arXiv:2109.10541v4 [math.ST] 28 Mar 2022

MINIMAX RATES FOR HIGH-DIMENSIONAL RANDOM TESSELLATION FORESTS

ELIZA O'REILLY AND NGOC MAI TRAN

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 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. In this work, we show that a large class of random forests with general split directions also achieve minimax rates in arbitrary dimension. This class includes STIT forests, a generalization of Mondrian forests to arbitrary split directions, as well as 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.

1. Introduction

Random forests are ensembles of randomized decision trees popularized by Breiman [7] and are widely applicable in classification and regression tasks in machine learning [11, 8]. However, statistical learning theorems for random forests are notoriously difficult to obtain in dimensions $d \geq 2$ [4]. To better understand the performance of random forests, simplified versions of the algorithm have been studied. In particular, purely random forests [5, 2] are built from trees grown independently of the data, and are much more amenable to theoretical analysis. Recently, Mourtada, Gaïffas and Scornet [19] were the first to obtain a minimax optimality theorem in arbitrary dimension for a particular class of purely random forests, namely, Mondrian forests [17, 18]. These are efficient and accurate online 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 [24]. In particular, the self-similarity of the Mondrian process is key to both the computational efficiency [17, 18, 31] and the analytical tractability of Mondrian forests [19].

One key limitation of the Mondrian forest, as well as Breiman's original random forest algorithm, is that the splits are axis-aligned. While this is computationally efficient, axis-aligned splits are unable to capture dependencies between features. In practice, this puts a large burden on the feature selection and representation step. Another limitation of the results of [19] is that their rates suffer from the curse of dimensionality. That is, the convergence rates depend on the ambient dimension of data and become uninformative in the presence of a high dimensional feature space, even though empirically random forests perform well in such regimes [8, 11, 30].

Attempts have been made to bridge these issues on both theoretical and practical fronts, with focus on random forests with sparse features, where locally, the target function to be learned depends on a small subset of features. In Breiman's orginal paper [7], he proposed a variant called Forest-RC that split data using linear combinations of features and showed improved empirical

2000 *Mathematics Subject Classification.* Primary 60D05; Secondary 62G07.

performance of this method. Many other models of random forests with oblique cuts have subsequently been proposed [6, 10, 12, 23, 30]. In particular, to address the computational trade-off, [30] studied random forests with oblique splits constructed with sparse projections. They demonstrated better empirical performance while retaining computational efficiency. On the theoretical front, Biau [3] studied a purely random forest variant called centered random forests and obtained (sub-optimal) convergence rates that depend on the sparsity level described by number of relevant features. Klusowksi [16] obtained an improved, but still sub-optimal, rate for this class of random forests depending on the sparsity level and showed this rate could not be improved in general. As of current, there is a lack of a comprehensive technique that can handle random forests with arbitrary split directions and that can yield optimal learning rates that adapt to a notion of sparsity.

This paper, to the best of our knowledge, gives the first results on minimax optimality for a large class of purely random forests that is defined for all dimensions with *general* split directions. In particular, we show that STIT forests, a significant generalization of Mondrian forests, also attain the minimax rates discovered in [19] under less technical assumptions. Furthermore, we incorporate an assumption of sparsity on the input data, which gives improved rates in high dimensional feature space.

The general class of STIT random forests are derived from the stable under iteration (STIT) processes of Nagel and Weiss [20, 21]. These stochastic processes generate the most general class of self-similar and stationary random partitions of $\mathbb{R}^d$ by hyperplane cuts possible [21]. Self-similarity in particular ensures that STIT random forests also enjoy the online construction that underpins popularity of the Mondrian forest in practice. The family of STIT processes is indexed by probability distributions on the unit sphere, denoted by ϕ, describing the distribution of directions of the hyperplane cuts in the random partition. The Mondrian process corresponds to the case where ϕ is the discrete uniform measure on the d coordinate vectors. Subsequent generalizations of the Mondrian process to oblique cuts [10, 12] are also special cases of STIT processes. The freedom in the choice of the directional distribution ϕ for the splits brings more flexibility in machine learning applications. For example, while the Mondrian process can only be used to approximate the Laplace kernel [18], the STIT process produces random features that approximate a much wider class of kernels, as characterized in [22]. Improved empirical performance of STIT forests built from STIT processes with a uniform directional distribution over Mondrian forests was also shown in [12] through a classification task and simulation study.

A major contribution of our work lies in the proof technique, as our approach relies on theorems in stochastic geometry that up until now have not been utilized in statistical learning theory. Going from the Mondrian to the STIT forests, the key difficulty is the geometry of the cells of the tessellations. The Mondrian process generates axis-aligned rectangular cells and the distribution of the cell of a given input can be characterized exactly from the construction. In contrast, the cells that STIT processes generate can be more complex random polytopes. A stochastic geometry viewpoint enables us to crucially exploit the self-similarity and stationarity of the underlying random partition, and thus to handle more general cell geometries. In particular, we obtain risk bounds on STIT forest estimators that depend on the distribution of a single random polytope, called the *typical cell*, defined using the corresponding stationary STIT tessellation on $\mathbb{R}^d$. Our analysis simplifies previous approaches for analyzing Mondrian forests as well, since we do not encounter the boundary issue that appears in Theorem 3 of [19]. That is, we achieve minimax rates for Mondrian forests *without* any conditioning for a larger class of functions.

Specifically, our first main result, Theorem 7, gives an upper bound on the quadratic risk of the STIT forest estimator for the regression of β-Hölder continuous functions for $\beta \in (0, 1]$, generalizing Theorem 3 of [19]. The theorem shows that *any* STIT forest with optimally tuned complexity parameter achieves the minimax rate for this function class. The directional distribution ϕ comes in

through the constant terms in the upper bound. Our proof method gives *geometric* interpretations to these constants in terms of moments of the diameter of the cell Z_0 containing the origin and the expected mixed volumes between the support of the input and the typical cell Z. *Because* of this explicit geometry, we were able to get the rates in terms of the sparsity dimension of the input. In the special case of the Mondrian forest, Z is the Minkowski sum of i.i.d. centered line segments parallel to the axes with exponential length. Taking the support to be $[0,1]^d$ recovers Theorem 2 in [19], see Example 5 for details. Similarly, our second main result, Theorem 9, extends Theorem 3 in [19], where additional smoothness assumptions are made on f. It says 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 in Theorem 3 of [19], an improved rate for STIT forests over STIT trees in this case is due to large enough forests having a smaller bias than single trees for smooth regression functions.

Our proof technique also takes us *beyond* the class of STIT forests. Any random partition can be used to define a tree estimator as in (7) and subsequently a random forest estimator (8). Since our rate upper bounds 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 by stationary random hyperplane partitions of $\mathbb{R}^d$. Our last main result, Theorem 11, is a demonstration of 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 ϕ and complexity parameter, and thus is also minimax optimal.

Random forests built from stationary random tessellations, of which STIT forests are a special case, are necessarily data agnostic. However, Breiman's original algorithm splits the feature space using a criterion that depends on the training data set. This dependence underpins the difficulty in the analysis, motivating the study of purely random forests. Recently, [9] obtained the first consistency rates for Breiman's original algorithm in arbitrary dimension, but theoretical understanding remains incomplete. In Section 4, we briefly discuss how one could approach the analysis of data-dependent random forest variants via non-stationary Poisson hyperplane and STIT processes. We conclude with a few concrete open problems at the intersection of stochastic geometry and statistical learning theory, and hope that our work will fuel further investigations in this interdisciplinary area.

Organization. Section 2 collects background on STIT processes, Poisson hyperplane processes, and essential results in stochastic geometry needed for our proofs. Section 3 states and proves our three main results, Theorems 7, 9 and 11. Section 4 concludes with discussions and open problems.

Acknowledgements. 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.

2. Preliminaries

2.1. Stochastic geometry background.

We recall the key concepts from stochastic geometry needed for our paper, which are *random tessellations, stationarity*, the *zero cell*, and the *typical cell*. For additional background, we recommend [27, Chapter 10].

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 take the view of the tessellation as the collection of cells, but will also discuss the union of cell boundaries to establish relevant definitions. Formally, we define a *random tessellation* as the 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 sets which satisfy:

- for all compact $K \subset \mathbb{R}^d$, a finite number of C_i's have non-empty intersection with K;
- for all $i \neq j$, $\text{int}(C_i) \cap \text{int}(C_j) = \emptyset$;
- $\cup_{i\in\mathbb{Z}} C_i = \mathbb{R}^d$;
- for all i, $\text{vol}_d(\partial C_i) = 0$.

A random tessellation is *stationary* if its distribution is invariant under translations. That is, for all $x \in \mathbb{R}^d$, $\{C_i + x\}_{i\in\mathbb{Z}} \stackrel{d}{=} \{C_i\}_{i\in\mathbb{Z}}$, where '$\stackrel{d}{=}$' denotes equality in distribution. 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 $Z_0 \stackrel{d}{=} Z_x - x$.

An important random object related to a stationary random tessellation $\mathcal{P}$ is the *typical cell*. To define this, first consider 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 [27, 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 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.

2.2. **STIT Tessellations.** For a random tessellation $\mathcal{P}$, we will denote by $\mathcal{Y}$ the union of cell boundaries, which forms a $d-1$-surface process in $\mathbb{R}^d$, see [27, Section 4.5]. For any stationary surface process, 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.

For two random tessellations $\mathcal{P}_1$ and $\mathcal{P}_2$, denote by $\mathcal{Y}_1$ and $\mathcal{Y}_2$ the cell boundaries of $\mathcal{P}_1$ and $\mathcal{P}_2$, 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 $n \in \mathbb{N}$,

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

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 $\mathcal{Y}(\lambda)$.

A consequence of property (iii) is that STIT processes have the scaling property, that is, for all $\lambda > 0$, $\mathcal{Y}(1) \stackrel{d}{=} \lambda\mathcal{Y}(\lambda)$ [21, Lemma 5]. Intuitively, this property says that one can swap time for

space. Specifically, if one fixes a compact observation window $W \subset \mathbb{R}^d$, then $\{\mathcal{Y}(\lambda) \cap W : \lambda > 0\}$ is a random partition process on 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 $\mathcal{Y}(\lambda) \mapsto \frac{1}{2}\mathcal{Y}(\lambda)$ to see more of the partition process $\mathcal{Y}(\lambda)$, or one can run the partition process for twice as long by mapping $\mathcal{Y}(\lambda) \mapsto \mathcal{Y}(2\lambda)$. The scaling property says that these two operations give the same random tessellation on W in distribution.

For the zero cell Z_0^λ and typical cell Z_λ of $\mathcal{P}(\lambda)$, the random tessellation induced by the boundaries $\mathcal{Y}(\lambda)$, 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 in [21] that the STIT partition process $\{\mathcal{Y}(\lambda) \cap W : \lambda > 0\}$ restricted to a fixed compact and convex window $W \subset \mathbb{R}^d$ can always be constructed by drawing random hyperplane cuts from a fixed distribution at exponential times. A special case of this construction was rediscovered by [24], 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 $\mathcal{H}^d$, defined by

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

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

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

Now, fix $\lambda > 0$ 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

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

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 lifetime expires.

Theorem 1 in [21] shows the existence of a STIT tessellation process $\mathcal{Y}(\lambda)$ on $\mathbb{R}^d$ such that $\mathcal{Y}(\lambda) \cap W \stackrel{d}{=} \mathcal{Y}(\lambda, W, \phi)$. Conversely, for any stationary tessellation $\mathcal{Y}$ in $\mathbb{R}^d$ satisfying (2) with directional distribution ϕ, Corollary 2 in [21] showed that there exists $\lambda > 0$ such that $\mathcal{Y} \stackrel{d}{=} \mathcal{Y}(\lambda)$, and $\mathcal{Y}(\lambda) \cap W \stackrel{d}{=} \mathcal{Y}(\lambda, W, \phi)$ 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 tessellations with the hierarchical construction that underpins the computational efficiency of the Mondrian process [24, 17, 18].

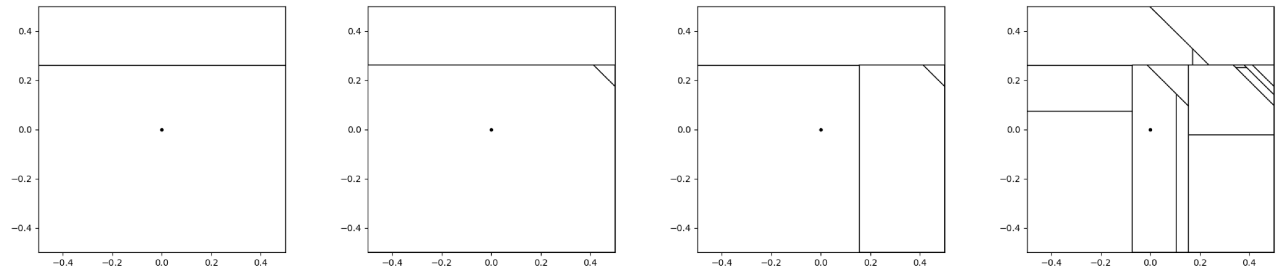


FIGURE 1. An example STIT process with three directions. Each cell W has an independent exponential clock with mean $\Lambda([W])$. When the clock rings, the cell is cut by a hyperplane drawn from the specified distribution Λ conditioned to hit this cell. In this simulation, at time $t = 0$ we started with the unit square $W = [-0.5, 0.5]^2$, and ran until time $t = 9$, called the lifetime of the STIT. The first three figures show the first three cuts, while the last figure shows the STIT at time $t = 9$, which has 14 cuts.

Example 1 (The Mondrian as a special case of STIT)**.** If ϕ is the uniform distribution over the positive and negative standard basis vectors, then the resulting STIT process $\mathcal{Y}(\lambda, W, \phi)$ is the Mondrian process [24]. See Figure 2(B) for a simulation.

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

2.3. **Poisson Hyperplane Tessellations.** We now describe another way to obtain stationary random tessellations of $\mathbb{R}^d$, and compare it to the STIT tessellation. 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 ϕ an even probability measure on $\mathbb{S}^{d-1}$ called the spherical directional distribution, see [27, Chapter 4.4]. To sample from X on a compact window W:

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

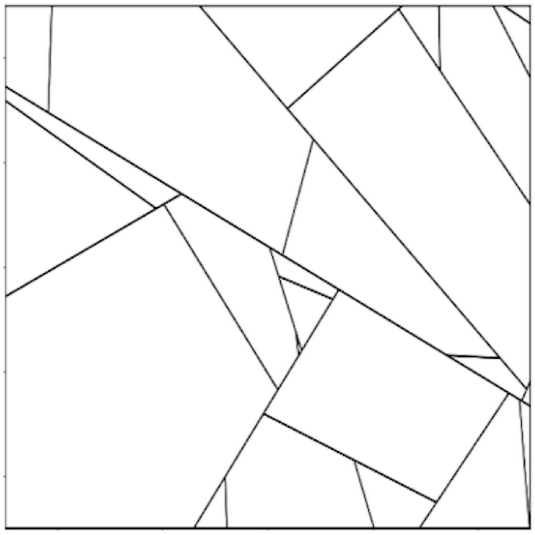
$$\Theta([W]) = \lambda \int_{\mathbb{R}} \int_{\mathbb{S}^{d-1}} 1_{\{H(u,t)\cap W\neq\emptyset\}} \mathrm{d}\phi(u)\mathrm{d}t.$$

(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

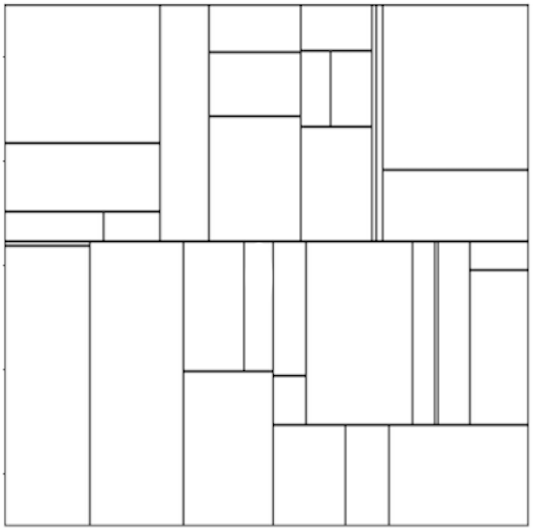
$$\mathrm{d}\Phi(u) := \frac{h(W,u) + h(W,-u)}{\Lambda([W])} \mathrm{d}\phi(u),$$

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

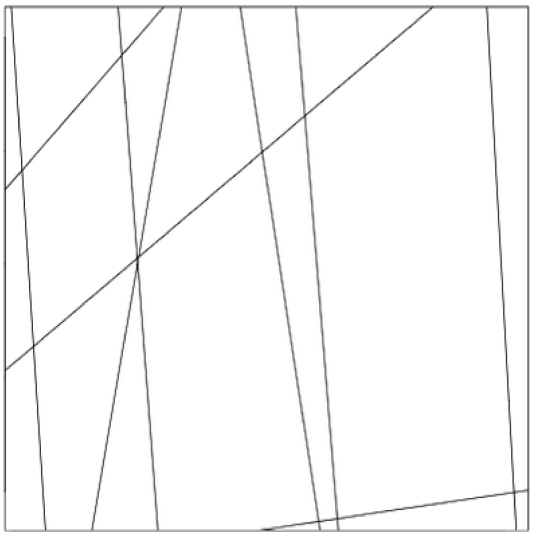
The Poisson hyperplane process induces a tessellation that is globally different that the STIT tessellation (cf. Figure 2). In particular, a Poisson hyperplane tessellation is face-to-face, meaning that the intersection of two cells is either non-empty, or is a face of both cells. This is not the case for a STIT tessellation, where, for example, a vertex of a cell can be an interior point of the



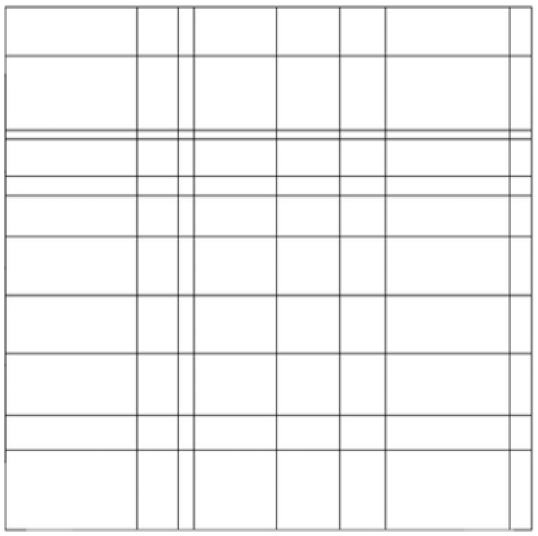
(A) Isotropic STIT



(B) Axis-aligned STIT (aka Mondrian)



(C) Isotropic Poisson hyperplane process



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

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 pair (A,C) (respectively (B,D)) are globally different tessellations, it was observed in [21] (see also Corollary 1 in [29]) 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 the random forests based on (A) and (C) (respectively (B) and (D)) have the same minimax rate.

facet of a neighbor cell. However, the typical cell of a STIT tessellation with lifetime parameter λ and directional distribution ϕ has the same distribution as the typical cell of a Poisson hyperplane tessellation with intensity λ and the same directional distribution [29, Corollary 1]. By Theorem 10.4.1 in [27], the typical cell determines the distribution of the zero cell, and so the zero cells of

a STIT tessellation and a Poisson hyperplane tessellation with corresponding parameters are also equal in distribution.

2.4. **Parameters of STIT and Poisson Hyperplane Tessellation Cells.** In this section we generalize the results in Section 4 of [19] on the diameter of the zero cell and the number of cells hitting a compact and convex domain. In combination, these observations show that STIT processes and 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 associated to a STIT or Poisson hyperplane process called the *associated zonoid*. If the process has directional distribution ϕ and lifetime/intensity λ, this is defined as the convex body Π_λ in $\mathbb{R}^d$ with support function

$$h(\Pi_\lambda, v) = \frac{\lambda}{2}\Lambda([[0,v]]) = \frac{\lambda}{2}\int_{\mathbb{S}^{d-1}} |\langle u, v\rangle| \mathrm{d}\phi(u), \quad v \in \mathbb{S}^{d-1}, \tag{4}$$

see [27, p.156]. We will denote by Π the associated zonoid of the process for lifetime/intensity $\lambda = 1$. In particular, note that $\Pi_\lambda = \lambda\Pi$. We also recall from [27, (10.4) and (10.44)] that

$$\mathbb{E}[\mathrm{vol}_d(Z)] = \frac{1}{\mathrm{vol}_d(\Pi)}. \tag{5}$$

For the Mondrian (see Example 1) or axis-aligned Poisson hyperplane process, $h(\Pi, v) = \frac{\|v\|_1}{2d}$, and so the associated zonoid is the ℓ^∞ ball $\Pi = \{x \in \mathbb{R}^d : \|x\|_\infty \leq \frac{1}{2d}\}$. For the isotropic STIT (see Example 2) or isotropic Poisson hyperplane process, $h(\Pi, v) = c_d\|v\|_2$, where $c_d := \frac{\Gamma\left(\frac{d}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{d+1}{2}\right)}$ and so the associated zonoid Π is an ℓ^2 ball centered at the origin with radius c_d.

2.4.1. *Diameter of cells.* The precise distribution of the diameter of the zero cell for a general directional distribution remains an open question in stochastic geometry. However, we will show an upper bound on the moments, which will be sufficient for proving the results in this paper.

Lemma 3. *Let Z_0^λ denote the zero cell of a STIT tessellation/Poisson hyperplane tessellation with lifetime parameter/intensity λ. Then, for all $k > 0$,*

$$\mathbb{E}[\mathrm{diam}(Z_0^\lambda)^k] \leq c_{k,\Pi}\lambda^{-k}.$$

Proof. In section 8 of [15], it is shown that for fixed $r > 0$ and $\tau \in (0,1)$, there exists a constant $c_r = c_r(\tau, \Pi)$ such that for all $a > r$,

$$\mathbb{P}(\mathrm{diam}(Z_0) \geq a) \leq c_r e^{-\tau a h_{\min}(\Pi)}, \tag{6}$$

where $h_{\min}(\Pi) := \min_{u\in\mathbb{S}^{d-1}} h(\Pi, u)$. Then, the expectation satisfies

$$\begin{aligned}
\mathbb{E}[\mathrm{diam}(Z_0)^k] &= \int_0^\infty kt^{k-1}\mathbb{P}(\mathrm{diam}(Z_0) \geq t)\mathrm{d}t \\
&\leq \int_0^r kt^{k-1}\mathrm{d}t + kc_r\int_r^\infty t^{k-1}e^{-\tau t h_{\min}(\Pi)}\mathrm{d}t \\
&= r^k + \frac{kc_r}{(\tau h_{\min}(\Pi))^k}\int_0^\infty y^{k-1}e^{-y}\mathrm{d}y \leq r^k + \frac{c_r\Gamma(k+1)}{\tau^k h_{\min}(\Pi)^k}.
\end{aligned}$$

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

$$\mathbb{E}[\mathrm{diam}(Z_0)^k] \leq \frac{2(1+c_r)\Gamma(k+1)}{h_{\min}(\Pi)^k},$$

and by (3),

$$\mathbb{E}[\mathrm{diam}(Z_0^\lambda)^k] = \frac{1}{\lambda^k}\mathbb{E}[\mathrm{diam}(Z_0)^k] \leq \frac{c_{k,\Pi}}{\lambda^k},$$

where $c_{k,\Pi} := \frac{2(1+c_r)\Gamma(k+1)}{h_{\min}(\Pi)^k}$ for the chosen $r = r(k,\Pi)$. □

Remark 1. For the isotropic model as in Example 2, the associated zonoid Π is a ball, and $h_{\min}(\Pi) = c_d$. In fact, $h_{\min}(\Pi)$ is maximal for the isotropic distribution since for any other ϕ and Π defined by (4), 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 (6) in the proof of the above Lemma is maximized by the isotropic directional distribution.

2.4.2. *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 W follows from equation (1).

Lemma 4. *Let $W \subset \mathbb{R}^d$ be a compact and convex set. Let $N_\lambda(W)$ be the number of cells of a STIT tessellation $\mathcal{P}(\lambda)$ in $\mathbb{R}^d$ that intersect W. Let Π be the associated zonoid of $\mathcal{P}(1)$. Then,*

$$\mathbb{E}[N_\lambda(W)] = \mathrm{vol}_d(\Pi)\sum_{k=0}^{d}\binom{d}{k}\lambda^k\mathbb{E}[V(W[k], Z[s-k])],$$

where $\mathbb{E}[V(W[k], Z[d-k])] := \mathbb{E}[V(\underbrace{W, \ldots, W}_{k}, \underbrace{Z, \ldots, Z}_{d-k})]$.

The mixed volume $V(K_1, \ldots, K_d)$ of convex bodies $K_1, \ldots, K_d$ is non-negative, translation-invariant, and multilinear and symmetric in its arguments [26, Section 5.1]. For $k \in \mathbb{N}$, let B^k denote the unit ball in $\mathbb{R}^k$, and define $\kappa_k := \mathrm{vol}_k(B^k)$. Then, 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. $V_d = \mathrm{vol}_d$.

Example 5. (1) If $W = RB^d$, 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 [27], $\mathbb{E}V_{d-k}(Z) = \frac{V_k(\Pi)}{\mathrm{vol}_d(\Pi)}$. Thus,

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

(2) For the isotropic STIT (see Example 2), Proposition 3 in [28] gives

$$\mathbb{E}[N_\lambda(W)] = \sum_{k=0}^{d}(\Pi_{j=1}^k\gamma_j)\frac{\lambda^k}{k!}V_k(W),$$

where the constant γ_j is defined as

$$\gamma_j := \frac{\Gamma(\frac{j+1}{2})\Gamma(\frac{d}{2})}{\Gamma(\frac{j}{2})\Gamma(\frac{d+1}{2})}.$$

(3) If $W = [0,1]^d$, and Z is the typical cell of a STIT with directional distribution $\phi = \frac{1}{2d}\sum_{i=1}^d(\delta_{e_i}+\delta_{-e_i})$ with lifetime parameter d (this is the case of the Mondrian in [19]), 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. From the formula for mixed volumes of zonoids on page 614 of [27],

$$V(W[k], Z[d-k]) \stackrel{d}{=} \prod_{i=1}^{d-j} T_i,$$

and $\mathbb{E}[V(W[k], Z[d-k])] = 1$. Thus, we recover Proposition 2 in [19] that

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

Proof. Using (1) with the indicator function $f(\cdot) = 1_{\{\cdot\cap W\neq\emptyset\}}$, (3), and (5) give that the expected number of cells of $\mathcal{P}(\lambda)$ intersecting $W\subset\mathbb{R}^d$ satisfies

$$\mathbb{E}[N_\lambda(W)] = \mathbb{E}\left[\sum_{C\in\mathcal{P}(\lambda)} 1_{\{C\cap W\neq\emptyset\}}\right] = \frac{1}{\mathbb{E}[\mathrm{vol}_d(Z_\lambda)]}\mathbb{E}\left[\int_{\mathbb{R}^d} 1_{\{Z_\lambda+y\cap W\neq\emptyset\}}\mathrm{d}y\right]$$

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

The last equality appears in [27, (5.16)]. The third equality follows from the fact that $Z_\lambda+y\cap W\neq\emptyset$ if and only if $y\in W-Z_\lambda$. By the scaling property of mixed volumes, (3), and the fact that $Z\stackrel{d}{=}-Z$,

$$\mathbb{E}[V(W[k],-Z_\lambda[d-k])] = \lambda^{-(d-k)}\mathbb{E}[V(W[k],Z[d-k])].$$

Thus,

$$\mathbb{E}[N_\lambda(W)] = \mathrm{vol}_d(\Pi)\sum_{k=0}^d\binom{d}{k}\lambda^k\mathbb{E}[V(W[k],Z[d-k])],$$

which proves the claim. □

3. Main results

3.1. **Statements of the main results.** Fix a compact and convex subset $W\subset\mathbb{R}^d$, and consider the following regression setting. The data set $\mathcal{D}_n := \{(X_1,Y_1),\dots,(X_n,Y_n)\}$ 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 partition of W. The regression tree estimator based on $\mathcal{P}$ is

$$\hat{f}_n(x,\mathcal{P}) := \sum_{i=1}^n \frac{1_{\{X_i\in Z_x\}}}{\mathcal{N}_n(x)}Y_i, \tag{7}$$

where Z_x is the cell of the partition $\mathcal{P}$ that contains x and $\mathcal{N}_n(x)$ 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.

$$\hat{f}_{n,M}(x) := \frac{1}{M}\sum_{m=1}^{M} \hat{f}_n(x, \mathcal{P}_m), \tag{8}$$

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 (7) and (8) respectively, where the random partition $\mathcal{P} := \mathcal{P}(\lambda) \cap W$ is the partition of W generated by a STIT tessellation with lifetime parameter λ and 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 $x, y \in W$. We assume here that W is a compact and convex d-dimensional domain $W \subset \mathbb{R}^d$. The minimax rate for the class $\mathcal{C}^{k,\beta}(L)$ is $n^{-2(k+\beta)/(2(k+\beta)+d)}$ [13, Theorem 3.2]. Our main results show that for an appropriate choice of λ, STIT forest estimators achieve the minimax rate of convergence for $\mathcal{C}^{0,\beta}(L)$ and $\mathcal{C}^{1,\beta}(L)$. Our rate is stated in terms of the intrinsic dimensionality of the input data, defined as follows.

Definition 6. The input X is *s-sparse* if its distribution μ is supported on the intersection of a compact and convex window $W \subset \mathbb{R}^d$ with an s-dimensional linear subspace S of $\mathbb{R}^d$.

Theorem 7. *Assume X is s-sparse and $f \in \mathcal{C}^{0,\beta}(L)$ for $\beta \in (0, 1]$ and $L > 0$. Then,*

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

where $W_{\mathcal{S}} := W \cap \mathcal{S}$.

Corollary 8. *In the setting of Theorem 7, as $n \to \infty$, letting $\lambda_n = L^{2/(s+2\beta)}n^{1/(s+2\beta)}$ yields*

$$R(\hat{f}_{\lambda,n,M}) = O\left(L^{2s/(s+2\beta)}n^{-2\beta/(s+2\beta)}\right), \tag{9}$$

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

Theorem 9. *Assume X is s-sparse and the distribution μ of X has a positive and Lipschitz density with respect to the Lebesgue measure on its support. Assume $f \in \mathcal{C}^{1,\beta}(L)$ for $\beta \in (0, 1]$ and $L > 0$. Then,*

$$R(\hat{f}_{\lambda,n,M}) \leq O\left(\frac{L^2}{\lambda^2 M} + \frac{L^2}{\lambda^{2\beta+2}} + \frac{\lambda^s}{n}\right).$$

Corollary 10. *In the setting of Theorem 9, choosing*

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

implies

$$R(\hat{f}_{\lambda,n,M}) = O(L^{2s/(s+2\beta+2)}n^{-(2\beta+2)/(s+2\beta+2)}), \tag{10}$$

which is the minimax rate for the class $\mathcal{C}^{1,\beta}(L)$ on $\mathbb{R}^s$.

Remark 2. Specialized to the case of the Mondrian, our rates are an improvement over the results of [19], where no notion of sparsity was considered. However, note that the rates are obtained through optimal choices of λ and M that depend on s and β. In practice, the sparsity dimension of the input and the regularity of f are not known *a priori*. An important open problem is 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 on the function f.

Remark 3. While the rates in Corollaries 8 and 10 depend only on the sparsity dimension, the ambient dimension does appear in the upper bounds of Theorems 7 and 9 through the geometric constants. We leave for future work a thorough study of how an "optimal" choice of directional distribution could be obtained via the geometric properties of the associated zonoid and the data set to improve the constants.

These rates follow from the following bias-variance decomposition of the risk of a tree estimator that is presented in [2]. 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:

$$R(\hat{f}_{\lambda,n}) := \mathbb{E}[(f(X) - \hat{f}_{\lambda,n}(X))^2] = \mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] + \mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2]. \tag{11}$$

The first term measures how far f is away 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 [19], 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 the diameter of the zero cell, and the variance is controlled by the expected number of cells that have non-empty intersection with the support of μ. Lemmas 3 and 4 provide the needed bounds, and choosing an optimal λ depending on the number of samples n and Lipschitz constant L gives the results.

Note that our proof technique only relies on statistics of the typical cell and the zero cell of the STIT tessellation $\mathcal{P}(\lambda)$. Since these statistics are identical for the STIT and the Poisson hyperplane process, it follows that regression estimators based on 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 using M i.i.d. random partitions induced by a Poisson hyperplane process with intensity λ and spherical directional distribution ϕ. Then, $\hat{f}_{\lambda,n,M}$ is consistent and the upper bounds* (9) *and* (10) *on the risk hold in each corresponding setting.*

3.2. **Variance Bound.** In the following, we see that we can control the variance term with the expected number of cells that intersect the support of μ, as in [19].

Lemma 12. *Let $N_\lambda(K)$ be the number of cells of $\mathcal{P}(\lambda)$ that have non-empty intersection with a compact 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 same ideas of Proposition 2 in [2] which relies crucially on Proposition 1 in [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 partition. Note that the assumption that $\mathcal{D}_n$ and $\mathcal{P}(\lambda)$ are independent allow us to take these expectations separately. 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

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

A closer look at the conditional expectation gives, by the assumptions on the noise,

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

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

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

Thus,

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

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

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

and by Lemma 4.1 in [13], $\mathbb{E}[B]\mathbb{E}[B^{-1}1_{\{B>0\}}] \leq \frac{2np_{\lambda,C}}{(n+1)p_{\lambda,C}} \leq 2$. Also, the upper bounds $1-x \leq e^{-x}$ and $xe^{-x} \leq e^{-1}$ for all $x \geq 0$ imply

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

Thus,

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

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

3.3. **Proof of Theorem 7.**

Proof of Theorem 7. As in the proof of Theorem 2 in [19], first use Jensen's inequality to reduce to the error of a single Mondrian tree estimator $\hat{f}_{\lambda,n} := \hat{f}_{\lambda,n,1}$. Then, using the bias-variance decomposition (11), we have

$$R(\hat{f}_{\lambda,n}) = \mathbb{E}[(f(X) - \hat{f}_{\lambda,n}(X))^2] = \mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] + \mathbb{E}[(\bar{f}(X) - \hat{f}_{\lambda,n}(X))^2]. \tag{12}$$

We first consider the bias term. For $x \in W$, by the assumption on f,

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

Then by Lemma 3,

$$\mathbb{E}[(f(X) - \bar{f}_\lambda(X))^2] \leq L\mathbb{E}[\text{diam}(Z_x^\lambda)^{2\beta}] \leq \frac{Lc_{\beta,\Pi}}{\lambda^{2\beta}}. \tag{13}$$

Recall that the assumption X is s-sparse means there exists an s-dimensional linear subspace $\mathcal{S} \subseteq \mathbb{R}^d$ such that $\text{supp}(\mu) = W \cap \mathcal{S}$. For the variance bound, Lemma 12 implies

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

Let $W_\mathcal{S} := W \cap \mathcal{S}$. By Lemma 4 and the fact that for $k > s$, $\mathbb{E}[V(W_\mathcal{S}[k], Z[d-k])] = 0$, we have for each $i = 1, \dots K$,

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

Then,

$$\mathbb{E}[(\bar{f}_\lambda(X) - \hat{f}_{\lambda,n}(X))^2] \leq \frac{(5\|f\|_\infty^2 + 2\sigma^2)\text{vol}_d(\Pi)}{n} \sum_{k=0}^{s} \binom{d}{k} \lambda^k \mathbb{E}[V(W_\mathcal{S}[k], Z[d-k])]. \tag{14}$$

Combining equations (13) and (14) gives the first claim. The right hand side above is of order $O\left(\frac{\lambda^s}{n}\right)$, and letting $\lambda = L^{2/(s+2\beta)}n^{1/(s+2\beta)}$ gives the second claim. □

3.4. **Proof of Theorem 9.** We first need the following technical lemma.

Lemma 13. *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 s. Then,*

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

Proof. First, we observe the intersection $\mathcal{P}(\lambda) \cap \mathcal{S}$ of $\mathcal{P}(\lambda)$ with a linear subspace is a STIT tessellation that is stationary with respect to $\mathcal{S}$. By stationarity and a change of variable, for $x \in \mathcal{S}$,

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

Then by (1),

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

By the fact that volume is translation invariant,

$$\mathbb{E}\left[\frac{\mathrm{vol}_d(Z_\lambda \cap Z_\lambda + y)}{\mathrm{vol}_s(Z_0^\lambda \cap \mathcal{S})}\right] = \mathbb{E}\left[\frac{\mathrm{vol}_d(Z_\lambda - y \cap Z_\lambda)}{\mathrm{vol}_s(Z_0^\lambda \cap \mathcal{S})}\right].$$

The integrand $y\mathbb{E}\left[\frac{\mathrm{vol}_d(Z_\lambda - y)}{\mathrm{vol}_s(Z_0^\lambda \cap \mathcal{S})}\right]$ is an odd function, and thus the integral is zero. □

Proof of Theorem 9. We begin the proof in the same way as the proof of Theorem 3 in [19]. For each m, define

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

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

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

The bias-variance decomposition for the risk of a tree estimator can be extended to the random forest estimator as in [2, Equation (1)]:

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

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,

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

Also, by Proposition 1 in [2],

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

We then have the following upper bound on the variance:

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

where the last inequality follows from Lemma 3 and stationarity. Also, by Jensen's inequality,

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

Thus, taking the expectation with respect to X,

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

We can use Lemma 12 to bound the second term on the right hand side, so it remains to control the bias term. By Taylor's theorem, for $f \in \mathcal{C}^{1,\beta}(L)$ with $\beta \in (0,1]$,

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

Then,

$$\begin{aligned}
&|\tilde{f}_\lambda(x) - f(x)| = \left|\mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}\int_{Z_x^\lambda}(f(z) - f(x))\mu(\mathrm{d}z)\right]\right| \\
&\le \left|\mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}\int_{Z_x^\lambda}\nabla f(x)^T(z-x)\mu(\mathrm{d}z)\right]\right| + \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}\int_{Z_x^\lambda}\left|f(z) - f(x) - \nabla f(x)^T(z-x)\right|\mu(\mathrm{d}z)\right] \\
&\le \left|\nabla f(x)^T\int_{Z_x^\lambda}(z-x)\mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}1_{\{z\in Z_x^\lambda\}}\right]\mu(\mathrm{d}z)\right| + \mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}\int_{\mathbb{R}^d} L\|z-x\|^{1+\beta}1_{\{z\in Z_x^\lambda\}}\mu(\mathrm{d}z)\right] \\
&\le \|\nabla f(x)\|\left\|\int_{Z_x^\lambda}(z-x)\mathbb{E}\left[\frac{1}{\mu(Z_x^\lambda)}1_{\{z\in Z_x^\lambda\}}\right]\mu(\mathrm{d}z)\right\| + L\mathbb{E}\left[\mathrm{diam}(Z_x^\lambda)^{1+\beta}\right] \\
&\le L\left\|\int_{\mathbb{R}^d}(z-x)\mathbb{E}\left[\frac{1_{\{z\in Z_x^\lambda\}}}{\mu(Z_x^\lambda)}\right]\mu(\mathrm{d}z)\right\| + \frac{Lc_{\beta,\Pi}}{\lambda^{1+\beta}}.
\end{aligned}$$

Up to this point, the proof has closely followed that of Theorem 3 in [19], with more general bounds for the parameters of STIT tessellations. Now, by the assumptions, there exists a constant $C_p > 0$ such that μ has a positive and C_p-Lipschitz density p w.r.t. the Lebesgue measure on its support. By the assumption that X is s-sparse, $\mathrm{supp}(\mu) = W \cap \mathcal{S}$, where $\mathcal{S}$ is an s-dimensional linear subspace. To bound the first term above, the authors of [19] 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 $F_{\lambda,unif}(z) := \mathbb{E}\left[\frac{1_{\{z\in Z_x^\lambda\cap[0,1]^d\}}}{\mathrm{vol}_d(Z_x^\lambda\cap[0,1]^d)}\right]$ (where p is the uniform density on the unit cube). They then apply their Lemma 1, the proof of which relies heavily on the rectangular geometry of the cells in the Mondrian process. One can generalize their strategy and obtain the same result for general STIT processes, but we can avoid the boundary issue that appears in their proof with the following modification. We instead compare $F_{\lambda,p}$ with the density

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

from Lemma 13. Now, by the assumptions, define $p_0 := \min_{x\in W\cap\mathcal{S}} p(x) > 0$. By Lemma 13, we add zero inside the norm to obtain the following upper bound on the first term above:

$$\begin{aligned}
\left\|\int_{\mathcal{S}}(z-x)F_{\lambda,p}(z)\mathrm{d}z\right\| &= \left\|\int_{\mathcal{S}}(z-x)\left(F_{\lambda,p}(z)-F_\lambda(z)\right)\mathrm{d}z\right\| \\
&\le \int_{\mathcal{S}}\|z-x\|\left|\mathbb{E}\left[\frac{p(z)}{\mu(Z_x^\lambda)}1_{\{z\in Z_x^\lambda\}}\right]-\mathbb{E}\left[\frac{1_{\{z\in Z_x^\lambda\}}}{\mathrm{vol}_s(Z_x^\lambda\cap\mathcal{S})}\right]\right|\mathrm{d}z \\
&\le \int_{\mathcal{S}}\|z-x\|\mathbb{E}\left[\frac{\int_{Z_x^\lambda\cap W\cap\mathcal{S}}|p(z)-p(y)|\mathrm{d}y}{\mu(Z_x^\lambda)\mathrm{vol}_s(Z_x^\lambda\cap\mathcal{S})}1_{\{z\in Z_x^\lambda\}}\right]\mathrm{d}z \\
&\le C_p\int_{\mathcal{S}}\|z-x\|\mathbb{E}\left[\frac{\int_{Z_x^\lambda\cap W\cap\mathcal{S}}\|z-y\|\mathrm{d}y}{\mu(Z_x^\lambda)\mathrm{vol}_s(Z_x^\lambda\cap\mathcal{S})}1_{\{z\in Z_x^\lambda\}}\right]\mathrm{d}z \\
&\le C_p\mathbb{E}\left[\frac{\mathrm{diam}(Z_x^\lambda)\mathrm{vol}_s(Z_x^\lambda\cap W\cap\mathcal{S})}{\mu(Z_x^\lambda)\mathrm{vol}_s(Z_x^\lambda\cap\mathcal{S})}\int_{\mathcal{S}}\|z-x\|1_{\{z\in Z_x^\lambda\}}\mathrm{d}z\right] \\
&\le C_p\mathbb{E}\left[\frac{\mathrm{diam}(Z_x^\lambda)^2\mathrm{vol}_s(Z_x^\lambda\cap W\cap\mathcal{S})}{\mu(Z_x^\lambda)}\right] \\
&\le \frac{C_p}{p_0}\mathbb{E}\left[\mathrm{diam}(Z_x^\lambda)^2\right] \le \frac{C_pc_{2,\Pi}}{\lambda^2p_0},
\end{aligned}$$

where the last inequality follows from Lemma 3 and stationarity. Thus,

$$\mathbb{E}[(\tilde{f}_\lambda(X)-f(X))^2] \le \left(\frac{LC_pc_{2,\Pi}}{\lambda^2p_0}+\frac{Lc_{\beta,\Pi}}{\lambda^{1+\beta}}\right)^2,$$

and the total risk satisfies

$$R(\hat{f}_{\lambda,n,M}(X)) = \mathbb{E}[\hat{f}_{\lambda,n,M}(X)-f(X))^2] \le O\left(\frac{L^2}{\lambda^2M}+\frac{L^2}{\lambda^{2(1+\beta)}}+\frac{\lambda^s}{n}\right).$$

Letting $\lambda=\lambda_n = L^{2/(s+2\beta+2)}n^{1/(s+2\beta+2)}$ and $M = M_n \succeq \lambda_n^{2\beta}$ gives the rate $O\left(L^{2s/(s+2\beta+2)}n^{-(2\beta+2)/(s+2\beta+2)}\right)$. □

3.5. **Proof of Theorem 11.**

Proof of Theorem 11. By Corollary 1 in [29], 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 associated zonoid/directional distribution. The distribution of the typical cell determines the distribution of the zero cell by Theorem 10.4.1 in [27], and thus, the same proof methods used in Theorems 7 and 9 can be applied in this setting and the results follow. □

4. Discussion and Future Work

This work expands and strengthens the theoretical basis for data independent random forests, and establishes stochastic geometry theory as an extremely promising tool set for analyzing random partition based regression and classification algorithms. In particular, we showed that a large class of random forests built from stationary hyperplane partitions all achieve the same minimax rates as Mondrian forests featured in [19], and we extended these rates to depend on the intrinsic dimension of the input as opposed to 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 notion of sparsity used in our assumptions about the input 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 sparsity of the input or the relevant feature space. Additionally, as mentioned in Remark 2, to obtain optimal rates in practice one needs an adaptive way of tuning the lifetime parameter and number of trees, since the sparsity dimension and regularity are not known *a priori*. For adaptation to regularity, a model aggregation method was proposed in [19] 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 direction distribution that depends on the input data density p_X may improve performance and decrease computational costs by decreasing the complexity of the partition needed to achieve optimal rates. It also remains an open question in general whether, under different assumptions about the underlying function f, one can obtain convergence rates that depend on the directional distribution, and whether optimal rates are achieved with an optimal choice of split directions.

A third research direction concerns data-dependent random forests. STIT or Poisson hyperplane tessellations with a non-stationary intensity measure Λ can yield inhomogeneous random partitions of space. The key open question is thus how to incorporate the given data set into Λ. In the stochastic geometry literature, some non-stationary random tessellation models have been studied. Sections 11.3 and 11.4 in [27] collect results on non-stationary flat processes and Poisson hyperplane tessellations. Many of the results there appear in [25], and [14] studies intersection densities and a generalization of the associated zonoid for non-stationary Poisson hyperplane tessellations.

References

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

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

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

[4] Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

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

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

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

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

[9] Yingying Fan Chien-Ming Chi, Patrick Vossler and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *arXiv:2004.13953*, 2021.

[10] Xuhui Fan, Bin Li, and Scott Sisson. The binary space partitioning-tree process. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1859–1867. PMLR, 09–11 Apr 2018.

[11] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[12] 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.

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

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

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

[16] Jason Klusowski. Sharp analysis of a simple model for random forests. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 757–765. PMLR, 13–15 Apr 2021.

[17] 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.
[18] 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.
[19] Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for Mondrian trees and forests. *Annals of Statistics*, 28(4):2253–2276, 2020.
[20] Werner Nagel and Viola Weiss. Limits of sequences of stationary planar tessellations. *Advances in Applied Probability*, 35:123–138, 2003.
[21] 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.
[22] Eliza O'Reilly and Ngoc Tran. Stochastic geometry to generalize the Mondrian process. *SIAM Journal on Mathematics of Data Science*, 2022.
[23] Tom Rainforth and Frank Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.
[24] 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.
[25] R. Schneider. Nonstationary Poisson hyperplanes and their induced tessellations. *Advances in Applied Probability*, 35:139–158, 2003.
[26] Rolf Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Cambridge University Press, 2013.
[27] Rolf Schneider and Wolfgang Weil. *Stochastic and Integral Geometry*. Probability and Its Applications. Springer-Verlag, Berlin, 2008.
[28] 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.
[29] Tomasz Schreiber and Christoph Thäle. Geometry of iteration stable tessellations: Connection with Poisson hyperplanes. *Bernoulli*, 19(5A):1637–1654, 2013.
[30] 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.
[31] 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.

Computing and Mathematical Sciences Department, California Institute of Technology, Pasadena, CA 91107
Email address: eoreilly@caltech.edu

Department of Mathematics, University of Texas at Austin, Austin, TX 78712
Email address: ntran@math.utexas.edu