 (1), the affine hull of $\{x^k\}_{k \in J} \cup \{z\}$ is a hyperplane $\pi^\top x = \gamma$ in $\mathbb{R}^p$. If $\{x^1, x^2, x^3\}$ are all on one side of the hyperplane $\pi^\top x = \gamma$, then Lemma 7 (2) implies that $\{x^k\}_{k \in I_L}$ and $\{x^k\}_{k \in I_R}$ are separable, a contradiction. Then, we assume w.l.o.g. that $\pi^\top x^2 > \gamma$, $\pi^\top x^3 > \gamma$ and $\pi^\top x^1 < \gamma$. Since $\pi^\top x^2 > \gamma$ and $\pi^\top x^3 > \gamma$, Lemma 7 (2) implies that $\{x^k\}_{k \in J_L \cup \{2,3\}}$ and $\{x^k\}_{k \in H_R}$ can be strictly separated by an hyperplane traversing z. Then, by Lemma 6, it follows that $C_L(1) \cap C_R = \{\mathbf{0}\}$. ◇

Let $j \in I_L$ be such that $C_L(j) \cap C_R = \{\mathbf{0}\}$. By Lemma 6, $\{x^k\}_{k \in I_L \setminus \{j\}}$ and $\{x^k\}_{k \in I_R}$ can be strictly separated by a hyperplane containing $z = x^i$. Therefore, there exist $a \in \mathbb{R}^p$, $b \in \mathbb{R}$ such that $a^\top x^i = b$, $a^\top x^k < b$ for all $k \in I_L \setminus \{j\}$ and $a^\top x^k > b$ for all $k \in I_R \cup \{j\}$. If $j \in I_R$ is such that $C_L \cap C_R(j) = \{\mathbf{0}\}$, through analogous arguments the lemma follows. □

The next corollary is directly implied by Theorems 1 and 2.

Corollary 1. *If the dataset is in general position, then the shattering inequalities* (2) *are facet-defining for* W_1.

5 Facets for arbitrary depth trees

Due to (1), any valid inequality for W_1 can be expressed using only the variables $(w_{i,2})_{i=1}^N$. Thus, every facet of W_1 can written in the form

$$\sum_{i=1}^{N} f_i w_{i,2} \leq g. \tag{15}$$

Of course, (15) is not only a valid inequality for W_1, but more generally for W_D.

Given a routing $w \in W_D$, one can obtain a "symmetric" routing $\bar{w}$ that satisfies $\bar{w}_{i,2} = 1 - w_{i,2}$ for all $i \in [N]$; this is done by flipping the orientation of the branching hyperplane at the root. This observation yields the following lemma.

Lemma 8. *For $D \in \mathbb{N}$, inequality* (15) *is a facet of W_D if and only if the "symmetric" inequality*

$$\sum_{i=1}^{N} f_i w_{i,2} \geq \sum_{i=1}^{N} f_i - g \tag{16}$$

is a facet of W_D.

Proof. If (15) is a facet of W_D, then there are $q = N(|\mathcal{L}|-1)$ affinely independent vectors $w^1, \ldots, w^q \in W_D$ that satisfy (15) at equality. Our goal is to construct q affinely independent vectors $\tilde{w}^1, \ldots, \tilde{w}^q \in W_D$ that satisfy (16) at equality.

For the depth $D = 1$ case, this can be easily seen using the vectors $\{\phi(\mathbf{1} - \hat{w}^j)\}_{j=1}^N$, which are contained in W_1, affinely independent, and satisfy (16) at equality. In other words, for each $j \in [N]$, we define $\tilde{w}^j$ such that $\tilde{w}_{it}^j = 1 - w_{it}^j$ for $t = 2, 3$, i.e., the routing $\tilde{w}^j$ sends the points $\{x^i : w_{i,2}^j = 1\}$ to the right leaf, and the points $\{x^i : w_{i,3}^j = 1\}$ to the left leaf.

The argument generalizes to the depth $D \geq 2$ case, where we begin with q affinely independent binary vectors $w^1, \ldots, w^q$, and for each $j \in [q]$, we construct a routing $\tilde{w}^j$ by sending the points $\{x^i : w_{i,2}^j = 1\}$ down the right subtree in the same manner they are sent down the left subtree by the routing w^j, and the points $\{x^i : w_{i,3}^j = 1\}$ down the left subtree in the same manner they are sent down the right subtree by the routing w^j Formally, for each $j \in [q]$, we define the routing $\tilde{w}^j$ (only the components $\tilde{w}_{it}^j$, $t \in \hat{\mathcal{L}}$) as follows:

$$\tilde{w}_{it} = \begin{cases} w_{i,t+2^{D-1}} & \text{if leaf } t \text{ is in the left subtree} \\ w_{i,t-2^{D-1}} & \text{if leaf } t \text{ is in the right subtree.} \end{cases}$$

□

We introduce our second main result. This result can be applied to shattering inequalities as well as to any other facet-defining inequality (15) of W_1.

Theorem 3. *A facet-defining inequality of W_1 involving at least two variables is also facet-defining for W_D, where $D \geq 2$.*

Proof. Let (15) be facet-defining for W_1. There exist N affinely independent binary vectors $w^1, \ldots, w^N$ in W_1 satisfying (15) at equality. Let $\hat{w}^1, \ldots, \hat{w}^N$ be their projections onto $\hat{W}_1$ We remark that $\text{aff}(\hat{w}^1, \ldots, \hat{w}^N) = \{\hat{w} \in \mathbb{R}^n : f^\top \hat{w} = g\}$. Thus: (i) $g = 0$ if and only if $\mathbf{0} \in \text{aff}(\hat{w}^1, \ldots, \hat{w}^N)$ and (ii) $\sum_{i=1}^N f_i = g$ if and only if $\mathbf{1} \in \text{aff}(\hat{w}^1, \ldots, \hat{w}^N)$. Our strategy is to construct $|\mathcal{L}| - 1$ routing vectors in $\hat{W}_D$ using $\hat{w}^i \in \hat{W}_1$ as a starting point, for $i \in [N]$. We partition $\tilde{w} \in \hat{W}_D$ by grouping the entries indexed by each $t \in \hat{\mathcal{L}}$ and ordering the variables as follows:

$$\tilde{w} = (\tilde{w}_{2^D+1}, \tilde{w}_{2^D+2}, \ldots, \tilde{w}_{2^{D+1}-1}), \text{ where } \tilde{w}_t = (w_{1,t}, \ldots, w_{N,t}) \; \forall \, t \in \hat{\mathcal{L}}.$$

We first show that $\sum_{i=1}^N f_i = g = 0$ cannot hold. By contradiction, suppose $\sum_{i=1}^N f_i = g = 0$, thus (15) is $\sum_{i=1}^N f_i w_{i,2} \leq 0$. By Lemma 8, $\sum_{i=1}^N f_i w_{i,2} \geq 0$ is also a facet of W_1. Thus $\sum_{i=1}^N f_i w_{i,2} = 0$ for all $w \in W_1$, contradicting that (15) and (16) are facets of W_1. We consider three disjoint cases separately: (1) $g \neq 0$ and $\sum_{i=1}^N f_i \neq g$; (2) $g = 0$ and $\sum_{i=1}^N f_i \neq g$; and (3) $g \neq 0$ and $\sum_{i=1}^N f_i = g$.

Case 1. Since $\mathbf{0} \notin \text{aff}(\hat{w}^1, \ldots, \hat{w}^N)$, the vectors $\hat{w}^1, \ldots, \hat{w}^N$ are actually linearly independent. Similarly, since $\mathbf{1} \notin \text{aff}(\hat{w}^1, \ldots, \hat{w}^N)$, we have $\mathbf{0} \notin \text{aff}(\mathbf{1}-\hat{w}^1, \ldots, \mathbf{1}-\hat{w}^N)$, thus the vectors $\mathbf{1}-\hat{w}^1, \ldots, \mathbf{1}-\hat{w}^N$ are also linearly independent. We define two $N \times N$ nonsingular matrices $L = (\hat{w}^1, \ldots, \hat{w}^N)^\top$ and $R = (\mathbf{1} - \hat{w}^1, \ldots, \mathbf{1} - \hat{w}^N)^\top$ and and we construct an $N(|\mathcal{L}|-1) \times N(|\mathcal{L}|-1)$ matrix M using as blocks L, R and the $N \times N$ zero matrix. For $i, j \in [|\mathcal{L}| - 1]$ we denote by M_{ij} the block of M in position (i, j). We define

$$M_{ij} = \begin{cases} L & \text{if } i \in [|\mathcal{L}|/2 - 1], j = i \\ R & \text{if } i \in [|\mathcal{L}|/2 - 1], j = |\mathcal{L}| - 1 \text{ or } i \in \{|\mathcal{L}|/2, \ldots, |\mathcal{L}| - 1\}, j = i \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Intuitively, for each routing w^i, $i \in [N]$ of the depth 1 decision tree, we construct $|\mathcal{L}| - 1$ routings for the depth D decision tree as follows. Denote by T_2^i and T_3^i the datapoints that are sent to leaf 2 and to leaf 3, respectively, by w^i. We construct $|\mathcal{L}|/2$ routings that send T_2^i to the leftmost leaf of the depth D decision tree and T_3^i to one of the leaves of the subtree routed at node 3. Similarly, we construct $|\mathcal{L}|/2 - 1$ routings that send T_3^i to the rightmost leaf of the depth D decision tree and T_2^i to one of the leaves of the subtree routed at node 2 that is not the leftmost leaf. We claim that each row $\tilde{w}$ of M is such that $\phi(\tilde{w}) \in W_D$ and $\phi(\tilde{w})$ satisfies (15) with equality. Indeed, $\phi(\tilde{w})$ partitions the dataset at the root identically to some routing among $\phi(\hat{w}^1), \ldots, \phi(\hat{w}^N)$, while at every other branch node the datapoints are all routed together either to the left or to the right. Note that M is a block triangular matrix and each block along the main diagonal is nonsingular, therefore M is nonsingular. By Lemma 1, there are $N(|\mathcal{L}| - 1)$ affinely independent vectors in W_D that satisfy (15) with equality.

Case 2. Both L and R have affinely independent rows. However, while R is nonsingular, as $\mathbf{0} \notin \text{aff}(\mathbf{1} - \hat{w}^1, \ldots, \mathbf{1} - \hat{w}^N)$, L is singular, as $\mathbf{0} \in \text{aff}(\hat{w}^1, \ldots, \hat{w}^N)$. W.l.o.g., assume that $\hat{w}^1 = \mathbf{0}$. Thus $\hat{w}^2, \ldots, \hat{w}^N$ are linearly independent. Observe that $H = \text{aff}(\hat{w}^1, \ldots, \hat{w}^N) = \{\hat{w} \in \mathbb{R}^N : f^\top \hat{w} = 0\}$ is a linear subspace of $\mathbb{R}^N$ of dimension $N - 1$, and the vectors $\hat{w}^2, \ldots, \hat{w}^N$ span H.

We note that the matrix L cannot have a column of all zeros. Suppose for a contradiction that the ith column of L was a column of zeros. Then the vectors $\hat{w}^1, \ldots, \hat{w}^N$ satisfy $w_{i,2} \geq 0$ at equality and thus lie on the face $\{w \in W_1 : w_{i,2} = 0\}$. Since (15) involves at least two variables, we conclude that $\hat{w}^1, \ldots, \hat{w}^N$ are at the intersection of two distinct faces of W_1, thus they cannot span a linear subspace of dimension $N-1$.

As $x^1, \ldots, x^N$ are all distinct, we can pick some $i \in \operatorname{supp}(f)$ such that x^i can be strictly separated from $\{x^j\}_{j \in \operatorname{supp}(f) \setminus \{i\}}$ using a hyperplane. Since L does not have a column of zeros, there exists $k \in [N]$ such that $w^k_{i,2} = 1$, i.e., i is routed to leaf 2 in w^k. Let $J = \{j \in [N] : w^k_{j,2} = 1\}$ be all the datapoints that are sent to leaf 2 in w^k, and consider a partition J_L, J_R of J such that $\{x^j\}_{j \in J_L}$ and $\{x^j\}_{j \in J_R}$ are linearly separable, $J \cap \operatorname{supp}(f) \setminus \{i\} \subseteq J_L$, and $i \in J_R$. Let $\bar{w} \in \{0,1\}^N$ where $\bar{w}_j = 1$ if $j \in J_R$ and $\bar{w}_j = 0$ otherwise. Observe that $\bar{w} \notin H$, as $f^\top \bar{w} = f_i \neq 0$, thus the vectors $\bar{w}, \hat{w}^2, \ldots, \hat{w}^N$ are linearly independent. We define the $N \times N$ nonsingular matrix $\bar{L} = (\bar{w}, \hat{w}^2, \ldots, w^N)^\top$, obtained from L by replacing the first row with $\bar{w}$. We also define a vector $w' \in \{0,1\}^N$ where $w'_j = 1$ if $j \in [N] \setminus J$ and $w'_j = 0$ otherwise and the matrix $R' = (w', \mathbf{1} - \hat{w}^2, \ldots, \mathbf{1} - w^N)^\top$. We construct a $N(|\mathcal{L}|-1) \times N(|\mathcal{L}|-1)$ matrix M using as blocks $\bar{L}$, R', R and the $N \times N$ zero matrix as follows

$$M_{ij} = \begin{cases} \bar{L} & \text{if } i \in [|\mathcal{L}|/2 - 1], j = i \\ R' & \text{if } i \in [|\mathcal{L}|/2 - 1], j = |\mathcal{L}| - 1 \\ R & \text{if } i \in \{|\mathcal{L}|/2, \ldots, |\mathcal{L}| - 1\}, j = i \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Intuitively, for each routing vector $w^2, \ldots, w^N$ of the depth 1 decision tree, we construct $|\mathcal{L}| - 1$ routing vectors for the depth D decision tree as in Case 1. We construct $|\mathcal{L}|/2$ additional routings that send all the datapoints to one of the leaves of the subtree routed at node 3. Finally, we construct $|\mathcal{L}|/2 - 1$ routings that send $[N] \setminus J$ to the rightmost leaf of the tree, J_L to the leftmost leaf of the tree, and J_R to a leaf of the subtree routed at node 2 that is not the leftmost leaf. Similar to Case 1, by construction, each row $\tilde{w}$ of M is such that $\phi(\tilde{w}) \in W_1$ and $\phi(\tilde{w})$ satisfies (15) with equality. Moreover, M is nonsingular and, by Lemma l: aff ind bijection, we obtain $N(|\mathcal{L}|-1)$ affinely independent vectors in W_D that satisfy (15) with equality.

Case 3. By Lemma 8, we can equivalently prove that the symmetric inequality (16) is a facet for W_D. Because $\sum_{i=1}^N f_i - g = 0$, we reduce to Case 2. □

We finally prove that every facet-defining inequality of W_D is also facet-defining for the convex hull of the feasible solutions of the MIP formulation (3).

Theorem 4. *A facet-defining inequality of W_D is also facet-defining for the convex hull of the feasible set of* (3)*.*

Proof. Constraints (3c) imply that for each $t \in \mathcal{L}$, the subvector $(c_{kt})_{k \in [K]}$ lies in a $(K-1)$-dimensional simplex, denoted by Δ^t. The Cartesian product

$\Delta = \prod_{t \in \mathcal{L}} \Delta^t$ has thus dimension $|\mathcal{L}|(K-1)$. Moreover, constraints (3d) imply that each variable z_{it} is at most either 0 or 1. Let S denote the feasible set of (3) where c is relaxed to be continuous and nonnegative and z is relaxed to be continuous. For $u \in \{0,1\}$, let $S_u = S_D \times \Delta \times (-\infty, u]^{[N] \times \mathcal{L}}$. Recalling that $W_D = \text{conv}(S_D)$, have that $\text{conv}(S_u) = W_D \times \Delta \times (-\infty, u]^{[N] \times \mathcal{L}}$. Moreover, $S_0 \subseteq S \subseteq S_1$ implies $\text{conv}(S_0) \subseteq \text{conv}(S) \subseteq \text{conv}(S_1)$. Thus $\dim(\text{conv}(S)) = \dim(W_D) + |\mathcal{L}|(K-1) + N|\mathcal{L}|$.

Let $\tilde{F}$ be a facet of W_D, and consider the corresponding face $F = \{(w,c,z) \in \text{conv}(S) : w \in \tilde{F}\}$ of $\text{conv}(S)$. We have that $\tilde{F} \times \Delta \times (-\infty, 0]^{[N] \times \mathcal{L}} \subseteq F \subseteq \tilde{F} \times \Delta \times (-\infty, 1]^{[N] \times \mathcal{L}}$, thus $\dim(F) = \dim(\tilde{F}) + |\mathcal{L}|(K-1) + N|\mathcal{L}| = \dim(\text{conv}(S)) - 1$. Therefore, F is a facet of $\text{conv}(S)$. $\square$

References

1. Aghaei, S., Azizi, M.J., Vayanos, P.: Learning optimal and fair decision trees for non-discriminative decision-making (2019)
2. Aghaei, S., Gómez, A., Jo, N., Vayanos, P.: Learning optimal prescriptive trees from observational data (August 2021), https://optimization-online.org/?p=17313, Optimization Online
3. Aghaei, S., Gómez, A., Vayanos, P.: Strong optimal classification trees (January 2021), https://optimization-online.org/?p=16925, Optimization Online
4. Aglin, G., Nijssen, S., Schaus, P.: Learning optimal decision trees using caching branch-and-bound search. Proceedings of the AAAI Conference on Artificial Intelligence **34**(04), 3146–3153 (Apr 2020)
5. Aglin, G., Nijssen, S., Schaus, P.: Pydl8.5: a library for learning optimal decision trees. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 5222–5224. International Joint Conferences on Artificial Intelligence Organization (7 2020), demos
6. Avellaneda, F.: Efficient inference of optimal decision trees. Proceedings of the AAAI Conference on Artificial Intelligence **34**(04), 3195–3202 (Apr 2020)
7. Bertsimas, D., Dunn, J.: Optimal classification trees. Machine Learning **106** (07 2017)
8. Boutilier, J., Michini, C., Zhou, Z.: Shattering inequalities for learning optimal decision trees. In: 19th International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research, June 20-23, 2022, Los Angeles, CA, USA (2022)
9. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
10. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984)
11. Dash, S., Günlük, O., Wei, D.: Boolean decision rules via column generation (2020)
12. Demirović, E., Lukina, A., Hebrard, E., Chan, J., Bailey, J., Leckie, C., Ramamohanarao, K., Stuckey, P.J.: Murtree: Optimal classification trees via dynamic programming and search (2021)
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
14. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a “right to explanation”. AI magazine **38**(3), 50–57 (2017)
15. Gunluk, O., Kalagnanam, J., Li, M., Menickelly, M., Scheinberg, K.: Optimal generalized decision trees via integer programming (2019)
16. Hu, H., Siala, M., Hebrard, E., Huguet, M.J.: Learning optimal decision trees with maxsat and its integration in adaboost. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 1170–1176. International Joint Conferences on Artificial Intelligence Organization (7 2020)
17. Janota, M., Morgado, A.: Sat-based encodings for optimal decision trees with explicit paths. In: Pulina, L., Seidl, M. (eds.) Theory and Applications of Satisfiability Testing – SAT 2020. pp. 501–518. Springer International Publishing, Cham (2020)
18. Justin, N., Aghaei, S., Gómez, A., Vayanos, P.: Optimal robust classification trees. In: The AAAI-22 Workshop on Adversarial Machine Learning and Beyond (2022), https://openreview.net/forum?id=HbasA9ysA3

19. Lawless, C., Günlük, O.: Fair decision rules for binary classification (2021). https://doi.org/10.48550/ARXIV.2107.01325, https://arxiv.org/abs/2107.01325
20. Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. R news **2**(3), 18–22 (2002)
21. Lin, J.J., Zhong, C., Hu, D., Rudin, C., Seltzer, M.I.: Generalized and scalable optimal sparse decision trees. In: ICML (2020)
22. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences **116**(44), 22071–22080 (2019)
23. Narodytska, N., Ignatiev, A., Pereira, F., Marques-Silva, J.: Learning optimal decision trees with sat. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 1362–1368. International Joint Conferences on Artificial Intelligence Organization (7 2018)
24. Nijssen, S., Fromont, E.: Mining optimal decision trees from itemset lattices. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 530–539. KDD '07, Association for Computing Machinery, New York, NY, USA (2007)
25. Schidler, A., Szeider, S.: Sat-based decision tree learning for large data sets. Proceedings of the AAAI Conference on Artificial Intelligence **35**(5), 3904–3912 (May 2021)
26. Vapnik, V.: Statistical learning theory. Wiley (1998)
27. Verhaeghe, H., Nijssen, S., Pesant, G., Quimper, C.G., Schaus, P.: Learning optimal decision trees using constraint programming. Constraints **25**, 1–25 (12 2020). https://doi.org/10.1007/s10601-020-09312-3
28. Verhaeghe, H., Nijssen, S., Pesant, G., Quimper, C.G., Schaus, P.: Learning optimal decision trees using constraint programming (extended abstract). In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 4765–4769. International Joint Conferences on Artificial Intelligence Organization (7 2020)
29. Verwer, S., Zhang, Y.: Learning optimal classification trees using a binary linear program formulation. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). pp. 1625–1632. AAAI Press (2019), 33rd AAAI Conference on Artificial Intelligence, AAAI-19 ; Conference date: 27-01-2019 Through 01-02-2019
30. Verwer, S., Zhang, Y.: Learning decision trees with flexible constraints and objectives using integer optimization. In: Salvagnin, D., Lombardi, M. (eds.) Integration of AI and OR Techniques in Constraint Programming. pp. 94–103. Springer International Publishing, Cham (2017)
31. Zhu, H., Murali, P., Phan, D.T., Nguyen, L.M., Kalagnanam, J.: A scalable mip-based method for learning optimal multivariate decision trees. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)