Journal of Machine Learning Research 25 (2024) 1-32 Submitted 5/22; Revised 4/23; Published 3/24

Minimax Rates for High-Dimensional Random Tessellation Forests

Eliza O'Reilly EOREILL2@JH.EDU
Applied Mathematics and Statistics Department
Johns Hopkins University
Baltimore, MD 21218, USA

Ngoc Mai Tran NTRAN@MATH.UTEXAS.EDU
Department of Mathematics
University of Texas at Austin
Austin, TX 78712, USA

Editor: Daniel Roy

Abstract

Random forests are a popular class of algorithms used for regression and classification. The algorithm introduced by Breiman in 2001 and many of its variants are ensembles of randomized decision trees built from *axis-aligned* partitions of the feature space. One such variant, called Mondrian forests, was proposed to handle the online setting and is the first class of random forests for which minimax optimal rates were obtained in arbitrary dimension. However, the restriction to axis-aligned splits fails to capture dependencies between features, and random forests that use oblique splits have shown improved empirical performance for many tasks. This work shows that a large class of random forests with general split directions also achieve minimax optimal rates in arbitrary dimension. This class includes STIT forests, a generalization of Mondrian forests to arbitrary split directions, and random forests derived from Poisson hyperplane tessellations. These are the first results showing that random forest variants with oblique splits can obtain minimax optimality in arbitrary dimension. Our proof technique relies on the novel application of the theory of stationary random tessellations in stochastic geometry to statistical learning theory.

Keywords: random forest regression, Mondrian process, STIT tessellation, Poisson hyperplane tessellation, minimax risk bound

1. Introduction

Random forests are ensembles of randomized decision trees popularized by Breiman (2001) and are broadly applicable in classification and regression tasks (Fernández-Delgado et al., 2014; Chen and Ishwaran, 2012). Despite their empirical success and widespread use, statistical learning theorems have been notoriously difficult to obtain in dimensions $D \geq 2$ (Biau and Scornet, 2016). There has been significant progress towards understanding the asymptotic behavior of the original algorithm proposed by Breiman (2001) (Scornet et al., 2015; Wager and Athey, 2015, 2018; Mentch and Hooker, 2016; Chien-Ming Chi and Lv, 2022; Klusowski and Tian, 2022), but much of the theory is still limited by strong assumptions and suboptimal rates. Some of the main difficulties in obtaining theoretical guarantees for this algorithm and its variants result from the complex dependence between the partition-

©2024 Eliza O'Reilly and Ngoc Mai Tran.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/22-0673.html.

ing process generating the tree and the underlying data set. In response, another line of research considers simplified and stylized versions of random forests. In particular, purely random forests (Biau et al., 2008; Arlot and Genuer, 2014) are models built from randomized hierarchical partitions of the input space that are independent of the data and are thus more amenable to theoretical analysis. Recently, Mourtada et al. (2020) obtained the first minimax optimal rates in arbitrary dimension for a particular class of purely random forests. Specifically, they proved that *Mondrian forests* (Lakshminarayanan et al., 2014, 2016) attain optimal rates for a properly tuned complexity parameter growing with the amount of data.

Mondrian forests are purely random forests based on the Mondrian process, a recursive random partition of $\mathbb{R}^D$ by axis-aligned cuts introduced by Roy and Teh (2008). This stochastic process enjoys an efficient Markov construction and the following self-consistency property: a sample of a Mondrian process in some domain $W_1 \subset \mathbb{R}^D$ has the same distribution as sampling a Mondrian process on a larger domain W_2 such that $W_1 \subset W_2$ and intersecting with W_1. These properties ensure the amenability of Mondrian forests to the online setting (Lakshminarayanan et al., 2014, 2016; Wang et al., 2018), where data arrives in a streaming manner, and the estimator is updated over time. This contrasts with Breiman's random forest algorithm and many of its variants which are restricted to the batch setting, where the entire data set is used at once to build the model. In addition to this practical advantage, the results of Mourtada et al. (2020) mentioned above highlight the theoretical advantages of Mondrian forests resulting from its construction.

One key limitation of the Mondrian forest and Breiman's original random forest algorithm comes from the constraint that only one feature of the input is used each time the data within a tree node is divided. While computationally efficient, these axis-aligned splits cannot capture dependencies between features and may produce complex step-wise decision boundaries that lead to high variance and overfitting. In practice, this places a large burden on the feature selection and representation process. To address these concerns, Breiman (2001) proposed a variant called Forest-RC that increases the expressiveness of the model by allowing splits using linear combinations of features, and it was shown to achieve improved empirical performance over the axis-aligned version. Many other models of random forests using oblique splits have subsequently been proposed (Menze et al., 2011; Blaser and Fryzlewicz, 2016; Fan et al., 2018; Ge et al., 2019; Rainforth and Wood, 2015). To mitigate the increased computational cost of using linear combinations of features, Tomita et al. (2020) studied random forests with oblique splits from sparse projections using only a small subset of features. However, oblique splits increase the already difficult task of proving theoretical guarantees for random forest algorithms, and theory justifying the empirical performance of these variants is extremely limited. Existing guarantees restricted to the setting of axis-aligned splits are not easily generalized to oblique splits, illuminating a need for a more flexible theoretical framework.

To the best of our knowledge, this paper gives the first results on minimax optimality for a large class of purely random forests that are defined for all dimensions and allow for *general* splits using linear combinations of features. In particular, we show that STIT forests, a significant generalization of Mondrian forests, also attain the minimax rates discovered in Mourtada et al. (2020). STIT forests are derived from the stable under iteration (STIT) processes introduced by Nagel and Weiss (2003, 2005), and these stochastic pro-

cesses all enjoy the self-consistency and online construction that underpin the popularity of the Mondrian forest in practice. The family of STIT processes is indexed by probability distributions on the unit sphere describing the distribution of directions of the hyperplane cuts in the random partition, and the Mondrian process corresponds to a STIT process where this directional distribution is the discrete uniform measure on the coordinate vectors. Subsequent generalizations of the Mondrian process to oblique cuts (Fan et al., 2018; Ge et al., 2019) are also special cases of STIT processes. The freedom in the choice of the directional distribution for the splits brings greater flexibility in building machine learning models. For example, while the Mondrian process can only be used to approximate the Laplace kernel (Lakshminarayanan et al., 2016), STIT processes produce random features that approximate a much broader class of kernels (O'Reilly and Tran, 2022). Improved empirical performance of STIT forests built from STIT processes with a uniform directional distribution over Mondrian forests was also shown by Ge et al. (2019) through a classification task and simulation study. Additionally, O'Reilly and Tran (2022) observed that a STIT process in $\mathbb{R}^D$ with discrete directional distribution having $N \geq D$ support vectors can be simulated by lifting to a subspace of $\mathbb{R}^N$ and running a Mondrian process. This observation can mitigate computational costs of the oblique splits and gives an interpretation of a STIT forest as implicitly generating a Mondrian forest in a higher dimensional feature space.

A significant contribution of our work lies in the proof technique, as our approach relies on theorems in stochastic geometry that have not previously been used in statistical learning theory. In extending the theory for the Mondrian to STIT forests, a fundamental difficulty is the geometry of the cells of the partitions. The Mondrian process generates axis-aligned rectangular cells, and the distribution of the cell that a given input is contained in can be characterized precisely from the construction. In contrast, STIT processes divide the input space into more general and complex convex polytopes. The theory of random tessellations in stochastic geometry provides a flexible and robust theoretical framework that enables us to handle these more general cell geometries. In particular, we crucially exploit the self-consistency property and the stationarity of the corresponding STIT tessellation on $\mathbb{R}^D$ to obtain risk bounds for STIT forest estimators that depend on the distribution of a single random polytope, called the *typical cell* of the random tessellation.

Additionally, our proof technique allows us to incorporate an assumption of intrinsic low-dimensionality on the input data, improving convergence rates in high-dimensional feature space. A well-known challenge in developing statistical learning guarantees for nonparametric regression is the curse of dimensionality, where, for instance, one needs $O(\varepsilon^{-D})$ number of samples to estimate general Lipschitz functions on $\mathbb{R}^D$ with ε accuracy. Indeed, the minimax optimal rates for Mondrian forests in Mourtada et al. (2020) depend on the ambient dimension of the input data and become very slow in the presence of high-dimensional feature space, even though empirically random forests perform well in such regimes (Chen and Ishwaran, 2012; Fernández-Delgado et al., 2014; Tomita et al., 2020). One approach to justifying such performance is to make additional structural assumptions on the input data source to attain improved convergence rates, as in the results of Kpotufe and Dasgupta (2012). For STIT forests, the theory of stationary hyperplane processes in stochastic geometry provides insight yielding optimal rates that also adapt to a notion of intrinsic dimensionality of the input.

Specifically, our first main result (see Theorem 6) gives an upper bound on the quadratic risk of a STIT forest regression estimator of a β-Hölder continuous function for $\beta \in (0, 1]$. The theorem implies that *any* STIT forest with optimally tuned complexity parameter achieves the minimax rate for this function class, and the choice of the directional distribution of the splits appears in the constant terms of the upper bound. Our proof method gives geometric interpretations to these constants in terms of moments of the diameter of the cell of the associated STIT tessellation containing the origin and the expected mixed volumes between the support of the input and a random polytope called the typical cell. This precise geometry enables us to obtain rates in terms of the intrinsic dimension of the input defined by the dimension of the subspace of $\mathbb{R}^D$ on which the input data is supported. In the particular case of the Mondrian forest, the typical cell is the Minkowski sum of i.i.d. centered line segments parallel to the axes with exponential length. Taking the support of the input to be $[0, 1]^D$ recovers Theorem 2 in Mourtada et al. (2020), see Example 5 for details. Our second main result (see Theorem 8) significantly generalizes Theorem 3 in Mourtada et al. (2020), where additional smoothness assumptions are made on f. We show that any STIT forest estimator achieves the minimax rate for the class of $(1 + \beta)$-Hölder functions for $\beta \in (0, 1]$ for both a large enough number of trees in the forest and an optimally tuned complexity parameter. As was the case for Mondrian forests, an improved rate for STIT forests over STIT trees is due to large enough forests having a smaller bias than single trees for smooth regression functions.

Finally, our proof technique also takes us *beyond* the class of STIT forests. Any random partition of the input space can be used to define a random tree estimator and subsequently a random forest estimator. Since the upper bounds we obtain on the convergence rates are explicitly derived in terms of geometric properties of the typical cell of the random partition, it can readily be applied to any random forest obtained from a stationary random tessellation of $\mathbb{R}^D$. Our last main result (see Theorem 11) demonstrates this principle. It states that a random forest derived from a Poisson hyperplane process achieves identical convergence rates as a STIT forest with the same directional distribution and complexity parameter and thus is also minimax optimal.

1.1 Organization

Section 2 collects background on STIT processes, Poisson hyperplane processes, and essential results in stochastic geometry needed for our proofs. Section 4 states and proves our three main results, Theorems 6, 8, and 11. Section 6 concludes with a discussion and open problems.

2. Preliminaries

We recall here the key concepts from stochastic geometry needed for our paper, including *random tessellations, stationarity*, the *zero cell*, and the *typical cell*. We recommend the book by Schneider and Weil (2008, Chapter 10) for additional background.

A tessellation is a locally finite random partition of $\mathbb{R}^D$ into compact and convex polytopes. It can be viewed as the collection of polytopes, or cells, of the tessellation or as the union of their boundaries. In this paper, we view tessellations as the collection of cells, but we will also discuss the union of cell boundaries to establish relevant definitions. Formally,

we define a *random tessellation* as a point process of cells $\mathcal{P} = \{C_i\}_{i\in\mathbb{Z}}$ taking values in the space $\mathcal{K}$ of non-empty compact and convex polytopes that satisfy the following:

- for all compact $K \subset \mathbb{R}^D$, a finite number of C_i's have non-empty intersection with K;
- for all $i \neq j$, $\text{int}(C_i) \cap \text{int}(C_j) = \emptyset$;
- $\bigcup_{i\in\mathbb{Z}} C_i = \mathbb{R}^D$.

A random tessellation $\mathcal{P} = \{C_i\}_{i\in\mathbb{Z}}$ is *stationary* if its distribution is invariant under translations. That is, for all $x \in \mathbb{R}^D$, $\{C_i + x\}_{i\in\mathbb{Z}} \overset{d}{=} \{C_i\}_{i\in\mathbb{Z}}$, where '$\overset{d}{=}$' denotes equality in distribution. A consequence of the definition of a random tessellation is that for all i, $\text{vol}_D(\partial C_i) = 0$. This fact, along with stationarity, implies that every $x \in \mathbb{R}^D$ almost surely belongs to a unique cell of the tessellation, which will be denoted Z_x. The zero cell of $\mathcal{P}$, denoted by Z_0, is defined as the unique cell of the tessellation containing the origin. Stationary also implies that $Z_0 \overset{d}{=} Z_x - x$.

An important random object related to a stationary random tessellation $\mathcal{P}$ is the *typical cell*. To define this, consider first a center function $c : \mathcal{K} \to \mathbb{R}^D$ such that $c(K+x) = c(K)+x$ for all $x \in \mathbb{R}^D$. Examples include the centroid or the center of the smallest ball containing K. We can then decompose $\mathcal{P}$ into a stationary marked point process $\{(c(C_j), C_j - c(C_j))\}_{j\in\mathbb{Z}}$ consisting of a ground point process in $\mathbb{R}^D$ of cell centers and elements from $\mathcal{K}_0 := \{K \in \mathcal{K} : c(K) = 0\}$ attached to each center. Following Schneider and Weil (2008, Section 4.1), there exists a random polytope Z in $\mathcal{K}_0$ such that for any non-negative measurable function f on $\mathcal{K}$,

$$\mathbb{E}\left[\sum_{C\in\mathcal{P}} f(C)\right] = \frac{1}{\mathbb{E}[\text{vol}_D(Z)]}\mathbb{E}\left[\int_{\mathbb{R}^D} f(Z+y)\mathrm{d}y\right]. \tag{1}$$

The random polytope Z is called the *typical cell* of $\mathcal{P}$. Its distribution can be understood as the limiting distribution of a cell chosen uniformly at random from a large ball, centered at the origin using the center function c, as the radius of the ball grows to infinity. The relation (1) also implies that the distribution of the centered zero cell $Z_0 - c(Z_0)$ has the same distribution as the volume-weighted typical cell, i.e.

$$\mathbb{E}[f(Z_0 - c(Z_0))] = \frac{1}{\mathbb{E}[\text{vol}_D(Z)]}\mathbb{E}\left[\text{vol}_D(Z)f(Z)\right].$$

For a random tessellation $\mathcal{P}$ in $\mathbb{R}^D$, we will denote by $\mathcal{Y}$ the union of cell boundaries, which forms a $d-1$-surface process in $\mathbb{R}^D$ (Schneider and Weil, 2008, Section 4.5). For the stationary surface process $\mathcal{Y}$ corresponding to a stationary random tessellation, we can define a directional distribution ϕ on $\mathbb{S}^{D-1}$ characterizing the 'rose of directions' for the $D-1$-dimensional facets generating the cell boundaries. We will say $\mathcal{P}$ has directional distribution ϕ if $\mathcal{Y}$ has directional distribution ϕ.

2.1 STIT Tessellations

For two random tessellations $\mathcal{P}_1$ and $\mathcal{P}_2$, denote the union of their cell boundaries by $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively. Associate to each cell c in $\mathcal{P}_1$ an independent copy $\mathcal{Y}_2(c)$ of $\mathcal{Y}_2$ and

assume the family $\{\mathcal{Y}_2(c) : c \in \mathcal{P}_1\}$ is independent of $\mathcal{P}_1$. Then, the *iteration* of $\mathcal{Y}_1$ and $\mathcal{Y}_2$ is defined as

$$\mathcal{Y}_1 \boxplus \mathcal{Y}_2 := \mathcal{Y}_1 \cup \bigcup_{c \in \mathcal{P}_1} (\mathcal{Y}_2(c) \cap c)\,.$$

That is, each cell c of the frame tessellation $\mathcal{P}_1$ is subdivided by the cells of $\mathcal{Y}_2(c) \cap c$. A random tessellation $\mathcal{P}$ is called *stable under iteration*, or STIT, if for the union of cell boundaries $\mathcal{Y}$, for all $n \in \mathbb{N}$,

$$\mathcal{Y} \stackrel{d}{=} n(\underbrace{\mathcal{Y} \boxplus \cdots \boxplus \mathcal{Y}}_{n \text{ times}}), \tag{2}$$

where $n\mathcal{Y} := \{nx : x \in \mathcal{Y}\}$ is the dilation of $\mathcal{Y}$ by the factor n.

A *STIT process* is a stochastic process $\{\mathcal{Y}(\lambda) : \lambda > 0\}$ of random tessellation cell boundaries in $\mathbb{R}^D$ with the following properties:

(i) Stationarity: $\mathcal{Y}(\lambda) + x \stackrel{d}{=} \mathcal{Y}(\lambda)$ for all $x \in \mathbb{R}^D$;

(ii) Markov Property: $\mathcal{Y}(\lambda_1 + \lambda_2) \stackrel{d}{=} \mathcal{Y}(\lambda_1) \boxplus \mathcal{Y}(\lambda_2)$ for all $\lambda_1, \lambda_2 > 0$;

(iii) STIT: for all $\lambda > 0$ and $n \in \mathbb{N}$, (2) holds for $\mathcal{Y}(\lambda)$.

We will call the parameter $\lambda > 0$ the *lifetime* of the process. A consequence of property (iii) is that STIT processes have the following scaling property: for all $\lambda > 0$, $\mathcal{Y}(1) \stackrel{d}{=} \lambda\mathcal{Y}(\lambda)$ (Nagel and Weiss, 2005, Lemma 5). Intuitively, this property says that one can swap time for space. If we fix a compact observation window $W \subset \mathbb{R}^D$, then $\{\mathcal{Y}(\lambda) \cap W : \lambda > 0\}$ is a stochastic process of random tessellation cell boundaries in W, which we think of as a visualization of $\mathcal{Y}(\lambda)$ through the window W. Now, one can fix λ and 'zoom out' on $\mathcal{Y}(\lambda)$ by mapping $\mathcal{Y}(\lambda) \cap W \mapsto \frac{1}{2}\mathcal{Y}(\lambda) \cap W$, or one can run the STIT process for twice as long by mapping $\mathcal{Y}(\lambda) \cap W \mapsto \mathcal{Y}(2\lambda) \cap W$. The scaling property says these two operations give the same random tessellation in W in distribution.

For a lifetime $\lambda > 0$, let $\mathcal{P}(\lambda)$ denote the STIT tessellation of $\mathbb{R}^D$ with cell boundaries given by a STIT process $\mathcal{Y}(\lambda)$. Denote the zero cell of $\mathcal{P}(\lambda)$ by Z_0^λ and the typical cell by Z_λ. The scaling property implies the following important facts that we will use in the remainder of the paper: for all $\lambda > 0$,

$$Z_0 \stackrel{d}{=} \lambda Z_0^\lambda \quad \text{and} \quad Z \stackrel{d}{=} \lambda Z_\lambda, \tag{3}$$

where $Z_0 := Z_0^1$ and $Z := Z_1$ are the zero cell and typical cell of $\mathcal{P}(1)$.

While seemingly abstract, it was proved by Nagel and Weiss (2005) that the local STIT process $\{\mathcal{Y}(\lambda) \cap W : \lambda > 0\}$ restricted to a fixed compact and convex window $W \subset \mathbb{R}^D$ can be simulated through a Markov process that generates a hierarchical partition of W over time. A special case of this construction was rediscovered by Roy and Teh (2008), which led to the Mondrian process.

Formally, let ϕ be an even probability measure on the unit sphere $\mathbb{S}^{D-1}$ with support containing d linearly independent directions. Then define Λ to be the stationary and locally finite measure on the space of hyperplanes in $\mathbb{R}^D$, denoted $\mathcal{H}^D$, such that

$$\Lambda(A) := \int_{\mathbb{R}} \int_{\mathbb{S}^{D-1}} 1_{\{H(u,t) \in A\}} \phi(\mathrm{d}u)\mathrm{d}t, \quad A \in \mathcal{B}(\mathcal{H}^D),$$

where $H(u,t) := \{x \in \mathbb{R}^D : \langle x, u\rangle = t\}$. The space $\mathcal{H}^D$ of hyperplanes is equipped with the hit-miss topology, which contains compact subsets of the following form: for compact $W \subset \mathbb{R}^D$,

$$[W] := \{H \in \mathcal{H}^D : H \cap W \neq \emptyset\}.$$

Now, fix a lifetime $\lambda > 0$ and consider the following procedure to construct a random partition $\mathcal{Y}(\lambda, W, \phi)$ of W.

1. Draw $\delta \sim \mathrm{Exp}(\Lambda([W]))$, where

$$\Lambda([W]) = \int_{\mathbb{R}} \int_{\mathbb{S}^{D-1}} 1_{\{H(u,t)\cap W \neq \emptyset\}} \mathrm{d}\phi(u)\mathrm{d}t = \int_{\mathbb{S}^{D-1}} \left(h(W,u) + h(W,-u)\right) \mathrm{d}\phi(u),$$

   and $h(W,u) := \sup_{x\in W}\langle u, x\rangle$ is the support function of W.

2. If $\delta > \lambda$, stop. Else, at time δ, generate a random hyperplane $H(U,T)$ where the direction U is drawn from the distribution

$$\mathrm{d}\Phi(u) := \frac{h(W,u) + h(W,-u)}{\Lambda([W])} \mathrm{d}\phi(u), \quad u \in \mathbb{S}^{D-1},$$

   and conditioned on U, T is drawn uniformly on the interval from $-h(W,-U)$ to $h(W,U)$. Split the window W into two cells W_1 and W_2 with $H(U,T)$.

3. Repeat steps (1) and (2) in each sub-window W_1 and W_2 independently with new lifetime parameter $\lambda - \delta$ until the lifetime expires.

In Figure 1, we show samples of the partition generated from a STIT process in the centered unit square $[-0.5, 0.5]^2$ using the above procedure over increasing lifetime λ. In this example, the directional distribution is uniform over the three directions $e_1 := (1,0)$, $e_2 := (0,1)$, $u := (1/\sqrt{2}, 1/\sqrt{2})$, and their reflections over the origin.

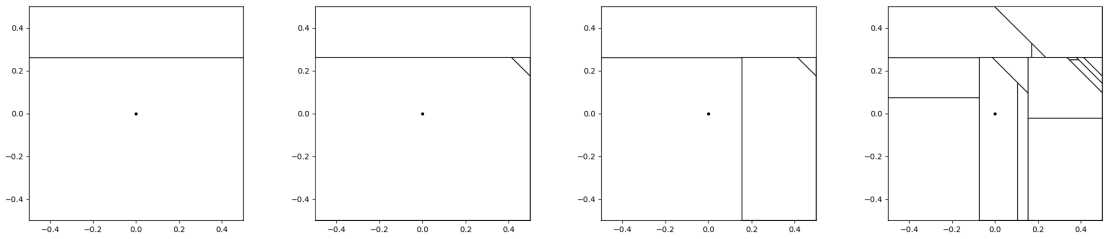


Figure 1: An example STIT process with three directions. When the process within a cell W begins, an independent exponential clock with mean $\Lambda([W])$ is started. When the clock rings, the cell is cut by a random hyperplane conditioned to hit this cell as in step 2 above. In this simulation, we start with the unit square $W = [-0.5, 0.5]^2$ and run the process until the lifetime $\lambda = 9$. The first three figures show the resulting partition after the first three cuts are made, and the last figure shows the final STIT tessellation at time $\lambda = 9$.

Theorem 1 in Nagel and Weiss (2005) shows the existence of a STIT tessellation process $\mathcal{Y}(\lambda)$ on $\mathbb{R}^D$ such that $\mathcal{Y}(\lambda) \cap W \overset{d}{=} \mathcal{Y}(\lambda, W, \phi)$. Conversely, for any stationary random tessellation in $\mathbb{R}^D$ with cell boundaries $\mathcal{Y}$ satisfying (2) with directional distribution ϕ, Corollary 2 in Nagel and Weiss (2005) shows that there exists $\lambda > 0$ such that $\mathcal{Y} \cap W \overset{d}{=} \mathcal{Y}(\lambda, W, \phi)$ for all compact $W \subset \mathbb{R}^D$. Together, these results imply that the class of STIT tessellations is the most general class of stationary random tessellations with the hierarchical construction that underpins the computational benefits of the Mondrian process (Roy and Teh, 2008; Lakshminarayanan et al., 2014, 2016).

Example 1 (The Mondrian process as a special case of STIT processes) *If ϕ is the uniform distribution over the positive and negative standard basis vectors, i.e.*

$$\phi = \frac{1}{2D} \sum_{i=1}^{D} (\delta_{e_i} + \delta_{-e_i}), \tag{4}$$

then the resulting STIT process $\mathcal{Y}(\lambda, W, \phi)$ is the Mondrian process (Roy and Teh, 2008). See Figure 2(b) for a simulation.

Example 2 (Isotropic STIT process) *If ϕ is the uniform distribution over $\mathbb{S}^{D-1}$, then the distribution of the corresponding STIT tessellation is invariant with respect to rotations about the origin. This model is called the isotropic STIT process. See Figure 2(a) for a simulation.*

2.2 Poisson Hyperplane Tessellations

We now define another class of stationary random tessellations in $\mathbb{R}^D$ and describe its relationship to the class of STIT tessellations. A *stationary Poisson hyperplane process* X is a stationary Poisson point process on the space of affine hyperplanes $\mathcal{H}^D$ in $\mathbb{R}^D$ with first moment measure

$$\Theta(\cdot) := \mathbb{E}[X(\cdot)] = \lambda \int_{\mathbb{R}} \int_{\mathbb{S}^{D-1}} 1_{\{H(u,t) \in \cdot\}} \mathrm{d}\phi(u) \mathrm{d}t,$$

for some constant $\lambda > 0$ called the *intensity*, and an even probability measure ϕ on $\mathbb{S}^{D-1}$ called the spherical directional distribution (Schneider and Weil, 2008, Chapter 4.4). The following procedure generates a sample from X on a compact window W:

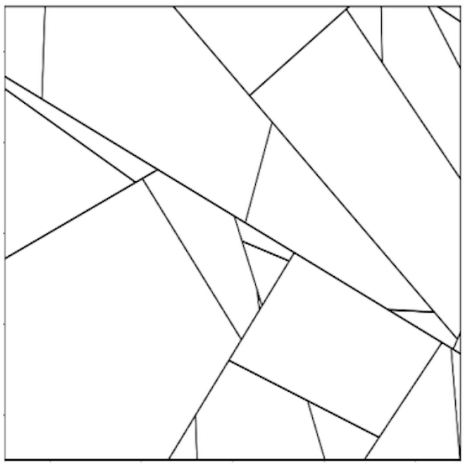
1. Sample $N \sim \text{Poisson}(\Theta([W]))$, where

$$\Theta([W]) = \lambda \int_{\mathbb{R}} \int_{\mathbb{S}^{D-1}} 1_{\{H(u,t) \cap W \neq \emptyset\}} \mathrm{d}\phi(u) \mathrm{d}t = \lambda \int_{\mathbb{S}^{D-1}} \left(h(W,u) + h(W,-u)\right) \mathrm{d}\phi(u).$$

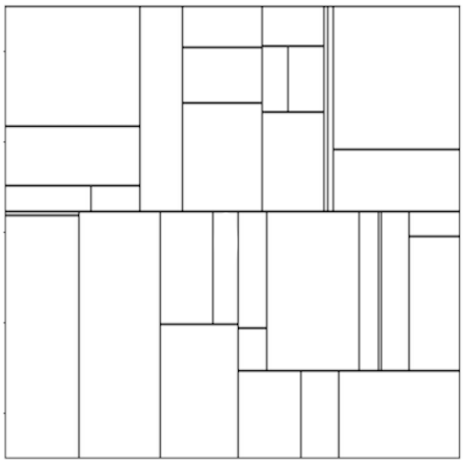
2. Conditioned on $N = n$, generate n i.i.d. random hyperplanes $\{H(U_i, T_i)\}_{i=1}^n$, where for each i, U_i has probability distribution

$$\mathrm{d}\Phi(u) := \frac{\lambda(h(W,u) + h(W,-u))}{\Theta([W])} \mathrm{d}\phi(u), \quad u \in \mathbb{S}^{D-1},$$

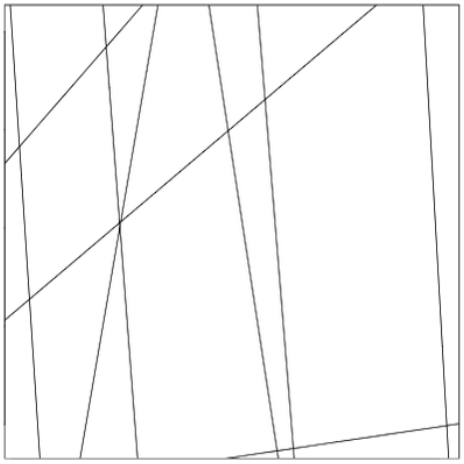
and conditioned on U_i, T_i is uniform in the interval from $-h(W, -U_i)$ to $h(W, U_i)$.



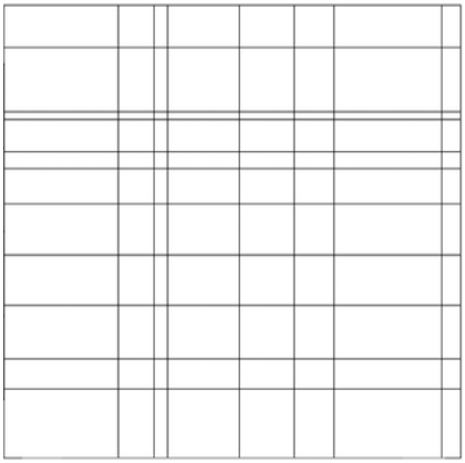
(a) Isotropic STIT process



(b) Axis-aligned STIT (aka Mondrian) process



(c) Isotropic Poisson hyperplane process



(d) Axis-aligned (aka Manhattan) Poisson hyperplane process

Figure 2: A simulation of STIT processes (top) vs. Poisson hyperplane processes (bottom) up to time $\lambda = 10$, with ϕ being the continuous uniform measure on the unit circle in (a) and (c), and the discrete uniform measure on the standard coordinate vectors in (b) and (d). Though the pairs (a,c) and (b,d) are globally different tessellations, it was observed in Nagel and Weiss (2005) (see also Schreiber and Thäle, 2013, Corollary 1) that the typical cell Z of these two tessellations have identical distribution. This fact is key to the proof of Theorem 11, which says that random forests based on (a) and (c) or random forests based on (b) and (d) achieve the same minimax rates.

Poisson hyperplane processes induce random tessellations on $\mathbb{R}^D$ called Poisson hyperplane tessellations that are globally different than STIT tessellations (cf. Figure 2). In particular, a Poisson hyperplane tessellation is face-to-face, meaning that the intersection of two cells is either empty, or is a face of both cells. This is not the case for a STIT tessellation. For example, a vertex of a cell in a STIT tessellation can be an interior point of the facet of a neighbor cell. However, the typical cell of a STIT tessellation with lifetime λ and directional distribution ϕ has the same distribution as the typical cell of a stationary Poisson hyperplane tessellation with intensity λ and the same directional distribution (Schreiber and Thäle, 2013, Corollary 1). By Theorem 10.4.1 in Schneider and Weil (2008), the typical cell determines the distribution of the zero cell, and so the zero cell of a STIT tessellation and a stationary Poisson hyperplane tessellation with corresponding parameters are also equal in distribution.

3. Parameters of STIT and Poisson Hyperplane Tessellation Cells

In this section, we generalize the results in Section 4 of Mourtada et al. (2020) on the diameter of the zero cell and the number of cells hitting a compact and convex domain. These observations show that STIT processes and stationary Poisson hyperplane processes produce partitions on a domain of unit volume that contain $O(\lambda^d)$ cells of diameter $O(1/\lambda)$, which is on the order of the $1/\lambda$-covering number for such a domain. Both bounds depend on an important parameter called the *associated zonoid* (Schneider and Weil, 2008, p.156). If the random tessellation has directional distribution ϕ and lifetime/intensity λ, this is defined as the convex body Π_λ in $\mathbb{R}^D$ with support function

$$h(\Pi_\lambda, v) = \frac{\lambda}{2}\Lambda([[0, v]]) = \frac{\lambda}{2}\int_{\mathbb{S}^{D-1}} |\langle u, v\rangle| \mathrm{d}\phi(u), \quad v \in \mathbb{S}^{D-1}. \tag{5}$$

We will denote by Π the associated zonoid of the process for lifetime/intensity $\lambda = 1$ and refer to this parameter that depends only on the directional distribution as the *normalized associated zonoid* of the random tessellation. In particular, note that $\Pi_\lambda = \lambda\Pi$. We also recall (Schneider and Weil, 2008, equations 10.4 and 10.44) that

$$\mathbb{E}[\mathrm{vol}_D(Z)] = \frac{1}{\mathrm{vol}_D(\Pi)}. \tag{6}$$

For the Mondrian (see Example 1) or axis-aligned Poisson hyperplane process, we can combine (4) and (5) to see that the support function of Π is given by

$$h(\Pi, v) = \frac{1}{D}\sum_{i=1}^{D} |\langle e_i, v\rangle| = \frac{1}{D}\|v\|_1, \quad u \in \mathbb{R}^D.$$

Thus the associated zonoid is the ℓ^∞ ball $\Pi = \{x \in \mathbb{R}^D : \|x\|_\infty \leq \frac{1}{D}\}$.

For the isotropic STIT (see Example 2) or isotropic Poisson hyperplane process, we let ϕ in (5) be the uniform distribution σ on the unit sphere to obtain

$$h(\Pi, v) = \frac{1}{2}\int_{\mathbb{S}^{D-1}} |\langle u, v\rangle| \mathrm{d}\sigma(u) = \frac{\|v\|_2}{2}\int_{\mathbb{S}^{D-1}} |u_n| \mathrm{d}\sigma(u) = c_D\|v\|_2,$$

where $c_D := \frac{\Gamma\left(\frac{D}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{D+1}{2}\right)}$. Thus the associated zonoid Π is an ℓ^2 ball centered at the origin with radius c_D.

The associated zonoid can also be used to understand the distribution of the intersection of a STIT or Poisson hyperplane tessellation in $\mathbb{R}^D$ with a linear subspace $\mathcal{S}$. We will repeatedly use the following important fact in the remainder of the paper.

Fact 1 *(Schneider and Weil, 2008) Let $\mathcal{P}$ be a STIT (or stationary Poisson hyperplane) tessellation in $\mathbb{R}^D$ with associated zonoid Π and let $\mathcal{S}$ be a linear subspace of $\mathbb{R}^D$. The intersection $\mathcal{P} \cap \mathcal{S}$ is a STIT (or stationary Poisson hyperplane) tessellation in $\mathcal{S}$ with associated zonoid given by the orthogonal projection $P_S\Pi$ of Π onto the subspace $\mathcal{S}$.*

In the following, let $\mathcal{P}(\lambda)$ denote a STIT tessellation in $\mathbb{R}^D$ with lifetime $\lambda \in (0, \infty)$ and normalized associated zonoid Π.

3.1 Diameter of Zero Cell

The precise distribution of the diameter of the zero cell of $\mathcal{P}(\lambda)$ with a general directional distribution remains an open question in stochastic geometry. However, we will provide an upper bound on the moments that is sufficient for proving the subsequent results.

Lemma 2 *Let Z_0^λ denote the zero cell of $\mathcal{P}(\lambda)$ in $\mathbb{R}^D$ and let $\mathcal{S}$ be a linear subspace of $\mathbb{R}^D$. Then, for all $k > 0$, there exists a constant $c := c(k, \Pi)$ depending only on k and Π such that*

$$\mathbb{E}[\mathrm{diam}(Z_0^\lambda \cap \mathcal{S})^k] \leq c\lambda^{-k} h_{\min}(P_{\mathcal{S}}\Pi)^{-k},$$

where $h_{\min}(P_{\mathcal{S}}\Pi) := \min_{u \in \mathbb{S}^{D-1} \cap \mathcal{S}} h(P_{\mathcal{S}}\Pi, u)$.

Proof We first note that by Fact 1 and (3), the random polytope $Z_0^\lambda \cap \mathcal{S}$ has the same distribution as $\lambda^{-1} Z_{0,\mathcal{S}}$, where $Z_{0,\mathcal{S}}$ is the zero cell of the STIT tessellation in $\mathcal{S}$ with associated zonoid $P_{\mathcal{S}}\Pi$. Next, it follows from Section 8 of Hug and Schneider (2007) that for fixed $r > 0$ and $\tau \in (0, 1)$, there exists a constant $c_r = c_r(\tau, \Pi)$ such that for all $a > r$,

$$\mathbb{P}(\mathrm{diam}(Z_{0,\mathcal{S}}) \geq a) \leq c_r e^{-\tau a h_{\min}(P_{\mathcal{S}}\Pi)}. \tag{7}$$

The moments of the diameter then satisfy

$$\begin{aligned}
\mathbb{E}[\mathrm{diam}(Z_{0,\mathcal{S}})^k] &= \int_0^\infty k t^{k-1} \mathbb{P}(\mathrm{diam}(Z_{0,\mathcal{S}}) \geq t) \mathrm{d}t \\
&\leq \int_0^r k t^{k-1} \mathrm{d}t + k c_r \int_r^\infty t^{k-1} e^{-\tau t h_{\min}(P_{\mathcal{S}}\Pi)} \mathrm{d}t \\
&= r^k + \frac{k c_r}{(\tau h_{\min}(P_{\mathcal{S}}\Pi))^k} \int_0^\infty y^{k-1} e^{-y} \mathrm{d}y \leq r^k + \frac{c_r \Gamma(k+1)}{\tau^k h_{\min}(P_{\mathcal{S}}\Pi)^k}.
\end{aligned}$$

Letting $\tau = 2^{-1/k}$ and $r = \frac{(\Gamma(k+1))^{1/k}}{\tau h_{\min}(P_{\mathcal{S}}\Pi)}$ gives

$$\mathbb{E}[\mathrm{diam}(Z_{0,\mathcal{S}})^k] \leq \frac{2(1 + c_r)\Gamma(k+1)}{h_{\min}(P_{\mathcal{S}}\Pi)^k},$$

and by (3),

$$\mathbb{E}[\mathrm{diam}(Z_0^\lambda)^k] = \frac{1}{\lambda^k}\mathbb{E}[\mathrm{diam}(Z_{0,\mathcal{S}})^k] \leq \frac{c(k,\Pi)}{\lambda^k h_{\min}(P_{\mathcal{S}}\Pi)^k},$$

where $c(k,\Pi) := 2(1+c_r)\Gamma(k+1)$ for the chosen $r = r(k,\Pi)$. ■

Remark 3 *For the isotropic model as in Example 2, the normalized associated zonoid is the ball $c_D B^D$, and $h_{\min}(\Pi) = c_D$. In fact, $h_{\min}(\Pi)$ is maximal in the isotropic case because for any other ϕ and Π defined by (5), there will be a direction v for which $h(\Pi, v) \leq \int_{\mathbb{S}^{D-1}} h(\Pi, u)\mathrm{d}\sigma_{D-1}(u) = c_D$, where σ_{D-1} is the uniform distribution on $\mathbb{S}^{D-1}$. Thus, the exponential rate of the tail bound (7) is maximal for the isotropic STIT tessellation.*

3.2 Number of Cells in a Compact Domain

The following upper bound on the number of cells of $\mathcal{P}(\lambda)$ that intersect a compact and convex subset of $\mathbb{R}^D$ follows from equation (1).

Lemma 4 *Let K be a compact and convex set contained in a d-dimensional subspace $\mathcal{S}$ of $\mathbb{R}^D$. Let $N_\lambda(K)$ be the number of cells of $\mathcal{P}(\lambda)$ that intersect K. Then,*

$$\mathbb{E}[N_\lambda(K)] = \mathrm{vol}_d(P_{\mathcal{S}}\Pi)\sum_{k=0}^{d}\binom{d}{k}\lambda^k\mathbb{E}[V(K[k], Z_{\mathcal{S}}[d-k])],$$

where $Z_{\mathcal{S}}$ is the typical cell of the STIT tessellation in $\mathcal{S}$ with associated zonoid $P_{\mathcal{S}}\Pi$ and $\mathbb{E}[V(K[k], Z[r-k])] := \mathbb{E}[V(\underbrace{K, \ldots, K}_{k}, \underbrace{Z, \ldots, Z}_{r-k})]$.

The mixed volume $V(K_1, \ldots, K_d)$ of a collection of convex bodies $K_1, \ldots, K_d$ is non-negative, translation-invariant, multilinear, and symmetric in its arguments (Schneider, 2013, Section 5.1). For $k \in \mathbb{N}$, let B^k denote the unit ball in $\mathbb{R}^k$, and define $\kappa_k := \mathrm{vol}_k(B^k)$. The intrinsic volumes of a convex body $K \subset \mathbb{R}^d$ are defined for $j = 0, \ldots, d$ by

$$V_j(K) := \frac{\binom{d}{j}}{\kappa_{d-j}} V(K[j], B^d[d-j]).$$

The case $j = d$ is the usual volume, i.e. $V_d = \mathrm{vol}_d$. If $K \subset \mathbb{R}^d$ has dimension $j < d$, the normalization ensures $V_j(K)$ is the usual j-dimensional volume $\mathrm{vol}_j(K)$.

Example 3 *If $K = RB^d$, a ball of radius R in $\mathbb{R}^d$, then*

$$\begin{aligned}\mathbb{E}[N_\lambda(RB^d)] &= \mathrm{vol}_d(\Pi)\sum_{k=0}^{d}\binom{d}{k}\lambda^k R^k\mathbb{E}[V(B^d[k], Z[d-k])] \\ &= \mathrm{vol}_d(\Pi)\sum_{k=0}^{d}\lambda^k R^k\kappa_k\mathbb{E}[V_{d-k}(Z)].\end{aligned}$$

By (10.3) and Theorem 10.3.3 in Schneider and Weil (2008), $\mathbb{E}V_{d-k}(Z) = \frac{V_k(\Pi)}{\mathrm{vol}_d(\Pi)}$. *Thus,*

$$\mathbb{E}[N_\lambda(RB^d)] = \sum_{k=0}^{d} (\lambda R)^k \kappa_k V_k(\Pi). \tag{8}$$

Example 4 *For the isotropic STIT (see Example 2) and any convex body* $K \subset \mathbb{R}^D$*, Proposition 3 in Schreiber and Thäle (2011) gives*

$$\mathbb{E}[N_\lambda(K)] = \sum_{k=0}^{D} \left(\prod_{j=1}^{k} \gamma_j \right) \frac{\lambda^k}{k!} V_k(K),$$

where the constant γ_j *is defined as* $\gamma_j := \frac{\Gamma(\frac{j+1}{2})\Gamma(\frac{D}{2})}{\Gamma(\frac{j}{2})\Gamma(\frac{D+1}{2})}$.

Example 5 *Suppose* $K = [0,1]^D$ *and* Z *is the typical cell of a STIT tessellation with directional distribution* $\phi = \frac{1}{2D}\sum_{i=1}^{D}(\delta_{e_i} + \delta_{-e_i})$ *and lifetime parameter* D*. This setting corresponds with the Mondrian process in Mourtada et al. (2020). Then,* $h(Z,u) = \frac{T_i}{2}\sum_{i=1}^{D}|\langle u, e_i\rangle|$*, where* $T_1, \dots T_d$ *are i.i.d. exponential random variables with unit mean. By the formula for mixed volumes of zonoids from Schneider and Weil (2008, p. 614),*

$$V(K[k], Z[d-k]) \overset{d}{=} \prod_{i=1}^{D} T_i,$$

and $\mathbb{E}[V(K[k], Z[d-k])] = 1$*. Lemma 4 then implies*

$$\mathbb{E}[N_\lambda([0,1]^D)] = \sum_{k=0}^{D} \binom{d}{k} \lambda^k = (1+\lambda)^D,$$

which recovers Proposition 2 in Mourtada et al. (2020).

Proof (of Lemma 4) Let Z_λ denote the typical cell of $\mathcal{P}(\lambda)$. We first note that by Fact 1 and (3), the random polytope $Z_\lambda \cap \mathcal{S}$ has the same distribution as $\lambda^{-1}Z_{\mathcal{S}}$, where $Z_{\mathcal{S}}$ is as defined in the Lemma. Then, applying (1) with the indicator function $f(\cdot) = 1_{\{\cdot \cap K \neq \emptyset\}}$ and (6) gives that the expected number of cells of $\mathcal{P}(\lambda)$ intersecting $K \subset \mathcal{S}$ satisfies

$$\begin{aligned}
\mathbb{E}[N_\lambda(K)] &= \mathbb{E}\left[\sum_{C \in \mathcal{P}(\lambda)} 1_{\{C \cap K \neq \emptyset\}}\right] = \mathbb{E}\left[\sum_{C \in \mathcal{P}(\lambda) \cap \mathcal{S}} 1_{\{C \cap K \neq \emptyset\}}\right] \\
&= \frac{\lambda^d}{\mathbb{E}[\mathrm{vol}_d(Z_{\mathcal{S}})]} \mathbb{E}\left[\int_{\mathcal{S}} 1_{\{\lambda^{-1}Z_{\mathcal{S}} + y \cap K \neq \emptyset\}} \mathrm{d}y\right] \\
&= \lambda^d \mathrm{vol}_d(P_{\mathcal{S}}\Pi) \mathbb{E}[\mathrm{vol}_d(K - \lambda^{-1}Z_{\mathcal{S}})] \\
&= \lambda^d \mathrm{vol}_d(P_{\mathcal{S}}\Pi) \sum_{k=0}^{d} \binom{d}{k} \mathbb{E}[V(K[k], -\lambda^{-1}Z_{\mathcal{S}}[d-k])].
\end{aligned}$$

The last equality follows from Steiner's formula; see equation (14.20) in Schneider and Weil (2008). The third equality follows from the fact that $\lambda^{-1}Z_{\mathcal{S}} + y \cap K \neq \emptyset$ if and only if $y \in K - \lambda^{-1}Z_{\mathcal{S}}$. By the scaling property of mixed volumes and the fact that $Z \overset{d}{=} -Z$,

$$\mathbb{E}[V(K[k], -\lambda^{-1}Z_{\mathcal{S}}[d-k])] = \lambda^{-(d-k)}\mathbb{E}[V(K[k], Z_{\mathcal{S}}[d-k])].$$

Thus,

$$\mathbb{E}[N_\lambda(K)] = \mathrm{vol}_d(P_{\mathcal{S}}\Pi)\sum_{k=0}^{d}\binom{d}{k}\lambda^k\mathbb{E}[V(K[k], Z_{\mathcal{S}}[d-k])].$$

■

4. Main Results

Fix a non-empty compact and convex D-dimensional domain $W \subset \mathbb{R}^D$, and consider the following regression setting. The data set $\mathcal{D}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consists of n i.i.d. samples from a random pair $(X, Y) \in W \times \mathbb{R}$ such that $\mathbb{E}[Y^2] < \infty$. Let μ denote the unknown distribution of X and

$$Y = f(X) + \varepsilon,$$

where $f(X) = \mathbb{E}[Y|X]$ is the conditional expectation of Y given X and ε is noise such that $\mathbb{E}[\varepsilon|X] = 0$ and $\mathrm{Var}(\varepsilon|X) = \sigma^2 < \infty$ almost surely.

Let $\mathcal{P}$ be a random tessellation of W. The regression tree estimator based on $\mathcal{P}$ is

$$\hat{f}_n(x, \mathcal{P}) := \sum_{i=1}^{n}\frac{1_{\{X_i \in Z_x\}}}{\mathcal{N}_n(x)}Y_i, \tag{9}$$

where Z_x is the cell of $\mathcal{P}$ that contains x and $\mathcal{N}_n(x) := \sum_{i=1}^{n} 1_{\{X_i \in Z_x\}}$ is the number of points in Z_x. If $\mathcal{N}_n(x) = 0$, then it is assumed that $\hat{f}_n(x, \mathcal{P}) = 0$. The random forest estimator based on $\mathcal{P}$ is defined by averaging M i.i.d. copies of the tree estimator, i.e.

$$\hat{f}_{n,M}(x) := \frac{1}{M}\sum_{m=1}^{M}\hat{f}_n(x, \mathcal{P}_m), \tag{10}$$

where $\mathcal{P}_1, \ldots, \mathcal{P}_M$ are M i.i.d. copies of $\mathcal{P}$.

We define the STIT regression tree estimator $\hat{f}_{\lambda,n}$ and the STIT regression forest estimator $\hat{f}_{\lambda,n,M}$ as in (9) and (10) respectively, where $\mathcal{P} := \mathcal{P}(\lambda) \cap W$ is the random tessellation on W generated by a STIT tessellation $\mathcal{P}(\lambda)$ with lifetime parameter λ and normalized associated zonoid Π. The quality of the estimator $\hat{f}_{\lambda,n,M}$ is measured by the quadratic risk

$$R(\hat{f}_{\lambda,n,M}) := \mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2].$$

We now define the function classes we will consider in our results. For $k \in \mathbb{N}$, $\beta \in (0, 1]$, and $L > 0$, define the $(k+\beta)$-Hölder ball of norm L, denoted by $\mathcal{C}^{k,\beta}(L) = \mathcal{C}^{k,\beta}(W, L)$, to

be the set of all k times differentiable functions $f : W \to \mathbb{R}$ such that for all multi-indices α with $|\alpha| \leq k$,

$$\|D^\alpha f(x) - D^\alpha f(y)\| \leq L\|x - y\|^\beta \text{ and } \|D^\alpha f(x)\| \leq L,$$

for all $x, y \in W$. The minimax rate for the class $\mathcal{C}^{k,\beta}(L)$ is $n^{-2(k+\beta)/(2(k+\beta)+D)}$ (Györfi et al., 2002, Theorem 3.2).

Our main results show that for an appropriate choice of lifetime λ, STIT forest estimators achieve the minimax rate of convergence for the function classes $\mathcal{C}^{0,\beta}(L)$ and $\mathcal{C}^{1,\beta}(L)$. The rates are also adaptive to the intrinsic dimension of the input data, defined as follows.

Definition 5 *The input X has intrinsic dimension d if the support of its distribution μ is contained in an d-dimensional linear subspace $\mathcal{S}$ of $\mathbb{R}^D$.*

We now state our first main result.

Theorem 6 *Assume X has intrinsic dimension d and $f \in \mathcal{C}^{0,\beta}(L)$ for $\beta \in (0, 1]$ and $L > 0$. Then,*

$$R(\hat{f}_{\lambda,n,M}) \leq \frac{Lc_{\beta,P_\mathcal{S}\Pi}}{\lambda^{2\beta} h_{\min}(P_\mathcal{S}\Pi)^{2\beta}} + \frac{(5\|f\|_\infty^2 + 2\sigma^2)\mathrm{vol}_d(P_\mathcal{S}\Pi)}{n} \sum_{k=0}^{d} \binom{d}{k} \lambda^k \mathbb{E}\left[V\left(W_\mathcal{S}[k], Z_\mathcal{S}[d-k]\right)\right], \tag{11}$$

where $W_\mathcal{S} := W \cap \mathcal{S}$ and $Z_\mathcal{S}$ is the typical cell of the STIT tessellation in $\mathcal{S}$ with associated zonoid $P_\mathcal{S}\Pi$. If $\mathrm{diam}(W) \leq 2R$, then

$$R(\hat{f}_{\lambda,n,M}) \leq \frac{Lc_{\beta,P_\mathcal{S}\Pi}}{\lambda^{2\beta} h_{\min}(P_\mathcal{S}\Pi)^{2\beta}} + \frac{(5\|f\|_\infty^2 + 2\sigma^2)}{n} \sum_{k=0}^{d} \lambda^k R^k \kappa_k V_k\left(P_\mathcal{S}\Pi\right). \tag{12}$$

Corollary 7 *In the setting of Theorem 6, letting $\lambda_n \sim L^{2/(d+2\beta)} n^{1/(d+2\beta)}$ as $n \to \infty$ yields*

$$R(\hat{f}_{\lambda,n,M}) = O\left(L^{2d/(d+2\beta)} n^{-2\beta/(d+2\beta)}\right), \tag{13}$$

which is the minimax rate for the class $\mathcal{C}^{0,\beta}(L)$ on $\mathbb{R}^d$.

The minimax rate above holds even for STIT tree estimators. To see the advantage of averaging multiple trees in a random forest estimator, the following result assumes additional smoothness on the function f. In this case, the minimax rate is only achieved when the forest size M is large enough.

Theorem 8 *Assume X has intrinsic dimension d and the distribution μ of X has a positive and Lipschitz density with respect to the Lebesgue measure on its d-dimensional convex support K. Let $r(K)$ denote the inradius of K and define $K_\varepsilon := \{x \in K : d(x, \partial K) \geq \varepsilon\}$. Assume $f \in \mathcal{C}^{1,\beta}(L)$ for $\beta \in (0, 1]$ and $L > 0$. Then, for $\varepsilon \in (0, r(K))$,*

$$\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in K_\varepsilon] \leq O\left(\frac{L^2}{\lambda^2 M} + \frac{L^2}{\lambda^{2\beta+2}} + \frac{\lambda^d}{n}\right).$$

In the unconditional case when $\varepsilon = 0$,

$$\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2] \leq O\left(\frac{L^2}{\lambda^2 M} + \frac{L^2}{\lambda^{\min\{3,2\beta+2)}\}} + \frac{\lambda^d}{n}\right).$$

Corollary 9 *In the setting of Theorem 8, choosing*

$$\lambda_n \sim L^{2/(d+2\beta+2)} n^{1/(d+2\beta+2)} \quad \text{and} \quad M_n \gtrsim L^{4\beta/(d+2\beta+2)} n^{2\beta/(d+2\beta+2)}$$

as $n \to \infty$ *implies*

$$\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in K_\varepsilon] = O\left(L^{2d/(d+2\beta+2)} n^{-(2\beta+2)/(d+2\beta+2)}\right), \tag{14}$$

which is the minimax rate for the class $\mathcal{C}^{1,\beta}(L)$ *on* $\mathbb{R}^d$. *In the unconditional case when* $\varepsilon = 0$, *the rate* (14) *is obtained when* $\beta \leq 1/2$. *When* $\beta > 1/2$, *choosing*

$$\lambda_n \sim L^{2/(d+3)} n^{1/(d+3)} \quad \text{and} \quad M_n \gtrsim L^{4/(d+3)} n^{2/(d+3)}$$

as $n \to \infty$ *implies*

$$\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2] = O\left(L^{2d/(d+3)} n^{-3/(d+3)}\right).$$

Remark 10 *Specialized to the case of the Mondrian, our rates are an improvement over the results of Mourtada et al. (2020), where no notion of the low dimensionality of the input was considered. However, note that the rates are obtained through optimal choices of* λ *and* M *that depend on* d *and* β. *In practice, the intrinsic dimension of the input and the regularity of* f *are not known* a priori, *and it is an open problem to find an adaptive way of choosing the lifetime parameter* λ *such that the random forest estimator achieves these optimal rates without this prior knowledge.*

The upper bounds on the risk in Theorem 6 and 8 only rely on statistics of the typical cell and the zero cell of the STIT tessellation $\mathcal{P}(\lambda)$. Since these values are identical for a STIT and a stationary Poisson hyperplane tessellation with matching parameters, it follows that regression estimators based on stationary Poisson hyperplane processes have *identical* risk bounds. We state this formally as follows.

Theorem 11 *Let* $\hat{f}_{\lambda,n,M}$ *be the random forest estimator defined as in* (10) *using* M *i.i.d. random tessellations induced by a stationary Poisson hyperplane process with intensity* λ *and normalized associated zonoid* Π. *Then, the minimax optimal convergence rates* (13) *and* (14) *for the risk of* $\hat{f}_{\lambda,n,M}$ *hold in each corresponding setting.*

Remark 12 *While the rates in Corollaries 7 and 9 depend only on the intrinsic dimension of* X, *the ambient dimension* D *appears in the upper bounds of Theorems 6 and 8 through the constants. To clarify the dependence on* D, *we insert* $\lambda = L^{1/(d+2\beta)} n^{1/(d+2\beta)}$ *into* (12) *to obtain*

$$R(\hat{f}_{\lambda,n,M}) \leq \left(\frac{c_{\beta,P_{\mathcal{S}}\Pi}}{h_{\min}(P_{\mathcal{S}}\Pi)^{2\beta}} + (5\|f\|_\infty^2 + 2\sigma^2) R^d \kappa_d \mathrm{vol}_d(P_{\mathcal{S}}\Pi)\right) L^{\frac{d}{d+2\beta}} n^{\frac{-2\beta}{d+2\beta}} + o\left(n^{\frac{-2\beta}{d+2\beta}}\right).$$

Since the leading order constant involves geometric properties of $P_{\mathcal{S}}\Pi$, the order with respect to D will depend on the subspace $\mathcal{S}$ and the shape of Π. We first note that the constant $\mathrm{vol}_d(P_{\mathcal{S}}\Pi)$ has a uniform upper bound that does not depend on D. Indeed, since $h(\Pi, u) \leq 1$ for all $u \in \mathbb{S}^{D-1}$, $\mathrm{vol}_d(P_{\mathcal{S}}\Pi) \leq \kappa_d$ for any subspace $\mathcal{S}$ of dimension d. However, the constant $h_{\min}(P_{\mathcal{S}}\Pi)^{-2\beta}$ could grow faster with D if the subspace $\mathcal{S}$ lies in "bad" directions with respect to Π. This observation highlights a potential downside of using axis-aligned splits. Indeed, for the Mondrian (see Example 1), the associated zonoid is $\Pi = \frac{1}{D}[-1,1]^D$ and if $\mathcal{S} = \mathit{span}(\mathbf{1})$, then $h_{\min}(P_{\mathcal{S}}\Pi)^{-2\beta} = O(D^{\beta})$. However, if $\mathcal{S} = \mathrm{span}(e_1)$, then $h_{\min}(P_{\mathcal{S}}\Pi)^{-2\beta} = O(D^{2\beta})$. Without further knowledge of the subspace one can instead choose an isotropic STIT (see Example 2), where the associated zonoid is $\Pi = c_D B^D$ and $h_{\min}(P_{\mathcal{S}}\Pi)^{-2\beta} = O(D^{\beta})$ for any subspace $\mathcal{S}$. Unfortunately, due to the unknown dependence of the constant $c_{\beta, P_{\mathcal{S}}\Pi}$ on D, we cannot currently make this analysis more precise. We leave for future work a more thorough study of the statistical advantages of oblique splits; see Section 6.

4.1 Rates for Binary Classification

We next show that we can extend all of the above results to binary classification as was done for Mondrian random forests in Section 5.5 of Mourtada et al. (2020). In this setting, we assume the data are i.i.d. samples from a random pair $(X, Y) \in W \times \{0, 1\}$. We then define the function $\eta(x) := \mathbb{P}(Y = 1 | X = x)$ and the optimal classifier $g(x) := 1_{\{\eta(x) \geq 1/2\}}$. The STIT (or Poisson hyperplane) forest classifier $\hat{g}_{\lambda,n,M}$ is defined by

$$\hat{g}_{\lambda,n,M}(x) := 1_{\{\hat{\eta}_{\lambda,n,M}(x) \geq 1/2\}}, \qquad x \in W,$$

where $\hat{\eta}_{\lambda,n,M}(x)$ is the STIT (or Poisson hyperplane) forest estimator of η as defined in the previous section. The risk of $\hat{g}_{\lambda,n,M}$ is given by the classification error

$$\mathcal{L}(\hat{g}_{\lambda,n,M}) := \mathbb{P}(\hat{g}_{\lambda,n,M}(X) \neq Y).$$

To evaluate the classification estimator, we compare $\mathcal{L}(\hat{g}_{\lambda,n,M})$ to the Bayes risk $\mathcal{L}(g) := \mathbb{P}(g(X) \neq Y)$ and consider the difference

$$R(\hat{g}_{\lambda,n,M}) := \mathcal{L}(\hat{g}_{\lambda,n,M}) - \mathcal{L}(g).$$

A general theorem (Devroye et al., 1996, Theorem 6.5) shows that $R(\hat{g}_{\lambda,n,M})$ is controlled by the square root of the risk of the regression estimator $\hat{\eta}_{\lambda,n,M}$ and immediately implies the following Corollary of Theorem 6.

Corollary 13 *Assume X has intrinsic dimension d and $\eta \in \mathcal{C}^{0,\beta}(L)$ for $\beta \in (0, 1]$ and $L > 0$. Then, letting $\lambda_n \sim n^{1/(d+2\beta)}$ as $n \to \infty$ yields*

$$R(\hat{g}_{\lambda_n,n,M}) = o\left(n^{-\beta/(d+2\beta)}\right).$$

We similarly obtain the following Corollary of Theorem 8 showing that we can obtain a faster rate with forest estimators than with single trees.

Corollary 14 *Assume X has intrinsic dimension d and the distribution μ of X has a positive and Lipschitz density with respect to Lebesgue measure on its d-dimensional convex*

support K. Let $r(K)$ denote the inradius of K and define $K_\varepsilon := \{x \in K : d(x, \partial K) \geq \varepsilon\}$. Assume $\eta \in \mathcal{C}^{1,\beta}(L)$ for $\beta \in (0,1]$ and $L > 0$. Then, for $\varepsilon \in (0, r(K))$, choosing

$$\lambda_n \sim n^{1/(d+2\beta+2)} \quad \text{and} \quad M_n \gtrsim n^{2\beta/(d+2\beta+2)}$$

as $n \to \infty$ implies

$$\mathbb{P}\left(\hat{g}_{\lambda_n,n,M_n}(X) \neq Y | X \in K_\varepsilon\right) - \mathbb{P}\left(g(X) \neq Y | X \in K_\varepsilon\right) = o\left(n^{-(\beta+1)/(d+2\beta+2)}\right).$$

5. Proofs

To prove the above results, we begin with the following bias-variance decomposition of the risk of a tree estimator presented by Arlot and Genuer (2014). A subtle difference between their setting and ours is that they view the partition as a finite partitioning of $[0,1]^d$, and here we consider the partition to be a stationary STIT tessellation on $\mathbb{R}^d$ which we view through the compact and convex window W that contains the support of μ. First, let Z_x^λ denote the cell of $\mathcal{P}(\lambda)$ that contains the vector $x \in \mathbb{R}^d$, and define

$$\bar{f}_\lambda(x) := \mathbb{E}_X[f(X) | X \in Z_x^\lambda], \quad x \in W.$$

Conditioned on $\mathcal{P}(\lambda)$, this is the orthogonal projection of $f \in L^2(W, \mu)$ onto the subspace of functions that are constant within the cells of $\mathcal{P}(\lambda) \cap W$.

Conditioned additionally on the data $\mathcal{D}_n$, the random tree estimator $\hat{f}_{\lambda,n}$ is in this subspace of piecewise functions, and hence $\mathbb{E}_X[(f(X) - \bar{f}_\lambda(X))\hat{f}_{\lambda,n}(X)] = 0$. Thus, given $\mathcal{P}(\lambda)$ and $\mathcal{D}_n$,

$$\begin{aligned}\mathbb{E}_X[(f(X) - \hat{f}_{\lambda,n}(X))^2] &= \mathbb{E}_X[(f(X) - \bar{f}_\lambda(X) + \bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2] \\ &= \mathbb{E}_X[(f(X) - \bar{f}_\lambda(X))^2] + \mathbb{E}_X[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2].\end{aligned}$$

Taking the expectation over $\mathcal{P}(\lambda)$ and $\mathcal{D}_n$ gives the following decomposition of the risk:

$$R(\hat{f}_{\lambda,n}) := \mathbb{E}[(f(X) - \hat{f}_{\lambda,n}(X))^2] = \mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] + \mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2]. \quad (15)$$

The first term measures how far away f is from the closest function in the hypothesis class that the estimators lie in and is called the approximation error or bias. The second term measures the estimation error, or variance, coming from the fact that we build the estimator from only a finite number of samples. As in the results of Mourtada et al. (2020), the bias and variance depend on the geometric properties of the cells of the tessellations from which the estimator is built. In particular, the bias is controlled by moments of the diameter of the zero cell, and the variance is controlled by the expected number of cells that have a non-empty intersection with the support of μ. Lemmas 2 and 4 provide the needed bounds, and choosing an optimal lifetime λ depending on the number of samples and Lipschitz constant gives the results.

5.1 Variance Bound

In the following, we see that the variance term can be controlled by the expected number of cells of the tessellation that intersect the support of μ.

Lemma 15 *Let $N_\lambda(K)$ be the number of cells of $\mathcal{P}(\lambda)$ that have a non-empty intersection with a bounded subset $K \subset \mathbb{R}^D$. Then,*

$$\mathbb{E}\left[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2\right] \leq \frac{5\|f\|_\infty^2 + 2\sigma^2}{n}\mathbb{E}[N_\lambda(\mathrm{supp}(\mu))].$$

The proof follows the ideas of Arlot and Genuer (2014, Proposition 2) which relies crucially on a result of Arlot (2008, Proposition 1). For completeness and clarity, a proof of this lemma appears below.

Proof We first condition on $\mathcal{P}(\lambda)$ and compute the variance of the tree estimator corresponding to a fixed tessellation. Note that the assumption $\mathcal{D}_n$ and $\mathcal{P}(\lambda)$ are independent allows us to take these expectations separately. Also, recall that if no points of $\{X_1, \ldots, X_n\}$ fall in Z_x^λ, then $\hat{f}_{\lambda,n}(x) = 0$. For each $C \in \mathcal{P}(\lambda)$, let $\mathcal{N}_n(C) = \sum_{i=1}^n 1_{\{X_i \in C\}}$ be the number of covariates inside C and let $p_{\lambda,C} := \mathbb{P}_X(X \in C)$. Then,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n,X}\left[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2\right] \\
&= \int_{\mathbb{R}^d} \sum_{C\in\mathcal{P}(\lambda)} 1_{\{x\in C\}}\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbb{E}_X[f(X)|X\in C] - \frac{\sum_{i=1}^n Y_i 1_{\{X_i\in C\}}}{\mathcal{N}_n(C)}\right)^2\right] \mathrm{d}\mu(x) \\
&= \sum_{C\in\mathcal{P}(\lambda):C\cap\mathrm{supp}(\mu)\neq\emptyset} p_{\lambda,C}\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbb{E}_X[f(X)|X\in C] - \frac{\sum_{i=1}^n Y_i 1_{\{X_i\in C\}}}{\mathcal{N}_n(C)}\right)^2\right].
\end{aligned}$$

The expectation in the sum satisfies

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbb{E}_X[f(X)|X\in C] - \frac{\sum_{i=1}^n Y_i 1_{\{X_i\in C\}}}{\mathcal{N}_n(C)}\right)^2\right] \\
&= \sum_{k=1}^n \mathbb{P}(\mathcal{N}_n(C)=k)\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbb{E}_X[f(X)|X\in C] - \frac{\sum_{i=1}^n Y_i 1_{\{X_i\in C\}}}{k}\right)^2 \middle| \mathcal{N}_n(C)=k\right] \\
&\quad + \mathbb{P}(\mathcal{N}_n(C)=0)\mathbb{E}_X[f(X)|X\in C]^2.
\end{aligned}$$

By the assumptions on the noise, the conditional expectation in the sum satisfies

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbb{E}_X[f(X)|X\in C]-\frac{\sum_{i=1}^n Y_i 1_{\{X_i\in C\}}}{k}\right)^2\,\middle|\,\mathcal{N}_n(C)=k\right]\\
&=k^{-2}\mathbb{E}_{\mathcal{D}_n}\left[\left(k\mathbb{E}_X[f(X)|X\in C]-\sum_{i=1}^n(f(X_i)+\varepsilon_i)1_{\{X_i\in C\}}\right)^2\,\middle|\,\mathcal{N}_n(C)=k\right]\\
&=k^{-2}\sum_{i_1<\cdots<i_k}\mathbb{P}(X_{i_1},\ldots,X_{i_k}\in C|\mathcal{N}_n(C)=k)\\
&\cdot\mathbb{E}_{\mathcal{D}_n}\left[\left(k\mathbb{E}_X[f(X)|X\in C]-\sum_{j=1}^k f(X_{i_j})-\sum_{j=1}^k\varepsilon_{i_j}\right)^2\,\middle|\,\mathcal{N}_n(C)=k,X_{i_1},\ldots,X_{i_k}\in C\right]\\
&=k^{-2}\mathbb{E}_{\mathcal{D}_n}\left[\left(k\mathbb{E}_X[f(X)|X\in C]-\sum_{i=1}^k f(X_i)-\sum_{i=1}^k\varepsilon_i\right)^2\,\middle|\,X_1,\ldots,X_k\in C\right]\\
&=k^{-2}\mathbb{E}_{\mathcal{D}_n}\left[\left(k\mathbb{E}_X[f(X)|\,X\in C]-\sum_{i=1}^k f(X_i)\right)^2\,\middle|\,X_1,\ldots,X_k\in C\right]+k^{-1}\sigma^2,
\end{aligned}
$$

and by the independence of the X_i's, the expectation in the first term simplifies to

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\left(k\mathbb{E}_X[f(X)|X\in C]-\sum_{i=1}^k f(X_i)\right)^2\,\middle|\,X_1,\ldots,X_k\in C\right]\\
&=k^2\mathbb{E}_X[f(X)|X\in C]^2-2k^2\mathbb{E}_X[f(X)|X\in C]^2\\
&\qquad+\mathbb{E}_{\mathcal{D}_n}\left[\sum_{i,j=1}^k f(X_i)f(X_j)\middle|\,X_1,\ldots,X_k\in C\right]\\
&=k\mathbb{E}_X[f(X)^2|X\in C]+(k^2-k)\mathbb{E}_X[f(X)|X\in C]^2-k^2\mathbb{E}_X[f(X)|X\in C]^2\\
&=k(\mathbb{E}_X[f(X)^2|X\in C]-\mathbb{E}_X[f(X)|X\in C]^2).
\end{aligned}
$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\left[\left(\mathbb{E}_X[f(X)|X\in C]-\frac{\sum_{i=1}^n Y_i 1_{\{X_i\in C\}}}{\mathcal{N}_n(C)}\right)^2\right]\\
&=\sum_{k=1}^n \mathbb{P}(\mathcal{N}_n(C)=k)k^{-1}\left(\mathbb{E}_X[f(X)^2|X\in C]-\mathbb{E}_X[f(X)|X\in C]^2+\sigma^2\right)\\
&\quad+\mathbb{P}(\mathcal{N}_n(C)=0)\mathbb{E}_X[f(X)|X\in C]^2\\
&=\left(\mathbb{E}_X[f(X)^2|X\in C]-\mathbb{E}_X[f(X)|X\in C]^2+\sigma^2\right)\sum_{k=1}^n\binom{n}{k}p_{\lambda,C}^k(1-p_{\lambda,C})^{n-k}k^{-1}\\
&\quad+\mathbb{E}_X[f(X)|X\in C]^2(1-p_{\lambda,C})^n\\
&\le\left(2\|f\|_\infty^2+\sigma^2\right)\sum_{k=1}^n\binom{n}{k}p_{\lambda,C}^k(1-p_{\lambda,C})^{n-k}k^{-1}+\|f\|_\infty^2(1-p_{\lambda,C})^n.
\end{aligned}$$

Now, note that for $B\sim\text{Binomial}(n,p_{\lambda,C})$,

$$\sum_{k=1}^n\binom{n}{k}np_{\lambda,C}^{k+1}(1-p_{\lambda,C})^{n-k}k^{-1}=\mathbb{E}[B]\mathbb{E}[B^{-1}1_{\{B>0\}}],$$

and $\mathbb{E}[B]\mathbb{E}[B^{-1}1_{\{B>0\}}]\le\frac{2np_{\lambda,C}}{(n+1)p_{\lambda,C}}\le 2$ (Györfi et al., 2002, Lemma 4.1). Also, the upper bounds $1-x\le e^{-x}$ and $xe^{-x}\le e^{-1}$ for all $x\ge 0$ imply

$$np_{\lambda,C}(1-p_{\lambda,C})^n\le e^{-1}\le 1.$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n,X}\left[(\bar{f}_\lambda(X)-\hat{f}_{\lambda,n}(X))^2\right]&\le\frac{1}{n}\sum_{\substack{C\in\mathcal{P}(\lambda):\\ C\cap\text{supp}(\mu)\neq\emptyset}}\left(2\|f\|_\infty^2+\sigma^2\right)\sum_{k=1}^n\binom{n}{k}np_{\lambda,C}^{k+1}(1-p_{\lambda,C})^{n-k}k^{-1}\\
&\quad+\frac{\|f\|_\infty^2}{n}\sum_{\substack{C\in\mathcal{P}(\lambda):\\ C\cap\text{supp}(\mu)\neq\emptyset}}np_{\lambda,C}(1-p_{\lambda,C})^n\\
&\le\frac{5\|f\|_\infty^2+2\sigma^2}{n}N_\lambda(\text{supp}(\mu)).
\end{aligned}$$

Taking the expectation with respect to $\mathcal{P}(\lambda)$ completes the proof. ■

5.2 Proof of Theorem 6 and Corollary 7

Following the proof of Theorem 2 by Mourtada et al. (2020), we first use Jensen's inequality to reduce the risk of $\hat{f}_{\lambda,n,M}$ to that of a single Mondrian tree estimator $\hat{f}_{\lambda,n}:=\hat{f}_{\lambda,n,1}$. Using the bias-variance decomposition (15), the risk of $\hat{f}_{\lambda,n}$ is given by

$$R(\hat{f}_{\lambda,n})=\mathbb{E}[(f(X)-\hat{f}_{\lambda,n}(X))^2]=\mathbb{E}[(f(X)-\bar{f}_\lambda(X))^2]+\mathbb{E}[(\bar{f}(X)-\hat{f}_{\lambda,n}(X))^2]. \quad (16)$$

We first consider the bias term. For $x \in \text{supp}(\mu)$, by the assumption on f,

$$\begin{aligned}|f(x) - \bar{f}_\lambda(x)| &\leq \frac{1}{\mu(Z_x^\lambda)} \int_{Z_x^\lambda} |f(x) - f(z)| \, \mu(\mathrm{d}z) \\ &\leq \frac{1}{\mu(Z_x^\lambda)} \int_{Z_x^\lambda} L\|x - z\|^\beta \mu(\mathrm{d}z) \leq L\text{diam}(Z_x^\lambda \cap \text{supp}(\mu))^\beta.\end{aligned}$$

Recall that the assumption X has intrinsic dimension d means there exists a d-dimensional linear subspace $\mathcal{S} \subseteq \mathbb{R}^d$ such that $\text{supp}(\mu) \subset \mathcal{S}$. Then by Lemma 2,

$$\mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] \leq L\mathbb{E}[\text{diam}(Z_x^\lambda \cap \mathcal{S})^{2\beta}] \leq \frac{Lc_{\beta, P_\mathcal{S}\Pi}}{\lambda^{2\beta} h_{\min}(P_\mathcal{S}\Pi)^{2\beta}}. \tag{17}$$

For the variance bound, Lemma 15 implies

$$\mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2] \leq \frac{5\|f\|_\infty^2 + 2\sigma^2}{n} \mathbb{E}[N_\lambda(\text{supp}(\mu))] \leq \frac{5\|f\|_\infty^2 + 2\sigma^2}{n} \mathbb{E}[N_\lambda(W \cap \mathcal{S})].$$

Let $W_\mathcal{S} := W \cap \mathcal{S}$. Finally by Lemma 4,

$$\mathbb{E}[N_\lambda(W_\mathcal{S})] = \text{vol}_d(P_\mathcal{S}\Pi) \sum_{k=0}^{d} \binom{d}{k} \lambda^k \mathbb{E}[V(W_\mathcal{S}[k], Z_\mathcal{S}[d-k])].$$

Then,

$$\mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2] \leq \frac{(5\|f\|_\infty^2 + 2\sigma^2)\text{vol}_d(P_\mathcal{S}\Pi)}{n} \sum_{k=0}^{d} \binom{d}{k} \lambda^k \mathbb{E}[V(W_\mathcal{S}[k], Z_\mathcal{S}[d-k])]. \tag{18}$$

Combining equations (17) and (18) gives the first claim. To prove (12), we note that if $\text{diam}(W) \leq 2R$, then $\mathbb{E}[N_\lambda(\text{supp}(\mu))] \leq \mathbb{E}[N_\lambda(RB^D \cap \mathcal{S})]$, and the bound follows from the same argument used in Example 3 to obtain (8).

Finally, letting $\lambda = \lambda_n \sim L^{2/(d+2\beta)} n^{1/(d+2\beta)}$ as $n \to \infty$ proves Corollary 7. ■

5.3 Proof of Theorem 8 and Corollary 9

We first need the following technical lemma.

Lemma 16 *Let Z_x^λ be the cell of $\mathcal{P}(\lambda)$ containing the point $x \in \mathbb{R}^D$. Assume $x \in \mathcal{S} \subseteq \mathbb{R}^D$ for a linear subspace $\mathcal{S}$ of dimension d. Then,*

$$\int_\mathcal{S} (z - x)\mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\text{vol}_d(Z_x^\lambda \cap \mathcal{S})}\right] \mathrm{d}z = 0.$$

Proof By the stationarity of $\mathcal{P}(\lambda)$ and a change of variable, for $x \in \mathcal{S}$,

$$\int_\mathcal{S} (z - x)\mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\text{vol}_d(Z_x^\lambda \cap \mathcal{S})}\right] \mathrm{d}z = \int_\mathcal{S} y\mathbb{E}\left[\frac{1_{\{y \in Z_0^\lambda\}}}{\text{vol}_d(Z_0^\lambda \cap \mathcal{S})}\right] \mathrm{d}y.$$

Then, for $y \in \mathcal{S}$, Fact 1 and (1) imply

$$\mathbb{E}\left[\frac{1_{\{y\in Z_0^\lambda\}}}{\mathrm{vol}_d(Z_0^\lambda \cap \mathcal{S})}\right] = \mathbb{E}\left[\frac{1_{\{y\in Z_0^\lambda\cap\mathcal{S}\}}}{\mathrm{vol}_d(Z_0^\lambda \cap \mathcal{S})}\right] = \frac{1}{\mathbb{E}[\mathrm{vol}_d(Z_\mathcal{S}^\lambda)]}\mathbb{E}\left[\frac{\mathrm{vol}_d(Z_\mathcal{S}^\lambda \cap Z_\mathcal{S}^\lambda - y)}{\mathrm{vol}_d(Z_\mathcal{S}^\lambda)}\right],$$

where $Z_\mathcal{S}^\lambda$ is the typical cell of $\mathcal{P}(\lambda)\cap\mathcal{S}$. By the fact that volume is translation invariant,

$$\mathbb{E}\left[\frac{\mathrm{vol}_d(Z_\mathcal{S}^\lambda \cap Z_\mathcal{S}^\lambda + y)}{\mathrm{vol}_d(Z_\mathcal{S}^\lambda \cap \mathcal{S})}\right] = \mathbb{E}\left[\frac{\mathrm{vol}_d(Z_\mathcal{S}^\lambda - y \cap Z_\mathcal{S}^\lambda)}{\mathrm{vol}_d(Z_\mathcal{S}^\lambda)}\right].$$

Thus the integrand $y\mathbb{E}\left[\frac{1_{\{y\in Z_0^\lambda\}}}{\mathrm{vol}_d(Z_0^\lambda\cap\mathcal{S})}\right]$ is an odd function, and the integral is zero. ■

Proof (of Theorem 8 and Corollary 9) Similarly to the proof of Theorem 3 by Mourtada et al. (2020), define for each m and $x \in \mathcal{S}$,

$$\bar{f}_\lambda^{(m)}(x) := \mathbb{E}_X[f(X)|X \in Z_x^{\lambda,(m)}],$$

and let $\bar{f}_{\lambda,M}(x) = \frac{1}{M}\sum_{m=1}^M \bar{f}_\lambda^{(m)}(x)$. Also define

$$\tilde{f}_\lambda(x) := \mathbb{E}[\bar{f}_\lambda^{(m)}(x)] = \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}\int_{Z_x^\lambda} f(z)\mu(dz)\right] = \int_\mathcal{S} f(z)\mathbb{E}\left[\frac{1_{\{z\in Z_x^\lambda\}}}{\mu(Z_x^\lambda)}\right]\mu(dz),$$

where we have used the fact that the support of μ is contained in a linear subspace $\mathcal{S}$. The bias-variance decomposition for the risk of a tree estimator can be extended to the random forest estimator as follows (Arlot and Genuer, 2014, equation 1):

$$\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2] = \mathbb{E}[(f(X) - \bar{f}_{\lambda,M}(X))^2] + \mathbb{E}[(\bar{f}_{\lambda,M}(X) - \hat{f}_{\lambda,n,M}(X))^2]. \quad (19)$$

This is due to the fact that $\mathbb{E}[\hat{f}_{\lambda,n,1}(x)|\mathcal{P}(\lambda)] = \bar{f}_{\lambda,1}(x)$. Indeed, by the independence of the X_i's,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}[\hat{f}_{\lambda,n,1}(x)] &= \frac{1}{n}\mathbb{E}_{\mathcal{D}_n}\left[\frac{\sum_{i=1}^n Y_i 1_{\{X_i\in Z_x\}}}{\mathcal{N}_n(Z_x)}\right] \\
&= \frac{1}{n}\sum_{k=1}^n \binom{n}{k}\mathbb{P}_{\mathcal{D}_n}(X_1,\dots,X_k \in Z_x | \mathcal{N}_n(Z_x) = k) \\
&\qquad \cdot \mathbb{E}_{\mathcal{D}_n}\left[\frac{\sum_{i=1}^k f(X_i)1_{\{X_i\in Z_x\}}}{k}\middle| X_1,\dots,X_k \in Z_x, \mathcal{N}_n(Z_x) = k\right] \\
&= \mathbb{E}_X\left[f(X)|X\in Z_x\right] = \bar{f}_{\lambda,n,1}(x).
\end{aligned}$$

For the bias term in (19), Proposition 1 of Arlot and Genuer (2014) implies

$$\mathbb{E}[(f(x) - \bar{f}_{\lambda,M}(x))^2] = \mathbb{E}[(f(x) - \tilde{f}_\lambda(x))^2] + \frac{\mathrm{Var}(\bar{f}_\lambda^{(1)}(x))}{M}.$$

We then have the following upper bound on the variance of $\bar{f}_\lambda^{(1)}$. For $x \in \mathcal{S}$,

$$\mathrm{Var}(\bar{f}_\lambda^{(1)}(x)) \le \mathbb{E}[(\bar{f}_\lambda^{(1)}(x) - f(x))^2] \le L^2\mathbb{E}[\mathrm{diam}(Z_x^\lambda \cap \mathcal{S})^2] \le \frac{L^2 c_{2,P_\mathcal{S}\Pi}}{\lambda^2},$$

where the last inequality follows from Lemma 2 and stationarity. For the variance term in (19), Jensen's inequality implies

$$\mathbb{E}[(\bar{f}_{\lambda,M}(x) - \hat{f}_{\lambda,n,M}(x))^2] \leq \mathbb{E}[(\bar{f}^{(1)}_{\lambda}(x) - \hat{f}_{\lambda,n,1}(x))^2].$$

Thus, taking the expectation with respect to X,

$$\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2] \leq \frac{L^2 c_{2,P_{\mathcal{S}}\Pi}}{M\lambda^2} + \mathbb{E}[(f(X) - \tilde{f}_{\lambda}(X))^2] + \mathbb{E}[(\bar{f}^{(1)}_{\lambda}(X) - \hat{f}_{\lambda,n,1}(X))^2].$$

This upper bound also holds when conditioning on $X \in K_\varepsilon$, that is,

$$\begin{aligned}\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in K_\varepsilon] &\leq \frac{L^2 c_{2,P_{\mathcal{S}}\Pi}}{M\lambda^2} \\ &+ \mathbb{E}[(f(X) - \tilde{f}_{\lambda}(X))^2 | X \in K_\varepsilon] + \mathbb{E}[(\bar{f}^{(1)}_{\lambda}(X) - \hat{f}_{\lambda,n,1}(X))^2 | X \in K_\varepsilon].\end{aligned}$$

We can then use Lemmas 15 and 4 to obtain the variance bound

$$\begin{aligned}&\mathbb{E}[(\bar{f}^{(1)}_{\lambda}(X) - \hat{f}_{\lambda,n,1}(X))^2] \\ &\quad\leq \frac{(5\|f\|_\infty^2 + 2\sigma^2)\mathrm{vol}_d(P_{\mathcal{S}}\Pi)}{n} \sum_{k=0}^{d} \binom{d}{k} \lambda^k \mathbb{E}[V(W_{\mathcal{S}}[k], Z_{\mathcal{S}}[d-k])],\end{aligned}$$

and the conditional variance satisfies

$$\begin{aligned}\mathbb{E}[(\bar{f}^{(1)}_{\lambda}(X) - \hat{f}_{\lambda,n,1}(X))^2 | X \in K_\varepsilon] &\leq \mathbb{P}(X \in K_\varepsilon)^{-1} \mathbb{E}[(\bar{f}^{(1)}_{\lambda}(X) - \hat{f}_{\lambda,n,1}(X))^2] \\ &\leq \frac{(5\|f\|_\infty^2 + 2\sigma^2)\mathrm{vol}_d(P_{\mathcal{S}}\Pi)}{n\mathbb{P}(X \in K_\varepsilon)} \sum_{k=0}^{d} \binom{d}{k} \lambda^k \mathbb{E}[V(W_{\mathcal{S}}[k], Z_{\mathcal{S}}[d-k])]. \qquad (20)\end{aligned}$$

It remains to control the remaining bias term. By Taylor's theorem, for $f \in \mathcal{C}^{1,\beta}(L)$ with $\beta \in (0, 1]$,

$$\begin{aligned}|f(z) - f(x) - \nabla f(x)^T(z-x)| &= \left| \int_0^1 [\nabla f(x + t(z-x)) - \nabla f(x)]^T (z-x) \mathrm{d}t \right| \\ &\leq \int_0^1 L(t\|z-x\|)^\beta \|z-x\| \mathrm{d}t \leq L\|z-x\|^{1+\beta}.\end{aligned}$$

Then, for $x \in \operatorname{supp}(\mu)$,

$$
\begin{aligned}
|\tilde{f}_\lambda(x) - f(x)| &= \left| \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)} \int_{Z_x^\lambda} (f(z) - f(x)) \mu(\mathrm{d}z) \right] \right| \\
&\leq \left| \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)} \int_{Z_x^\lambda} \nabla f(x)^T (z-x) \mu(\mathrm{d}z) \right] \right| \\
&\quad + \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)} \int_{Z_x^\lambda} \left| f(z) - f(x) - \nabla f(x)^T (z-x) \right| \mu(\mathrm{d}z) \right] \\
&\leq \left| \nabla f(x)^T \int_{\mathcal{S}} (z-x) \mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\mu(Z_x^\lambda)} \right] \mu(\mathrm{d}z) \right| + \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)} \int_{\mathcal{S}} L \|z-x\|^{1+\beta} 1_{\{z \in Z_x^\lambda\}} \mu(\mathrm{d}z) \right] \\
&\leq \|\nabla f(x)\| \left\| \int_{\mathcal{S}} (z-x) \mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\mu(Z_x^\lambda)} \right] \mu(\mathrm{d}z) \right\| + L \mathbb{E}\left[\operatorname{diam}(Z_x^\lambda \cap \mathcal{S})^{1+\beta} \right] \\
&\leq L \left\| \int_{\mathcal{S}} (z-x) \mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\mu(Z_x^\lambda)} \right] \mu(\mathrm{d}z) \right\| + \frac{L c_{\beta, P_{\mathcal{S}}\Pi}}{\lambda^{1+\beta}}.
\end{aligned}
$$

Up to this point, we have closely followed the proof of Theorem 3 of Mourtada et al. (2020) with more general bounds for the parameters of STIT tessellations. For the next step, recall that by the assumptions, μ has a positive and Lipschitz density p w.r.t. the Lebesgue measure on its support. To bound the first term above, Mourtada et al. (2020) compare the density $F_{\lambda,p}(z) := \mathbb{E}\left[\frac{p(z)}{\mu(Z_x^\lambda)} 1_{\{z \in Z_x^\lambda\}} \right]$ with the density $F_{\lambda,\mathrm{unif}}(z) := \mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda \cap [0,1]^D\}}}{\mathrm{vol}_D(Z_x^\lambda \cap [0,1]^D)} \right]$ (where p is the uniform density on the unit cube). We instead compare $F_{\lambda,p}$ with the density

$$
F_\lambda(z) := \mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\mathrm{vol}_d\left(Z_x^\lambda \cap \mathcal{S}\right)} \right], \quad z \in \mathcal{S},
$$

from Lemma 16. In particular, using Lemma 16, we add zero inside the norm to obtain the following upper bound

$$
\begin{aligned}
\left\| \int_{\mathcal{S}} (z-x) F_{\lambda,p}(z) \mathrm{d}z \right\| &= \left\| \int_{\mathcal{S}} (z-x) \left(F_{\lambda,p}(z) - F_\lambda(z) \right) \mathrm{d}z \right\| \\
&\leq \int_{\mathcal{S}} \|z-x\| \left| \mathbb{E}\left[\frac{p(z) 1_{\{z \in Z_x^\lambda\}}}{\mu(Z_x^\lambda)} \right] - \mathbb{E}\left[\frac{1_{\{z \in Z_x^\lambda\}}}{\mathrm{vol}_d(Z_x^\lambda \cap \mathcal{S})} \right] \right| \mathrm{d}z \\
&\leq \int_{\mathcal{S}} \|z-x\| \mathbb{E}\left[\frac{\int_{Z_x^\lambda \cap \mathcal{S}} |p(z) - p(y)| \mathrm{d}y}{\mu(Z_x^\lambda) \mathrm{vol}_d(Z_x^\lambda \cap \mathcal{S})} 1_{\{z \in Z_x^\lambda\}} \right] \mathrm{d}z \\
&\leq \mathbb{E}\left[\operatorname{diam}(Z_x^\lambda \cap \mathcal{S}) \frac{\int_{Z_x^\lambda \cap \mathcal{S}} \int_{Z_x^\lambda \cap \mathcal{S}} |p(z) - p(y)| \mathrm{d}y \mathrm{d}z}{\mu(Z_x^\lambda) \mathrm{vol}_d(Z_x^\lambda \cap \mathcal{S})} \right].
\end{aligned}
$$

Recall from the assumptions that the density p of μ has a finite Lipschitz constant $C_p > 0$ on its compact and convex d-dimensional support $K := \operatorname{supp}(\mu) \subset \mathcal{S}$ and we can define $p_0 := \min_{x \in K} p(x) > 0$ and $p_1 := \max_{x \in K} p(x) < \infty$. Also note that the integrand above is zero when $z, y \notin K$. In the following, we denote by $K^c := \mathbb{R}^d \backslash K$ the complement of K.

Thus,

$$\begin{aligned}
\left\|\int_{\mathcal{S}}(z-x)F_{\lambda,p}(z)\mathrm{d}z\right\| &\leq \mathbb{E}\left[\mathrm{diam}(Z_x^\lambda\cap\mathcal{S})\frac{\int_{Z_x^\lambda\cap\mathcal{S}}\int_{Z_x^\lambda\cap\mathcal{S}}|p(z)-p(y)|\mathrm{d}y\mathrm{d}z}{\mu(Z_x^\lambda)\mathrm{vol}_d(Z_x^\lambda\cap\mathcal{S})}\right] \\
&\leq \frac{C_p}{p_0}\mathbb{E}\left[\mathrm{diam}(Z_x^\lambda\cap\mathcal{S})\frac{\int_{Z_x^\lambda\cap K}\int_{Z_x^\lambda\cap K}\|z-y\|\mathrm{d}y\mathrm{d}z}{\mathrm{vol}_d(Z_x^\lambda\cap K)\mathrm{vol}_d(Z_x^\lambda\cap\mathcal{S})}\right] \\
&\quad+\frac{2C_pp_1}{p_0}\mathbb{E}\left[\mathrm{diam}(Z_x^\lambda\cap\mathcal{S})\frac{\int_{Z_x^\lambda\cap K}\int_{Z_x^\lambda\cap\mathcal{S}\cap K^c}\mathrm{d}y\mathrm{d}z}{\mathrm{vol}_d(Z_x^\lambda\cap K)\mathrm{vol}_d(Z_x^\lambda\cap\mathcal{S})}\right] \\
&\leq \frac{C_p}{p_0}\mathbb{E}\left[\mathrm{diam}(Z_x^\lambda\cap\mathcal{S})^2\right]+\frac{2C_pp_1}{p_0}\mathbb{E}\left[\frac{\mathrm{diam}(Z_x^\lambda\cap\mathcal{S})\mathrm{vol}_d\left(Z_x^\lambda\cap\mathcal{S}\cap K^c\right)}{\mathrm{vol}_d(Z_x^\lambda\cap\mathcal{S})}\right] \\
&\leq \frac{C_pc_{2,P_{\mathcal{S}}\Pi}}{\lambda^2p_0}+\frac{2C_pp_1}{p_0}\mathbb{E}\left[\frac{\mathrm{diam}(Z_x^\lambda\cap\mathcal{S})\mathrm{vol}_d\left(Z_x^\lambda\cap\mathcal{S}\cap K^c\right)}{\mathrm{vol}_d(Z_x^\lambda\cap\mathcal{S})}\right],
\end{aligned}$$

where the last inequality follows from Lemma 2 and stationarity. Now, using stationarity, Fact 1, and (3),

$$\frac{\mathrm{diam}(Z_x^\lambda)\mathrm{vol}_d(Z_x^\lambda\cap K^c)}{\mathrm{vol}_d(Z_x^\lambda\cap\mathcal{S})}\overset{d}{=}\frac{\mathrm{diam}(Z_{0,\mathcal{S}})\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-x))}{\lambda\mathrm{vol}_d(Z_{0,\mathcal{S}})}.$$

Thus we have the following upper bound on the conditional bias

$$\begin{aligned}
&\mathbb{E}[(\tilde{f}_\lambda(X)-f(X))^2|X\in K_\varepsilon]\leq L^2\mathbb{E}\left[\left(\left\|\int_{\mathcal{S}}(z-X)F_{\lambda,p}(z)\mathrm{d}z\right\|+\frac{c_{\beta,P_{\mathcal{S}}\Pi}}{\lambda^{1+\beta}}\right)^2\middle|X\in K_\varepsilon\right] \\
&\leq L^2\mathbb{E}\left[\left(\frac{c_{\beta,P_{\mathcal{S}}\Pi}}{\lambda^{1+\beta}}+\frac{C_pc_{2,P_{\mathcal{S}}\Pi}}{\lambda^2p_0}+\frac{2C_pp_1}{\lambda p_0}\mathbb{E}\left[\frac{\mathrm{diam}(Z_{0,\mathcal{S}})\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-x))}{\mathrm{vol}_d(Z_{0,\mathcal{S}})}\right]\right)^2\middle|X\in K_\varepsilon\right] \\
&\leq L^2\left(\frac{c_{\beta,P_{\mathcal{S}}\Pi}}{\lambda^{1+\beta}}+\frac{C_pc_{2,P_{\mathcal{S}}\Pi}}{\lambda^2p_0}\right)^2+\frac{4C_p^2p_1^2}{\lambda^2p_0^2}\mathbb{E}\left[\frac{\mathrm{diam}(Z_{0,\mathcal{S}})^2\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-X))^2}{\mathrm{vol}_d(Z_{0,\mathcal{S}})^2}\middle|X\in K_\varepsilon\right] \\
&\quad+\frac{4L^2C_pp_1}{\lambda p_0}\left(\frac{c_{\beta,P_{\mathcal{S}}\Pi}}{\lambda^{1+\beta}}+\frac{C_pc_{2,P_{\mathcal{S}}\Pi}}{\lambda^2p_0}\right)\mathbb{E}\left[\frac{\mathrm{diam}(Z_{0,\mathcal{S}})\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-X))}{\mathrm{vol}_d(Z_{0,\mathcal{S}})}\middle|X\in K_\varepsilon\right].
\end{aligned}\tag{21}$$

Conditioned on $X\in K_\varepsilon$, $\varepsilon B^d\subseteq K-X$. This implies that $\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-X))=0$ if $\mathrm{diam}(Z_{0,\mathcal{S}})\leq\lambda\varepsilon$. Then, for $k\in\{1,2\}$,

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathrm{diam}(Z_{0,\mathcal{S}})^k\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-X))^k}{\mathrm{vol}_d(Z_{0,\mathcal{S}})^k}\middle|X\in K_\varepsilon\right] \\
&\qquad\leq\mathbb{E}\left[\frac{\mathrm{diam}(Z_{0,\mathcal{S}})^k1_{\{\mathrm{diam}(Z_{0,\mathcal{S}})\geq\lambda\varepsilon\}}}{\mathrm{vol}_d(Z_{0,\mathcal{S}})^2\mathbb{P}(X\in K_\varepsilon)}\mathbb{E}_X\left[\mathrm{vol}_d(Z_{0,\mathcal{S}}\cap\lambda(K^c-X))^k\right]\right].
\end{aligned}$$

To bound the inner expectation with respect to X, we see that

$$\begin{aligned}
\mathbb{E}_X[\mathrm{vol}_d(Z_{0,\mathcal{S}} \cap \lambda(K^c - X))^k] &= \int_K p(x) \left(\int_{\mathbb{R}^d} 1_{\{y \in Z_{0,\mathcal{S}} \cap \lambda(K^c - x)\}} \right)^k \mathrm{d}x \\
&\leq p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^{k-1} \int_K \int_{\mathbb{R}^d} 1_{\{y \in Z_{0,\mathcal{S}} \cap \lambda(K^c - x)\}} \mathrm{d}y \mathrm{d}x \\
&= p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^{k-1} \int_{Z_{0,\mathcal{S}}} \int_K 1_{\{x \in K^c - \frac{y}{\lambda}\}} \mathrm{d}x \mathrm{d}y \\
&= p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^{k-1} \int_{Z_{0,\mathcal{S}}} \mathrm{vol}_d \left(K \cap K^c - \frac{y}{\lambda} \right) \mathrm{d}y \\
&= p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^{k-1} \int_{Z_{0,\mathcal{S}}} \mathrm{vol}_d \left(K \cup K - \frac{y}{\lambda} \right) \mathrm{d}y - p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^k \mathrm{vol}_d(K),
\end{aligned}$$

where we have used that $\mathrm{vol}_d(K \cap K^c - y/\lambda) = \mathrm{vol}_d(K) - \mathrm{vol}_d(K \cap K - y/\lambda)$ and $\mathrm{vol}_d(K \cap K - y/\lambda) = 2\mathrm{vol}_d(K) - \mathrm{vol}_d(K \cup K - y/\lambda)$. We now observe that the union $K \cup K - \frac{y}{\lambda}$ is a subset of the Minkowski sum $K + \frac{\|y\|}{\lambda} B^d$. By Steiner's formula (Schneider and Weil, 2008, equation 14.5),

$$\begin{aligned}
\mathrm{vol}_d \left(K \cup K - \frac{y}{\lambda} \right) \leq \mathrm{vol}_d \left(K + \frac{\|y\|}{\lambda} B^d \right) &= \sum_{j=0}^{d} \left(\frac{\|y\|}{\lambda} \right)^{d-j} \kappa_{d-j} V_j(K) \\
&= \mathrm{vol}_d(K) + \sum_{j=0}^{d-1} \left(\frac{\|y\|}{\lambda} \right)^{d-j} \kappa_{d-j} V_j(K).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}_X[\mathrm{vol}_d(Z_{0,\mathcal{S}} \cap \lambda(K^c - X))^k] &\leq p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^k \sum_{j=0}^{d-1} \left(\frac{\mathrm{diam}(Z_{0,\mathcal{S}})}{\lambda} \right)^{d-j} \kappa_{d-j} V_j(K) \\
&= 2\lambda^{-1} p_1 \mathrm{vol}_d(Z_{0,\mathcal{S}})^k \mathrm{diam}(Z_{0,\mathcal{S}}) V_{d-1}(K) + O(\lambda^{-2}).
\end{aligned}$$

Putting the above bounds together gives

$$\begin{aligned}
&\mathbb{E} \left[\frac{\mathrm{diam}(Z_{0,\mathcal{S}})^k \mathrm{vol}_d(Z_{0,\mathcal{S}} \cap \lambda(K^c - X))^k}{\mathrm{vol}_d(Z_{0,\mathcal{S}})^k} \middle| X \in K_\varepsilon \right] \\
&\qquad \leq \left(\frac{p_1 V_{d-1}(K)}{\lambda \mathbb{P}(X \in K_\varepsilon)} + O(\lambda^{-2}) \right) \mathbb{E} \left[\mathrm{diam}(Z_{0,\mathcal{S}})^{k+1} 1_{\{\mathrm{diam}(Z_{0,\mathcal{S}}) \geq \lambda\varepsilon\}} \right].
\end{aligned}$$

By Holder's inequality, Lemma 2 and (7),

$$\begin{aligned}
\mathbb{E} \left[\mathrm{diam}(Z_{0,\mathcal{S}})^{k+1} 1_{\{\mathrm{diam}(Z_{0,\mathcal{S}}) \geq \lambda\varepsilon\}} \right] &\leq \mathbb{E} \left[\mathrm{diam}(Z_{0,\mathcal{S}})^{2k+2} \right]^{1/2} \mathbb{P} \left(\mathrm{diam}(Z_{0,\mathcal{S}}) \geq \lambda\varepsilon \right)^{1/2} \\
&\leq c_{k,\varepsilon,P_{\mathcal{S}}\Pi} e^{-\lambda\varepsilon c_{P_{\mathcal{S}}\Pi}}.
\end{aligned}$$

Combining the above bounds implies

$$\begin{aligned}
&\mathbb{E}[(\tilde{f}_\lambda(X) - f(X))^2 | X \in K_\varepsilon] \\
&\leq L^2 \left(\frac{c_{\beta,P_\mathcal{S}\Pi}}{\lambda^{1+\beta}} + \frac{C_p c_{2,P_\mathcal{S}\Pi}}{\lambda^2 p_0} \right)^2 + \frac{4C_p^2 p_1^3 V_{d-1}(K)}{\lambda^3 p_0^2 \mathbb{P}(X \in K_\varepsilon)} c_{2,\varepsilon,P_\mathcal{S}\Pi} e^{-\lambda\varepsilon c_{P_\mathcal{S}\Pi}} \\
&\quad + \frac{4L^2 C_p p_1^2}{\lambda^2 p_0} \left(\frac{c_{\beta,P_\mathcal{S}\Pi}}{\lambda^{1+\beta}} + \frac{C_p c_{2,P_\mathcal{S}\Pi}}{\lambda^2 p_0} \right) \frac{V_{d-1}(K)}{\mathbb{P}(X \in K_\varepsilon)} c_{1,\varepsilon,P_\mathcal{S}\Pi} e^{-\lambda\varepsilon c_{P_\mathcal{S}\Pi}} + O(\lambda^{-4} e^{-\lambda\varepsilon c_{P_\mathcal{S}\Pi}}).
\end{aligned}$$

Next observe that $\mathbb{P}(X \in K_\varepsilon) \geq \frac{p_0 \mathrm{vol}_d(K_\varepsilon)}{\mathrm{vol}_d(K)} > 0$. The total conditional risk then satisfies

$$\begin{aligned}
&\mathbb{E}[(\hat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in K_\varepsilon] \leq \frac{L^2 c_{2,P_\mathcal{S}\Pi}}{M\lambda^2} + L^2 \left(\frac{c_{\beta,P_\mathcal{S}\Pi}}{\lambda^{1+\beta}} + \frac{C_p c_{2,P_\mathcal{S}\Pi}}{\lambda^2 p_0} \right)^2 \\
&+ \frac{4C_p^2 p_1^3 V_{d-1}(K)\mathrm{vol}_d(K)}{\lambda^3 p_0^2 p_0 \mathrm{vol}_d(K_\varepsilon)} c_{2,\varepsilon,P_\mathcal{S}\Pi} e^{-\lambda\varepsilon c_{P_\mathcal{S}\Pi}} \\
&+ \frac{4L^2 C_p p_1^2}{\lambda^2 p_0} \left(\frac{c_{\beta,P_\mathcal{S}\Pi}}{\lambda^{1+\beta}} + \frac{C_p c_{2,P_\mathcal{S}\Pi}}{\lambda^2 p_0} \right) \frac{V_{d-1}(K)\mathrm{vol}_d(K)}{p_0 \mathrm{vol}_d(K_\varepsilon)} c_{1,\varepsilon,P_\mathcal{S}\Pi} e^{-\lambda\varepsilon c_{P_\mathcal{S}\Pi}} + O\left(\lambda^{-4} e^{-\lambda\varepsilon c_{P_\mathcal{S}\Pi}}\right) \\
&+ \frac{(5\|f\|_\infty^2 + 2\sigma^2)\mathrm{vol}_d(P_\mathcal{S}\Pi)\mathrm{vol}_d(K)}{np_0 \mathrm{vol}_d(K_\varepsilon)} \sum_{k=0}^{d} \binom{d}{k} \lambda^k \mathbb{E}[V(W_\mathcal{S}[k], Z_\mathcal{S}[d-k])].
\end{aligned}$$

Extracting leading order terms, we have

$$\mathbb{E}[\hat{f}_{\lambda,n,M}(X) - f(X))^2 | X \in K_\varepsilon] \leq O\left(\frac{L^2}{\lambda^2 M} + \frac{L^2}{\lambda^{2(1+\beta)}} + \frac{\lambda^d}{n} \right).$$

Finally, letting $\lambda = \lambda_n \sim L^{2/(d+2\beta+2)} n^{1/(d+2\beta+2)}$ and $M = M_n \gtrsim \lambda_n^{2\beta}$ as $n \to \infty$ gives the rate $O\left(L^{2d/(d+2\beta+2)} n^{-(2\beta+2)/(d+2\beta+2)}\right)$ in Corollary 9. In the unconditional case,

$$\mathbb{E}[\hat{f}_{\lambda,n,M}(X) - f(X))^2] \leq O\left(\frac{L^2}{\lambda^2 M} + \frac{L^2}{\lambda^{\min\{3,2(1+\beta)\}}} + \frac{\lambda^d}{n} \right).$$

This bound gives the same rate as above when $\beta \leq 1/2$. When $\beta > 1/2$, this bound gives the suboptimal rate $O\left(L^{2d/(d+3)} n^{-3/(d+3)}\right)$ by letting $\lambda = \lambda_n \sim L^{2/(d+3)} n^{1/(d+3)}$ and $M = M_n \gtrsim \lambda_n$ as $n \to \infty$. ■

5.4 Proof of Theorem 11

By Corollary 1 of Schreiber and Thäle (2013), the typical cell of a STIT tessellation with lifetime parameter λ has the same distribution as the typical cell of a Poisson hyperplane tessellation with intensity λ and the same normalized associated zonoid, or equivalently, the same directional distribution. The distribution of the typical cell determines the distribution of the zero cell (Schneider and Weil, 2008, Theorem 10.4.1), and thus the same proof methods used in Theorems 6 and 8 can be applied in this setting, and the results follow. ■

6. Discussion and Future Work

This work expands and strengthens the theoretical basis for purely random forests and establishes stochastic geometry as a promising toolkit for analyzing regression and classification algorithms based on random partitions. In particular, we showed that a large class of random forests built from stationary hyperplane partitions with oblique splits all achieve the same minimax rates as Mondrian forests proved by Mourtada et al. (2020). We also extended these rates to depend on a notion of the intrinsic dimension of the input instead of the ambient dimension of the feature space. This work motivates many more questions at the intersection of stochastic geometry and machine learning. We outline a few future research directions here.

First, the definition of low intrinsic dimension used in our assumptions has limited applicability. However, we hope that our results and proof techniques can form a basis for future work to obtain optimal rates under more general notions of low dimensionality of the input. Additionally, as mentioned in Remark 10, one needs an adaptive way of tuning the lifetime parameter and number of trees to obtain these rates in practice since the intrinsic dimension and regularity are not known *a priori*. For adaptation to regularity, a model aggregation method was proposed by Mourtada et al. (2020) for Mondrian forests, which could potentially be extended to STIT forests.

Another open question is whether the flexibility of the directional distribution allows us to find "optimal" split directions, or directional distribution ϕ, for a given data set. In particular, a directional distribution that depends on the covariate distribution may improve performance and decrease computational costs by decreasing the complexity of the partition needed to achieve optimal rates. The flexibility of these models also motivates a study of whether, under different assumptions about the underlying function f, one can obtain convergence rates that depend on the directional distribution and show optimal rates are achieved with an appropriate choice of this distribution. In particular, we might expect that with an appropriate choice of directional distribution, STIT random forests will adapt to the same types of low dimensional structure that other purely random forest variants have been shown to adapt to under a modification of the split direction probabilities. For example, centered random forests have been shown to obtain convergence rates that depend on the sparsity level of the regression function described by the number of relevant features (Biau, 2012; Klusowski, 2021).

A third research direction concerns using the toolkit of stochastic geometry to study random forests built from random tessellations where split locations depend on the data set. For example, we may consider STIT or Poisson hyperplane tessellations associated with a non-stationary intensity measure and somehow incorporate the given data set into this measure to improve the performance of this class of random forests. In the stochastic geometry literature, some non-stationary random tessellation models have been studied. Sections 11.3 and 11.4 in Schneider and Weil (2008) collect results on non-stationary flat processes and Poisson hyperplane tessellations, many of which are due to Schneider (2003). Also of note is work by Hoffmann (2007), who studied a generalization of the associated zonoid for non-stationary Poisson hyperplane tessellations.

Acknowledgments

Ngoc Mai Tran is supported by NSF Grant DMS-2113468 and the NSF IFML 2019844 award to the University of Texas at Austin. Eliza O'Reilly is supported by NSF MSPRF Award 2002255 with additional funding from ONR Award N00014-18-1-2363.

References

Sylvain Arlot. V-fold cross-validation improved: V-fold penalization. *arXiv:0802.0566*, 2008.

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv:1407.3939*, 2014.

Gerard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095, 2012.

Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016.

Gérard Biau, Luc Devroye, and Gäbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

Rico Blaser and Piotr Fryzlewicz. Random rotation ensembles. *Journal of Machine Learning Research*, 17(1):126–151, 2016.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99 (6):323–329, 2012.

Yingying Fan Chien-Ming Chi, Patrick Vossler and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

Xuhui Fan, Bin Li, and Scott Sisson. The binary space partitioning-tree process. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1859–1867, April 2018.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Shufei Ge, Shijia Wang, Yee Whye Teh, Liangliang Wang, and Lloyd Elliott. Random tessellation forests. In *Advances in Neural Information Processing Systems 32*, pages 9571–9581. 2019.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Lars Michael Hoffmann. Intersection densities of nonstationary Poisson processes of hypersurfaces. *Advances in Applied Probability*, 39(2):307–317, 2007.

Daniel Hug and Rolf Schneider. Asymptotic shapes of large cells in random tessellations. *Geometric and Functional Analysis*, 17:156–191, 2007.

Jason Klusowski. Sharp analysis of a simple model for random forests. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 757–765, April 2021.

Jason M. Klusowski and Peter Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2022.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012. JCSS Special Issue: Cloud Computing 2011.

Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems*, pages 3140–3148, 2014.

Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. Mondrian forests for large-scale regression when uncertainty matters. In *Artificial Intelligence and Statistics*, pages 1478–1487, 2016.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016.

Bjoern H. Menze, B. Michael Kelm, Daniel N. Splitthoff, Ullrich Koethe, and Fred A. Hamprecht. On oblique random forests. In *Machine Learning and Knowledge Discovery in Databases*, pages 453–469, Berlin, Heidelberg, 2011.

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for Mondrian trees and forests. *Annals of Statistics*, 28(4):2253–2276, 2020.

Werner Nagel and Viola Weiss. Limits of sequences of stationary planar tessellations. *Advances in Applied Probability*, 35:123–138, 2003.

Werner Nagel and Viola Weiss. Crack STIT tessellations: Characterization of stationary random tessellations stable with respect to iteration. *Advances in Applied Probability*, 37: 859–883, 2005.

Eliza O'Reilly and Ngoc Mai Tran. Stochastic geometry to generalize the Mondrian process. *SIAM Journal on Mathematics of Data Science*, 4(2):531–552, 2022.

Tom Rainforth and Frank Wood. Canonical correlation forests. *arXiv:1507.05444*, 2015.

Daniel M Roy and Yee Whye Teh. The Mondrian process. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 1377–1384, 2008.

Rolf Schneider. Nonstationary Poisson hyperplanes and their induced tessellations. *Advances in Applied Probability*, 35:139–158, 2003.

Rolf Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Cambridge University Press, 2013.

Rolf Schneider and Wolfgang Weil. *Stochastic and Integral Geometry*. Probability and Its Applications. Springer-Verlag, Berlin, 2008.

Tomasz Schreiber and Christoph Thäle. Intrinsic volumes of the maximal polytope process in higher dimensional STIT tessellations. *Stochastic Processes and their Applications*, 121(5):989–1012, May 2011.

Tomasz Schreiber and Christoph Thäle. Geometry of iteration stable tessellations: Connection with Poisson hyperplanes. *Bernoulli*, 19(5A):1637–1654, 2013.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Tyler M. Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L. Patsolic, Benjamin Falk, Carey E. Priebe, Jason Yim, Randal Burns, Mauro Maggioni, and Joshua T. Vogelstein. Sparse projection oblique randomer forests. *Journal of Machine Learning Research*, 21(104):1–39, 2020.

Stefan Wager and Guenther Athey. Adaptive concentration of regression trees, with application to random forests. *arXiv:1503.06388*, 2015.

Stefan Wager and Guenther Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

Zi Wang, Clement Gehring, Pushmeet Kohli, and Stefanie Jegelka. Batched large-scale Bayesian optimization in high-dimensional spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 745–754, 2018.