Random forest estimation of conditional distribution functions and conditional quantiles

Kevin Elie-Dit-Cosaque, Véronique Maume-Deschamps

Random forest estimation of conditional distribution functions and conditional quantiles

Kévin Elie-Dit-Cosaque[1, 2] and Véronique Maume-Deschamps[1]

[1]Institut Camille Jordan, Université Claude Bernard Lyon 1, Lyon, FRANCE
[2]Actuarial Modelling, SCOR, Paris, FRANCE

June 2, 2020

Abstract

We propose a theoretical study of two realistic estimators of conditional distribution functions and conditional quantiles using random forests. The estimation process uses the bootstrap samples generated from the original dataset when constructing the forest. Bootstrap samples are reused to define the first estimator, while the second requires only the original sample, once the forest has been built. We prove that both proposed estimators of the conditional distribution functions are consistent uniformly a.s. To the best of our knowledge, it is the first proof of consistency including the bootstrap part. We also illustrate the estimation procedures on a numerical example.

1 Introduction

Conditional distribution functions and conditional quantiles estimation is an important task in several domains including environment, insurance or industry. It is also an important tool for Quantile Oriented Sensitivity Analysis (QOSA), see e.g., Fort et al. (2016); Maume-Deschamps and Niang (2018); Browne et al. (2017). In order to estimate conditional quantiles, various methods exist such as kernel based estimation or quantile regression (Koenker and Hallock, 2001) but they present some limitations. Indeed, the performance of kernel methods strongly depends on the bandwidth parameter selection and quickly break down as the number of covariates increases. On one other hand, quantile regression is not adapted in a non-gaussian setting since the true conditional quantile is not necessarily a linear comination of the input variables (Maume-Deschamps et al., 2017). To overcome these issues, we propose to explore the Random Forest estimation of conditional quantiles (Meinshausen, 2006).

Random forest algorithms allow a flexible modeling of interactions in high dimension by building a large number of regression trees and averaging their predictions. The most famous random forest algorithm is that of Breiman (2001) whose construction is based on the seminal work of Amit and Geman (1997); Ho (1998); Dietterich (2000). Breiman's random forest estimate is a combination of two essential components: Bagging and Classification And Regression Trees (CART)-split criterion (Breiman et al., 1984). Bagging for *bootstrap-aggregating* was proposed by Breiman (1996) in order to improve the performance of weak or unstable learners.

Random forests are also related to some local averaging algorithms such as nearest neighbors methods (Lin and Jeon, 2006; Biau and Devroye, 2010) or kernel estimators (Scornet, 2016c). More precisely, thanks to Lin and Jeon (2006), the random forest method can be seen as an adaptive neighborhood regression procedure and therefore the prediction (estimation of the conditional mean) can be formulated as a weighted average of the observed response variables.

Based on that approach, we develop a Weighted Conditional Empirical Cumulative Distribution Function (W_C_ECDF) approximating the Conditional Cumulative Distribution Function (C_CDF). Then, α-quantile estimates are obtained by using W_C_ECDF instead of C_CDF. Meinshausen (2006) defined a W_C_ECDF with weights using the original dataset whereas we allow to construct the weights using the bootstrap samples, as it is done practically in regression random forests. We prove the almost sure consistency of these estimators. Both estimators have several advantages over methods such as kernel methods (Nadaraya, 1964; Watson, 1964). Due to the intrinsic tree building process, random forest estimators can easily handle both univariate and multivariate data with few parameters to tune. Besides, these methods have good predictive power and can outperform standard kernel methods (Davies and Ghahramani, 2014; Scornet, 2016c). Lastly, being based on the random forest algorithm, they are also easily parallelizable and can handle large dataset. A implementation of both algorithms is made available within a `Julia` package called `ConditionalDistributionForest` (Fabrège and Maume-Deschamps, 2020) as well as a `python` package named `qosa-indices`, (Elie-Dit-Cosaque, 2020).

The C_CDF can be seen as a regression function. On this basis, we were interested in the literature dealing with the consistency of random forest estimates in order to show the convergence of ours estimators.
Several authors such as Breiman (2004); Biau (2012); Wager and Walther (2015); Scornet et al. (2015b); Mentch and Hooker (2016); Wager and Athey (2018); Goehry (2019) have established asymptotic properties of particular variants and simplifications of the original Breiman's random forest algorithm. Facing some theoretical issues with the bootstrap, most studies replace it by subsampling, assuming that each tree is grown with $s_n < n$ observations randomly chosen without replacement from the original dataset. Most of the time, in order to ensure the convergence of the simplified model, the subsampling rate s_n/n is assumed tend to zero at some prescribed rate, assumption that excludes the bootstrap mode. Besides, the consistency is generally showed by assuming that the number of trees goes to infinity which is not fully relevant in practice. Under some conditions, Scornet (2016a) showed that if the infinite random forest regression estimator is $\mathbb{L}^2$ consistent then so does the finite random forest regression estimator when the number of trees goes to infinity in a controlled way.
Recent attempts to bridge the gap between theory and practice provide some results on random forest algorithms at the price of fairly strong conditions. For example, Scornet et al. (2015b) showed the $\mathbb{L}^2$ consistency of random forests in an additive regression framework by replacing the bootstrap step by subsampling. Their result rests on a fundamental lemma developped in Scornet et al. (2015a) which reviews theoretical random forest. Highlighted by a counterexample developed in Section 4, assumptions are required to get out of the additive framework. Furthermore, consistency and asymptoctic normality of the whole algorithm were recently proved under strong conditions by Wager and Athey (2018) replacing bootstrap by subsampling and simplifying the splitting step. One of the strong conditions used in the Theorem 3.1. of Wager and Athey (2018) is that the individual trees satisfy a condition called *honesty*. An example of an honest tree given by the authors is *one where the tree is grown using one subsample, while the predictions at the leaves of the tree are estimated using a different subsample.* Due to this assumption, the authors admit that their theorems are not valid for the practical applications most of the time because almost all implementations of random forests use the training sample twice.

Thus, despite an active investigation during the last decade, the consistency of the original (i.e. with the bootstrap samples) Breiman's random forest method is not fully proved. This motivated our work.
Our major contribution is the proof of the almost everywhere uniform convergence of the es-

timator W_C_ECDF both using the bootstrap samples (Theorem 4.1) or the original one (Theorem 4.2). To the best of our knowledge, this is the first consistency result under realistic assumptions for a method based on bootstrap samples in the random forest field. Remark that Meinshausen (2006) gave of proof of the consistency in probability of the W_C_ECDF for a simplified model where the weights are considered as constant while they are indeed random variables heavily data-dependent.

The paper is organized as follows. Breiman's random forest algorithm is detailed in Section 2 and notations are stated. The random forest estimations of C_CDF based both on bootstrap samples and the original dataset are introduced in Section 3 as a natural generalization of regression random forests. The main consistency results are presented in Section 4 and the proofs of those are gathered in Section 5. Section 6 is devoted to a short simulation study and a conclusion is given in Section 7.

2 Breiman's random forest

The aim of this section is to present the Breiman's random forest algorithm as well as notations used throughout this paper.

Random forest is a generic term to name an aggregation scheme of decision trees allowing to deal with both supervised classification and regression tasks. We are only concerned with the regression task.
The general framework is the nonparametric regression estimation where an input random vector $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ is observed and a response $Y \in \mathbb{R}$ is predicted by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\,\mathbf{X} = \mathbf{x}]$. We assume that we are given a training sample $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ of independent random variables distributed as the prototype pair $(\mathbf{X}, Y)$ which is a $(d+1)$-dimensional random vector. The purpose is to use the dataset $\mathcal{D}_n$ to construct an estimator $m_n : \mathcal{X} \mapsto \mathbb{R}$ of the function m.

Random forests proposed by Breiman (2001) build a predictor consisting of a collection of k randomized regression trees grown based on the CART algorithm.
The CART-split criterion of Breiman et al. (1984) is used in the construction of the individual trees to recursively partition the input space $\mathcal{X}$ in a dyadic manner. More precisely, at each step of the partitioning, a part of the space is divided into two sub-parts according to the best cut perpendicular to the axes. This best cut is selected in each node of the tree by optimizing the CART-split criterion over the d variables, i.e. minimizing the prediction squared error in the two child nodes. The trees are thus grown until reaching a stopping rule. There are several, but one that is generally proposed is that the tree construction continues while leaves contain at least $min\_samples\_leaf$ elements. This criterion is implemented in the `RandomForestRegressor` class of the `python` package `Scikit-Learn` (Pedregosa et al., 2011) or in the `build_forest` function of the `Julia` (Bezanson et al., 2017) package `DecisionTree`.

Building several different trees from a single dataset requires to randomize the tree building process. Breiman (2001) proposed to inject some randomness both in the dataset and in the tree construction. First of all, prior to the construction of each tree, a resampling step is done by bootstrapping (Efron, 1979) from the original dataset, that is, by choosing uniformly at random n times from n observations with replacement. Only these bootstrap observations are taken into account in the tree building. Accordingly, the $min\_samples\_leaf$ hyperparameter introduced previously refers in the random forest method to the minimum number of bootstrap observations contained in each leaf of a tree. Secondly, at each step of the tree construction, instead of optimizing the CART-split criterion over the d variables, a number of variables called

$max\_features$ is selected uniformly at random among the d variables. Then, the best split is chosen as the one optimizing the CART-split criterion only along the $max\_features$ preselected variables in each node.

For any query point $\mathbf{x} \in \mathcal{X}$, the ℓ-th tree estimates $m(\mathbf{x})$ as follows:

$$m_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right) = \sum_{j \in \mathcal{D}_n^\star(\Theta_\ell)} \frac{\mathbb{1}_{\{\mathbf{X}^j \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\}}}{N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)} Y^j \tag{2.1}$$

where:

- $\Theta_\ell, \ell = 1, \ldots, k$ are independent random vectors, distributed as a generic random vector $\Theta = (\Theta^1, \Theta^2)$ and independent of $\mathcal{D}_n$. Θ^1 contains indexes of observations that are used to build each tree, i.e. the bootstrap sample and Θ^2 indexes of splitting candidate variables in each node.
- $\mathcal{D}_n^\star(\Theta_\ell)$ is the bootstrap sample selected prior to the tree construction.
- $A_n\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$ is the tree cell (subspace of $\mathcal{X}$) containing $\mathbf{x}$.
- $N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$ is the number of elements of $\mathcal{D}_n^\star(\Theta_\ell)$ that fall into $A_n\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$.

The trees are then combined to form the finite forest estimator:

$$m_{k,n}^b\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n\right) = \frac{1}{k} \sum_{\ell=1}^{k} m_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right) \tag{2.2}$$

We may now present the conditional distribution function estimators.

3 Conditional Distribution Forests

We aim to estimate $F(y|\mathbf{x}) = \mathbb{P}(Y \leqslant y|\mathbf{X} = \mathbf{x})$. Two estimators may be defined. One uses the bootstrap samples both in the forest construction and in the estimation. The other uses the original sample in the estimation part. Once the distribution function has been estimated, the conditional quantiles may be estimated straightforwardly.

3.1 Bootstrap samples based estimator

First of all, let us define the random variable $B_j\left(\Theta_\ell^1, \mathcal{D}_n\right)$ as the number of times that the observation $(\mathbf{X}^j, Y^j)$ has been drawn from the original dataset for the ℓ-th tree construction. Thanks to it, the conditional mean estimator in Equation (2.2) may be rewritten as:

$$\begin{aligned} m_{k,n}^b\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n\right) &= \sum_{j=1}^{n} \left(\frac{1}{k} \sum_{\ell=1}^{k} \frac{B_j\left(\Theta_\ell^1, \mathcal{D}_n\right) \mathbb{1}_{\{\mathbf{X}^j \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\}}}{N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)} \right) Y^j \\ &= \sum_{j=1}^{n} w_{n,j}^b\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n\right) Y^j \end{aligned} \tag{3.1}$$

where the weights are defined by:

$$w_{n,j}^b\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n\right) = \frac{1}{k} \sum_{\ell=1}^{k} \frac{B_j\left(\Theta_\ell^1, \mathcal{D}_n\right) \mathbb{1}_{\{\mathbf{X}^j \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\}}}{N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)}. \tag{3.2}$$

Note that the weights $w_{n,j}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)$ are nonnegative random variables as functions of $\Theta_1,\ldots,\Theta_k,\mathcal{D}_n$ and their sum for $j=1,\ldots,n$ equals 1.

The random forest estimator (3.1) can be seen as a local averaging estimate. Indeed, as mentioned by Scornet (2016b), the regression trees make an average of the observations located in a neighborhood of $\mathbf{x}$, this neighborhood being defined as the leaf of the tree containing $\mathbf{x}$. The forest, which aggregates several trees, also operates by calculating a weighted average of the observations in a neighborhood of $\mathbf{x}$. However, in the case of forests, this neighborhood results from the superposition of the neighborhoods of each tree, and therefore has a more complex shape. Several works have tried to study the random forest algorithm from this point of view (local averaging estimate) such as Lin and Jeon (2006) who was the first to point out the connection between the random forest and the adaptive nearest-neighbors methods, further developped by Biau and Devroye (2010). Some works such as Scornet (2016c) have also studied random forests through their link with the kernel methods.

We are interested in the Conditional Cumulative Distribution Function (C_CDF) of Y given $\mathbf{X}=\mathbf{x}$ in order to obtain the conditional quantiles. Pairing the following equality:

$$F\left(y|\,\mathbf{X}=\mathbf{x}\right)=\mathbb{P}\left(Y\leqslant y|\,\mathbf{X}=\mathbf{x}\right)=\mathbb{E}\left[\mathbb{1}_{\{Y\leqslant y\}}\middle|\,\mathbf{X}=\mathbf{x}\right] \tag{3.3}$$

with the weighted approach described above, we propose to estimate the C_CDF as follows:

$$F_{k,n}^{b}\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\sum_{j=1}^{n}w_{n,j}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)\mathbb{1}_{\{Y^j\leqslant y\}} \tag{3.4}$$

Hence, given a level $\alpha\in\left]0,1\right[$, the conditional quantile estimator $\widehat{q}^{\,\alpha}\left(Y|\,\mathbf{X}=\mathbf{x}\right)$ is defined as follows:

$$\widehat{q}^{\,\alpha}\left(Y|\,\mathbf{X}=\mathbf{x}\right)=\inf\left\{Y^p,p=1,\ldots,n:F_{k,n}^{b}\left(Y^p|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)\geqslant\alpha\right\}$$

Let us turn now to the estimator using the original sample.

3.2 Original sample based estimator

Trees are still grown with their respective bootstrap sample $\mathcal{D}_n^{\star}\left(\Theta_\ell\right),\ell=1,\ldots,k$. But instead of considering them in the estimation, we may use the original sample $\mathcal{D}_n$. Consider the weights:

$$w_{n,j}^{o}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\frac{1}{k}\sum_{\ell=1}^{k}\frac{\mathbb{1}_{\{\mathbf{X}^j\in A_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)\}}}{N_n^{o}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}. \tag{3.5}$$

where $N_n^{o}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)$ is the number of points of $\mathcal{D}_n$ that fall into $A_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)$. As previously, the weights $w_{n,j}^{o}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)$ are nonnegative random variables as functions of $\Theta_1,\ldots,\Theta_k,\mathcal{D}_n$ and their sum over $j=1,\ldots,n$ equals 1.

It was proposed in Meinshausen (2006) to estimate the C_CDF with:

$$F_{k,n}^{o}\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\sum_{j=1}^{n}w_{n,j}^{o}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)\mathbb{1}_{\{Y^j\leqslant y\}} \tag{3.6}$$

The conditional quantiles are then estimated by plugging $F_{k,n}^{o}\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)$ instead of $F\left(Y|\,\mathbf{X}=\mathbf{x}\right)$ as before.

Algorithm 1: Conditional Distribution Forest algorithm

Input:

- Training sample: $\mathcal{D}_n$
- Number of trees: $k \in \mathbb{N}^\star$
- Number of features to be considered for the best split in a node: $max\_features \in \{1, \ldots, d\}$
- Minimum number of samples required in a leaf node: $min\_samples\_leaf \in \{1, \ldots, n\}$
- Point where the conditional distribution function or the conditional quantile is required: $\mathbf{x} \in \mathcal{X}$
- The order of the conditional quantile: $\alpha \in [0, 1]$

Output: Estimated value of the conditional quantile of $\mathbf{x}$ at the α-order.

1 **for** $\ell = 1, \ldots, k$ **do**
2 Select uniformly with replacement n data points among $\mathcal{D}_n$. Only these observations will be used in the tree construction.
3 **begin** Tree construction
4 Consider the whole space $\mathcal{X}$ as root node.
5 Select uniformly without replacement $max\_features$ coordinates among $\{1, \ldots, d\}$.
6 Select the split minimizing the CART-split criterion (Breiman et al., 1984; Biau and Scornet, 2016) along the pre-selected $max\_features$ directions.
7 Cut the current node at the selected split in two child nodes.
8 Repeat the lines (5)-(7) for the two resulting nodes until each node of the tree contains at least $min\_samples\_leaf$ observations.
9 **end**
10 Save in which leaf node of the tree fall each observation of the training sample $\mathcal{D}_n$.
11 **end**
12 Drop $\mathbf{x}$ through all trees and calculate for each observation in $\mathcal{D}_n$ its weigthed average through the forest as in Equation (3.2) or (3.5) according to the estimator used.
13 Sort the calculated weights according to $\left(Y^{(j)}\right)_{j=1,\ldots,n}$ that are the order statistics of $\left(Y^j\right)_{j=1,\ldots,n}$.
14 Compute the cumulative sum of the sorted weigths which gives us a Weighted Conditional Empirical Cumulative Distribution Function (W_C_ECDF).
15 Get the α-conditional quantile of $\mathbf{x}$ thanks to the W_C_ECDF.

A complete description of the procedure for computing conditional quantile estimates via the C_CDF with both previous estimators can be found in Algorithm 1. A `python` library named `qosa-indices` has also been developed to perform the numerical estimations of conditional distributions and quantiles for both methods. It is available at Elie-Dit-Cosaque (2020) and uses `Scikit-Learn, Numpy, Numba`. Both approaches are also implemented in a `Julia` package based on the library `DecisionTree` and that is available at Fabrège and Maume-Deschamps (2020).

It has to be noted that a package called `quantregForest` has been made available in `R` (R Core Team, 2019) and can be found at Meinshausen (2019). The estimation method currently implemented in `quantregForest` is different from the method described in Meinshausen (2006). It does the following. For a new observation $\mathbf{x}$ and the ℓ-th tree, one element of $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ falling into in the leaf node $A_n\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$ is chosen at random. This gives, k values of Y and allows to estimate the conditional distribution function with the classical Empirical Cumulative Distribution Function associated with the empirical measure. The performance of this method seems weak and no theoretical guarantees are available.

4 Consistency results

In this section, we state our main results, which are the uniform a.s. consistency of both estimators $F^b_{k,n}$ and $F^o_{k,n}$ of the conditional distribution function. It constitutes the most interesting result of this paper because it handles the bootstrap component and gives the almost sure uniform convergence. Indeed, most of the studies (Scornet et al., 2015b; Wager and Athey, 2018; Goehry, 2019) replace the bootstrap by subsampling without replacement in order to avoid the mathematical difficulties induced by this one and therefore differ slightly from the procedure used in practice.

Meinshausen (2006) showed the uniform convergence in probability of a simplified version of the estimator $F^o_{k,n}$. In Meinshausen (2006), the weights $w^o_{n,j}\left(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n\right)$ are in fact considered to be non-random while they are indeed random variables depending on $(\Theta_\ell)_{\ell=1,\dots,k}$ and $\mathcal{D}_n$.

Overall, proving the consistency of the forest methods whose construction depends both on the $\mathbf{X}^j$'s and on the Y^j's is a difficult task. This feature makes the resulting estimate highly data-dependent, and therefore difficult to analyze. A simplification widely used by most authors from a theoretical point of view is to work with random forest estimates whose form of the tree depends only on $\mathbf{X}^j$'s which Devroye et al. (2013) called the $\mathbf{X}$-property but the Y^j's are still used to compute the prediction, either the conditional mean or the conditional distribution function for example. One of the first results dealing with data-dependent random forest estimator of the regression function is Scornet et al. (2015b) who showed the $\mathbb{L}^2$ consistency in an additive regression framework by replacing the bootstrap by subsampling. Thanks to the following assumptions, we go further by showing the consistency of our estimators in a general framework and not only in the additive regression scheme.

Assumption 4.1.
For all $\ell \in [\![1, k]\!]$, we assume that the variation of the conditional cumulative distribution function within any cell goes to 0:

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} \left| F\left(y \middle| \mathbf{z}\right) - F\left(y \middle| \mathbf{x}\right) \right| \xrightarrow[n \to \infty]{a.s.} 0$$

We shall discuss further on Assumption 4.1 but let us remark that it is satisfied, for example, provided that the diameter of each tree cell goes to zero and for all y, $F(y|\cdot)$ is continuous.

Assumption 4.2.
We shall make the following assumptions on k (number of trees) and $N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)$ (number of bootstrap observations in a leaf node):

1. $k = \mathcal{O}(n^\alpha)$, with $\alpha > 0$.
2. $\forall \mathbf{x} \in \mathcal{X}, \quad N_n^b(\mathbf{x};\Theta,\mathcal{D}_n) = \Omega^1\left(\sqrt{n}\,(\ln(n))^\beta\right)$, with $\beta > 1$, *a.s.*

   or
3. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)\right] = \Omega\left(\sqrt{n}\,(\ln(n))^\beta\right)$, with $\beta > 1$, and
   $\forall \mathbf{x} \in \mathcal{X}, \quad \mathrm{CV}^2\left(N_n^b(\mathbf{x};\Theta,\mathcal{D}_n)\right) = \mathcal{O}\left(\dfrac{1}{\sqrt{n}\,(\ln(n))^{\gamma/2}}\right)$, with $\gamma > 1$.

Remark 4.1.
In order to prove our main consistency result, either Assumption 4.2 item 2. or item 3. is needed. Item 2. may seem much stronger than item 3. but it has to be noted that the number of bootstrap observations in a tree leaf is a construction parameter of the forest, so that it can be controlled. Using item 2. simplifies the proof but item 3. is sufficient.

Assumption 4.3.
For every $\mathbf{x} \in \mathcal{X}$, the conditional cumulative distribution function $F(y|\,\mathbf{X}=\mathbf{x})$ is continuous and strictly increasing in y.

The two theorems below give the uniform a.s. consistency of our two estimators.

Theorem 4.1.
Consider a random forest which satisfies Assumptions 4.1 to 4.3. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{y\in\mathbb{R}} \left|F_{k,n}^b(y|\,\mathbf{X}=\mathbf{x}) - F(y|\,\mathbf{X}=\mathbf{x})\right| \xrightarrow[n\to\infty]{a.s.} 0$$

Theorem 4.2.
Consider a random forest which satisfies Assumptions 4.1 to 4.3. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{y\in\mathbb{R}} \left|F_{k,n}^o(y|\,\mathbf{X}=\mathbf{x}) - F(y|\,\mathbf{X}=\mathbf{x})\right| \xrightarrow[n\to\infty]{a.s.} 0$$

Remark 4.2.
Using standard arguments, the consistency of quantile estimates stems from Assumption 4.3 as well as the uniform convergence of the conditional distribution function estimators obtained above.

Let us make some comments on the assumptions above.

Assumption 4.1 ensures a control on the approximation error of the estimators. It is drawn from the Proposition 2 of Scornet et al. (2015b) who shows the consistency of Breiman's random

[1] $f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$
[2] $\mathrm{CV}(X) = \sigma_X / \mathbb{E}[X]$

forest estimate in an additive regression framework. Their Proposition 2 allows to manage the approximation error of the estimator by showing that the variation of the regression function m within a cell of a random empirical tree is small provided n is large enough. This result is based on the fundamental Lemma 1 of Scornet et al. (2015b) which states that the variation of the regression function m within a cell of a random theoretical tree goes to zero for an additive regression model. A random theoretical tree is grown as a random empirical tree, except that the theoretical equivalent of the empirical CART-split criterion (Biau and Scornet, 2016) defined in any node A below is used to choose the best split:

$$\begin{aligned} L_A^\star(i,z) =& \operatorname{Var}\left(Y \middle| \mathbf{X} \in A\right) \\ & - \mathbb{P}\left(X_i < z \middle| \mathbf{X} \in A\right) \operatorname{Var}\left(Y \middle| X_i < z, \mathbf{X} \in A\right) \\ & - \mathbb{P}\left(X_i \geqslant z \middle| \mathbf{X} \in A\right) \operatorname{Var}\left(Y \middle| X_i \geqslant z, \mathbf{X} \in A\right) \end{aligned}$$

Hence, a theoretical tree is obtained thanks to the best consecutive cuts $(i^\star, z^\star)$ optimizing the previous criterion $L^\star(\cdot,\cdot)$.

General results on standard partitioning estimators whose construction is independent of the label in the training set (see Chapter 4 in Györfi et al. (2006) or Chapter 6 in Devroye et al. (2013)) state that a necessary condition to prove the consistency is that the diameter of the cells tend to zero as $n \to \infty$. Instead of such a geometrical assumption, Proposition 2 in Scornet et al. (2015b) ensures that the variation of m inside a node is small thanks to their Lemma 1. But the cornerstone of the Lemma 1 is the Technical Lemma 1 of Scornet et al. (2015a) recalled below for completness.

Technical Lemma.
Assume that:

- $Y = m\left(\mathbf{X}\right) + \varepsilon$ *with* $m\left(\mathbf{X}\right) = \sum_{i=1}^{d} m_i\left(X_i\right), \mathbf{X} \sim \mathcal{U}\left([0,1]^d\right)$ *and* $\varepsilon \sim \mathcal{N}(0,\sigma^2)$,
- $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i] \;\; (0 \leqslant a_i < b_i \leqslant 1)$

Then the regression function m is constant on A.

This lemma states that if the theoretical split criterion is zero for all cuts in a node, then the regression function m is constant on this node, i.e. the variation of this one within the cell is zero. But, we will see in the sequel a counterexample where $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i]$ and yet, the regression function is not constant.

First of all, let us look under which conditions $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i]$ in a node A. Remark that:

$$\begin{aligned} \operatorname{Var}\left(Y \middle| \mathbf{X} \in A\right) = & \mathbb{P}\left(X_i < z \middle| \mathbf{X} \in A\right) \operatorname{Var}\left(Y \middle| X_i < z, \mathbf{X} \in A\right) \\ & + \mathbb{P}\left(X_i \geqslant z \middle| \mathbf{X} \in A\right) \operatorname{Var}\left(Y \middle| X_i \geqslant z, \mathbf{X} \in A\right) \\ & + \mathbb{P}\left(X_i < z \middle| \mathbf{X} \in A\right) \left(\mathbb{E}\left[Y \middle| X_i < z, \mathbf{X} \in A\right] - \mathbb{E}\left[Y \middle| \mathbf{X} \in A\right]\right)^2 \\ & + \mathbb{P}\left(X_i \geqslant z \middle| \mathbf{X} \in A\right) \left(\mathbb{E}\left[Y \middle| X_i \geqslant z, \mathbf{X} \in A\right] - \mathbb{E}\left[Y \middle| \mathbf{X} \in A\right]\right)^2 \end{aligned}$$

Thus, $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i]$ if and only if $\mathbb{E}\left[Y \middle| \mathbf{X} \in A\right] = \mathbb{E}\left[Y \middle| X_i < z, \mathbf{X} \in A\right] = \mathbb{E}\left[Y \middle| X_i \geqslant z, \mathbf{X} \in A\right] \quad \forall i, \forall z \in [a_i, b_i]$.

Within a standard cell of the form $A = \prod_{i=1}^{d} A_i = \prod_{i=1}^{d} [a_i, b_i]$ as well as for a generic model of the type $Y = m\left(\mathbf{X}\right) + \varepsilon$ with $\mathbf{X}$, independent random inputs and ε, an independent centered

noise of $\mathbf{X}$, the condition above is equivalent to:

$$\mathbb{E}\left[m\left(\mathbf{X}\right)\mathbb{1}_{\{\mathbf{X}\in A\}}\right] = \mathbb{P}\left(X_i \in A_i\right)\mathbb{E}\left[m\left(\mathbf{X}_{-i}, z\right)\mathbb{1}_{\{\mathbf{X}_{-i}\in A_{-i}\}}\right] \quad \forall i, \forall z \in [a_i, b_i]$$

This result is obtained by deriving with respect to z the following function:

$$\Phi(z) = \mathbb{P}\left(X_i < z, \mathbf{X} \in A\right)\mathbb{E}\left[m\left(\mathbf{X}\right)\mathbb{1}_{\{\mathbf{X}\in A\}}\right] - \mathbb{P}\left(\mathbf{X} \in A\right)\mathbb{E}\left[m\left(\mathbf{X}\right)\mathbb{1}_{\{X_i<z, \mathbf{X}\in A\}}\right].$$

Let us consider a two-dimensional example, let $A = A_1 \times A_2 = [a_1, b_1] \times [a_2, b_2]$ and suppose that the response Y is:

$$Y = X_1X_2 + c_1X_1 + c_2X_2 + \varepsilon$$

with:

- $\mathbf{X} = (X_1, X_2)$ independent random inputs,
- $c_1 = -\dfrac{\mathbb{E}\left[X_2\mathbb{1}_{\{X_2\in A_2\}}\right]}{\mathbb{P}\left(X_2 \in A_2\right)}$ and $c_2 = -\dfrac{\mathbb{E}\left[X_1\mathbb{1}_{\{X_1\in A_1\}}\right]}{\mathbb{P}\left(X_1 \in A_1\right)}$,
- and ε a centered noise in independent of $\mathbf{X}$.

It can be shown for this model that within the node A, $L^\star \equiv 0$ for all $i \in \{1, 2\}$, for all $z \in [a_i, b_i]$ and yet the regression function m is not constant.

Accordingly, the technical lemma above is well-designed for an additive regression framework. But this context is far from reality for many concrete examples. Outside this framework and as highlighted by our counterexample, an additional assumption is necessary in order to control the approximation error of the estimator. Theorem 1 in Meinshausen (2006) handles the approximation error of its estimator based on a simplified random forest model thanks to a restrictive assumption on the proportion of observations selected in each split and for each direction (see Assumption 3 in Meinshausen (2006)) and a Lipschitz assumption on the conditional distribution function, which is not always true as mentionned in Biau and Scornet (2016). We use Assumption 4.1 instead.

On one other hand, Assumption 4.2 allows us to control the estimation error of our estimators and expresses that cells should contain a sufficiently large number of points so that averaging among the observations is effective.

Finally, Assumption 4.3 is necessary to get uniform convergence of the estimators.

Hence, thanks to all these suitable assumptions we get the consistency of our estimators in Theorems 4.1 and 4.2. As far as we know, Theorem 4.1 is the first consistency result for a method based on the original Breiman's random forest algorithm (i.e. using the bootstrap samples).

The next section is devoted to the proofs of the two Theorems 4.1 and 4.2

5 Proofs of the main theorems

The proofs of Theorems 4.1 and 4.2 are close. We begin with that of Theorem 4.1 and then sketch that of Theorem 4.2 which is a bit simpler.

5.1 Proof of Theorem 4.1

The main ingredient of the proof is to use a second sample $\mathcal{D}_n^\diamond$ in order to deal with the data-dependent aspect. Thus, we first define a dummy estimator based on two samples $\mathcal{D}_n$ and $\mathcal{D}_n^\diamond$ which will be used below. The trees are grown as in Algorithm 1 using $\mathcal{D}_n$, but we consider another sample $\mathcal{D}_n^\diamond$ (independent of $\mathcal{D}_n$ and Θ) which is used to define a dummy estimator:

$$F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n^\diamond,\mathcal{D}_n\right)=\sum_{j=1}^{n} w_{n,j}^\diamond\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}} \tag{5.1}$$

where the weights are:

$$w_{n,j}^\diamond\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)=\frac{1}{k}\sum_{\ell=1}^{k}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)},\ j=1,\ldots,n$$

with $N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)$, the number of elements of $\mathcal{D}_n^\diamond$ that fall into $A_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)$. Throughout this section, we shall use the convention $\frac{0}{0}=0$ in case $N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)=0$ and thus $\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}=0$ for $j=1,\ldots n$.
The weights $w_{n,j}^\diamond\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)$ are nonnegative random variables, as function of $\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n$. To lighten the notation in the sequel, we will simply write $F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)=\sum_{j=1}^{n} w_j^\diamond\left(\mathbf{x}\right)\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}$ instead of (5.1).

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have:

$$\left|F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\leqslant\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|+\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|$$

The convergence of the two right-hand terms is handled separately into the following Proposition 5.1 and Lemma 5.2.

Proposition 5.1.
Consider a random forest which satisfies Assumptions 4.1 and 4.2. Then,

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)\xrightarrow[n\to\infty]{a.s.}F\left(y|\,\mathbf{X}=\mathbf{x}\right)$$

Hence, Proposition 5.1 establishes the consistency for a random forest estimator based on a second sample $\mathcal{D}_n^\diamond$ independent of $\mathcal{D}_n$ and Θ. Wager and Athey (2018) proved that estimators build from honest forests are asymptotically Gaussian. Remark that in Wager and Athey (2018), it is also required to control the proportion of chosen observations at each split and in each direction. In our case, going through a kind of honest trees is just a theoretical tool. We go one step further with the following lemma by showing that the estimators build with honest and non-honest trees are close.

Lemma 5.2.
Consider a random forest which satisfies Assumption 4.2. Then,

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\xrightarrow[n\to\infty]{a.s.}0$$

Hence, according to Proposition 5.1 and Lemma 5.2, we get

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)\xrightarrow[n\to\infty]{a.s.}F\left(y|\,\mathbf{X}=\mathbf{x}\right) \tag{5.2}$$

Now, thanks to Dini's second theorem, let us sketch how to obtain the almost sure uniform convergence relative to y of the estimator.

Note that $\{Y^j \leqslant y\} = \left\{U_j \leqslant F_{Y|\mathbf{X}=\mathbf{x}}(y)\right\}$ under Assumption 4.3 with $U_j = F_{Y|\mathbf{X}=\mathbf{x}}(Y^j), j = 1, \ldots, n$ which are i.i.d random variables. Then, (5.2) is equivalent to:

$$\forall \mathbf{x} \in \mathcal{X}, \forall s \in [0,1], \quad \sum_{j=1}^{n} w_j^b(\mathbf{x}) \mathbb{1}_{\{U_j \leqslant s\}} \xrightarrow[n\to\infty]{a.s.} s$$

As in the proof of Glivenko–Cantelli's Theorem, using that $s \mapsto \sum_{j=1}^{n} w_j^b(\mathbf{x}) \mathbb{1}_{\{U_j(\omega) \leqslant s\}}$ is increasing and Dini's second theorem, we get the uniform convergence almost everywhere, which concludes the proof of the theorem. ∎

We now turn to the proofs of Proposition 5.1 and Lemma 5.2. To that aim, the following lemma, based on Vapnik-Chervonenkis classes (Vapnik and Chervonenkis, 1971) is a key tool.

Lemma 5.3.
Consider $\mathcal{D}_n$ and $\mathcal{D}_n^\diamond$, two independent datasets of independent n samples of $(\mathbf{X}, Y)$. Build a tree using $\mathcal{D}_n$ with bootstrap and bagging procedure driven by Θ. As before, $N^b(A_n(\Theta)) = N_n^b(\mathbf{x}; \Theta, \mathcal{D}_n)$ is the number of bootstrap observations of $\mathcal{D}_n$ that fall into in $A_n(\Theta) = A_n(\mathbf{x}; \Theta, \mathcal{D}_n)$ and $N^\diamond(A_n(\Theta)) = N_n^\diamond(\mathbf{x}; \Theta, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, \mathcal{D}_n)$, the number of observations of $\mathcal{D}_n^\diamond$ that fall into in $A_n(\Theta)$. Then:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|N^b(A_n(\Theta)) - N^\diamond(A_n(\Theta))\right| > \varepsilon\right) \leqslant 24(n+1)^{2d} e^{-\varepsilon^2/288n}$$

Proof of Lemma 5.3.

Let $\varepsilon > 0$ and $\mathbf{x} \in \mathcal{X}$, we have:

$$\begin{aligned}
&\mathbb{P}\left(\left|N^b(A_n(\Theta)) - N^\diamond(A_n(\Theta))\right| > \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\left|\frac{N^b(A_n(\Theta))}{n} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{X}^j \in A_n(\Theta)\}}\right| > \frac{\varepsilon}{3n}\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{X}^j \in A_n(\Theta)\}} - \mathbb{P}_{\mathbf{X}}(\mathbf{X} \in A_n(\Theta))\right| > \frac{\varepsilon}{3n}\right) \\
&\quad + \mathbb{P}\left(\left|\frac{N^\diamond(A_n(\Theta))}{n} - \mathbb{P}_{\mathbf{X}}(\mathbf{X} \in A_n(\Theta))\right| > \frac{\varepsilon}{3n}\right) \\
&\leqslant \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta^1, \mathcal{D}_n\right) \mathbb{1}_{\{\mathbf{X}^j \in A\}} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{X}^j \in A\}}\right| > \frac{\varepsilon}{3n}\right) \\
&\quad + \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{X}^j \in A\}} - \mathbb{P}_{\mathbf{X}}(\mathbf{X} \in A)\right| > \frac{\varepsilon}{3n}\right) + \mathbb{P}\left(\sup_{A \in \mathcal{B}} \left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{X}^{\diamond j} \in A\}} - \mathbb{P}_{\mathbf{X}}(\mathbf{X} \in A)\right| > \frac{\varepsilon}{3n}\right)
\end{aligned}$$

with $\mathcal{B} = \left\{\prod_{i=1}^{d} [a_i, b_i] : a_i, b_i \in \overline{\mathbb{R}}\right\}$. The last two right-hand terms are handled thanks to a direct application of the Theorem of Vapnik and Chervonenkis (1971) over the class $\mathcal{B}$ whose Vapnik-Chervonenkis dimension is $2d$. This class is nothing more than an extension of the class $\mathcal{R}$ of rectangles in $\mathbb{R}^d$. Following the lines of the proof of Theorem 13.8 in Devroye et al. (2013), one sees that the classes $\mathcal{R}$ and $\mathcal{B}$ have the same Vapnik-Chervonenkis dimension.

A special attention should be given to the first right hand-term. The bootstrap component is represented with the random vector $\left(B_j\left(\Theta^1, \mathcal{D}_n\right)\right)_{j=1,\ldots,n}$ referring to the number of times that

the observation $(\mathbf{X}^j, Y^j)$ has been chosen from the original dataset. Conditionally to $\mathcal{D}_n$, this random vector has a multinomial distribution with parameters $\mathcal{M}(n; 1/n, \ldots, 1/n)$. As stated in Arenal-Gutiérrez et al. (1996), the bootstrap component can also be represented thanks to the variables selected with replacement from the set $\mathcal{D}_n = \{(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^n, Y^n)\}$. Let $Z^1, \ldots, Z^n$ be these elements which are distributed as $Z = (Z_1, Z_2)$ that has a discrete uniform distribution over $\mathcal{D}_n$ conditionally to $\mathcal{D}_n$. The first right hand-term is rewritten as:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\varepsilon}{3n}\right)\\
&= \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\varepsilon}{3n}\right)\\
&= \mathbb{E}\left[\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\varepsilon}{3n}\middle|\mathcal{D}_n\right)\right]\\
&= \mathbb{E}\left[\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j\in A\}} - \mathbb{P}\left(Z_1\in A|\,\mathcal{D}_n\right)\right| > \frac{\varepsilon}{3n}\middle|\mathcal{D}_n\right)\right]
\end{aligned}$$

By applying Vapnik-Chervonenkis' Theorem under the conditional distribution given $\mathcal{D}_n$, we get:

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j\in A\}} - \mathbb{P}\left(Z_1\in A|\,\mathcal{D}_n\right)\right| > \frac{\varepsilon}{3n}\middle|\mathcal{D}_n\right) \leqslant 8(n+1)^{2d}e^{-\varepsilon^2/288n}$$

Therefore,

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\varepsilon}{3n}\right) \leqslant 8(n+1)^{2d}e^{-\varepsilon^2/288n}$$

Finally, we get the overall upper bound:

$$\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^\diamond\left(A_n\left(\Theta\right)\right)\right| > \varepsilon\right) \leqslant 24(n+1)^{2d}e^{-\varepsilon^2/288n}$$

■

Lemma 5.3 is the main ingredient of the proof of Proposition 5.1.

Proof of Proposition 5.1.

We aim to prove:

$$\forall \mathbf{x}\in\mathcal{X}, \forall y\in\mathbb{R}, \quad \mathbb{P}\left(\omega\in\Omega:\ F_{k,n}^{\diamond}\left(y|\,\mathbf{X}=\mathbf{x}\right) \underset{n\to\infty}{\longrightarrow} F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right) = 1$$

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have:

$$\left|F_{k,n}^{\diamond}\left(y|\,\mathbf{x}\right) - F\left(y|\,\mathbf{x}\right)\right| \leqslant \left|\sum_{j=1}^{n} w_j^{\diamond}\left(\mathbf{x}\right)\left(\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}} - F\left(y|\,\mathbf{X}^{\diamond j}\right)\right)\right| + \left|\sum_{j=1}^{n} w_j^{\diamond}\left(\mathbf{x}\right)\left(F\left(y|\,\mathbf{X}^{\diamond j}\right) - F\left(y|\,\mathbf{x}\right)\right)\right|$$

Define $W_n = \sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})\left(\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}} - F\left(y|\,\mathbf{X}^{\diamond j}\right)\right) = \sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})\, Z_j^{\diamond}$ with $Z_j^{\diamond} = \mathbb{1}_{\{Y^{\diamond j}\leqslant y\}} - F\left(y|\,\mathbf{X}^{\diamond j}\right)$, n i.i.d random variables and $V_n = \sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})\left(F\left(y|\,\mathbf{X}^{\diamond j}\right) - F\left(y|\,\mathbf{x}\right)\right)$. Remark that $\mathbb{E}\left[Z_j^{\diamond}\middle|\,\mathbf{X}^{\diamond j}\right] = 0$.

1 We first show that $(W_n)_{n\geqslant 1}$ goes to 0 a.s. in the case of Assumption 4.2 item 2. This is achieved by adapting Hoeffding inequality's proof to our random weighted sum context.

For any $\varepsilon > 0$, $t \in \mathbb{R}_+^*$, we have

$$\mathbb{P}\left(W_n > \varepsilon\right) \leq \mathbb{E}\left[e^{tW_n}\right]\cdot e^{-t\varepsilon}.$$

We shall make use of the folklore lemma below.

Lemma.
Let X be a centred random variable, a.s. bounded by 1. Then, for any $t \in \mathbb{R}$, $\mathbb{E}\left[e^{tX}\right] \leqslant e^{\frac{t^2}{2}}$.

Let $t > 0$, we have:

$$\mathbb{E}\left[e^{tW_n}\right] = \mathbb{E}\left[\prod_{j=1}^{n} e^{tw_j^{\diamond}(\mathbf{x})Z_j^{\diamond}}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{n} e^{tw_j^{\diamond}(\mathbf{x})Z_j^{\diamond}}\middle|\,\mathcal{D}_n, \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}\right]\right]$$

conditionally to $\mathcal{D}_n, \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}$, the $w_j^{\diamond}$ are constant and the $Z_j^{\diamond}$ are centred, independent and bounded by 1. Thus, using the folklore lemma,

$$\mathbb{E}\left[e^{tW_n}\right] = \mathbb{E}\left[\prod_{j=1}^{n}\mathbb{E}\left[e^{tw_j^{\diamond}(\mathbf{x})Z_j^{\diamond}}\middle|\,\mathcal{D}_n, \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}\right]\right] \leqslant \mathbb{E}\left[\prod_{j=1}^{n} e^{t^2 w_j^{\diamond}(\mathbf{x})^2/2}\right].$$

Let $K > 0$ be such that for all $\ell = 1, \ldots, k$, $N_n^b\left(A_n(\ell)\right) = N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right) \geqslant K\sqrt{n}\left(\ln\left(n\right)\right)^{\beta}$ a.s. by using Assumption 4.2 item 2. Denote $\Gamma(\ell)$ the event $\left\{N_n^{\diamond}(A_n(\ell)) < \frac{K\sqrt{n}(\ln(n))^{\beta}}{2}\right\}$. Remark that $\Gamma(\ell) \subset \left\{\left|N_n^{\diamond}(A_n(\ell)) - N_n^b\left(A_n(\ell)\right)\right| > \frac{K\sqrt{n}(\ln(n))^{\beta}}{2}\right\}$. Thus, using Lemma 5.3, we have that $\mathbb{P}\left(\Gamma(\ell)\right) \leq 24\left(n+1\right)^{2d}\exp\left[-\frac{K^2(\ln(n))^{2\beta}}{1152}\right]$.

We have

$$\begin{aligned}\sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})^2 &= \sum_{j=1}^{n}\frac{w_j^{\diamond}(\mathbf{x})}{k}\left(\sum_{\ell=1}^{k}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\ell)\}}}{N_n^{\diamond}(A_n(\ell))}\left(\mathbb{1}_{\{\Gamma(\ell)^c\}} + \mathbb{1}_{\{\Gamma(\ell)\}}\right)\right)\\ &\leq \sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})\left(\frac{2}{K\sqrt{n}(\ln n)^{\beta}} + \frac{1}{k}\sum_{\ell=1}^{k}\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\ell)\}}\mathbb{1}_{\{\Gamma(\ell)\}}\right)\end{aligned}$$

So that,

$$\mathbb{E}\left[\prod_{j=1}^{n} e^{t^2 w_j^\diamond(\mathbf{x})^2/2}\right] \leqslant \exp\left[t^2/\left(K\sqrt{n}\left(\ln(n)\right)^\beta\right)\right] \times \mathbb{E}\left[\exp\left(\frac{t^2}{2}\cdot \mathbb{1}_{\left\{\bigcup_{\ell=1}^{k}\Gamma(\ell)\right\}}\right)\right]$$
$$\leqslant \exp\left[t^2/\left(K\sqrt{n}\left(\ln(n)\right)^\beta\right)\right] \times \left(1+e^{t^2/2}\sum_{\ell=1}^{k}\mathbb{P}\left(\Gamma(\ell)\right)\right)$$
$$\leqslant \exp\left[t^2/\left(K\sqrt{n}\left(\ln(n)\right)^\beta\right)\right] \times \left(1+24k\left(n+1\right)^{2d}\exp\left[\frac{t^2}{2}-\frac{K^2\left(\ln(n)\right)^{2\beta}}{1152}\right]\right)$$

Taking $t^2 = \frac{K^2(\ln(n))^{2\beta}}{576}$ leads to

$$\mathbb{P}\left(W_n > \varepsilon\right) \leqslant \left(1+24k\left(n+1\right)^{2d}\right)\exp\left[\frac{K\left(\ln(n)\right)^\beta}{576\sqrt{n}}-\frac{\varepsilon K\left(\ln(n)\right)^\beta}{24}\right]$$

By using Assumption 4.2, item 1., $k = O(n^\alpha)$ so that the right hand side is summable, then we conclude that W_n goes to 0 almost surely.

2 Let us now show that $(W_n)_{n\geqslant 1}$ goes to 0 a.s. in the case where Assumption 4.2 item 3 is satified.

★ Let us first show that $(W_{n^2})_{n\geqslant 1}$ goes to 0 a.s.

$$\mathbb{E}\left[\left(W_n\right)^2\right] = \mathbb{E}\left[\left(\sum_{j=1}^{n} w_j^\diamond\left(\mathbf{x}\right) Z_j^\diamond\right)^2\right]$$
$$= \sum_{j=1}^{n}\sum_{m=1}^{n}\mathbb{E}\left[w_j^\diamond\left(\mathbf{x}\right) w_m^\diamond\left(\mathbf{x}\right) Z_j^\diamond Z_m^\diamond\right]$$
$$= \sum_{j=1}^{n}\mathbb{E}\left[w_j^{\diamond^2}\left(\mathbf{x}\right) Z_j^{\diamond^2}\right] + \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}}\mathbb{E}\left[w_j^\diamond\left(\mathbf{x}\right) w_m^\diamond\left(\mathbf{x}\right) Z_j^\diamond Z_m^\diamond\right]$$
$$\stackrel{\text{def}}{=} I_n + J_n$$

$$I_n = \mathbb{E}\left[\sum_{j=1}^{n} w_j^{\diamond^2}\left(\mathbf{x}\right) Z_j^{\diamond^2}\right]$$
$$\leqslant \mathbb{E}\left[\sum_{j=1}^{n} w_j^{\diamond^2}\left(\mathbf{x}\right)\right]$$
$$\leqslant \mathbb{E}\left[\frac{1}{\min_{\ell=1,\ldots,k} N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)}\right] \text{ (recall that } \sum_{j=1}^{n} w_j^\diamond\left(\mathbf{x}\right)=1\text{)}$$
$$\leqslant \mathbb{E}\left[\frac{1}{\min_{\ell=1,\ldots,k} N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)}\mathbb{1}_{\left\{\exists \ell \setminus \left|N^b\left(A_n\left(\Theta_\ell\right)\right)-N^\diamond\left(A_n\left(\Theta_\ell\right)\right)\right|>\lambda\right\}}\right]$$
$$+\mathbb{E}\left[\frac{1}{\min_{\ell=1,\ldots,k} N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)}\mathbb{1}_{\left\{\forall \ell \setminus \left|N^b\left(A_n\left(\Theta_\ell\right)\right)-N^\diamond\left(A_n\left(\Theta_\ell\right)\right)\right|\leqslant\lambda\right\}}\right]$$

where $\lambda = \dfrac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{4}$.

$$I_n \leqslant k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^\diamond\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) + \mathbb{E}\left[\frac{1}{N^b\left(A_n\left(\Theta\right)\right) - \lambda}\right]$$

We have:

$$\mathbb{E}\left[\frac{1}{N^b\left(A_n\left(\Theta\right)\right) - \lambda}\right] \leqslant \frac{4}{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]} + \mathbb{P}\left(N^b\left(A_n\left(\Theta\right)\right) - \lambda \leqslant \lambda\right)$$

Using Bienaymé-Tchebychev's inequality, we get:

$$\begin{aligned}\mathbb{P}\left(N^b\left(A_n\left(\Theta\right)\right) \leqslant 2\lambda\right) &\leqslant 4\frac{\mathrm{Var}\left(N^b\left(A_n\left(\Theta\right)\right)\right)}{\left(\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]\right)^2} \\ &\leqslant 4\left(\mathrm{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2\end{aligned}$$

Finally, thanks to Lemma 5.3 and Assumption 4.2 items 1. and 3., there exist C, K and M positive constants such that:

$$\begin{aligned}I_n &\leqslant k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^\diamond\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) + \frac{4}{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]} + 4\left(\mathrm{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 \\ &\leqslant 24Cn^\alpha(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{4608}\right] + \frac{4}{K\sqrt{n}\left(\ln\left(n\right)\right)^\beta} + \frac{4M^2}{n\left(\ln\left(n\right)\right)^\gamma}\end{aligned}$$

The trick of using a second sample $\mathcal{D}_n^\diamond$ independent of the first-one and the random variable Θ is really important to handle the J_n term. Indeed, we have $J_n = 0$ while the equivalent term encountered in the proof of the Theorem 2 developed by Scornet et al. (2015b) is handled using a conjecture regarding the correlation behavior of the CART algorithm that is difficult to verify (cf. assumption (H2) of Scornet et al. (2015b)). Indeed:

$$\begin{aligned}J_n &= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[w_j^\diamond\left(\mathbf{x}\right) w_m^\diamond\left(\mathbf{x}\right) Z_j^\diamond Z_m^\diamond\right] \\ &= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\mathbb{E}\left[w_j^\diamond\left(\mathbf{x}\right) w_m^\diamond\left(\mathbf{x}\right) Z_j^\diamond Z_m^\diamond \middle| \Theta_1,\ldots,\Theta_k,\mathcal{D}_n,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},Y^{\diamond j}\right]\right] \\ &= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[w_j^\diamond\left(\mathbf{x}\right) w_m^\diamond\left(\mathbf{x}\right) Z_j^\diamond \mathbb{E}\left[Z_m^\diamond \middle| \mathbf{X}^{\diamond m}\right]\right] \text{ using that } w_j^\diamond\left(\mathbf{x}\right), w_m^\diamond\left(\mathbf{x}\right) \text{ and} \\ &Z_j^\diamond \text{ are } \Theta_1,\ldots,\Theta_k,\mathcal{D}_n,\left(\mathbf{X}^{\diamond j}\right)_{j=1,\ldots,n},Y^{\diamond j} \text{ measurable.} \\ &= 0 \text{ because } \mathbb{E}\left[Z_m^\diamond \middle| \mathbf{X}^{\diamond m}\right] = 0\end{aligned}$$

Finally:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(|W_n| \geqslant \varepsilon\right) \leqslant \frac{\mathbb{E}\left[\left(W_n\right)^2\right]}{\varepsilon^2} = \frac{I_n}{\varepsilon^2}$$

Hence, since $\sum_{n\geq 1} I_{n^2} < \infty$, Borel–Cantelli Lemma gives:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\limsup_{n\to\infty}\left\{|W_{n^2}| \geqslant \varepsilon\right\}\right) = 0$$

which implies that $W_{n^2} \xrightarrow[n\to\infty]{a.s.} 0$.

★ Let us now show that $(W_n)_{n\geqslant 1}$ converges almost surely to 0.

Let $p = p(n) = \lfloor \sqrt{n} \rfloor$, we have $W_n - W_{p^2} = \sum_{j=p^2+1}^{n} w_j^{\diamond}(\mathbf{x}) Z_j^{\diamond}$. Fix $\varepsilon > 0$ and consider again $\lambda = \dfrac{\mathbb{E}\left[N^b(A_n(\Theta))\right]}{4}$,

$$
\begin{aligned}
&\mathbb{P}\left(\left|W_n - W_{p^2}\right| \geqslant \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\sum_{j=p^2+1}^{n} w_j^{\diamond}(\mathbf{x})\left|Z_j^{\diamond}\right| \geqslant \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\frac{2\sqrt{n}}{\min\limits_{\ell=1,\dots,k} N_n^{\diamond}\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)} \geqslant \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\frac{2\sqrt{n}}{\min\limits_{\ell=1,\dots,k} N_n^{\diamond}\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)} \geqslant \varepsilon, \exists \ell \setminus \left|N^b(A_n(\Theta_\ell)) - N^{\diamond}(A_n(\Theta_\ell))\right| > \lambda\right) \\
&\quad + \mathbb{P}\left(\frac{2\sqrt{n}}{\min\limits_{\ell=1,\dots,k} N_n^{\diamond}\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)} \geqslant \varepsilon, \forall \ell \setminus \left|N^b(A_n(\Theta_\ell)) - N^{\diamond}(A_n(\Theta_\ell))\right| \leqslant \lambda\right) \\
&\leqslant k\mathbb{P}\left(\left|N^b(A_n(\Theta)) - N^{\diamond}(A_n(\Theta))\right| > \lambda\right) + \mathbb{P}\left(\frac{2\sqrt{n}}{N^b(A_n(\Theta)) - \lambda} \geqslant \varepsilon\right)
\end{aligned}
$$

Using Bienaymé-Tchebychev's inequality and that $\mathbb{E}\left[N^b(A_n(\Theta))\right] \geqslant \dfrac{8\sqrt{n}}{\varepsilon}$ for n large enough, thanks to Assumption 4.2 item 3., we have:

$$
\begin{aligned}
\mathbb{P}\left(N^b(A_n(\Theta)) \leqslant \lambda + \frac{2\sqrt{n}}{\varepsilon}\right) &\leqslant 4\frac{\mathrm{Var}\left(N^b(A_n(\Theta))\right)}{\left(\mathbb{E}\left[N^b(A_n(\Theta))\right]\right)^2} \\
&\leqslant 4\left(\mathrm{CV}\left(N^b(A_n(\Theta))\right)\right)^2
\end{aligned}
$$

Finally, thanks to Lemma 5.3 and Assumption 4.2 items 1. and 3., we have for n large enough:

$$
\begin{aligned}
\mathbb{P}\left(\left|W_n - W_{p^2}\right| \geqslant \varepsilon\right) &\leqslant k\mathbb{P}\left(\left|N^b(A_n(\Theta)) - N^{\diamond}(A_n(\Theta))\right| > \lambda\right) + 4\left(\mathrm{CV}\left(N^b(A_n(\Theta))\right)\right)^2 \\
&\leqslant 24Cn^{\alpha}(n+1)^{2d}\exp\left[-\frac{K^2(\ln(n))^{2\beta}}{4608}\right] + \frac{4M^2}{n(\ln(n))^{\gamma}}
\end{aligned}
$$

Hence, using Borel–Cantelli Lemma:

$$
\forall \varepsilon > 0, \quad \mathbb{P}\left(\limsup_{n\to\infty}\left\{\left|W_n - W_{p^2}\right| \geqslant \varepsilon\right\}\right) = 0
$$

which implies that the random variable $W_n - W_{p^2} \xrightarrow[n\to\infty]{a.s.} 0$.

From this, we deduce that $(W_n)_{n\geqslant 1}$ goes to 0 a.s.

3 Finally, we show that $(V_n)_{n\geqslant 1}$ goes to 0 a.s.

$$
\begin{aligned}
|V_n| &= \left|\sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})\left(F\left(y|\,\mathbf{X}^{\diamond j}\right) - F(y|\,\mathbf{x})\right)\right| \\
&\leqslant \frac{1}{k}\sum_{\ell=1}^{k}\left(\sum_{j=1}^{n}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^{\diamond}(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n)}\left|F\left(y|\,\mathbf{X}^{\diamond j}\right) - F(y|\,\mathbf{x})\right|\right) \\
&\leqslant \frac{1}{k}\sum_{\ell=1}^{k}\left(\sum_{j=1}^{n}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^{\diamond}(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n)}\sup_{\mathbf{z}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}|F(y|\,\mathbf{z}) - F(y|\,\mathbf{x})|\right) \\
&\leqslant \frac{1}{k}\sum_{\ell=1}^{k}\sup_{\mathbf{z}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}|F(y|\,\mathbf{z}) - F(y|\,\mathbf{x})|
\end{aligned}
$$

Hence, by using Assumption 4.1, we have:

$$
\left|\sum_{j=1}^{n} w_j^{\diamond}(\mathbf{x})\left(F\left(y|\,\mathbf{X}^{\diamond j}\right) - F(y|\,\mathbf{x})\right)\right| \xrightarrow[n\to\infty]{a.s.} 0
$$

This allows us to conclude that

$$
\forall \mathbf{x}\in\mathcal{X}, \forall y\in\mathbb{R}, \quad F_{k,n}^{\diamond}(y|\,\mathbf{X}=\mathbf{x}) \xrightarrow[n\to\infty]{a.s.} F(y|\,\mathbf{X}=\mathbf{x})
$$

■

Let us now turn to the proof of Lemma 5.2 which shows that the dummy estimator $F_{k,n}^{\diamond}$ is close to the interesting one $F_{k,n}^{b}$.

Proof of Lemma 5.2.

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have that:

$$
\begin{aligned}
&\left|F_{k,n}^{\diamond}(y|\,\mathbf{X}=\mathbf{x}) - F_{k,n}^{b}(y|\,\mathbf{X}=\mathbf{x})\right| \\
&= \left|\frac{1}{k}\sum_{\ell=1}^{k}\left(\sum_{j=1}^{n}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}}{N_n^{\diamond}(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n)} - \sum_{j=1}^{n}\frac{B_j(\Theta_\ell^1,\mathcal{D}_n)\,\mathbb{1}_{\{\mathbf{X}^{j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}\mathbb{1}_{\{Y^{j}\leqslant y\}}}{N_n^{b}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right)\right| \\
&= \left|\frac{1}{k}\sum_{\ell=1}^{k}\left(\frac{\#\left\{j\leqslant J^{\diamond}\ /\ \mathbf{X}^{\diamond(j)}\in A_n(\Theta_\ell)\right\}}{N^{\diamond}(A_n(\Theta_\ell))} - \frac{\sum_{j\in\mathcal{S}}B_j(\Theta_\ell^1,\mathcal{D}_n)}{N^{b}(A_n(\Theta_\ell))}\right)\right| \text{ with } \mathcal{S} = \left\{j\leqslant J\ /\ \mathbf{X}^{(j)}\in A_n(\Theta_\ell)\right\}
\end{aligned}
$$

where we denote $A_n(\Theta_\ell) = A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$, $N^{\diamond}(A_n(\Theta_\ell)) = N_n^{\diamond}(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n)$ and $N^{b}(A_n(\Theta_\ell)) = N_n^{b}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$. $J, J^{\diamond}$ are such that $Y^{\diamond(J^{\diamond})}\leqslant y < Y^{\diamond(J^{\diamond}+1)}$ and $Y^{(J)}\leqslant y < Y^{(J+1)}$, with $Y^{\diamond(j)}$ (resp. $Y^{(j)}$) the order statistics of $(Y^{\diamond 1},\ldots,Y^{\diamond n})$ (resp. $(Y^1,\ldots,Y^n)$) and the $\mathbf{X}^{\diamond(j)}$ (resp. $\mathbf{X}^{(j)}$) the corresponding $\mathbf{X}^{\diamond p}$'s (resp. $\mathbf{X}^{p}$'s).

Let us consider for some $\ell\in[\![1,k]\!]$,

$$
G = \frac{\#\left\{j\leqslant J^{\diamond}\ /\ \mathbf{X}^{\diamond(j)}\in A_n(\Theta_\ell)\right\}}{N^{\diamond}(A_n(\Theta_\ell))} - \frac{\sum_{j\in\mathcal{S}}B_j(\Theta_\ell^1,\mathcal{D}_n)}{N^{b}(A_n(\Theta_\ell))} \overset{\text{def}}{=} \frac{N_{J^{\diamond}}^{\diamond}(A_n(\Theta_\ell))}{N^{\diamond}(A_n(\Theta_\ell))} - \frac{N_J(A_n(\Theta_\ell))}{N^{b}(A_n(\Theta_\ell))}.
$$

We have,

$$|G| \leqslant \frac{\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^b\left(A_n\left(\Theta_\ell\right)\right)} + \frac{\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^b\left(A_n\left(\Theta_\ell\right)\right)}$$
$$\stackrel{\text{def}}{=} |G_1| + |G_2|$$

We continue the proof below in the case where Assumption 4.2 item 3. is satisfied. The case where item 2. is verified is done easier following the same lines. Let $\varepsilon > 0$.

1 We are now going to show the almost everywhere convergence to 0 for each term G_1 and G_2. Let us start with G_1.

$$\begin{aligned}
\mathbb{P}\left(|G_1| > \varepsilon\right) &= \mathbb{P}\left(\frac{\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^b\left(A_n\left(\Theta_\ell\right)\right)} > \varepsilon\right) \\
&= \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^b\left(A_n\left(\Theta_\ell\right)\right), N^b\left(A_n\left(\Theta_\ell\right)\right) > \lambda\right) \\
&\quad + \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^b\left(A_n\left(\Theta_\ell\right)\right), N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)
\end{aligned}$$

where $\lambda = \dfrac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{2}$

$$\leqslant \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + \mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)$$

Again, thanks to the Bienaymé-Tchebychev's inequality:

$$\begin{aligned}
\mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right) &\leqslant 4\frac{\operatorname{Var}\left(N^b\left(A_n\left(\Theta_\ell\right)\right)\right)}{\left(\mathbb{E}\left[N^b\left(A_n\left(\Theta_\ell\right)\right)\right]\right)^2} \\
&\leqslant 4\left(\operatorname{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 \qquad (5.3)
\end{aligned}$$

Now, using Lemma 5.3 and Assumption 4.2, we get:

$$\begin{aligned}
\mathbb{P}\left(|G_1| > \varepsilon\right) &\leqslant \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + 4\left(\operatorname{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 \\
&\leqslant 24(n+1)^{2d}\exp\left[-\frac{\varepsilon^2 K^2\left(\ln\left(n\right)\right)^{2\beta}}{1152}\right] + \frac{4M^2}{n\left(\ln\left(n\right)\right)^{\gamma}}
\end{aligned}$$

Then, thanks to Borel–Cantelli Lemma:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\limsup_{n\to\infty}\left\{|G_1| > \varepsilon\right\}\right) = 0$$

which implies $G_1 \xrightarrow[n\to\infty]{a.s.} 0$.

2 Now, consider the G_2 term:

$$\begin{aligned}
\mathbb{P}\left(|G_2| > \varepsilon\right) &= \mathbb{P}\left(\frac{\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^b\left(A_n\left(\Theta_\ell\right)\right)} > \varepsilon\right) \\
&= \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N\left(A_n\left(\Theta_\ell\right)\right), N^b\left(A_n\left(\Theta_\ell\right)\right) > \lambda\right) \\
&\quad + \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^b\left(A_n\left(\Theta_\ell\right)\right), N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)
\end{aligned}$$

where $\lambda = \dfrac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{2}$

$$\leqslant \mathbb{P}\left(\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + \mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)$$

We are going to bound the first term by using again the Vapnik-Chervonenkis theory. By considering the class $\mathcal{B} = \left\{\prod_{i=1}^{d}\left[a_i, b_i\right] \times \left]-\infty, y\right] : a_i, b_i \in \overline{\mathbb{R}}\right\}$, we have:

$$\begin{aligned}
&\mathbb{P}\left(\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right)\\
&= \mathbb{P}\left(\frac{\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{n} > \frac{\varepsilon\lambda}{n}\right)\\
&\leqslant \mathbb{P}\left(\left|\frac{N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)}{n} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A_n\left(\Theta_\ell\right) \times \left]-\infty, y\right]\right)\right| > \frac{\varepsilon\lambda}{3n}\right)\\
&\quad + \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{(\mathbf{X}^j,Y^j)\in A_n(\Theta_\ell)\times\left]-\infty,y\right]\}} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A_n\left(\Theta_\ell\right) \times \left]-\infty, y\right]\right)\right| > \frac{\varepsilon\lambda}{3n}\right)\\
&\quad + \mathbb{P}\left(\left|\frac{N_J\left(A_n\left(\Theta_\ell\right)\right)}{n} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{(\mathbf{X}^j,Y^j)\in A_n(\Theta_\ell)\times\left]-\infty,y\right]\}}\right| > \frac{\varepsilon\lambda}{3n}\right)\\
&\leqslant \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{(\mathbf{X}^{\diamond j},Y^{\diamond j})\in A\}} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A\right)\right| > \frac{\varepsilon\lambda}{3n}\right) \qquad (5.4)\\
&\quad + \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{(\mathbf{X}^{j},Y^{j})\in A\}} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A\right)\right| > \frac{\varepsilon\lambda}{3n}\right)\\
&\quad + \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}B_j\left(\Theta_\ell^1, \mathcal{D}_n\right)\mathbb{1}_{\{(\mathbf{X}^{j},Y^{j})\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{(\mathbf{X}^{j},Y^{j})\in A\}}\right| > \frac{\varepsilon\lambda}{3n}\right)
\end{aligned}$$

Let us make some comments on Vapnik-Chervonenkis dimension of the class $\mathcal{B}$. If we had the class $\left\{\prod_{i=1}^{d}\left[a_i, b_i\right] \times \left]-\infty, c\right] : a_i, b_i \in \overline{\mathbb{R}}, c \in \mathbb{R}\right\}$, it could be shown by calculations similar to those from Theorem 13.8 in Devroye et al. (2013) that the Vapnik-Chervonenkis dimension is $2d+1$. But in our case, the element c is fixed as y, then all the possibilities to break the points are related to the elements a_i, b_i, which thus gives us a Vapnik-Chervonenkis dimension equals to $2d$. Therefore, the first two right-hand terms are handled thanks to a direct application of the Theorem of Vapnik and Chervonenkis (1971) over the class $\mathcal{B}$. The latter deserves special attention.

The third term is treated as in the proof of Lemma 5.3. We use as before, the representation of the bootstrap component with the random variables $Z_\ell^1, \ldots, Z_\ell^n$. We apply Vapnik-Chervonenkis' Theorem under the conditional distribution given $\mathcal{D}_n$ and get:

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_\ell^j\in A\}} - \mathbb{P}\left(Z \in A \middle| \mathcal{D}_n\right)\right| > \frac{\varepsilon\lambda}{3n}\middle| \mathcal{D}_n\right) \leqslant 8(n+1)^{2d}e^{-\varepsilon^2\lambda^2/288n}$$

Therefore,

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_\ell^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{(\mathbf{X}^{j},Y^{j})\in A\}}\right| > \frac{\varepsilon\lambda}{3n}\right) \leqslant 8(n+1)^{2d}e^{-\varepsilon^2\lambda^2/288n}$$

Hence, at last

$$\mathbb{P}\left(\left|N_{J^\diamond}^\diamond\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\lambda\right)\leqslant 24(n+1)^{2d}e^{-\varepsilon^2\lambda^2/288n}$$

As a consequence, Equation (5.3) and Assumption 4.2 lead to:

$$\begin{aligned}\mathbb{P}\left(|G_2|>\varepsilon\right)&\leqslant\mathbb{P}\left(\left|N_{J^\diamond}^\diamond\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\lambda\right)+4\left(\mathrm{CV}\left(N\left(A_n\left(\Theta\right)\right)\right)\right)^2\\&\leqslant 24(n+1)^{2d}\exp\left[-\frac{\varepsilon^2K^2\left(\ln\left(n\right)\right)^{2\beta}}{1152}\right]+\frac{4M^2}{n\left(\ln\left(n\right)\right)^\gamma}\end{aligned}$$

Thanks to Borel–Cantelli Lemma, we get:

$$\forall\varepsilon>0,\quad\mathbb{P}\left(\limsup_{n\to\infty}\left\{|G_2|>\varepsilon\right\}\right)=0$$

which implies that $G_2\xrightarrow[n\to\infty]{a.s.}0$.

We conclude that G goes to 0 for all ℓ, thus

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\xrightarrow[n\to\infty]{a.s.}0$$

In the case where Assumption 4.2 item 2. is verified, it exists $K>0$ such that

$$N^b\left(A_n\left(\Theta_\ell\right)\right)\geqslant K\sqrt{n}\left(\ln\left(n\right)\right)^\beta\ \ a.s.$$

So that $\mathbb{P}(|G_1|>\varepsilon)$ and $\mathbb{P}(|G_2|>\varepsilon)$ are bounded above respectively by

- $\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta_\ell\right)\right)-N^\diamond\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon K\sqrt{n}\left(\ln\left(n\right)\right)^\beta\right)$
- and $\mathbb{P}\left(\left|N_{J^\diamond}^\diamond\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon K\sqrt{n}\left(\ln\left(n\right)\right)^\beta\right)$.

A simple application of Lemma 5.3 and an adaptation of it to $N_J\left(A_n\left(\Theta_\ell\right)\right)$ show that G_1 and G_2 go to 0 a.s. ■

This conclude the proof of Theorem 4.1, we now sketch the proof of Theorem 4.2 which is a bit simpler.

5.2 Proof of Theorem 4.2

The different steps are similar to those of the proof of Theorem 4.1 and the dummy estimator $F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)$ introduced in the proof of Theorem 4.1 will be reused.

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have:

$$\left|F_{k,n}^o\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\leqslant\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|+\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F_{k,n}^o\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|$$

As in the proof of Theorem 4.1, the first right-hand term is handled thanks to Proposition 5.1 which gives:

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)\xrightarrow[n\to\infty]{a.s.}F\left(y|\,\mathbf{X}=\mathbf{x}\right)$$

The convergence of the second right-hand term is handled into the following Lemma 5.4.

Lemma 5.4.
Consider a random forest which satisfies Assumption 4.2. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \left| F^{\diamond}_{k,n}\left(y | \mathbf{X}=\mathbf{x}\right) - F^{o}_{k,n}\left(y | \mathbf{X}=\mathbf{x}\right) \right| \xrightarrow[n\to\infty]{a.s.} 0$$

This allows us to conclude that

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad F^{o}_{k,n}\left(y | \mathbf{X}=\mathbf{x}\right) \xrightarrow[n\to\infty]{a.s.} F\left(y | \mathbf{X}=\mathbf{x}\right)$$

As in the proof of Theorem 4.1, the almost sure uniform convergence relative to y of the estimator is achieved using Dini's second theorem, which concludes the proof.

■

Proof of Lemma 5.4.

The proof is done by following the same steps as the proof of Lemma 5.2 and with the same notations. Let $\mathbf{x} \in \mathcal{X}$ and $y \in \mathbb{R}$, we have

$$\begin{aligned}
&\left| F^{\diamond}_{k,n}\left(y | \mathbf{X}=\mathbf{x}\right) - F^{o}_{k,n}\left(y | \mathbf{X}=\mathbf{x}\right) \right| \\
&= \left| \frac{1}{k} \sum_{\ell=1}^{k} \left(\frac{\#\left\{ j \leqslant J^{\diamond} \ / \ \mathbf{X}^{\diamond(j)} \in A_n\left(\Theta_\ell\right) \right\}}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{\#\left\{ j \leqslant J \ / \ \mathbf{X}^{(j)} \in A_n\left(\Theta_\ell\right) \right\}}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} \right) \right|
\end{aligned}$$

For any $\ell \in [\![1, k]\!]$,

$$G = \frac{\#\left\{ j \leqslant J^{\diamond} \ / \ \mathbf{X}^{\diamond(j)} \in A_n\left(\Theta_\ell\right) \right\}}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{\#\left\{ j \leqslant J \ / \ \mathbf{X}^{(j)} \in A_n\left(\Theta_\ell\right) \right\}}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} \overset{\text{def}}{=} \frac{N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{N_{J}\left(A_n\left(\Theta_\ell\right)\right)}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)}.$$

We have,

$$\begin{aligned}
|G| &\leqslant \frac{\left| N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right) \right|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} + \frac{\left| N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_{J}\left(A_n\left(\Theta_\ell\right)\right) \right|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} \\
&\overset{\text{def}}{=} |G_1| + |G_2|
\end{aligned}$$

We prove that G_1 and G_2 go to 0 a.s. in the case where Assumption 4.2 item 3. is verified. The case where item 2. is satisfied is done easier following the same lines as in the proof of Lemma 5.2. Let $\varepsilon > 0$.

1 Let us start by proving the a.s. convergence to 0 of G_1.

$$\begin{aligned}
&\mathbb{P}\left(|G_1| > \varepsilon\right) \\
&= \mathbb{P}\left(\frac{\left| N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right) \right|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} > \varepsilon \right) \\
&= \mathbb{P}\left(\left| N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right) \right| > \varepsilon N^{o}\left(A_n\left(\Theta_\ell\right)\right), \exists \ell \setminus \left| N^{b}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right) \right| > \lambda \right) \\
&\quad + \mathbb{P}\left(\left| N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right) \right| > \varepsilon N^{o}\left(A_n\left(\Theta_\ell\right)\right), \forall \ell \setminus \left| N^{b}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right) \right| \leqslant \lambda \right)
\end{aligned}$$

$$\text{with } \lambda = \frac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{4}$$

$$\begin{aligned}
&\leqslant k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^o\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \\
&\quad + \mathbb{P}\left(\left|N^\diamond\left(A_n\left(\Theta_\ell\right)\right) - N^o\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\left(N^b\left(A_n\left(\Theta_\ell\right)\right) - \lambda\right), N^b\left(A_n\left(\Theta_\ell\right)\right) > \delta\right) \\
&\quad + \mathbb{P}\left(\left|N^\diamond\left(A_n\left(\Theta_\ell\right)\right) - N^o\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\left(N^b\left(A_n\left(\Theta_\ell\right)\right) - \lambda\right), N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \delta\right)
\end{aligned}$$

$$\text{with } \delta = \frac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{2}$$

$$\begin{aligned}
&\leqslant k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^o\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) + \mathbb{P}\left(\left|N^\diamond\left(A_n\left(\Theta_\ell\right)\right) - N^o\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) \\
&\quad + \mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \delta\right)
\end{aligned}$$

The first two right-hand terms will be bounded by using the Vapnik-Chervonenkis' theory with the class $\mathcal{B} = \left\{\prod_{i=1}^{d}\left[a_i, b_i\right] : a_i, b_i \in \overline{\mathbb{R}}\right\}$. Let us start with the first right hand-term as follows:

$$\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^o\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \leqslant \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\lambda}{n}\right)$$

This term is handled as in the proof of Lemma 5.3 by rewriting the bootstrap component thanks to the variables selected with replacement from the set $\mathcal{D}_n = \left\{\left(\mathbf{X}^1, Y^1\right), \ldots, \left(\mathbf{X}^n, Y^n\right)\right\}$ instead of the random vector $\left(B_j\left(\Theta^1, \mathcal{D}_n\right)\right)_{j=1,\ldots,n}$.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\lambda}{n}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}} - \mathbb{P}\left(Z_1 \in A\middle|\, \mathcal{D}_n\right)\right| > \frac{\lambda}{n}\middle|\, \mathcal{D}_n\right)\right]
\end{aligned}$$

By applying Vapnik-Chervonenkis' Theorem under the conditional distribution given $\mathcal{D}_n$, we get:

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}} - \mathbb{P}\left(Z_1 \in A\middle|\, \mathcal{D}_n\right)\right| > \frac{\lambda}{n}\middle|\, \mathcal{D}_n\right) \leqslant 8(n+1)^{2d}e^{-\lambda^2/32n}$$

Therefore,

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\lambda}{n}\right) \leqslant 8(n+1)^{2d}e^{-\lambda^2/32n}$$

Finally, we get the overall upper bound:

$$\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^o\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \leqslant 8(n+1)^{2d}e^{-\lambda^2/32n}$$

Regarding the second right hand-term, we have:

$$\begin{aligned}\mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)-N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\lambda\right) &= \mathbb{P}\left(\frac{\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)-N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right|}{n}>\frac{\varepsilon\lambda}{n}\right)\\ &\leqslant \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A\}}-\mathbb{P}_{\mathbf{X}}\left(\mathbf{X}\in A\right)\right|>\frac{\varepsilon\lambda}{2n}\right)\\ &\quad+\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^{j}\in A\}}-\mathbb{P}_{\mathbf{X}}\left(\mathbf{X}\in A\right)\right|>\frac{\varepsilon\lambda}{2n}\right)\\ &\leqslant 16(n+1)^{2d}e^{-\varepsilon^2\lambda^2/128n}\end{aligned}$$

Finally, using Equation (5.3) and Assumption 4.2, we have

$$\mathbb{P}\left(|G_1|>\varepsilon\right)\leqslant 8Cn^{\alpha}(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{512}\right]+16(n+1)^{2d}\exp\left[-\frac{\varepsilon^2K^2\left(\ln\left(n\right)\right)^{2\beta}}{2048}\right]+\frac{4M^2}{n\left(\ln\left(n\right)\right)^{\gamma}}$$

Then, thanks to Borel–Cantelli Lemma: $G_1 \xrightarrow[n\to\infty]{a.s.} 0$.

2 Now, consider the G_2 term:

$$\begin{aligned}&\mathbb{P}\left(|G_2|>\varepsilon\right)\\ &=\mathbb{P}\left(\frac{\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)}>\varepsilon\right)\\ &=\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon N^{o}\left(A_n\left(\Theta_\ell\right)\right),\exists\ell\setminus\left|N^{b}\left(A_n\left(\Theta_\ell\right)\right)-N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right|>\lambda\right)\\ &\quad+\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon N^{o}\left(A_n\left(\Theta_\ell\right)\right),\forall\ell\setminus\left|N^{b}\left(A_n\left(\Theta_\ell\right)\right)-N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right|\leqslant\lambda\right)\end{aligned}$$

where $\lambda=\dfrac{\mathbb{E}\left[N^{b}\left(A_n\left(\Theta\right)\right)\right]}{4}$

$$\begin{aligned}&\leqslant k\mathbb{P}\left(\left|N^{b}\left(A_n\left(\Theta\right)\right)-N^{o}\left(A_n\left(\Theta\right)\right)\right|>\lambda\right)\\ &\quad+\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\left(N^{b}\left(A_n\left(\Theta_\ell\right)\right)-\lambda\right),N^{b}\left(A_n\left(\Theta_\ell\right)\right)>\delta\right)\\ &\quad+\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\left(N^{b}\left(A_n\left(\Theta_\ell\right)\right)-\lambda\right),N^{b}\left(A_n\left(\Theta_\ell\right)\right)\leqslant\delta\right)\end{aligned}$$

where $\delta=\dfrac{\mathbb{E}\left[N^{b}\left(A_n\left(\Theta\right)\right)\right]}{2}$

$$\begin{aligned}&\leqslant k\mathbb{P}\left(\left|N^{b}\left(A_n\left(\Theta\right)\right)-N^{o}\left(A_n\left(\Theta\right)\right)\right|>\lambda\right)+\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\lambda\right)\\ &\quad+\mathbb{P}\left(N^{b}\left(A_n\left(\Theta_\ell\right)\right)\leqslant\delta\right)\end{aligned}$$

The middle term is treated as in Equation (5.4), we get:

$$\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\lambda\right)\leqslant 16(n+1)^{2d}e^{-\varepsilon^2\lambda^2/128n}$$

As a consequence, Equation (5.3) and Assumption 4.2 give:

$$\mathbb{P}\left(|G_2|>\varepsilon\right)\leqslant 8Cn^{\alpha}(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{512}\right]+16(n+1)^{2d}\exp\left[-\frac{\varepsilon^2K^2\left(\ln\left(n\right)\right)^{2\beta}}{2048}\right]+\frac{4M^2}{n\left(\ln\left(n\right)\right)^{\gamma}}$$

Thanks to Borel–Cantelli Lemma, we get $G_2 \xrightarrow[n\to\infty]{a.s.} 0$.

We conclude that G goes to 0 a.s. for all ℓ, thus

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \left| F^{\diamond}_{k,n}\left(y|\,\mathbf{X}=\mathbf{x}\right) - F^{o}_{k,n}\left(y|\,\mathbf{X}=\mathbf{x}\right)\right| \xrightarrow[n\to\infty]{a.s.} 0$$

■

In order to illustrate the theoretical results, we provide a numerical example.

6 Numerical example

The convergence of the estimators, introduced in Section 3, is illustrated on the following toy example:

$$Y = X_1 + X_2 + X_3 + \varepsilon \tag{6.1}$$

where $\mathbf{X} = (X_1, X_2, X_3)$ are three independent random variables with $X_1 \sim GPD(1.5, 0.25)$ (a Generalised Pareto Distribution), $X_2 \sim \mathcal{LN}(1.1, 0.6)$ (a Log Normal Distribution), $X_3 \sim \Gamma(2, 0.6)$ (a Gamma Distribution) and ε is an independent centered Gaussian noise with variance $\sigma^2 = 4$.

The accuracy of the conditional distribution function estimators will be evaluated first, then that of the conditional quantile estimators.

6.1 Conditional distribution function

Let us start by assessing the performance of the C_CDF estimators using the Kolmogorov-Smirnov distance recalled below:

$$KS(\mathbf{x}) = \max_{y} \left| \widehat{F}\left(y|\,\mathbf{x}\right) - F\left(y|\,\mathbf{x}\right)\right|$$

with $\widehat{F}\left(y|\,\mathbf{x}\right)$ being either $F^{b}_{k,n}\left(y|\,\mathbf{x}\right)$ or $F^{o}_{k,n}\left(y|\,\mathbf{x}\right)$ here. $F\left(y|\,\mathbf{x}\right)$ has the following expression in our example:

$$F\left(y|\,\mathbf{X}=\mathbf{x}\right) = \Phi\left(\frac{y-(x_1+x_2+x_3)}{\sigma}\right)$$

with Φ, the distribution function of the standard normal distribution $\mathcal{N}(0,1)$.

For two randomly chosen points $\mathbf{x}$, the C_CDF estimates are built with a sample of size $n = 10^4$, a forest grown with $n_{trees} = 500$ and the minimum number of samples required to be at a leaf node set to $min\_samples\_leaf = \lfloor \sqrt{n} \cdot \log(n)^{1.5}/250 \rfloor$ for each tree. These experiments are replicated $s = 500$ times. Figure 1 below shows the C_CDF approximations computed with the estimator using the original dataset on the left-hand side and those of the estimator calculated with the bootstrap samples on the right-hand side. On each graph, the orange plain line is the true C_CDF, while the blue ones represent the 95% quantiles of the replications. Figure 1 therefore allows to display and compare approximated curves visually.

From a quantitative perspective, the quality of the estimators is measured at each point $\mathbf{x}$ using the following average Kolmogorov-Smirnov distance:

$$\overline{KS}(\mathbf{x}) = \frac{1}{s}\sum_{j=1}^{s} KS_j(\mathbf{x})$$

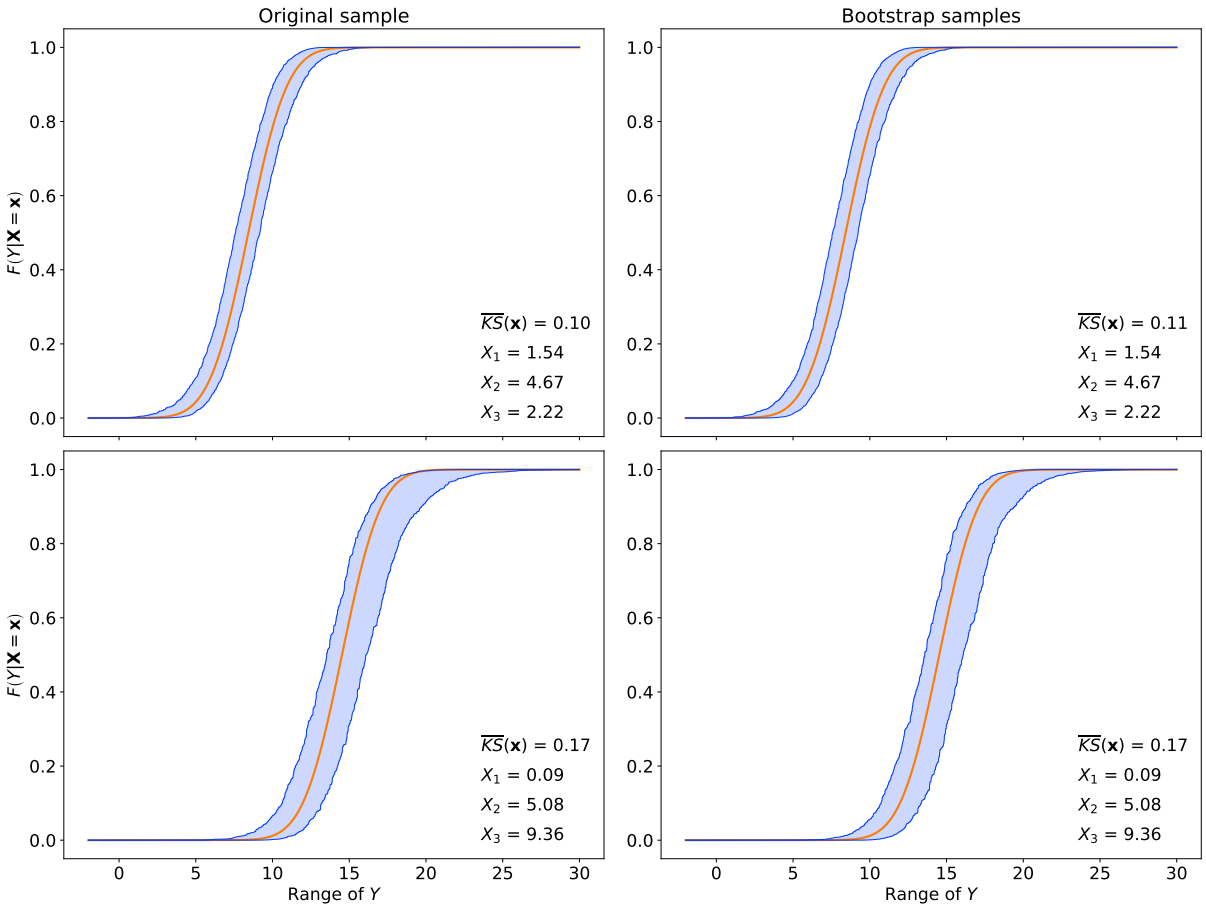

Figure 1: *Estimation of the conditional distribution function for two different values* $\mathbf{x}$ *by using the original sample (on the left side) and the bootstrap samples (on the right side). On each graph, the orange line is the true value along with the* 95% *confidence bands in blue.*

According to the numerical results displayed in Figure 1, estimators perform well for points that are well represented in the training sample but, the performance decreases for extreme points. In order to reflect the overall performance of the estimators, we define in the sequel an averaged version of the previous metric and compute it with $p = 5 \times 10^4$ randomly chosen points $\mathbf{x}$:

$$M\_\overline{KS} := \frac{1}{p} \sum_{j=1}^{p} \overline{KS}(\mathbf{x}^j)$$

We get $M\_\overline{KS} = 0.1344$ for the estimator $F^b_{k,n}(y|\,\mathbf{x})$ and $M\_\overline{KS} = 0.1295$ for $F^o_{k,n}(y|\,\mathbf{x})$. Thus, it seems that both estimators have a good accuracy for estimating the C_CDF of most points $\mathbf{x}$.

Let us now assess the performance of the conditional quantile estimators.

6.2 Conditional quantiles

The analytic value of the α-quantile conditionally to $\mathbf{x} = (x_1, x_2, x_3)$ is easy to calculate:

$$q^\alpha(Y|\,\mathbf{x}) = x_1 + x_2 + x_3 + \sigma \times z_\alpha$$

with z_α, α-quantile of the standard normal distribution $\mathcal{N}(0,1)$.

Figure 2 shows for two specific points $\mathbf{x}$ and for several levels α ranging from 0.1 to 0.9, the distribution of the estimators of the conditional quantiles computed with the original dataset on the left-hand side and with the bootstrap samples on the right-hand side. The estimates have been calculated with the following setting: a sample of size $n = 10^4$, a forest grown with $n_{trees} = 500$ and the minimum number of samples required to be at a leaf node set to $min\_samples\_leaf = \lfloor \sqrt{n} \cdot \log(n)^{1.5}/250 \rfloor$. In order to assess the quality of the estimators at these points, the following indicators are computed by repeating the experiment $s = 500$ times.

$$RMSE(\mathbf{x}) = \sqrt{\frac{1}{s} \sum_{j=1}^{s} \left(\hat{q}^\alpha_j(Y|\,\mathbf{x}) - q^\alpha(Y|\,\mathbf{x}) \right)^2}$$

$$Bias(\mathbf{x}) = \left| \frac{1}{s} \sum_{j=1}^{s} \hat{q}^\alpha_j(Y|\,\mathbf{x}) - q^\alpha(Y|\,\mathbf{x}) \right|$$

$$Variance(\mathbf{x}) = \frac{1}{s} \sum_{j=1}^{s} \left(\hat{q}^\alpha_j(Y|\,\mathbf{x}) - \frac{1}{s} \sum_{j=1}^{s} \hat{q}^\alpha_j(Y|\,\mathbf{x}) \right)^2$$

where $\hat{q}^\alpha_j(\mathbf{x})$ is the estimator on the j's sample, $j = 1, \ldots, s$.

Based on the graphs obtained in Figure 2, it seems difficult to say if one estimator is better than the other. It appears that the performance of the two estimators differs a bit depending on the observation $\mathbf{x}$.

In order to get global measures of both estimators, we define an averaged version of the previous indicators computed with $p = 5 \times 10^4$ randomly chosen points $\mathbf{x}$ according to the

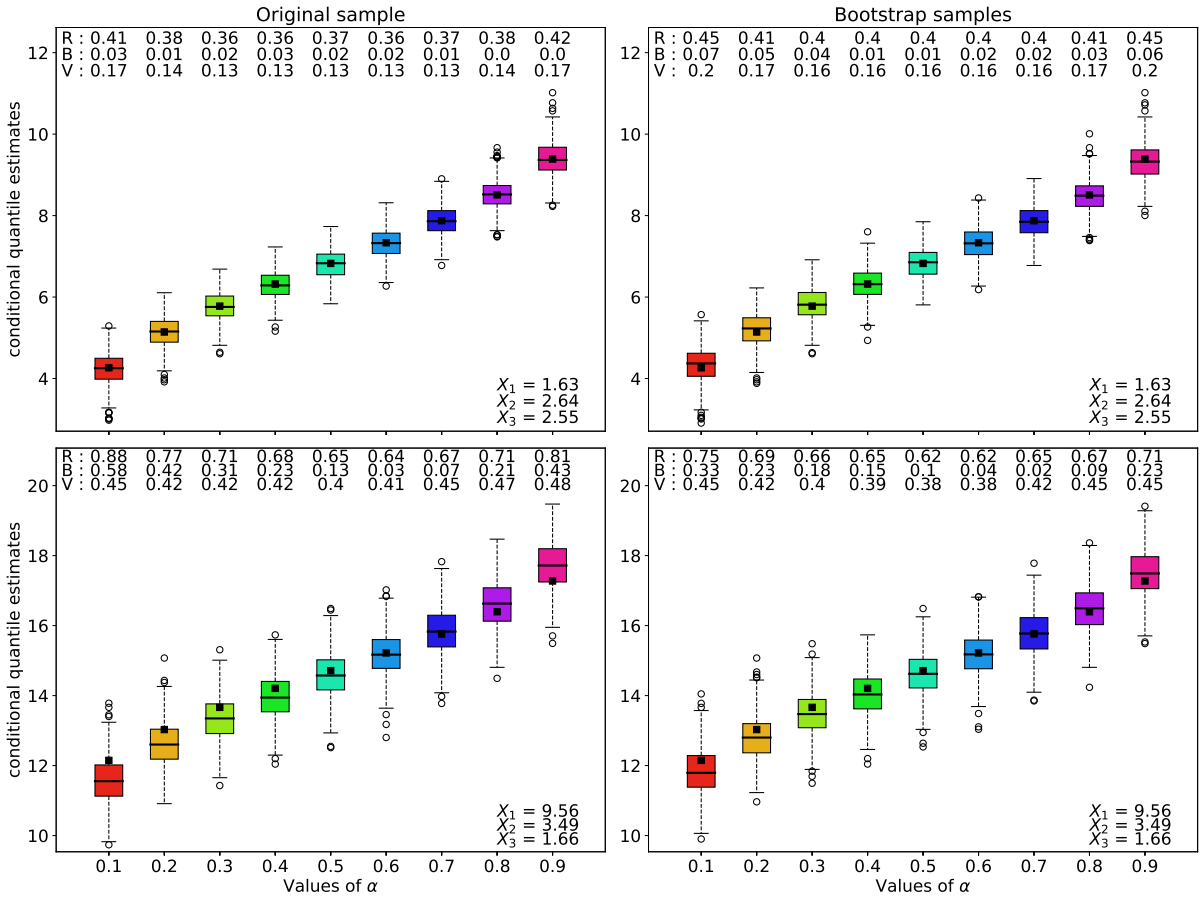

Figure 2: *Distribution of the conditional quantile approximations computed for three different values* $\mathbf{x}$ *by using the original sample (on the left side) and the bootstrap samples (on the right side). On each graph, the values above the boxplots are* `R` *for* $RMSE(\mathbf{x})$, `B` *for* $Bias(\mathbf{x})$ *and* `V` *for* $Variance(\mathbf{x})$.

	Original sample			Bootstrap samples		
	M_RMSE	M_Bias	$M\_Variance$	M_RMSE	M_Bias	$M\_Variance$
$\alpha = 0.1$	0.6382	0.2382	0.2826	0.6410	0.1926	0.3115
$\alpha = 0.2$	0.5868	0.2011	0.2565	0.6008	0.1669	0.2846
$\alpha = 0.3$	0.5640	0.1791	0.2519	0.5837	0.1490	0.2791
$\alpha = 0.4$	0.5521	0.1638	0.2544	0.5748	0.1351	0.2808
$\alpha = 0.5$	0.5470	0.1530	0.2615	0.5714	0.1274	0.2874
$\alpha = 0.6$	0.5489	0.1482	0.2758	0.5736	0.1276	0.3010
$\alpha = 0.7$	0.5602	0.1530	0.3053	0.5836	0.1360	0.3298
$\alpha = 0.8$	0.5901	0.1766	0.3659	0.6085	0.1562	0.3890
$\alpha = 0.9$	0.6786	0.2443	0.6526	0.6842	0.2074	0.6837

Table 1: *Results of the averaged* $RMSE(\mathbf{x})$, $Bias(\mathbf{x})$ *and* $Variance(\mathbf{x})$ *computed over* $p = 5 \times 10^4$ *observations of* $\mathbf{x}$.

following formulas:

$$M\_RMSE := \frac{1}{p}\sum_{j=1}^{p} RMSE\left(\mathbf{x}^j\right)$$

$$M\_Bias := \frac{1}{p}\sum_{j=1}^{p} Biais\left(\mathbf{x}^j\right)$$

$$M\_Variance := \frac{1}{p}\sum_{j=1}^{p} Variance\left(\mathbf{x}^j\right)$$

By using the same setting as previously for the estimators, the numerical results for these three measures are listed in Table 1 for several levels α. First of all, both estimators have an equivalent RMSE, whereas the estimator computed with the bootstrap samples has the smallest bias for all α. Concerning the variance, the original dataset based estimator is the one with the smallest variance, which may seem surprising. Indeed, random forest method is an ensemble learning method that begins with bagging (the bootstrapped aggregation of regression tree predictions) in order to reduce the variance of the prediction function. Thus, for the particular case of the conditional quantile approximation, it is observed the opposite phenomenon on our example. Finally, it has to be noted that the performance of the two estimators depends on the level α.

7 Conclusion

This article proposes two conditional distribution functions and conditional quantiles approximations based on random forests. The former is a natural generalisation of the random forest estimator of the regression function making use of the bootstrap samples, while the latter is based on a variant using only the original dataset.

The consistency of the bootstrap samples based estimator is shown under realistic assumptions and constitutes the major contribution of this paper. Indeed, this is the first consistency result handling the bootstrap component in a random forest method whereas it is usually replaced by subsampling. As for the second estimator, the consistency proof established in Meinshausen (2006) for a simplified random forest model is extended to a realistic one by taking into account all the randomness used in the procedure. The two estimators have close performances on our toy example. A specific interest of the bootsrap estimation is that the Out-Of-Bag samples could be used for cross-validation and / or back-testing procedures.

The estimators developed in this paper rest on trees grown with the CART-split criterion. But the assumptions providing the consistency results are detached from the split procedure used. Thus, the theoretical tools developed here could be useful for a large class of methods by just changing the splitting scheme. An ambitious additional work would be to develop a theoretical analysis for obtaining convergence rates and also to construct confidence intervals.

Acknowledgments

We are grateful to Andrés Cuberos, Ecaterina Nisipasu, Mathieu Poulin and Przemyslaw Sloma from SCOR for their valuable comments and support. We are also much indebted to Roland Denis and Benoit Fabrèges for intensive support on computational aspects.

References

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

Arenal-Gutiérrez, E., Matrán, C., and Cuesta-Albertos, J. A. (1996). Unconditional glivenko-cantelli-type theorems and weak laws of large numbers for bootstrap. *Statistics & probability letters*, 26(4):365–375.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98.

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

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. *Journal of Multivariate Analysis*, 101(10):2499–2518.

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

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

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

Breiman, L. (2004). Consistency for a simple model of random forests.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.

Browne, T., Fort, J.-C., Iooss, B., and Le Gratiet, L. (2017). Estimate of quantile-oriented sensitivity indices.

Davies, A. and Ghahramani, Z. (2014). The random forest kernel and other kernels for big data from random partitions. *arXiv preprint arXiv:1402.4293*.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Elie-Dit-Cosaque, K. (2020). qosa-indices, a python package available at: `https://gitlab.com/qosa_index/qosa`.

Fabrège, B. and Maume-Deschamps, V. (2020). Conditional distribution forest: a julia package available at `https://github.com/bfabreges/conditionaldistributionforest.jl`.

Fort, J.-C., Klein, T., and Rachdi, N. (2016). New sensitivity analysis subordinated to a contrast. *Communications in Statistics-Theory and Methods*, 45(15):4349–4364.

Goehry, B. (2019). Random forests for time-dependent processes.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

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

Maume-Deschamps, V. and Niang, I. (2018). Estimation of quantile oriented sensitivity indices. *Statistics & Probability Letters*, 134:122–127.

Maume-Deschamps, V., Rullière, D., and Usseglio-Carleve, A. (2017). Quantile predictions for elliptical random fields. *Journal of Multivariate Analysis*, 159:1–17.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Meinshausen, N. (2019). Quantile regression forests, a r package available at `https://cran.r-project.org/package=quantregforest`.

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

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Scornet, E. (2016b). Promenade en forêts aléatoires. *MATAPLI*, 111.

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

Scornet, E., Biau, G., and Vert, J.-P. (2015a). Supplementary materials for: Consistency of random forests. *arXiv*, 1510.

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

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

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

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

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.