The Annals of Statistics
2015, Vol. 43, No. 4, 1716–1741
DOI: 10.1214/15-AOS1321
© Institute of Mathematical Statistics, 2015

CONSISTENCY OF RANDOM FORESTS[1]

BY ERWAN SCORNET*, GÉRARD BIAU* AND JEAN-PHILIPPE VERT†

Sorbonne Universités and MINES ParisTech, PSL-Research University†*

Random forests are a learning algorithm proposed by Breiman [*Mach. Learn.* **45** (2001) 5–32] that combines several randomized decision trees and aggregates their predictions by averaging. Despite its wide usage and outstanding practical performance, little is known about the mathematical properties of the procedure. This disparity between theory and practice originates in the difficulty to simultaneously analyze both the randomization process and the highly data-dependent tree structure. In the present paper, we take a step forward in forest exploration by proving a consistency result for Breiman's [*Mach. Learn.* **45** (2001) 5–32] original algorithm in the context of additive regression models. Our analysis also sheds an interesting light on how random forests can nicely adapt to sparsity.

1. Introduction.

Random forests are an ensemble learning method for classification and regression that constructs a number of randomized decision trees during the training phase and predicts by averaging the results. Since its publication in the seminal paper of Breiman (2001), the procedure has become a major data analysis tool, that performs well in practice in comparison with many standard methods. What has greatly contributed to the popularity of forests is the fact that they can be applied to a wide range of prediction problems and have few parameters to tune. Aside from being simple to use, the method is generally recognized for its accuracy and its ability to deal with small sample sizes, high-dimensional feature spaces and complex data structures. The random forest methodology has been successfully involved in many practical problems, including air quality prediction (winning code of the EMC data science global hackathon in 2012, see http://www.kaggle.com/c/dsg-hackathon), chemoinformatics [Svetnik et al. (2003)], ecology [Cutler et al. (2007), Prasad, Iverson and Liaw (2006)], 3D object recognition [Shotton et al. (2013)] and bioinformatics [Díaz-Uriarte and Alvarez de Andrés (2006)], just to name a few. In addition, many variations on the original algorithm have been proposed to improve the calculation time while maintaining good prediction accuracy; see, for example, Amaratunga, Cabrera and Lee (2008), Geurts, Ernst and Wehenkel (2006). Breiman's forests have also been

Received May 2014; revised February 2015.

[1] Supported by the European Research Council [SMAC-ERC-280032].

MSC2010 subject classifications. Primary 62G05; secondary 62G20.

Key words and phrases. Random forests, randomization, consistency, additive model, sparsity, dimension reduction.

extended to quantile estimation [Meinshausen (2006)], survival analysis [Ishwaran et al. (2008)] and ranking prediction [Clémençon, Depecker and Vayatis (2013)].

On the theoretical side, the story is less conclusive, and regardless of their extensive use in practical settings, little is known about the mathematical properties of random forests. To date, most studies have concentrated on isolated parts or simplified versions of the procedure. The most celebrated theoretical result is that of Breiman (2001), which offers an upper bound on the generalization error of forests in terms of correlation and strength of the individual trees. This was followed by a technical note [Breiman (2004)] that focuses on a stylized version of the original algorithm. A critical step was subsequently taken by Lin and Jeon (2006), who established lower bounds for nonadaptive forests (i.e., independent of the training set). They also highlighted an interesting connection between random forests and a particular class of nearest neighbor predictors that was further worked out by Biau and Devroye (2010). In recent years, various theoretical studies [e.g., Biau (2012), Biau, Devroye and Lugosi (2008), Genuer (2012), Ishwaran and Kogalur (2010), Zhu, Zeng and Kosorok (2012)] have been performed, analyzing consistency of simplified models, and moving ever closer to practice. Recent attempts toward narrowing the gap between theory and practice are by Denil, Matheson and Freitas (2013), who proves the first consistency result for online random forests, and by Wager (2014) and Mentch and Hooker (2014) who study the asymptotic sampling distribution of forests.

The difficulty in properly analyzing random forests can be explained by the black-box nature of the procedure, which is actually a subtle combination of different components. Among the forest essential ingredients, both bagging [Breiman (1996)] and the classification and regression trees (CART)-split criterion [Breiman et al. (1984)] play a critical role. Bagging (a contraction of bootstrap-aggregating) is a general aggregation scheme which proceeds by generating subsamples from the original data set, constructing a predictor from each resample and deciding by averaging. It is one of the most effective computationally intensive procedures to improve on unstable estimates, especially for large, high-dimensional data sets where finding a good model in one step is impossible because of the complexity and scale of the problem [Bühlmann and Yu (2002), Kleiner et al. (2014), Wager, Hastie and Efron (2014)]. The CART-split selection originated from the most influential CART algorithm of Breiman et al. (1984), and is used in the construction of the individual trees to choose the best cuts perpendicular to the axes. At each node of each tree, the best cut is selected by optimizing the CART-split criterion, based on the notion of Gini impurity (classification) and prediction squared error (regression).

Yet, while bagging and the CART-splitting scheme play a key role in the random forest mechanism, both are difficult to analyze, thereby explaining why theoretical studies have, thus far, considered simplified versions of the original procedure. This is often done by simply ignoring the bagging step and by replacing the CART-split selection with a more elementary cut protocol. Besides, in Breiman's forests, each leaf (i.e., a terminal node) of the individual trees contains a fixed pre-specified number of observations (this parameter, called `nodesize` in the R package `randomForests`, is usually chosen between 1 and 5). There is also an extra parameter in the algorithm which allows one to control the total number of leaves (this parameter is called `maxnode` in the R package and has, by default, no effect on the procedure). The combination of these various components makes the algorithm difficult to analyze with rigorous mathematics. As a matter of fact, most authors focus on simplified, data-independent procedures, thus creating a gap between theory and practice.

Motivated by the above discussion, we study in the present paper some asymptotic properties of Breiman's (2001) algorithm in the context of additive regression models. We prove the $\mathbb{L}^2$ consistency of random forests, which gives a first basic theoretical guarantee of efficiency for this algorithm. To our knowledge, this is the first consistency result for Breiman's (2001) original procedure. Our approach rests upon a detailed analysis of the behavior of the cells generated by CART-split selection as the sample size grows. It turns out that a good control of the regression function variation inside each cell, together with a proper choice of the total number of leaves (Theorem 1) or a proper choice of the subsampling rate (Theorem 2) are sufficient to ensure the forest consistency in a $\mathbb{L}^2$ sense. Also, our analysis shows that random forests can adapt to a sparse framework, when the ambient dimension p is large (independent of n), but only a smaller number of coordinates carry out information.

The paper is organized as follows. In Section 2, we introduce some notation and describe the random forest method. The main asymptotic results are presented in Section 3 and further discussed in Section 4. Section 5 is devoted to the main proofs, and technical results are gathered in the supplemental article [Scornet, Biau and Vert (2015)].

2. Random forests.

The general framework is $\mathbb{L}^2$ regression estimation, in which an input random vector $\mathbf{X} \in [0,1]^p$ is observed, and the goal is to predict the square integrable random response $Y \in \mathbb{R}$ by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. To this end, we assume given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $[0,1]^p \times \mathbb{R}$-valued independent random variables distributed as the independent prototype pair $(\mathbf{X}, Y)$. The objective is to use the data set $\mathcal{D}_n$ to construct an estimate $m_n : [0,1]^p \to \mathbb{R}$ of the function m. In this respect, we say that a regression function estimate m_n is $\mathbb{L}^2$ consistent if $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \to 0$ as $n \to \infty$ (where the expectation is over $\mathbf{X}$ and $\mathcal{D}_n$).

A random forest is a predictor consisting of a collection of M randomized regression trees. For the jth tree in the family, the predicted value at the query point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ and independent of $\mathcal{D}_n$. In practice, this variable is used to resample the training set prior to the growing of individual trees and to select the successive candidate directions for splitting. The trees are combined to form the (finite) forest estimate

$$m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n). \tag{1}$$

Since in practice we can choose M as large as possible, we study in this paper the property of the infinite forest estimate obtained as the limit of (1) when the number of trees M grows to infinity as follows:

$$m_n(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_\Theta[m_n(\mathbf{x}; \Theta, \mathcal{D}_n)],$$

where $\mathbb{E}_\Theta$ denotes expectation with respect to the random parameter Θ, conditional on $\mathcal{D}_n$. This operation is justified by the law of large numbers, which asserts that, almost surely, conditional on $\mathcal{D}_n$,

$$\lim_{M\to\infty} m_{n,M}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = m_n(\mathbf{x}; \mathcal{D}_n);$$

see, for example, Breiman (2001), Scornet (2014) for details. In the sequel, to lighten notation, we will simply write $m_n(\mathbf{x})$ instead of $m_n(\mathbf{x}; \mathcal{D}_n)$.

In Breiman's (2001) original forests, each node of a single tree is associated with a hyper-rectangular cell. At each step of the tree construction, the collection of cells forms a partition of $[0, 1]^p$. The root of the tree is $[0, 1]^p$ itself, and each tree is grown as explained in Algorithm 1.

This algorithm has three parameters:

(1) $m_{\text{try}} \in \{1, \dots, p\}$, which is the number of pre-selected directions for splitting;
(2) $a_n \in \{1, \dots, n\}$, which is the number of sampled data points in each tree;
(3) $t_n \in \{1, \dots, a_n\}$, which is the number of leaves in each tree.

By default, in the original procedure, the parameter m_{try} is set to $p/3$, a_n is set to n (resampling is done with replacement) and $t_n = a_n$. However, in our approach, resampling is done without replacement and the parameters a_n, and t_n can be different from their default values.

In words, the algorithm works by growing M different trees as follows. For each tree, a_n data points are drawn at random without replacement from the original data set; then, at each cell of every tree, a split is chosen by maximizing the CART-criterion (see below); finally, the construction of every tree is stopped when the total number of cells in the tree reaches the value t_n (therefore, each cell contains exactly one point in the case $t_n = a_n$).

We note that the resampling step in Algorithm 1 (line 2) is done by choosing a_n out of n points (with $a_n \leq n$) without replacement. This is slightly different from the original algorithm, where resampling is done by bootstrapping, that is, by choosing n out of n data points with replacement.

Selecting the points "without replacement" instead of "with replacement" is harmless—in fact, it is just a means to avoid mathematical difficulties induced by the bootstrap; see, for example, Efron (1982), Politis, Romano and Wolf (1999).

Algorithm 1: Breiman's random forest predicted value at **x**

```
Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $m_{\mathrm{try}} \in \{1, \dots, p\}$,
    $a_n \in \{1, \dots, n\}$, $t_n \in \{1, \dots, a_n\}$, and $\mathbf{x} \in [0, 1]^p$.
Output: Prediction of the random forest at $\mathbf{x}$.
for $j = 1, \dots, M$ do
    Select $a_n$ points, without replacement, uniformly in $\mathcal{D}_n$.
    Set $\mathcal{P}_0 = \{[0, 1]^p\}$ the partition associated with the root of the tree.
    For all $1 \le \ell \le a_n$, set $\mathcal{P}_\ell = \varnothing$.
    Set $n_{\mathrm{nodes}} = 1$ and level $= 0$.
    while $n_{\mathrm{nodes}} < t_n$ do
        if $\mathcal{P}_{\mathrm{level}} = \varnothing$ then
            level $=$ level $+ 1$
        else
            Let $A$ be the first element in $\mathcal{P}_{\mathrm{level}}$.
            if $A$ contains exactly one point then
                $\mathcal{P}_{\mathrm{level}} \leftarrow \mathcal{P}_{\mathrm{level}} \setminus \{A\}$
                $\mathcal{P}_{\mathrm{level}+1} \leftarrow \mathcal{P}_{\mathrm{level}+1} \cup \{A\}$
            else
                Select uniformly, without replacement, a subset
                $\mathcal{M}_{\mathrm{try}} \subset \{1, \dots, p\}$ of cardinality $m_{\mathrm{try}}$.
                Select the best split in $A$ by optimizing the CART-split
                criterion along the coordinates in $\mathcal{M}_{\mathrm{try}}$ (see details below).
                Cut the cell $A$ according to the best split. Call $A_L$ and $A_R$
                the two resulting cell.
                $\mathcal{P}_{\mathrm{level}} \leftarrow \mathcal{P}_{\mathrm{level}} \setminus \{A\}$
                $\mathcal{P}_{\mathrm{level}+1} \leftarrow \mathcal{P}_{\mathrm{level}+1} \cup \{A_L\} \cup \{A_R\}$
                $n_{\mathrm{nodes}} = n_{\mathrm{nodes}} + 1$
            end
        end
    end
    Compute the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ equal to the average of
    the $Y_i$'s falling in the cell of $\mathbf{x}$ in partition $\mathcal{P}_{\mathrm{level}} \cup \mathcal{P}_{\mathrm{level}+1}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$ at the query
point $\mathbf{x}$ according to (1).
```

On the other hand, letting the parameters a_n and t_n depend upon n offers several degrees of freedom which opens the route for establishing consistency of the method. To be precise, we will study in Section 3 the random forest algorithm in two different regimes. The first regime is when $t_n < a_n$, which means that trees are not fully developed. In this case, a proper tuning of t_n ensures the forest's consistency (Theorem 1). The second regime occurs when $t_n = a_n$, that is, when trees are fully grown. In this case, consistency results from an appropriate choice of the subsample rate a_n/n (Theorem 2).

So far, we have not made explicit the CART-split criterion used in Algorithm 1. To properly define it, we let A be a generic cell and $N_n(A)$ be the number of data points falling in A. A cut in A is a pair (j, z), where j is a dimension in $\{1, \dots, p\}$ and z is the position of the cut along the jth coordinate, within the limits of A. We let $\mathcal{C}_A$ be the set of all such possible cuts in A. Then, with the notation $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(p)})$, for any $(j, z) \in \mathcal{C}_A$, the CART-split criterion [Breiman et al. (1984)] takes the form

$$\begin{aligned} L_n(j,z) = {} & \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\ & - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z})^2 \mathbb{1}_{\mathbf{X}_i \in A}, \end{aligned} \tag{2}$$

where $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$, and $\bar{Y}_A$ (resp., $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) is the average of the Y_i's belonging to A (resp., A_L, A_R), with the convention $0/0 = 0$. At each cell A, the best cut $(j_n^\star, z_n^\star)$ is finally selected by maximizing $L_n(j, z)$ over $\mathcal{M}_{\text{try}}$ and $\mathcal{C}_A$, that is,

$$(j_n^\star, z_n^\star) \in \mathop{\arg\max}_{\substack{j \in \mathcal{M}_{\text{try}} \\ (j,z) \in \mathcal{C}_A}} L_n(j, z).$$

To remove ties in the argmax, the best cut is always performed along the best cut direction $j_n^\star$, at the middle of two consecutive data points.

3. Main results. We consider an additive regression model satisfying the following properties:

(H1) *The response Y follows*

$$Y = \sum_{j=1}^{p} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

where $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(p)})$ is uniformly distributed over $[0, 1]^p$, ε is an independent centered Gaussian noise with finite variance $\sigma^2 > 0$ and each component m_j is continuous.

Additive regression models, which extend linear models, were popularized by Stone (1985) and Hastie and Tibshirani (1986). These models, which decompose the regression function as a sum of univariate functions, are flexible and easy to interpret. They are acknowledged for providing a good trade-off between model complexity and calculation time, and accordingly, have been extensively studied for the last thirty years. Additive models also play an important role in the context of high-dimensional data analysis and sparse modeling, where they are successfully involved in procedures such as the Lasso and various aggregation schemes; for an overview, see, for example, Hastie, Tibshirani and Friedman (2009). Although random forests fall into the family of nonparametric procedures, it turns out that the analysis of their properties is facilitated within the framework of additive models.

Our first result assumes that the total number of leaves t_n in each tree tends to infinity more slowly than the number of selected data points a_n.

THEOREM 1. *Assume that* (H1) *is satisfied. Then, provided* $a_n \to \infty$, $t_n \to \infty$ *and* $t_n(\log a_n)^9/a_n \to 0$, *random forests are consistent, that is,*

$$\lim_{n\to\infty} \mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 = 0.$$

It is noteworthy that Theorem 1 still holds with $a_n = n$. In this case, the subsampling step plays no role in the consistency of the method. Indeed, controlling the depth of the trees via the parameter t_n is sufficient to bound the forest error. We note in passing that an easy adaptation of Theorem 1 shows that the CART algorithm is consistent under the same assumptions.

The term $(\log a_n)^9$ originates from the Gaussian noise and allows us to control the noise tail. In the easier situation where the Gaussian noise is replaced by a bounded random variable, it is easy to see that the term $(\log a_n)^9$ turns into $\log a_n$, a term which accounts for the complexity of the tree partition.

Let us now examine the forest behavior in the second regime, where $t_n = a_n$ (i.e., trees are fully grown), and as before, subsampling is done at the rate a_n/n. The analysis of this regime turns out to be more complicated, and rests upon assumption (H2) below. We denote by $Z_i = \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}$ the indicator that $\mathbf{X}_i$ falls into the same cell as $\mathbf{X}$ in the random tree designed with $\mathcal{D}_n$ and the random parameter Θ. Similarly, we let $Z'_j = \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}$, where Θ' is an independent copy of Θ. Accordingly, we define

$$\psi_{i,j}(Y_i, Y_j) = \mathbb{E}[Z_i Z'_j | \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \dots, \mathbf{X}_n, Y_i, Y_j]$$

and

$$\psi_{i,j} = \mathbb{E}[Z_i Z'_j | \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \dots, \mathbf{X}_n].$$

Finally, for any random variables W_1, W_2, Z, we denote by $\operatorname{Corr}(W_1, W_2|Z)$ the conditional correlation coefficient (whenever it exists).

(H2) *Let* $Z_{i,j} = (Z_i, Z'_j)$. *Then one of the following two conditions holds*:

(H2.1) *One has*

$$\lim_{n\to\infty}(\log a_n)^{2p-2}(\log n)^2\mathbb{E}\Big[\max_{\substack{i,j\\ i\neq j}}|\psi_{i,j}(Y_i,Y_j)-\psi_{i,j}|\Big]^2=0.$$

(H2.2) *There exist a constant $C>0$ and a sequence $(\gamma_n)_n\to 0$ such that, almost surely*,

$$\max_{\ell_1,\ell_2=0,1}\frac{|\operatorname{Corr}(Y_i-m(\mathbf{X}_i),\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i,\mathbf{X}_j,Y_j)|}{\mathbb{P}^{1/2}[Z_{i,j}=(\ell_1,\ell_2)|\mathbf{X}_i,\mathbf{X}_j,Y_j]}\leq\gamma_n$$

and

$$\max_{\ell_1=0,1}\frac{|\operatorname{Corr}((Y_i-m(\mathbf{X}_i))^2,\mathbb{1}_{Z_i=\ell_1}|\mathbf{X}_i)|}{\mathbb{P}^{1/2}[Z_i=\ell_1|\mathbf{X}_i]}\leq C.$$

Despite their technical aspect, statements (H2.1) and (H2.2) have simple interpretations. To understand the meaning of (H2.1), let us replace the Gaussian noise by a bounded random variable. A close inspection of Lemma 4 shows that (H2.1) may be simply replaced by

$$\lim_{n\to\infty}\mathbb{E}\Big[\max_{\substack{i,j\\ i\neq j}}|\psi_{i,j}(Y_i,Y_j)-\psi_{i,j}|\Big]^2=0.$$

Therefore, (H2.1) means that the influence of two Y-values on the probability of connection of two couples of random points tends to zero as $n\to\infty$.

As for assumption (H2.2), it holds whenever the correlation between the noise and the probability of connection of two couples of random points vanishes quickly enough, as $n\to\infty$. Note that, in the simple case where the partition is independent of the Y_i's, the correlations in (H2.2) are zero, so that (H2) is trivially satisfied. This is also verified in the noiseless case, that is, when $Y=m(\mathbf{X})$. However, in the most general context, the partitions strongly depend on the whole sample $\mathcal{D}_n$, and unfortunately, we do not know whether or not (H2) is satisfied.

THEOREM 2. *Assume that* (H1) *and* (H2) *are satisfied, and let $t_n=a_n$. Then, provided $a_n\to\infty$, $t_n\to\infty$ and $a_n\log n/n\to 0$, random forests are consistent, that is*,

$$\lim_{n\to\infty}\mathbb{E}[m_n(\mathbf{X})-m(\mathbf{X})]^2=0.$$

To our knowledge, apart from the fact that bootstrapping is replaced by subsampling, Theorems 1 and 2 are the first consistency results for Breiman's (2001) forests. Indeed, most models studied so far are designed independently of $\mathcal{D}_n$ and are, consequently, an unrealistic representation of the true procedure. In fact, understanding Breiman's random forest behavior deserves a more involved mathematical treatment. Section 4 below offers a thorough description of the various mathematical forces in action.

Our study also sheds some interesting light on the behavior of forests when the ambient dimension p is large but the true underlying dimension of the model is small. To see how, assume that the additive model (H1) satisfies a sparsity constraint of the form

$$Y = \sum_{j=1}^{S} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

where $S < p$ represents the true, but unknown, dimension of the model. Thus, among the p original features, it is assumed that only the first (without loss of generality) S variables are informative. Put differently, Y is assumed to be independent of the last $(p - S)$ variables. In this dimension reduction context, the ambient dimension p can be very large, but we believe that the representation is sparse, that is, that few components of m are nonzero. As such, the value S characterizes the sparsity of the model: the smaller S, the sparser m.

Proposition 1 below shows that random forests nicely adapt to the sparsity setting by asymptotically performing, with high probability, splits along the S informative variables.

In this proposition, we set $m_{\mathrm{try}} = p$ and, for all k, we denote by $j_{1,n}(\mathbf{X}), \ldots, j_{k,n}(\mathbf{X})$ the first k cut directions used to construct the cell containing $\mathbf{X}$, with the convention that $j_{q,n}(\mathbf{X}) = \infty$ if the cell has been cut strictly less than q times.

PROPOSITION 1. *Assume that* (H1) *is satisfied. Let* $k \in \mathbb{N}^\star$ *and* $\xi > 0$. *Assume that there is no interval* $[a, b]$ *and no* $j \in \{1, \ldots, S\}$ *such that* m_j *is constant on* $[a, b]$. *Then, with probability* $1 - \xi$, *for all* n *large enough, we have, for all* $1 \leq q \leq k$,

$$j_{q,n}(\mathbf{X}) \in \{1, \ldots, S\}.$$

This proposition provides an interesting perspective on why random forests are still able to do a good job in a sparse framework. Since the algorithm selects splits mostly along informative variables, everything happens as if data were projected onto the vector space generated by the S informative variables. Therefore, forests are likely to only depend upon these S variables, which supports the fact that they have good performance in sparse framework.

It remains that a substantial research effort is still needed to understand the properties of forests in a high-dimensional setting, when $p = p_n$ may be substantially larger than the sample size. Unfortunately, our analysis does not carry over to this context. In particular, if high-dimensionality is modeled by letting $p_n \to \infty$, then assumption (H2.1) may be too restrictive since the term $(\log a_n)^{2p-2}$ will diverge at a fast rate.

4. Discussion.

One of the main difficulties in assessing the mathematical properties of Breiman's (2001) forests is that the construction process of the individual trees strongly depends on both the X_i's and the Y_i's. For partitions that are independent of the Y_i's, consistency can be shown by relatively simple means via Stone's (1977) theorem for local averaging estimates; see also Györfi et al. (2002), Chapter 6. However, our partitions and trees depend upon the Y-values in the data. This makes things complicated, but mathematically interesting too. Thus, logically, the proof of Theorem 2 starts with an adaptation of Stone's (1977) theorem tailored for random forests, whereas the proof of Theorem 1 is based on consistency results of data-dependent partitions developed by Nobel (1996).

Both theorems rely on Proposition 2 below, which stresses an important feature of the random forest mechanism. It states that the variation of the regression function m within a cell of a random tree is small provided n is large enough. To this end, we define, for any cell A, the variation of m within A as

$$\Delta(m, A) = \sup_{\mathbf{x},\mathbf{x}' \in A} \left|m(\mathbf{x}) - m(\mathbf{x}')\right|.$$

Furthermore, we denote by $A_n(\mathbf{X}, \Theta)$ the cell of a tree built with random parameter Θ that contains the point $\mathbf{X}$.

PROPOSITION 2. *Assume that* (H1) *holds. Then, for all* $\rho, \xi > 0$, *there exists* $N \in \mathbb{N}^\star$ *such that, for all* $n > N$,

$$\mathbb{P}\left[\Delta\left(m, A_n(\mathbf{X}, \Theta)\right) \leq \xi\right] \geq 1 - \rho.$$

It should be noted that in the standard, Y-independent analysis of partitioning regression function estimates, the variance is controlled by letting the diameters of the tree cells tend to zero in probability. Instead of such a geometrical assumption, Proposition 2 ensures that the variation of m inside a cell is small, thereby forcing the approximation error of the forest to asymptotically approach zero.

While Proposition 2 offers a good control of the approximation error of the forest in both regimes, a separate analysis is required for the estimation error. In regime 1 (Theorem 1), the parameter t_n allows us to control the structure of the tree. This is in line with standard tree consistency approaches; see, for example, Devroye, Györfi and Lugosi (1996), Chapter 20. Things are different for the second regime (Theorem 2), in which individual trees are fully grown. In this case, the estimation error is controlled by forcing the subsampling rate a_n/n to be $\mathrm{o}(1/\log n)$, which is a more unusual requirement and deserves some remarks.

At first, we note that the $\log n$ term in Theorem 2 is used to control the Gaussian noise ε. Thus if the noise is assumed to be a bounded random variable, then the $\log n$ term disappears, and the condition reduces to $a_n/n \to 0$. The requirement $a_n \log n/n \to 0$ guarantees that every single observation $(\mathbf{X}_i, Y_i)$ is used in the tree construction with a probability that becomes small with n. It also implies that the query point $\mathbf{x}$ is not connected to the same data point in a high proportion of trees. If not, the predicted value at $\mathbf{x}$ would be influenced too much by one single pair $(\mathbf{X}_i, Y_i)$, making the forest inconsistent. In fact, the proof of Theorem 2 reveals that the estimation error of a forest estimate is small as soon as the maximum probability of connection between the query point and all observations is small. Thus the assumption on the subsampling rate is just a convenient way to control these probabilities, by ensuring that partitions are dissimilar enough (i.e., by ensuring that $\mathbf{x}$ is connected with many data points through the forest). This idea of diversity among trees was introduced by Breiman (2001), but is generally difficult to analyze. In our approach, the subsampling is the key component for imposing tree diversity.

Theorem 2 comes at the price of assumption (H2), for which we do not know if it is valid in all generality. On the other hand, Theorem 2, which mimics almost perfectly the algorithm used in practice, is an important step toward understanding Breiman's random forests. Contrary to most previous works, Theorem 2 assumes that there is only one observation per leaf of each individual tree. This implies that the single trees are eventually not consistent, since standard conditions for tree consistency require that the number of observations in the terminal nodes tends to infinity as n grows; see, for example, Devroye, Györfi and Lugosi (1996), Györfi et al. (2002). Thus the random forest algorithm aggregates rough individual tree predictors to build a provably consistent general architecture.

It is also interesting to note that our results (in particular Lemma 3) cannot be directly extended to establish the pointwise consistency of random forests; that is, for almost all $\mathbf{x} \in [0, 1]^d$,

$$\lim_{n\to\infty} \mathbb{E}\big[m_n(\mathbf{x}) - m(\mathbf{x})\big]^2 = 0.$$

Fixing $\mathbf{x} \in [0, 1]^d$, the difficulty results from the fact that we do not have a control on the diameter of the cell $A_n(\mathbf{x}, \Theta)$, whereas, since the cells form a partition of $[0, 1]^d$, we have a global control on their diameters. Thus, as highlighted by Wager (2014), random forests can be inconsistent at some fixed point $\mathbf{x} \in [0, 1]^d$, particularly near the edges, while being $\mathbb{L}^2$ consistent.

Let us finally mention that all results can be extended to the case where ε is a heteroscedastic and sub-Gaussian noise, with for all $\mathbf{x} \in [0, 1]^d$, $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \le \sigma'^2$, for some constant σ'^2. All proofs can be readily extended to match this context, at the price of easy technical adaptations.

5. Proof of Theorems 1 and 2.

For the sake of clarity, proofs of the intermediary results are gathered in the supplemental article [Scornet, Biau and Vert (2015)]. We start with some notation.

5.1. *Notation.*

In the sequel, to clarify the notation, we will sometimes write $d = (d^{(1)}, d^{(2)})$ to represent a cut (j, z).

Recall that, for any cell A, $\mathcal{C}_A$ is the set of all possible cuts in A. Thus, with this notation, $\mathcal{C}_{[0,1]^p}$ is just the set of all possible cuts at the root of the tree, that is, all possible choices $d = (d^{(1)}, d^{(2)})$ with $d^{(1)} \in \{1, \ldots, p\}$ and $d^{(2)} \in [0, 1]$.

More generally, for any $\mathbf{x} \in [0, 1]^p$, we call $\mathcal{A}_k(\mathbf{x})$ the collection of all possible $k \geq 1$ consecutive cuts used to build the cell containing $\mathbf{x}$. Such a cell is obtained after a sequence of cuts $\mathbf{d}_k = (d_1, \ldots, d_k)$, where the dependency of $\mathbf{d}_k$ upon $\mathbf{x}$ is understood. Accordingly, for any $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$, we let $A(\mathbf{x}, \mathbf{d}_k)$ be the cell containing $\mathbf{x}$ built with the particular k-tuple of cuts $\mathbf{d}_k$. The proximity between two elements $\mathbf{d}_k$ and $\mathbf{d}'_k$ in $\mathcal{A}_k(\mathbf{x})$ will be measured via

$$\|\mathbf{d}_k - \mathbf{d}'_k\|_\infty = \sup_{1 \leq j \leq k} \max\bigl(|d_j^{(1)} - d_j'^{(1)}|, |d_j^{(2)} - d_j'^{(2)}|\bigr).$$

Accordingly, the distance d_∞ between $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$ and any $\mathcal{A} \subset \mathcal{A}_k(\mathbf{x})$ is

$$d_\infty(\mathbf{d}_k, \mathcal{A}) = \inf_{\mathbf{z} \in \mathcal{A}} \|\mathbf{d}_k - \mathbf{z}\|_\infty.$$

Remember that $A_n(\mathbf{X}, \Theta)$ denotes the cell of a tree containing $\mathbf{X}$ and designed with random parameter Θ. Similarly, $A_{k,n}(\mathbf{X}, \Theta)$ is the same cell but where only the first k cuts are performed ($k \in \mathbb{N}^\star$ is a parameter to be chosen later). We also denote by $\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta) = (\hat{d}_{1,n}(\mathbf{X}, \Theta), \ldots, \hat{d}_{k,n}(\mathbf{X}, \Theta))$ the k cuts used to construct the cell $A_{k,n}(\mathbf{X}, \Theta)$.

Recall that, for any cell A, the empirical criterion used to split A in the random forest algorithm is defined in (2). For any cut $(j, z) \in \mathcal{C}_A$, we denote the following theoretical version of $L_n(\cdot, \cdot)$ by

$$\begin{aligned} L^\star(j, z) = {} & \mathbb{V}[Y | \mathbf{X} \in A] - \mathbb{P}\bigl[\mathbf{X}^{(j)} < z | \mathbf{X} \in A\bigr] \mathbb{V}\bigl[Y | \mathbf{X}^{(j)} < z, \mathbf{X} \in A\bigr] \\ & - \mathbb{P}\bigl[\mathbf{X}^{(j)} \geq z | \mathbf{X} \in A\bigr] \mathbb{V}\bigl[Y | \mathbf{X}^{(j)} \geq z, \mathbf{X} \in A\bigr]. \end{aligned}$$

Observe that $L^\star(\cdot, \cdot)$ does not depend upon the training set and that, by the strong law of large numbers, $L_n(j, z) \to L^\star(j, z)$ almost surely as $n \to \infty$ for all cuts $(j, z) \in \mathcal{C}_A$. Therefore, it is natural to define the best theoretical split $(j^\star, z^\star)$ of the cell A as

$$\bigl(j^\star, z^\star\bigr) \in \mathop{\arg\min}_{\substack{(j,z) \in \mathcal{C}_A \\ j \in \mathcal{M}_{\text{try}}}} L^\star(j, z).$$

In view of this criterion, we define the theoretical random forest as before, but with consecutive cuts performed by optimizing $L^\star(\cdot, \cdot)$ instead of $L_n(\cdot, \cdot)$. We note that this new forest does depend on Θ through $\mathcal{M}_{\text{try}}$, but not on the sample $\mathcal{D}_n$. In particular, the stopping criterion for dividing cells has to be changed in the theoretical random forest; instead of stopping when a cell has a single training point, we impose that each tree of the theoretical forest is stopped at a fixed level $k \in \mathbb{N}^\star$. We also let $A_k^\star(\mathbf{X}, \Theta)$ be a cell of the theoretical random tree at level k, containing $\mathbf{X}$, designed with randomness Θ, and resulting from the k theoretical cuts $\mathbf{d}_k^\star(\mathbf{X},\Theta) = (d_1^\star(\mathbf{X},\Theta),\ldots,d_k^\star(\mathbf{X},\Theta))$. Since there can exist multiple best cuts at, at least, one node, we call $\mathcal{A}_k^\star(\mathbf{X},\Theta)$ the set of all k-tuples $\mathbf{d}_k^\star(\mathbf{X},\Theta)$ of best theoretical cuts used to build $A_k^\star(\mathbf{X},\Theta)$.

We are now equipped to prove Proposition 2. For reasons of clarity, the proof has been divided in three steps. First, we study in Lemma 1 the theoretical random forest. Then we prove in Lemma 3 (via Lemma 2) that theoretical and empirical cuts are close to each other. Proposition 2 is finally established as a consequence of Lemma 1 and Lemma 3. Proofs of these lemmas are to be found in the supplemental article [Scornet, Biau and Vert (2015)].

5.2. *Proof of Proposition* 2. We first need a lemma which states that the variation of $m(\mathbf{X})$ within the cell $A_k^\star(\mathbf{X},\Theta)$ where $\mathbf{X}$ falls, as measured by $\Delta(m, A_k^\star(\mathbf{X},\Theta))$, tends to zero.

LEMMA 1. *Assume that* (H1) *is satisfied. Then, for all* $\mathbf{x}\in[0,1]^p$,

$$\Delta\big(m, A_k^\star(\mathbf{x},\Theta)\big) \to 0 \qquad \textit{almost surely, as } k\to\infty.$$

The next step is to show that cuts in theoretical and original forests are close to each other. To this end, for any $\mathbf{x}\in[0,1]^p$ and any k-tuple of cuts $\mathbf{d}_k\in\mathcal{A}_k(\mathbf{x})$, we define

$$\begin{aligned} L_{n,k}(\mathbf{x},\mathbf{d}_k) &= \frac{1}{N_n(A(\mathbf{x},\mathbf{d}_{k-1}))}\sum_{i=1}^n (Y_i - \bar{Y}_{A(\mathbf{x},\mathbf{d}_{k-1})})^2 \mathbb{1}_{\mathbf{X}_i\in A(\mathbf{x},\mathbf{d}_{k-1})} \\ &\quad - \frac{1}{N_n(A(\mathbf{x},\mathbf{d}_{k-1}))}\sum_{i=1}^n (Y_i - \bar{Y}_{A_L(\mathbf{x},\mathbf{d}_{k-1})}\mathbb{1}_{\mathbf{X}_i^{(d_k^{(1)})}<d_k^{(2)}} \\ &\qquad\qquad - \bar{Y}_{A_R(\mathbf{x},\mathbf{d}_{k-1})}\mathbb{1}_{\mathbf{X}_i^{(d_k^{(1)})}\ge d_k^{(2)}})^2 \mathbb{1}_{\mathbf{X}_i\in A(\mathbf{x},\mathbf{d}_{k-1})}, \end{aligned}$$

where $A_L(\mathbf{x},\mathbf{d}_{k-1}) = A(\mathbf{x},\mathbf{d}_{k-1})\cap\{\mathbf{z}:\mathbf{z}^{(d_k^{(1)})}<d_k^{(2)}\}$ and $A_R(\mathbf{x},\mathbf{d}_{k-1}) = A(\mathbf{x},\mathbf{d}_{k-1})\cap\{\mathbf{z}:\mathbf{z}^{(d_k^{(1)})}\ge d_k^{(2)}\}$, and where we use the convention $0/0=0$ when $A(\mathbf{x},\mathbf{d}_{k-1})$ is empty. Besides, we let $A(\mathbf{x},\mathbf{d}_0)=[0,1]^p$ in the previous equation. The quantity $L_{n,k}(\mathbf{x},\mathbf{d}_k)$ is nothing but the criterion to maximize in d_k to find the best kth cut in the cell $A(\mathbf{x},\mathbf{d}_{k-1})$. Lemma 2 below ensures that $L_{n,k}(\mathbf{x},\cdot)$ is stochastically equicontinuous, for all $\mathbf{x}\in[0,1]^p$. To this end, for all $\xi>0$, and for all $\mathbf{x}\in[0,1]^p$, we denote by $\mathcal{A}_{k-1}^\xi(\mathbf{x})\subset\mathcal{A}_{k-1}(\mathbf{x})$ the set of all $(k-1)$-tuples $\mathbf{d}_{k-1}$ such that the cell $A(\mathbf{x},\mathbf{d}_{k-1})$ contains a hypercube of edge length ξ. Moreover, we let $\bar{\mathcal{A}}_k^\xi(\mathbf{x}) = \{\mathbf{d}_k : \mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^\xi(\mathbf{x})\}$ equipped with the norm $\|\mathbf{d}_k\|_\infty$.

LEMMA 2. *Assume that* (H1) *is satisfied. Fix* $\mathbf{x}\in[0,1]^p$, $k\in\mathbb{N}^\star$, *and let* $\xi>0$. *Then* $L_{n,k}(\mathbf{x},\cdot)$ *is stochastically equicontinuous on* $\bar{\mathcal{A}}_k^\xi(\mathbf{x})$; *that is, for all*

$\alpha, \rho > 0$, *there exists* $\delta > 0$ *such that*

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{\substack{\|\mathbf{d}_k - \mathbf{d}'_k\|_\infty \le \delta \\ \mathbf{d}_k, \mathbf{d}'_k \in \bar{\mathcal{A}}_k^{\xi}(\mathbf{x})}} \big|L_{n,k}(\mathbf{x}, \mathbf{d}_k) - L_{n,k}(\mathbf{x}, \mathbf{d}'_k)\big| > \alpha \Big] \le \rho.$$

Lemma 2 is then used in Lemma 3 to assess the distance between theoretical and empirical cuts.

LEMMA 3. *Assume that* (H1) *is satisfied. Fix* $\xi, \rho > 0$ *and* $k \in \mathbb{N}^\star$. *Then there exists* $N \in \mathbb{N}^\star$ *such that, for all* $n \ge N$,

$$\mathbb{P}\big[d_\infty\big(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)\big) \le \xi\big] \ge 1 - \rho.$$

We are now ready to prove Proposition 2. Fix $\rho, \xi > 0$. Since almost sure convergence implies convergence in probability, according to Lemma 1, there exists $k_0 \in \mathbb{N}^\star$ such that

$$\mathbb{P}\big[\Delta\big(m, A_{k_0}^\star(\mathbf{X}, \Theta)\big) \le \xi\big] \ge 1 - \rho. \tag{3}$$

By Lemma 3, for all $\xi_1 > 0$, there exists $N \in \mathbb{N}^\star$ such that, for all $n \ge N$,

$$\mathbb{P}\big[d_\infty\big(\hat{\mathbf{d}}_{k_0,n}(\mathbf{X}, \Theta), \mathcal{A}_{k_0}^\star(\mathbf{X}, \Theta)\big) \le \xi_1\big] \ge 1 - \rho. \tag{4}$$

Since m is uniformly continuous, we can choose ξ_1 sufficiently small such that, for all $\mathbf{x} \in [0, 1]^p$, for all $\mathbf{d}_{k_0}, \mathbf{d}'_{k_0}$ satisfying $d_\infty(\mathbf{d}_{k_0}, \mathbf{d}'_{k_0}) \le \xi_1$, we have

$$\big|\Delta\big(m, A(\mathbf{x}, \mathbf{d}_{k_0})\big) - \Delta\big(m, A(\mathbf{x}, \mathbf{d}'_{k_0})\big)\big| \le \xi. \tag{5}$$

Thus, combining inequalities (4) and (5), we obtain

$$\mathbb{P}\big[\big|\Delta\big(m, A_{k_0,n}(\mathbf{X}, \Theta)\big) - \Delta\big(m, A_{k_0}^\star(\mathbf{X}, \Theta)\big)\big| \le \xi\big] \ge 1 - \rho. \tag{6}$$

Using the fact that $\Delta(m, A) \le \Delta(m, A')$ whenever $A \subset A'$, we deduce from (3) and (6) that, for all $n \ge N$,

$$\mathbb{P}\big[\Delta\big(m, A_n(\mathbf{X}, \Theta)\big) \le 2\xi\big] \ge 1 - 2\rho.$$

This completes the proof of Proposition 2.

5.3. *Proof of Theorem* 1. We still need some additional notation. The partition obtained with the random variable Θ and the data set $\mathcal{D}_n$ is denoted by $\mathcal{P}_n(\mathcal{D}_n, \Theta)$, which we abbreviate as $\mathcal{P}_n(\Theta)$. We let

$$\Pi_n(\Theta) = \big\{\mathcal{P}\big((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n), \Theta\big) : (\mathbf{x}_i, y_i) \in [0, 1]^d \times \mathbb{R}\big\}$$

be the family of all achievable partitions with random parameter Θ. Accordingly, we let

$$M\big(\Pi_n(\Theta)\big) = \max\big\{\mathrm{Card}(\mathcal{P}) : \mathcal{P} \in \Pi_n(\Theta)\big\}$$

be the maximal number of terminal nodes among all partitions in $\Pi_n(\Theta)$. Given a set $\mathbf{z}_1^n = \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \subset [0,1]^d$, $\Gamma(\mathbf{z}_1^n, \Pi_n(\Theta))$ denotes the number of distinct partitions of $\mathbf{z}_1^n$ induced by elements of $\Pi_n(\Theta)$, that is, the number of different partitions $\{\mathbf{z}_1^n \cap A : A \in \mathcal{P}\}$ of $\mathbf{z}_1^n$, for $\mathcal{P} \in \Pi_n(\Theta)$. Consequently, the partitioning number $\Gamma_n(\Pi_n(\Theta))$ is defined by

$$\Gamma_n\big(\Pi_n(\Theta)\big) = \max\big\{\Gamma\big(\mathbf{z}_1^n, \Pi_n(\Theta)\big) : \mathbf{z}_1, \dots, \mathbf{z}_n \in [0,1]^d\big\}.$$

Let $(\beta_n)_n$ be a positive sequence, and define the truncated operator T_{β_n} by

$$\begin{cases} T_{\beta_n} u = u, & \text{if } |u| < \beta_n, \\ T_{\beta_n} u = \operatorname{sign}(u)\beta_n, & \text{if } |u| \geq \beta_n. \end{cases}$$

Hence $T_{\beta_n} m_n(\mathbf{X}, \Theta)$, $Y_L = T_L Y$ and $Y_{i,L} = T_L Y_i$ are defined unambiguously. We let $\mathcal{F}_n(\Theta)$ be the set of all functions $f : [0,1]^d \to \mathbb{R}$ piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$. [Notice that $\mathcal{F}_n(\Theta)$ depends on the whole data set.] Finally, we denote by $\mathcal{I}_{n,\Theta}$ the set of indices of the data points that are selected during the subsampling step. Thus the tree estimate $m_n(\mathbf{x}, \Theta)$ satisfies

$$m_n(\cdot, \Theta) \in \mathop{\arg\min}_{f \in \mathcal{F}_n(\Theta)} \frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} \big|f(\mathbf{X}_i) - Y_i\big|^2.$$

The proof of Theorem 1 is based on ideas developed by Nobel (1996), and worked out in Theorem 10.2 in Györfi et al. (2002). This theorem, tailored for our context, is recalled below for the sake of completeness.

THEOREM 3 [Györfi et al. (2002)]. *Let m_n and $\mathcal{F}_n(\Theta)$ be as above. Assume that*:

(i) $\lim_{n\to\infty} \beta_n = \infty$;
(ii) $\lim_{n\to\infty} \mathbb{E}[\inf_{f \in \mathcal{F}_n(\Theta), \|f\|_\infty \leq \beta_n} \mathbb{E}_{\mathbf{X}}[f(\mathbf{X}) - m(\mathbf{X})]^2] = 0$;
(iii) *for all $L > 0$*,

$$\lim_{n\to\infty} \mathbb{E}\Bigg[\sup_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \leq \beta_n}} \bigg|\frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} \big[f(\mathbf{X}_i) - Y_{i,L}\big]^2 - \mathbb{E}\big[f(\mathbf{X}) - Y_L\big]^2\bigg|\Bigg] = 0.$$

Then

$$\lim_{n\to\infty} \mathbb{E}\big[T_{\beta_n} m_n(\mathbf{X}, \Theta) - m(\mathbf{X})\big]^2 = 0.$$

Statement (ii) [resp., statement (iii)] allows us to control the approximation error (resp., the estimation error) of the truncated estimate. Since the truncated estimate $T_{\beta_n} m_n$ is piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$, $T_{\beta_n} m_n$ belongs to the set $\mathcal{F}_n(\Theta)$. Thus the term in (ii) is the classical approximation error.

We are now equipped to prove Theorem 1. Fix $\xi > 0$, and note that we just have to check statements (i)–(iii) of Theorem 3 to prove that the truncated estimate of the random forest is consistent. Throughout the proof, we let $\beta_n = \|m\|_\infty + \sigma\sqrt{2}(\log a_n)^2$. Clearly, statement (i) is true.

Approximation error. To prove (ii), let

$$f_{n,\Theta} = \sum_{A \in \mathcal{P}_n(\Theta)} m(\mathbf{z}_A) \mathbb{1}_A,$$

where $\mathbf{z}_A \in A$ is an arbitrary point picked in cell A. Since, according to (H1), $\|m\|_\infty < \infty$, for all n large enough such that $\beta_n > \|m\|_\infty$, we have

$$\begin{aligned}
\mathbb{E} \inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \beta_n}} \mathbb{E}_{\mathbf{X}} [f(\mathbf{X}) - m(\mathbf{X})]^2 &\le \mathbb{E} \inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \|m\|_\infty}} \mathbb{E}_{\mathbf{X}} [f(\mathbf{X}) - m(\mathbf{X})]^2 \\
&\le \mathbb{E}[f_{\Theta,n}(\mathbf{X}) - m(\mathbf{X})]^2 \\
&\quad \bigl(\text{since } f_{\Theta,n} \in \mathcal{F}_n(\Theta)\bigr) \\
&\le \mathbb{E}[m(\mathbf{z}_{A_n(\mathbf{X},\Theta)}) - m(\mathbf{X})]^2 \\
&\le \mathbb{E}[\Delta(m, A_n(\mathbf{X}, \Theta))]^2 \\
&\le \xi^2 + 4\|m\|_\infty^2 \mathbb{P}[\Delta(m, A_n(\mathbf{X}, \Theta)) > \xi].
\end{aligned}$$

Thus, using Proposition 2, we see that for all n large enough,

$$\mathbb{E} \inf_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \beta_n}} \mathbb{E}_{\mathbf{X}} [f(\mathbf{X}) - m(\mathbf{X})]^2 \le 2\xi^2.$$

This establishes (ii).

Estimation error. To prove statement (iii), fix $L > 0$. Then, for all n large enough such that $L < \beta_n$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X},\mathcal{D}_n}\Biggl(\sup_{\substack{f \in \mathcal{F}_n(\Theta) \\ \|f\|_\infty \le \beta_n}} \biggl| \frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} [f(\mathbf{X}_i) - Y_{i,L}]^2 - \mathbb{E}[f(\mathbf{X}) - Y_L]^2 \biggr| > \xi \Biggr) \\
&\quad \le 8 \exp\biggl[\log \Gamma_n(\Pi_n(\Theta)) + 2M(\Pi_n(\Theta)) \log\Bigl(\frac{333e\beta_n^2}{\xi}\Bigr) - \frac{a_n \xi^2}{2048 \beta_n^4} \biggr] \\
&\qquad [\text{according to Theorem 9.1 in Györfi et al. (2002)}] \\
&\quad \le 8 \exp\biggl[-\frac{a_n}{\beta_n^4} \Bigl(\frac{\xi^2}{2048} - \frac{\beta_n^4 \log \Gamma_n(\Pi_n)}{a_n} - \frac{2\beta_n^4 M(\Pi_n)}{a_n} \log\Bigl(\frac{333e\beta_n^2}{\xi}\Bigr)\Bigr)\biggr].
\end{aligned}$$

Since each tree has exactly t_n terminal nodes, we have $M(\Pi_n(\Theta)) = t_n$, and simple calculations show that

$$\Gamma_n(\Pi_n(\Theta)) \le (da_n)^{t_n}.$$

Hence

$$\begin{aligned}
&\mathbb{P}\Bigl(\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\Bigl|\frac{1}{a_n}\sum_{i\in\mathcal{I}_{n,\Theta}}[f(\mathbf{X}_i)-Y_{i,L}]^2-\mathbb{E}[f(\mathbf{X})-Y_L]^2\Bigr|>\xi\Bigr)\\
&\quad\le 8\exp\Bigl(-\frac{a_nC_{\xi,n}}{\beta_n^4}\Bigr),
\end{aligned}$$

where

$$\begin{aligned}
C_{\xi,n} &= \frac{\xi^2}{2048}-4\sigma^4\frac{t_n(\log(da_n))^9}{a_n}-8\sigma^4\frac{t_n(\log a_n)^8}{a_n}\log\Bigl(\frac{666e\sigma^2(\log a_n)^4}{\xi}\Bigr)\\
&\to \frac{\xi^2}{2048}\qquad \text{as } n\to\infty,
\end{aligned}$$

by our assumption. Finally, observe that

$$\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\Bigl|\frac{1}{a_n}\sum_{i\in\mathcal{I}_{n,\Theta}}[f(\mathbf{X}_i)-Y_{i,L}]^2-\mathbb{E}[f(\mathbf{X})-Y_L]^2\Bigr|\le 2(\beta_n+L)^2,$$

which yields, for all n large enough,

$$\begin{aligned}
&\mathbb{E}\Bigl[\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\Bigl|\frac{1}{a_n}\sum_{i=1}^{a_n}[f(\mathbf{X}_i)-Y_{i,L}]^2-\mathbb{E}[f(\mathbf{X})-Y_L]^2\Bigr|\Bigr]\\
&\quad\le \xi+2(\beta_n+L)^2\mathbb{P}\Bigl[\sup_{\substack{f\in\mathcal{F}_n(\Theta)\\ \|f\|_\infty\le\beta_n}}\Bigl|\frac{1}{a_n}\sum_{i=1}^{a_n}[f(\mathbf{X}_i)-Y_{i,L}]^2-\mathbb{E}[f(\mathbf{X})-Y_L]^2\Bigr|>\xi\Bigr]\\
&\quad\le \xi+16(\beta_n+L)^2\exp\Bigl(-\frac{a_nC_{\xi,n}}{\beta_n^4}\Bigr)\\
&\quad\le 2\xi.
\end{aligned}$$

Thus, according to Theorem 3,

$$\mathbb{E}[T_{\beta_n}m_n(\mathbf{X},\Theta)-m(\mathbf{X})]^2\to 0.$$

Untruncated estimate. It remains to show the consistency of the nontruncated random forest estimate, and the proof will be complete. For this purpose, note that, for all n large enough,

$$\begin{aligned}
\mathbb{E}[m_n(\mathbf{X})-m(\mathbf{X})]^2 &= \mathbb{E}[\mathbb{E}_\Theta[m_n(\mathbf{X},\Theta)]-m(\mathbf{X})]^2\\
&\le \mathbb{E}[m_n(\mathbf{X},\Theta)-m(\mathbf{X})]^2\\
&\qquad \text{(by Jensen's inequality)}
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}\big[m_n(\mathbf{X},\Theta) - T_{\beta_n} m_n(\mathbf{X},\Theta)\big]^2 \\
&\quad + \mathbb{E}\big[T_{\beta_n} m_n(\mathbf{X},\Theta) - m(\mathbf{X})\big]^2 \\
&\leq \mathbb{E}\big[\big[m_n(\mathbf{X},\Theta) - T_{\beta_n} m_n(\mathbf{X},\Theta)\big]^2 \mathbb{1}_{m_n(\mathbf{X},\Theta)\geq\beta_n}\big] + \xi \\
&\leq \mathbb{E}\big[m_n^2(\mathbf{X},\Theta)\mathbb{1}_{m_n(\mathbf{X},\Theta)\geq\beta_n}\big] + \xi \\
&\leq \mathbb{E}\big[\mathbb{E}\big[m_n^2(\mathbf{X},\Theta)\mathbb{1}_{m_n(\mathbf{X},\Theta)\geq\beta_n}|\Theta\big]\big] + \xi.
\end{aligned}$$

Since $|m_n(\mathbf{X},\Theta)| \leq \|m\|_\infty + \max_{1\leq i\leq n}|\varepsilon_i|$, we have

$$\begin{aligned}
&\mathbb{E}\big[m_n^2(\mathbf{X},\Theta)\mathbb{1}_{m_n(\mathbf{X},\Theta)\geq\beta_n}|\Theta\big] \\
&\quad\leq \mathbb{E}\Big[\Big(2\|m\|_\infty^2 + 2\max_{1\leq i\leq a_n}\varepsilon_i^2\Big)\mathbb{1}_{\max_{1\leq i\leq a_n}\varepsilon_i\geq\sigma\sqrt{2}(\log a_n)^2}\Big] \\
&\quad\leq 2\|m\|_\infty^2\mathbb{P}\Big[\max_{1\leq i\leq a_n}\varepsilon_i \geq \sigma\sqrt{2}(\log a_n)^2\Big] \\
&\qquad + 2\Big(\mathbb{E}\Big[\max_{1\leq i\leq a_n}\varepsilon_i^4\Big]\mathbb{P}\Big[\max_{1\leq i\leq a_n}\varepsilon_i\geq\sigma\sqrt{2}(\log a_n)^2\Big]\Big)^{1/2}.
\end{aligned}$$

It is easy to see that

$$\mathbb{P}\Big[\max_{1\leq i\leq a_n}\varepsilon_i \geq \sigma\sqrt{2}(\log a_n)^2\Big] \leq \frac{a_n^{1-\log a_n}}{2\sqrt{\pi}(\log a_n)^2}.$$

Finally, since the ε_i's are centered i.i.d. Gaussian random variables, we have, for all n large enough,

$$\begin{aligned}
&\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \\
&\quad\leq \frac{2\|m\|_\infty^2 a_n^{1-\log a_n}}{2\sqrt{\pi}(\log a_n)^2} + \xi + 2\Big(3a_n\sigma^4\frac{a_n^{1-\log a_n}}{2\sqrt{\pi}(\log a_n)^2}\Big)^{1/2} \\
&\quad\leq 3\xi.
\end{aligned}$$

This completes the proof of Theorem 1.

5.4. *Proof of Theorem* 2. Recall that each cell contains exactly one data point. Thus, letting

$$W_{ni}(\mathbf{X}) = \mathbb{E}_\Theta[\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)}],$$

the random forest estimate m_n may be rewritten as

$$m_n(\mathbf{X}) = \sum_{i=1}^n W_{ni}(\mathbf{X})Y_i.$$

We have in particular that $\sum_{i=1}^{n} W_{ni}(\mathbf{X}) = 1$. Thus

$$\begin{aligned}\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 &\leq 2\mathbb{E}\Bigg[\sum_{i=1}^{n} W_{ni}(\mathbf{X})\big(Y_i - m(\mathbf{X}_i)\big)\Bigg]^2 \\ &\quad + 2\mathbb{E}\Bigg[\sum_{i=1}^{n} W_{ni}(\mathbf{X})\big(m(\mathbf{X}_i) - m(\mathbf{X})\big)\Bigg]^2 \\ &\overset{\text{def}}{=} 2I_n + 2J_n.\end{aligned}$$

Approximation error. Fix $\alpha > 0$. To upper bound J_n, note that by Jensen's inequality,

$$\begin{aligned}J_n &\leq \mathbb{E}\Bigg[\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}\big(m(\mathbf{X}_i) - m(\mathbf{X})\big)^2\Bigg] \\ &\leq \mathbb{E}\Bigg[\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)}\Delta^2\big(m, A_n(\mathbf{X}, \Theta)\big)\Bigg] \\ &\leq \mathbb{E}\big[\Delta^2\big(m, A_n(\mathbf{X}, \Theta)\big)\big].\end{aligned}$$

So, by definition of $\Delta(m, A_n(\mathbf{X}, \Theta))^2$,

$$\begin{aligned}J_n &\leq 4\|m\|_\infty^2 \mathbb{E}[\mathbb{1}_{\Delta^2(m, A_n(\mathbf{X},\Theta)) \geq \alpha}] + \alpha \\ &\leq \alpha\big(4\|m\|_\infty^2 + 1\big),\end{aligned}$$

for all n large enough, according to Proposition 2.

Estimation error. To bound I_n from above, we note that

$$\begin{aligned}I_n &= \mathbb{E}\Bigg[\sum_{i,j=1}^{n} W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})\big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big)\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{i=1} W_{ni}^2(\mathbf{X})\big(Y_i - m(\mathbf{X}_i)\big)^2\Bigg] + I_n',\end{aligned}$$

where

$$I_n' = \mathbb{E}\Bigg[\sum_{\substack{i,j \\ i \neq j}} \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big)\Bigg].$$

The term I_n', which involves the double products, is handled separately in Lemma 4 below. According to this lemma, and by assumption (H2), for all n large enough,

$$|I_n'| \leq \alpha.$$

Consequently, recalling that $\varepsilon_i = Y_i - m(\mathbf{X}_i)$, we have, for all n large enough,

$$\begin{aligned}|I_n| &\le \alpha + \mathbb{E}\Bigg[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})\big(Y_i - m(\mathbf{X}_i)\big)^2\Bigg]\\ &\le \alpha + \mathbb{E}\Bigg[\max_{1\le \ell\le n} W_{n\ell}(\mathbf{X}) \sum_{i=1}^{n} W_{ni}(\mathbf{X})\varepsilon_i^2\Bigg] \qquad (7)\\ &\le \alpha + \mathbb{E}\Big[\max_{1\le \ell\le n} W_{n\ell}(\mathbf{X}) \max_{1\le i\le n} \varepsilon_i^2\Big].\end{aligned}$$

Now, observe that in the subsampling step, there are exactly $\binom{a_n-1}{n-1}$ choices to pick a fixed observation $\mathbf{X}_i$. Since $\mathbf{x}$ and $\mathbf{X}_i$ belong to the same cell only if $\mathbf{X}_i$ is selected in the subsampling step, we see that

$$\mathbb{P}_\Theta[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i] \le \frac{\binom{a_n-1}{n-1}}{\binom{a_n}{n}} = \frac{a_n}{n},$$

where $\mathbb{P}_\Theta$ denotes the probability with respect to Θ, conditional on $\mathbf{X}$ and $\mathcal{D}_n$. So,

$$\max_{1\le i\le n} W_{ni}(\mathbf{X}) \le \max_{1\le i\le n} \mathbb{P}_\Theta[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i] \le \frac{a_n}{n}. \qquad (8)$$

Thus, combining inequalities (7) and (8), for all n large enough,

$$|I_n| \le \alpha + \frac{a_n}{n}\mathbb{E}\Big[\max_{1\le i\le n} \varepsilon_i^2\Big].$$

The term inside the brackets is the maximum of n χ^2-squared distributed random variables. Thus, for some positive constant C,

$$\mathbb{E}\Big[\max_{1\le i\le n} \varepsilon_i^2\Big] \le C\log n;$$

see, for example, Boucheron, Lugosi and Massart (2013), Chapter 1. We conclude that for all n large enough,

$$I_n \le \alpha + C\frac{a_n \log n}{n} \le 2\alpha.$$

Since α was arbitrary, the proof is complete.

LEMMA 4. *Assume that* (H2) *is satisfied. Then, for all* $\varepsilon > 0$, *and all* n *large enough*, $|I_n'| \le \alpha$.

PROOF. First, assume that (H2.2) is verified. Thus we have for all $\ell_1, \ell_2 \in \{0, 1\}$,

$$\begin{aligned}&\operatorname{Corr}\big(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)} | \mathbf{X}_i, \mathbf{X}_j, Y_j\big)\\ &\quad= \frac{\mathbb{E}[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}]}{\mathbb{V}^{1/2}[Y_i - m(\mathbf{X}_i)|\mathbf{X}_i, \mathbf{X}_j, Y_j]\mathbb{V}^{1/2}[\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j]}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j]}{\sigma(\mathbb{P}[Z_{i,j} = (\ell_1, \ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j] - \mathbb{P}[Z_{i,j} = (\ell_1, \ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j]^2)^{1/2}} \\
&\geq \frac{\mathbb{E}[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j]}{\sigma\mathbb{P}^{1/2}[Z_{i,j} = (\ell_1, \ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j]},
\end{aligned}
$$

where the first equality comes from the fact that, for all $\ell_1, \ell_2 \in \{0, 1\}$,

$$
\begin{aligned}
&\operatorname{Cov}\big(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\big) \\
&\qquad = \mathbb{E}\big[\big(Y_i - m(\mathbf{X}_i)\big)\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j\big],
\end{aligned}
$$

since $\mathbb{E}[Y_i - m(\mathbf{X}_i)|\mathbf{X}_i, \mathbf{X}_j, Y_j] = 0$. Thus, noticing that, almost surely,

$$
\begin{aligned}
&\mathbb{E}\big[Y_i - m(\mathbf{X}_i)|Z_{i,j}, \mathbf{X}_i, \mathbf{X}_j, Y_j\big] \\
&\quad = \sum_{\ell_1,\ell_2=1}^{2} \frac{\mathbb{E}[(Y_i - m(\mathbf{X}_i))\mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j]}{\mathbb{P}[Z_{i,j} = (\ell_1, \ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j]} \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)} \\
&\quad \leq 4\sigma \max_{\ell_1,\ell_2=0,1} \frac{|\operatorname{Corr}(Y_i - m(\mathbf{X}_i), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)}|\mathbf{X}_i, \mathbf{X}_j, Y_j)|}{\mathbb{P}^{1/2}[Z_{i,j} = (\ell_1, \ell_2)|\mathbf{X}_i, \mathbf{X}_j, Y_j]} \\
&\quad \leq 4\sigma\gamma_n,
\end{aligned}
$$

we conclude that the first statement in (H2.2) implies that, almost surely,

$$
\mathbb{E}\big[Y_i - m(\mathbf{X}_i)|Z_{i,j}, \mathbf{X}_i, \mathbf{X}_j, Y_j\big] \leq 4\sigma\gamma_n.
$$

Similarly, one can prove that the second statement in assumption (H2.2) implies that, almost surely,

$$
\mathbb{E}\big[\big|Y_i - m(\mathbf{X}_i)\big|^2|\mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\big] \leq 4C\sigma^2.
$$

Returning to the term I_n', and recalling that $W_{ni}(\mathbf{X}) = \mathbb{E}_\Theta[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}]$, we obtain

$$
\begin{aligned}
I_n' &= \mathbb{E}\Big[\sum_{\substack{i,j \\ i \neq j}} \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big)\Big] \\
&= \sum_{\substack{i,j \\ i \neq j}} \mathbb{E}[\mathbb{E}[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \big(Y_i - m(\mathbf{X}_i)\big) \\
&\qquad\qquad \times \big(Y_j - m(\mathbf{X}_j)\big)|\mathbf{X}_i, \mathbf{X}_j, Y_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}]] \\
&= \sum_{\substack{i,j \\ i \neq j}} \mathbb{E}[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \big(Y_i - m(\mathbf{X}_i)\big) \\
&\qquad\qquad \times \mathbb{E}\big[Y_j - m(\mathbf{X}_j)|\mathbf{X}_i, \mathbf{X}_j, Y_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}\big]].
\end{aligned}
$$

Therefore, by assumption (H2.2),

$$\begin{aligned}
|I_n'| &\le 4\sigma\gamma_n \sum_{\substack{i,j\\ i\neq j}} \mathbb{E}\big[\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\mathbf{X}\overset{\Theta'}{\leftrightarrow}\mathbf{X}_j}|Y_i - m(\mathbf{X}_i)|\big]\\
&\le \gamma_n \sum_{i=1}^{n} \mathbb{E}\big[\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}|Y_i - m(\mathbf{X}_i)|\big]\\
&\le \gamma_n \sum_{i=1}^{n} \mathbb{E}\big[\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{E}\big[|Y_i - m(\mathbf{X}_i)||\mathbf{X}_i, \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\big]\big]\\
&\le \gamma_n \sum_{i=1}^{n} \mathbb{E}\big[\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{E}^{1/2}\big[|Y_i - m(\mathbf{X}_i)|^2|\mathbf{X}_i, \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\big]\big]\\
&\le 2\sigma C^{1/2}\gamma_n.
\end{aligned}$$

This proves the result, provided (H2.2) is true. Let us now assume that (H2.1) is verified. The key argument is to note that a data point $\mathbf{X}_i$ can be connected with a random point $\mathbf{X}$ if $(\mathbf{X}_i, Y_i)$ is selected via the subsampling procedure and if there are no other data points in the hyperrectangle defined by $\mathbf{X}_i$ and $\mathbf{X}$. Data points $\mathbf{X}_i$ satisfying the latter geometrical property are called *layered nearest neighbors* (LNN); see, for example, Barndorff-Nielsen and Sobel (1966). The connection between LNN and random forests was first observed by Lin and Jeon (2006), and later worked out by Biau and Devroye (2010). It is known, in particular, that the number of LNN $L_{a_n}(\mathbf{X})$ among a_n data points uniformly distributed on $[0, 1]^d$ satisfies, for some constant $C_1 > 0$ and for all n large enough,

$$\begin{aligned}
\mathbb{E}\big[L_{a_n}^4(\mathbf{X})\big] &\le a_n\mathbb{P}\big[\mathbf{X}\underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}}\mathbf{X}_j\big] + 16a_n^2\mathbb{P}\big[\mathbf{X}\underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}}\mathbf{X}_i\big]\mathbb{P}\big[\mathbf{X}\underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}}\mathbf{X}_j\big]\\
&\le C_1(\log a_n)^{2d-2};
\end{aligned} \tag{9}$$

see, for example, Bai et al. (2005), Barndorff-Nielsen and Sobel (1966). Thus we have

$$I_n' = \mathbb{E}\Big[\sum_{\substack{i,j\\ i\neq j}} \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\mathbf{X}\overset{\Theta'}{\leftrightarrow}\mathbf{X}_j}\mathbb{1}_{\mathbf{X}_i\underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}}\mathbf{X}}\mathbb{1}_{\mathbf{X}_j\underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}}\mathbf{X}}\big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big)\Big].$$

Consequently,

$$\begin{aligned}
I_n' = \mathbb{E}\Big[&\sum_{\substack{i,j\\ i\neq j}} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big)\mathbb{1}_{\mathbf{X}_i\underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}}\mathbf{X}}\mathbb{1}_{\mathbf{X}_j\underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}}\mathbf{X}}\\
&\times \mathbb{E}\big[\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\mathbf{X}\overset{\Theta'}{\leftrightarrow}\mathbf{X}_j}|\mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \dots, \mathbf{X}_n, Y_i, Y_j\big]\Big],
\end{aligned}$$

where $\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}$ is the event where $\mathbf{X}_i$ is selected by the subsampling and is also a LNN of $\mathbf{X}$. Next, with the notation of assumption (H2),

$$\begin{aligned}
I_n' &= \mathbb{E}\Big[\sum_{\substack{i,j\\ i\neq j}} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \psi_{i,j}(Y_i, Y_j)\Big] \\
&= \mathbb{E}\Big[\sum_{\substack{i,j\\ i\neq j}} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \psi_{i,j}\Big] \\
&\quad + \mathbb{E}\Big[\sum_{\substack{i,j\\ i\neq j}} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \big(\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}\big)\Big].
\end{aligned}$$

The first term is easily seen to be zero since

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{\substack{i,j\\ i\neq j}} \big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big) \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \psi\big(\mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n\big)\Big] \\
&\quad = \sum_{\substack{i,j\\ i\neq j}} \mathbb{E}\big[\mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \psi_{i,j} \\
&\qquad\qquad \times \mathbb{E}\big[\big(Y_i - m(\mathbf{X}_i)\big)\big(Y_j - m(\mathbf{X}_j)\big) | \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n, \Theta, \Theta'\big]\big] \\
&\quad = 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|I_n'| &\leq \mathbb{E}\Big[\sum_{\substack{i,j\\ i\neq j}} |Y_i - m(\mathbf{X}_i)| |Y_j - m(\mathbf{X}_j)| \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\Big] \\
&\leq \mathbb{E}\Big[\max_{1\leq \ell \leq n} |Y_i - m(\mathbf{X}_i)|^2 \max_{\substack{i,j\\ i\neq j}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}| \sum_{\substack{i,j\\ i\neq j}} \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}}\Big].
\end{aligned}$$

Now, observe that

$$\sum_{\substack{i,j\\ i\neq j}} \mathbb{1}_{\mathbf{X}_i \underset{\text{LNN}}{\overset{\Theta}{\leftrightarrow}} \mathbf{X}} \mathbb{1}_{\mathbf{X}_j \underset{\text{LNN}}{\overset{\Theta'}{\leftrightarrow}} \mathbf{X}} \leq L_{a_n}^2(\mathbf{X}).$$

Consequently,

$$\begin{aligned}
|I_n'| &\leq \mathbb{E}^{1/2}\Big[L_{a_n}^4(\mathbf{X}) \max_{1\leq \ell \leq n} |Y_i - m(\mathbf{X}_i)|^4\Big] \\
&\quad \times \mathbb{E}^{1/2}\Big[\max_{\substack{i,j\\ i\neq j}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\Big]^2.
\end{aligned} \tag{10}$$

Simple calculations reveal that there exists $C_1 > 0$ such that, for all n,

$$\mathbb{E}\Big[\max_{1\le\ell\le n}|Y_i - m(\mathbf{X}_i)|^4\Big] \le C_1(\log n)^2. \tag{11}$$

Thus, by inequalities (9) and (11), the first term in (10) can be upper bounded as follows:

$$\begin{aligned}
&\mathbb{E}^{1/2}\Big[L_{a_n}^4(\mathbf{X})\max_{1\le\ell\le n}|Y_i - m(\mathbf{X}_i)|^4\Big]\\
&\quad= \mathbb{E}^{1/2}\Big[L_{a_n}^4(\mathbf{X})\mathbb{E}\Big[\max_{1\le\ell\le n}|Y_i - m(\mathbf{X}_i)|^4|\mathbf{X},\mathbf{X}_1,\dots,\mathbf{X}_n\Big]\Big]\\
&\quad\le C'(\log n)(\log a_n)^{d-1}.
\end{aligned}$$

Finally,

$$|I_n'| \le C'(\log a_n)^{d-1}(\log n)^{\alpha/2}\mathbb{E}^{1/2}\Big[\max_{\substack{i,j\\ i\ne j}}|\psi_{i,j}(Y_i,Y_j) - \psi_{i,j}|\Big]^2,$$

which tends to zero by assumption. □

Acknowledgments. We greatly thank two referees for valuable comments and insightful suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Consistency of random forests" (DOI: 10.1214/15-AOS1321SUPP; .pdf). Proofs of technical results.

REFERENCES

AMARATUNGA, D., CABRERA, J. and LEE, Y.-S. (2008). Enriched random forests. *Bioinformatics* **24** 2010–2014.

BAI, Z.-D., DEVROYE, L., HWANG, H.-K. and TSAI, T.-H. (2005). Maxima in hypercubes. *Random Structures Algorithms* **27** 290–309. MR2162600

BARNDORFF-NIELSEN, O. and SOBEL, M. (1966). On the distribution of the number of admissible points in a vector random sample. *Teor. Verojatnost. i Primenen.* **11** 283–305. MR0207003

BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. MR2930634

BIAU, G. and DEVROYE, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivariate Anal.* **101** 2499–2518. MR2719877

BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9** 2015–2033. MR2447310

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L. (2004). Consistency for a simple model of random forests. Technical Report 670, Univ. California, Berkeley, CA.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961. MR1926165

CLÉMENÇON, S., DEPECKER, M. and VAYATIS, N. (2013). Ranking forests. *J. Mach. Learn. Res.* **14** 39–73. MR3033325

CUTLER, D. R., EDWARDS, T. C. JR, BEARD, K. H., CUTLER, A., HESS, K. T., GIBSON, J. and LAWLER, J. J. (2007). Random forests for classification in ecology. *Ecology* **88** 2783–2792.

DENIL, M., MATHESON, D. and FREITAS, N. D. (2013). Consistency of online random forests. In *Proceedings of the ICML Conference*. Available at arXiv:1302.4853.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093

DÍAZ-URIARTE, R. and ALVAREZ DE ANDRÉS, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7** 1–13.

EFRON, B. (1982). *The Jackknife*, *the Bootstrap and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia. MR0659849

GENUER, R. (2012). Variance reduction in purely random forests. *J. Nonparametr. Stat.* **24** 543–562. MR2968888

GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Mach. Learn.* **63** 3–42.

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. MR1920390

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. MR0858512

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

ISHWARAN, H. and KOGALUR, U. B. (2010). Consistency of random survival forests. *Statist. Probab. Lett.* **80** 1056–1064. MR2651045

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677

LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. MR2256176

MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

MENTCH, L. and HOOKER, G. (2014). Ensemble trees and clts: Statistical inference for supervised learning. Available at arXiv:1404.6473.

NOBEL, A. (1996). Histogram regression estimation using data-dependent partitions. *Ann. Statist.* **24** 1084–1105. MR1401839

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York. MR1707286

PRASAD, A. M., IVERSON, L. R. and LIAW, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* **9** 181–199.

SCORNET, E. (2014). On the asymptotics of random forests. Available at arXiv:1409.2090.

SCORNET, E., BIAU, G. and VERT, J. (2015). Supplement to "Consistency of random forests." DOI:10.1214/15-AOS1321SUPP.

SHOTTON, J., SHARP, T., KIPMAN, A., FITZGIBBON, A., FINOCCHIO, M., BLAKE, A., COOK, M. and MOORE, R. (2013). Real-time human pose recognition in parts from single depth images. *Comm. ACM* **56** 116–124.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

SVETNIK, V., LIAW, A., TONG, C., CULBERSON, J. C., SHERIDAN, R. P. and FEUSTON, B. P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43** 1947–1958.

WAGER, S. (2014). Asymptotic theory for random forests. Available at arXiv:1405.0352.

WAGER, S., HASTIE, T. and EFRON, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **15** 1625–1651. MR3225243

ZHU, R., ZENG, D. and KOSOROK, M. R. (2012). Reinforcement learning trees. Technical report, Univ. North Carolina.

E. SCORNET
G. BIAU
SORBONNE UNIVERSITÉS
UPMC UNIV PARIS 06
PARIS F-75005
FRANCE
E-MAIL: erwan.scornet@upmc.fr
gerard.biau@upmc.fr

J.-P. VERT
MINES PARISTECH, PSL-RESEARCH UNIVERSITY
CBIO-CENTRE FOR COMPUTATIONAL BIOLOGY
FONTAINEBLEAU F-77300
FRANCE
AND
INSTITUT CURIE
PARIS F-75248
FRANCE
AND
U900, INSERM
PARIS F-75248
FRANCE
E-MAIL: jean-philippe.vert@mines-paristech.fr