Supplemental Material for "Classification Trees for Imbalanced and Sparse Data: Surface-to-Volume Regularization"

Yichen Zhu
Department of Statistical Science, Duke University
and
David B. Dunson
Department of Statistical Science, Duke University

April 26, 2020

The supplementary material contains proofs and algorithms that are not included in the main paper. Without loss of generality, we assume $\Omega = [0,1]^d$ in all proofs. We use $\mathbb{E}_\alpha$, $\mathbb{E}_{n,\alpha}$ to denote the expectation over probability measures $\mathbb{P}_\alpha$, $\mathbb{P}_{n,\alpha}$, respectively. For all $A \subset \Omega$, we let $p_\alpha(A) = \mathbb{E}_\alpha(Y|X \in A)$ and $p_\alpha(A) = \mathbb{E}_{n,\alpha}(Y|X \in A)$.

1 Proof of Lemma 1-4

1.1 Proof of Lemma 1

Proof. Decompose the difference between $\tilde{I}_\alpha(T_n, \mathbb{P})$ and I^*_α as:

$$\tilde{I}_\alpha(T_n, \mathbb{P}) - I^*_\alpha = \tilde{I}_\alpha(T_n, \mathbb{P}) - I_\alpha(T_n, \mathbb{P}) + I_\alpha(T_n, \mathbb{P}) - I^*_\alpha.$$

The first term is nonnegative by definition. Now we show the second term is also nonnegative. Let $A_1, A_2, \dots A_m$ be all the leaf nodes of tree T_n. Then the impurity can be computed as

$$I_\alpha(T_n, \mathbb{P}) = \sum_{l=1}^{m} 2 \frac{\int_{A_l} p_\alpha(x) d\mathbb{P}_\alpha(x) \cdot \int_{A_l} (1 - p_\alpha(x)) d\mathbb{P}_\alpha(x)}{\mathbb{P}_\alpha(A_l)}.$$

Thus we have

$$\begin{aligned}
I_\alpha(T_n, \mathbb{P}) - I^*_\alpha &= \sum_{l=1}^{m} \left[2 \frac{\int_{A_l} p_\alpha(x) d\mathbb{P}_\alpha(x) \cdot \int_{A_l} (1 - p_\alpha(x)) d\mathbb{P}_\alpha(x)}{\mathbb{P}_\alpha(A_l)} - \int_{A_l} 2 p_\alpha(x)(1 - p_\alpha(x)) d\mathbb{P}_\alpha(x) \right] \\
&= \sum_{l=1}^{m} \frac{2}{\mathbb{P}_\alpha(A_l)} \left[\int_{A_l} p_\alpha(x) d\mathbb{P}_\alpha(x) \cdot \int_{A_l} (1 - p_\alpha(x)) d\mathbb{P}_\alpha(x) - \int_{A_l} p_\alpha(x)(1 - p_\alpha(x)) d\mathbb{P}_\alpha(x) \cdot \mathbb{P}_\alpha(A_l) \right] \\
&= \sum_{l=1}^{m} \frac{2}{\mathbb{P}_\alpha(A_l)} \left[-\Big(\int_{A_l} p_\alpha(x) d\mathbb{P}_\alpha(x) \Big)^2 + \int_{A_l} p^2_\alpha(x) d\mathbb{P}_\alpha(x) \cdot \mathbb{P}_\alpha(A_l) \right] \geq 0,
\end{aligned}$$

where the last inequality is obtained by Jensen's inequality. Therefore, $\tilde{I}_\alpha(T_n, \mathbb{P}) \to I^*_\alpha$ in probability implies

$$\tilde{I}_\alpha(T_n, \mathbb{P}) - I_\alpha(T_n, \mathbb{P}) \to 0 \text{ in probability and} \tag{1}$$

$$I_\alpha(T_n, \mathbb{P}) - I^*_\alpha \to 0 \text{ in probability.} \tag{2}$$

We first consider (2). Denote $\psi_n(x)$ as the leaf node of T_n that contains x. Define $\bar{p}_{n,\alpha}(x)$ as

$$\bar{p}_{n,\alpha}(x) = \frac{\int_{\psi_n(x)} p_\alpha(t) d\mathbb{P}_\alpha(t)}{\mathbb{P}_\alpha(\psi_n(x))},$$

the average value of $p_\alpha(x)$ at the leaf node of T_n that contains x. Then we can rewrite $I_\alpha(T_n, \mathbb{P}) - I_\alpha^*$ as

$$I_\alpha(T_n, \mathbb{P}) - I_\alpha^* = \sum_{l=1}^{m} 2 \int_{A_l} (p_\alpha^2(x) - \bar{p}_{n,\alpha}^2(x)) \mathbb{P}_\alpha(x) = 2 \int_\Omega (p_\alpha^2(x) - \bar{p}_{n,\alpha}^2(x)) \mathbb{P}_\alpha(x).$$

Therefore (2) implies $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(X \in \Omega : |p_\alpha(X) - \bar{p}_{\alpha,n}(X)| > \epsilon\right) = 0.$$

For any $\epsilon > 0$, denote $A_\epsilon \subset \Omega$ as: $A_\epsilon = \{X \in \Omega : |p_\alpha(X) - 1/2| > \epsilon\}$. Then we have

$$\lim_{n\to\infty} \mathbb{P}\left(\{|p_\alpha(X) - \bar{p}_{\alpha,n}(X)| \le \epsilon\} \cap \ A_{2\epsilon}\right) = \mathbb{P}(A_{2\epsilon}). \tag{3}$$

Define a classifier: $\tilde{f}_n : \Omega \to \{0, 1\}$ such that:

$$\tilde{f}_n(x) = \begin{cases} 1, & \mathbb{P}(Y = 1 | X \in \psi_n(x)) \ge 1/(1+\alpha), \\ 0, & \text{otherwise.} \end{cases}$$

That is, $\tilde{f}_n$ achieves the minimal error $R(\tilde{f}_n)$ among all the classifiers that are piecewise constant on all leaf nodes of T_n. Because $|p_\alpha(X) - 1/2| > 2\epsilon$, $|p_\alpha(X) - \bar{p}_{\alpha,n}(X)| \le \epsilon$ implies: 1. $p_\alpha(X) - 1/2$ has the same sign as $\bar{p}_{\alpha,n}(X) - 1/2$; 2. $|\bar{p}_{\alpha,n}(X) - 1/2| \ge \epsilon$. So by (3), we have for any oracle classifier $f^* \in F^*$,

$$\lim_{n\to\infty} \mathbb{P}\left(\{\tilde{f}_n(X) = f^*(X)\} \cap \{|\bar{p}_{\alpha,n}(X) - 1/2| \ge \epsilon\} \cap A_{2\epsilon}\right) = \mathbb{P}(A_{2\epsilon}). \tag{4}$$

Denote the $B_\epsilon = \{\tilde{f}(X) = f^*(X)\} \cap \{|\bar{p}_{\alpha,n}(X) - 1/2| \ge \epsilon\} \cap A_{2\epsilon}$. We then consider (1). We have

$$\begin{aligned} \tilde{I}_\alpha(T_n, \mathbb{P}) - I_\alpha(T_n, \mathbb{P}) &= \int_\Omega |1 - 4\bar{p}_{\alpha,n}(x)(1 - \bar{p}_{\alpha,n}(x))| \mathbb{1}_{\{f_n(x) \ne \tilde{f}_n(x)\}} d\mathbb{P}_\alpha(x) \\ &\ge \int_{B_\epsilon} |1 - 4\bar{p}_{\alpha,n}(x)(1 - \bar{p}_{\alpha,n}(x))| \mathbb{1}_{\{f_n(x) \ne \tilde{f}_n(x)\}} d\mathbb{P}_\alpha(x) \\ &\ge \int_{B_\epsilon} 4\epsilon^2 \mathbb{1}_{\{f_n(x) \ne f^*(x)\}} d\mathbb{P}_\alpha(x) \end{aligned} \tag{5}$$

Combining (1) and (5), we have

$$\lim_{n\to\infty} \mathbb{P}(B_\epsilon \cap \{f_n(X) \ne f^*(X)\}) = 0. \tag{6}$$

Combining equation (4) and (6), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\{f_n(X) \ne f^*(X)\} \cap A_{2\epsilon}\right) = 0.$$

Since ϵ is arbitrary, and $\lim_{n\to\infty} \mathbb{P}(A_{2\epsilon}) = \mathbb{P}(\{\mathbb{E}(Y|X) \ne 1/(1+\alpha)\}) = 1$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\{f_n(X) \ne f^*(X)\}\right) = 0.$$

Therefore

$$\lim_{n\to\infty} \mathbb{E}|f_n - f^*| = 0.$$

□

1.2 Proof of Lemma 2

Proof. Since $\mathscr{T}_n$ is a finite set, we can find $\tilde{T}_n \in \mathscr{F}_n$ satisfying

$$|\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) + \lambda_n r(\tilde{T}_n) - I_\alpha^*| = \inf_{T \in \mathscr{T}_n} |\tilde{I}_\alpha(T, \mathbb{P}) + \lambda_n r(T) - I_\alpha^*|.$$

Therefore, we have

$$\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}) + \lambda_n r(\hat{T}_n) - I_\alpha^* = [\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}) + \lambda_n r(\hat{T}_n)] - [\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) + \lambda_n r(\tilde{T}_n)] + [\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) + \lambda_n r(\tilde{T}_n)] - I_\alpha^*.$$

For the first two terms,

$$\begin{aligned}
[\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}) + \lambda_n r(\hat{T}_n)] - [\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) + \lambda_n r(\tilde{T}_n)] =& [\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}) + \lambda_n r(\hat{T}_n)] - [\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}_n) + \lambda_n r(\hat{T}_n)] \\
&+ [\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}_n) + \lambda_n r(\hat{T}_n)] - [\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}_n) + \lambda_n r(\tilde{T}_n)] \\
&+ [\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}_n) + \lambda_n r(\tilde{T}_n)] - [\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) + \lambda_n r(\tilde{T}_n)] \\
\leq& 2 \sup_{T \in \mathscr{T}_n} |\tilde{I}_\alpha(T, \mathbb{P}) - \tilde{I}_\alpha(T, \mathbb{P}_n)|,
\end{aligned}$$

where we use the fact that $\hat{T}_n$ is the minimizer of $\tilde{I}(T, \mathbb{P}_n) + \lambda_n(T)$ for all $T \in \mathscr{T}_n$. Recalling the definition of $\tilde{T}_n$, by the triangle inequality, we have

$$\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}) + \lambda_n(\hat{T}_n) - I_\alpha^* \leq 2 \sup_{T \in \mathscr{T}_n} |\tilde{I}_\alpha(T, \mathbb{P}) - \tilde{I}_\alpha(T, \mathbb{P}_n)| + \inf_{T \in \mathscr{T}_n} |\tilde{I}_\alpha(T, \mathbb{P}) + \lambda_n r(T) - I_\alpha^*|.$$

Since $\lambda_n r(\hat{T}_n) \geq 0$, we have

$$\tilde{I}_\alpha(\hat{T}_n, \mathbb{P}) - I_\alpha^* \leq 2 \sup_{T \in \mathscr{T}_n} |\tilde{I}_\alpha(T, \mathbb{P}) - \tilde{I}_\alpha(T, \mathbb{P}_n)| + \inf_{T \in \mathscr{T}_n} |\tilde{I}_\alpha(T, \mathbb{P}) + \lambda_n r(T) - I_\alpha^*|.$$

□

1.3 Proof of Lemma 3

Let $A_1, A_2, \ldots A_m (m \leq \bar{a}_n)$ be all the leaf nodes of T. Define $p_\alpha(A_j) = \mathbb{E}_\alpha(Y|X \in A_j)$, $p_{n,\alpha}(A_j) = \mathbb{E}_\alpha(Y|X \in A_j)$. We first introduce a technical lemma which says the maximal difference of $|p_\alpha(A_j) - p_{n,\alpha}(A_j)|$ over all leaf nodes goes to zero in probability.

Technical Lemma 1. *If $\frac{\bar{a}_n d \log n}{n} = o(1)$, then $\forall \epsilon > 0$, regardless whether optional steps of Algorithm 1 are enabled, we have*

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{1 \leq j \leq m} |\mathbb{P}_\alpha(A_j) - \mathbb{P}_{n,\alpha}(A_j)| > \epsilon \right) = 0,$$

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{1 \leq j \leq m} |p_\alpha(A_j) - p_{n,\alpha}(A_j)| > \epsilon \right) = 0.$$

Proof. This technical lemma is a special case of Lemma 3 of Nobel (1996). We let the feature space be $\Omega_0 = \Omega \times \{0, 1\} \subset \mathbb{R}^{d+1}$. Let Π_n be the partitions that agree with tree T_n in the first d dimensions and do not partition on the last dimension. Define three real valued functions $g_0, g_1, g_2 : \Omega \to \mathbb{R}$ such that $g_0(x) = 0, g_1(x) = 1, g_2(x) = x_{d+1}, \forall x \in \Omega_0 \subset \mathbb{R}^{d+1}$. Let $\mathscr{G} = \{g_0, g_1, g_2\}$. It suffices to verify two conditions in Lemma 3 of Nobel (1996): $m(\Pi_n : V) = m \leq \bar{a}_n = o(n)$, $\log \Delta_n^*(\Pi_n) = \log(n^{(m-1)(d+1)}) \leq \log(n^{\bar{a}_n(d+1)}) = o(n)$. □

Proof of Lemma 3. For all $\epsilon \in (0, 1)$, by Technical Lemma 1, the event

$$H_0 = \left\{ \sup_{1 \leq j \leq m} |\mathbb{P}_\alpha(A_j) - \mathbb{P}_{n,\alpha}(A_j)| \leq \epsilon, \sup_{1 \leq j \leq m} |p_\alpha(A_j) - p_{n,\alpha}(A_j)| \leq \epsilon \right\}$$

holds with probability tending to one. Therefore it suffices to prove $\sup_{T\in\mathscr{T}_n}|\tilde{I}_\alpha(T,\mathbb{P})-\tilde{I}_\alpha(T,\mathbb{P}_n)|\to 0$ under event H_0. For all $T\in\mathscr{T}_n$, define two collection of leaf nodes $\mathscr{A}$, $\mathscr{B}$ as:

$$\mathscr{A}=\{A_j,1\le j\le m:\tilde{I}_\alpha(A_j,\mathbb{P})-\tilde{I}_\alpha(A_j,\mathbb{P}_n)=I_\alpha(A_j,\mathbb{P})-I_\alpha(A_j,\mathbb{P}_n)\},$$

$$\mathscr{B}=\{A_j,1\le j\le m:|\tilde{I}_\alpha(A_j,\mathbb{P})-\tilde{I}_\alpha(A_j,\mathbb{P}_n)|=|1-I_\alpha(A_j,\mathbb{P})-I_\alpha(A_j,\mathbb{P}_n)|\}.$$

That is, $\mathscr{A}$ contains all leaf nodes where both $p_\alpha(A_j)$ and $p_{n,\alpha}(A_j)$ are no greater than $1/2$ (or both $p_\alpha(A_j)$ and $p_{n,\alpha}(A_j)$ are no less than $1/2$) while $\mathscr{B}$ contains all leaf nodes where one and only one of $p_\alpha(A_j)$ and $p_{n,\alpha}(A_j)$ is less than $1/2$. For all $A_j\in\mathscr{A}$, we have

$$\begin{aligned}|\tilde{I}_\alpha(A_j,\mathbb{P})-\tilde{I}_\alpha(A_j,\mathbb{P}_n)|&=|I_\alpha(A_j,\mathbb{P})-I_\alpha(A_j,\mathbb{P}_n)\}|\\&=|2p_\alpha(A_j)(1-p_\alpha(A_j))-2p_{n,\alpha}(A_j)(1-p_{n,\alpha}(A_j))|\\&=2|(p_\alpha(A_j)-p_{n,\alpha}(A_j))(1-p_\alpha(A_j)-p_{n,\alpha}(A_j))|\le 2(\epsilon+\epsilon^2)\end{aligned}$$

For all $A_j\in\mathscr{B}$, since $p_{n,\alpha}(A_j)<1/2<p_\alpha(A_j)$ or $p_\alpha(A_j)<1/2<p_{n,\alpha}(A_j)$, recalling $|p_{n,\alpha}(A_j)-p_\alpha(A_j)|<\epsilon$, we have $|p_{n,\alpha}(A_j)-1/2|<\epsilon$ and $|p_\alpha(A_j)-1/2|<\epsilon$. Therefore

$$\begin{aligned}|\tilde{I}_\alpha(A_j,\mathbb{P})-\tilde{I}_\alpha(A_j,\mathbb{P}_n)|&=|1-I_\alpha(A_j,\mathbb{P})-I_\alpha(A_j,\mathbb{P}_n)|\\&\le|1/2-I_\alpha(A_j,\mathbb{P})|+|1/2-I_\alpha(A_j,\mathbb{P}_n)|\\&\le|1/2-2p_\alpha(A_j)(1-p_\alpha(A_j)|+|1/2-2p_{n,\alpha}(A_j)(1-p_{n,\alpha}(A_j))|\\&=\Big|1/2-2[1/2-(1/2-p_\alpha(A_j))][1/2+(1/2-p_\alpha(A_j)]\Big|\\&+\Big|1/2-2[1/2-(1/2-p_{n,\alpha}(A_j))][1/2+(1/2-p_{n,\alpha}(A_j)]\Big|\\&\le 4\epsilon^2.\end{aligned}$$

We can finally compute the difference of two signed tree impurity as

$$\begin{aligned}|\tilde{I}_\alpha(T,\mathbb{P})-\tilde{I}_\alpha(T,\mathbb{P}_n)|\le&|\sum_j\mathbb{P}_\alpha(A_j)\tilde{I}_\alpha(A_j,\mathbb{P})-\sum_j\mathbb{P}_{n,\alpha}(A_j)\tilde{I}_\alpha(A_j,\mathbb{P}_n)|\\\le&\sum_j[\mathbb{P}_\alpha(A_j)|\tilde{I}_\alpha(A_j,\mathbb{P})-\tilde{I}_\alpha(A_j,\mathbb{P}_n)|+|\mathbb{P}_\alpha(A_j)-\mathbb{P}_{n,\alpha}(A_j)|]\\\le&m[\max(2\epsilon+2\epsilon^2,4\epsilon^2)+\epsilon]\\\le&(3+2\epsilon^2)m\epsilon.\end{aligned}$$

Since the above equation holds for all $T\in\mathscr{T}_n$ and $\epsilon\in(0,1)$, $\sup_{T\in\mathscr{T}_n}|\tilde{I}_\alpha(T,\mathbb{P})-\tilde{I}_\alpha(T,\mathbb{P}_n)|$ goes to zero in probability. □

1.4 Proof of Lemma 4

It is not easy to directly compute the approximation error, hence we introduce "theoretical tree" as a bridge connecting our estimator $\hat{T}_n$ and the oracle lower bound T^*_α. Algorithm 2 describes how we construct theoretical trees.

Algorithm 2: Steps for building the theoretical tree

Result: Output the fitted tree

Input distribution $\mathbb{P}$, impurity function $f(\cdot)$, weight for minority class α, and maximal number of leaf nodes $k \in \mathbb{N}$. Set SVR penalty parameter $\lambda = 0$. Let the root node be Ω, and node.X $= \Omega$;

while node_queue *is not empty and number of leaf nodes* $\leq k$ **do**

  Dequeue the first entity in node_queue, denoting it as node;

  for j *in* $1 : d$ **do**

    Find the best partition $\hat{x}_j$ and its corresponding class label assignments inside the current node, such that if we divide node into two nodes $X[j] \leq \hat{x}$ and $X[j] > \hat{x}$ and assign class labels to two left, right child node as lab_j^l, lab_j^r, tree impurity is minimized;

  end

  Find the best $\hat{x}_j$ among $\hat{x}_1, \hat{x}_2, \cdots \hat{x}_d$ that minimizes tree impurity. Denote it as $\hat{x}$, its class label assignments for left and right child node as lab^l, lab^r, respectively;

  if *tree impurity is decreased* **then**

    Let node.left.X $= \{X \in \text{node} : X[j] \leq \hat{x}\}$, and node.right.X $= \{X \in \text{node} : X[j] > \hat{x}\}$. Assign node.Y to node.left, and node.right according to the assignment of node.X. Let the class label of node.right and node.left be lab^l, lab^r, respectively. Enqueue node.left, node.right to the end of node_queue;

  else

    Reject the partition;

  end

end

Theoretical trees are computed with SVR penalty parameter $\lambda = 0$, because no regularization is needed if we know the true distribution. Consequently, the theoretical tree always "correctly" assigns class labels, resulting in $I_\alpha(T_k^*, \mathbb{P}) = \tilde{I}_\alpha(T_k^*, \mathbb{P})$. For simplicity, in all the following proofs, we assume for each $k \in \mathbb{N}$, the theoretical tree T_k^* is unique. Our proofs can be easily generalized to the case that there exist multiple theoretical trees, where the distance between SVR-Tree and theoretical tree T_k^* is replaced with the infimum between SVR-Tree and all theoretical trees with k leaf nodes.

The proof of Lemma 4 is mainly built on three technical lemmas. The first one shows the sequence of theoretical trees is consistent, the second one shows two trees with similar structures are also close in impurity, and the last one shows as n goes to infinity, theoretical trees and our estimated tree are close in their partitions structures. We begin with the consistency of theoretical trees.

Technical Lemma 2. *The sequence of theoretical trees* $T_1^*, T_2^*, \ldots$ *satisfies*

$$\lim_{k\to\infty} \tilde{I}_\alpha(T_k^*, \mathbb{P}) = I_\alpha^*.$$

Proof. $\forall x \in \Omega$, let $\psi_k(x)$ be the leaf node of T_k^* that contains x, then we have $\forall k_1 < k_2$, $\psi_{k_2}(x) \subset \psi_{k_1}(x)$. Define $\eta_k(x)$ as the maximal variation of $\mathbb{E}(Y|X)$ in $\psi_k(x)$:

$$\eta_k(x) = \sup_{x\in\psi_k(x)} \mathbb{E}(Y|X = x) - \inf_{x\in\psi_k(x)} \mathbb{E}(Y|X = x).$$

By Lemma 1 in Scornet et al. (2015), we have

$$\lim_{k\to\infty} \eta_k(X) = 0, \;\; a.s.$$

Therefore, for all $\epsilon \in (0, 1)$, $\exists K \in \mathbb{N}$, $\forall k > K$, $\mathbb{P}(\eta_k(x) > \epsilon) < \epsilon$. Let $X_0 = \{x \in X : \eta_k(x) \leq \epsilon\}$. Define $q_{k,\alpha}(x) = \mathbb{E}_\alpha(Y|\psi_k(x))$. By the definition of a theoretical tree, we have

$$\begin{aligned} |I_\alpha(T_k^*, \mathbb{P}) - I_\alpha^*| =& \Big(\int_{X_0} + \int_{X_0^c}\Big)|2q_{k,\alpha}(x)(1 - q_{k,\alpha}(x)) - 2p_\alpha(x)(1 - p_\alpha(x))|d\mathbb{P}_\alpha(x) \\ \leq& \int_{X_0} |2q_{k,\alpha}(x)(1 - q_{k,\alpha}(x)) - 2p_\alpha(x)(1 - p_\alpha(x))|d\mathbb{P}_\alpha(x) + \alpha\epsilon \end{aligned}$$

Since $\eta_k(x) \le \epsilon$ for $x \in X_0$, we have $|q_{k,\alpha}(x) - p_\alpha(x)| \le \alpha\eta_k(x) \le \alpha\epsilon$. Thus

$$\begin{aligned}|I_\alpha(T_k^*, \mathbb{P}) - I_\alpha^*| \le& 2\int_{X_0} |[q_{k,\alpha}(x) - p_\alpha(x)][1 - q_{k,\alpha}(x) - p_\alpha(x)]| d\mathbb{P}_\alpha(x) + \alpha\epsilon \\ \le& 2 \cdot \alpha\epsilon \cdot 1 + \alpha\epsilon = 3\alpha\epsilon.\end{aligned}$$

Because $\tilde{I}_\alpha(T_k^*, \mathbb{P}) = I_\alpha(T_k^*, \mathbb{P})$ for theoretical trees, we have $|\tilde{I}_\alpha(T_k^*, \mathbb{P}) - I_\alpha^*| < 3\alpha\epsilon$ for all $k > K$. Since ϵ is arbitrary, we finish the proof. □

We now consider the relation between tree structures and tree impurity.

Technical Lemma 3. *Let T, T' be two trees both having k leaf nodes and $\mathbb{P}$ be a probability measure. Denote all the leaf nodes of T as $A_1, A_2, \dots A_k$, all the leaf nodes of T' as $A_1', A_2', \dots A_k'$. Then if*

$$\sup_{1 \le j \le k} \mathbb{P}(A_j \Delta A_j') \le \epsilon$$

we have $|I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| \le 5\alpha k\epsilon$.

Proof. We first consider nodes A_1 and A_1'.

$$\begin{aligned}|I_\alpha(A_1, \mathbb{P}) - I_\alpha(A_1 \cup A_1', \mathbb{P})| =& |2p_\alpha(A_1)(1 - p_\alpha(A_1)) - 2p_\alpha(A_1 \cup A_1')(1 - p_\alpha(A_1 \cup A'))| \\ =& 2|(p_\alpha(A_1) - p_\alpha(A_1 \cup A_1'))(1 - p_\alpha(A_1) - p_\alpha(A_1 \cup A_1'))| \\ \le& 2|p_\alpha(A_1) - p_\alpha(A_1 \cup A_1')| \\ \le& 2\frac{\mathbb{P}_\alpha(A_1 \Delta (A_1 \cup A_1'))}{\mathbb{P}_\alpha(A_1)} \le \frac{2\alpha\epsilon}{\mathbb{P}_\alpha(A_1)}.\end{aligned}$$

Similarly, we have $|I_\alpha(A_1', \mathbb{P}) - I_\alpha(A_1 \cup A_1', \mathbb{P})| \le 2\alpha\epsilon / \mathbb{P}_\alpha(A_1')$. Therefore

$$\begin{aligned}|I_\alpha(A_1, \mathbb{P})\mathbb{P}_\alpha(A_1) - I_\alpha(A_1', \mathbb{P})\mathbb{P}_\alpha(A_1')| =& \Big|[I_\alpha(A_1, \mathbb{P}) - I_\alpha(A_1 \cup A_1', \mathbb{P})]\mathbb{P}_\alpha(A_1) - \\ & [I_\alpha(A_1', \mathbb{P}) - I_\alpha(A_1 \cup A_1', \mathbb{P})]\mathbb{P}_\alpha(A_1') + [\mathbb{P}_\alpha(A_1) - \mathbb{P}_\alpha(A_1')]I_\alpha(A_1 \cup A_1', \mathbb{P})\Big| \\ \le& 2\alpha\epsilon + 2\alpha\epsilon + \alpha\epsilon \le 5\alpha\epsilon.\end{aligned}$$

Therefore, the difference in tree impurity can be computed as

$$\begin{aligned}|I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| =& |\sum_{l=1}^k [I_\alpha(A_l, \mathbb{P})\mathbb{P}_\alpha(A_l) - I(A_l', \mathbb{P})\mathbb{P}_\alpha(A_l')]| \\ \le& \sum_{l=1}^k |[I_\alpha(A_l, \mathbb{P})\mathbb{P}_\alpha(A_l) - I(A_l', \mathbb{P})\mathbb{P}_\alpha(A_l')]| \le 5k\alpha\epsilon\end{aligned}$$

□

We then define a generalized distance metric[1] for partition of trees and discuss its properties. For any classification tree T built by algorithm 1 or 2, let $s(T) = ((x_1, j_1), (x_2, j_2), \dots, (x_k, j_k))$ denote all the partitions of T. That is, T is obtained by first partitioning Ω at $X[j_1] = x_1$, then partitioning the left child of the root at $X[j_2] = x_2$, the right child of the root at $X[j_3] = x_3$ and so on. If no partition is accepted at the lth step, we let $x_l = 0, j_l = 0$. For two possible partitions (x_l, j_l) and (x_l', j_l') at the same step l, we define their distance as

$$D((x_l, j_l), (x_l', j_l')) = \max(|x_l - x_l'|, |j_l - j_l'|).$$

Recall that we let the sample space $\Omega = [0, 1]^d$, so $D((x_l, j_l), (x_l', j_l')) \ge 1$ if and only if $j_l \ne j_l'$, which means either they partition at different features, or one of the j_l, j_l' indicates no partition is accepted at that

[1] In fact, $D(\cdot, \cdot)$ is a metric on space $\{s(T) : T \in \mathscr{F}_k\}$. As a result, $D(s(\cdot), s(\cdot))$ becomes a generalized metric on $\mathscr{F}_k$ where all trees share the same partition structures are treated equally in terms of metric.

step. If two trees T, T' are both obtained with $k-1, (k \in \mathbb{N})$ steps of partitions, we define the distance of $s(T) = ((x_1, j_1), (x_2, j_2), \ldots, (x_k, j_k)), s(T) = ((x'_1, j'_1), (x'_2, j'_2), \ldots, (x'_k, j'_k))$ as

$$D(s(T_1), s(T_2)) = \max_{1 \le i \le k-1} D((x_i, j_i), (x'_i, j'_i)).$$

Let $\hat{T}_{n,k}$ be the tree obtained by only the first $k-1$ partitions of SVR-Tree $\hat{T}_n$. The following technical lemma shows the distance between $\hat{T}_{n,k}$ and the theoretical tree T_k^* goes to zero as n goes to infinity.

Technical Lemma 4. *For all $k \in \mathbb{N}$ and $\delta > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(D(s(\hat{T}_{n,k}), s(T_k^*)) > \delta \right) = 0$$

Proof. We first prove this technical lemma with optional steps in Algorithm 1 disabled. The proof can be broken into 3 steps. In the first step, we establish the uniform continuity of $I_\alpha(T, \mathbb{P})$ and $I_\alpha(T, \mathbb{P}_n)$ with respect to the generalized distance metric $D(s(\cdot), s(\cdot))$; In the second step, we show $I_\alpha(T, \mathbb{P}_n)$ and $I_\alpha(T, \mathbb{P})$ are very close as n goes to infinity; The final step utilizes the optimality of $\hat{T}_{n,k}$ in each partition step and finishes the proof. In the end of the proof, we show the case with optional steps enabled are essentially the same.

Step 1 In this step, we will show for all $\epsilon > 0$, $\sigma > 0$, $k \in \mathbb{N}$, there exists $\delta \in (0, 1)$, such that for any two trees T, T' obtained with $k-1$ partitions, if $D(s(T), s(T')) < \sigma$, then

$$|I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| < \epsilon$$

and

$$\mathbb{P}\left(\sup_{\substack{T, T' \in \mathscr{T}_k \\ D(s(T), s(T')) < \delta}} |I_\alpha(T, \mathbb{P}_n) - I_\alpha(T', \mathbb{P}_n)| < \epsilon \right) \ge 1 - \sigma.$$

We assume both T, T' have k leaf nodes, i.e., the partition of each step is accepted. If one of T, T' does not have k leaf nodes, then some partitions are not accepted. Because $D(s(T), s(T')) < \delta < 1$, a partition of T is accepted if and only if the corresponding partition of T' is accepted. So T, T' still have the same number of leaf nodes, but the number of leaf nodes is smaller than k, which still follows the same proof.

We first consider $|I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})|$. Because the marginal probability measure of $\mathbb{P}$ on Ω is absolutely continuous with respect to the Lebesgue measure μ, for all $\epsilon_0 > 0$, there exists $\delta_0 \in (0, 1)$, for any Borel set $A \subset \Omega$ with $\mu(A) < \delta_0$, we have $\mathbb{P}(A) < \epsilon_0$. Let $\delta = \delta_0 / d$. Because $D(s(T), s(T')) < \delta$, denote all the leaf nodes of T as $A_1, A_2, \ldots A_k$, there exists an order of all the leaf nodes of T: $A'_1, A'_2, \ldots, A'_k$, such that $\forall 1 \le l \le k$, $\mu(A_l \Delta A'_l) < d\delta = \delta_0$. Therefore $\mathbb{P}(A_l \Delta A'_l) < \epsilon_0$, $\forall 1 \le l \le k$. By Technical Lemma 3, we have

$$|I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| < 5\alpha k \epsilon_0.$$

Since ϵ_0 is arbitrary, let $\epsilon_0 = \epsilon / (5\alpha k)$, we have $|I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| < \epsilon$.

We then consider $|I_\alpha(T, \mathbb{P}_n) - I_\alpha(T', \mathbb{P}_n)|$, where we still show the condition of Technical Lemma 3 is satisfied, but with probability greater than $1 - \sigma$. By absolute continuity of the marginal of $\mathbb{P}$, for all $\epsilon_0 > 0$, there exists δ_0, for any Borel set A with Lebesgue measure $\mu(A) < \delta_0$, such that we have $\mathbb{P}(A) < \epsilon_0$. Let $m = \lceil \frac{1}{\delta_0} \rceil$. Then we define a collection of subsets of Ω as

$$\mathscr{A} = \left\{ [0,1] \times [0,1] \times \ldots \times \underset{j}{[l/m, (l+1)/m]} \times \ldots \times [0,1] \middle| j, l \in \mathbb{N}, 1 \le j \le d, 1 \le l \le m-1 \right\}.$$

That is, each $A \in \mathscr{A}$ is a hyperrectangle with jth dimension equaling to interval $[[l/m, (l+1)/m]$ and the other $d-1$ dimensions equaling to $[0, 1]$. The set $\mathscr{A}$ consists of all such hyperrectangles where j varies from 1 to d and l varies from 0 to $m-1$. Since m is usually very large, $\mathscr{A}$ can be viewed as the collection of all $d-1$-dimensional regularly spaced "slabs" with thickness $1/m$. These "slabs" will help us control the difference of two corresponding leaf nodes of T and T'.

Because there are finitely many elements in $\mathscr{A}$, and for each $A \in \mathscr{A}$, $\mathbb{P}_n(A)$ converges to $\mathbb{P}(A)$ almost surely, we have for the prementioned ϵ_0 and all $\sigma > 0$, there exists $N \in \mathbb{N}$, such that $\forall n > N_1$,

$$\mathbb{P}\left(\sup_{A\in\mathscr{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| \geq \epsilon_0\right) < \sigma, \tag{7}$$

where the outer probability is taking over the distribution of $\mathscr{D}_n$. Combine equation (7) and definition of $\mathscr{A}$, we have

$$\mathbb{P}(\sup_{A\in\mathscr{A}} \mathbb{P}_n(A) \geq 2\epsilon_0) < \sigma.$$

Define an event $\mathcal{E} = \{\sup_{A\in\mathscr{A}} \mathbb{P}_n(A) < 2\epsilon_0\}$. Then $\mathbb{P}(\mathcal{E}) \geq 1 - \sigma$. We then consider a larger collection of "slabs". Denote the set $\mathscr{B}$ as:

$$\mathscr{B} = \left\{ [0,1] \times [0,1] \times \ldots \times \underset{j}{[b_1, b_2]} \times \ldots \times [0,1] \middle| j \in \mathbb{N}, 1 \leq j \leq d, b_2 - b_1 \leq 1/m \right\}.$$

The collection $\mathscr{B}$ consists of all the "slabs" whose thickness is no more than $1/m$. By definition, we have $\forall B \in \mathscr{B}$, $\exists A_1, A_2 \in \mathscr{A}$, such that $B \subset A_1 \cup A_2$. Let $\delta = 1/m \leq \delta_0$. For all T, T' obtained by $k-1$ partitions and satisfying $D(s(T), s(T')) < \delta$, denoting all leaf nodes of T as $A_1, A_2, \ldots, A_k$, there exists an order of all the leaf nodes of T': $A_1', A_2', \ldots A_k'$, such that $\forall 1 \leq l \leq k$, the symmetrical set difference between A_1, A_1' is contained in the union of d "slabs" in the set $\mathscr{B}$. Formally,

$$A_l \Delta A_l' \subset \bigcup_{j=1}^{d} B_j, \quad \forall 1 \leq l \leq k$$

where $B_j \in \mathscr{B}$. Since each $B_j \in \mathscr{B}$ is included in the union of two elements of $\mathscr{A}$, we have on event $\mathcal{E}$,

$$\mathbb{P}_n(A_l \Delta A_l') \leq 2d \sup_{A\in\mathscr{A}} \mathbb{P}_n(A) \leq 4d\epsilon_0, \quad \forall 1 \leq l \leq k. \tag{8}$$

Combining equation (8) with Technical Lemma 3, we have on event $\mathcal{E}$,

$$|I_\alpha(T, \mathbb{P}_n) - I_\alpha(T', \mathbb{P}_n)| < 20\alpha k d \epsilon_0.$$

Since ϵ_0 is arbitrary, let $\epsilon_0 = \epsilon/(20\alpha k d)$, we have with probability greater than $1 - \sigma$

$$\sup_{\substack{T,T'\in\mathscr{F}_k \\ D(s(T),s(T'))<\delta}} |I_\alpha(T, \mathbb{P}_n) - I_\alpha(T', \mathbb{P}_n)| < \epsilon.$$

Step 2 By step 1, for all $\epsilon > 0, \sigma > 0, k \in \mathbb{N}$, let $\mathscr{F}_k$ be the set of all trees obtained by $k-1$ partitions, there exists $N_1 > 0, \delta > 0$, $\forall n > N_1$, we have

$$\sup_{\substack{T,T'\in\mathscr{F}_k \\ D(s(T),s(T'))<\delta}} |I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| < \epsilon/3 \tag{9}$$

and

$$\mathbb{P}\left(\sup_{\substack{T,T'\in\mathscr{F}_k \\ D(s(T),s(T'))<\delta}} |I_\alpha(T, \mathbb{P}_n) - I_\alpha(T', \mathbb{P}_n)| \geq \epsilon/3\right) < \sigma/2 \tag{10}$$

Let $\mathscr{G}_k \subset \mathscr{F}_k$ be a $\epsilon/3$-cover on $\mathscr{F}$ with respect to the generalized metric $D(s(\cdot), s(\cdot))$. Because the covering number is no greater than $(d\epsilon/3)^k$, without loss of generality, we let the $\text{card}(\mathscr{G}) = (d\epsilon/3)^k$. Since each element of $\mathscr{G}$ is a fixed tree with at most k leaf nodes and the cardinality of $\mathscr{G}$ is also finite, there exists $N_2 \in \mathbb{N}$, $\forall n > N_2$,

$$\mathbb{P}\left(\sup_{T\in\mathscr{G}_k} |I_\alpha(T, \mathbb{P}) - I_\alpha(T, \mathbb{P}_n)| > \epsilon/3\right) < \sigma/2 \tag{11}$$

Combining equations (9), (10) and (11) and applying triangle inequality, we have $\forall n > \max(N_1, N_2)$

$$\mathbb{P}\left(\sup_{T\in\mathscr{F}_k} |I_\alpha(T, \mathbb{P}_n) - I_\alpha(T, \mathbb{P})| \geq \epsilon\right) < \sigma.$$

Step 3 We now finish the proof of Technical Lemma 4 by induction on k. For $k = 1$, there is only one tree: the root node. So the technical lemma holds naturally. Now supposing the technical lemma holds for $1, 2, \ldots, k-1$, We then consider it for k. By step 1, $\forall \epsilon/3$, $\forall \sigma > 0$, for $k \in \mathbb{N}$, there exists $\delta_0 \in (0, \delta)$,

$$\sup_{\substack{T,T' \in \mathscr{F}_k \\ D(s(T),s(T'))<\delta_0}} |I_\alpha(T, \mathbb{P}) - I_\alpha(T', \mathbb{P})| < \epsilon/3. \tag{12}$$

with probability $1 - \sigma$. Since the technical lemma holds for $k-1$, for all $\sigma > 0$, for the prementioned δ_0, $\exists N_1 > 0$, $\forall n > N_1$, we have

$$\mathbb{P}\left(D(s(\hat{T}_{n,k-1}), s(T^*_{k-1})) > \delta_0\right) < \sigma. \tag{13}$$

Denote $\hat{s}_{n,k-1}, s^*_{k-1}$ as the $(k-1)$th partition of $\hat{T}_{n,k}, T^*_k$, respectively. Define two auxiliary trees $T^*_{k-1}(\hat{s}_{n,k-1})$, $\hat{T}_{n,k-1}(s^*_{k-1})$. $T^*_{k-1}(\hat{s}_{n,k-1})$ is obtained by applying partition $\hat{s}_{n,k-1}$ on theoretical tree T^*_{k-1}, while $\hat{T}_{n,k-1}(s^*_{k-1})$ is obtained by applying partition s^*_{k-1} on SVR-Tree $\hat{T}_{n,k-1}$. We reassign class labels on all leaf nodes of both $T^*_{k-1}(\hat{s}_{n,k-1})$ and $\hat{T}_{n,k-1}(s^*_{k-1})$. For any leaf node A of $T^*_{k-1}(\hat{s}_{n,k-1})$ or $\hat{T}_{n,k-1}(s^*_{k-1})$, its class label is 1 if $\mathbb{P}(Y|X \in A) \geq 1/(1+\alpha)$ and 0 otherwise. By doing this, both $T^*_{k-1}(\hat{s}_{n,k-1})$ and $\hat{T}_{n,k-1}(s^*_{k-1})$ estimates the class label with the true probability distribution, while the partition structure is prefixed. Combine equation (12) and (13), we have with probability greater than $1 - \sigma$,

$$|I_\alpha(\hat{T}_{n,k}, \mathbb{P}) - I(\hat{T}^*_{k-1}(\hat{s}_{n,k-1}), \mathbb{P})| < \epsilon/3 \tag{14}$$

$$|I_\alpha(\hat{T}_{n,k-1}(s^*_{k-1}), \mathbb{P}) - I(T^*_k, \mathbb{P})| < \epsilon/3 \tag{15}$$

By the optimality of $\hat{T}_{n,k}$, we have

$$I_\alpha(\hat{T}_{n,k}, \mathbb{P}_n) + \lambda_n r(\hat{T}_{n,k}) \leq I_\alpha(\hat{T}_{n,k-1}(s^*_{k-1}), \mathbb{P}_n) + \lambda_n r(\hat{T}_{n,k-1}(s^*_{k-1})).$$

By Technical Lemma 2, the sequence of theoretical trees is consistent. Therefore, there exists $V_0 > 0, k_0 \in \mathbb{N}$, such that $V(T_k) \geq V_0, \forall k > k_0$. Without loss of generality, we assume here $k - 1 > k_0$. Thus with probability greater than $1 - \sigma$, the volume of the theoretical tree satisfies

$$V(T^*_{k-1}) \geq V_0.$$

Therefore, the surface-to-volume ratio of $\hat{T}_{n,k-1}(s^*_{k-1})$ is bounded by:

$$r\left(\hat{T}_{n,k-1}(s^*_{k-1})\right) \leq \frac{2dk}{V_0 - (k-1)\delta_0}.$$

Therefore, we have

$$I_\alpha(\hat{T}_{n,k}, \mathbb{P}_n) \leq I_\alpha(\hat{T}_{n,k-1}(s^*_{k-1}), \mathbb{P}_n) + \lambda_n \frac{2dk}{V_0 - (k-1)\delta_0}. \tag{16}$$

Combining equation (16) and step 2, there exists $N_2 \in \mathbb{N}$, $\forall n > N_2$, such that with probability greater than $1 - \sigma$,

$$I_\alpha(\hat{T}_{n,k}, \mathbb{P}) \leq I_\alpha(\hat{T}_{n,k-1}(s^*_{k-1}), \mathbb{P}) + \lambda_n \frac{2dk}{V_0 - (k-1)\delta_0} + \frac{2}{3}\epsilon. \tag{17}$$

Combining equations (14), (15) and (17), $\forall n > \max\{N_1, N_2\}$, we have with probability greater than $1 - 2\sigma$,

$$I_\alpha(T^*_{k-1}(\hat{s}_{n,k}), \mathbb{P}) \leq I_\alpha(T^*_k, \mathbb{P}) + \lambda_n \frac{2dk}{V_0 - (k-1)\delta_0} + \frac{4}{3}\epsilon. \tag{18}$$

Denote a collection of trees as

$$\mathscr{H}_\varepsilon = \{T \in \mathscr{F}_k : s(T) = (s^*_1, s^*_2, \ldots, s^*_{k-2}, s'),\ s' \text{ arbitrary},\ I_\alpha(T, \mathbb{P}) - I_\alpha(T^*_k, \mathbb{P}) < \epsilon\}.$$

Therefore, $\mathscr{H}_\varepsilon$ is the collection of trees whose first $k-2$ partitions are the same as T_k^* and the tree impurity is within ε distance of T_k^*. Letting $\epsilon_n = \lambda_n \frac{2dk}{V_0-(k-1)\delta_0} + 4\epsilon/3$, equation (18) implies

$$D(s(T_{k-1}^*(\hat{s}_{n,k})), s(T_k^*)) \leq \sup_{T \in \mathscr{H}_{\epsilon_n}} D(s(T), s(T_k^*)). \tag{19}$$

Because T_k^* is the unique theoretical tree with k leaf nodes and $\cap_{\epsilon>0}\mathscr{H}_\epsilon$ is the collection of theoretical trees with k leaf nodes,[2] we have

$$\lim_{\varepsilon \to 0} \sup_{T \in \mathscr{H}_\varepsilon} D(s(T), s(T_k^*)) = 0.$$

This implies for $\delta > 0$, there exists $\epsilon_0 > 0$, such that $\forall \varepsilon < \epsilon_0$, we have $\sup_{T\in\mathscr{H}_\varepsilon} D(s(T), s(T_k^*)) < \delta$. Recall equation (19), noting $\lim_{n\to\infty}\lambda_n = 0$, there exists $N_3 > 0$, such that $\forall n > N_3$, $\lambda_n \frac{2dk}{V_0-(k-1)\delta_0} < \epsilon_0/2$. Since ϵ is arbitrary, we can let $4\epsilon/3 < \epsilon_0/2$. Therefore, we have $\forall n > \max\{N_1, N_2, N_3\}$, with probability greater than $1-2\sigma$,

$$D(s(T_{k-1}^*(\hat{s}_{n,k})), s(T_k^*)) < \delta. \tag{20}$$

Noting on the same event where equation (20) holds, the induction hypothesis (13) also holds. We finally have $\forall n > \max\{N_1, N_2, N_3\}$

$$\mathbb{P}\left(D(s(\hat{T}_{n,k}), s(T_k^*)) \geq \delta\right) < 2\sigma.$$

Since σ is arbitrary, we finish the proof.

If Optional Steps Are Enabled Now suppose optional steps in Algorithm 1 are enabled. We will reject a new feature if the impurity decrease at a new feature is no greater than $c_0\lambda_n$ plus the maximal impurity decrease at features that are already partitioned. Because λ_n goes to zero as n goes to infinity, for any fixed $k \in \mathbb{N}$, the probability to for a partition to be rejected by optional steps[3] in first k partitions goes to zero. Therefore for all $k \in \mathbb{N}$, as n goes to zero, the probability for $\hat{T}_{n,k}$ to remain invariant with respect to optional steps goes to one. We finish the proof.

□

We are now prepared to prove Lemma 4.

Proof of Lemma 4. For all $\epsilon > 0$, by Technical Lemma 2, there exists $K \in \mathbb{N}$, such that $\forall k > K$, the theoretical tree T_k^* satisfies

$$|\tilde{I}_\alpha(T_k^*, \mathbb{P}) - I_\alpha^*| < \epsilon. \tag{21}$$

Because the probability measure $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure μ, for $\epsilon/5\alpha k > 0$, there exists $\delta > 0$, such that $\forall A \subset \Omega$, $\mu(A) < \delta$ implies $\mathbb{P}(A) < \epsilon/(5\alpha k)$. Fix $k > K$. Since $\bar{a}_n \to \infty$, as n goes to infinity, $\hat{T}_n$ has no less than k partitions, and $\hat{T}_{n,k}$ is well-defined. By Technical Lemma 4, for $\delta/d > 0$, $\forall \sigma > 0$, $\exists N \in \mathbb{N}$, $\forall n > N$, we have

$$\mathbb{P}\left(D(s(\hat{T}_{n,k}), s(T_k^*)) > \delta/d\right) < \sigma.$$

Thus by Technical Lemma 3, we have

$$|I_\alpha(\hat{T}_{n,k}, \mathbb{P}) - I_\alpha(T_k^*, \mathbb{P})| \leq 5\alpha k \cdot \epsilon/(5\alpha k) = \epsilon.$$

Let $\tilde{T}_n$ be the tree that has the same partition as $\hat{T}_{n,k}$. On each leaf node A of $\tilde{T}_n$, let the class label be 1 if and only if $\mathbb{E}(Y|X \in A) \geq \alpha/(1+\alpha)$. Since $\hat{T}_{n,k}$ is formed by the first k partitions of $\hat{T}_n$, $\tilde{T} \in \mathscr{T}_n$. By definition of $\tilde{T}_n$ and T_k^*, we have $\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) = I_\alpha(\tilde{T}_n, \mathbb{P})$ and $\tilde{I}_\alpha(T_k^*, \mathbb{P}) = I_\alpha(T_k^*, \mathbb{P})$. Therefore we have

$$|\tilde{I}_\alpha(\hat{T}_{n,k}, \mathbb{P}) - \tilde{I}_\alpha(T_k^*, \mathbb{P})| \leq \epsilon. \tag{22}$$

[2] As we have mentioned before, if the theoretical tree is not unique, we can replace $D(s(T), s(T_k^*))$ with the infimum distance to all theoretical trees. Since the space of partitions of all theoretical trees forms a closed subspace of $\mathscr{H}_1$, the same argument still holds.

[3] This means, a partition is rejected by these optional step but will be accepted as risk minimizer if these steps are not enabled.

By Technical Lemma 2, there exists $k_0 \in \mathbb{N}$, such that for all $k > k_0$, $V_k^* \geq V_0$. Without loss of generality, we assume $k > k_0$. Combining it with Technical Lemma 4, the surface to volume ratio of $\tilde{T}_n$ can be bounded by

$$r(\tilde{T}_n) = r(\hat{T}_{n,k}) \leq \frac{2dk}{V_0 - k\delta} = \frac{2dk}{V_0 - \epsilon/(5d)}. \tag{23}$$

Combining equation (21), (22) and (23), we have with probability greater than $1 - \sigma$,

$$|\tilde{I}_\alpha(\tilde{T}_n, \mathbb{P}) + \lambda_n r(\tilde{T}_n) - I_\alpha^*| \leq 2\epsilon + \lambda_n \frac{2dk}{V_0 - \epsilon/(5d)}.$$

Since ϵ, σ are arbitrary and $\lim_{n\to\infty} \lambda_n = 0$, we finish the proof. □

2 Proof of Lemma 5-6

2.1 Proof of Lemma 5

Proof. We utilize M_A^* as a bridge to compare the values between $\sup_{1\leq j\leq q} M_{A,j}$ and $\sup_{q+1\leq l\leq d} M_{A,l}$. By condition 1, we have $\sup_{1\leq j\leq q} M_{A,j} \geq c_1 M_A^*$. It remains to compare the values between $\sup_{q+1\leq l\leq d} M_{A,l}$ and M_A^*. Suppose hyperrectangle A is partitioned into two measurable (but not necessarily hyperrectangular) sets A_1 and A_2, such that $A_1 \cap A_2 = \emptyset$ and $A_1 \cup A_2 = A$.

For simplicity of notation, $\forall X \in A$, we denote $p_\alpha(X) = \mathbb{E}_\alpha(Y|X)$. Recall for all $A \subset \Omega$, we let $p_\alpha(A) = \mathbb{E}_\alpha(Y|X \in A)$. Further denote $p_{1,\alpha} = \mathbb{E}_\alpha(Y|X \in A_1)$, $p_{2,\alpha} = \mathbb{E}_\alpha(Y|X \in A_2)$, $V_1 = \mathbb{P}_\alpha(A_1|A)$, $V_2 = \mathbb{P}_\alpha(A_2|A)$, then the impurity decrease on node A can be computed as

$$\Delta I_\alpha(A, \mathbb{P}) = 2p_\alpha(A)(1 - p_\alpha(A)) - 2V_1 p_{1,\alpha}(1 - p_{1,\alpha}) - 2V_2 p_{2,\alpha}(1 - p_{2,\alpha}). \tag{24}$$

By definition of conditional expectation, we also have

$$V_1 p_{1,\alpha} + V_2 p_{2,\alpha} = p_\alpha(A). \tag{25}$$

Combining equation (24), (25) with the fact that $V_1 + V_2 = 1$, we have

$$\Delta I_\alpha(A, \mathbb{P}) = V_1 V_2 (p_{1,\alpha} - p_{2,\alpha})^2. \tag{26}$$

By equation (26), when V_1, V_2 are fixed, the impurity decrease on A is proportional to the squared difference between $p_{1,\alpha}$ and $p_{2,\alpha}$. Let A_{y1}, A_{y2} be a pair of partitioned sets that achieves impurity decrease $h_A^*(V_1)$; i.e.,

$$\mathbb{P}_\alpha(A_{y_1}) = V_1 \mathbb{P}_\alpha(A),\ A_{y_1} \cap A_{y_2} = \emptyset,\ A_{y_1} \cup A_{y_2} = A,$$
$$\forall X_1 \in A_{y_1}, X_2 \in A_{y_2},\ \mathbb{E}(Y|X_1) \leq \mathbb{E}(Y|X_2).$$

Let $y_{1,\alpha}, y_{2,\alpha}$ be the conditional expectation of Y at A_{y_1}, A_{y_2} under $\mathbb{P}_\alpha$:

$$y_{1,\alpha} = \mathbb{E}(Y|X \in A_{y1}), \quad y_{2,\alpha} = \mathbb{E}(Y|X \in A_{y2}).$$

The critical value of conditional expectation (under $\mathbb{P}_\alpha$) between A_{y_1}, A_{y_2} is denoted as y_t:

$$y_t \in [0,1]: \quad \sup_{X\in A_{y_1}} p_\alpha(X) \leq y_t \leq \inf_{X\in A_{y_2}} p_\alpha(X).$$

In the rest of the proof, supppose we partition at a redundant feature to obtain A_1, A_2. For all fixed $V_1 \in [0,1]$, we compare $|p_{1,\alpha} - p_{2,\alpha}|$ and $|y_{1,\alpha} - y_{2,\alpha}|$, thus comparing the values between $M_{A,l}$ and M_A^*, $\forall l \geq q+1$. For hyperrectangle A, we use $\mathrm{Proj}(A, X')$, $\mathrm{Proj}(A, X'')$ to denote the projection of A on X', X'', respectively. By definition, $A = \mathrm{Proj}(A, X') \times \mathrm{Proj}(A, X'')$. We first consider the properties of $p_{1,\alpha}$ and $p_{2,\alpha}$.

$$\begin{aligned}
p_{1,\alpha} V_1 \mathbb{P}_\alpha(A) =& \int_{A_1} \rho_{1,\alpha}(X')\rho_{2,\alpha}(X'')p_\alpha(X)dX'dX'' + \int_{A_1} \rho_{3,\alpha}(X', X'')p_\alpha(X)dX'dX'' \\
=& \frac{\int_{\mathrm{Proj}(A_1,X')} p_\alpha(X)\rho_{1,\alpha}(X')dX'}{\int_{\mathrm{Proj}(A_1,X')} \rho_{1,\alpha}(X')dX'} \int_{\mathrm{Proj}(A_1,X')} \rho_{1,\alpha}(X')dX' \int_{\mathrm{Proj}(A_1,X'')} \rho_{2,\alpha}(X'')dX'' + \\
& \frac{\int_{A_1} \rho_{3,\alpha}(X', X'')p_\alpha(X)dX'dX''}{\int_{A_1} \rho_{3,\alpha}(X', X'')dX'dX''} \int_{A_1} \rho_{3,\alpha}(X', X'')dX'dX''
\end{aligned}$$

Denote

$$y_A \triangleq \frac{\int_{\mathrm{Proj}(A,X')} p_\alpha(X)\rho_{1,\alpha}(X')dX'}{\int_{\mathrm{Proj}(A,X')} \rho_{1,\alpha}(X')dX'},$$

$$\gamma_1 = \int_{A_1} \rho_{3,\alpha}(X', X'')dX'dX''/\mathbb{P}_\alpha(A_1).$$

Noting $\mathrm{Proj}(A_1, X') = \mathrm{Proj}(A, X')$, $p_\alpha(X)$ is independent of X'' and y_A being well-defined, we have

$$p_{1,\alpha}V_1\mathbb{P}_\alpha(A) = \left[y_A(1-\gamma_1) + \frac{\int_{A_1} \rho_{3,\alpha}(X', X'')p_\alpha(X)dX'dX''}{\int_{A_1} \rho_{3,\alpha}(X', X'')dX'dX''}\gamma_1\right] V_1\mathbb{P}_\alpha(A).$$

Therefore,

$$[p_{1,\alpha} - (1-\gamma_1)y_A - \gamma_1 y_t]V_1\mathbb{P}_\alpha(A) = \int_{A_1} \rho_{3,\alpha}(X', X'')[p_\alpha(X) - y_t]dX'dX''.$$

Similarly, define

$$\gamma_2 = \int_{A_2} \rho_{3,\alpha}(X', X'')dX'dX''/\mathbb{P}_\alpha(A_2),$$

we have

$$[p_{2,\alpha} - (1-\gamma_2)y_A - \gamma_2 y_t]V_1\mathbb{P}_\alpha(A) = \int_{A_2} \rho_{3,\alpha}(X', X'')[p_\alpha(X) - y_t]dX'dX''.$$

Without loss of generality, we assume $p_{1,\alpha} - (1-\gamma_1)y_A - \gamma_1 y_t \le p_{2,\alpha} - (1-\gamma_2)y_A - \gamma_2 y_t$. Then

$$\begin{aligned}
[p_{1,\alpha} - (1-\gamma_1)y_A - \gamma_1 y_t]V_1\mathbb{P}_\alpha(A) =& \int_{A_1} \rho_{3,\alpha}(X', X'')[p_\alpha(X) - y_t]dX'dX'' \\
\ge& \int_{A_1} \rho_{3,\alpha}(X', X'')\min\{p_\alpha(X) - y_t, 0\}dX'dX'' \\
\ge& \int_{A} \rho_{3,\alpha}(X', X'')\min\{p_\alpha(X) - y_t, 0\}dX'dX'' \\
=& \int_{\{X\in A: p_\alpha(X) < y_t\}} \rho_{3,\alpha}(X', X'')[p_\alpha(X) - y_t]dX'dX'' \\
\ge& c_2 \int_{\{X\in A: p_\alpha(X) < y_t\}} \rho_{\alpha}(X', X'')[p_\alpha(X) - y_t]dX'dX''
\end{aligned}$$

By the definition of y_t,

$$\{X \in A : p_\alpha(X) < y_t\} \subset A_{y_1} \subset \{X \in A : p_\alpha(X) \le y_t\}.$$

Therefore

$$\begin{aligned}
[p_{1,\alpha} - (1-\gamma_1)y_A - \gamma_1 y_t]V_1\mathbb{P}_\alpha(A) \ge& c_2 \int_{A_{y_1}} \rho_{3,\alpha}(X', X'')[p_\alpha(X) - y_t]dX'dX'' \\
=& c_2\mathbb{P}_\alpha(A)V_1(y_{1,\alpha} - y_t),
\end{aligned}$$

Thus

$$p_{1,\alpha} - (1-\gamma_1)y_A - \gamma_1 y_t \ge c_2(y_{1,\alpha} - y_t). \tag{27}$$

Similarly, we have

$$p_{2,\alpha} - (1-\gamma_2)y_A - \gamma_2 y_t \le c_2(y_{2,\alpha} - y_t). \tag{28}$$

We then consider the difference between y_A and y_t.

$$\begin{aligned}
&|y_A - y_t| \int_A \rho_{1,\alpha}(X')\rho_{2,\alpha}(X'')dX'dX'' \\
=&\left| \int_A [p_\alpha(X) - y_t]\rho_{1,\alpha}(X')\rho_{2,\alpha}(X'')dX'dX'' \right| \\
\leq& \int_A |p_\alpha(X) - y_t|\rho_{1,\alpha}(X')\rho_{2,\alpha}(X'')dX'dX'' \\
\leq& \int_A |p_\alpha(X) - y_t|\rho_\alpha(X', X'')dX'dX'' \\
=& \int_{A_{y_1}} |p_\alpha(X) - y_t|\rho_\alpha(X', X'')dX'dX'' + \int_{A_{y_2}} |p_\alpha(X) - y_t|\rho_\alpha(X', X'')dX'dX'' \\
=& \left| \int_{A_{y_1}} [p_\alpha(X) - y_t]\rho_\alpha(X', X'')dX'dX'' \right| + \left| \int_{A_{y_2}} [p_\alpha(X) - y_t]\rho_\alpha(X', X'')dX'dX'' \right| \\
=& (|y_{1,\alpha} - y_t|V_1 + |y_{2,\alpha} - y_t|V_2)\, \mathbb{P}_\alpha(A).
\end{aligned}$$

Because

$$\int_A \rho_{1,\alpha}(X')\rho_{2,\alpha}(X'')dX'dX'' \geq \int_A (1 - c_2)dX'dX'' \geq (1 - c_2)\mathbb{P}_\alpha(A),$$
$$V_1 < 1, \;\; V_2 < 1,$$

we have

$$|y_A - y_t| \leq (|y_{1,\alpha} - y_t| + |y_{2,\alpha} - y_t|)\, \frac{1}{1 - c_2} = \frac{|y_{1,\alpha} - y_{2,\alpha}|}{1 - c_2}. \tag{29}$$

Combining equations (27), (28) and (29), we have

$$\begin{aligned}
|p_{2,\alpha} - p_{1,\alpha}| =& |[p_{2,\alpha} - (1 - \gamma_2)y_A - \gamma_2 y_t] - [p_{1,\alpha} - (1 - \gamma_1)y_A - \gamma_1 y_t] + (\gamma_1 - \gamma_2)(y_A - y_t)| \\
\leq& |[p_{2,\alpha} - (1 - \gamma_2)y_A - \gamma_2 y_t] - [p_{1,\alpha} - (1 - \gamma_1)y_A - \gamma_1 y_t]| + |(\gamma_1 - \gamma_2)(y_A - y_t)| \\
\leq& c_2(y_{2,\alpha} - y_{1,\alpha}) + c_2 \frac{|y_{1,\alpha} - y_{2,\alpha}|}{1 - c_2} \\
\leq& \frac{c_2(2 - c_2)}{1 - c_2} |y_{2,\alpha} - y_{1,\alpha}|.
\end{aligned} \tag{30}$$

Because equation (30) holds for all $V \in [0, 1]$, recalling equation (26), we have

$$M_{A,l} \leq \frac{c_2(2 - c_2)}{1 - c_2} M_A^*, \;\; \forall l \geq q + 1.$$

Recalling that $c_1 > \frac{c_2(2-c_2)}{1-c_2}$, we finish the proof. □

2.2 Proof of Lemma 6

Proof. Since $0 \leq \Delta I_\alpha(A, A_1, \mathbb{P}), \Delta I_\alpha(A, A_1, \mathbb{P}_n) \leq 1$, the lemma automatically holds when $c_0 \lambda_n n/n' > 1$. We only need to consider the case when $c_0 \lambda_n n/n' \leq 1$. The proof can be divided into three steps. In the first step, we define an ϵ-net on the space of all possible A_1, with the distance being the symmetric set difference. We prove this ϵ-net also corresponds to a $5\alpha\epsilon$-net of $\Delta I_\alpha(A, A_1, \mathbb{P})$, while with high probability corresponds to a $10\alpha\epsilon$-net of $\Delta I_\alpha(A, A_1, \mathbb{P}_n)$. In the second step, we show with high probability, the difference between $\Delta I_\alpha(A, A_1, \mathbb{P})$ and $\Delta I_\alpha(A, A_1, \mathbb{P}_n)$ can be bounded uniformly for all A_1 in the ϵ-net. The last step combines the results of the previous two steps, calculating tree impurity decrease from impurity decrease at node A, giving a uniform bound of $|\Delta I_\alpha(T, \mathbb{P}) - \Delta I_\alpha(T, \mathbb{P}_n)|$.

Step 1 Denote the hyperrectangle A as $A = [x_{1,l}, x_{1,r}] \times [x_{2,l}, x_{2,r}] \times \cdots \times [x_{d,l}, x_{d,r}]$. Without loss of generality, we assume $(x_{1,l}, x_{2,l}, \ldots, x_{d,l}) \in A_1$, i.e., after a partition at A, A_1 always contains the "left corner" of A. Let n' be the number of training samples inside A. Let $\epsilon = c\lambda_n n/n'$, and $m = \lceil 1/\epsilon \rceil$. For each feature $X[j]$, define a series of points $x_{j,0} < x_{j,1} < \ldots < x_{j,m}$ such that $x_{j,0} = x_{j,l}$, $x_{j,m} = x_{j,r}$,

$$\mathbb{P}_\alpha(X[j] \in [x_{j,k}, x_{j,k+1}] | X \in A) = 1/m, \ \ \forall 0 \le k \le m-1.$$

That is, $x_{j_1}, x_{j_2}, \ldots, x_{j_{m-1}}$ are all the mth quantiles of the marginal probability measure. Define a collection of subsets of A as:

$$\mathscr{A} = \big\{A = [x_{1,l}, x_{1,r}] \times [x_{2,l}, x_{2,r}] \times \cdots \times [x_{j-1,l}, x_{j-1,r}] \times [x_{j,0}, x_{j,k}] \times [x_{j+1,l}, x_{j+1,r}] \times \cdots \times [x_{d,0}, x_{d,1}] : \ 1 \le j \le d, 1 \le k \le m-1\big\}$$

That is, $\mathscr{A}$ contains all the possible A_1 that are obtained by partitioning at $X[j] = x_{j,k}$, $1 \le j \le d, 1 \le k \le m-1$. By the definition of $x_{j,k}$, for all A_1 that are obtained by a single partition at A, there exists $A_1' \in \mathscr{A}$, such that $\mathbb{P}(A_1 \Delta A_1' | A) \le 1/m \le \epsilon$. So $\mathscr{A}$ forms a ϵ-net for all the possible A_1, with the distance function being the symmetric set difference under $\mathbb{P}$. By technical lemma 3, the set

$$\{\Delta I_\alpha(A, A_1, \mathbb{P}) : A_1 \in \mathscr{A}\}$$

forms a $5\alpha\epsilon$-net in the space of $\Delta I_\alpha(A, A_1, \mathbb{P})$.

Define another collection of subsets of A as:

$$\mathscr{B} = \big\{A = [x_{1,l}, x_{1,r}] \times [x_{2,l}, x_{2,r}] \times \cdots \times [x_{j-1,l}, x_{j-1,r}] \times [x_{j,k}, x_{j,k+1}] \times [x_{j+1,l}, x_{j+1,r}] \times \cdots \times [x_{d,0}, x_{d,1}] : \ 1 \le j \le d, 0 \le k \le m-1$$

We have for all $B \in \mathscr{B}$, $\mathbb{P}(B|A) = 1/m < \epsilon$. By Hoeffding's inequality, $\forall B \in \mathscr{B}$,

$$\mathbb{P}\left(\Big|\frac{n\mathbb{P}_n(B)}{n'} - \frac{1}{m}\Big| \ge \epsilon\right) \le 2\exp(-2n'\epsilon^2) = 2\exp(-2c^2\lambda_n^2 n^2/n') \le 2\exp(-2c^2\lambda_n^2 n) \le 2\exp(-2c^2 n^{2\beta}).$$

Because $1/\epsilon = (c\lambda_n n/n')^{-1} \le n^{1/2-\beta}/c$, applying a union bound for all the $B \in \mathscr{B}$, we have with probability greater than $1 - 2n^{1/2-\beta}\exp(-2c^2n^{2\beta})/c$ that

$$\sup_{B \in \mathscr{B}} \mathbb{P}_n(B|A) \le 2\epsilon.$$

That is, $\mathscr{A}$ forms a 2ϵ-net for all possible A_1 measured by $\mathbb{P}_n$. Therefore, by Technical Lemma 3, the set

$$\{\Delta I_\alpha(A, A_1, \mathbb{P}_n) : A_1 \in \mathscr{A}\}$$

forms a $10\alpha\epsilon$-net in the space of $\Delta I_\alpha(A, A_1, \mathbb{P}_n)$.

Step 2 Denote $p_{1,\alpha} = \mathbb{E}_\alpha(Y|X \in A_1)$, $\hat{p}_{1,\alpha} = \mathbb{E}_{n,\alpha}(Y|X \in A_1)$, $p_{2,\alpha} = \mathbb{E}_\alpha(Y|X \in A_2)$, $\hat{p}_{2,\alpha} = \mathbb{E}_{n,\alpha}(Y|X \in A_2)$, $p_\alpha = \mathbb{E}_\alpha(Y|X \in A)$, $\hat{p}_\alpha = \mathbb{E}_{n,\alpha}(Y|X \in A)$, then we have

$$\begin{aligned}
|\Delta I_\alpha(A, A_1, \mathbb{P}) - \Delta I_\alpha(A, A_1, \mathbb{P}_n)| =& \big|2p_\alpha(1-p_\alpha) - 2p_{1,\alpha}(1-p_{1,\alpha})\mathbb{P}_\alpha(A_1|A) - 2p_{2,\alpha}(1-p_{2,\alpha})\mathbb{P}_\alpha(A_2|A) \\
&+ 2\hat{p}_\alpha(1-\hat{p}_\alpha) + 2\hat{p}_{1,\alpha}(1-\hat{p}_{1,\alpha})\mathbb{P}_{n,\alpha}(A_1|A) - 2\hat{p}_{2,\alpha}(1-\hat{p}_{2,\alpha})\mathbb{P}_{n,\alpha}(A_2|A)\big| \\
\le& 2|(p_\alpha - \hat{p}_\alpha)(1 - p_\alpha - \hat{p}_\alpha)| + \\
& 2|p_{1,\alpha}(1-p_{1,\alpha})\mathbb{P}_\alpha(A_1|A) - \hat{p}_{1,\alpha}(1-\hat{p}_{1,\alpha})\mathbb{P}_{n,\alpha}(A_1|A)| + \\
& 2|p_{2,\alpha}(1-p_{2,\alpha})\mathbb{P}_\alpha(A_2|A) - \hat{p}_{2,\alpha}(1-\hat{p}_{2,\alpha})\mathbb{P}_{n,\alpha}(A_2|A)| \\
\le& 2|p_\alpha - \hat{p}_\alpha| + \frac{1}{2}|\mathbb{P}_\alpha(A_1|A) - \mathbb{P}_{n,\alpha}(A_1|A)| + 2\mathbb{P}_\alpha(A_1|A)|p_{1,\alpha} - \hat{p}_{1,\alpha}| + \\
& \frac{1}{2}|\mathbb{P}_\alpha(A_2|A) - \mathbb{P}_{n,\alpha}(A_2|A)| + 2\mathbb{P}_\alpha(A_2|A)|p_{2,\alpha} - \hat{p}_{2,\alpha}| \qquad (31)
\end{aligned}$$

For all $A_1 \in \mathscr{A}$, by Hoeffding's inequality,

$$\mathbb{P}\left(\Big|\frac{n\mathbb{P}_n(A_1)}{n'} - \frac{\mathbb{P}(A_1)}{\mathbb{P}(A)}\Big| \ge \epsilon\right) \le 2\exp(-2n'\epsilon^2) \le 2\exp(-2c^2n^{2\beta}). \qquad (32)$$

Noting $\mathbb{P}_\alpha(A_1|A) - \mathbb{P}_{n,\alpha}(A_1|A) + \mathbb{P}_\alpha(A_2|A) - \mathbb{P}_{n,\alpha}(A_2|A) = 1 - 1 = 0$, we have $|\mathbb{P}_\alpha(A_2|A) - \mathbb{P}_{n,\alpha}(A_2|A)| = |\mathbb{P}_\alpha(A_1|A) - \mathbb{P}_{n,\alpha}(A_1|A)|$. Thus the same bound as equation (31) applies to $|\mathbb{P}_\alpha(A_2|A) - \mathbb{P}_{n,\alpha}(A_2|A)|$. It remains to bound the term $\mathbb{P}_\alpha(A_1|A)|p_{1,\alpha} - \hat{p}_{1,\alpha}|$ and $\mathbb{P}_\alpha(A_2|A)|p_{2,\alpha} - \hat{p}_{2,\alpha}|$.

For all $A_1 \in \mathscr{A}$, let $\epsilon_{A_1} = \epsilon/\sqrt{\mathbb{P}(A_1|A)}$, then by Hoeffding's inequality,

$$\mathbb{P}\left(\left|\mathbb{E}_n(Y|A_1) - \mathbb{E}(Y|A_1)\right| \geq \epsilon_{A_1}\right) \leq 2\exp(-2n'\mathbb{P}_n(A_1|A)\epsilon_{A_1}^2) = 2\exp(-2n'\epsilon^2) \leq 2\exp(-2c^2n^{2\beta}).$$

Therefore

$$\begin{aligned}
\mathbb{P}\left(\mathbb{P}_\alpha(A_1|A)\left|\hat{p}_{1,\alpha} - p_{1,\alpha}\right| \geq \alpha^2\epsilon\right) &\leq \mathbb{P}\left(\mathbb{P}(A_1|A)\left|\hat{p}_{1,\alpha} - p_{1,\alpha}\right| \geq \alpha\epsilon\right) \\
&\leq \mathbb{P}\left(\sqrt{\mathbb{P}(A_1|A)}\left|\hat{p}_{1,\alpha} - p_{1,\alpha}\right| \geq \alpha\epsilon\right) \\
&= \mathbb{P}\left(\left|\hat{p}_{1,\alpha} - p_{1,\alpha}\right| \geq \alpha\epsilon_{A_1}\right) \\
&\leq 2\exp(-2c^2n^{2\beta}). \qquad (33)
\end{aligned}$$

Similarly, we have

$$\mathbb{P}\left(\mathbb{P}_\alpha(A_1|A)\left|\hat{p}_{1,\alpha} - p_{1,\alpha}\right| \geq \alpha^2\epsilon\right) \leq 2\exp(-2c^2n^{2\beta}), \quad \mathbb{P}\left(\left|\hat{p}_\alpha - p_\alpha\right| \geq \alpha\epsilon\right) \leq 2\exp(-2c^2n^{2\beta}). \qquad (34)$$

Combining equation (31), (32), (33) and (34), and applying a union bound, we have with probability greater than

$$1 - \left[\frac{6}{c}n^{1/2-\beta} + 2\right]\exp(-2c^2n^{2\beta}),$$

for all $A_1 \in \mathscr{A}$, the difference between $\Delta I_\alpha(A, A_1, \mathbb{P})$ and $\Delta I_\alpha(A, A_1, \mathbb{P}_n)$ is bounded by

$$\sup_{A_1 \in \mathscr{A}} |\Delta I_\alpha(A, A_1, \mathbb{P}) - \Delta I_\alpha(A, A_1, \mathbb{P}_n)| \leq (2 + 1/2 + 2\alpha + 1/2 + 2\alpha)\epsilon\alpha = \epsilon\alpha(3 + 4\alpha).$$

Step 3 Let $\sigma(n)$ be

$$\sigma(n) = 1 - \left[\frac{8}{c}n^{1/2-\beta} + 2\right]\exp(-c^2n^\beta).$$

Combining the results of step 1 and step 2, with probability greater than $1 - \sigma(n)$, $\forall A_1$, there exists $A_1' \in \mathscr{A}$, such that the following events hold simultaneously

$$|\Delta I_\alpha(A, A_1, \mathbb{P}) - \Delta I_\alpha(A, A_1', \mathbb{P})| < 5\alpha\epsilon,$$

$$|\Delta I_\alpha(A, A_1, \mathbb{P}_n) - \Delta I_\alpha(A, A_1', \mathbb{P}_n)| < 10\alpha\epsilon,$$

$$|\Delta I_\alpha(A, A_1', \mathbb{P}) - \Delta I_\alpha(A, A_1', \mathbb{P}_n)| < (3 + 4\alpha)\alpha\epsilon,$$

Therefore

$$|\Delta I_\alpha(A, A_1, \mathbb{P}) - \Delta I_\alpha(A, A_1, \mathbb{P}_n)| < (18 + 4\alpha)\alpha\epsilon.$$

Letting $c = \frac{c_0}{(36+8\alpha)\alpha}$ and noticing $\lim_{n\to\infty}\sigma(n) = 0$, we finish the proof. □

3 Proof of Corollary 1

We first prove the results of Lemma 4 still hold under the condition of corollary 1.

Technical Lemma 5. *If Condition 1 holds with $c_1 = 0$, for all hyperrectangles $A \subset \Omega$, we have*

$$\sup_{1\leq j\leq q} M_{A,j} \geq \sup_{q+1\leq l\leq d} M_{A,l},$$

where $M_{A,j}$ is the maximal impurity decrease of feature j at node A.

Proof. By definition of M_A^*, we have $M_A^* \geq M_{A,l}$, $\forall q+1 \leq l \leq d$. Since $c_1 = 1$, we have $\sup_{1\leq j\leq q} M_{A,j} = M_A^*$, therefore $\sup_{1\leq j\leq q} M_{A,j} \geq M_{A,l}$, $\forall q+1 \leq l \leq d$. □

The proof of corollary 1 using Technical Lemma 5 and Lemma 6 is the same as the proof of Theorem 2 using Lemma 5 and Lemma 6.

4 Proof of c_1 Values in Examples 1-2

4.1 Example 1

Proof. Using similar notation as in the proof of Theorem 2, let $\mathbb{P}_\alpha(A_{y1}) = V_1$, $\mathbb{P}_\alpha(A_{y2}) = V_2$, and $p_\alpha = \mathbb{E}_\alpha(Y|X \in A)$, $y_{1,\alpha} = \mathbb{E}_\alpha(Y|X \in A_{y1})$, $y_{2,\alpha} = \mathbb{E}_\alpha(Y|X \in A_{y2})$. Without loss of generality, we assume $\mathbb{P}_\alpha(A_{y2}) = h \le \mathbb{P}_\alpha(A_{y1})$, and $y_{1,\alpha} \le y_{2,\alpha}$. Since $\mathbb{E}(Y|X)$ is a monotonically increasing function of $a^T + b$, there exists $t \in \mathbb{R}$, such that $\forall X_1 \in A_{y1}, X_2 \in A_{y2}$, we have $a^T X_1 + b \le t \le a^T X_2 + b$.

Let A_1, A_2 be two hyperrectangles obtained by partitioning at one of the features, which satisfies $\mathbb{P}(A_1) = V_1$ and $\mathbb{P}(A_2) = V_2$. The feature to partition is selected by minimizing $\mathbb{P}(A_1 \backslash A_{y1})$. Denote $\mathbb{P}(A_1 \backslash A_{y1})$ as V'. The expectation of Y (under $\mathbb{P}_\alpha$) in A_1, A_2 are denoted as $p_{1,\alpha} = \mathbb{E}_\alpha(Y|A_1)$, $p_{2,\alpha} = \mathbb{E}_\alpha(Y|A_2)$. Denote the set difference as $A_{y1} \backslash A_1 \triangleq B$, $A_1 \backslash A_{y1} \triangleq C$. Further denote the expectation of Y (under $\mathbb{P}_\alpha$) in B, C as

$$p_{B,\alpha} = \mathbb{E}_\alpha(Y|X \in B), \quad p_{C,\alpha} = \mathbb{E}_\alpha(Y|X \in C).$$

Because $\mathbb{P}_\alpha(A_1) = V_1 = \mathbb{P}_\alpha(A_{y1})$ and $\mathbb{P}_\alpha(A_2) = V_2 = \mathbb{P}_\alpha(A_{y2})$, we have

$$p_{1,\alpha} = y_{1,\alpha} + (p_{C,\alpha} - p_{B,\alpha})\frac{V'}{V_1}, \;\; p_{2,\alpha} = y_{2,\alpha} + (p_{B,\alpha} - p_{C,\alpha})\frac{V'}{V_2}.$$

Therefore

$$|p_{2,\alpha} - p_{1,\alpha}| \ge |y_{2,\alpha} - y_{1,\alpha}| - |p_{B,\alpha} - p_{C,\alpha}|\Big(\frac{V'}{V_1} + \frac{V'}{V_2}\Big). \tag{35}$$

We first show $|p_{B,\alpha} - p_{C,\alpha}| \le |y_{2,\alpha} - y_{1,\alpha}|$. Because $\mathbb{E}(Y|X)$ is a monotonically increasing function of $a^T X + b$, we have $p_{B,\alpha} \le y_{2,\alpha}$. We then compare $p_{B,\alpha}$ and $y_{1,\alpha}$. Let $t_B = \inf_{X \in B}(a^T X + b)$ and A'_{y1} be

$$A'_{y1} = \left\{X \in A_{y1} : a^T X + b \ge t_B\right\}.$$

That is, A'_{y1} is the subset of A_{y1} whose $a^T X + b$ values are no smaller than the infimum of $a^T X + b$ values in B. Let $y'_{1,\alpha} = \mathbb{E}_\alpha(Y|X \in A'_{y1})$, Since $\mathbb{E}(Y|X)$ is a monotonically increasing function of $a^T X + b$, we have $y_{1,\alpha} \le y'_{1,\alpha}$. Because the marginal of $\mathbb{P}_\alpha$ is uniform on Ω, we have

$$p_{B,\alpha} = \frac{1}{\mathbb{P}_\alpha(B)} \int_{t_B}^{t} \frac{\alpha\phi(z)}{1 - \phi(z) + \alpha\phi(z)} \rho_B(z) dz,$$

where

$$\rho_B(z) = \lim_{\Delta z \to 0} \frac{\mathbb{P}_\alpha(X \in B : a^T X + b \in (z - \Delta z, z))}{\Delta z}.$$

Similarly, we have

$$y'_{1,\alpha} = \frac{1}{\mathbb{P}_\alpha(A'_{y1})} \int_{t_B}^{t} \frac{\alpha\phi(z)}{1 - \phi(z) + \alpha\phi(z)} \rho'_1(z) dz,$$

where

$$\rho'_1(z) = \lim_{\Delta z \to 0} \frac{\mathbb{P}_\alpha(X \in A'_{y1} : a^T X + b \in (z - \Delta z, z))}{\Delta z}.$$

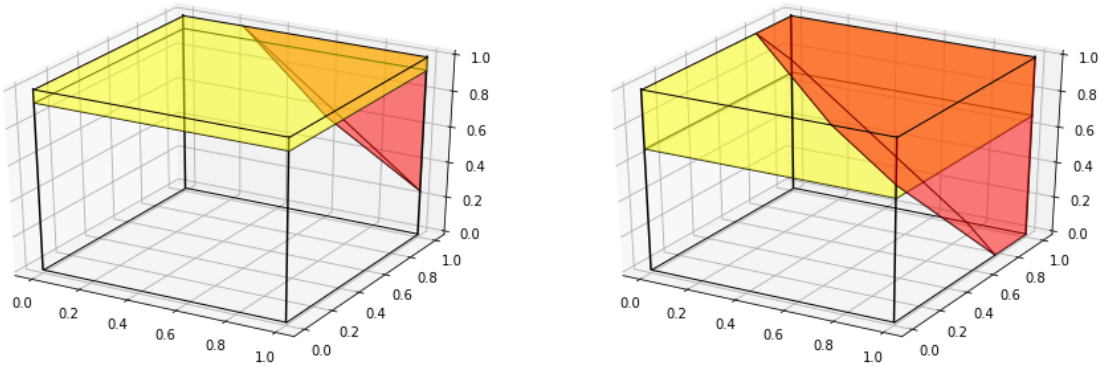
Both $p_{B,\alpha}$ and $y'_{1,\alpha}$ are weighted averages of $\alpha\phi(z)/[1 - \phi(z) + \alpha\phi(z)]$ from t_B to t, with the weights being $\rho_B(z)/\mathbb{P}_\alpha(B)$ and $\rho'_1(z)/\mathbb{P}_\alpha(A'_{y1})$, respectively. By the construction of sets B and A'_{y1}, we can see $[\rho'_1(z)/\mathbb{P}_\alpha(A'_{y1})]/[\rho_B(z)/\mathbb{P}_\alpha(B)]$ is decreasing at (t_B, t). Thus we have $y'_{1,\alpha} \le p_{B,\alpha}$. Recalling that $y_{1,\alpha} \le y'_{1,\alpha}$ and $p_{B,\alpha} \le y_{2,\alpha}$, we have

$$y_{1,\alpha} \le p_{B,\alpha} \le y_{2,\alpha}.$$

Similarly, we have $y_{1,\alpha} \le p_{C,\alpha} \le y_{2,\alpha}$. Thus we have proved $|p_{B,\alpha} - p_{C,\alpha}| \le |y_{2,\alpha} - y_{1,\alpha}|$. Combining this with equation (26) and (35), c_1 can be lower bounded by

$$c_1 \ge 1 - \frac{|p_{B,\alpha} - p_{C,\alpha}|(V'/V_1 + V'/V_2)}{|y_{2,\alpha} - y_{1,\alpha}|} \ge 1 - \frac{V'}{V_1} - \frac{V'}{V_2}. \tag{36}$$

We then consider $V'/V_1 + V'/V_2$. For simplicity of notation, we can assume A is the hypercube $[0,1]^d$. If A is not this hypercube, we can make a linear transformation on X so that A becomes hypercube $[0,1]^d$ under the transformed X. All the conditions in this example are invariant to linear transformation. It suffices to consider the worse case scenarios when $V'/V_1 + V'/V_2$ are largest, which is when a is parallel to $(1, 1, \dots 1)^T$. In that case, the angle between a and each feature is equally large and the least angle between a and all features reaches the maximum. We discuss $V'/V_1 + V'/V_2$ of the worst case scenario in two situations: $h \leq 1/d!$ and $h > 1/d!$.



(a) $h \leq 1/d!$ (b) $h > 1/d!$

Figure 1: Diagrams for two situations when $d = 3$. The set A_{y2} is colored in red while the set A_2 is colored in yellow. The set $A_{y2} \cap A_2$ is colored in orange.

Situation One: $h \leq 1/d!$ In the worst case when a is parallel to $(1, 1, \dots 1)^T$, the set A_{y2} is a hyper-pyramid as shown in Figure 1(a). Denote the length of edge of A_{y2} which is on the edge of hypercube $[0,1]^d$ as e. Because $\mathbb{P}_\alpha(A_2) = \mathbb{P}_\alpha(A_{y2})$, we have

$$h = \frac{e^d}{d!}.$$

Therefore

$$\frac{V'}{V_1} + \frac{V'}{V_2} = \frac{V'}{(1-h)h} = \frac{\left[1 - h^{\frac{d-1}{d}}(d!)^{-\frac{1}{d}}\right]^d}{1-h} = 1 - d(d!)^{-\frac{1}{d}} h^{\frac{d-1}{d}} + h + o(h).$$

Because $h \in (0, 1/2]$, we have $h^{d-1}d \geq g$. Let $g(d) = d(d!)^{-\frac{1}{d}}$. Then

$$\log g(d) = \log d - \frac{1}{d}\sum_{i=1}^{d} \log i = \frac{1}{d}\sum_{i=1}^{d} (\log d - \log i).$$

Therefore $g(d)$ is an increasing function. The minimal number of dimensions to consider feature selection is $d = 2$, so we have $g(d) \geq g(2) = \sqrt{2}$. Therefore we have

$$\frac{V'}{V_1} + \frac{V'}{V_2} \leq 1 - (\sqrt{2} - 1)h + o(h).$$

Thus when $h \leq 1/d!$ we have

$$c_1 \geq (\sqrt{2} - 1)h + o(h).$$

Situation Two: $h > 1/d!$ Because $h \in (0, 1/2]$, situation two can only happen if $d \geq 3$. We still consider the worst case scenario when a is parallel to $(1, 1, \dots 1)^T$ as shown in Figure 1(b). Suppose A is partitioned at feature j to obtain A_1, A_2, then we say the feature j is the direction of **height**, while all the other features

form the space of **base**. Denote the volume of cross section between hyperplane $X[j] = t$ and A_{y2} as $S(t)$. Then we have

$$h = \int_0^1 S(t)dt, \;\; V' = \int_0^{1-h} S(t)dt.$$

Denote S_{up}, S_{low} as

$$S_{up} = \frac{1}{h}\int_{1-h}^1 S(t)dt, \;\; S_{low} = \frac{1}{1-h}\int_0^{1-h} S(t)dt.$$

That is, S_{up} is the average base area of the set $A_{y2} \cap A_2$, and S_{low} is the average base area of the set C. We have

$$c_1 \geq 1 - \frac{V'}{V_1} - \frac{V'}{V_2} = 1 - \frac{S_{low}(1-h)}{h(1-h)} = 1 - \frac{S_{low}}{h}.$$

Noting $S_{low}(1-h) + S_{up}h = h$, we have

$$c_1 \geq S_{up} - S_{low}. \tag{37}$$

Let $F_d(x)$ be the cumulative distribution function of the Irwin-Hall distribution with parameter d. By definition $F_d(x)$ is the probability that the sum of d independent uniform random variables is no greater than x. Let $z = F_d^{-1}(h)$, then we have $S(1) = F_{d-1}(z)$. Thus $S(t) = F_{d-1}(z+t-1)$, $\forall t \in [0,1]$. Because $S(t)$ is increasing with t, we have $S(1) = F_{d-1}(z) \geq h \geq S(0) = F_{d-1}(z-1)$. Thus $z \in (F_{d-1}^{-1}(h)-1, F_{d-1}^{-1}(h)+1)$. To compare S_{up} and S_{low}, define a new quantity as

$$S_m \triangleq \frac{1}{h}\int_{1-2h}^{1-h} F_{d-1}(z-1+t)dt.$$

Because $F_{d-1}(z)$ is increasing with z, we have $S_m \geq S_{low}$. The difference of S_{up} and S_m can be written as

$$\begin{aligned} S_{up} - S_m =& \frac{1}{h}\int_0^h [F_{d-1}(z-h+t) - F_{d-1}(z-2h+t)]dt \\ \geq& \inf_{t\in[0,h]} [F_{d-1}(z-h+t) - F_{d-1}(z-2h+t)] \\ =& \inf_{t\in[0,h]} \int_0^h F'_{d-1}(z-2h+s+t)ds, \end{aligned}$$

where $F'_{d-1}(z)$ is the density function of the Irwin-Hall distribution with parameter $d-1$. Because $F'_{d-1}(z)$ increases at $[0,(d-1)/2]$ and decreases at $[(d-1)/2, d-1]$, the minimum of $\int_0^h F'_{d-1}(z-2h+s+t)ds$ is achieved at either $t=0$ or $t=h$. For $t=0$,

$$\int_0^h F'_{d-1}(z-2h+s+t)ds = \int_{z-2h}^{z-h} F'_{d-1}(z)ds.$$

For $t=h$,

$$\int_0^h F'_{d-1}(z-2h+s+t)ds = \int_{z-h}^{z} F'_{d-1}(z)ds.$$

Because $F'_{d-1}(z)$ is symmetric with respect to $z = (d-1)/2$, $\int_{z-2h}^{z-h} F'_{d-1}(z)ds$ will be no greater than $\int_{z-h}^{z} F'_{d-1}(z)ds$ if $z-h \leq (d-1)/2$. Noting $z-h = z-F_d(z)$, $\frac{\partial}{\partial z}(z-h) = 1-F'_d(z) \geq 0$ and $[z-F_d(z)]|_{z=d/2} = (d-1)/2$, we have proved

$$S_{up} - S_m \geq \int_{z-2h}^{z-h} F'_{d-1}(z)ds = F_{d-1}(F_d^{-1}(h)-h) - F_{d-1}(F_d^{-1}(h)-2h).$$

Thus

$$S_{up} - S_{low} \geq F_{d-1}(F_d^{-1}(h)-h) - F_{d-1}(F_d^{-1}(h)-2h). \tag{38}$$

Combining equation (37) and (38), we have

$$c_1 \geq F_{d-1}(F_d^{-1}(h)-h) - F_{d-1}(F_d^{-1}(h)-2h).$$

We finish the proof. □

4.2 Example 2

Proof. Let $A = \Omega$, and rectangles A_1, A_2 be obtained by an arbitrary partition at feature $X[1]$ or $X[2]$. Then it's easy to see $\mathbb{P}_\alpha(Y|A_1) = \mathbb{P}_\alpha(Y|A_2) = 0.5$.. Noting $\mathbb{P}_\alpha(Y|A) = \mathbb{P}_\alpha(Y) = 0.5$, the impurity does not change after partition, thus $c_1 = 0$. □

5 Details of Algorithm 1

Algorithm 1 in the main paper is represented in details here.

References

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *Annals of Statistics 24*(3), 1084–1105.

Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. *Annals of Statistics 43*(4), 1716–1741.

Algorithm 3: Detailed Steps of SVR-Tree

Result: Output the fitted tree

Input training data $\{(X_i, Y_i)\}_{i=1}^n$, impurity function $f(\cdot)$, weight for minority class α, SVR penalty parameter λ_n, and maximal number of leaf nodes $\bar{a}_n \in \mathbb{N}$. Let the root node be Ω, and node.X $= \{X_i\}_{i=1}^n$, node.Y $= \{Y_i\}_{i=1}^n$. Let node_queue = [root];

Denote the jth coordinate of the ith sample as $X[j]_i$. For $j = 1, \ldots, d$, sort $\{X_i\}, 1 \le i \le n$ by their jth feature $\{X[j]_i\}, 1 \le i \le n$. Denote the sorted increasing subscripts of $\{X_i\}_{i=1}^n$ as $(j_1, j_2, \cdots j_n)$, i.e., $X[j]_{j_1} \le X[j]_{j_2} \cdots \le X[j]_{j_n}$;

Set $R' = +\infty$. For each partition, we compare the risk of the partitioned tree to R', accepting the new partition if the risk after the new partition is smaller than R';

while node_queue *is not empty and number of leaf nodes* $\le \bar{a}_n$ **do**

  Dequeue the first entity in node_queue, denoting it as node. Denote the sample size in node as n';

  Denote the number of features that have already been partitioned as d'. Rearrange $1, 2, \cdots d$ into a list J_0, such that the first d' elements in J_0 corresponds to indices of features that have already been partitioned. Let $\Delta I_0 = 0$;

  for j *in* J_0 **do**

    For jth feature, denote the pre-sorted subscripts of node.X as $(j'_1, j'_2, \cdots j'_{n'})$;

    for i *in* $1 : n' - 1$ **do**

      Partition the current tree at $X_j = (X[j]_{j'_i} + X[j]_{j'_{i+1}})/2$;

      (Optional) compute the (unsigned) tree impurity decrease $\Delta I_\alpha(T, \mathbb{P}_n)$ for the current partition. If jth feature has already been partitioned, let $\Delta I_0 = \min\{\Delta I_0, \Delta I_\alpha(T, \mathbb{P}_n)\}$; Otherwise, reject the current partition if $\Delta I_\alpha(T, \mathbb{P}_n) < \Delta I_0 + \lambda_n$;

      Compute the risk $\tilde{I}_\alpha(T, \mathbb{P}_n) + \lambda_n r(T)$ for all 4 ways of class label assignment after this partition. Denote the smallest risk of the four trees as $R_{i,j}$ and the corresponding class labels for left and right child as $\text{lab}^l_{i,j}$, $\text{lab}^r_{i,j}$;

    end

    Let $i_0 = \arg\min_i R_{i,j}$, $\hat{x}_j = (X[j]_{j'_{i_0}} + X[j]_{j'_{i_0+1}})/2$, and $R_j = R_{i_0,j}$, $\text{lab}^l_j = \text{lab}^l_{i_0,j}$, $\text{lab}^r_j = \text{lab}^r_{i_0,j}$;

  end

  Let $j_0 = \arg\min_j R_j$, $\hat{x} = \hat{x}_{j_0}$, and $R = R_{j_0}$, $\text{lab}^l = \text{lab}^l_{j_0}$, $\text{lab}^r = \text{lab}^r_{j_0}$;

  if $R < R'$ **then**

    Accept the partition. Let node.left.X $= \{X \in \text{node} : X[j] \le \hat{x}\}$, and node.right.X $= \{X \in \text{node} : X[j] > \hat{x}\}$. Assign node.Y to node.left.Y and node.right.Y according to the assignment of node.X. Assign class labels to two child nodes as: node.left.lab $= \text{lab}^l$, node.right.lab $= \text{lab}^r$;

    Update R': $R' := R$;

    Enqueue node.left, node.right to the end of node_queue;

  else

    Reject the partition;

  end

end