Journal of Multivariate Analysis 146 (2016) 72–83

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the asymptotics of random forests

Erwan Scornet

Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France

ARTICLE INFO

Article history:
Received 7 September 2014
Available online 24 June 2015

AMS 2010 subject classifications:
62G05
62G20

Keywords:
Random forests
Randomization
Consistency
Central limit theorem
Empirical process
Number of trees
q-quantile

ABSTRACT

The last decade has witnessed a growing interest in random forest models which are recognized to exhibit good practical performance, especially in high-dimensional settings. On the theoretical side, however, their predictive power remains largely unexplained, thereby creating a gap between theory and practice. In this paper, we present some asymptotic results on random forests in a regression framework. Firstly, we provide theoretical guarantees to link finite forests used in practice (with a finite number M of trees) to their asymptotic counterparts (with $M = \infty$). Using empirical process theory, we prove a uniform central limit theorem for a large class of random forest estimates, which holds in particular for Breiman's (2001) original forests. Secondly, we show that infinite forest consistency implies finite forest consistency and thus, we state the consistency of several infinite forests. In particular, we prove that q quantile forests – close in spirit to Breiman's (2001) forests but easier to study – are able to combine inconsistent trees to obtain a final consistent prediction, thus highlighting the benefits of random forests compared to single trees.

1. Introduction

Random forests are a class of algorithms used to solve classification and regression problems. As ensemble methods, they grow several trees as base estimates and aggregate them to make a prediction. In order to obtain many different trees based on a single training set, random forests procedures introduce randomness in the tree construction. For instance, trees can be built by randomizing the set of features [14,21], the data set [5,6], or both at the same time [7,10].

Among all random forest algorithms, the most popular one is that of Breiman [7], which relies on CART procedure (Classification and Regression Trees, [8]) to grow the individual trees. As highlighted by several applied studies (see, e.g., [20,13]), Breiman's [7] random forests can outperform state-of-the-art methods. They are recognized for their ability to handle high-dimensional data sets, thus being useful in fields such as genomics [31] and pattern recognition [33], just to name a few. On the computational side, Breiman's [7] forests are easy to run and robust to changes in the parameters they depend on [27,16]. Besides, extensions have been developed in ranking problems [9], quantile estimation [28], and survival analysis [25]. Interesting new developments in the context of massive data sets have been achieved. For instance, Geurts et al. [17] modified the procedure to reduce calculation time, while other authors extended the procedure to online settings [11,26, and the references therein].

While Breiman's [7] forests are extensively used in practice, some of their mathematical properties remain under active investigation. In fact, most theoretical studies focus on simplified versions of the algorithm, where the forest construction is independent of the training set. Consistency of such simplified models has been proved (e.g., [2,24,11]). However, these results do not extend to Breiman's [7] original forests whose construction critically depends on the whole training set. Recent attempts to bridge the gap between theoretical forest models and Breiman's [7] forests have been made by Wager [37] and Scornet et al. [34] who establish consistency of the original algorithm under suitable assumptions.

E-mail address: erwan.scornet@upmc.fr.

http://dx.doi.org/10.1016/j.jmva.2015.06.009
0047-259X/© 2015 Elsevier Inc. All rights reserved.

Apart from the dependence of the forest construction on the data set, there is another fundamental difference between existing forest models and ones implemented. Indeed, in practice, a forest can only be grown with a finite number M of trees although most theoretical works assume, by convenience, that $M = \infty$. Since the predictor with $M = \infty$ does not depend on the specific tree realizations that form the forest, it is therefore more amenable to analysis. However, surprisingly, no study aims at clarifying the link between finite forests (finite M) and infinite forests ($M = \infty$) even if some authors [29,38] proved results on finite forest predictions at a fixed point $\mathbf{x}$.

In the present paper, our goal is to study the connection between infinite forest models and finite forests used in practice in the context of regression. We start by proving a uniform central limit theorem for various random forests estimates, including Breiman's [7] ones. In Section 3, assuming some regularity on the regression model, we point out that the $\mathbb{L}^2$ risk of any infinite forest is bounded above by the risk of the associated finite forests. Thus infinite forests are better estimate than finite forests in terms of $\mathbb{L}^2$ risk. Under the same assumptions, our analysis also shows that the risks of infinite and finite forests are close, if the number of trees is chosen to be large enough. An interesting corollary of this result is that infinite forest consistency implies finite forest consistency. Finally, in Section 4, we prove the consistency of several infinite random forests. In particular, taking one step towards the understanding of Breiman's [7] forests, we prove that q quantile forests, a variety of forests whose construction depends on the positions $\mathbf{X}_i$'s of the data, are consistent. As for Breiman's [7] forests, each leaf of each tree in q quantile forests contains a small number of points that does not grow to infinity with the sample size. Thus, q quantile forests average inconsistent trees estimate to build a consistent prediction.

We start by giving some notation in Section 2. All proofs are postponed to Section 5.

2. Notation

Throughout the paper, we assume to be given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $[0, 1]^d \times \mathbb{R}$-valued independent random variables distributed as the prototype pair $(\mathbf{X}, Y)$, where $\mathbb{E}[Y^2] < \infty$. We aim at predicting the response Y, associated with the random variable $\mathbf{X}$, by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. In this context, we use random forests to build an estimate $m_n : [0, 1]^d \to \mathbb{R}$ of m, based on the data set $\mathcal{D}_n$.

A random forest is a collection of M randomized regression trees (for an overview on tree construction, see Chapter 20 in [19]). For the jth tree in the family, the predicted value at point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}, \Theta_j)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ, independent of the sample $\mathcal{D}_n$. This random variable can be used to sample the training set or to select the candidate directions or positions for splitting. The trees are combined to form the finite forest estimate

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m). \tag{1}$$

By the law of large numbers, for any fixed $\mathbf{x}$, conditionally on $\mathcal{D}_n$, the finite forest estimate tends to the infinite forest estimate

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}, \Theta) \right].$$

The risk of $m_{\infty,n}$ is defined by

$$R(m_{\infty,n}) = \mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2, \tag{2}$$

while the risk of $m_{M,n}$ equals

$$R(m_{M,n}) = \mathbb{E}[m_{M,n}(\mathbf{X}, \Theta_1, \ldots, \Theta_M) - m(\mathbf{X})]^2. \tag{3}$$

It is stressed that both risks $R(m_{\infty,n})$ and $R(m_{M,n})$ are deterministic since the expectation in (2) is over $\mathbf{X}$, $\mathcal{D}_n$, and the expectation in (3) is over $\mathbf{X}$, $\mathcal{D}_n$ and $\Theta_1, \ldots, \Theta_M$. Throughout the paper, we say that $m_{\infty,n}$ (resp. $m_{M,n}$) is $\mathbb{L}^2$ consistent if $R(m_{\infty,n})$ (resp. $R(m_{M,n})$) tends to zero as $n \to \infty$.

As mentioned earlier, there is a large variety of forests, depending on how trees are grown and how the randomness Θ influences the tree construction. For instance, tree construction can be independent of $\mathcal{D}_n$ [1], depend only on the $\mathbf{X}_i$'s [2] or depend on the whole training set [10,17,39]. Throughout the paper, we use Breiman's [7] forests and uniform forests to exemplify our results. In Breiman's [7] original procedure, splits depend on the whole sample and are performed to minimize variance within the two resulting cells. The algorithm stops when each cell contains less than a small pre-specified number of points (typically, 5 in regression). On the other hand, uniform forests are a simpler procedure since, at each node, a coordinate is uniformly selected among $\{1, \ldots, d\}$ and a split position is uniformly chosen in the range of the cell, along the pre-chosen coordinate. The algorithm stops when a full binary tree of level k is built, that is if each cell has been cut exactly k times, where $k \in \mathbb{N}$ is a parameter of the algorithm.

In the rest of the paper, we will repeatedly use the random forest connection function K_n, defined as

$$\begin{aligned} K_n : [0, 1]^d \times [0, 1]^d &\to [0, 1] \\ (\mathbf{x}, \mathbf{z}) &\mapsto \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z} \right], \end{aligned}$$

where $\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{z}$ is the event where $\mathbf{x}$ and $\mathbf{z}$ belong to the same cell in the tree $\mathcal{T}_n(\Theta)$ designed with Θ and $\mathcal{D}_n$. Moreover, notation $\mathbb{P}_\Theta$ denotes the probability with respect to Θ, conditionally on $\mathcal{D}_n$. The same notational convention holds for the expectation $\mathbb{E}_\Theta$ and the variance $\mathbb{V}_\Theta$. Thus, if we fix the training set $\mathcal{D}_n$, we see that the connection $K_n(\mathbf{x}, \mathbf{z})$ is just the probability that $\mathbf{x}$ and $\mathbf{z}$ are connected in the forest.

We say that a forest is discrete (resp. continuous) if, keeping $\mathcal{D}_n$ fixed, its connection function $K_n(\bullet, \bullet)$ is piecewise constant (resp. continuous). In fact, most existing forest models fall in one of these two categories. For example, if, at each cell, the number of possible splits is finite, then the forest is discrete. This is the case of Breiman's [7] forests, where splits can only be performed at the middle of two consecutive data points along any coordinate. However, if splits are drawn according to some density along each coordinate, the resulting forest is continuous. For instance, uniform forests are continuous.

3. Finite and infinite random forests

Contrary to finite forests which depend upon the particular Θ_j's used to design trees, infinite forests do not and are therefore more amenable to mathematical analysis. Besides, finite forests predictions can be difficult to interpret since they depend on the random parameters Θ_j's. In addition, the Θ_j's are independent of the data set and thus unrelated to the particular prediction problem.

In this section, we study the link between finite forests and infinite forests. More specifically, assuming that the data set $\mathcal{D}_n$ is fixed, we examine the asymptotic behavior of the finite forest estimate $m_{M,n}(\bullet, \Theta_1, \ldots, \Theta_M)$ as M tends to infinity. This setting is consistent with practical problems, where the $\mathcal{D}_n$ is fixed, and one can grow as many trees as possible.

Clearly, by the law of large numbers, we know that conditionally on $\mathcal{D}_n$, for all $\mathbf{x} \in [0, 1]^d$, almost surely,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) \underset{M\to\infty}{\to} m_{\infty,n}(\mathbf{x}). \tag{4}$$

The following theorem extends the pointwise convergence in (4) to the convergence of the whole functional estimate $m_{M,n}(\bullet, \Theta_1, \ldots, \Theta_M)$, towards the functional estimate $m_{\infty,n}(\bullet)$.

Theorem 3.1. *Consider a continuous or discrete random forest. Then, conditionally on $\mathcal{D}_n$, almost surely, for all $\mathbf{x} \in [0, 1]^d$, we have*

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) \underset{M\to\infty}{\to} m_{\infty,n}(\mathbf{x}).$$

Remark 1. Since the set $[0, 1]^d$ is not countable, we cannot reverse the "almost sure" and "for all $\mathbf{x} \in [0, 1]^d$" statements in (4). Thus, Theorem 3.1 is not a consequence of (4).

Theorem 3.1 is a first step to prove that infinite forest estimates can be uniformly approximated by finite forest estimates. To pursue the analysis, a natural question is to determine the rate of convergence in Theorem 3.1. The pointwise rate of convergence is provided by the central limit theorem which says that, conditionally on $\mathcal{D}_n$, for all $\mathbf{x} \in [0, 1]^d$,

$$\sqrt{M}\big(m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M) - m_{\infty,n}(\mathbf{x})\big) \overset{\mathcal{L}}{\underset{M\to\infty}{\to}} \mathcal{N}\big(0, \tilde{\sigma}^2(\mathbf{x})\big), \tag{5}$$

where

$$\tilde{\sigma}^2(\mathbf{x}) = \mathbb{V}_\Theta\left(\frac{1}{N_n(\mathbf{x}, \Theta)} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right) \le 4 \max_{1\le i\le n} Y_i^2$$

(as before, $\mathbb{V}_\Theta$ denotes with respect to Θ, conditionally on $\mathcal{D}_n$), and $N_n(\mathbf{x}, \Theta)$ is the number of data points falling into the cell of the tree $\mathcal{T}_n(\Theta)$ which contains $\mathbf{x}$.

Eq. (5) is not sufficient to determine the asymptotic distribution of the functional estimate $m_{M,n}(\bullet, \Theta_1, \ldots, \Theta_M)$. To make it explicit, we need to introduce the empirical process $\mathbb{G}_M$ (see [36]) defined by

$$\mathbb{G}_M = \sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M} \delta_{\Theta_m} - \mathbb{P}_\Theta\right),$$

where δ_{Θ_m} is the Dirac function at Θ_m. We also let $\mathcal{F}_2 = \{g_{\mathbf{x}} : \theta \mapsto m_n(\mathbf{x}, \theta); \mathbf{x} \in [0, 1]^d\}$ be the collection of all possible tree estimates in the forest. In order to prove that a uniform central limit theorem holds for random forest estimates, we need to show that there exists a Gaussian process $\mathbb{G}$ such that

$$\sup_{g\in\mathcal{F}_2}\left\{\int_\Theta |g(\theta)|\mathrm{d}\mathbb{G}_M(\theta) - \int_\Theta |g(\theta)|\mathrm{d}\mathbb{G}(\theta)\right\} \underset{M\to\infty}{\to} 0, \tag{6}$$

where the first part on the left side can be written as

$$\int_\Theta |g(\theta)|\mathrm{d}\mathbb{G}_M(\theta) = \sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M} |g(\Theta_m)| - \mathbb{E}_\Theta\big[|g(\Theta)|\big]\right).$$

For more clarity, instead of (6), we will write

$$\sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M} m_n(\bullet, \Theta_m) - \mathbb{E}_{\Theta}\left[m_n(\bullet, \Theta)\right]\right) \xrightarrow{\mathcal{L}} \mathbb{G}g_{\bullet}. \tag{7}$$

To establish identity (7), we first define, for all $\varepsilon > 0$, the random forest grid step $\delta(\varepsilon)$ by

$$\delta(\varepsilon) = \sup\left\{\eta \in \mathbb{R} : \sup_{\substack{\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d \\ \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \le \eta}} \left|1 - K_n(\mathbf{x}_1, \mathbf{x}_2)\right| \le \frac{\varepsilon^2}{8}\right\},$$

where K_n is the connection function of the forest. The function δ can be seen as the modulus of continuity of K_n in the sense that it is the distance such that $K_n(\mathbf{x}_1, \mathbf{x}_2)$ does not vary of much that $\varepsilon^2/8$ if $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \le \delta(\varepsilon)$. We will also need the following assumption.

(H1). One of the following properties is satisfied:

- The random forest is discrete,
- There exist $C, A > 0$, $\alpha < 2$ such that, for all $\varepsilon > 0$,

$$\delta(\varepsilon) \ge C \exp(-A/\varepsilon^{\alpha}).$$

Observe that (H1) is mild since most forests are discrete and the only continuous forest we have in mind, the uniform forest, satisfies (H1), as stated in Lemma 1.

Lemma 1. *Let $k \in \mathbb{N}$. Then, for all $\varepsilon > 0$, the grid step $\delta(\varepsilon)$ of uniform forests of level k satisfies*

$$\delta(\varepsilon) \ge \exp\left(-\frac{A_{k,d}}{\varepsilon^{2/3}}\right),$$

where $A_{k,d} = (8de(k+2)!)^{1/3}$.

The following theorem states that a uniform central limit theorem is valid over the class of random forest estimates, providing that (H1) is satisfied.

Theorem 3.2. *Consider a random forest which satisfies (H1). Then, conditionally on $\mathcal{D}_n$,*

$$\sqrt{M}\left(m_{M,n}(\bullet) - m_{\infty,n}(\bullet)\right) \xrightarrow{\mathcal{L}} \mathbb{G}g_{\bullet},$$

where $\mathbb{G}$ is a Gaussian process with mean zero and a covariate function

$$\mathrm{Cov}_{\Theta}(\mathbb{G}g_{\mathbf{x}}, \mathbb{G}g_{\mathbf{z}}) = \mathrm{Cov}_{\Theta}\left(\sum_{i=1}^{n} Y_i \frac{\mathbb{1}_{\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{x}, \Theta)}, \sum_{i=1}^{n} Y_i \frac{\mathbb{1}_{\mathbf{z} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{z}, \Theta)}\right).$$

According to the discussion above, Theorem 3.2 holds for uniform forests (by Lemma 1) and Breiman's [7] forests (since they are discrete). Moreover, according to this theorem, the finite forest estimates tend uniformly to the infinite forest estimates, with the standard rate of convergence $\sqrt{M}$. This result contributes to bridge the gap between finite forests used in practice and infinite theoretical forests.

The proximity between two estimates can also be measured in terms of their $\mathbb{L}^2$ risk. In this respect, Theorem 3.3 states that the risk of infinite forests is lower than the one of finite forests and provides a bound on the difference between these two risks. We first need an assumption on the regression model.

(H2). One has

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centered Gaussian noise with finite variance σ^2, independent of $\mathbf{X}$, and $\|m\|_\infty = \sup_{\mathbf{x} \in [0,1]^d} |m(\mathbf{x})| < \infty$.

Theorem 3.3. *Assume that (H2) is satisfied. Then, for all $M, n \in \mathbb{N}^\star$,*

$$R(m_{M,n}) = R(m_{\infty,n}) + \frac{1}{M}\mathbb{E}_{\mathbf{X}, \mathcal{D}_n}\left[\mathbb{V}_{\Theta}\left[m_n(\mathbf{X}, \Theta)\right]\right].$$

In particular,

$$0 \le R(m_{M,n}) - R(m_{\infty,n}) \le \frac{8}{M} \times \left(\|m\|_\infty^2 + \sigma^2(1 + 4\log n)\right).$$

Theorem 3.3 reveals that the prediction accuracy of infinite forests is better than that of finite forests. In practice however, there is no simple way to implement infinite forests and, in fact, finite forests are nothing but Monte Carlo approximations of infinite forests. But, since the difference of risks between both types of forests is bounded (by Theorem 3.3), the prediction accuracy of finite forests is almost as good as that of infinite forests providing the number of trees is large enough. More precisely, under (H2), for all $\varepsilon > 0$, if

$$M \geq \frac{8(\|m\|_\infty^2 + \sigma^2)}{\varepsilon} + \frac{32\sigma^2 \log n}{\varepsilon},$$

then $R(m_{M,n}) - R(m_{\infty,n}) \leq \varepsilon$.

Another interesting consequence of Theorem 3.3 is that, assuming that (H2) holds and that $M/\log n \to \infty$ as $n \to \infty$, finite random forests are consistent as soon as infinite random forests are. This allows to extend all previous consistency results regarding infinite forests (see, e.g., [28,2]) to finite forests. It must be stressed that the "$\log n$" term comes from the Gaussian noise, since, if $\varepsilon_1, \ldots, \varepsilon_n$ are independent and distributed as a Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, we have,

$$\mathbb{E}\left[\max_{1\leq i\leq n} \varepsilon_i^2\right] \leq \sigma^2(1 + 4\log n),$$

(see, e.g., Chapter 1 in [4]). Therefore, the required number of trees depends on the noise in the regression model. For instance, if Y is bounded, then the condition turns into $M \to \infty$.

4. Consistency of some random forest models

Section 3 was devoted to the connection between finite and infinite forests. In particular, we proved in Theorem 3.3 that the consistency of infinite forests implies that of finite forests, as soon as (H2) is satisfied and $M/\log n \to \infty$. Thus, it is natural to focus on the consistency of infinite forest estimates, which can be written as

$$m_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) Y_i, \tag{8}$$

where

$$W_{ni}^{\infty}(\mathbf{X}) = \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{X}, \Theta)}\right]$$

are the random forest weights.

4.1. Totally non adaptive forests

Proving consistency of infinite random forests is in general a difficult task, mainly because forest construction can depend on both the $\mathbf{X}_i$'s and the Y_i's. This feature makes the resulting estimate highly data-dependent, and therefore difficult to analyze (this is particularly the case for Breiman's [7] forests). To simplify the analysis, we investigate hereafter infinite random forest estimates whose weights depend only on $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ which is called the X-property. The good news is that when infinite forest estimates have the X-property, they fall in the general class of local averaging estimates, whose consistency can be addressed using Stone's [35] theorem.

Therefore, using Stone's theorem as a starting point, we first prove the consistency of random forests whose construction is independent of $\mathcal{D}_n$, which is the simplest case of random forests satisfying the X-property. For such forests, the construction is based on the random parameter Θ only. As for now, we say that a forest is totally non adaptive of level k ($k \in \mathbb{N}$, with k possibly depending on n) if each tree of the forest is built independently of the training set and if each cell is cut exactly k times. The resulting cell containing $\mathbf{X}$, designed with randomness Θ, is denoted by $A_n(\mathbf{X}, \Theta)$.

Theorem 4.1. *Assume that $\mathbf{X}$ is distributed on $[0, 1]^d$ and consider a totally non adaptive forest of level k. In addition, assume that for all $\rho, \varepsilon > 0$, there exists $N > 0$ such that, with probability $1 - \rho$, for all $n > N$,*

$$\mathrm{diam}(A_n(\mathbf{X}, \Theta)) \leq \varepsilon.$$

Then, providing $k \to \infty$ and $2^k/n \to 0$, the infinite random forest is $\mathbb{L}^2$ consistent, that is

$$R(m_{\infty,n}) \to 0 \quad \text{as } n \to \infty.$$

Theorem 4.1 is a generalization of some consistency results in [2] for the case of totally non adaptive random forest. Together with Theorem 3.3, we see that if (H2) is satisfied and $M/\log n \to \infty$ as $n \to \infty$, then the finite random forest is $\mathbb{L}^2$ consistent.

According to Theorem 4.1, a totally non adaptive forest of level k is consistent if the cell diameters tend to zero as $n \to \infty$ and if the level k is properly tuned. This is in particular true for uniform random forests, as shown in the following corollary.

Corollary 1. *Assume that $\mathbf{X}$ is distributed on $[0, 1]^d$ and consider a uniform forest of level k. Then, providing that $k \to \infty$ and $2^k/n \to 0$, the uniform random forest is $\mathbb{L}^2$ consistent.*

4.2. q quantile forests

For totally non adaptive forests, the main difficulty that consists in using the data set to build the forest and to predict at the same time, vanishes. However, because of their simplified construction, these forests are far from accurately modeling Breiman's [7] forest. To take one step further into the understanding of Breiman's [7] forest behavior, we study the q ($q \in [1/2, 1)$) quantile random forest, which satisfies the X-property. Indeed, their construction depends only on the X_i's which is a good trade off between the complexity of Breiman's [7] forests and the simplicity of totally non adaptive forests. As an example of q quantile trees, the median tree ($q = 1/2$) has already been studied by Devroye et al. [12], such as the k-spacing tree [12] whose construction is based on quantiles.

In the spirit of Breiman's [7] algorithm, before growing each tree, data are subsampled, that is a_n points ($a_n < n$) are selected without replacement. Then, each split is performed on an empirical q'-quantile (where $q' \in [1 - q, q]$ can be pre-specified by the user or randomly chosen) along a coordinate, chosen uniformly at random among the d coordinates. Recall that the q'-quantile ($q' \in [1 - q, q]$) of $\mathbf{X}_1, \dots, \mathbf{X}_n$ is defined as the only $\mathbf{X}_{(\ell)}$ satisfying $F_n(\mathbf{X}_{(\ell-1)}) \leq q' < F_n(\mathbf{X}_{(\ell)})$, where the $\mathbf{X}_{(i)}$'s are ordered increasingly. Note that data points on which splits are performed are not sent down to the resulting cells. This is done to ensure that data points are uniformly distributed on the resulting cells (otherwise, there would be at least one data point on the edge of the resulting cell, and thus the data point distribution would not be uniform on this cell). Finally, the algorithm stops when each cell contains exactly one point. The full procedure is described in Algorithm 1.

Algorithm 1: q quantile forest predicted value at $\mathbf{x}$.

```
Input: Fix $a_n \in \{1, \dots, n\}$, and $\mathbf{x} \in [0, 1]^d$.
Data: A training set $\mathcal{D}_n$.
for $\ell = 1, \dots, M$ do
    Select $a_n$ points, without replacement, uniformly in $\mathcal{D}_n$.
    Set $\mathcal{P} = \{[0, 1]^d\}$ the partition associated with the root of the tree.
    while there exists $A \in \mathcal{P}$ which contains strictly more than two points do
        Select uniformly one dimension $j$ within $\{1, \dots, d\}$.
        Let $N$ be the number of data points in $A$ and select $q' \in [1 - q, q] \cap (1/N, 1 - 1/N)$.
        Cut the cell $A$ at the position given by the $q'$ empirical quantile (see definition above) along the $j$th coordinate.
        Call $A_L$ and $A_R$ the two resulting cell.
        Set $\mathcal{P} \leftarrow (\mathcal{P} \backslash \{A\}) \cup A_L \cup A_R$.
    end
    for each $A \in \mathcal{P}$ which contains exactly two points do
        Select uniformly one dimension $j$ within $\{1, \dots, d\}$.
        Cut along the $j$th direction, in the middle of the two points.
        Call $A_L$ and $A_R$ the two resulting cell.
        Set $\mathcal{P} \leftarrow (\mathcal{P} \backslash \{A\}) \cup A_L \cup A_R$.
    end
    Compute the predicted value $m_n(\mathbf{x}, \Theta_\ell)$ at $\mathbf{x}$ equal to the single $Y_i$ falling in the cell of $\mathbf{x}$, with respect to the
    partition $\mathcal{P}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}, \Theta_1, \dots, \Theta_M)$ at the query point $\mathbf{x}$ according to equality (1).
```

Since the construction of q quantile forests depends on the $\mathbf{X}_i$'s and is based on subsampling, it is a more realistic modeling of Breiman's [7] forests than totally non adaptive forests. It also provides a good understanding on why random forests are still consistent even when there is exactly one data point in each leaf. Theorem 4.2 states that with a proper subsampling rate of the training set, the q quantile random forests are consistent.

(H3). One has

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a centered noise such that $\mathbb{V}[\varepsilon | \mathbf{X} = \mathbf{x}] \leq \sigma^2$, where $\sigma^2 < \infty$ is a constant. Moreover, $\mathbf{X}$ has a density on $[0, 1]^d$ and m is continuous.

Theorem 4.2. *Assume that (H3) is satisfied. Then, providing $a_n \to \infty$ and $a_n/n \to 0$, the infinite q quantile random forest is $\mathbb{L}^2$ consistent.*

4.3. Discussion

Some remarks are in order. At first, observe that each tree in the q quantile forest is inconsistent (see Problem 4.3 in [19]), because each leaf contains exactly one data point, a number which does not grow to infinity as $n \to \infty$. Thus, Theorem 4.2 shows that q quantile forest combines inconsistent trees to form a consistent estimate.

Secondly, many random forests can be seen as quantile forests if they satisfy the X-property and if splits do not separate a small fraction of data points from the rest of the sample (indeed, for each split in the q quantile forests, the resulting cells contain at least a fraction q of the observations falling into the parent node). The last assumption is true, for example, if $\mathbf{X}$ has a density on $[0, 1]^d$ bounded from below and from above, and if some splitting rule forces splits to be performed far away from the cell edges. This assumption is explicitly made in the analysis of Meinshausen [28] and Wager [37] to ensure that cell diameters tend to zero as $n \to \infty$, which is a necessary condition to prove the consistency of partitioning estimates (see Chapter 4 in [19]). Unfortunately, there are no results stating that splits in Breiman's [7] forests are performed far from the edges (see [23] for an analysis of the splitting criterion in Breiman's forests).

In addition, we note that Theorem 4.2 does not cover the bootstrap case since in that case, $a_n = n$ data points are selected with replacement. However, the condition on the subsampling rate can be replaced by the following one: for all $\mathbf{x}$,

$$\max_i \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i \right] \to 0 \quad \text{as } n \to \infty. \tag{9}$$

Condition (9) can be interpreted by saying that a point $\mathbf{x}$ should not be connected too often to the same data point in the forest, thus meaning that trees have to be various enough to ensure the forest consistency. This idea of diversity among trees has already been suggested by Breiman [7]. In bootstrap case, a single data point is selected in about 63% of trees. Thus, the term $\max_i \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i \right]$ is roughly upper bounded by 0.63 which is not sufficient to prove (9). It does not mean that random forests based on bootstrap are inconsistent but that a more detailed analysis is required. A possible, but probably difficult, route is an in-depth analysis of the connection function $K_n(\mathbf{x}, \mathbf{X}_i) = \mathbb{P}_\Theta \left[\mathbf{x} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i \right]$.

Finally, a natural question is how to extend random forests to the case of functional data (see, e.g., [32,15,22,3] for an overview of functional data analysis). A first attempt may be done by expanding each variable in a particular truncated functional basis. Each curve is then represented by a finite number of coefficient and any standard random forest procedure can be applied (see, e.g., [30,18] for practical applications). Since this method mainly consists in projecting functional variables onto finite dimensional spaces, it suffers from several drawbacks (for example, it depends on the basis and on the truncated procedure which are arbitrarily chosen in most cases). Unfortunately, we are not aware of functional random forest procedures that can directly handle functional data. Given the good performance of random forests in high dimensional settings and the numerous applications involving functional data, developing such functional forests is certainly is an interesting research topic.

5. Proofs

For the sake of clarity, proofs of Section 3 are gathered in the supplementary material (see Appendix A).

5.1. Proof of Theorem 4.1 and Corollary 1

The proof of Theorem 4.1 is based on Stone's theorem [35].

Proof of Theorem 4.1. We check the assumptions of Stone's theorem. For every non negative measurable function f satisfying $\mathbb{E}f(\mathbf{X}) < \infty$ and for any n, almost surely,

$$\mathbb{E}_{\mathbf{X},\mathcal{D}_n} \left(\sum_{i=1}^{n} W_{ni}(\mathbf{X}, \Theta) f(\mathbf{X}_i) \right) \leq \mathbb{E}_{\mathbf{X}} \left(f(\mathbf{X}) \right),$$

where

$$W_{ni}(\mathbf{X}, \Theta) = \frac{\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{X}, \Theta)}$$

are the weights of the random tree $\mathcal{T}_n(\Theta)$ (see the proof of Theorem 4.2 in [19]). Taking expectation with respect to Θ from both sides, we have

$$\mathbb{E}_{\mathbf{X},\mathcal{D}_n} \left(\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) f(\mathbf{X}_i) \right) \leq \mathbb{E}_{\mathbf{X}} \left(f(\mathbf{X}) \right),$$

which proves the first condition of Stone's theorem.

According to the definition of random forest weights W_{ni}^{∞}, since $\sum_{i=1}^{n} W_{ni}(\mathbf{X}, \Theta) \leq 1$ almost surely, we have

$$\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) = \mathbb{E}_\Theta \left[\sum_{i=1}^{n} W_{ni}(\mathbf{X}, \Theta) \right] \leq 1.$$

To check condition (iii), note that, for all $a > 0$,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right]\\ &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a} \times \mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))\geq a/2}\right],\end{aligned}$$

because $\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))<a/2} = 0$ (where $A_n(\mathbf{X},\Theta)$ is the cell containing $\mathbf{X}$ built with the randomness Θ). Thus,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X})\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right] &\leq \mathbb{E}\left[\mathbb{1}_{\mathrm{diam}(A_n(\mathbf{X},\Theta))\geq a/2} \times \sum_{i=1}^{n}\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\|\mathbf{X}-\mathbf{X}_i\|_\infty > a}\right]\\ &\leq \mathbb{P}\left[\mathrm{diam}(A_n(\mathbf{X},\Theta)) \geq a/2\right],\end{aligned}$$

which tends to zero, as $n \to \infty$, by assumption.

To prove assumption (iv), we follow the arguments developed by Biau et al. [2]. For completeness, these arguments are recalled here. Let us consider the partition associated with the random tree $\mathcal{T}_n(\Theta)$. By definition, this partition has 2^k cells, denoted by $A_1, \ldots, A_{2^k}$. For $1 \leq i \leq 2^k$, let N_i be the number of points among $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ falling into A_i. Finally, set $\mathcal{S} = \{\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Since these points are independent and identically distributed, fixing the set $\mathcal{S}$ (but not the order of the points) and Θ, the probability that $\mathbf{X}$ falls in the ith cell is $N_i/(n+1)$. Thus, for every fixed $t > 0$,

$$\begin{aligned}\mathbb{P}\left[N_n(\mathbf{X},\Theta) < t\right] &= \mathbb{E}\left[\mathbb{P}\left[N_n(\mathbf{X},\Theta) < t \,\middle|\, \mathcal{S}, \Theta\right]\right]\\ &= \mathbb{E}\left[\sum_{i:N_i<t+1} \frac{N_i}{n+1}\right]\\ &\leq \frac{2^k}{n+1}t.\end{aligned}$$

Thus, by assumption, $N_n(\mathbf{X},\Theta) \to \infty$ in probability, as $n \to \infty$. Consequently, observe that

$$\begin{aligned}\sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) &= \mathbb{E}_\Theta\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X},\Theta)\right]\\ &= \mathbb{E}_\Theta\left[\mathbb{1}_{N_n(\mathbf{X},\Theta)\neq 0}\right]\\ &= \mathbb{P}_\Theta\left[N_n(\mathbf{X},\Theta) \neq 0\right]\\ &\to 1 \quad \text{as } n \to \infty.\end{aligned}$$

At last, to prove (v), note that,

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}^{\infty}(\mathbf{X})\right] &\leq \mathbb{E}\left[\max_{1\leq i\leq n} \frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)}}{N_n(\mathbf{X},\Theta)}\right]\\ &\leq \mathbb{E}\left[\frac{\mathbb{1}_{N_n(\mathbf{X},\Theta)>0}}{N_n(\mathbf{X},\Theta)}\right]\\ &\to 0 \quad \text{as } n \to \infty,\end{aligned}$$

since $N_n(\mathbf{X},\Theta) \to \infty$ in probability, as $n \to \infty$. □

Proof of Corollary 1. We check conditions of Theorem 4.1. Let us denote by $V_{nj}(\mathbf{X},\Theta)$ the length of the jth side of the cell containing $\mathbf{X}$ and $K_{nj}(\mathbf{X},\Theta)$ the number of times the cell containing $\mathbf{X}$ is cut along the j-coordinate. Note that, if $U_1, \ldots, U_n$ are independent uniform on $[0, 1]$,

$$\begin{aligned}\mathbb{E}\left[V_{nj}(\mathbf{X},\Theta)\right] &\leq \mathbb{E}\left[\mathbb{E}\left[\prod_{l=1}^{K_{nj}(\mathbf{X},\Theta)} \max(U_i, 1-U_i) | K_{nj}(\mathbf{X},\Theta)\right]\right]\\ &= \mathbb{E}\left[\left[\mathbb{E}\left[\max(U_1, 1-U_1)\right]\right]^{K_{nj}(\mathbf{X},\Theta)}\right]\\ &= \mathbb{E}\left[\left(\frac{3}{4}\right)^{K_{nj}(\mathbf{X},\Theta)}\right].\end{aligned}$$

Since $K_{nj}(\mathbf{X}, \Theta)$ is distributed as a binomial $\mathcal{B}(k_n, 1/d)$, $K_{nj}(\mathbf{X}, \Theta) \to +\infty$ in probability, as n tends to infinity. Thus $\mathbb{E}\left[V_{nj}(\mathbf{X}, \Theta)\right] \to 0$ as $n \to \infty$. □

5.2. Proof of Theorem 4.2

Consider a theoretical q quantile tree where cuts are made similarly as in the q quantile tree (defined in Algorithm 1) but by selecting $q' \in [1-q, q]$ and by performing the cut at the q' theoretical quantile (instead of empirical one). The tree is then stopped at level k, where $k \in \mathbb{N}$ is a parameter to be chosen later. Denote by $A_k^\star(\mathbf{X}, \Theta)$ the cell of the theoretical q quantile tree of level k containing $\mathbf{X}$ and built with the randomness Θ. Finally, we let $\mathbf{d}_k^\star = (d_1^\star(\mathbf{X}, \Theta), \ldots, d_k^\star(\mathbf{X}, \Theta))$ be the k cuts used to construct the cell $A_k^\star(\mathbf{X}, \Theta)$.

To prove Theorem 4.2, we need the following lemma which states that the cell diameter of a theoretical q quantile tree tends to zero.

Lemma 2. *Assume that $\mathbf{X}$ has a density over $[0, 1]^d$, with respect to the Lebesgue measure. Thus, for all $q \in [1/2, 1)$, the theoretical q quantile tree defined above satisfies, for all γ,*

$$\mathbb{P}\Big[\mathrm{diam}(A_k^\star(\mathbf{X}, \Theta)) > \gamma\Big] \underset{k\to\infty}{\to} 0.$$

Proof of Lemma 2. Set $q \in [1/2, 1)$ and consider a theoretical q quantile tree. For all $A \subset [0, 1]^d$, let

$$\mu(A) = \int_A f d\nu,$$

where ν is the Lebesgue measure, and f the density of $\mathbf{X}$. Take $z \in [0, 1]$, $\ell \in \{1, \ldots, d\}$ and let Δ be the hyperplane such that $\Delta = \{\mathbf{x} : x^{(\ell)} = z\}$. At last, we denote by $\mathrm{D} = \{A : A \cap \Delta \neq \emptyset\}$ the set of cells of the theoretical q quantile tree that have a non-empty intersection with Δ.

If a cell $A_k^\star(\mathbf{X}, \Theta)$ belongs to D, then:

- **Case 1**. Either the next split in $A_k^\star(\mathbf{X}, \Theta)$ is performed along the ℓ-th coordinate and, in that case, one of the two resulting cell has an empty intersection with Δ. Note that the measure of this cell is, at least, $(1-q)\mu(A_k^\star(\mathbf{X}, \Theta))$.
- **Case 2**. Or the next split is performed along the jth coordinate (with $j \neq \ell$) and, in that case, the two resulting cells have a non-empty intersection with Δ.

Since the splitting directions are chosen uniformly over $\{1, \ldots, d\}$, for each cell *Case 1* occurs with probability $1/d$ and *Case 2* with probability $1 - 1/d$. Let $j_k(\mathbf{X}, \Theta)$ be the random variable equal to the coordinate along which the split in the cell $A_{k-1}^\star(\mathbf{X}, \Theta)$ is performed. Thus,

$$\begin{aligned}
\mathbb{P}\left[A_{k+1}^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right] &\leq \mathbb{E}\Big[\mathbb{P}\left[A_{k+1}^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right] \Big| j_{k+1}(\mathbf{X}, \Theta)\Big] \\
&\leq \mathbb{E}\Big[\mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right](1-q)\mathbb{1}_{j_{k+1}(\mathbf{X},\Theta)=\ell} + \mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right]\mathbb{1}_{j_{k+1}(\mathbf{X},\Theta)\neq\ell}\Big] \\
&\leq \mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right] \times \Big((1-q)\mathbb{P}\left[j_{k+1}(\mathbf{X}, \Theta) = \ell\right] + \mathbb{P}\left[j_{k+1}(\mathbf{X}, \Theta) \neq \ell\right]\Big) \\
&\leq \left(1 - \frac{q}{d}\right)\mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right].
\end{aligned}$$

Consequently, for all k,

$$\mathbb{P}\left[A_{k+1}^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right] \leq \left(1 - \frac{q}{d}\right)^k \mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right], \tag{10}$$

that is

$$\mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \in \mathrm{D}\right] \underset{k\to\infty}{\to} 0. \tag{11}$$

To finish the proof, take $\varepsilon > 0$ and consider a $\varepsilon \times \cdots \times \varepsilon$ grid. Within a grid cell, all points are distant from, at most, $\varepsilon d^{1/2}$. Thus, if a cell A of the median tree is contained in a grid cell, it satisfies

$$\mathrm{diam}(A) \leq \varepsilon d^{1/2}.$$

Consider the collection of hyperplanes that correspond to the grid, that is all hyperplanes of the form $\{x : x^{(\ell)} = j\varepsilon\}$ for $\ell \in \{1, \ldots, d\}$ and $j \in \{0, \ldots, \lfloor 1/\varepsilon \rfloor\}$. Denote by Δ_{grid} the collection of these hyperplanes. Since the number of hyperplanes is finite, according to (11), we have

$$\mathbb{P}\left[\mathrm{diam}(A_k^\star(\mathbf{X}, \Theta)) \geq \varepsilon d^{1/2}\right] \leq \mathbb{P}\left[A_k^\star(\mathbf{X}, \Theta) \cap \Delta_{grid} \neq \emptyset\right] \underset{k\to\infty}{\to} 0,$$

which concludes the proof. □

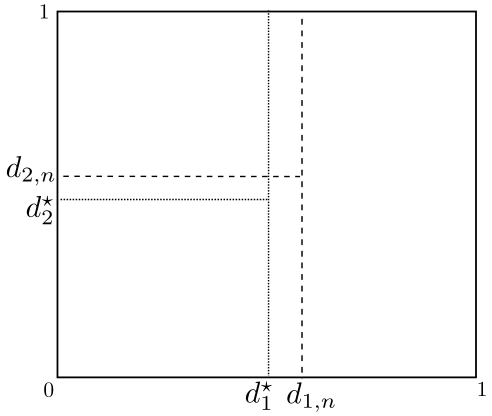

Fig. 1. Respective positions of theoretical and empirical splits in a median tree.

Recall that $A_n(\mathbf{X},\Theta)$ is the cell of the q quantile tree containing $\mathbf{X}$. Similarly, $A_{k,n}(\mathbf{X},\Theta)$ is the cell of the q quantile tree containing $\mathbf{X}$ where only the first k cuts ($k \in \mathbb{N}^\star$) are performed. We denote by $\mathbf{d}_{k,n} = (d_{1,n}(\mathbf{X},\Theta), \ldots, d_{k,n}(\mathbf{X},\Theta))$ the k cuts used to construct the cell $A_{k,n}(\mathbf{X},\Theta)$.

Lemma 3. *Assume that* $\mathbf{X}$ *has a density over* $[0,1]^d$*, with respect to the Lebesgue measure. Thus, for all* $k \in \mathbb{N}$*, a.s.*

$$\|\mathbf{d}_{k,n} - \mathbf{d}_k^\star\|_\infty \underset{n\to\infty}{\to} 0.$$

Proof of Lemma 3. To keep the argument simple, we fix $\mathbf{X} \in [0,1]^d$ and assume that the first and second splits are performed at the empirical median along the first (resp. second) coordinate. Since $\mathbf{X}$ and Θ are fixed, we omit the dependency in $\mathbf{X}$ and Θ in the rest of the proof. Let $d_{1,n}$ (resp. $d_{2,n}$) be the position of the first (resp. second) splits along the first (resp. second) axis. We denote by $d_1^\star$ (resp. $d_2^\star$) the position of the theoretical median of the distribution (see Fig. 1).

Fix $\varepsilon > 0$. Since the X_i's are i.i.d., the empirical median tends to the theoretical median almost surely. With our notation, a.s., $d_{1,n} \to d_1^\star$, as n tends to infinity. Therefore, Lemma 3 holds for $k = 1$. We now prove Lemma 3 for $k = 2$. To this end, we define, for all $0 \le a < b \le 1$, the subset $H_{a,b}$ of the cell $A_1^\star$ by

$$H_{a,b} = [0,1] \times [a,b] \times [0,1] \times \cdots \times [0,1] \cap A_1^\star.$$

Let $\alpha = \min(\mu(H_{d_2^\star-\varepsilon,d_2^\star}), \mu(H_{d_2^\star,d_2^\star+\varepsilon}))$. Denote by $d_{2,n}(d_1^\star)$ the empirical median of data points falling into the cell $A_1^\star$. Since $\mathbf{X}$ has a density on $[0,1]^d$, one can find ε_1 such that, for all n large enough, a.s.,

$$\begin{cases} |d_{2,n}(d_1^\star) - d_2^\star| \le \varepsilon_1 \\ \min(\mu(H_{d_2^\star-\varepsilon_1,d_2^\star}), \mu(H_{d_2^\star,d_2^\star+\varepsilon_1})) \le \alpha/100. \end{cases}$$

By the same argument, one can find ε_2 such that, for all n large enough, a.s.,

$$\begin{cases} |d_{1,n} - d_1^\star| \le \varepsilon_2 \\ \min(\mu(H_{d_1^\star-\varepsilon_2,d_1^\star}), \mu(H_{d_1^\star,d_1^\star+\varepsilon_2})) \le \alpha/100. \end{cases}$$

A direct consequence of the law of the iterated logarithm applied to cumulative distribution function is that, for all n large enough, a.s.,

$$\max\big(N_n(H_{d_2^\star-\varepsilon_1,d_2^\star}), N_n(H_{d_2^\star,d_2^\star+\varepsilon_1})\big) \le 0.02\alpha n, \tag{12}$$

$$\max\big(N_n(H_{d_1^\star-\varepsilon_2,d_1^\star}), N_n(H_{d_1^\star,d_1^\star+\varepsilon_2})\big) \le 0.02\alpha n, \tag{13}$$

$$\text{and} \quad \min\big(N_n(H_{d_2^\star-\varepsilon,d_2^\star-\varepsilon_1}), N_n(H_{d_2^\star+\varepsilon_1,d_2^\star+\varepsilon})\big) \ge 0.98\alpha n. \tag{14}$$

The empirical median in the cell $A_{1,n}$ is given by $X^{(2)}_{(\lfloor N_n(A_{1,n})/2\rfloor)}$, where the $\mathbf{X}_i$'s are sorted along the second coordinate. According to (13), the cell $A_1^\star$ contains at most $N_n(A_{1,n}) + 0.02\alpha n$. Therefore, the empirical median $d_{2,n}(d_1^\star)$ in the cell $A_1^\star$ is at most $X^{(2)}_{(\lfloor (N_n(A_{1,n})+0.02\alpha n)/2\rfloor)}$. Thus, according to (12) and (14), for all n large enough, a.s.,

$$d_{2,n} \le X^{(2)}_{(\lfloor (N_n(A_{1,n})+0.02\alpha n)/2\rfloor)} \le d_2^\star + \varepsilon.$$

Similarly, for all n large enough, a.s.,

$$d_{2,n} \geq X^{(2)}_{(\lfloor (N_n(A_{1,n})-0.02\alpha n)/2\rfloor)} \geq d_2^\star - \varepsilon.$$

Consequently, for all n large enough, a.s., $|d_{2,n} - d_2^\star| \leq \varepsilon$. The extension for arbitrary k is straightforward. □

Lemma 4. *Assume that* $\mathbf{X}$ *has a density over* $[0,1]^d$, *with respect to the Lebesgue measure. Thus, for all* $q \in [1/2, 1)$, *the theoretical* q *quantile tree defined above satisfies, for all* γ,

$$\mathbb{P}\Big[\mathrm{diam}(A_n(\mathbf{X},\Theta)) > \gamma\Big] \underset{n\to\infty}{\to} 0.$$

Proof. Now, consider the empirical q quantile tree as defined in Algorithm 1 but stopped at level k. Thus, for n large enough, at each step of the algorithm, q' is selected in $[1-q, q]$. Set $\varepsilon, \gamma > 0$. By Lemma 2, there exists $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$,

$$\mathbb{P}\Big[\mathrm{diam}(A_k(\mathbf{X},\Theta)) > \gamma\Big] \leq \varepsilon.$$

Thus, according to Lemma 3, for all n large enough, a.s.,

$$\mathbb{P}\Big[\mathrm{diam}(A_{k_0,n}(\mathbf{X},\Theta)) > \gamma/2\Big] \leq \varepsilon.$$

Since, for all n large enough, a.s.,

$$\mathrm{diam}(A_{k_0,n}(\mathbf{X},\Theta)) \geq \mathrm{diam}(A_n(\mathbf{X},\Theta)),$$

the proof is complete. □

Proof of Theorem 4.2. We check the conditions of Stone's theorem. Condition (i) is satisfied since the regression function is uniformly continuous and $\mathrm{Var}[Y|\mathbf{X}=\mathbf{x}] \leq \sigma^2$ (see remark after Stone's theorem in [19]).

Condition (ii) is always satisfied for random trees. Condition (iii) is verified since

$$\mathbb{P}\left[\mathrm{diam}(A_n(\mathbf{X},\Theta)) > \gamma\right] \underset{n\to\infty}{\to} 0,$$

according to Lemma 4.

Since each cell contains exactly one data point,

$$\begin{aligned}\sum_{i=1}^n W_{ni}(x) &= \sum_{i=1}^n \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)}}{N_n(\mathbf{X},\Theta)}\right]\\ &= \mathbb{E}_\Theta\left[\frac{1}{N_n(\mathbf{X},\Theta)}\sum_{i=1}^n \mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)}\right]\\ &= 1.\end{aligned}$$

Thus, conditions (iv) of the Stone theorem is satisfied.

To check (v), 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

$$\mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{\binom{a_n-1}{n-1}}{\binom{a_n}{n}} = \frac{a_n}{n}.$$

So,

$$\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(\mathbf{X})\right] \leq \mathbb{E}\left[\max_{1\leq i\leq n} \mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right]\right] \leq \frac{a_n}{n},$$

which tends to zero by assumption. □

Acknowledgments

We greatly thank two referees for valuable comments and insightful suggestions. We also thank Luc Devroye for a substantial improvement of the proof of Lemma 2.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2015.06.009.

References

[1] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[2] G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, J. Mach. Learn. Res. 9 (2008) 2015–2033.
[3] E.G. Bongiorno, A. Goia, E. Salinelli, P. Vieu, Contributions in Infinite-dimensional Statistics and Related Topics, Società Editrice Esculapio, 2014.
[4] S. Boucheron, G. Lugosi, P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, 2013.
[5] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[6] L. Breiman, Some Infinity Theory for Predictor Ensembles. Technical Report 577, UC Berkeley, 2000.
[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[8] L. Breiman, J. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.
[9] S. Clémençon, M. Depecker, N. Vayatis, Ranking forests, J. Mach. Learn. Res. 14 (2013) 39–73.
[10] A. Cutler, G. Zhao, Pert–perfect random tree ensembles, Comput. Sci. Stat. 33 (2001) 490–497.
[11] M. Denil, D. Matheson, N. de Freitas, Consistency of online random forests, 2013. arXiv:1302.4853.
[12] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[13] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, BMC Bioinform. 7 (2006) 1–13.
[14] T.G. Dietterich, E.B. Kong, Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms. Technical Report, Department of Computer Science, Oregon State University, 1995.
[15] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer Science & Business Media, 2006.
[16] R. Genuer, J.-M. Poggi, C. Tuleau, Random forests: some methodological insights, 2008. arXiv:0811.3619.
[17] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[18] B. Gregorutti, B. Michel, P. Saint-Pierre, Grouped variable importance with random forests and application to multivariate functional data analysis, 2014. arXiv:1411.4170.
[19] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer, New York, 2002.
[20] M. Hamza, D. Laroque, An empirical comparison of ensemble methods based on classification trees, J. Stat. Comput. Simul. 75 (2005) 629–643.
[21] T. Ho, The random subspace method for constructing decision forests, Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[22] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer Science & Business Media, 2012.
[23] H. Ishwaran, The effect of splitting on random forests, Mach. Learn. (2013) 1–44.
[24] H. Ishwaran, U.B. Kogalur, Consistency of random survival forests, Statist. Probab. Lett. 80 (2010) 1056–1064.
[25] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, Random survival forest, Ann. Appl. Stat. 2 (2008) 841–860.
[26] B. Lakshminarayanan, D.M. Roy, Y.W. Teh, Mondrian forests: Efficient online random forests, 2014. arXiv:1406.2673.
[27] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (2002) 18–22.
[28] N. Meinshausen, Quantile regression forests, J. Mach. Learn. Res. 7 (2006) 983–999.
[29] L. Mentch, G. Hooker, Ensemble trees and clts: Statistical inference for supervised learning, 2014. arXiv:1404.6473.
[30] J.-M. Poggi, C. Tuleau, Classification supervisée en grande dimension. application à l'agrément de conduite automobile, Rev. Stat. Appl. 54 (2006) 41–60.
[31] Y. Qi, Ensemble Machine Learning, Springer, 2012, pp. 307–323. Chapter Random forest for bioinformatics.
[32] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, 2005.
[33] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P.H. Torr, Randomized trees for human pose detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008, IEEE, 2008, pp. 1–8.
[34] E. Scornet, G. Biau, J.-P. Vert, Consistency of random forests, 2014. arXiv:1405.2881.
[35] C.J. Stone, Consistent nonparametric regression, Ann. Statist. 5 (1977) 595–645.
[36] Aad W. van der Vaart, Jon A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer, New York, 1996.
[37] S. Wager, Asymptotic theory for random forests, 2014. arXiv:1405.0352.
[38] S. Wager, T. Hastie, B. Efron, Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, J. Mach. Learn. Res. 15 (2014) 1625–1651.
[39] R. Zhu, D. Zeng, M.R. Kosorok, Reinforcement Learning Trees. Technical Report, University of North Carolina, 2012.