

Minimax Decision Trees via Martingale Approximations

Zhenyuan Zhang*

Hengrui Luo†

Abstract

We develop a martingale-based approach to constructing decision trees that efficiently approximate a target variable through recursive conditioning. We introduce MinimaxSplit, a novel splitting criterion that minimizes the worst-case variance at each step, and analyze its cyclic variant, proving an exponential error decay rate under mild conditions. Our analysis builds upon partition-based martingale approximations, providing new insights into their convergence behavior. Unlike traditional variance-based methods, MinimaxSplit avoids end-cut preference and performs well in noisy settings. We derive empirical risk bounds and also explore its integration into random forests.

Keywords: martingale approximation; convergence rates; tree-based models; non-parametric regression

1 Introduction

Let $d \geq 1$ and (\mathbf{X}, Y) be a coupling (or a joint distribution) on \mathbb{R}^{d+1} , where $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. Throughout, vectors are denoted by bold symbols. We study the general problem of efficiently approximating Y by conditioning on the value of \mathbf{X} using finite partitions. If we adopt the L^2 norm as the minimization criterion, the problem can be formulated as

$$\min_{m \in \mathcal{M}} \mathbb{E}[(Y - m(\mathbf{X}))^2], \quad (1)$$

where \mathcal{M} denotes some class of candidate functions defined on \mathbb{R}^d . In this paper, we focus on the case where \mathcal{M} consists of piecewise constant functions. Generally speaking, the motivation is algorithmic: in many applications, such as decision trees (Breiman et al. (1984)), the goal is to provide an efficient approximation of Y by a function of \mathbf{X} . For instance, if the conditioning on \mathbf{X} is based on a partition of cardinality at most K , we have

$$\mathcal{M} = \left\{ m : \mathbb{R}^d \rightarrow \mathbb{R} \mid m(\mathbf{x}) = \sum_{A \in \pi} m_A \mathbb{1}_{\{\mathbf{x} \in A\}}, \pi \in \mathfrak{P}(\mathbb{R}^d), |\pi| \leq K, m_A \in \mathbb{R} \right\},$$

where $\mathfrak{P}(\mathbb{R}^d)$ denotes the set of all partitions of \mathbb{R}^d . Clearly, the choice $m_A = \mathbb{E}[Y \mid \mathbf{X} \in A]$ is always optimal in (1), so it remains to optimize the partition π . However, finding a direct solution to (1) is generally computationally infeasible, so recursive optimizers are favorable. The common approach is to construct a sequence of nested partitions $\{\pi_k\}_{k \geq 0}$ in $\mathfrak{P}(\mathbb{R}^d)$ that generate the conditioning on \mathbf{X} , where the construction is recursive by sequentially optimizing a certain decision criterion, starting from $\pi_0 = \{\mathbb{R}^d\}$. It does not hurt to consider binary partitions, as the general case follows analogously. We then formulate our main problem as follows.

(P-general). Construct nested binary partitions $\{\pi_k\}_{k \geq 0}$ of \mathbb{R}^d such that $\mathbb{E}[(Y - M_k)^2]$ is small for each k , where

$$M_k(\omega) := \mathbb{E}[Y \mid \mathbf{X} \in A], \quad \text{for } \mathbf{X}(\omega) \in A, A \in \pi_k. \quad (2)$$

*Department of Mathematics, Stanford University. Email: zzy@stanford.edu

†Department of Statistics, Rice University; Computational Research Division, Lawrence Berkeley National Laboratory. Email: hrluo@rice.edu

Clearly, if Y is not a function of \mathbf{X} , the quantity $\mathbb{E}[(Y - M_k)^2]$ has a strictly positive lower bound $\mathbb{E}[\text{Var}(Y | \mathbf{X})]$. Another obstruction arises from atoms of \mathbf{X} , which prevents an effective split of the partitions. The relation (2) and its analogues will be essential to this paper.

The main theme of this paper is providing solutions to **(P-general)** in various settings. Roughly speaking, our main contribution can be summarized as follows.

(Solving P-general). If $\mathbb{E}[Y | \mathbf{X}]$ is well-behaved, then modulo the lower bound $\mathbb{E}[\text{Var}(Y | \mathbf{X})]$ and the atoms of \mathbf{X} , there exists an explicit construction of $\{\pi_k\}_{k \geq 0}$, not depending on the law of (\mathbf{X}, Y) , such that $\mathbb{E}[(Y - M_k)^2]$ decays exponentially in k , where M_k is defined in (2).

An instance of **(Solving P-general)** is Theorem 9 below, where the construction is given by the cyclic MinimaxSplit algorithm detailed in Section 3.1.

In the rest of the Introduction, we illustrate the motivations of **(P-general)** and summarize our contribution via two concrete applications: partition-based martingale approximations and regression trees.

1.1 Partition-based martingale approximations

Let U be a real-valued atomless random variable. A *partition-based martingale approximation* of U is a discrete-time martingale $\{M_k\}_{k \geq 0}$ such that

$$M_k(\omega) := \mathbb{E}[U | U \in A], \quad \text{for } U(\omega) \in A, \quad A \in \pi_k, \quad (3)$$

for some nested partitions $\{\pi_k\}_{k \geq 0}$ of \mathbb{R} , where we assume that the partitions are binary and $\pi_0 = \{\mathbb{R}\}$. The martingale property follows from the tower property of conditional expectations.

The connection to **(P-general)** is as follows. Suppose that $d = 1$ (i.e., \mathbf{X} is real-valued, and will be denoted by X in the following), and the coupling (X, Y) satisfies that X is atomless and Y is a strictly increasing function of X . Since in this case $\mathbb{E}[\text{Var}(Y | X)] = 0$ and X is atomless, we expect that the approximation error $\mathbb{E}[(U - M_k)^2]$ vanishes as $k \rightarrow \infty$. Moreover, each partition in X bijectively maps to a partition in Y , and hence (2) reduces to (3). In other words, in the current setting, **(P-general)** is equivalent to the following problem:

(P-martingale). Given a real atomless random variable U , construct a partition-based martingale approximation $\{M_k\}_{k \geq 0}$ of U such that $\mathbb{E}[(U - M_k)^2] \rightarrow 0$ exponentially in k .

Motivated by a martingale embedding problem, Simons (1970) first introduced the Simons martingale and established the a.s. convergence $M_k \rightarrow U$ (and hence also in L^2), but did not analyze the convergence rate. Another recent motivation for studying the convergence rate of the Simons martingale arises from the construction of powerful e-values in hypothesis testing (Ramdas and Wang, 2024). Zhang et al. (2024) (Lemma 5.6) proved that if $U \in L^{2+\delta}$ for some $\delta > 0$ and $\{M_k\}$ is the Simons martingale, then there exist $C > 0$ and $r \in (0, 1)$ such that $\mathbb{E}[(U - M_k)^2] \leq Cr^k$, and $r < 0.827$ is feasible if U is bounded. Our Theorem 3(ii) improves the rate to $r = 1/2$ and hence provides a tighter theoretical bound. We also show by example that the rate $r = 1/2$ is optimal.

The terminology *martingale approximation* has been used extensively in the probability literature with different meanings. In Rüschendorf (1985), it refers to the best approximation of a random vector by a (single) martingale based on ideas from optimal transport. In Borovskikh and Korolyuk (1997) and Hall and Heyde (2014), it refers to techniques from martingale theory (such as inequalities and CLT rates) with various applications in statistics. Similar techniques are also used in the study of stationary ergodic sequences (Wu and Woodroffe, 2004; Zhao and Woodroffe, 2008) and Markovian walks (Grama et al., 2018). Note that in our setting, Y is a sum of martingale differences, but whose variance is bounded, and hence one cannot apply the martingale CLT and its convergence rates.

1.2 Regression trees

Consider a regression problem where we have a data set $(\mathbf{X}_i, Y_i)_{1 \leq i \leq N}$, where $\mathbf{X}_i \in \mathbb{R}^d$ represents the covariates (or features) and $Y_i \in \mathbb{R}$ represents the responses. We follow van der Vaart and Wellner (2013) and assume the regression model:

$$Y_i = g_*(\mathbf{X}_i) + \varepsilon_i, \quad (4)$$

where $g_* : \mathbb{R}^d \rightarrow \mathbb{R}$ is the true signal function, $\{\mathbf{X}_i\}_{1 \leq i \leq N}$ are i.i.d. sampled from a certain law, and ε_i 's are i.i.d. random (Gaussian) errors with zero mean. In other words, the data set is sampled from some coupling (\mathbf{X}_*, Y_*) where $Y_* | \mathbf{X}_* = g_*(\mathbf{X}_*) + \varepsilon$ with \mathbf{X}_* and ε independent. The function g_* can be estimated by minimizing the empirical L^2 risk (i.e., sum of squares; also called L^2 loss or mean-squared error (MSE) in this paper):

$$\hat{m}_N = \arg \min_{m \in \mathcal{M}} \frac{1}{N} \sum_{i=1}^N (Y_i - m(\mathbf{X}_i))^2, \quad (5)$$

where \mathcal{M} denotes the class of candidate functions for the mean m . The quantity (5) serves as an approximation to $\mathbb{E}[(Y_* - m(\mathbf{X}_*))^2]$. We usually establish the consistency of the estimator \hat{m}_N by showing the convergence in probability:

$$\|\hat{m}_N - g_*\|_2 \xrightarrow{P} 0 \text{ as } N \rightarrow \infty. \quad (6)$$

Here, the underlying probability measure is the true data generating mechanism \mathbb{P} , the joint law of (\mathbf{X}_*, Y_*) .

A decision tree regression model for (4) is a non-parametric model for constructing \hat{m}_N by partitioning the input space into distinct regions A_1, A_2, \dots and fitting a simple model $\hat{m}_{N,i}$ to each region A_i (Breiman et al., 1984). In this case, \hat{m}_N takes the form of

$$\hat{m}_N = \sum_i \hat{m}_{N,i} \mathbb{1}_{A_i} \quad (7)$$

and the \mathcal{M} in (5) is the class of all piecewise constant functions. The construction procedures of A_i 's and $\hat{m}_{N,i}$'s are data dependent, where the regions are associated to a tree grown sequentially by maximizing the chosen decision criteria, and hence the parameters associated with each tree node are chosen recursively.

The recursive nature of the decision tree also indicates the connection to **(P-general)**. Let \mathbb{P}_N denote the empirical measure of the training data and $(\mathbf{X}, Y) \stackrel{\text{law}}{\sim} \mathbb{P}_N$. A (simplified variant of the) decision tree algorithm produces nested binary partitions $\{\pi_k\}_{k \geq 0}$ with axis-aligned borders of the input space \mathbb{R}^d based on the law of (\mathbf{X}, Y) , where $\pi_0 = \{\mathbb{R}^d\}$ and the index k is the *depth* of the tree (Breiman et al., 1984). The prediction M_k given by a tree of depth k is defined by (2):

$$M_k(\omega) = M_{k(N)}(\omega) := \mathbb{E}[Y | \mathbf{X} \in A], \quad \text{for } \mathbf{X}(\omega) \in A, A \in \pi_k.$$

Note that this notation M_k can be interpreted as substituting a random input \mathbf{X} into (7). The estimator of g_* is then given by $\hat{m}_k(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} \in A]$ for $\mathbf{x} \in A, A \in \pi_k$.

To facilitate our discussion, we assume that the number of samples in any $A_i \in \pi_k$ is nonzero and tends to ∞ as $k \rightarrow \infty$ uniformly, and moreover, $N/2^k$ is exponentially growing in k .¹ Such an assumption will not be required in any of the main results in this paper, but is convenient for the explanation here. In this case, we are interested in obtaining the fast convergence rates of (6), a problem formulated as follows.

(P-regression). Under the model (7) and the above assumption, construct a decision tree algorithm such that if the true signal function g_* belongs to a certain wide function class, $\mathbb{E}[\|\hat{m}_k - g_*\|_2^2]$ decays exponentially in k .

The classic CART algorithm (Breiman et al. (1984)) for decision trees involves the following widely adopted greedy algorithm that recursively constructs partitions $\{\pi_k\}_{k \geq 0}$. To partition a set $A \in \pi_k$, the greedy algorithm selects a dimension and a threshold where the set A is split into two subsets such that the total remaining risk is minimized. See (11) below for a formal definition. A simplified variant of CART without pruning, called VarianceSplit, is described in Section 3.1.

In the theoretical analysis of the convergence rate of CART in the regression setting, researchers often assume a lower bound on the variance decay when the algorithm splits each node in the decision tree (Chi et al., 2022; Syrgkanis and Zampetakis, 2020; Mazumder and Wang, 2024; Cattaneo et al., 2024). For example, a prevalent variance decay assumption is known as the *sufficient impurity decrease (SID)* condition introduced by Chi et al. (2022), which requires that each split in the CART algorithm decreases the variance

¹Gordon and Olshen (1984) (Theorem 3.6) imposed a similar set of stronger assumptions that require specific growth rates.

of the parent node by at least a constant factor, leading to exponential decay rates of the MSE of the form r^k for some $r \in (0, 1)$.

The first general result without the variance decay assumption was obtained by Klusowski and Tian (2024). Their main result asserts that modulo model mis-specification, the MSE of CART is upper bounded by $1/k$ at depth k ; see (23) below. In a similar vein, Cattaneo et al. (2024) studied the convergence rates of the *oblique* CART and obtained a $1/k$ upper bound for MSE in the general setting (Theorem 1 therein) as well as an r^k upper bound under further assumptions such as node sizes (Theorem 4 therein).

The unconditional rate guarantee of $1/k$ appears significantly slower than the exponential decay rates under variance decay assumptions. On the other hand, Cattaneo et al. (2022) discussed why CART may have slower-than-polynomial convergence. This phenomenon was known as the *end-cut preference (ECP)* in the literature (see Section 11.8 of Breiman et al. (1984)).

We have chosen to work on the model (7) in the simple regression setting (4), where sequential construction is prevalent (Luo and Pratola, 2023; Luo et al., 2024; Liu et al., 2023). Although there are multiple approaches to construct \hat{m}_N in (7) sequentially (Loh, 2014), we focus on the CART-like decision tree model (Breiman et al., 1984) in the regression setting above. The classification setting has been well studied in the computational complexity and encoding literature (Blanc et al., 2020; O’Donnell et al., 2005) in terms of using uniform bounds on Boolean functions between partial trees and optimal trees.

1.3 Our contribution

Partition-based martingale approximations. We show that various martingale constructions are feasible and enjoy exponential convergence rates. For instance, a classic construction by Simons (1970) is the following: start from $\pi_0 = \{\mathbb{R}\}$ and for each $k \geq 0$ and $A = [a, b) \in \pi_k$ where $-\infty \leq a < b \leq \infty$, split A into $[a, \mathbb{E}[U | U \in A])$ and $[\mathbb{E}[U | U \in A], b)$ to form the sets in π_{k+1} . We show that if U is bounded, the convergence rate of the Simons martingale is $C2^{-k}$, where C does not depend on the law of U . Moreover, the rate 2^{-k} can be shown to be asymptotically optimal.

Other constructions of partition-based martingale approximations involve replacing the conditional mean in the Simons martingale by other criteria such as the median. The explicit constructions are given formally by Definition 1, whose convergence rates are summarized in Theorem 3. All constructions yield explicit exponential rates of convergence in both the bounded and the unbounded cases under moment constraints (Theorem 4). Under further assumptions on the law of U , we also establish exponential rates of Cr^k where r can be made arbitrarily close to $1/4$ (a threshold that cannot be surpassed with binary partitions) while allowing C to depend on the law of U (Theorem 5).

Regression trees. We introduce new splitting criteria, namely the MinimaxSplit algorithm and its multivariate variation, the cyclic MinimaxSplit algorithm (see (14) and (15) below for formal definitions). On one hand, the VarianceSplit algorithm is based on variance reduction; that is, for each split, the remaining *total variance* within the children nodes is minimized (among the splits over all dimensions). On the other hand, by definition, the MinimaxSplit algorithm selects the split that minimizes the *maximum variance* within the children nodes (among splits over all dimensions), hence the term *minimax*. The cyclic MinimaxSplit algorithm selects the same active dimension on each level of the tree and the active dimension cycles over all dimensions.

We show that for the cyclic MinimaxSplit algorithm, the MSE at depth k decays exponentially with rate $2^{-2k/(3d)}$ (given sufficient samples), without any further assumption (Theorem 9). We further develop empirical risk bounds for the cyclic MinimaxSplit algorithm (Theorem 10).

Paper outline. The respective solutions to **(P-general)** in the partition-based martingale approximation and the regression tree settings will be detailed in Sections 2 and 3. All technical proofs are given in Appendix A.

2 On partition-based martingale approximations

In this section, we develop the framework of partition-based martingale approximations and show that a number of constructions feature exponential convergence rates, thus answering **(P-martingale)**.

2.1 Basic concepts

We say that a sequence of finite partitions $\{\pi_k\}_{k \geq 0}$ of \mathbb{R}^d is *nested* if for each $k \geq 0$ and $A \in \pi_{k+1}$, there exists (a unique) $B \in \pi_k$ such that $A \subseteq B$. Consider nested partitions $\{\pi_k\}_{k \geq 0}$ of \mathbb{R} and a real-valued random variable U with a finite second moment. Recall (3):

$$M_k(\omega) := \mathbb{E}[U \mid U \in A], \quad \text{for } U(\omega) \in A, \quad A \in \pi_k, \quad (8)$$

where the MSE $\mathbb{E}[(U - M_k)^2]$ is non-increasing in k by the nested property of $\{\pi_k\}_{k \geq 0}$ and the total variance formula.

A discrete-time real-valued stochastic process $\{M_k\}_{k \geq 0}$ is a *martingale* if for any $0 \leq \ell < k$, $\mathbb{E}[M_k \mid \sigma(M_0, \dots, M_\ell)] = M_\ell$. We use the abbreviation $\sigma(\pi_k)$ to denote the σ -algebra generated by events of the form $\{U \in A\}_{A \in \pi_k}$. If $\Pi := \{\pi_k\}_{k \geq 0}$ is a nested sequence of partitions, $\{\sigma(\pi_k)\}_{k \geq 0}$ is a filtration generated by indicator functions of sets in π_k 's (Doob, 1953). It follows from the tower property of conditional expectations that the sequence $\{M_k\}_{k \geq 0} = \{\mathbb{E}[U \mid \sigma(\pi_k)]\}_{k \geq 0}$ is a martingale (in fact, a Doob martingale (Doob, 1940, 1953)). We call this martingale the Π -based *martingale approximation* of the random variable U , or in general a *partition-based martingale approximation* that approximates the random variable in terms of MSE.

Unless U is a constant, there are different kinds of partition, so there are different partition-based martingale approximations depending on the distribution of U . Our goal in this section is to identify a few partition-based martingale approximations that efficiently approximate U , where the construction algorithm is universal; see **(P-martingale)**. The efficiency criterion is given by the decay of the MSE $\mathbb{E}[(U - M_k)^2]$.

Without loss of generality, we give our construction of a partition of a generic interval $[a, b) \subset \mathbb{R}$, where $a, b \in \mathbb{R} \cup \{\pm\infty\}$.² The sequence of partitions Π in the response space will be constructed recursively, where $\pi_0 = \{\mathbb{R}\}$ and for every $k \geq 0$, each interval $A \in \pi_k$ splits into two intervals, by following the same construction, forming the elements in π_{k+1} . In the following, we introduce four distinct splitting rules that define partition-based martingale approximations. For simplicity, we assume that U is atomless, so that the endpoints of the intervals do not matter and that no trivial split occurs.

Definition 1. Suppose we are given an atomless law of U and a non-empty interval $I = [a, b)$, where $a, b \in \mathbb{R} \cup \{\pm\infty\}$.

(i) Define

$$u_{\text{var}} = \arg \min_{u \in I} (\mathbb{P}(U \in [a, u)) \text{Var}(U \mid U \in [a, u)) + \mathbb{P}(U \in [u, b)) \text{Var}(U \mid U \in [u, b))). \quad (9)$$

If the minimizer is not unique, we pick the largest minimizer. The *variance* splitting rule (corresponding to the VarianceSplit algorithm in Section 3.1) splits I into the two sets $[a, u_{\text{var}})$ and $[u_{\text{var}}, b)$.

(ii) Define $u_{\text{Simons}} = \mathbb{E}[U \mid U \in I]$. The *Simons* splitting rule splits I into the two sets $[a, u_{\text{Simons}})$ and $[u_{\text{Simons}}, b)$.

(iii) Define

$$u_{\text{minimax}} = \arg \min_{u \in I} \max \{ \mathbb{P}(U \in [a, u)) \text{Var}(U \mid U \in [a, u)), \mathbb{P}(U \in [u, b)) \text{Var}(U \mid U \in [u, b)) \}.$$

If the minimizer is not unique, we pick the largest minimizer. The *minimax* splitting rule (corresponding to the MinimaxSplit algorithm in Section 3.1) splits I into the two sets $[a, u_{\text{minimax}})$ and $[u_{\text{minimax}}, b)$.

(iv) Define

$$u_{\text{median}} = \sup \{ u \in I : \mathbb{P}(a \leq U < u) = \mathbb{P}(u \leq U < b) \}.$$

The *median* splitting rule splits I into two sets $[a, u_{\text{median}})$ and $[u_{\text{median}}, b)$.

²Here we slightly abuse notation that $[a, b) = (-\infty, b)$ if $a = -\infty$.

In turn, when applying recursively the variance (resp. Simons, minimax, median) splitting rule starting from $\pi_0 = \{\mathbb{R}\}$, we obtain a nested sequence of partitions $\Pi = \{\pi_k\}_{k \geq 0}$ depending on the law of U . We call the resulting Π -based martingale approximation $\{M_k\}_{k \geq 0} = \{\mathbb{E}[U \mid \sigma(\pi_k)]\}_{k \geq 0}$ the variance (resp. Simons, minimax, median) martingale (with respect to U).

Example 2. If U is uniformly distributed on a compact interval, all four martingales coincide. For example, if $U \sim \text{Uni}[0, 1]$, each of the four martingales from Definition 1 satisfy $M_k = \mathbb{E}[U \mid \sigma(\pi_k)]$, where $\pi_k := \{[j/2^k, (j+1)/2^k) : 0 \leq j < 2^k\}$ and it follows that $\mathbb{E}[(U - M_k)^2] \asymp 4^{-k}$ (where \asymp means up to universal constants).

The intuition for the minimax and median martingales is that at each splitting step, we balance the “sizes” of $U \mathbb{1}_{\{U \in I\}}$ on the two sets. The “size” corresponds to the (unconditional) variance for the minimax martingale and the total probability for the median martingale. Intending to minimize the MSE $\mathbb{E}[(U - M_k)^2]$ at each step k , the variance martingale naturally arises as an algorithm that greedily reduces the remaining risk within U in each iteration through layers.

In a partition-based martingale, the nested partitions $\{\pi_k\}_{k \geq 0}$ can be naturally identified as the vertices of a binary tree. A binary tree is the unique infinite tree (V, E) such that all vertices have degree 3, except for a unique vertex \emptyset called the root, which has degree 2. Denote by V_k the set of all vertices in V with the graph distance from the root equal to k . Then, each vertex $v \in V_k$ can be identified as a set $A \in \pi_k$. With each edge $e = (A_k, A_{k+1}) \in E$ where $A_k \in \pi_k$ and $A_{k+1} \in \pi_{k+1}$, we may associate a coefficient $p_e := \mathbb{P}(U \in A_{k+1}) \in [0, 1]$. With each vertex $v = A_k$, we may associate a location $\ell_v := \mathbb{E}[U \mid U \in A_k]$. It follows that M_k is supported on the discrete points $\{\ell_v\}_{v \in V_k}$ and furthermore,

$$\mathbb{E}[(M_k - M_{k+1})^2] = \sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p_{(v_k, v_{k+1})} (\ell_{v_k} - \ell_{v_{k+1}})^2. \quad (10)$$

This binary tree representation (10) of the risk will be frequently used in this paper. In the next two sections, we discuss two types of results: *uniform* rates for bounded U (Section 2.2) and *non-uniform* rates (Section 2.3) where the asymptotic constant may depend on the possibly unbounded law of U .

2.2 Uniform convergence rates

Theorem 3. *Let U be a $[0, 1]$ -valued atomless random variable and $\{M_k\}_{k \geq 0}$ be one of the four martingale approximations with respect to U given by Definition 1. The following statements hold.*

- (i) *If $\{M_k\}_{k \geq 0}$ is the variance martingale, $\mathbb{E}[(U - M_k)^2] \leq 2.71 \cdot 2^{-2k/3}$.*
- (ii) *If $\{M_k\}_{k \geq 0}$ is the Simons martingale, $\mathbb{E}[(U - M_k)^2] \leq 2^{1-k}$.*
- (iii) *If $\{M_k\}_{k \geq 0}$ is the minimax martingale, $\mathbb{E}[(U - M_k)^2] \leq 0.4 \cdot 2^{-2k/3}$.*
- (iv) *If $\{M_k\}_{k \geq 0}$ is the median martingale, $\mathbb{E}[(U - M_k)^2] \leq 2^{-k}$.*

The general case of a bounded U can be derived by a scaling argument, since the constructions in Definition 1 are scale-invariant. Let us sketch the arguments for the minimax martingale, as the same idea becomes crucial when applied to minimax decision trees we develop in Section 3.

Assuming that U is atomless and supported in $[0, 1]$, we apply the representation (10). By construction, $\sum p_{(v_k, v_{k+1})} = 1$ and $\sum |\ell_{v_k} - \ell_{v_{k+1}}| \leq \sup \text{supp } U - \inf \text{supp } U = 1$. Moreover, by the minimax property and the law of total variance, it is straightforward to verify (with details in Section 3.2) that the variances decay at least geometrically by half at each split at u_{minimax} , i.e., for any $v \in V_k$ with $v' \sim w$ for $v' \in V_{k-1}$,

$$\max_{w \in V_{k+1}, v \sim w} p_{(v, w)} (\ell_v - \ell_w)^2 \leq \frac{1}{2} p_{(v', v)} (\ell_{v'} - \ell_v)^2.$$

Therefore, by induction,

$$\max_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p_{(v_k, v_{k+1})} (\ell_{v_k} - \ell_{v_{k+1}})^2 \leq L 2^{-k} \text{Var}(U) \leq L 2^{-k},$$

where $L > 0$ is a universal constant that may not be the same on each occurrence. By Hölder's inequality,

$$\begin{aligned}
\mathbb{E}[(M_k - M_{k+1})^2] &= \sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) (\ell_{v_k} - \ell_{v_{k+1}})^2 \\
&\leq \left(\sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) \right)^{1/3} \\
&\quad \times \left(\max_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) (\ell_{v_k} - \ell_{v_{k+1}})^2 \sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} |\ell_{v_k} - \ell_{v_{k+1}}| \right)^{2/3} \\
&\leq \left(\max_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) (\ell_{v_k} - \ell_{v_{k+1}})^2 \right)^{2/3} \leq L2^{-2k/3}.
\end{aligned}$$

The rest then follows from the martingale property: $\mathbb{E}[(U - M_k)^2] \leq \sum_{\ell \geq k} \mathbb{E}[(M_\ell - M_{\ell+1})^2] \leq L2^{-2k/3}$.

Finding the *optimal* rate $r \in (0, 1)$ (where $\mathbb{E}[(U - M_k)^2] \leq Lr^k$ for some universal constant L) is also an intriguing question. Observe that one cannot have $r < 1/4$ by Example 2 above. In Examples 11 and 12 below, we show that the rate r in Theorem 3 is indeed optimal for the Simons and median martingales.

2.3 Non-uniform convergence rates

The above Theorem 3 requires that the support of U is bounded and the asymptotic constant depends only on the range of U , $\sup U - \inf U$. We show in the next result that, without the bounded range assumption, non-uniform convergence rates can still be guaranteed.

Theorem 4. *Let U be an atomless random variable with a finite $(2 + \varepsilon)$ -th moment for some $\varepsilon > 0$, and $\{M_k\}_{k \geq 0}$ be one of the four partition-based martingale approximations given by Definition 1. Then there exist constants $r \in (0, 1)$ and $C > 0$, both possibly depending on the law of U , such that*

$$\mathbb{E}[(U - M_k)^2] \leq Cr^k.$$

In particular, $M_k \rightarrow U$ both a.s. and in L^2 .

If U is bounded, Theorem 4 is a special case of Theorem 3. The case of the Simons martingale has been established by Zhang et al. (2024), and our proof of Theorem 4 will follow a similar route. Since the asymptotic constant is allowed to depend on the law of U , the optimal non-uniform rate r might be smaller than the uniform ones (in Theorem 3).

Recall from Section 2.2 that the rates obtained in Theorem 3 are asymptotically optimal for the Simons and median martingales (see Examples 11 and 12). For the variance and minimax martingales, we show in the next result that the rate $r = 1/4$ is optimal under certain regularity conditions on the law of U .

Theorem 5. *Let $\{M_k\}_{k \geq 0}$ be either the minimax or variance martingale converging to U . Suppose that U is bounded and has a bounded and continuous density f with $\inf f > 0$ on $\text{supp } U$, which we assume is a connected interval. Then for any $r > 1/4$, there exists a constant $C > 0$ (depending on r and the law of U) such that*

$$\mathbb{E}[|U - M_k|^2] \leq Cr^k.$$

In other words, under the assumptions in Theorem 5, the MSE has an asymptotic convergence rate of $1/4 + \varepsilon$. Unfortunately, we are unable to remove the assumptions for the density of U in Theorem 5 nor give counterexamples. We conjecture that the same conclusion of Theorem 5 holds without those assumptions.

3 The MinimaxSplit algorithm

The goal of this section is to provide a solution to **(P-regression)** by constructing decision tree algorithms with exponential convergence. Section 3.1 gives the constructions inspired by the minimax martingale construction in Section 2. In traditional regression decision tree settings that minimize (1), models greedily split in order to minimize the sum of variances over the responses of the children nodes (Liu et al., 2023; Breiman et al., 1984), yet we propose to greedily minimize the maximum variance among the children nodes.

In the following, we say that a law on \mathbb{R}^d is *marginally atomless* if its projection onto any of the d dimensions is atomless. We will start from a marginally atomless coupling (that may not arise from empirical measures \mathbb{P}_N) and prove exponential rates in Section 3.2. Section 3.3 builds upon this case, covering more generally the non-marginally atomless setting and, in particular, the empirical risk bounds. Further numerical analysis will be presented in Section 3.4. All proofs can be found in Appendix A.4.

3.1 Formulation of the algorithms

We begin by recapping the greedy splitting regime in the VarianceSplit algorithm to construct an efficient decision tree. Subsequently, we propose alternative CART algorithms, namely the MinimaxSplit algorithm, along with its variation, the cyclic MinimaxSplit algorithm, and make comparisons against the VarianceSplit algorithm. Roughly speaking, the (cyclic) MinimaxSplit algorithms are analogues of the minimax martingale from Definition 1. Hereafter, we denote by $[d] = \{1, \dots, d\}$ for $d \in \mathbb{N}$. We also need the following notion of a splittable set. If the coupling is defined on a continuous space, the split can always be performed; yet, for atomic measures like empirical measures, we may halt decision splits when there is no probability mass remained. Intuitively, in the following decision tree algorithms, the decision trees only split a node if it corresponds to a splittable set. A similar notion is also introduced by Definition 3.1 of Klusowski and Tian (2024).

Definition 3.1. Consider a coupling (\mathbf{X}, Y) on \mathbb{R}^{d+1} . We say a set $A \subseteq \mathbb{R}^d$ is (\mathbf{X}, Y) -non-splittable, or non-splittable, if any of the following occurs:

- $\mathbb{P}(\mathbf{X} \in A) = 0$;
- $\mathbb{P}(\mathbf{X} \in A) > 0$ and $Y \mid \mathbf{X} \in A$ is a constant;
- $\mathbb{P}(\mathbf{X} \in A) > 0$ and $\mathbf{X} \mid \mathbf{X} \in A$ is a constant.

Otherwise, we say that A is (\mathbf{X}, Y) -splittable, or splittable.

Revisiting the VarianceSplit algorithm. The greedy VarianceSplit construction (Section 8.4 of Breiman et al. (1984) and Luo and Li (2024)) proceeds by introducing nested partitions $\{\pi_k\}_{k \geq 0}$ of \mathbb{R}^d with axis-aligned borders, in a way that sequentially and greedily minimizes the total risk $\mathbb{E}[(Y - M_k)^2]$ at each step k (while splitting from the fixed partition π_{k-1}), where M_k is taken as the conditional mean of Y given the σ -algebra generated by the collection $\mathbb{1}_{\{\mathbf{X} \in A\}}$, $A \in \pi_k$. So, the partitions $\{\pi_k\}_{k \geq 0}$ of the input space are nested and we require that each π_k consists of sets that split each element $A \in \pi_{k-1}$ into two hyper-rectangles with axis-aligned borders, unless all covariates and/or all response values on the event $\{\mathbf{X} \in A\}$ are the same, in which case we do not perform split and $A \in \pi_k$.

Formally, suppose that a hyper-rectangle $A = [a_1, b_1] \times \dots \times [a_d, b_d]$ belongs to the partition π_{k-1} and is splittable. In particular, $M_{k-1}(\omega) = \mathbb{E}[Y \mid \mathbf{X} \in A]$ if $\mathbf{X}(\omega) \in A$. To find a split of A at depth k , the VarianceSplit algorithm looks for a covariate $j \in [d]$ and $x_j \in [a_j, b_j]$ such that the remaining risk

$\mathbb{E}[(Y - M_k)^2 \mathbb{1}_{\{\mathbf{X} \in A\}}]$ is the smallest after splitting the A at x_j in covariate j . In other words, one seeks for

$$\begin{aligned}
(j, x_j) &= \arg \min_{\substack{j \in [d] \\ x_j \in [a_j, b_j]}} \left(\mathbb{P}(\mathbf{X} \in A, X_j < x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j < x_j) \right. \\
&\quad \left. + \mathbb{P}(\mathbf{X} \in A, X_j \geq x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j \geq x_j) \right) \\
&= \arg \min_{\substack{j \in [d] \\ x_j \in [a_j, b_j]}} \left(\mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X} \in A, X_j < x_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j < x_j\}}] \right. \\
&\quad \left. + \mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X} \in A, X_j \geq x_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j \geq x_j\}}] \right) \\
&=: \arg \min_{\substack{j \in [d] \\ x_j \in [a_j, b_j]}} (V_{\text{left}} + V_{\text{right}}),
\end{aligned} \tag{11}$$

where we break ties arbitrarily.³ Consequently, the splittable set A splits into the sets $A_L := [a_1, b_1] \times \cdots \times [a_j, x_j] \times \cdots \times [a_d, b_d]$ and $A_R := [a_1, b_1] \times \cdots \times [x_j, b_j] \times \cdots \times [a_d, b_d]$ to form its two descendants in the set π_k , and define

$$M_k(\omega) = \begin{cases} \mathbb{E}[Y \mid \mathbf{X} \in A, X_j < x_j] & \text{if } \mathbf{X}(\omega) \in A_L; \\ \mathbb{E}[Y \mid \mathbf{X} \in A, X_j \geq x_j] & \text{if } \mathbf{X}(\omega) \in A_R. \end{cases} \tag{12}$$

In other words, at depth $k \geq 0$ and after constructing the partition π_k , define M_k by

$$M_k(\omega) = \mathbb{E}[Y \mid \mathbf{X} \in A], \quad \text{if } \mathbf{X}(\omega) \in A, \quad \text{for } A \in \pi_k. \tag{13}$$

This leads to the desired coupling $(\mathbf{X}, Y, \{M_k\}_{k \geq 0})$, where $\{M_k\}_{k \geq 0}$ is a martingale. It is also common practice to weigh the terms V_{left} and V_{right} in (11) by the sample sizes of the left and right children nodes (resulting in a minimization of $n_{\text{left}} V_{\text{left}} + n_{\text{right}} V_{\text{right}}$), for which we call the Weighted Variance algorithm.

The MinimaxSplit algorithm. A common feature of the MinimaxSplit and VarianceSplit algorithms is that they both start from a nested sequence of partitions $\{\pi_k\}_{k \geq 0}$ of \mathbb{R}^d consisting of (at-most) binary splits into hyper-rectangles with axis-aligned borders. The k -th approximation M_k is defined by (13). The two algorithms differ in the way the partition is constructed: in the MinimaxSplit setting, the split is no longer the best split that reduces total variance but instead is the best split that minimizes the maximum variance within the two descendants. Formally, for a splittable set $A = [a_1, b_1] \times \cdots \times [a_d, b_d] \in \pi_{k-1}$, define the split location as

$$\begin{aligned}
(j, \hat{x}_j) &= \arg \min_{\substack{j \in [d] \\ x_j \in [a_j, b_j]}} \max \left\{ \mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X} \in A, X_j < x_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j < x_j\}}], \right. \\
&\quad \left. \mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X} \in A, X_j \geq x_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j \geq x_j\}}] \right\} \\
&= \arg \min_{\substack{j \in [d] \\ x_j \in [a_j, b_j]}} \max \left\{ \mathbb{P}(\mathbf{X} \in A, X_j < x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j < x_j), \right. \\
&\quad \left. \mathbb{P}(\mathbf{X} \in A, X_j \geq x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j \geq x_j) \right\} \\
&= \arg \min_{\substack{j \in [d] \\ x_j \in [a_j, b_j]}} \max \left\{ V_{\text{left}}, V_{\text{right}} \right\},
\end{aligned} \tag{14}$$

³Strictly speaking, the arg min in (11) may not always be attained for general couplings (\mathbf{X}, Y) , unless we assume that \mathbf{X} is marginally either atomless or has a finite support, which will always be the case in our analysis. However, in the general setting, if we allow splits in which atoms can be duplicated and assigned to both nodes (instead of only on the right node as in (11)), arg min can be attained. In this paper, we will implicitly assume that arg min is always attained.

where we break ties arbitrarily. Consequently, the splittable set A splits into the two hyper-rectangles $[a_1, b_1) \times \cdots \times [a_j, \hat{x}_j) \times \cdots \times [a_d, b_d)$ and $[a_1, b_1) \times \cdots \times [\hat{x}_j, b_j) \times \cdots \times [a_d, b_d)$ to form its two descendants in the set π_k . We call this the *MinimaxSplit rule*. After obtaining the partition π_k , define M_k through (13). This leads to a coupling $(\mathbf{X}, Y, \{M_k\}_{k \geq 0})$, where $\{M_k\}_{k \geq 0}$ is a martingale.

The cyclic MinimaxSplit algorithm. The cyclic MinimaxSplit algorithm is a variation of the MinimaxSplit algorithm. Instead of optimizing over all covariates $j \in [d]$ in (14), we cycle through the d covariates as the tree grows. That is, for $k \geq 1$, let $j = j_k = (k - 1 \pmod{d}) + 1 \in [d]$. For all splittable sets $A \in \pi_{k-1}$, we split A in the j -th coordinate to form its two descendants at depth k . The split location is then defined as

$$\begin{aligned} \hat{x}_j &= \arg \min_{x_j \in [a_j, b_j)} \max \left\{ \mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X} \in A, X_j < x_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j < x_j\}}], \right. \\ &\quad \left. \mathbb{E}[(Y - \mathbb{E}[Y \mid \mathbf{X} \in A, X_j \geq x_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j \geq x_j\}}] \right\} \\ &= \arg \min_{x_j \in [a_j, b_j)} \max \left\{ \mathbb{P}(\mathbf{X} \in A, X_j < x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j < x_j), \right. \\ &\quad \left. \mathbb{P}(\mathbf{X} \in A, X_j \geq x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j \geq x_j) \right\}, \end{aligned} \tag{15}$$

where we break ties arbitrarily. Note that if \mathbf{X} is marginally atomless, the minimizer \hat{x}_j is essentially unique, in the sense that $\mathbb{P}(\mathbf{X} \in [a_1, b_1) \times \cdots \times (\hat{x}_j, \hat{x}'_j) \times \cdots \times [a_d, b_d)) = 0$ between two minimizers $\hat{x}_j < \hat{x}'_j$. Define M_k by (13). This leads to the desired coupling $(\mathbf{X}, Y, \{M_k\}_{k \geq 0})$, where again, $\{M_k\}_{k \geq 0}$ is a martingale aiming to approximate Y .

If $d = 1$, the cyclic MinimaxSplit algorithm coincides with the MinimaxSplit algorithm. If $d > 1$, compared to the MinimaxSplit algorithm, the cyclic MinimaxSplit algorithm is technically more tractable. Moreover, since all features are considered equally, the cyclic MinimaxSplit algorithm appears more robust in recovering the underlying geometry of the regression function in the setting of (4) in low dimensions, which we illustrate numerically in Section 3.4. On the other hand, the cyclic MinimaxSplit algorithm may be less effective in higher dimensions because the major dimensions may not be easily identified or reached in the first few splits.

(At-most-binary) tree representation. In the context of regression trees, the same development of the tree representation (10) holds, except that the tree may not be a binary tree due to possible non-splittable sets. Instead, there exists a tree (V, E) that is a sub-tree of a binary tree whose vertices are identified as the sets in the nested partitions. With the same notation leading to (10), we have

$$\mathbb{E}[(M_k - M_{k+1})^2] = \sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p_{(v_k, v_{k+1})} (\ell_{v_k} - \ell_{v_{k+1}})^2.$$

The goal of the next section is to provide a partial answer to **(P-general)** under the cyclic MinimaxSplit rule. Assuming marginal non-atomicity, we show that up to universal constants and model mis-specification, the approximation error decays exponentially with rate r^k for some explicit $r \in (0, 1)$ that depends only on dimension d .

3.2 Cyclic MinimaxSplit risk bound

We first claim that if \mathbf{X} is marginally atomless and $j \in [d]$, the L^2 loss

$$\mathbb{P}(\mathbf{X} \in A, X_j < x_j) \text{Var}(Y \mid \mathbf{X} \in A, X_j < x_j) \tag{16}$$

is non-decreasing and continuous in x_j . To see this, we take $a < b$ and define events $E = E_1 \cup E_2$, where $E_1 = \{\mathbf{X} \in A, X_j < a\}$ and $E_2 = E_1^c = \{\mathbf{X} \in A, a \leq X_j < b\}$. By the total variance formula,

$$\begin{aligned} \text{Var}(Y \mid E) &= \mathbb{E}[\text{Var}(Y \mid \sigma(E_1, E_2)) \mid E] + \text{Var}(\mathbb{E}[Y \mid \sigma(E_1, E_2)] \mid E) \\ &\geq \mathbb{E}[\text{Var}(Y \mid \sigma(E_1, E_2)) \mid E] \geq \mathbb{P}(E_1 \mid E) \text{Var}(Y \mid E_1). \end{aligned} \tag{17}$$

Observe that $\text{Var}(Y | E) = \text{Var}(Y | \mathbf{X} \in A, X_j < b)$ and $\text{Var}(Y | E_1) = \text{Var}(Y | \mathbf{X} \in A, X_j < a)$. Inserting these two expressions back into (17), we obtain

$$\text{Var}(Y | \mathbf{X} \in A, X_j < b) \geq \frac{\mathbb{P}(\mathbf{X} \in A, X_j < a)}{\mathbb{P}(\mathbf{X} \in A, X_j < b)} \text{Var}(Y | \mathbf{X} \in A, X_j < a).$$

Rearranging gives

$$\mathbb{P}(\mathbf{X} \in A, X_j < a) \text{Var}(Y | \mathbf{X} \in A, X_j < a) \leq \mathbb{P}(\mathbf{X} \in A, X_j < b) \text{Var}(Y | \mathbf{X} \in A, X_j < b)$$

and therefore (16) (or the first term inside the maximum in (15)) is non-decreasing; and similarly the second term inside the maximum is non-increasing. Moreover, by the marginally atomless property, both quantities are continuous. Therefore, if the arg min is attained in (15) in the cyclic MinimaxSplit setting, the two terms inside the maximum are equal. Namely,

$$\mathbb{P}(\mathbf{X} \in A, X_j < \hat{x}_j) \text{Var}(Y | \mathbf{X} \in A, X_j < \hat{x}_j) = \mathbb{P}(\mathbf{X} \in A, X_j \geq \hat{x}_j) \text{Var}(Y | \mathbf{X} \in A, X_j \geq \hat{x}_j). \quad (18)$$

Since the events $A_L = \{\mathbf{X} \in A, X_j < \hat{x}_j\}$ and $A_R = \{\mathbf{X} \in A, X_j \geq \hat{x}_j\}$ form a partition of $\{\mathbf{X} \in A\}$, we have again by the total variance formula,

$$\begin{aligned} & \mathbb{P}(\mathbf{X} \in A, X_j < \hat{x}_j) \text{Var}(Y | \mathbf{X} \in A, X_j < \hat{x}_j) + \mathbb{P}(\mathbf{X} \in A, X_j \geq \hat{x}_j) \text{Var}(Y | \mathbf{X} \in A, X_j \geq \hat{x}_j) \\ & \leq \mathbb{P}(\mathbf{X} \in A) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A])^2 \mathbb{1}_{\{\mathbf{X} \in A\}}]. \end{aligned} \quad (19)$$

Combining (18) and (19), we have

$$\begin{aligned} & \max \left\{ \mathbb{P}(\mathbf{X} \in A, X_j < \hat{x}_j) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A, X_j < \hat{x}_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j < \hat{x}_j\}}], \right. \\ & \quad \left. \mathbb{P}(\mathbf{X} \in A, X_j \geq \hat{x}_j) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A, X_j \geq \hat{x}_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j \geq \hat{x}_j\}}] \right\} \\ & \leq \frac{1}{2} \mathbb{P}(\mathbf{X} \in A) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A])^2 \mathbb{1}_{\{\mathbf{X} \in A\}}]. \end{aligned} \quad (20)$$

In other words, the remaining risk on each descendant is at most half the risk on the parent node. Inductively, we gain a uniform geometric decay of the maximum risk among the nodes at depth k :

$$\max_{A \in \pi_k} \mathbb{P}(\mathbf{X} \in A) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A])^2 \mathbb{1}_{\{\mathbf{X} \in A\}}] \leq 2^{-k} \text{Var}(Y). \quad (21)$$

This is the key to controlling the “sizes of the nodes”.

In the case of the VarianceSplit construction, such uniform control is reminiscent—counterexamples exist due to ECP, which explains why assumptions such as the SID condition are favorable to ensure risk decay when using loss functions like (11) (Chi et al., 2022; Mazumder and Wang, 2024). To see the intuition, we consider the event $\{\mathbf{X} \in A\}$ and sub-events $A_L = \{\mathbf{X} \in A, X_j < \hat{x}_j\}$ and $A_R = \{\mathbf{X} \in A, X_j \geq \hat{x}_j\}$. Applying the total variance formula to this partition yields

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}[\text{Var}(Y | \sigma(A_L))] + \text{Var}(\mathbb{E}[Y | \sigma(A_L)]) \\ &= \text{Var}(Y | A_L) \mathbb{P}(A_L) + \text{Var}(Y | A_R) \mathbb{P}(A_R) \\ & \quad + (\mathbb{E}[Y | A_L] - \mathbb{E}[Y | A_R])^2 (1 - \mathbb{P}(A_L)) \mathbb{P}(A_L). \end{aligned} \quad (22)$$

From this, we see that when ECP occurs, $\mathbb{P}(A_L) \approx 1$ or 0 , the improvement is dominated by the first two terms of (22): $\text{Var}(Y | A_L) \mathbb{P}(A_L) + \text{Var}(Y | A_R) \mathbb{P}(A_R) = V_{\text{left}} + V_{\text{right}}$. Therefore, when ECP occurs, the sum of variances does not necessarily decay fast enough for the VarianceSplit. However, for cyclic MinimaxSplit, $\max\{V_{\text{left}}, V_{\text{right}}\}$ always displays geometric decay (21), hence “avoiding” the ECP phenomenon. In other words, *the fast decay of the sum of variance is not easily guaranteed from the VarianceSplit construction due to ECP; but the geometric decay of the maximum variance can be assured by the MinimaxSplit construction.*

Consider the additive function class

$$\mathcal{G} := \{g(\mathbf{x}) := g_1(x_1) + \dots + g_d(x_d)\},$$

where $\mathbf{x} = (x_1, \dots, x_d)$. For $g \in \mathcal{G}$, define $\|g\|_{\text{TV}}$ as the infimum of $\sum_{i=1}^d \|g_i\|_{\text{TV}}$ over all such additive representations of g . (A continuous version of) Theorem 4.2 of [Klusowski and Tian \(2024\)](#) states that if $(\mathbf{X}, Y, \{M_k\}_{k \geq 0})$ is constructed from the VarianceSplit algorithm, then for each $k \geq 1$,

$$\mathbb{E}[(Y - M_k)^2] \leq \inf_{g \in \mathcal{G}} \left(\mathbb{E}[(Y - g(\mathbf{X}))^2] + \frac{\|g\|_{\text{TV}}^2}{k+3} \right). \quad (23)$$

In the next result, we show that the $1/k$ rate in (23) can be significantly improved to an exponential rate for the cyclic MinimaxSplit algorithm. This also partially answers **(P-general)** from the Introduction. Note that the splitting here is binary because of the marginal non-atomicity condition.

Theorem 6. *Suppose that \mathbf{X} is marginally atomless. Let the coupling $(\mathbf{X}, Y, \{M_k\}_{k \geq 0})$ be constructed from the cyclic MinimaxSplit algorithm. Then uniformly for any $\delta > 0$ and $k \geq 0$,*

$$\begin{aligned} \mathbb{E}[(Y - M_k)^2] \leq \inf_{g \in \mathcal{G}} & \left(\left((1 + \delta) + \frac{2(1 + \delta)(1 + \delta^{-1})}{3 \cdot 2^{\lfloor k/d \rfloor / 3}} \right) \mathbb{E}[(Y - g(\mathbf{X}))^2] \right. \\ & \left. + (1 + \delta^{-1}) \left(\frac{1}{3} + \left(\frac{1 + \delta^{-1}}{4} \right)^{2/3} \right) 2^{-2 \lfloor k/d \rfloor / 3} \|g\|_{\text{TV}}^2 \right). \end{aligned} \quad (24)$$

Remark 7. Since the bound (24) is uniform over δ, k, d , one can further optimize over δ for fixed k, d . For example, taking $\delta = 2^{-\lfloor k/d \rfloor / 10}$, the right-hand side of (24) then has the more explicit (but cruder) upper bound

$$\inf_{g \in \mathcal{G}} \left(5 \mathbb{E}[(Y - g(\mathbf{X}))^2] + 2^{1 - \lfloor k/d \rfloor / 2} \|g\|_{\text{TV}}^2 \right).$$

Remark 8. In the case where $\|g\|_{\text{TV}}$ is much larger than the range $\Delta g := \sup g - \inf g$, (24) can be further improved to

$$\begin{aligned} \mathbb{E}[(Y - M_k)^2] \leq \inf_{g \in \mathcal{G}} & \left((1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2] + 2^{1/3} (1 + \delta^{-1}) 2^{-2 \lfloor k/d \rfloor / 3} \|g\|_{\text{TV}}^{2/3} \mathbb{E}[(Y - g(\mathbf{X}))^2]^{2/3} \right. \\ & \left. + 2^{-2/3} (1 + \delta^{-1}) 2^{-2 \lfloor k/d \rfloor / 3} \|g\|_{\text{TV}}^{2/3} (\Delta g)^{4/3} \right). \end{aligned} \quad (25)$$

Similar considerations also apply for Theorems 9 and 10 below.

Let us explain the intuition behind Theorem 6 when $d = 1$ and $Y = g(X)$ for some $g \in \mathcal{G}$. Roughly speaking, (24) amounts to

$$\mathbb{E}[(Y - M_k)^2] \leq L 2^{-2k/3} \|g\|_{\text{TV}}^2. \quad (26)$$

To verify (26), observe that $\{M_k\}_{k \geq 0}$ is a martingale converging to Y . We employ the tree representation to control $\mathbb{E}[(M_k - M_{k+1})^2]$. By Hölder's inequality and (21),

$$\begin{aligned} \mathbb{E}[(M_k - M_{k+1})^2] &= \sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) (\ell_{v_k} - \ell_{v_{k+1}})^2 \\ &\leq \left(\sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) \right)^{1/3} \\ &\quad \times \left(\max_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} p(v_k, v_{k+1}) (\ell_{v_k} - \ell_{v_{k+1}})^2 \sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} |\ell_{v_k} - \ell_{v_{k+1}}| \right)^{2/3} \\ &\leq \text{Var}(Y)^{2/3} 2^{-2k/3} \left(\sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} |\ell_{v_k} - \ell_{v_{k+1}}| \right)^{2/3}. \end{aligned}$$

Since each split introduces a difference term $|\ell_{v_k} - \ell_{v_{k+1}}|$, which is bounded by the values of Y conditioned on a partition in the covariate space, we have

$$\sum_{\substack{v_k \in V_k, v_{k+1} \in V_{k+1} \\ v_k \sim v_{k+1}}} |\ell_{v_k} - \ell_{v_{k+1}}| \leq \sum_{A \in \pi_{k+1}} \left(\sup_A g - \inf_A g \right) \leq \|g\|_{\text{TV}}.$$

On the other hand, $\text{Var}(Y) = \text{Var}(g(\mathbf{X})) \leq \|g\|_{\text{TV}}^2$. Altogether (and using the martingale property), we obtain $\mathbb{E}[(Y - M_k)^2] \leq L2^{-2k/3} \|g\|_{\text{TV}}^2$.

Let us compare the rates between (23) and (24). For $Y = g(\mathbf{X})$ with $g \in \mathcal{G}$ and d fixed, we consider $\varepsilon > 0$. For the VarianceSplit algorithm, (23) requires $2^k \approx e^{L\varepsilon^{-1}}$ nodes in the decision tree to guarantee that the MSE is smaller than ε for some $L > 0$. On the other hand, for the cyclic MinimaxSplit algorithm, (24) requires only $2^k \approx \varepsilon^{-Ld}$ nodes. Therefore, as $\varepsilon \rightarrow 0$, the cyclic MinimaxSplit algorithm requires fewer nodes to attain the same risk bound and thus facilitates computation.

Note that Theorem 6 exhibits the curse of dimensionality, which cannot be circumvented in the current setting to the best of our knowledge (Cattaneo et al., 2022; Tan et al., 2022).

3.3 Empirical risk bound for cyclic MinimaxSplit

We now focus on the regression setting and apply Theorem 9 to derive finite-sample performance guarantees. Suppose that under the original law \mathbb{P}_* , \mathbf{X}_* is marginally atomless and

$$Y_* | \mathbf{X}_* = g_*(\mathbf{X}_*) + \varepsilon := \mathbb{E}[Y_* | \mathbf{X}_*] + \varepsilon,$$

where the error $\varepsilon = Y_* - g_*(\mathbf{X}_*)$ is sub-Gaussian, i.e., for some $\sigma > 0$,

$$\mathbb{P}_*(|\varepsilon| \geq u) \leq 2 \exp\left(-\frac{u^2}{2\sigma^2}\right), \quad u \geq 0.$$

Also, we consider the empirical law of (\mathbf{X}, Y) of N samples from \mathbb{P}_* .

A limitation of Theorem 6 is that it applies only if the covariate \mathbf{X} is marginally atomless (for (20) to hold). Therefore, we need the following result that deals with measures whose marginals may contain atoms. This also provides a more complete answer to **(P-general)**.

Theorem 9. *Suppose that \mathbf{X} is purely atomic⁴ and for some $N > 0$,*

$$\max_{j \in [d]} \max_{u \in \mathbb{R}} \mathbb{P}(X_j = u) \leq \frac{1}{N}. \quad (27)$$

Let $M := \sup \text{supp } Y - \inf \text{supp } Y$ and the coupling $(\mathbf{X}, Y, \{M_k\}_{k \geq 0})$ be constructed from the cyclic MinimaxSplit algorithm. Then uniformly for any $\delta > 0$ and $k \geq 0$,

$$\begin{aligned} \mathbb{E}[(Y - M_k)^2] \leq (1 + \delta^{-1})2^{-2\lfloor k/d \rfloor/3} \frac{2^{k+2}M^2}{3N} + \inf_{g \in \mathcal{G}} \left(\left((1 + \delta) + \frac{2(1 + \delta)(1 + \delta^{-1})}{3 \cdot 2^{2\lfloor k/d \rfloor/3}} \right) \mathbb{E}[(Y - g(\mathbf{X}))^2] \right. \\ \left. + (1 + \delta^{-1}) \left(\frac{2}{3} + \left(\frac{1 + \delta^{-1}}{4} \right)^{2/3} \right) 2^{-2\lfloor k/d \rfloor/3} \|g\|_{\text{TV}}^2 \right). \end{aligned} \quad (28)$$

The compensation term

$$(1 + \delta^{-1})2^{-2\lfloor k/d \rfloor/3} \frac{2^{k+2}M^2}{3N}$$

appearing in (28) guarantees that if the number of samples N is much larger than $2^k M^2$, the asymptotic upper bounds (24) and (28) are comparable.

⁴The atomicity condition is only required such that the arg min in (15) is well-defined; see Footnote 3.

Our next result is the following oracle inequality for decision trees under model mis-specification (i.e., when g_* does not belong to \mathcal{G}). This result can be compared with Theorem 4.3 of [Klusowski and Tian \(2024\)](#), where our rate in k is sharper. Let g_k be the output of the cyclic MinimaxSplit construction (that is, $M_k = g_k(\mathbf{X})$). Note that g_k is random under the law \mathbb{P}_* . We use $\|\cdot\|$ to denote the L^2 norm in \mathbb{P}_* , e.g. $\|g_k - g_*\|^2 = \mathbb{E}^{\mathbb{P}_*}[(g_k(\mathbf{X}_*) - g_*(\mathbf{X}_*))^2]$.

Theorem 10. *Assume that \mathbf{X}_* is marginally atomless. We have for $\delta \geq 2^{-2\lfloor k/d \rfloor/3}$,*

$$\mathbb{E}[\|g_k - g_*\|^2] \leq \frac{C2^k(\log N)^2 \log(Nd)}{N} + 2 \inf_{g \in \mathcal{G}} \left(\left((1 + \delta) + \frac{2(1 + \delta)(1 + \delta^{-1})}{3 \cdot 2^{2\lfloor k/d \rfloor/3}} \right) \|g - g_*\|^2 + (1 + \delta^{-1}) \left(\frac{2}{3} + \left(\frac{1 + \delta^{-1}}{4} \right)^{2/3} \right) 2^{-2\lfloor k/d \rfloor/3} \|g\|_{\text{TV}}^2 \right),$$

where $C > 0$ is a constant depending only on $\|g_*\|_\infty$ and σ^2 .

For a fixed sample size N , if we assume that $g_* \in \mathcal{G}$ and $k = 3d \log_2 N / (3d + 2)$ is a multiple of d , Theorem 10 has the further consequence that (under the same assumptions)

$$\mathbb{E}[\|g_k - g_*\|^2] \leq CN^{-\frac{2}{3d+2}} (\|g_*\|_{\text{TV}} + (\log N)^2 \log(Nd)). \quad (29)$$

The proof of Theorem 10 follows essentially the same path as Theorem 4.3 for CART of [Klusowski and Tian \(2024\)](#) while replacing Theorem 4.2 therein by our Theorem 9.

3.4 Numerical experiments

In this section, we provide experiments that verify the advantages of MinimaxSplit and cyclic MinimaxSplit methods in one- and two-dimensional domains, as well as integrated ensemble methods applied to approximating a real-valued function, and therefore it can be used to the application of denoising images. Further numerics (including those in high dimensions) are provided in Appendix B.

3.4.1 One-dimensional input domain

As shown in Figure 1, our first experiment compares decision tree regression methods under varying noise conditions and offers notable insights into their relative performance and characteristics. The optimal tree is obtained using global optimization over all possible split values when conditioning on the tree structure and split coordinates (in our case, split coordinate can only be chosen along one coordinate since $x \in \mathbb{R}$). Specifically, for `max_depth=3`, it employs a fixed structure with $2^3 - 1 = 7$ split points, optimized simultaneously using sequential least squares programming to minimize the overall MSE on the training data. The topology of the tree is predetermined and enforced through constraints in the optimization process. The optimal tree approach contrasts with traditional variance-based, top-down decision tree algorithms by optimizing all splits at once, potentially capturing global data patterns more effectively. However, its fixed structure may limit adaptability to varying data complexities compared to conventional adaptive tree methods. The optimal tree serves as a baseline, particularly in high-noise scenarios, and suggests that its optimization criteria may be more effective at capturing the underlying function while minimizing overfitting.

On the other hand, our MinimaxSplit's competitive performance, especially under high noise, aligns with the theoretical strengths of the minimax optimization in decision trees. Its ability to maintain fidelity to the underlying function in noisy conditions underscores the value of robust splitting criteria in challenging environments. This observation may motivate further exploration of minimax-based approaches in domains with uncertain data quality or high noise levels. As another reference, the standard Scikit-learn DecisionTreeRegressor and the decision tree splitting based on variance minimization show similar performance patterns, with notable degradation in high-noise scenarios. This similarity is expected given their likely use of variance reduction as the splitting criterion. However, their performance lags behind that of MinimaxSplit in noisy conditions, highlighting the flexibility of MinimaxSplit in the presence of significant noise. The varying relative performance suggests that ensemble methods leveraging the strengths of different approaches could yield more robust and adaptive models across diverse data conditions, which we detail in Appendix B. In conclusion, this study provides valuable insight into the behavior of decision tree regression methods under

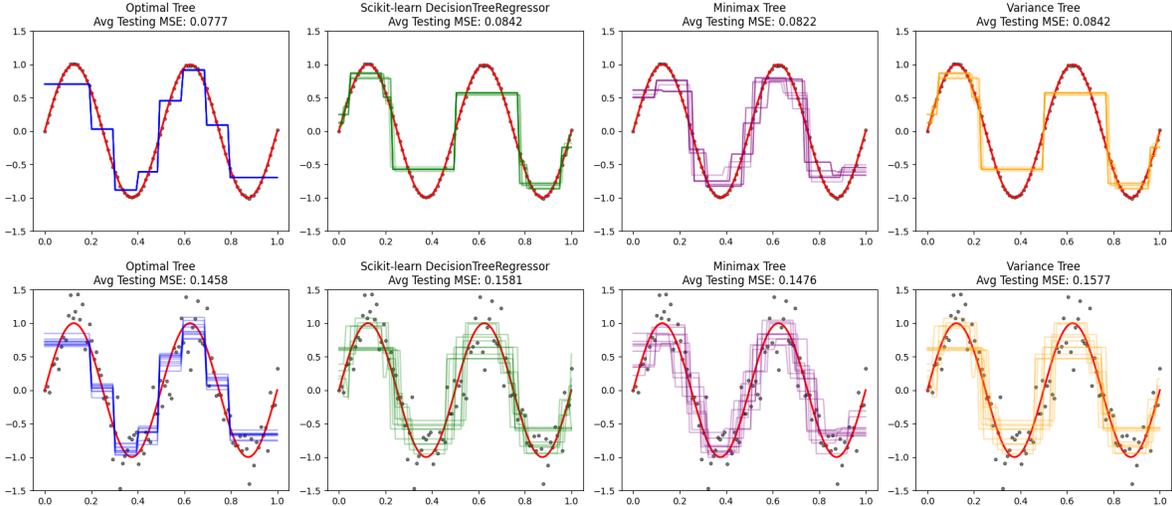


Figure 1: The comparison of the optimal tree, Scikit-learn DecisionTreeRegressor, our MinimaxSplit tree, and VarianceSplit tree on a sinusoidal target function $y = f(x) + \varepsilon = \sin(4\pi x) + \varepsilon$, $x \in [0, 1]$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ with low (first row, noise variance $\sigma_\varepsilon^2 = 0.01$) and high (second row, noise variance $\sigma_\varepsilon^2 = 0.25$) noise levels. We illustrate the true function f using solid red curves and use the same training set for each of the low-noise and high-noise cases; the training set represented by the solid dots is one of the 10 different training sets of size 100. We display 10 different model fits based on 10 different batches of training sets, and the averaged mean-squared error evaluated on a different training set across 10 batches.

varying noise conditions, highlighting the potential for improved performance through custom optimization criteria and robust splitting methods.

3.4.2 Two-dimensional input domain

For this experiment, we use a synthetic dataset generated from a complex, non-linear function of two features, $\mathbf{x} = (x_1, x_2)$. The function is designed to simulate realistic challenges encountered in regression tasks, such as non-linearity and interaction between features. Specifically, the true function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ used for generating the target values is defined as:

$$\begin{aligned}
 f(\mathbf{x}) &= \left(2 - 2.1u_1^2 + \frac{u_1^4}{3}\right) u_1^2 + u_1 u_2 + (-4 + 4u_2^2) u_2^2; \\
 u_1 &= 3(x_1 - 0.5), \\
 u_2 &= 3(x_2 - 0.5),
 \end{aligned}
 \tag{30}$$

where x_1, x_2 are independently and uniformly sampled from $[0, 1]$. This function incorporates both quadratic and higher-order polynomial terms, creating a complex surface with multiple peaks and valleys. The visualizations of the prediction surfaces, as shown in Figure 2, further illustrate the improved accuracy and smoother transitions in the predicted values when using the proposed methods.

In this example (as well as in Sections 3.4.3 and 3.4.4 below), we also incorporate an L^1 variation of the VarianceSplit and MinimaxSplit algorithms, where in the splitting rule, we replace the variances in (11), (14), and (15) by the L^1 distance from the mean. For instance, the L^1 variation of (11) is

$$\begin{aligned}
 (j, x_j) &= \arg \min_{(j, x_j)} \left(\mathbb{P}(\mathbf{X} \in A, X_j < x_j) \mathbb{E}[|Y - \mathbb{E}[Y | \mathbf{X} \in A, X_j < x_j]| | \mathbf{X} \in A, X_j < x_j] \right. \\
 &\quad \left. + \mathbb{P}(\mathbf{X} \in A, X_j \geq x_j) \mathbb{E}[|Y - \mathbb{E}[Y | \mathbf{X} \in A, X_j \geq x_j]| | \mathbf{X} \in A, X_j \geq x_j] \right).
 \end{aligned}
 \tag{31}$$

The optimal L^1 split point may not be unique, in which case we break ties arbitrarily.

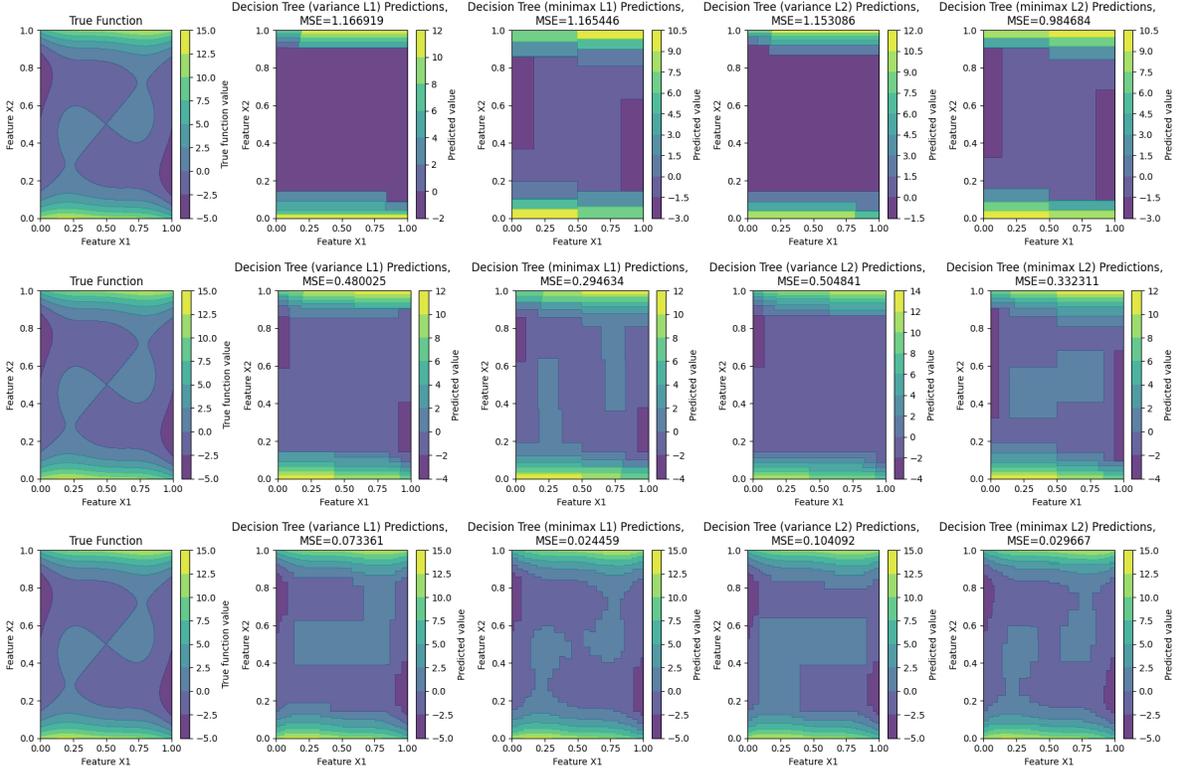


Figure 2: The predictions and true function values are visualized using heatmaps, allowing for a clear comparison of how well each model approximates the underlying function (30). The first subplot shows the heatmap of the true function values across the input space, serving as the benchmark for evaluating the models. The next three subplots depict the predictions of the L^1 VarianceSplit, L^1 MinimaxSplit, L^2 VarianceSplit, and L^2 MinimaxSplit. Top Row: max depth 2; Middle Row: max depth 6; Bottom Row: max depth 10.

3.4.3 Application to denoising images

Another notable experiment involves applying decision trees to denoise noisy images (Luo et al., 2024). We apply the different tree variants to predict pixel values based on their locations, effectively treating this as a regression problem. The experiments include the following configurations:

1. VarianceSplit with L^1 (see (31)) and L^2 (see (11)) norms: These trees use either L^1 or L^2 norms to measure errors. Variance L^1 trees tend to capture more outlier noise, while L^2 norm-based trees provide smoother approximations.
2. MinimaxSplit and cyclic MinimaxSplit: These methods, especially when paired with cyclic padding, are effective in balancing feature splits in high-dimensional image data. The cyclic MinimaxSplit method shows robustness against feature dominance, which often occurred when a few features had much stronger correlations than others.

The results show that mixed methods, where the error calculation alternates between minimax and variance-based splits, lead to a balance between local feature capturing and overall smoothness.

The experiment shown in Figure 3 compares decision tree regression methods for denoising the classic Astronaut (Van der Walt et al., 2014) grayscale image, which has been preprocessed to 128 by 128. Subsequently, Gaussian noise with a standard deviation of 0.25 was added to create the noisy version. Examining the output plot reveals distinct characteristics of each method. The L^1 VarianceSplit approach (RMSE=0.140421) produces a piecewise constant approximation with visible rectangular artifacts, indicating a tendency to oversegment the image. The L^1 MinimaxSplit method (RMSE=0.114668) yields a

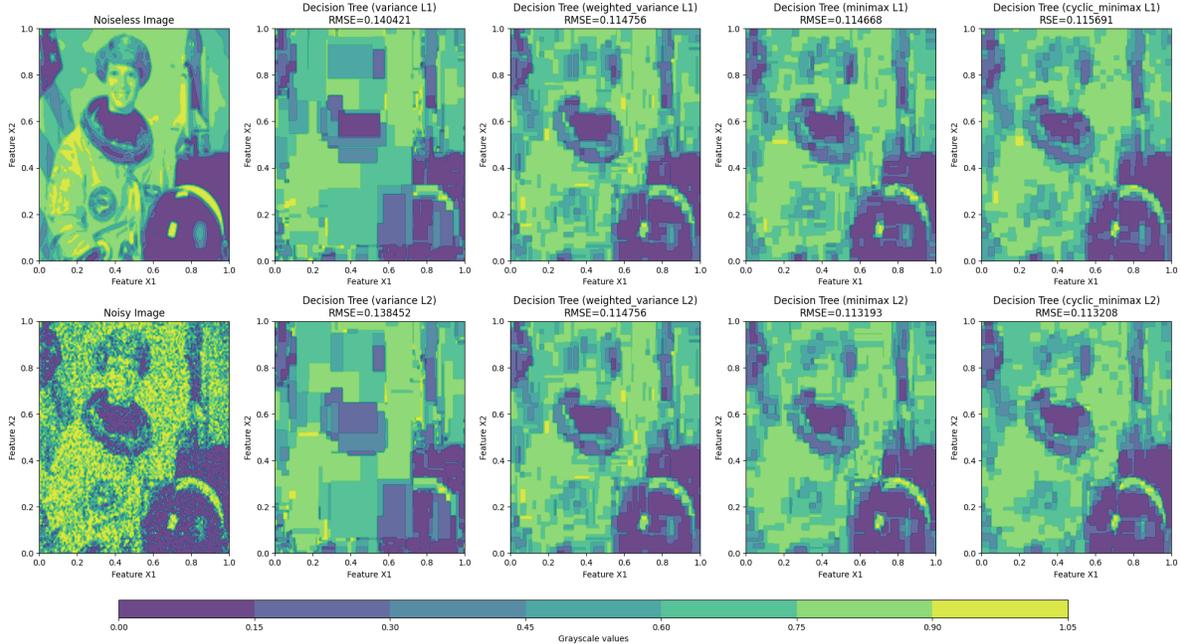


Figure 3: The noiseless image, noisy image and denoised images using different kinds of splitting criteria for a single tree, allowing for a detailed comparison of how well each model approximates the geometric structure. The first subplot shows the heatmap of the true function values across the input space, serving as the benchmark for evaluating the models. The next four subplots depict the predictions of the VarianceSplit, Weighted VarianceSplit, MinimaxSplit, cyclic MinimaxSplit for L^1 (top row) and L^2 (bottom row) decision trees. All trees are fitted with a max depth of 10.

smoother reconstruction with better preservation of large-scale features and edges, particularly evident in the feather of the subject’s outline. The L^2 VarianceSplit method (RMSE=0.138452) results in a slightly smoother output compared to its L^1 counterpart, but still exhibits noticeable blockiness. The L^2 MinimaxSplit approach (RMSE=0.113193) achieves the lowest RMSE, producing a reconstruction that balances smoothness with feature preservation, especially apparent in the gradual shading of the astronaut’s shoulder and the nuanced details around her head. The superior performance of the MinimaxSplit criteria, especially with the L^2 norm, can be attributed to its focus on minimizing worst-case errors. This strategy is particularly effective for image denoising, where it helps preserve important structural elements while smoothing out noise.

Unlike traditional methods that select the best feature at each split based solely on immediate error reduction, the cyclic MinimaxSplit method systematically cycles through the available features as the tree grows. At different levels of the tree, different features are chosen for splitting, based on a pre-determined order or cyclic pattern. This ensures that all features are considered fairly and thus reduces the risk of the model becoming overly dependent on a small subset of features. The cyclic nature emphasizes on modeling each dimension of the underlying function while avoiding heterogeneity across dimensions, which can be particularly beneficial in multivariate cases where feature correlations are complex.

3.4.4 From single tree to ensemble

Although a single decision tree can capture patterns in data through hierarchical partitioning, it often suffers from high variance in predictions, which makes it sensitive to small changes in the training dataset. To address these limitations, an ensemble approach can be utilized to combine multiple decision trees, leading to improved stability and predictive accuracy. In this section, we extend the MinimaxSplit and cyclic MinimaxSplit methods to an ensemble context.

The ensemble method we consider here is the random forest (Breiman et al., 1984), which aggregates

multiple decision trees to improve the precision and robustness of the prediction. Each tree is grown to a fixed depth k , and different combinations of splitting criteria are used to determine how each subset of the data is partitioned. The ensemble predictions are then averaged to provide the final output (see Algorithm 1 in Appendix C).

Specifically, we use the MinimaxSplit, cyclic MinimaxSplit, and VarianceSplit algorithms as the base learners in a random forest setting. Each tree in the ensemble is trained on a bootstrap sample of the original dataset and predictions are made by averaging the output of individual trees, effectively reducing the overall variance of the model. The model aims to maintain a balance between bias and variance by taking advantage of the strengths of each base model while averaging individual weaknesses.

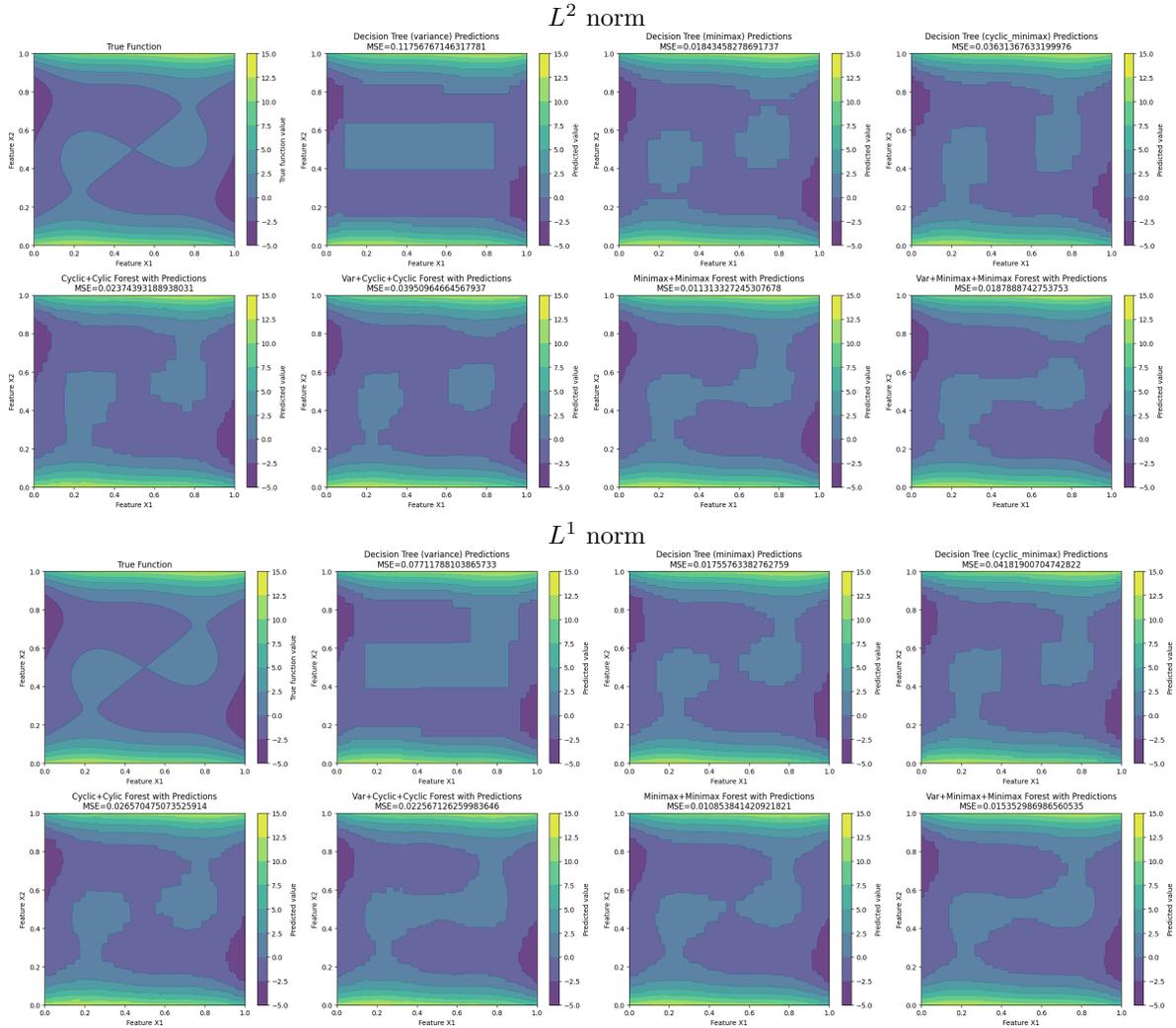


Figure 4: Predictions and true function values of (30), visualized using heatmaps. Top Row: The first subplot shows the heatmap of the true function values across the input space, serving as the benchmark for evaluating the models. The next three subplots depict the predictions of the variance-based decision tree, minimax decision tree, and cyclic minimax decision tree (depth $k = 10$ with L^1 (top) and L^2 (bottom) norm), respectively. Each plot is annotated with the corresponding MSE to quantitatively assess the accuracy. Bottom Row: The final four subplots visualize the predictions from four different random forest models (using Algorithm 1 of Appendix C), each built with varying combinations of the three error methods. Again, each plot is labeled with the MSE to facilitate direct comparison.

We apply the above ensemble approach in the same setting as in Section 3.4.2. The experiment in Figure 4 involves training multiple regression models (VarianceSplit, MinimaxSplit, and cyclic MinimaxSplit) and

ensemble models on the generated dataset. The results demonstrate that the MinimaxSplit+MinimaxSplit ensemble outperforms the traditional variance-based tree in terms of MSE, particularly in scenarios where the data distribution is non-uniform. Further discussions on a weighted aggregation regime will be provided in Appendix B.3.

4 Contribution and Future Work

In this study, we introduced a novel decision tree splitting strategy, the MinimaxSplit algorithm, along with its multivariate variant, the cyclic MinimaxSplit algorithm. Unlike traditional VarianceSplit methods, which aim to reduce overall variance in the children nodes, our MinimaxSplit algorithm seeks to minimize the maximum variance within the split partitions, thereby reducing the risk of overly biased partitions. The cyclic MinimaxSplit algorithm further ensures that each dimension is used in a balanced manner throughout the tree construction process, avoiding dominance by a subset of features.

A key theoretical result of this study is that the cyclic MinimaxSplit algorithm achieves an exponential MSE decay rate of given sufficient data samples, without requiring additional variance decay assumptions (Theorem 9). Furthermore, we derived empirical risk bounds for this method, establishing its robustness in different settings (Theorem 10). Along the same lines of techniques, we prove novel results on the convergence rates of univariate partition-based martingale approximations, which are of their own interest.

In addition to single-tree regression analysis, we explored ensemble learning approaches, where we combined decision trees constructed using MinimaxSplit, cyclic MinimaxSplit, and VarianceSplit methods within a random forest framework. The results demonstrated that a hybrid approach leveraging different splitting techniques yields superior performance, particularly when dealing with non-uniform data distributions (Section 3.4.4). These findings suggest that the MinimaxSplit approach can provide more stable and adaptive decision trees compared to traditional methods, especially in high-dimensional settings.

Future research could explore the MinimaxSplit algorithm in classification tasks, particularly for imbalanced datasets. We expect that an adaptive splitting strategy could replace cyclic selection, dynamically adjusting to feature importance. These directions may refine the theoretical foundations of MinimaxSplit and broaden its practical applications in high-dimensional and complex learning environments.

Acknowledgment

The authors thank Ruodu Wang for his valuable input. We provide our reproducible code at github.com/hrluo. HL was supported by U.S. Department of Energy under Contract DE-AC02-05CH11231 and U.S. National Science Foundation NSF-DMS 2412403.

References

- Guy Blanc, Neha Gupta, Jane Lange, and Li-Yang Tan. Universal guarantees for decision tree induction via a higher-order splitting criterion. *Advances in Neural Information Processing Systems*, 33:9475–9484, 2020.
- Yu V Borovskikh and VS Korolyuk. *Martingale Approximation*. Walter de Gruyter GmbH & Co KG, 1997.
- Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Wadsworth, Inc., 1984.
- Matias D Cattaneo, Jason M Klusowski, and Peter M Tian. On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.10805*, 2022.
- Matias D Cattaneo, Rajita Chandak, and Jason M Klusowski. Convergence rates of oblique regression trees for flexible function libraries. *The Annals of Statistics*, 52(2):466–490, 2024.

- Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.
- Joseph L Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3):455–486, 1940.
- Joseph L Doob. *Stochastic Processes*. John Wiley & Sons, New York, 1953.
- Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.
- Louis Gordon and Richard A Olshen. Almost surely consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 15(2):147–163, 1984.
- Ion Grama, Ronan Lauvergnat, and Émile Le Page. Limit theorems for Markov walks conditioned to stay positive under a spectral gap assumption. *The Annals of Probability*, 46(4):1807–1877, 2018.
- Peter Hall and Christopher C Heyde. *Martingale Limit Theory and Its Application*. Academic press, 2014.
- Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99:75–118, 2015.
- Jason M Klusowski and Peter M Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024.
- Linxi Liu, Dangna Li, and Wing Hung Wong. Convergence rates of a class of multivariate density estimation methods based on adaptive partitioning. *Journal of Machine Learning Research*, 24(50):1–64, 2023.
- Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.
- Hengrui Luo and Meng Li. Ranking perspective for tree-based methods with applications to symbolic feature selection. *arXiv preprint arXiv:2410.02623*, 2024.
- Hengrui Luo and Matthew T Pratola. Sharded Bayesian additive regression trees. *arXiv preprint arXiv:2306.00361*, 2023.
- Hengrui Luo, Akira Horiguchi, and Li Ma. Efficient decision trees for tensor regressions. *arXiv preprint arXiv:2408.01926*, 2024.
- Rahul Mazumder and Haoyue Wang. On the convergence of CART under sufficient impurity decrease condition. *Advances in Neural Information Processing Systems*, 36, 2024.
- Ryan O’Donnell, Michael Saks, Oded Schramm, and Rocco A Servedio. Every decision tree has an influential variable. In *46th annual IEEE symposium on foundations of computer science (FOCS’05)*, pages 31–39. IEEE, 2005.
- Aaditya Ramdas and Ruodu Wang. Hypothesis Testing with E-values. *arXiv preprint arXiv:2410.23614*, 2024.
- Ludger Rüschendorf. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.
- Gordon Simons. A martingale decomposition theorem. *The Annals of Mathematical Statistics*, 41(3):1102–1104, 1970.
- Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on Learning Theory*, pages 3453–3454. PMLR, 2020.
- Yan Shuo Tan, Abhineet Agarwal, and Bin Yu. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 9663–9685. PMLR, 2022.

Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 2013.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

Wei Biao Wu and Michael Woodroffe. Martingale approximations for sums of stationary processes. *The Annals of Probability*, 32(2):1674–1690, 2004.

Zhenyuan Zhang, Aaditya Ramdas, and Ruodu Wang. On the existence of powerful p-values and e-values for composite hypotheses. *The Annals of Statistics*, 52(5):2241–2267, 2024.

Ou Zhao and Michael Woodroffe. On martingale approximations. *The Annals of Applied Probability*, 18(5):1831–1847, 2008.

A Proofs and further results

A.1 Proof of Theorem 3

Variance martingale

Proof for the variance martingale. The first observation is that the split location is precisely the middle point of the two means of the two descendants. This is a special case of Theorem 1 of [Ishwaran \(2015\)](#) applied with f monotone and $X \sim U(0, 1)$, but for completeness, we provide a full argument here with our notation. To see this, consider (without loss of generality by shifting) the law of U with mean zero, and let $u = u_*$ be the optimal split location. Let p_1, p_2 and m_1, m_2 be the associated one-step probabilities and locations in the binary tree representation. Since the second moment $\mathbb{E}[M_1^2]$ is maximized, we must have

$$0 = \frac{d}{du} \mathbb{E}[M_1^2] |_{u_*} = \frac{d}{du} (p_1 m_1^2 + p_2 m_2^2) |_{u_*} . \quad (32)$$

Observe that

$$\frac{dp_1}{du} = -\frac{dp_2}{du} = \frac{d\mathbb{P}_U}{du}$$

and

$$\frac{dm_1}{du} |_{u_*} = \frac{(u_* - m_1)}{p_1} \frac{d\mathbb{P}_U}{du} |_{u_*}; \quad \frac{dm_2}{du} |_{u_*} = \frac{(m_2 - u_*)}{p_2} \frac{d\mathbb{P}_U}{du} |_{u_*} .$$

Inserting into (32) leads to

$$\frac{d\mathbb{P}_U}{du} |_{u_*} \left(m_1^2 - m_2^2 + 2m_1(u_* - m_1) - 2m_2(u_* - m_2) \right) = 0,$$

which simplifies into $m_1 + m_2 = 2u_*$ (in the case $d\mathbb{P}_U/du(u_*) = 0$, we replace u_* with the closest point of u_* to $\text{supp } U$ and this does not change the values of p_1, p_2, m_1, m_2). This fact will be frequently used in the following.

At level k , $\mathbb{E}[(M_{k-1} - M_k)^2]$ is a sum of 2^k terms in the binary tree representation (10), which we may rewrite as $\sum_j p_j d_j^2$ using a change of variable. This corresponds to the 2^k edges (indexed by $[2^k]$) in the binary tree between levels $k-1$ and k , and p_j represents the probabilities of the edge j and d_j is the location difference between the end vertices of the edge, or the *length* of the edge. Our goal is to bound from above

$$\mathbb{E}[(M_{k-1} - M_k)^2] = \sum_{j=1}^{2^k} p_j d_j^2.$$

The key is first to bound the quantity

$$\max_{1 \leq j \leq 2^{k-1}} (p_{2j-1} d_{2j-1}^2 + p_{2j} d_{2j}^2).$$

That is, the variance increases by splitting the j -th mass at level $k-1$ (which is the common ancestor of the $(2j-1)$ -th and $(2j)$ -th masses). By the martingale property,

$$p_{2j-1} d_{2j-1} = p_{2j} d_{2j}.$$

It follows that

$$p_{2j-1} d_{2j-1}^2 + p_{2j} d_{2j}^2 = (p_{2j-1} + p_{2j}) d_{2j-1} d_{2j}. \quad (33)$$

The general strategy to analyze the quantity (33) is to start from an arbitrary mass/node \mathbf{M} at level $k-1$, trace back its ancestors, and bound from above the probability of the mass (which is $p_{2j-1} + p_{2j}$ in the above expression) and the upper bounds for the length of the edges (which is d_{2j-1}, d_{2j}).

The path leading to a mass \mathbf{M} at level $k-1$ in the binary tree representation (10) consists of k nodes: $\emptyset = \mathbf{N}_0, \mathbf{N}_1, \dots, \mathbf{N}_{k-1} = \mathbf{M}$, where $\mathbf{N}_j \in V_j$. We denote the lengths of the edges connecting the node \mathbf{N}_{j-1} to its two descendants on level j by a_j, b_j , $j \in [k-1]$, where $b_j = |\ell_{\mathbf{N}_{j-1}} - \ell_{\mathbf{N}_j}|$. It follows by the martingale property that the mass \mathbf{M} has probability

$$\prod_{j=1}^{k-1} q_j := \prod_{j=1}^{k-1} \frac{a_j}{a_j + b_j}. \quad (34)$$

The lengths of the edges of \mathbf{M} are the mean distances of its children, which are bounded by the distances from $\sum_{j=1}^{k-1} b_j$ (the location of \mathbf{M}) to the endpoints of the interval that \mathbf{M} carries in the splitting process. Therefore, it is natural to update these two distances as we trace the path from the root of the tree down to the node \mathbf{M} .

Let $a_0 = b_0 = 1 \geq \sup U - \inf U$. We consider the following algorithm, which starts from the tuple $(a_0, b_0, a_1, b_1, q_1)$ and updates it $k-1$ times for each $j \in [k-1]$ when we trace the path down to \mathbf{M} . At step j , the tuple represents the two pairs $(a_k, b_k), (a_j, b_j)$ that govern the maximum length of the edges of the j -th node (that is, the distance from its location to the boundaries of the intervals this node represents) for a certain $j \in [k-1]$, as well as the probability weight q_j multiplied at this step. In a generic setting, suppose that we have $(a_k, b_k, a_j, b_j, q_j)$ updated at the node $\mathbf{N}_j \in V_j$. This means that the edges forming a_{k+1}, \dots, a_j are all on one side of the path since only the last one of these contributes to the bound. Now when updating the node at level $j+1$ there are two cases:

- (i) a_{j+1} and a_j are on the same side. We update $(a_k, b_k, a_j, b_j, q_j)$ with $(a_k, b_k, a_{j+1}, b_{j+1}, q_{j+1})$, where $a_{j+1} \leq (a_j + b_j)/2$ and $b_{j+1} \leq (a_k + b_k)/2$ (they are the two edges) and $q_{j+1} = a_{j+1}/(a_{j+1} + b_{j+1})$;
- (ii) a_{j+1} and a_j are on distinct sides. We update $(a_k, b_k, a_j, b_j, q_j)$ with $(a_j, b_j, a_{j+1}, b_{j+1}, q_{j+1})$, where $b_{j+1} \leq (a_j + b_j)/2$, $a_{j+1} \leq (a_k + b_k)/2$, and $q_{j+1} = a_{j+1}/(a_{j+1} + b_{j+1})$.

Here, we have used (34) and the fact that $m_1 + m_2 = 2x_*$ where x_* is the split location.

Let (a, b, a', b', q) be generic entries of the tuple and $(\hat{a}, \hat{b}, \hat{a}', \hat{b}', \hat{q})$ be the updated tuple, where we recall $\hat{q} = \hat{a}'/(\hat{a}' + \hat{b}')$. We claim that

$$\frac{(\hat{a} + \hat{b})(\hat{a}' + \hat{b}')}{(a + b)(a' + b')} \leq \frac{1}{2\hat{q}}. \quad (35)$$

To see this, we separately consider the two cases above.

- In case (i), $\hat{a} = a$ and $\hat{b} = b$, and

$$\hat{a}' + \hat{b}' = a_{j+1} + b_{j+1} = \frac{a_{j+1}}{\hat{q}} \leq \frac{a_j + b_j}{2\hat{q}} = \frac{a' + b'}{2\hat{q}}.$$

- In case (ii), $\hat{a} = a'$ and $\hat{b} = b'$, and

$$\hat{a}' + \hat{b}' = a_{j+1} + b_{j+1} = \frac{a_{j+1}}{\hat{q}} \leq \frac{a_k + b_k}{2\hat{q}} = \frac{a + b}{2\hat{q}}.$$

This proves the claim.

Recall that our goal is to bound (33). The probability of the mass at each step is multiplied by q , and the edge lengths are bounded by $(a + b)/2$ and $(a' + b')/2$. By the above claim, their product multiplies by at most $1/2$ each time we update the tuple. We have thus shown that

$$\max_{1 \leq j \leq 2^{k-1}} (p_{2j-1} d_{2j-1}^2 + p_{2j} d_{2j}^2) \leq 2^{1-k}. \quad (36)$$

Observe also that $\sum_j p_j = 1$ and $\sum_j d_j \leq \sup U - \inf U \leq 1$ (because the edges at a single level cannot intersect each other since they represent the conditional means of U in disjoint intervals). By Hölder's inequality,

$$\begin{aligned} \mathbb{E}[(M_{k-1} - M_k)^2] &= \sum_{j=1}^{2^k} p_j d_j^2 \leq \left(\sum_{j=1}^{2^k} p_j d_j^3 \right)^{2/3} \left(\sum_{j=1}^{2^k} p_j \right)^{1/3} \\ &\leq \left(\max_{j \in [2^k]} (p_j d_j^2) \sum_{j=1}^{2^k} d_j \right)^{2/3} \leq \max_{j \in [2^k]} (p_j d_j^2)^{2/3} \leq 2^{-2(k-1)/3}. \end{aligned} \quad (37)$$

By the martingale property, we then have

$$\mathbb{E}[(U - M_k)^2] = \sum_{j=k}^{\infty} \mathbb{E}[(M_j - M_{j+1})^2] \leq \sum_{j=k}^{\infty} 2^{-2j/3} \leq \frac{2^{-2k/3}}{1 - 2^{-2/3}} \leq 2.71 \cdot 2^{-2k/3},$$

as desired. \square

Simons martingale

Proof for the Simons martingale. It is noted in Zhang et al. (2024) that the Simons martingale satisfies the separated tree condition (i.e., the branches of the tree representation from all different levels, when projected onto the real line as intervals, are either disjoint or have a containment relationship), which is a consequence of the construction. We follow a similar argument to the variance martingale. Consider a node \mathbf{M} at level $k - 1$ and a path $(\emptyset = \mathbf{N}_0, \dots, \mathbf{N}_{k-1} = \mathbf{M})$ leading to \mathbf{M} , where \emptyset denotes the root of a binary tree. We provide an algorithm that recursively defines the tuples (A_j, B_j, q_j) , $0 \leq j \leq k - 1$ for each node \mathbf{N}_j , $0 \leq j \leq k - 1$ on that path, from root to \mathbf{M} . The quantities A_j and B_j represent the maximum possible lengths of the two edges emanating from a node \mathbf{N}_j , and q represents the one-step splitting probability leading to \mathbf{N}_j , based on the constraints from the separated tree property.

We start from $(A_1, B_1, q_1) = (1, 1, 1)$. Again, denote the lengths of the edges connecting the node \mathbf{N}_{j-1} to its two descendants at level j by a_j, b_j , $j \in [k - 1]$, where $b_j = |\ell_{\mathbf{N}_{j-1}} - \ell_{\mathbf{N}_j}|$. Consider the node \mathbf{N}_j at level j . Then (A_j, B_j, q_j) must be of the form (while not distinguishing the order of A_j and B_j)

$$(A_j, B_j, q_j) = \left(b_i - \sum_{\ell=i+1}^j b_\ell, b_j, \frac{a_j}{a_j + b_j} \right)$$

for some $i \leq j - 1$. This happens when the edges a_{i+1}, \dots, a_j lie on the same side of the path $(\mathbf{N}_1, \dots, \mathbf{N}_k)$. For the descendant \mathbf{N}_{j+1} of \mathbf{N}_j , there are two possibilities:

- (i) if a_{j+1} is on the same side of the path as a_j , we define

$$(A_{j+1}, B_{j+1}, q_{j+1}) = \left(b_i - \sum_{\ell=i+1}^{j+1} b_\ell, b_{j+1}, \frac{a_{j+1}}{a_{j+1} + b_{j+1}} \right),$$

where $a_{j+1} \leq b_j$;

(ii) otherwise, we define

$$(A_{j+1}, B_{j+1}, q_{j+1}) = \left(b_j - b_{j+1}, b_{j+1}, \frac{a_{j+1}}{a_{j+1} + b_{j+1}} \right),$$

where $a_{j+1} \leq b_i - \sum_{\ell=i+1}^j b_\ell$.

The choice of $(A_{j+1}, B_{j+1}, q_{j+1})$ is justified by the separated tree property.

Let $q_{j+1} = a_{j+1}/(a_{j+1} + b_{j+1})$. We next express the ratio $A_{j+1}B_{j+1}/(A_jB_j)$ after the update in step $j + 1$ using q_{j+1} . We have in case (i) that

$$\frac{(b_i - \sum_{\ell=i+1}^{j+1} b_\ell)b_{j+1}}{(b_i - \sum_{\ell=i+1}^j b_\ell)b_j} \leq \frac{b_{j+1}}{b_j} \leq \frac{b_{j+1}}{a_{j+1}} = \frac{1}{q_{j+1}} - 1,$$

and in case (ii),

$$\frac{(b_j - b_{j+1})b_{j+1}}{(b_i - \sum_{\ell=i+1}^j b_\ell)b_j} \leq \frac{b_{j+1}}{b_i - \sum_{\ell=i+1}^j b_\ell} \leq \frac{b_{j+1}}{a_{j+1}} = \frac{1}{q_{j+1}} - 1.$$

In both cases, $A_{j+1}B_{j+1}/(A_jB_j) \leq 1/q_{j+1} - 1$. In other words, the contribution of the two edges of \mathbf{M} is upper bounded by

$$A_k B_k \prod_{\ell=1}^{k-1} q_\ell = \prod_{\ell=1}^{k-1} \frac{A_\ell B_\ell}{A_{\ell-1} B_{\ell-1}} \prod_{\ell=1}^{k-1} q_\ell \leq \prod_{\ell=1}^{k-1} \left(\frac{1}{q_\ell} - 1 \right) q_\ell = \prod_{\ell=1}^{k-1} (1 - q_\ell),$$

where we recall that $\{q_\ell\}_{1 \leq \ell \leq k-1}$ are the one-step probabilities leading to \mathbf{M} . On the other hand, we also know that the probability that an edge of \mathbf{M} carries must be bounded by $\prod_{\ell=1}^{k-1} q_\ell$. Therefore, we conclude that, with the decomposition

$$\mathbb{E}[(M_{k-1} - M_k)^2] = \sum_{j=1}^{2^k} p_j d_j^2,$$

it holds that for each $j \in [2^k]$, there exist $\{q_\ell\}_{1 \leq \ell \leq k-1}$ such that

$$p_j d_j^2 \leq \prod_{\ell=1}^{k-1} (1 - q_\ell) \quad \text{and} \quad p_j \leq \prod_{\ell=1}^{k-1} q_\ell.$$

Multiplying the two inequalities leads to

$$p_j d_j \leq \sqrt{\prod_{\ell=1}^{k-1} q_\ell (1 - q_\ell)} \leq 2^{1-k}.$$

Since this holds uniformly in j , we conclude that

$$\mathbb{E}[(M_{k-1} - M_k)^2] = \sum_{j=1}^{2^k} p_j d_j^2 \leq \max_{j \in [2^k]} (p_j d_j) \sum_{j=1}^{2^k} d_j \leq 2^{1-k}.$$

By the martingale property,

$$\mathbb{E}[(M_k - U)^2] = \sum_{j=k}^{\infty} \mathbb{E}[(M_j - M_{j+1})^2] \leq \sum_{j=k}^{\infty} 2^{-j} \leq 2^{1-k}.$$

This completes the proof. \square

Minimax martingale

Proof for the minimax martingale. At level k , there are 2^k vertices (denoted by $j \in [2^k]$) in the binary tree representation of the partition-based martingale approximation. Denote the probabilities of the vertices by p_j (which correspond to $\mathbb{P}(U \in A_j)$, $A_j \in \pi_k$) and the locations by ℓ_j . It follows that

$$\mathbb{E}[(U - M_k)^2] = \sum_{j=1}^{2^k} p_j \mathbb{E}[(U - \ell_j)^2 \mid U \in A_j] =: \sum_{j=1}^{2^k} p_j d_j^2.$$

Note that $\sum_j p_j = 1$ and $\sum_j d_j \leq \sup U - \inf U \leq 1$. It follows analogously as the Hölder's inequality argument in (37) that

$$\mathbb{E}[(U - M_k)^2] \leq \max_{j \in [2^k]} (p_j d_j^2)^{2/3}.$$

Let $\phi(u) = \mathbb{P}(U < u) \text{Var}(U \mid U < u)$ and $\psi(u) = \mathbb{P}(U \geq u) \text{Var}(U \mid U \geq u)$. Note that ϕ is increasing and ψ is decreasing in u , and both functions are continuous if we assume that U is atomless. Recall by the minimax property that at each step we pick a $u \in \mathbb{R}$ such that $\max\{\phi(u), \psi(u)\}$ is minimized. This means that $\phi(u) = \psi(u)$. Since $\text{Var}(U) \geq \phi(u) + \psi(u)$, we have both $\phi(u) \leq \text{Var}(U)/2$ and $\psi(u) \leq \text{Var}(U)/2$. Inductively, we have for any k , $\max(p_j d_j^2) \leq 2^{-k} \text{Var}(U) \leq 2^{-2-k}$. Combining the above, we arrive at

$$\mathbb{E}[(U - M_k)^2] \leq \max_{j \in [2^k]} (p_j d_j^2)^{2/3} \leq 2^{-4/3} 2^{-2k/3} \leq 0.4 \cdot 2^{-2k/3},$$

as desired. \square

Median martingale

Proof for the median martingale. In the same setting as in the minimax case, we write

$$\mathbb{E}[(U - M_k)^2] = \sum_{j=1}^{2^k} p_j d_j^2,$$

where $\sum_j p_j = 1$ and $\sum_j d_j \leq 1$. We further know that $\max_j p_j = 2^{-k}$ by the median splitting construction. Therefore,

$$\mathbb{E}[(U - M_k)^2] = \sum_{j=1}^{2^k} p_j d_j^2 \leq 2^{-k} \sum_{j=1}^{2^k} d_j^2 \leq 2^{-k},$$

as desired. \square

A.2 Examples of optimal rates for partition-based martingale approximations

In the following two examples, we show that the geometric rates in Theorem 3 are optimal for the Simons and median martingales.

Example 11 (Rate $r = 1/2$ is optimal for the Simons martingale). Fix an arbitrary $s \in (1/2, 1)$ and let U satisfy $\mathbb{P}(U = -1) = 1 - s$ and for $J \geq 0$,

$$\mathbb{P}\left(U = \sum_{j=1}^J \left(\frac{1-s}{s}\right)^j\right) = s^{J+1}(1-s).$$

It follows that U is a bounded random variable. Next, we compute that

$$\mathbb{E}\left[U \mid U \geq \sum_{j=1}^J \left(\frac{1-s}{s}\right)^j\right] = \left(\sum_{\ell=J}^{\infty} s^{\ell+1}(1-s)\right)^{-1} \sum_{\ell=J}^{\infty} s^{\ell+1}(1-s) \sum_{j=1}^{\ell} \left(\frac{1-s}{s}\right)^j$$

$$\begin{aligned}
&= s^{-(J+1)} \sum_{\ell=J}^{\infty} s^{\ell+1} (2s-1) \left(1 + \left(\frac{1-s}{s}\right)^\ell\right) \\
&= \frac{s^{-(J+1)} (2s-1) s^{J+1}}{1-s} + \frac{s^{-(J+1)} (2s-1) s}{1-(1-s)} \\
&= \sum_{j=1}^{J+1} \left(\frac{1-s}{s}\right)^j.
\end{aligned}$$

It is then straightforward to verify that in the Simons martingale $\{M_k\}_{k \geq 0}$, each $M_{k+1} - M_k$ is supported on the set $\{0, -((1-s)/s)^k, ((1-s)/s)^{k+1}\}$, where $\mathbb{P}(M_{k+1} - M_k = -((1-s)/s)^k) = (1-s)s^k$ and $\mathbb{P}(M_{k+1} - M_k = ((1-s)/s)^{k+1}) = s^{k+1}$. It follows that

$$\mathbb{E}[(M_{k+1} - M_k)^2] = (1-s)^{2k+1} s^{-(k+1)} \asymp \left(\frac{(1-s)^2}{s}\right)^k.$$

In other words, the rate r is given by $r = (1-s)^2/s$. If s is picked close to $1/2$, r can be made close enough to $1/2$. Therefore, $r = 1/2$ is the optimal rate parameter. This example is illustrated with an atomic distribution U , but a slight twist also leads to an example for an absolutely continuous law U .

Example 12 (Rate $r = 1/2$ is optimal for the median martingale). Let $s \in (0, 1)$ and a bounded random variable U satisfy

$$\mathbb{P}\left(U = \sum_{j=1}^{k-1} s^{j-1} - s^{k-1}\right) = 2^{-k}, \quad k \geq 1.$$

Next, we compute that

$$\begin{aligned}
\mathbb{E}\left[U \mid U \geq \sum_{j=1}^k s^{j-1} - s^k\right] &= 2^k \sum_{\ell=k+1}^{\infty} 2^{-\ell} \left(\sum_{j=1}^{\ell-1} s^{j-1} - s^{\ell-1}\right) \\
&= 2^k \sum_{\ell=k+1}^{\infty} 2^{-\ell} \left(\frac{1}{1-s} - \frac{s^{\ell-1}(2-s)}{1-s}\right) \\
&= \frac{1}{1-s} - \frac{2-s}{s(1-s)} \frac{s^{k+1}/2}{1-s/2} \\
&= \sum_{j=1}^k s^{j-1}.
\end{aligned}$$

Note also that

$$\sum_{j=1}^k s^{j-1} - s^k < \sum_{j=1}^k s^{j-1} < \sum_{j=1}^{k+1} s^{j-1} - s^{k+1}.$$

It is then straightforward to verify that the law of $M_{k+1} - M_k$ is supported on $\{0, \pm s^k\}$ with $\mathbb{P}(M_{k+1} - M_k = s^k) = \mathbb{P}(M_{k+1} - M_k = -s^k) = 2^{-(k+1)}$. We then obtain that

$$\mathbb{E}[(M_{k+1} - M_k)^2] = 2^{-k} s^{2k}.$$

In other words, the rate $r = s^2/2$ has a limit $1/2$ as $s \rightarrow 1$ from below; hence, the optimal rate is $1/2$. Again, a simple twist yields similar examples, where the law of U is absolutely continuous.

A.3 Proofs of Theorems 4 and 5

Proof of Theorem 4. Let $\text{supp } M_1 = \{x, y\}$ where $x < y$. Take $M > \sup\{x, y\}$. Let τ_M^+ be the first hitting time to $\{x : x \geq M\}$, τ_M^- be the first hitting time to $\{x : x \leq -M\}$, and $\tau_M = \min\{\tau_M^+, \tau_M^-\}$. It is straightforward to prove (see Zhang et al. (2024)) that for some $r \in (0, 1)$ and $C > 0$,

$$\mathbb{E}[(U - M_k)^2] \leq \mathbb{E}[\mathbb{1}_{\{\tau_M < \infty\}} U^2] + CM^2 r^k.$$

Since the supports of $\{M_k\}_{k \geq 0}$ correspond to the means of U conditioned on nested partitions of \mathbb{R} , we have $\tau_M^+ < \infty$ implies $M_k \geq x$ for all $k \geq 0$ and $M_1 = y$. Therefore, conditional on the event $\tau_M^+ < \infty$, X is a martingale that is bounded from below by x . By Ville's inequality,

$$\begin{aligned} \mathbb{P}(\tau_M^+ < \infty) &\leq \mathbb{P}(\tau_M^+ < \infty \mid M_1 > x) \\ &= \mathbb{P}\left(\sup_{k \geq 0} M_k \geq M \mid M_1 > x\right) \leq \frac{\mathbb{E}[U \mid M_1 > x] - x}{M - x} = O(M^{-1}). \end{aligned}$$

Similarly, the same analysis holds for τ_M^- . Hence, we conclude that $\mathbb{P}(\tau_M < \infty) = O(M^{-1})$. The rest follows line by line as in [Zhang et al. \(2024\)](#), which we sketch for completeness. By Hölder's inequality, $\mathbb{E}[\mathbb{1}_{\{\tau_M < \infty\}} U^2] = O(M^{-\delta'})$ for some $\delta' > 0$. With the choices $M = r^{-k/(2+\delta')}$ and $q = r^{\delta'/(2+\delta')} \in (0, 1)$, we have that for some $C > 0$,

$$\mathbb{E}[(U - M_k)^2] \leq C(M^{-\delta'} + M^2 r^k) \leq Cq^{-k}.$$

The final statement follows immediately from the martingale convergence theorem. \square

Proof of Theorem 5. First, we claim that the split points for the nested partitions $\{\pi_k\}$ are dense. Indeed, for both martingales, we have shown that the variances on each component in the partition π_k go to zero uniformly as $k \rightarrow \infty$. If the split points were not dense, some element in the partition would cover an interval of positive length, on which the variance of U cannot vanish since we assumed $\inf f > 0$ and $\sup f < \infty$. Therefore, for any $\varepsilon > 0$, there exists k such that the k -th partition π_k is such that on each interval $A \in \pi_k$, $\sup_A f \leq (1 + \varepsilon) \inf_A f$.

Since one-step variance is always better than one-step minimax, it remains to show that for an interval A with $\sup_A f \leq (1 + \varepsilon) \inf_A f$, there exists $C(\varepsilon) \downarrow 1/4$ as $\varepsilon > 0$ such that at each step the total variance reduces by a factor of $C(\varepsilon)$ for the minimax martingale.

To see this, without loss of generality we start from U supported on $[0, 1]$ (meaning that $\mathbb{P}(U \in [0, 1]) = 1$) whose density takes values in $[1, 1 + \varepsilon]$ (we may normalize afterwards to a probability measure, which we omit). Then for any interval $[a, b] \subseteq [0, 1]$,

$$\mathbb{E}[U \mid U \in [a, b]] \in \left[a + \frac{1}{2 + \varepsilon}(b - a), a + \frac{1 + \varepsilon}{2 + \varepsilon}(b - a) \right].$$

Therefore,

$$\begin{aligned} \mathbb{P}(U \in [a, b]) \text{Var}(U \mid U \in [a, b]) &\leq \sup_{m \in [a + \frac{1}{2 + \varepsilon}(b - a), a + \frac{1 + \varepsilon}{2 + \varepsilon}(b - a)]} \int_a^b (1 + \varepsilon)(x - m)^2 dx \\ &\leq \frac{2(1 + \varepsilon)^4}{3(2 + \varepsilon)^3} (b - a)^3. \end{aligned}$$

Similarly,

$$\mathbb{P}(U \in [a, b]) \text{Var}(U \mid U \in [a, b]) \geq \inf_{m \in [a + \frac{1}{2 + \varepsilon}(b - a), a + \frac{1 + \varepsilon}{2 + \varepsilon}(b - a)]} \int_a^b (x - m)^2 dx \geq \frac{1}{12} (b - a)^3.$$

It follows that the split point $x_* \in [0, 1]$ must satisfy

$$\frac{1}{12} x_*^3 \leq \frac{2(1 + \varepsilon)^4}{3(2 + \varepsilon)^3} (1 - x_*)^3 \quad \text{and} \quad \frac{1}{12} (1 - x_*)^3 \geq \frac{2(1 + \varepsilon)^4}{3(2 + \varepsilon)^3} x_*^3.$$

Therefore, for some $a_\varepsilon \rightarrow 0$, $|x_* - 1/2| \leq a_\varepsilon$.

A direct computation yields $\text{Var}(U) \geq \frac{1}{12}$. The total remaining variance $\mathbb{E}[(U - M_1)^2]$ is bounded by

$$\mathbb{E}[(U - M_1)^2] \leq \frac{2(1 + \varepsilon)^4}{3(2 + \varepsilon)^3} (x_*^3 + (1 - x_*)^3) \leq \frac{4(1 + \varepsilon)^4}{3(2 + \varepsilon)^3} \left(\frac{1}{2} + a_\varepsilon\right)^3.$$

Hence, as $\varepsilon \rightarrow 0$, the ratio becomes

$$\frac{\mathbb{E}[(U - M_1)^2]}{\text{Var}(U)} \leq \frac{2(1 + \varepsilon)^4}{(2 + \varepsilon)^3} (1 + 2a_\varepsilon)^3 \rightarrow \frac{1}{4},$$

as desired. \square

A.4 Proofs of results from Section 3

Lemma A.1. *Let $\delta > 0$. Then for any coupling (X, Y) with $X \geq \mathbb{E}[Y]$, it holds that*

$$\text{Var}(Y) \leq (1 + \delta) \mathbb{E}[(X - Y)^2] + \frac{(1 + \delta)^3}{\delta^3} (\sup \text{supp } X - \inf \text{supp } X)^2. \quad (38)$$

Proof. Let $C = (1 + \delta)^3 / \delta^3$. Without loss of generality, we assume that $\mathbb{E}[Y] = 0$. Equivalent to (38) is

$$(1 + \delta) \mathbb{E}[X^2] - 2(1 + \delta) \mathbb{E}[XY] + \delta \mathbb{E}[Y^2] + C(\sup \text{supp } X - \inf \text{supp } X)^2 \geq 0.$$

After completing the square, it remains to show

$$\mathbb{E} \left[\left(\sqrt{\delta} Y - \frac{1 + \delta}{\sqrt{\delta}} X \right)^2 \right] + \left(C(\sup \text{supp } X - \inf \text{supp } X)^2 - \frac{1 + \delta}{\delta} \mathbb{E}[X^2] \right) \geq 0. \quad (39)$$

If $\sup \text{supp } X / \inf \text{supp } X \geq (1 - \sqrt{(1 + \delta)/(C\delta)})^{-1}$, the second term is always non-negative and hence (39) holds. Otherwise, since $C = (1 + \delta)^3 / \delta^3$,

$$(1 + \delta)(\inf \text{supp } X)^2 \geq (1 + \delta)(1 - \sqrt{(1 + \delta)/(C\delta)})(\sup \text{supp } X)^2 \geq (\sup \text{supp } X)^2.$$

It follows that

$$\begin{aligned} \mathbb{E} \left[\left(\sqrt{\delta} Y - \frac{1 + \delta}{\sqrt{\delta}} X \right)^2 \right] &\geq \mathbb{E} \left[\sqrt{\delta} Y - \frac{1 + \delta}{\sqrt{\delta}} X \right]^2 \\ &= \frac{(1 + \delta)^2}{\delta} \mathbb{E}[X^2] \geq \frac{(1 + \delta)^2}{\delta} (\inf \text{supp } X)^2 \geq \frac{(1 + \delta)}{\delta} (\sup \text{supp } X)^2 \geq \frac{1 + \delta}{\delta} \mathbb{E}[X^2]. \end{aligned}$$

This shows (39) and thus proves (38). \square

Proof of Theorem 6. We divide the proof into three steps.

Step I: reducing to an inequality involving $\text{Var}(Y)$. We claim that it suffices to prove

$$\mathbb{E}[(Y - M_k)^2] \leq \inf_{g \in \mathcal{G}} \left((1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2] + (1 + \delta^{-1}) 2^{-2\lfloor k/d \rfloor / 3} (\|g\|_{\text{TV}} \text{Var}(Y))^{2/3} \right). \quad (40)$$

To this end, we need an upper bound of $\text{Var}(Y)$. Fix $g \in \mathcal{G}$ and let $g_* := (\sup g + \inf g)/2$. By Minkowski's inequality,

$$\begin{aligned} \text{Var}(Y)^{1/2} &= \min_{y \in \mathbb{R}} \mathbb{E}[(Y - y)^2]^{1/2} \leq \mathbb{E}[(Y - g_*)^2]^{1/2} \leq \mathbb{E}[(Y - g(\mathbf{X}))^2]^{1/2} + \mathbb{E}[(g(\mathbf{X}) - g_*)^2]^{1/2} \\ &\leq \mathbb{E}[(Y - g(\mathbf{X}))^2]^{1/2} + \frac{1}{2} \|g\|_{\text{TV}}. \end{aligned} \quad (41)$$

Squaring both sides yields that for any $\delta > 0$,

$$\text{Var}(Y) \leq (1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2] + \frac{1 + \delta^{-1}}{4} \|g\|_{\text{TV}}^2,$$

where we have used $(a + b)^2 \leq (1 + \delta)a^2 + (1 + \delta^{-1})b^2$. It follows that

$$\begin{aligned} \|g\|_{\text{TV}}^{2/3} \text{Var}(Y)^{2/3} &\leq \|g\|_{\text{TV}}^{2/3} \left((1 + \delta)^{2/3} \mathbb{E}[(Y - g(\mathbf{X}))^2]^{2/3} + \left(\frac{1 + \delta^{-1}}{4} \right)^{2/3} \|g\|_{\text{TV}}^{4/3} \right) \\ &\leq \frac{2(1 + \delta)}{3} \mathbb{E}[(Y - g(\mathbf{X}))^2] + \left(\frac{1}{3} + \left(\frac{1 + \delta^{-1}}{4} \right)^{2/3} \right) \|g\|_{\text{TV}}^2, \end{aligned} \quad (42)$$

where we have used the inequality $x^{1/3} y^{2/3} \leq x/3 + 2y/3$, which follows from the concavity of logarithm and Jensen's inequality. Inserting (42) into (40) yields (24).

Step II: decomposition of the risk. It remains to establish (40). Denote by $\pi_k = \{I_j\}_{j \in J_k}$ the partition of \mathbb{R}^d formed at level k from the cyclic minimax construction. Note that by construction, $\#J_k \leq 2^k$ and $M_k | \mathbf{X} \in I_j$ is a piecewise constant random variable, which takes a constant value $y_{k,j}$ inside the interval I_j , so $M_k \mathbb{1}_{\{\mathbf{X} \in I_j\}} = y_{k,j} \mathbb{1}_{\{\mathbf{X} \in I_j\}}$. By definition, $y_{k,j} = \mathbb{E}[Y | \mathbf{X} \in I_j]$. Define $g_j = g|_{I_j}$, so that $\|g\|_{\text{TV}} = \sum_j \|g_j\|_{\text{TV}}$. Define $\text{Ran}(g_j) = [\inf g_j, \sup g_j]$ and recall that $\Delta g_j = \sup g_j - \inf g_j \leq \|g_j\|_{\text{TV}}$.

The key step is to decompose the local mean-squared error $\mathbb{E}[(Y - M_k)^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}]$ into two parts for each j , with the help of Lemma A.1. Next, we claim that

$$\mathbb{E}[(Y - M_k)^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}] = \mathbb{E}[(Y - y_{k,j})^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}] \leq U_j + V_j, \quad (43)$$

where U_j, V_j are defined in the following. We divide into two separate cases depending on the location of $y_{k,j}$:

- (i) $y_{k,j} \in \text{Ran}(g_j)$. Define $U_j = (1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}]$ and $V_j = (1 + \delta^{-1}) \mathbb{E}[(g(\mathbf{X}) - y_{k,j})^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}]$.
- (ii) $y_{k,j} \notin \text{Ran}(g_j)$. Define $U_j = (1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}]$ and $V_j = p_j (\Delta g_j)^2 (1 + \delta)^3 / \delta^3$.

In case (i), (43) follows immediately from the polarization identity $(a + b)^2 \leq (1 + \delta)a^2 + (1 + \delta^{-1})b^2$. On the other hand, for case (ii), (43) follows by applying Lemma A.1 to the coupling $(g(\mathbf{X}), Y) | \mathbf{X} \in I_j$ and using that $y_{k,j} = \mathbb{E}[Y | \mathbf{X} \in I_j]$. This completes the proof of (43).

Step III: controlling $\sum_j U_j$ and $\sum_j V_j$. We have by construction and (43) that

$$\mathbb{E}[(Y - M_k)^2] = \sum_{j \in J_k} \mathbb{E}[(Y - M_k)^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}] \leq \sum_{j \in J_k} U_j + \sum_{j \in J_k} V_j. \quad (44)$$

The first term is easy to control by definition:

$$\sum_{j \in J_k} U_j = \sum_{j \in J_k} (1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}] = (1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2]. \quad (45)$$

To control $\sum_j V_j$, we apply the same Hölder's inequality argument as in the proof of Theorem 3 in the minimax case. We have

$$\begin{aligned} \sum_{j \in J_k} V_j &\leq \left(\sum_{j \in J_k} p_j \left(\frac{V_j}{p_j} \right)^{3/2} \right)^{2/3} \left(\sum_{j \in J_k} p_j \right)^{1/3} \\ &\leq \left(\left(\max_{j \in J_k} V_j \right) \sum_{j \in J_k} \sqrt{\frac{V_j}{p_j}} \right)^{2/3} \leq \left(\left(\max_{j \in J_k} \mathbb{E}[(Y - M_k)^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}] \right) \sum_{j \in J_k} \sqrt{\frac{V_j}{p_j}} \right)^{2/3}. \end{aligned} \quad (46)$$

To further bound the right-hand side of (46), we recall from (21) that as a consequence of the marginally atomless property,

$$\max_{j \in J_k} \mathbb{E}[(Y - M_k)^2 \mathbb{1}_{\{\mathbf{X} \in I_j\}}] \leq 2^{-k} \text{Var}(Y). \quad (47)$$

Next, we claim that

$$\sqrt{\frac{V_j}{p_j}} \leq (1 + \delta^{-1})^{3/2} \Delta g_j. \quad (48)$$

This is immediate for case (ii) above (when $y_{k,j} \notin \text{Ran}(g_j)$). By definition and using $y_{k,j} \in \text{Ran}(g_j)$, we have for case (i) above (when $y_{k,j} \in \text{Ran}(g_j)$),

$$V_j \leq (1 + \delta^{-1}) p_j (\Delta g_j)^2 < p_j (1 + \delta^{-1})^3 (\Delta g_j)^2,$$

as desired. On the other hand, suppose that $k = dq + r$ for some integer q and $0 \leq r < d$, and denote by π_k the resulting partition of \mathbb{R}^d . Observe (by induction on q, r , and elementary geometry) that

$$\sum_{A \in \pi_k} \left(\sup_{\mathbf{x} \in A} g(\mathbf{x}) - \inf_{\mathbf{x} \in A} g(\mathbf{x}) \right) \leq 2^{q(d-1)+r} \|g\|_{\text{TV}}.$$

In other words, we have that for $k \in \mathbb{N}$,

$$\sum_{j \in J_k} \Delta g_j = \sum_{A \in \pi_k} \left(\sup_{\mathbf{x} \in A} g(\mathbf{x}) - \inf_{\mathbf{x} \in A} g(\mathbf{x}) \right) \leq 2^{k - \lfloor k/d \rfloor} \|g\|_{\text{TV}}, \quad (49)$$

where $\lfloor \cdot \rfloor$ is the floor function. Combined with (48), we arrive at

$$\sum_{j \in J_k} \sqrt{\frac{V_j}{p_j}} \leq (1 + \delta^{-1})^{3/2} 2^{k - \lfloor k/d \rfloor} \|g\|_{\text{TV}}.$$

Inserting into (46) gives

$$\sum_{j \in J_k} V_j \leq ((1 + \delta^{-1})^{3/2} 2^{k - \lfloor k/d \rfloor} \|g\|_{\text{TV}} 2^{-k} \text{Var}(Y))^{2/3} \leq (1 + \delta^{-1}) 2^{-2 \lfloor k/d \rfloor / 3} (\|g\|_{\text{TV}} \text{Var}(Y))^{2/3}. \quad (50)$$

Finally, inserting (45) and (50) into (44) yields

$$\mathbb{E}[(Y - M_k)^2] \leq (1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2] + (1 + \delta^{-1}) 2^{-2 \lfloor k/d \rfloor / 3} (\|g\|_{\text{TV}} \text{Var}(Y))^{2/3}.$$

This proves (40) and hence concludes the proof. \square

Proof of Remark 8. The bound (41) can be improved to

$$\text{Var}(Y)^{1/2} \leq \mathbb{E}[(Y - g(\mathbf{X}))^2]^{1/2} + \frac{1}{2} \Delta g.$$

Inserting the following inequality (which follows from $(a + b)^p \leq 2^{p-1}(a^p + b^p)$ for $a, b \geq 0$ and $p \geq 1$)

$$\text{Var}(Y)^{2/3} \leq 2^{1/3} \mathbb{E}[(Y - g(\mathbf{X}))^2]^{2/3} + 2^{-2/3} (\Delta g)^{4/3}$$

into (40) yields (25). \square

Proof of Theorem 9. The only difference from the proof of Theorem 6 is (47), since for non-marginally atomless measures (20) and (21) may not hold. Instead, by our assumptions, (20) is replaced by

$$\begin{aligned} & \max \left\{ \mathbb{P}(\mathbf{X} \in A, X_j < \hat{x}_j) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A, X_j < \hat{x}_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j < \hat{x}_j\}}], \right. \\ & \quad \left. \mathbb{P}(\mathbf{X} \in A, X_j \geq \hat{x}_j) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A, X_j \geq \hat{x}_j])^2 \mathbb{1}_{\{\mathbf{X} \in A, X_j \geq \hat{x}_j\}}] \right\} \\ & \leq \frac{1}{2} \mathbb{P}(\mathbf{X} \in A) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A])^2 \mathbb{1}_{\{\mathbf{X} \in A\}}] + \frac{M^2}{N}, \end{aligned}$$

because adding an atom of weight $\leq 1/N$ to a node increases the risk by at most M^2/N . As a consequence, if we denote by

$$u_k := \max_{A \in \pi_k} \mathbb{P}(\mathbf{X} \in A) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A])^2 \mathbb{1}_{\{\mathbf{X} \in A\}}],$$

then u_k satisfies the recursive inequalities

$$u_{k+1} \leq \frac{u_k}{2} + \frac{M^2}{N}, \quad k \geq 0; \quad u_0 = \text{Var}(Y).$$

To solve this, let $v_k := u_k - 2M^2/N$. Suppose that $u_0 \geq 2M^2/N$. Then v_k satisfies $v_{k+1} \leq v_k/2$ and $v_0 \leq \text{Var}(Y)$. It follows that $v_k \leq 2^{-k} \text{Var}(Y)$ for all $k \geq 0$, and hence

$$\max_{A \in \pi_k} \mathbb{P}(\mathbf{X} \in A) \mathbb{E}[(Y - \mathbb{E}[Y | \mathbf{X} \in A])^2 \mathbb{1}_{\{\mathbf{X} \in A\}}] = u_k \leq 2^{-k} \text{Var}(Y) + \frac{2M^2}{N}. \quad (51)$$

It is also easy to check that (51) is also satisfied for the case $u_0 < 2M^2/N$, that is, (51) holds in general. Replacing (47) by (51) leads to

$$\mathbb{E}[(Y - M_k)^2] \leq \inf_{g \in \mathcal{G}} \left((1 + \delta) \mathbb{E}[(Y - g(\mathbf{X}))^2] + (1 + \delta^{-1}) 2^{-2\lfloor k/d \rfloor / 3} \left(\|g\|_{\text{TV}} (\text{Var}(Y) + 2^{k+1} \frac{M^2}{N}) \right)^{2/3} \right).$$

Applying concavity of the function $x \mapsto x^{2/3}$ for $x > 0$, (42), and that

$$\|g\|_{\text{TV}}^{2/3} \left(2^{k+1} \frac{M^2}{N} \right)^{2/3} \leq \frac{\|g\|_{\text{TV}}^2}{3} + \frac{2^{k+2} M^2}{3N},$$

we obtain (28). \square

Proof of Theorem 10. Note that if \mathbf{X}_* is marginally atomless, the assumption (27) for the coupling (\mathbf{X}, Y) in Theorem 9 holds \mathbb{P}_* -almost surely. Consequently, the proof is almost verbatim compared to Theorem 4.3 of Klusowski and Tian (2024), by applying our Theorem 9 instead of Theorem 4.2 therein. The only difference here is the extra term

$$(1 + \delta^{-1}) 2^{-2\lfloor k/d \rfloor / 3} \frac{2^{k+2} M^2}{3N}$$

appearing in (28), where we remind the reader that $M = \sup \text{supp } Y - \inf \text{supp } Y$. Nevertheless, since the proof of Theorem 4.3 of Klusowski and Tian (2024) deals first with the case with bounded data, we may simply replace M by that bound. Specifically, the first part of the proof of Theorem 4.3 of Klusowski and Tian (2024) assumes $\max_i |Y_i| \leq U$. We may set $M = 2U$ and add the term

$$(1 + \delta^{-1}) 2^{-2\lfloor k/d \rfloor / 3} \frac{2^{k+2} M^2}{3N} \leq \frac{2^{k+5} U^2}{3N}$$

in (B.28) therein (where we used $\delta \geq 2^{-2\lfloor k/d \rfloor / 3}$). This extra term is carried until (B.34) therein, where it can be absorbed by the term $U^4 2^k \log(Nd)/N$ therein. The rest of the proof remains unchanged. \square

B Further numerics

B.1 Convergence rates of partition-based martingale approximations

The experiments in Figure 5 are designed to compare the performance of four different partition-based martingale approximation methods defined in Definition 1—minimax, median, Simons, and variance—based on their ability to approximate a target function using nested partitions. The methods are applied to different probability densities over a specified interval, and the results are visualized in a series of plots. We want to evaluate how differently martingale approximation methods reduce the variance (or the second moment) of a random variable Y that is distributed according to various density functions f . In Figure 5a, we consider the density f of Y given by

$$f(x) = \begin{cases} 1/51 & \text{if } x \in [0, 0.9]; \\ (1 + 10^4(x - 0.9))/51 & \text{if } x \in [0.9, 1]; \\ 0 & \text{otherwise,} \end{cases} \quad (52)$$

with four types of partition-based martingale approximations. The minimax method generally shows a steeper decline in MSE in the initial stages, suggesting that it is effective at quickly reducing the worst-case variance.

Furthermore, the second plot in Figure 5b shows the ratio of the remaining risk after each step normalized by the variance of the previous step, as an indicator of rates of improvement as the tree grows deeper. The behavior of these curves can be used to infer the asymptotic rates of convergence for each method, which align with the theoretical analysis of Section 2.3.

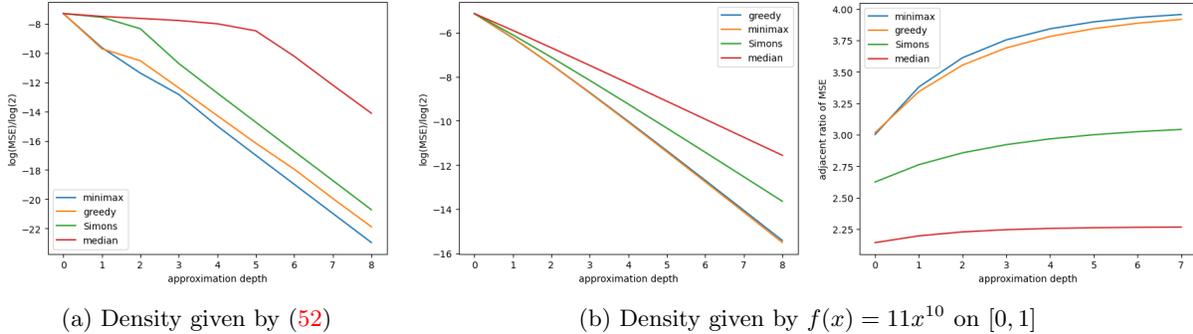


Figure 5: Plots of the log-MSE, $\log_2(\mathbb{E}[(Y - M_k)^2])$ versus the approximation depth k for four different methods, where the density of the law Y is given: (a) by (52) and (b) by $f(x) = 11x^{10}$ on $[0, 1]$ and zero otherwise. The right panel of (b) plots the ratios $\mathbb{E}[(Y - M_{k+1})^2]/\mathbb{E}[(Y - M_k)^2]$ for the four methods as a function of k .

B.2 Higher dimensional input domain

Consider a dimension $d \geq 1$ that is a multiple of 4. The Powell function on n ,

$$f(x) = \sum_{i=1}^{d/4} [(x_{4i-3} + 10x_{4i-2})^2 + 5(x_{4i-1} - x_{4i})^2 + (x_{4i-2} - 2x_{4i-1})^4 + 10(x_{4i-3} - x_{4i})^4],$$

with uniformly random samples on $[0, 1]^d$, serves as an exemplary testbed for decision tree methods in high-dimensional regression problems, offering a nuanced landscape of challenges that mirror real-world complexities. In the dense design scenario ($n > d > 100$), the function’s inherent non-linearity and intricate variable interactions push the boundaries of decision trees’ capabilities. The quartic terms in the function introduce sharp curvatures in the response surface, necessitating sophisticated splitting strategies to approximate these non-linear relationships accurately. Moreover, the coupling of variables in quartets creates a web of local interactions that tests the trees’ ability to discern and model multivariable dependencies efficiently.

Table 1 presents a comprehensive comparison of three decision tree methods—Scikit-learn’s standard implementation, VarianceSplit and MinimaxSplit—across various sample sizes n and dimensions d for the Powell function. This data offers valuable insights into the performance characteristics of these methods in both dense and sparse design scenarios. In the dense design regime ($n > d$), particularly evident in the cases where the sample size (100, 1000 or 10000) exceeds the dimensionality (4, 16 or 64), we observe relatively consistent performance across all three methods. This suggests that when data is abundant relative to the input space, the choice of splitting criterion (be it the standard impurity measure used by Scikit-learn, MinimaxSplit, or VarianceSplit) has less impact on the overall performance. However, as we transition into sparse designs ($d > n$), exemplified by cases where dimensions (64, 256 or 1024) exceed the sample size (10 or 100), more pronounced differences emerge. The MinimaxSplit method often demonstrates superior performance in these high-dimensional, data-scarce scenarios. For instance, with 10 samples and 256 dimensions, the MinimaxSplit approach achieves an MSE of 8.91×10^4 , outperforming both Scikit-learn (1.07×10^5) and the VarianceSplit method (9.58×10^4). This aligns with the theoretical strengths of the minimax criterion in handling sparse, high-dimensional data, where it can more robustly identify informative splits with limited samples. The VarianceSplit method generally performs comparably to Scikit-learn’s implementation, which is expected given that both likely use variants of variance reduction as their splitting criteria. However, in some sparse cases, the VarianceSplit method shows slight improvements, possibly due to specific implementation details or hyperparameter choices. As dimension increases for a fixed sample size, we observe a consistent increase in MSE across all methods, illustrating the challenges posed by the curse of dimensionality. This effect is particularly pronounced in the transition from 64 to 256 dimensions, where the MSE often increases by an order of magnitude or more. Interestingly, the performance gap between methods narrows as the sample size increases, even in high-dimensional settings. For instance, with 1000 samples and 256 dimensions, all three methods achieve comparable MSE values (around 6.4×10^4). This

Sample Size	Dimension	Sklearn	VarianceSplit	MinimaxSplit
10	4	4.48×10^2	2.01×10^2	2.57×10^2
10	16	1.60×10^4	1.52×10^4	1.52×10^4
10	64	3.70×10^4	4.24×10^4	3.79×10^4
10	256	1.35×10^5	1.24×10^5	1.14×10^5
10	1024	9.31×10^5	8.04×10^5	4.28×10^5
100	4	5.49×10^1	5.21×10^1	5.31×10^1
100	16	2.70×10^3	2.67×10^3	2.57×10^3
100	64	1.99×10^4	1.95×10^4	2.20×10^4
100	256	1.01×10^5	1.02×10^5	9.39×10^4
100	1024	4.65×10^5	4.29×10^5	4.02×10^5
1000	4	4.68×10^1	4.65×10^1	4.49×10^1
1000	16	1.85×10^3	1.85×10^3	2.03×10^3
1000	64	1.62×10^4	1.62×10^4	1.59×10^4
1000	256	6.46×10^4	6.47×10^4	6.30×10^4
1000	1024	2.94×10^5	2.96×10^5	2.88×10^5
10000	4	4.61×10^1	4.61×10^1	4.63×10^1
10000	16	1.87×10^3	1.87×10^3	1.97×10^3
10000	64	1.36×10^4	1.36×10^4	1.43×10^4
10000	256	6.22×10^4	6.22×10^4	6.11×10^4
10000	1024	2.80×10^5	2.80×10^5	2.78×10^5

Table 1: Comparison of Decision Tree Methods (max_depth=3) using the Powell function without noise. Scikit-learn’s DecisionTreeRegressor uses mean-squared Error (MSE) as its default splitting criterion, which is equivalent to minimizing the variance of the target variable within each split and serve as a baseline comparison here.

suggests that with sufficient data, the choice of splitting criterion becomes less critical and all methods can effectively capture the underlying structure of the Powell function.

B.3 A weighted aggregation approach

The variance approach aims to create homogeneous subsets by minimizing the average squared difference between the observed values and the mean value in each node. In contrast, the minimax criterion focuses on minimizing the maximum error within each split, potentially leading to more balanced trees. The variance criterion, similar to Scikit-learn’s MSE, seeks to minimize the overall variance in child nodes but can be implemented with different loss norms (L^1 or L^2). Cyclic minimax, a variant of minimax, alternates through features (i.e., the first and second coordinates) in a predetermined order for splitting, which can be beneficial in high-dimensional spaces when the depth is larger than the input dimension, or when feature importance is known a priori (e.g., the first coordinate is more heterogeneous).

These alternative criteria offer different trade-offs: minimax may be more robust to outliers and preserve edges better in image processing tasks, variance (especially with L^2 norm) often provides good average-case performance, and cyclic minimax can ensure a more diverse use of features. Unlike Scikit-learn’s implementation, which primarily optimizes for average error reduction, these custom criteria allow for more specialized tree structures tailored to specific problem characteristics or performance goals.

Our novel random forest (see Algorithm 1 of Appendix C) implementation diverges from conventional practices by introducing diverse splitting criteria and weighted aggregation. This approach is motivated by our previous discussion that the recognition that different splitting criteria can capture various aspects of data structure, while weighted aggregation can emphasize more accurate models in the final prediction. The implementation incorporates Variance, Minimax, and cyclic Minimax splitting methods for different weak learners within a single ensemble. The variance method optimizes for overall error reduction, minimax focuses on worst-case performance at each split, and cyclic minimax introduces a deterministic feature selection process. This diversity in the splitting criteria aims to mitigate feature importance bias present in standard

random forests and enhance ensemble diversity beyond what is typically achieved through bootstrapping and random feature subset selection (Friedman, 2001).

The weighted aggregation scheme, based on the reciprocal of each tree’s training RMSE, addresses the limitation that not all trees in the ensemble are equally reliable or informative. This adaptive ensemble aggregation allows the model to adjust to the varying quality of its constituent trees, potentially leading to more robust predictions. These modifications are primarily motivated by addressing limitations of standard random forests: feature importance bias, limited ensemble diversity, and non-adaptive aggregation. By introducing varied perspectives on feature importance, adding layers of diversification, and allowing for adaptive weighting, our approach aims to create a more flexible and robust ensemble.

B.4 Heterogeneous splitting

We also develop *heterogeneous splitting* strategies based on our findings. The heterogeneous splitting strategy in decision trees involves using different error metrics (splitting criteria) at various depths, unlike the homogeneous strategy that uses the same metric throughout. This approach allows flexibility and adaptability, essential for handling complex datasets where different regions of the input space may benefit from different strategies. Early splits might use variance reduction for effective data division, while deeper splits could switch to minimax to minimize worst-case errors, enhancing model robustness against outliers and reducing overfitting risks; alternatively we can also alternate between VarianceSplit and MinimaxSplit to control errors (Figure 6). Additionally, strategies like cyclic MinimaxSplit leverage periodic features, potentially improving performance in time series or cyclic data. This tailored approach adapts the splitting criterion based on depth, potentially leading to more accurate and effective splits as different tree levels might require different considerations. In fact, the minimax/variance alternating strategy with L^2 norm (RMSE=0.112334) further improves the lowest RMSE compared to a homogeneous splitting criteria. Consequently, a heterogeneous strategy can lead to more accurate and robust models by optimally addressing the unique characteristics of the data at each level of the tree, making it particularly beneficial in complex datasets with variable distributions.

C Algorithms

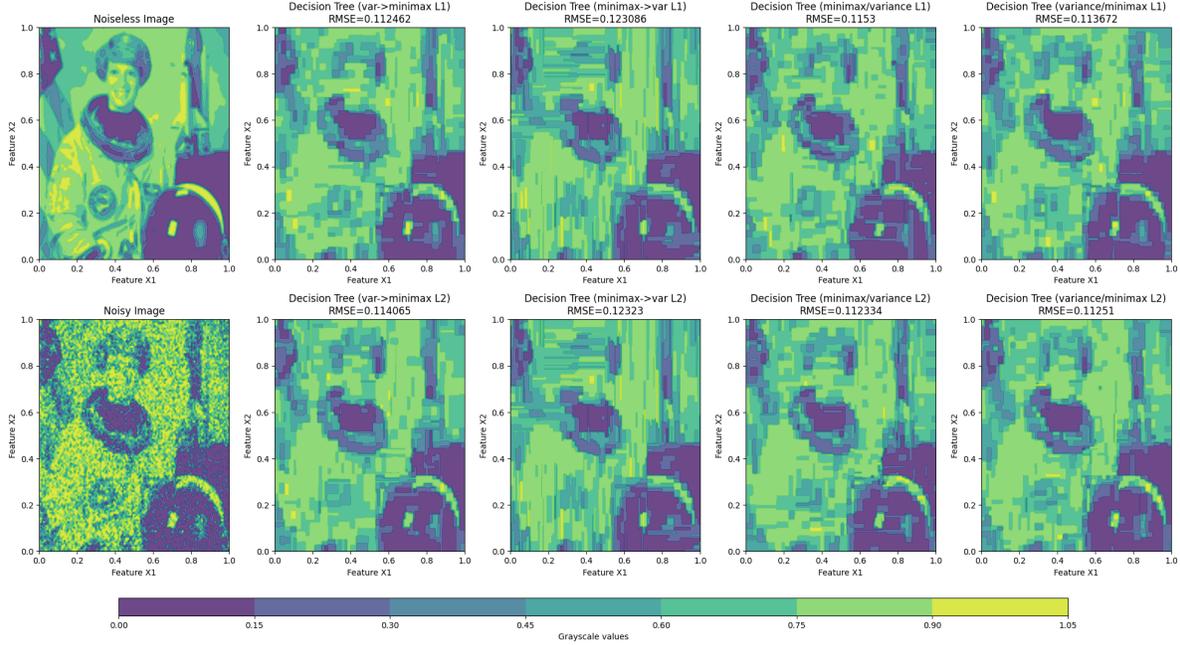


Figure 6: The noiseless image, noisy image and denoised images using different kinds of heterogeneous splitting strategies for a single tree with a max depth of 10. There are four different strategies: (1) var→minimax: we complete 5 layers with VarianceSplit, followed by 5 layers with MinimaxSplit; (2) minimax→var: we complete 5 layers with MinimaxSplit, followed by 5 layers with VarianceSplit; (3) minimax/variance: we apply VarianceSplit in the odd layers and MinimaxSplit in the even layers alternatively; (4) variance/minimax: we apply MinimaxSplit in the odd layers and VarianceSplit in the even layers alternatively.

Algorithm 1 Weighted Random Forest with Multiple Splitting Criteria

Require: Training data (\mathbf{X}, Y) , number of trees N , minimum samples to split, maximum depth, error method, loss norm, use weights flag

Ensure: Trained Random Forest model

- 1: Initialize N decision trees with specified parameters
 - 2: **for** $i = 1$ to N **do**
 - 3: Set tree's cyclic padding to i
 - 4: Create bootstrap sample (X_i, Y_i) from (X, Y)
 - 5: Train tree using VarianceSplit, MinimaxSplit or CyclicMinimaxSplit
 - 6: **if** use_weights is True **then**
 - 7: Predict on (X_i, Y_i) and calculate RMSE
 - 8: Set tree weight $w_i = 1/\text{RMSE}$ (when we set equal weights this becomes the regular random forest)
 - 9: **end if**
 - 10: **end for**
 - 11: **if** use_weights is True **then**
 - 12: Normalize weights: $w_i = w_i / \sum_{j=1}^N w_j$
 - 13: **end if**
 - 14: **function** PREDICT(X_{test})
 - 15: **for** $i = 1$ to N **do**
 - 16: Get predictions p_i from tree i on X_{test}
 - 17: **end for**
 - 18: **if** use_weights = True **then return** $\sum_{i=1}^N w_i p_i$
 - 19: **elsereturn** $\frac{1}{N} \sum_{i=1}^N p_i$
 - 20: **end if**
 - 21: **end function**
-

Algorithm 2 Single Tree fitting with different Criteria

```
1: function VARIANCESPLIT(node)
2:   for each feature  $j$  do
3:     Sort data points in node by feature  $j$ 
4:     for each potential split point do
5:       Calculate  $\text{MSE}_{\text{left}} \leftarrow \sum_{x \in \text{left}} (y_x - \bar{y}_{\text{left}})^2$ 
6:       Calculate  $\text{MSE}_{\text{right}} \leftarrow \sum_{x \in \text{right}} (y_x - \bar{y}_{\text{right}})^2$ 
7:       Calculate total error  $\leftarrow \text{MSE}_{\text{left}} + \text{MSE}_{\text{right}}$ 
8:     end for
9:     Find split point with minimum total error (with minimal number of splitting samples)
10:  end for
11:  return feature and split point with overall minimum error
12: end function
13: function MINIMAXSPLIT(node)
14:   for each feature  $j$  do
15:     Sort data points in node by feature  $j$ 
16:     for each potential split point do
17:       Calculate  $\text{MSE}_{\text{left}} \leftarrow \sum_{x \in \text{left}} (y_x - \bar{y}_{\text{left}})^2$ 
18:       Calculate  $\text{MSE}_{\text{right}} \leftarrow \sum_{x \in \text{right}} (y_x - \bar{y}_{\text{right}})^2$ 
19:       Calculate max error  $\leftarrow \max(\text{MSE}_{\text{left}}, \text{MSE}_{\text{right}})$ 
20:     end for
21:     Find split point with minimum max error (with minimal number of splitting samples)
22:  end for
23:  return feature and split point with overall minimum max error
24: end function
25: function CYCLICMINIMAXSPLIT(node, depth, cyclic_padding)
26:   feature_index  $\leftarrow (\text{cyclic\_padding} + \text{depth}) \bmod d$ 
27:   Sort data points in node by feature feature_index
28:   for each potential split point do
29:     Calculate  $\text{MSE}_{\text{left}} \leftarrow \sum_{x \in \text{left}} (y_x - \bar{y}_{\text{left}})^2$ 
30:     Calculate  $\text{MSE}_{\text{right}} \leftarrow \sum_{x \in \text{right}} (y_x - \bar{y}_{\text{right}})^2$ 
31:     Calculate max error  $\leftarrow \max(\text{MSE}_{\text{left}}, \text{MSE}_{\text{right}})$ 
32:   end for
33:   Find split point with minimum max error (with minimal number of splitting samples)
34:   return (feature_index, best split point)
35: end function
```
