Sharp Analysis of a Simple Model for Random Forests

Jason M. Klusowski
Department of Operations Research and Financial Engineering
Princeton University
Princeton, New Jersey 08544

Abstract

Random forests have become an important tool for improving accuracy in regression and classification problems since their inception by Leo Breiman in 2001. In this paper, we revisit a historically important random forest model, called *centered random forests*, originally proposed by Breiman in 2004 and later studied by Gérard Biau in 2012, where a feature is selected at random and the splits occurs at the midpoint of the node along the chosen feature. If the regression function is d-dimensional and Lipschitz, we show that, given access to n observations, the mean-squared prediction error is $O((n(\log n)^{(d-1)/2})^{-\frac{1}{d\log 2+1}})$. This positively answers an outstanding question of Biau about whether the rate of convergence for this random forest model could be improved beyond $O(n^{-\frac{1}{d(4/3)\log 2+1}})$. Furthermore, by a refined analysis of the approximation and estimation errors for linear models, we show that our new rate cannot be improved in general. Finally, we generalize our analysis and improve current prediction error bounds for another random forest model, called *median random forests*, in which each tree is constructed from subsampled data and the splits are performed at the empirical median along a chosen feature.

1 INTRODUCTION

Random forests are ubiquitous among ensemble averaging algorithms because of their ability to reduce overfitting, handle high-dimensional sparse settings, and efficient implementation. Due to these attractive features, they have been widely adopted and applied to various prediction and classification problems, such as those encountered in bioinformatics and computer vision.

One of the most widely used random forests is Breiman's random forest algorithm (Breiman, 2001), which was inspired by the random subspace method of (Ho, 1995), spatial feature selection of (Amit and Geman, 1997), and random decision method of (Dietterich, 2000). To this date, researchers have spent a great deal of effort in understanding theoretical properties of various streamlined versions of Breiman's original algorithm (Arlot and Genuer, 2014; Biau et al., 2008; Denil et al., 2014; Genuer, 2010, 2012; Geurts et al., 2006; Mentch and Hooker, 2016; Scornet et al., 2015; Wager and Walther, 2015). See (Biau and Scornet, 2016) for a comprehensive overview of current theoretical and practical understanding. The present paper is an effort to add to this body of work.

We assume the training data is $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i)$, $1 \le i \le n$ are i.i.d. with common joint distribution $\mathbb{P}_{\mathbf{X},Y}$. Here, $\mathbf{X}_i \in [0,1]^d$ is the feature or covariate and $Y_i \in \mathbb{R}$ is a continuous response variable. The j^{th} feature of $\mathbf{X}$ will be denoted by $\mathbf{X}^{(j)}$. We make the following assumptions on the statistical regression model.

Assumption 1. *The response variable can be written as $Y_i = f(\mathbf{X}_i) + \varepsilon_i$, for $i = 1, \dots, n$ where $f(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ is an unknown regression function and $\{\varepsilon_i\}_{1\le i\le n}$ are i.i.d. errors. Furthermore, $Var(Y \mid \mathbf{X}) \equiv \sigma^2$, for some positive constant σ^2, and $\mathbf{X}$ is uniformly distributed on $[0,1]^d$.*

Assumption 2. *The regression function $f(\cdot)$ is bounded in magnitude by a positive constant B and has bounded first-order partial derivates, i.e., $\|\partial f_j\|_\infty := \sup_{\mathbf{x}\in[0,1]^d} |\partial_j f(\mathbf{x})| < \infty$ for $j = 1, 2, \dots, d$. The largest infinity norm of the partial derivatives is denoted by $L = \max_j \|\partial f_j\|_\infty$.*

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

The efficacy of a predictor $\widehat{Y}(\mathbf{x}) = \widehat{Y}(\mathbf{x}; \mathcal{D})$ of $f(\cdot)$ will be measured in terms of its *mean squared prediction error*, $\mathbb{E}[(\widehat{Y}(\mathbf{X}) - f(\mathbf{X}))^2]$, where the expectation is with respect to the new input $\mathbf{X}$ and the training data $\mathcal{D}$. Throughout this paper, λ is the Lebesgue measure and log is the natural logarithm.

As mentioned earlier, many scholars have proposed and studied idealized versions of Breiman's original algorithm (Breiman, 2001), largely with the intent of reducing the complexity of their theoretical analysis. Unlike Breiman's random forests, these stylized versions are typically analyzed under the assumption that the probabilistic mechanism Θ that governs the construction of each tree *does not depend* on the training sample $\mathcal{D}$ (e.g., the splits are not data dependent). Such models are referred to as *purely random forests* (Genuer, 2012). On the other hand, recent works have proved properties like asymptotic normality (Mentch and Hooker, 2016; Wager, 2014) or consistency (Denil et al., 2014; Scornet, 2016a; Scornet et al., 2015), where the data may be bootstrapped or the splits determined by optimizing some empirical objective. However, these results are asymptotic in nature, and it is difficult to determine the quality of convergence as a function of the parameters of the random forest (e.g., sample size, dimension, and depth to which the individual trees are grown).

In this paper, we focus on another historically significant model that was proposed by Breiman in a technical report (Breiman, 2004). Here, importantly, the individual trees are grown *independently of the training sample* $\mathcal{D}$ (although subsequent work allows the trees to depend on a second sample $\mathcal{D}'$, independent of $\mathcal{D}$). Despite its simplicity, this random forest model captures a few of the attractive features of Breiman's original algorithm (Breiman, 2004), i.e., variance reduction by randomization. While feature selection (by random subset selection) is not directly related to or incorporated in vanilla centered random forests, it has nonetheless been considered by others (Biau, 2012). This model also allows one to provide a non-asymptotic prediction error bound that reveals the dependence on the parameters of the random forest.

Later, in an influential paper, Biau (2012) considered the same model and rigorously established some informal, heuristic-based claims made by Breiman. Both works of Breiman and Biau will serve as the basis for this article, whose primary purpose is to strengthen the analysis of this model and offer a full picture of its fundamental limits. Borrowing the terminology of (Scornet, 2016b), we shall refer to this model henceforth as a *centered random forest*.

New contributions. In his seminal paper, Biau (2012, Corollary 6) showed that the mean squared prediction error of a centered random forest is

$$O(n^{-\frac{1}{d(4/3)\log 2+1}}). \tag{1}$$

Biau also raised the question (Biau, 2012, Remark 7) as to whether this rate could be improved. We will answer this in the affirmative and show that the error (1) can indeed be improved to

$$O((n\log^{(d-1)/2} n)^{-r}), \tag{2}$$

where

$$r \coloneqq \frac{2\log_2(1-d^{-1}/2)}{2\log_2(1-d^{-1}/2)-1} = \frac{1}{d\log 2+1}(1+\delta),$$

and δ is some positive quantity that decreases to zero as d becomes large. In particular,

(a) We improve the rate in the exponent from $\frac{1}{d(4/3)\log 2+1}$ to $\frac{1}{d\log 2+1}$ and, due to the presence of the logarithmic term in (2), improve the convergence by a factor of $O((\log n)^{-\frac{1}{2\log 2}})$. Note that the rate (2) is *not* minimax optimal for the class of Lipschitz regression functions in d dimensions, unless $d = 1$.

(b) We generalize our proof techniques and use them to improve the convergence rates of other random forest models. In particular, for *median random forests* (Duroux and Scornet, 2018), we improve the rate from $O(n^{-\frac{\log_2(1-3d^{-1}/4)}{\log_2(1-3d^{-1}/4)-1}})$ to $O(n^{-\frac{2\log_2(1-d^{-1}/2)}{2\log_2(1-d^{-1}/2)-1}})$.

(c) We show that the rate (2) is not generally improvable for centered random forests. To accomplish this, we show that the approximation error is tight for all linear models with nonzero parameter vector. We also characterize the estimation error, which is, surprisingly, nearly the smallest among *all* purely random forests with splitting schemes that are not data dependent.

Additional comparisons between our work and (Biau, 2012) and (Duroux and Scornet, 2018) are provided in Table 1. The improvements in (a) and (b) stem from a novel analysis of the estimation and approximation errors of the random forest.

Related results. We now mention a few related results. Scornet (2016b) slightly altered the definition of random forests so that they could be rewritten as kernel methods. Scornet (2016b, Theorem 1) showed that *centered kernel random forests*, where the trees

are grown according to the same selection and splitting procedure as centered random forests, have mean squared prediction error $O(n^{-\frac{1}{d\log 2+3}}\log^2 n)$. In addition to the computational advantages of centered random forests when n and d are moderately sized, note that (2) is strictly better. The improved rate (2) is obtained by growing the trees to a shallower depth than the depth used by Scornet, and this may explain why he found centered kernel random forests to empirically outperform centered random forests for certain regression models (Scornet, 2016b, Model 1, Figure 5).

Other results have been established for function classes with additional smoothness assumptions. For example, a multivariate function on $[0,1]^d$ is of class $\mathcal{C}_k([0,1]^d)$ if all its k^{th} order partial derivatives exist and are bounded on $[0,1]^d$. Then, for regression functions in $\mathcal{C}_2([0,1]^d)$, (Arlot and Genuer, 2014, p. 21) obtained a similar rate of $O(n^{-r})$ for $d = d \geq 4$ under the so-called *balanced purely random forest* model, where all nodes are split at each stage (in contrast to single splits with centered random forests). However, in addition to requiring that the regression function is of class $\mathcal{C}_2([0,1]^d)$ (instead of just Lipschitz), it is unclear whether these random forest models can be modified to adapt to sparsity.

Finally, there are other versions of random forests, albeit defined somewhat differently than centered random forests, which have better theoretical guarantees. Recently, Mourtada et al. (2019) have shown that a type of random forest called *Mondrian forests* (for batch or online learning) achieve minimax optimal rates when $f(\cdot)$ belongs to $\mathcal{C}_1([0,1]^d)$ or $\mathcal{C}_2([0,1]^d)$, i.e., $\Theta(n^{-\frac{2}{d+2}})$ or $\Theta(n^{-\frac{4}{d+4}})$, respectively (Yang and Barron, 1999, Example 6.5).

2 RANDOM FORESTS

In general terms, a random forest is a predictor that is built from an ensemble of randomized base regression trees $\{\widehat{Y}(\mathbf{x};\Theta_m,\mathcal{D})\}_{1\leq m\leq M}$. The sequence $\{\Theta_m\}_{1\leq m\leq M}$ consists of i.i.d. realizations of a random variable Θ, which governs the probabilistic mechanism that builds each tree. These individual random trees are aggregated to form the final output

$$\widehat{Y}_M(\mathbf{X};\Theta_1,\dots,\Theta_M,\mathcal{D}) \coloneqq \frac{1}{M}\sum_{m=1}^{M}\widehat{Y}(\mathbf{X};\Theta_m,\mathcal{D}).$$

When M is sufficiently large, Theorem 3.3 from (Scornet, 2016a) justifies using

$$\widehat{Y}(\mathbf{X}) = \widehat{Y}(\mathbf{X},\mathcal{D}) \coloneqq \mathbb{E}_\Theta[\widehat{Y}(\mathbf{X};\Theta,\mathcal{D})]$$

in lieu of $\widehat{Y}(\mathbf{X};\Theta_1,\dots,\Theta_M,\mathcal{D})$, where $\mathbb{E}_\Theta$ denotes expectation with respect to Θ, conditionally on $\mathbf{X}$ and $\mathcal{D}$. We henceforth work with this asymptotic random forest.

The randomized base regression tree $\widehat{Y}(\mathbf{X};\Theta,\mathcal{D})$ is a local weighted average of all Y_i for which the corresponding $\mathbf{X}_i$ falls into the same node of the random partition as $\mathbf{X}$. For concreteness, let $\mathbf{t} = \mathbf{t}(\mathbf{X},\Theta,\mathcal{D})$ be the terminal node of the random partition containing $\mathbf{X}$ and define the individual tree predictor via

$$\widehat{Y}(\mathbf{X};\Theta,\mathcal{D}) \coloneqq \frac{\sum_{i=1}^n Y_i\mathbf{1}(\mathbf{X}_i\in\mathbf{t})}{\sum_{i=1}^n \mathbf{1}(\mathbf{X}_i\in\mathbf{t})}\,\mathbf{1}(\mathcal{E}), \qquad (3)$$

where $\mathcal{E}$ is the event that $\sum_{i=1}^n \mathbf{1}(\mathbf{X}_i\in\mathbf{t})$ is nonzero. We then take the expectation of these individual predictors with respect to the randomizing variable Θ yielding

$$\widehat{Y}(\mathbf{X}) = \sum_{i=1}^n \mathbb{E}_\Theta[W_i]Y_i,$$

where $W_i = W_i(\mathbf{t}) \coloneqq \frac{\mathbf{1}(\mathbf{X}_i\in\mathbf{t})}{N(\mathbf{t})}\,\mathbf{1}(\mathcal{E})$ are the weights corresponding to each observed output and $N(\mathbf{t}) \coloneqq \sum_{i=1}^n \mathbf{1}(\mathbf{X}_i\in\mathbf{t})$ is the total number of observations that fall into the same box of the random partition as $\mathbf{X}$. The node $\mathbf{t}$ is a Cartesian product and thus can be decomposed into the product of its sides $\prod_{j=1}^d [a_j(\mathbf{X}),b_j(\mathbf{X})]$, where $a_j(\mathbf{X}) = a_j(\mathbf{X},\Theta,\mathcal{D})$ and $b_j(\mathbf{X}) = b_j(\mathbf{X},\Theta,\mathcal{D})$ are its left and right endpoints, respectively, along the j^{th} axis.

Let us now formally define how each base tree $\widehat{Y}(\mathbf{x};\Theta_m,\mathcal{D})$ of a centered random forest and median random forest are constructed. We first describe the centered random forest from (Breiman, 2004) and (Biau, 2012).

Centered random forest:

(i) Initialize with $[0,1]^d$ as the root.

(ii) At each node, select one feature j in $\{1,2,\dots,d\}$ with probability $(p_j)_{1\leq j\leq d}$, where $\sum_{j=1}^d p_j = 1$.

(iii) Split the node at the midpoint of the interval along the direction of the selected feature.

(iv) Repeat steps (ii) and (iii) for the two daughter nodes until each node has been split exactly $\lceil \log_2 k_n \rceil$ times.

Remark 1. *Let us briefly mention that this model is similar in spirit to a recent random forest model proposed by (Basu et al., 2018), coined* iterative random forests. *Iterative random forests explicitly learn feature sampling probabilities, and so the results from the present paper could be useful for studying a simplified variant of the model.*

The split probabilities $(p_j)_{1\le j\le d}$ determine how frequently a particular direction is split. The agnostic choice $p_j = 1/d$ leads to the aforementioned rate (2). On the other hand, if the regression function $f(\cdot)$ is sparse and depends on only a small subset of the d variables, then by tuning these probabilities to be large for relevant variables and small otherwise, one can show convergence rates that do not degrade severely with the ambient dimension. In Section 3.2, we will consider data-driven choices of $(p_j)_{1\le j\le d}$ with the aide of a second sample $\mathcal{D}'$, independent of $\mathcal{D}$. In this case, the probabilities are *data-dependent*, i.e., $p_j = p_j(\mathcal{D}')$, and therefore our forthcoming prediction error bounds are written conditional on them.

The next random forest model we study is similar to centered random forests, though there are two important differences. First, each tree is constructed from subsampled data and, second, the splits are performed at the empirical median in an interval along a randomly chosen feature—thus making the splits *data-dependent*. As we will see, if the split probability sequence $(p_j)_{1\le j\le d}$ from centered random forests are uniform over the d features, i.e., $p_j = 1/d$, then these two random forest models have nearly identical convergence rates.

Median random forest:

(i) Select, uniformly without replacement, $n_0 < n$ data points among $\mathcal{D}_n$. Only these n_0 observations will be used in the tree construction.

(ii) Initialize with $[0,1]^d$ as the root.

(iii) At each node, select uniformly at random one feature j among $\{1, 2, \dots, d\}$.

(iv) Split the node at the empirical median of the $\mathbf{X}_i^{(j)}$ in the interval along the selected feature.

(v) Repeat steps (iii) and (iv) for the two daughter nodes until each node has been split exactly $\lceil \log_2 k_n \rceil$ times.

Though each tree in a median random forest is built from subsampled data, as with centered random forests, the output is computed with *all* the response values Y_i in $\mathcal{D}$, per (3).

Remark 2. *Since $\mathbf{X}^{(j)}$ is uniformly distributed on $[0,1]$, it has a binary expansion*

$$\mathbf{X}^{(j)} \stackrel{d}{=} \sum_{k\ge 1} B_k 2^{-k},$$

where $\{B_k\}_{k=1}^{\infty}$ are i.i.d. Bern(1/2)*. Thus, for the centered random forest model, if $K_j = K_j(\mathbf{X}, \Theta)$ is the number of times the nodes are split along the j^{th} feature, each endpoint of $[a_j(\mathbf{X}), b_j(\mathbf{X})]$ is a randomly stopped binary expansion of $\mathbf{X}^{(j)}$, viz.,*

$$a_j(\mathbf{X}) \stackrel{d}{=} \sum_{k=1}^{K_j} B_k 2^{-k}, \qquad b_j(\mathbf{X}) \stackrel{d}{=} 2^{-K_j} + \sum_{k=1}^{K_j} B_k 2^{-k}. \tag{4}$$

The representations (4) will also prove to be useful when we derive converse results for this random forest model.

Armed with these concepts and notation, we are now ready to present our main results.

3 MAIN RESULTS

We begin our analysis with the standard approximation/estimation error decomposition of the mean squared prediction error:

$$\underbrace{\mathbb{E}[(\overline{Y}(\mathbf{X}) - f(\mathbf{X}))^2]}_{\text{approximation error}} + \underbrace{\mathbb{E}[(\widehat{Y}(\mathbf{X}) - \overline{Y}(\mathbf{X}))^2]}_{\text{estimation error}}, \tag{5}$$

where $\overline{Y}(\mathbf{X}) \coloneqq \mathbb{E}[\widehat{Y}(\mathbf{X}) \mid \mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{X}]$. As is generally true with purely random forests, the estimation error is typically of order $\sigma^2 k_n/n$. What does vary with the specific random forest model, however, is the approximation error. Below we give a general upper bound on the approximation error that is valid for any random forest model. Due to space constraints, the proof of Theorem 1 is given in the supplementary material.

Theorem 1. *For any random forest model whose construction depends on the data only through $\mathbf{X}_1, \dots, \mathbf{X}_n$, under Assumption 2,*

$$\begin{aligned}&\mathbb{E}[(\overline{Y}(\mathbf{X}) - f(\mathbf{X}))^2] \\ &\le d \sum_{j=1}^{d} \|\partial f_j\|_\infty^2 \mathbb{E}[(\mathbb{E}_\Theta[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2] + B^2 \mathbb{P}(\mathcal{E}^c).\end{aligned} \tag{6}$$

Despite its simple proof, Theorem 1 leads to nontrivial improvements over past work. It is now easy to isolate precisely where our improvements manifest. In standard analysis of random forest models, the quantity $\mathbb{E}_\Theta[(b_j(\mathbf{X}) - a_j(\mathbf{X}))^2]$ is typically analyzed directly, where the Θ-averaging occurs on the outside of the square. On the other hand, the bound (6) allows the Θ-averaging to occur *inside* the square, and thus by Jensen's inequality, it represents a uniform improvement, i.e.,

$$(\mathbb{E}_\Theta[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2 \le \mathbb{E}_\Theta[(b_j(\mathbf{X}) - a_j(\mathbf{X}))^2].$$

Another interpretation of this improvement is that infinite forests perform better than single trees. Indeed, for single trees, the rate would depend on $\mathbb{E}_{\Theta,\mathcal{D}}[(b_j(\mathbf{X}) - a_j(\mathbf{X}))^2]$ instead of $(\mathbb{E}_{\Theta}[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2$, leading to the same rate as (Biau, 2012). The Θ-averaging can partly, though not fully, mitigate the suboptimality of the tree construction.

Both Biau (2012) and Duroux and Scornet (2018) bound the approximation error by $O(k_n^{\log_2(1-3d^{-1}/4)}) = O(k_n^{-\frac{1}{d(4/3)\log 2}})$. We will use (6) to improve this bound to $O(k_n^{2\log_2(1-d^{-1}/2)}) = O(k_n^{-\frac{1}{d\log 2}})$. Note that this bound is the same as (Arlot and Genuer, 2014, Corollary 9) when $d \geq 4$, though the authors analyze the balanced purely random forest model and make a stronger assumption that $f(\cdot)$ has bounded second-order partial derivatives.

3.1 Centered Random Forests

In this subsection, we derive bounds on the mean squared prediction error of a centered random forest in terms of k_n and the probability sequence $(p_j)_{1\leq j\leq d}$. As a consequence, we also obtain rates of convergence.

Theorem 2 (Centered random forests)**.** *Let $\mathcal{P} \coloneqq \{j : p_j \neq 0\}$ and $d_0 \coloneqq \#\mathcal{P}$. Under Assumption 1 and Assumption 2 and conditional on $(p_j)_{1\leq j\leq d}$,*

$$\mathbb{E}[(\widehat{Y}(\mathbf{X}) - f(\mathbf{X}))^2] \leq d\sum_{j=1}^{d} \|\partial_j f\|_\infty^2 k_n^{2\log_2(1-p_j/2)} + \frac{12\sigma^2 k_n}{n}\frac{8^{d_0}}{\sqrt{\prod_{j\in\mathcal{P}} p_j \times \log_2^{d_0-1}(k_n)}} + B^2 e^{-n/(2k_n)}.$$

Consequently, if $p \coloneqq \min_j p_j$, $r \coloneqq \frac{2\log_2(1-p/2)}{2\log_2(1-p/2)-1}$, and $k_n = c(n(\log_2^{d_0-1} n)^{1/2})^{1-r}$ for some constant $c > 0$ independent of n, then, conditional on p, there exists a constant $C > 0$ independent of n, such that

$$\mathbb{E}[(\widehat{Y}(\mathbf{X}) - f(\mathbf{X}))^2] \leq C(n(\log_2^{d_0-1} n)^{1/2})^{-r}. \quad (7)$$

Remark 3. *Since theoretically favorable choices of k_n depend on unknown quantities, in practice, good values can be chosen using cross-validation.*

Remark 4. *We emphasize that d_0 is simply the number of nonzero feature selection probabilities and so it need not correspond to any sparsity assumptions in the regression model. However, if the regression function is sparse and the set of nonzero feature selection probabilities $\mathcal{P}$ corresponds to the set of relevant variables, then d_0 equals the sparsity level.*

Proof. First, Biau (2012, Section 5.3, p. 1089) shows that $\mathbb{P}(\mathcal{E}^c) \leq e^{-n/(2k_n)}$. Next, let $K_j = K_j(\mathbf{X}, \Theta)$ be the number of times the nodes are split along the j^{th} feature and note that K_j is conditionally distributed $\text{Bin}(\lceil \log_2 k_n \rceil, p_j)$ given $\mathbf{X}$. Then, conditional on $(p_j)_{1\leq j\leq d}$,

$$\begin{aligned}\mathbb{E}_\Theta[b_j(\mathbf{X}) - a_j(\mathbf{X})] &= \mathbb{E}_\Theta[2^{-K_j}] \\ &= \mathbb{E}_{K\sim\text{Bin}(\lceil\log_2 k_n\rceil, p_j)}[2^{-K}] \\ &= (1 - p_j/2)^{\lceil\log_2 k_n\rceil} \\ &\leq k_n^{\log_2(1-p_j/2)}.\end{aligned}$$

Thus, by Theorem 1, the approximation error $\mathbb{E}[(\overline{Y}(\mathbf{X}) - f(\mathbf{X}))^2]$ is bounded by

$$d\sum_{j=1}^{d} \|\partial_j f\|_\infty^2 k_n^{2\log_2(1-p_j/2)} + B^2 e^{-n/(2k_n)}. \quad (8)$$

Next, we bound the estimation error of the random forest. In particular, we show that, conditional on $(p_j)_{1\leq j\leq d}$, $\mathbb{E}[(\widehat{Y}(\mathbf{X}) - \overline{Y}(\mathbf{X}))^2]$ is at most

$$\frac{12\sigma^2 k_n}{n}\frac{8^{d_0}}{\sqrt{\prod_{j\in\mathcal{P}} p_j \times \log_2^{d_0-1} k_n}}. \quad (9)$$

Henceforth, we let K_j', $[a_j'(\mathbf{X}), b_j'(\mathbf{X})']$, and $\mathbf{t}'$ denote the feature selection frequency, terminal node side, and terminal node, respectively, from an independent copy Θ' of Θ. It is shown in (Biau, 2012, Section 5.2, p. 1085) that

$$\mathbb{E}[(\widehat{Y}(\mathbf{X}) - \overline{Y}(\mathbf{X}))^2] \leq \frac{12\sigma^2 k_n^2}{n}\mathbb{E}_{\Theta,\Theta'}[\lambda(\mathbf{t}\cap\mathbf{t}')]. \quad (10)$$

We can use the representations (4) to show that for any Θ and Θ', the sides of the node are nested according to $[a_j'(\mathbf{X}), b_j'(\mathbf{X})] \subseteq [a_j(\mathbf{X}), b_j(\mathbf{X})]$ if and only if $K_j' \geq K_j$ and hence

$$\lambda([a_j(\mathbf{X}), b_j(\mathbf{X})] \cap [a_j'(\mathbf{X}), b_j'(\mathbf{X})]) = 2^{-\max\{K_j, K_j'\}}. \quad (11)$$

Using this, we have that $\lambda(\mathbf{t}\cap\mathbf{t}')$ equals

$$\begin{aligned}&\prod_{j=1}^{d} \lambda([a_j(\mathbf{X}), b_j(\mathbf{X})] \cap [a_j'(\mathbf{X}), b_j'(\mathbf{X})]) \\ &= 2^{-\sum_{j=1}^d \max\{K_j, K_j'\}} \\ &= 2^{-\lceil\log_2 k_n\rceil - \frac{1}{2}\sum_{j=1}^d |K_j - K_j'|}, \quad (12)\end{aligned}$$

where the equality in (12) follows from the identity

$$\begin{aligned}\sum_{j=1}^{d}\max\{K_j, K_j'\} &= \frac{1}{2}\sum_{j=1}^{d} K_j + \frac{1}{2}\sum_{j=1}^{d} K_j' \\ &\quad + \frac{1}{2}\sum_{j=1}^{d}|K_j - K_j'| \\ &= \lceil\log_2 k_n\rceil + \frac{1}{2}\sum_{j=1}^{d}|K_j - K_j'|.\end{aligned}$$

Next, note that conditional on $\mathbf{X}$, $(K_1, \ldots, K_d)$ has a multinomial distribution with $\lceil \log_2 k_n \rceil$ trials and event probabilities $(p_j)_{1 \le j \le d}$. We take the expected value of (12) and use Lemma S.1 from the supplementary material, yielding

$$\mathbb{E}_{\Theta,\Theta'}[2^{-\frac{1}{2}\sum_{j=1}^{d}|K_j - K'_j|}] = \mathbb{E}_{\Theta,\Theta'}[2^{-\frac{1}{2}\sum_{j\in\mathcal{P}}|K_j - K'_j|}] \le \frac{8^{d_0}}{\sqrt{\prod_{j\in\mathcal{P}} p_j \times \log_2^{d_0-1} k_n}}. \tag{13}$$

Combining (10), (12), and (13) proves (9). The choice of k_n that leads to the prediction error bound (7) is determined by (approximately) optimizing the sum of the bounds (8) and (9) on the approximation and estimation errors. □

Remark 5. *In proving the estimation error bound (9), we depart from the strategy of (Biau, 2012), which we now briefly outline. Biau's approach consists of applying Hölder's inequality to the expectation of (12) and resultant expected product, i.e.,*

$$\mathbb{E}_{\Theta,\Theta'}[2^{-\sum_{j=1}^{d}\max\{K_j, K'_j\}}] \le k_n^{-1} \prod_{j\in\mathcal{P}} (\mathbb{E}_{\Theta,\Theta'}[2^{-\frac{d}{2}|K_j - K'_j|}])^{1/d}. \tag{14}$$

With K_j conditionally distributed $\mathrm{Bin}(\lceil \log_2 k_n \rceil, p_j)$ *given* $\mathbf{X}$*, Biau uses the previous inequality together with the fact that, for $d \ge 2$, $\mathbb{E}_{\Theta,\Theta'}[2^{-\frac{d}{2}|K_j - K'_j|}] \le \frac{12}{\sqrt{\pi p_j (1-p_j) \log_2 k_n}}$ (Biau, 2012, Proposition 13), to conclude that the estimation error is of order $O((k_n/n)(\log_2 k_n)^{-d_0/(2d)})$. Our approach is different. Instead of reducing the calculations so that the expectations involve only their marginals K_j and K'_j, we work with their joint multinomial distribution.*

Remark 6. *In the fully grown case when $k_n = n$ (i.e., when there is on average one observation per terminal node), (9) shows that the estimation error still decays as $O((\log n)^{-(d_0-1)/2})$, due to the correlation between trees.*

Remark 7. *It is a standard result for partitioning based regression predictors that the estimation error is of order $\sigma^2 k_n/n$ and hence our improvement (9) is only in terms of the logarithmic factor $(\log n)^{(d_0-1)/2}$. Note that if the split probabilities $(p_j)_{1\le j\le d}$ are uniform over the d input features, the logarithmic factor multiplying $12\sigma^2 k_n/n$ is small if the tree depth $\lceil \log_2 k_n \rceil$ is greater than a constant multiple of d, i.e.,*

$$\lceil \log_2 k_n \rceil \gg d \quad \Longrightarrow \quad \frac{8^d}{\sqrt{(1/d)^d \log_2^{d-1} k_n}} \ll 1.$$

Thus, the improvement to the estimation error manifests only for deep trees. However, with these specifications for $(p_j)_{1\le j\le d}$, the leading term in the approximation error bound (8) is $d^2 L^2 k_n^{-d^{-1}/\log 2}$—which is small precisely when $\lceil \log_2 k_n \rceil \gg d$—so the improvement to the estimation error is in fact always present in the regime of interest for small mean squared prediction error.

3.2 Data-driven Approach for Split Probabilities

To avoid the curse of dimensionality—which plagues high-dimensional regression models—and the associated undesirable consequences (e.g., overfitting and large sample requirements), it is typically assumed that $f(\cdot)$ is sparse in the sense that it only depends on a small subset $\mathcal{S}$ of the d features, where $s := |\mathcal{S}| \ll d$. In other words, $f(\cdot)$ is almost surely equal to its restriction to the subspace of its *relevant* features in $\mathcal{S}$. Conversely, the output of $f(\cdot)$ does not dependent on *irrelevant* features that belong to $\mathcal{S}^c$. Of course, the set $\mathcal{S}$ is not known a priori and must be learned from the data.

The approximation error upper bound (8) involves a subtle interplay between the split probabilities $(p_j)_{1\le j\le d}$ and the size of the partial derivatives of the regression function—directions that have larger variability require more splits—and thus have higher selection probabilities. If each direction contributes equally to the variability of the regression function, then (by a Lagrange multipliers argument),

$$\sum_{j=1}^{d} \|\partial_j f\|_\infty^2 k_n^{2\log_2(1-p_j/2)} \approx L^2 \sum_{j\in\mathcal{S}} k_n^{2\log_2(1-p_j/2)}$$

are minimized when the $(p_j)_{1\le j\le d}$ are uniform over the set of relevant features, i.e., $p_j = 1/s$ for $j \in \mathcal{S}$ and $p_j = 0$ otherwise. Similarly, the factor $(\prod_{j\in\mathcal{P}} p_j)^{-1/2}$ from the estimation error (9) is separately minimized when the $(p_j)_{1\le j\le d}$ are uniform over the set of relevant features. When this is the case, Theorem 2 yields the rate

$$(n(\log^{s-1} n)^{1/2})^{-\frac{1}{s\log 2+1}},$$

which beats the minimax optimal rate $\Theta(n^{-\frac{2}{d+2}})$ (Yang and Barron, 1999, Example 6.5) for Lipschitz regression models in d dimensions roughly when $s \le \lfloor 0.72d \rfloor$ (cf., $d \le \lfloor 0.54s \rfloor$ from (Biau, 2012, p. 1069)).

Since the set $\mathcal{S}$ is not known a priori, how can one learn these split probabilities from the data? To avoid entanglement with the same data used to train the random forest, one solution is to adaptively select candidate strong features using a second sample $\mathcal{D}' = \{(\mathbf{X}'_1, Y'_1), \ldots (\mathbf{X}'_n, Y'_n)\}$, independent of $\mathcal{D}$ (which can

be done, for example, by sample-splitting). Here, candidate strong features are those that maximize the *decrease in variance* (the impurity) that would be obtained if the root node $[0,1]^d$ is split along the direction j at position z, denoted by $\widehat{\Delta}(j,z)$ (Breiman et al., 1984, Definition 8.13) and constructed from the second sample $\mathcal{D}'$. Indeed, it was recently shown in (Klusowski and Tian, 2021) that if Y is an additive model with smooth component functions, then one can, with high probability, identify the relevant variables according to the size of $\widehat{\Delta}(j,z)$. This is similar in spirit to Breiman's random forests, except that the candidate features are chosen from a random subset and the decrease in variance $\widehat{\Delta}(j,z)$ depends on a current node of the tree.

3.3 Median Random Forests

Following the same path as the previous subsection, here we derive bounds and rates of convergence for the mean squared prediction error of a median random forest.

Theorem 3 (Median random forests)**.** *Suppose $n_0 \geq 2^{\lceil \log_2 k_n \rceil}$. Then, under Assumption 1 and Assumption 2,*

$$\mathbb{E}[(\widehat{Y}(\mathbf{X})-f(\mathbf{X}))^2] \leq 256 d^2 L^2 k_n^{2\log_2(1-d^{-1}/2)} + 2\sigma^2 k_n/n.$$

Consequently, if $r \coloneqq \frac{2\log_2(1-d^{-1}/2)}{2\log_2(1-d^{-1}/2)-1}$ and $k_n = cn^{1-r}$ for some constant $c>0$ independent of n, then there exists a constant $C>0$ independent of n, such that

$$\mathbb{E}[(\widehat{Y}(\mathbf{X}) - f(\mathbf{X}))^2] \leq Cn^{-r}.$$

Proof. We follow the proof of (Duroux and Scornet, 2018, Lemma 6.1), but with some important modifications. Let $\mathbf{x} \in [0,1]^d$ and let $\mathcal{C} = \{N_0, N_1, \ldots, N_{2^{\lceil \log_2 k_n \rceil}}\}$ be the number of points in the successive nodes containing $\mathbf{x}$ (for example, N_0 is the number of points in the root node of the tree, i.e., $N_0 = n_0$). We also let j_k denote the feature index selected at the k^{th} step. The counts in $\mathcal{C}$ implicitly depend on $\mathcal{D}_n$ and Θ, but we suppress these dependencies for clarity. Then $b_j(\mathbf{X}) - a_j(\mathbf{X})$ can be written as a product of independent beta distributions:

$$b_j(\mathbf{X}) - a_j(\mathbf{X}) \overset{d}{=} \prod_{k=1}^{\lceil \log_2 k_n \rceil} B_k^{\mathbf{1}(j_k=j)},$$

where B_k are independent and distributed $\text{Beta}(n_k+1, n_{k-1}-n_k)$, conditional on $N_k = n_k$. Consequently,

$$\begin{aligned}\mathbb{E}_{\Theta|\mathcal{C}}[b_j(\mathbf{X}) - a_j(\mathbf{X})] &= \prod_{k=1}^{\lceil \log_2 k_n \rceil} \mathbb{E}_{\Theta|\mathcal{C}}\Big[B_k^{\mathbf{1}(j_k=j)}\Big] \\ &= \prod_{k=1}^{\lceil \log_2 k_n \rceil} \Big(\frac{d-1}{d} + \frac{1}{d}B_k\Big), \end{aligned} \tag{15}$$

since $\mathbb{P}_{\Theta|\mathcal{C}}(j_k = j) = 1/d$. Now, by Jensen's inequality for the square function,

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n}[(\mathbb{E}_\Theta[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2] \\ &\leq \mathbb{E}[(\mathbb{E}_{\Theta|\mathcal{C}}[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2] \\ &= \mathbb{E}_{\mathcal{C}}[\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_d|\mathcal{C}}[(\mathbb{E}_{\Theta|\mathcal{C}}[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2]].\end{aligned}$$

Furthermore, using (15), we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathcal{C}}[(\mathbb{E}_{\Theta|\mathcal{C}}[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2] \\ &= \prod_{k=1}^{\lceil \log_2 k_n \rceil} \mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathcal{C}}\Big[\Big(\frac{d-1}{d} + \frac{1}{d}B_k\Big)^2\Big]. \end{aligned} \tag{16}$$

We must calculate the first and second moments of a $\text{Beta}(n_k+1, n_{k-1}-n_k)$ distribution in (16). Doing so yields

$$\begin{aligned}\prod_{k=1}^{\lceil \log_2 k_n \rceil} \Big(1 - \frac{2}{d} + \frac{1}{d^2} + 2\frac{n_k+1}{n_{k-1}+1}\Big(\frac{1}{d} - \frac{1}{d^2}\Big) \\ + \frac{(n_k+1)(n_k+2)}{(n_{k-1}+1)(n_{k-1}+2)}\frac{1}{d^2}\Big).\end{aligned}$$

Next, we use the relation $n_k \leq \lceil n_{k-1}/2 \rceil \leq (n_{k-1}+1)/2$ to further bound the above expression by

$$\begin{aligned}&\prod_{k=1}^{\lceil \log_2 k_n \rceil} \Big(1 - \frac{2}{d} + \frac{1}{d^2} + \frac{n_{k-1}+3}{n_{k-1}+1}\Big(\frac{1}{d} - \frac{1}{d^2}\Big) \\ &\qquad + \frac{(n_{k-1}+3)(n_{k-1}+5)}{(n_{k-1}+1)(n_{k-1}+2)}\frac{1}{4d^2}\Big) \\ &\leq \prod_{k=1}^{\lceil \log_2 k_n \rceil} \Big(\Big(1 - \frac{1}{2d}\Big)^2 + \frac{2}{d(n_{k-1}+1)}\Big). \end{aligned} \tag{17}$$

Now, $n_k \geq \lfloor n_{k-1}/2 \rfloor \geq (n_{k-1}-1)/2$ and hence by induction, $n_k \geq (1/2)^k n_0 - 1$. Furthermore, by assumption, $n_0 2^{-\lceil \log_2 k_n \rceil} \geq 1$. Putting these facts together, we have

$$\frac{1}{n_{k-1}+1} \leq \frac{1}{(1/2)^{k-1}n_0} \leq 2^{k-1-\lceil \log_2 k_n \rceil}. \tag{18}$$

Continuing from (17) and using (18), we have

$$\begin{aligned}
&\log_2 \prod_{k=1}^{\lceil \log_2 k_n \rceil} \left(\left(1 - \frac{1}{2d}\right)^2 + \frac{1}{d(n_{k-1}+1)} \right) \\
&\leq \log_2 \prod_{k=1}^{\lceil \log_2 k_n \rceil} \left(\left(1 - \frac{1}{2d}\right)^2 + \frac{2^{k-\lceil \log_2 k_n \rceil}}{d} \right) \\
&\leq 2\lceil \log_2 k_n \rceil \log_2(1 - d^{-1}/2) + 8.
\end{aligned}$$

This shows that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n}[(\mathbb{E}_\Theta[b_j(\mathbf{X}) - a_j(\mathbf{X})])^2] \\
&\quad \leq 2^8 (1 - d^{-1}/2)^{2\lceil \log_2 k_n \rceil} \\
&\quad \leq 256 k_n^{2\log_2(1-d^{-1}/2)},
\end{aligned}$$

and hence by Theorem 1, the approximation error $\mathbb{E}[(\overline{Y}(\mathbf{X}) - f(\mathbf{X}))^2]$ is at most $d\sum_{j=1}^d \|\partial_j f\|_\infty^2 256 k_n^{2\log_2(1-d^{-1}/2)}$. This quantity is further upper bounded by $256 d^2 L^2 k_n^{2\log_2(1-d^{-1}/2)}$. Finally, it is shown in (Duroux and Scornet, 2018, Section 6.2) that the estimation error has the bound $\mathbb{E}[(\widehat{Y}(\mathbf{X}) - \overline{Y}(\mathbf{X}))^2] \leq 2\sigma^2 k_n/n$. □

In Table 1, we catalogue our improvements in Theorem 2 and Theorem 3 to (Biau, 2012) and (Duroux and Scornet, 2018) in terms of the estimation, approximation, and prediction errors of a k_n that optimizes our upper bounds on the tradeoff between the goodness-of-fit and complexity. To make more the comparisons between the two random forest models easier to see, for centered random forests, we consider the agnostic choice $p_j = 1/d$, producing $p = 1/d$. For the sake of clarity, we also ignore logarithmic factors in n and replace the rate $\frac{2\log_2(1-d^{-1}/2)}{2\log_2(1-d^{-1}/2)-1}$ with the more palatable lower bound $(d\log 2 + 1)^{-1}$.

Remark 8. *According to (Yang and Barron, 1999, Example 6.5), the minimax rate for Lipschitz regression models in d dimensions is* $\Theta(n^{-\frac{2}{d+2}})$. *Thus, we see our rate* $\frac{2\log_2(1-d^{-1}/2)}{2\log_2(1-d^{-1}/2)-1}$ *for median and centered random forests is minimax optimal* only *when* $d = 1$.

Remark 9. *Compare our choice* $k_n = \Theta(n^{\frac{d\log 2}{d\log 2+1}})$ *with that of (Biau, 2012, Corollary 6) and (Duroux and Scornet, 2018, Theorem 3.1), namely,* $k_n = \Theta(n^{\frac{d(4/3)\log 2}{d(4/3)\log 2+1}})$. *Thus, a better prediction error bound is achieved if the trees are shallower.*

4 TIGHTNESS OF BOUNDS

In this section, we show that the approximation error bound (8) for centered random forests we derived in Theorem 2 cannot be improved in general. To see this, consider the linear model $Y = \langle \boldsymbol{\beta}, \mathbf{X} \rangle + \varepsilon$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(d)})$ is a d-dimensional parameter vector. Then we have the following lower bound on the approximation error of a centered random forest. This lower bound decays with k_n at the same rate as the approximation error upper bound in Theorem 2, regardless of the split probabilities $(p_j)_{1\leq j\leq d}$. We provide the proof of Theorem 4 in the supplementary material.

Table 1: The old approximation error, estimation error, and mean squared error (MSE) bounds from (Biau, 2012) (for centered random forests) and (Duroux and Scornet, 2018) (for median random forests) and the new bounds from Theorem 2 and Theorem 3.

	APPROX. ERR.	EST. ERR.	MSE
OLD	$k_n^{-\frac{1}{d(4/3)\log 2}}$	k_n/n	$n^{-\frac{1}{d(4/3)\log 2+1}}$
NEW	$k_n^{-\frac{1}{d\log 2}}$	k_n/n	$n^{-\frac{1}{d\log 2+1}}$

Theorem 4. *Suppose* $Y = \langle \boldsymbol{\beta}, \mathbf{X} \rangle + \varepsilon$, *where* $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(d)})$ *is a d-dimensional parameter vector. Also, assume* $n \geq 2^{\lceil \log_2 k_n \rceil}$. *Then, under Assumption 1 and conditional on* $(p_j)_{1\leq j\leq d}$,

$$\mathbb{E}[(\overline{Y}(\mathbf{X}) - f(\mathbf{X}))^2] \geq \frac{1}{96} \sum_{j=1}^d |\boldsymbol{\beta}^{(j)}|^2 k_n^{2\log_2(1-p_j/2)}.$$

We also argue that the estimation error bound (9) derived in the proof of Theorem 2 is nearly tight when the split probabilities are uniform over all d features. To this end, Lin and Jeon (2006, Theorem 3) showed that if w_{max} is the maximum number of observations per terminal node, the estimation error for *any* purely random forest (with uniformly distributed input $\mathbf{X}$) is at least a constant multiple of[1]

$$\frac{\sigma^2}{w_{max}} \times \frac{(d-1)!}{2^d \log^{d-1} n}. \tag{19}$$

Now, the number of observations per terminal node of a centered random forest is on average about $w_{avg} = n/k_n$ and hence from (9), centered random forests nearly achieve the best-case estimation error (19), namely,

$$\frac{\sigma^2}{w_{avg}} \times \sqrt{\frac{(8d)^d}{\log^{d-1}(n/w_{avg})}}. \tag{20}$$

Taken together, (19) and (20) imply that centered random forests have nearly the lowest estimation error

[1]The lower bound in (Lin and Jeon, 2006, Theorem 3) is actually for the mean squared prediction error, but the proof therein is for the estimation error.

among all purely random forests with splitting schemes that are not data dependent. More rigorously, we can prove the following estimation error lower bound, which is valid for any probability sequence $(p_j)_{1\le j\le d}$. The proof of Theorem 5 is furnished in the supplementary material.

Theorem 5. *Let $\mathcal{P} \coloneqq \{j : p_j \neq 0\}$ and $d_0 \coloneqq \#\mathcal{P}$. Suppose $\lceil \log_2 k_n \rceil p_j \geq 1$ for $j \in \mathcal{P}$ and $n \geq 2^{\lceil \log_2 k_n \rceil}$. Then, under Assumption 1 and conditional on $(p_j)_{1\le j\le d}$,*

$$\mathbb{E}[(\widehat{Y}(\mathbf{X}) - \overline{Y}(\mathbf{X}))^2] \geq \frac{\sigma^2 k_n}{5n} \frac{(47)^{-d_0}}{\prod_{j\in\mathcal{P}} p_j \times (\lceil \log_2 k_n \rceil)^{d_0 - 1}}.$$

Combining the sharpness of our approximation and estimation error bounds for linear models, we conclude that the rate (2) is not generally improvable and hence centered random forests do not achieve the d-dimensional minimax optimal rate $\Theta(n^{-\frac{2}{d+2}})$ for d-dimensional Lipschitz regression functions. While centered random forests enjoy near optimal estimation error (19) (among all purely random forests), their $O(k_n^{-\frac{1}{d\log 2}})$ approximation error is far from the optimal $\Theta(k_n^{-2/d})$ required to achieve the minimax rate. Only in the one-dimensional setting do centered or median random forests achieve the minimax optimal rate $\Theta(n^{-2/3})$ for Lipschitz regression functions in one dimension (Yang and Barron, 1999, Example 6.5)—in the multi-dimensional setting, the rate is suboptimal. These converse statements shed light on the importance of more sophisticated tree construction mechanisms—like Mondrian random forests (Mourtada et al., 2019)—if optimality is to be guaranteed.

5 Conclusion

As explained in the introduction, centered random forests were conceived by Leo Breiman in 2004 as a way to theoretically explain some of the heuristics of his original algorithm (which uses CART methodology for the trees) from 2001, such as the ensemble principle. Because of these connections, centered random forests have since then become a sort of theoretical benchmark model in the literature and have inspired the development of other variants that more closely resemble the original algorithm. Continuing from (Biau, 2012), our work completes the history of this influential random forest model by providing a full characterization of its fundamental limits. In doing so, we also reveal the importance of more sophisticated tree constructions (such as Mondrian random forests (Mourtada et al., 2019) or constructions that use both the input and output data, since otherwise ensembles of such trees may have suboptimal performance.

Acknowledgements

A portion of this work was completed while the author was a visiting graduate student at The Wharton School Department of Statistics. He is grateful to Matthew Olson for suggesting relevant literature to review and Edgar Dobriban for helpful discussions. Financial support was provided in part by NSF DMS-1915932 and NSF HDR TRIPODS DATA-INSPIRE DCCF-1934924.

References

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588.

Arlot, S. and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, page 201711236.

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

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

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

Breiman, L. (2004). Consistency for a simple model of random forests. *Technical Report 670, UC Berkeley*.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Denil, M., Matheson, D., and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning (ICML)*.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128.

Genuer, R. (2010). Risk bounds for purely uniformly random forests. *arXiv preprint arXiv:1006.2980*.

Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1)-Volume 1*, page 278. IEEE Computer Society.

Klusowski, J. M. and Tian, P. (2021). Nonparametric variable screening with optimal decision stumps. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(1):841–881.

Mourtada, J., Gaïffas, S., and Scornet, E. (2019). Minimax optimal rates for Mondrian trees and forests. *To appear, Annals of Statistics*.

Scornet, E. (2016a). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83.

Scornet, E. (2016b). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500.

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

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599.