ESAIM: PS 24 (2020) 801–826
https://doi.org/10.1051/ps/2020015

ESAIM: Probability and Statistics
www.esaim-ps.org

RANDOM FORESTS FOR TIME-DEPENDENT PROCESSES

Benjamin Goehry*

Abstract. Random forests were introduced by Breiman in 2001. We study theoretical aspects of both original Breiman's random forests and a simplified version, the centred random forests. Under the independent and identically distributed hypothesis, Scornet, Biau and Vert proved the consistency of Breiman's random forest, while Biau studied the simplified version and obtained a rate of convergence in the sparse case. However, the i.i.d hypothesis is generally not satisfied for example when dealing with time series. We extend the previous results to the case where observations are weakly dependent, more precisely when the sequences are stationary $\beta-$mixing.

Mathematics Subject Classification. 62M10.

Received 20 November 2018. Accepted 9 April 2020.

1. Introduction

Random forests were introduced in 2001 by Breiman in [6] and are since then extremely successful as a regression and classification method. The popularity comes from the wide range of applications in which they are used and the accuracy they offer in high-dimensional problems. They are also easy to implement, can be easily parallelizable and require only few tuning parameters. We can cite as successful applications: chemo-informatics [24], ecology [9, 19], 3D object recognition [23] and time series prediction [12, 14].

Let a stationary random sequence $(X_t, Y_t)_{t\in\mathbb{Z}} \in \mathbb{R}^p \times \mathbb{R}$ be such that

$$Y_t = f(X_t) + \epsilon_t \tag{1.1}$$

and the error ϵ_t is such that $\mathbb{E}[\epsilon_t|X_t] = 0$. The purpose of random forests is to estimate the regression function

$$\forall x \in \mathbb{R}^p, f(x) = \mathbb{E}[Y_t \mid X_t = x].$$

In the statistical context we only observe a training sample $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ used to build the random forest estimator denoted by $\hat{f}_n$.

Random forests can be related to two main sources, regression trees [8] and bagging [5]. Regression trees are constructed by a recursive partitioning of the input space based on some criterion, dependent or independent of the data (we detail precisely two in the following), to estimate the regression function f. At each step of the tree construction, a split is selected (a variable and a location on the variable) based on the evaluation of the criterion

Keywords and phrases: Statistics, random forests, time-dependent processes.

Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France

* Corresponding author: benjamin.goehry@math.u-psud.fr

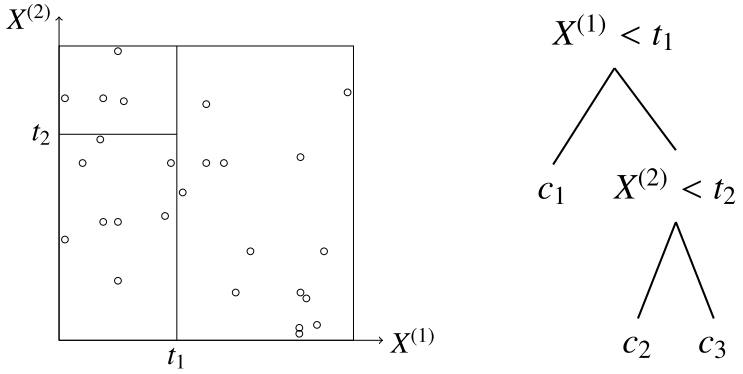

FIGURE 1. A partitioning of $[0,1]^2$ and the associated binary tree. c_1, c_2, c_3 are the constants associated to each cell.

among all the admissible splits based on all the variables. The cell is cut in two on the selected split and the previous step is reiterated on the new cells. A tree is then a piecewise constant decomposition of the input space. We can associated to the input space partitioning a binary tree where each node corresponds to a test matching how the input space was cut. An illustration is given in Figure 1 of a partitioning in the two-dimensional space and its associated binary tree. The principle of bagging (short form of *bootstrap aggregating*) is to create M randomly generated training sets by randomly sampling α_n observations with or without replacement from the set $\mathcal{D}_n$ and to construct on each set a predictor. Once the predictors are constructed, the bagging prediction for a new observation x is an aggregation, generally the empirical mean, of the predictions given by the M predictors for the point x. This procedure aims to improve stability and accuracy of the base predictor. In the context of random forests, the predictors are regression trees.

We study two variants of random forests, the random forest-random input and the centred forest. By construction of the bagging, each predictor is computed in the same way. In order to explain the different procedures we then have to explicit the construction of one predictor. Let us begin with the variant which remains to this day the most commonly used and referred to as the original Breiman's random forest, the random forest-random input (RF-RI). For a given generated training set of α_n points, a tree is computed using the CART [8] criterion: at each node of the tree the best split is selected by minimising the intra-node variance. This criterion is detailed in Section 2.1. A subtlety of the RF-RI is to restrict at each node the minimisation of the criterion on a random subset of m_{try} variables rather than on the p variables and thereby increase the diversity of the predictors by adding randomness in the construction. This is then recursively repeated until a stopping criterion is met, typically when the number of nodes reached a given number or when the number of observations in each node is below a given threshold.

The RF-RI have received increasing attention in recent years regarding theoretical analysis and we can cite for example the works described in [18, 21, 22, 25]. Since notations are only set later on for ease of readability, we decide to develop Section 2.1 with the exception of the result in [22] on which the present work relies on and doesn't require additional notations. Assuming that the observations $(X_i, Y_i)_{1\leq i\leq n}$ are independent and identically distributed as (X, Y), they establish the consistency of the pruned version (that is, the depth of the trees is controlled by a parameter) of the RF-RI, *i.e.* that $\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 \to 0$ as $n \to +\infty$, for trees where points are selected without replacement and the regression function is an additive model. Under an additional assumption, yet hard to verify in general, they also established consistency of the unprunned version (that is, the depth of a tree is not controlled) which is almost the algorithm commonly used in practice.

The second variant of random forests we study belongs to the so-called *purely random forests*'s family. The RF-RI is based on the CART criterion which is heavily data-dependent, the criterion depends on both the

position of the X_i and the value of the Y_i to choose the best split, while the purely random forests are based on criteria which are independent of the data. The variant we consider is called *centred forest* which was introduced in [7]. The first difference with the RF-RI is that there is no re-sampling step, meaning that the set used to compute the trees is $\mathcal{D}_n$. A tree is then recursively constructed as follows. At each node, a coordinate is chosen uniformly or according to some probability independent of the data and the split is performed in the middle of the cell along the selected coordinate. This kind of variants has been preferred for statistical analysis since they are easier to define, provide non-asymptotic risk bound giving insight in the choice of the parameters of the forest but also capture some attractive features of the original random forest as the variance reduction by randomisation and adaptive variable selection. Under the hypothesis that $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d, [2] established that if the splits concentrate on the relevant variables then the procedure adapts to sparsity by giving a rate of convergence which depends on the number of strong features. We refer to [3] for a complete theoretical survey on random forests.

The aforementioned theoretical results are established under the condition that the observations are independent and identically distributed. However, in applications, it is very common to have dependent data instead of independent one such as in time series and random forests are proven to perform well on these kind of observations. We may cite as an example of successful applications of random forests in time series [11, 12, 14, 15]. In this regard, many algorithms were studied in the case of weakly dependent observations, and in particular, when dealing with $\beta-$mixing sequences. The $\beta-$mixing provides some kind of measure of how the dependence between observations decreases as the distance between them increases. It is usually difficult to estimate the mixing rates in practice. However, $\beta-$mixing sequences can be theoretically well-studied and estimated for various classes of random processes as Gaussian or Markov processes. We refer to [10] and [20] for more details about dependent processes. The general problem of one-step ahead predicting of time series was considered in [17] when the time series satisfies $\beta-$mixing and stationary condition, establishing consistency and rates of convergence for a certain class of functions which complexity and memory are determined by the data and minimising the structural risk. Consistency and a rate of convergence are also established for the boosting algorithm in [16] when the observations are stationary $\beta-$mixing. Their rate of convergence has an additional term, we also find in our analysis, which can be viewed as a penalty when considering $\beta-$mixing sequences instead of independent observations, $\mathcal{O}\left(n^{1-a(r_\beta+1)}\right)$ with $a \in [0,1)$ and where r_β measures the dependence of the mixing sequence we precise later on.

The paper is organised as follows: we first formalise the models studied and then set the statistical framework together with the notion of $\beta-$mixing sequences. We then state our contribution, including the extension of the aforementioned results to the case where observations are weakly dependent, namely the consistency of the RF-RI when trees are not fully grown and the rate of convergence of centred random forests. The proofs are postponed to the appendices for ease of readability.

2. Models

In this section, we formalise the previous mentioned models, namely the RF-RI and the centred random forest.

Recall that a random forest (either RF-RI or simpler models) is a collection of M random trees, computed in the same way, and the trees are constructed from a recursive partitioning of the input space $\mathcal{X}$ to which a binary tree can be associated matching how the input space was cut. We denote for the jth random tree, the predicted value at the point x, $\hat{f}_n\left(x;\Theta_j;\mathcal{D}_n\right)$ where $(\Theta_1,\ldots,\Theta_M)$ are independent and identically distributed as Θ and independent of $\mathcal{D}_n$. The random variable Θ is defined later on depending on the variant. The jth random tree is defined as follows

$$\hat{f}_n\left(x;\Theta_j;\mathcal{D}_n\right) = \sum_{i\in\mathcal{D}_n(\Theta_j)} \frac{\mathbb{1}_{X_i\in A_n(x;\Theta_j;\mathcal{D}_n)}Y_i}{N_n\left(x;\Theta_j;\mathcal{D}_n\right)}\mathbb{1}_{E_n(X,\Theta_j)}$$

where $\mathcal{D}_n(\Theta_j)$ is the data set which can be dependent on the random variable Θ_j for example if re-sampling or sub-sampling is used to construct the jth tree. The cell containing the point x is denoted $A_n(x, \Theta_j, \mathcal{D}_n)$,

$$N_n(x; \Theta_j; \mathcal{D}_n) = \sum_{i=1}^{n} \mathbb{1}_{X_i \in A_n(x, \Theta_j, \mathcal{D}_n)} = \#\left\{i \in \{1, \dots, n\}, X_i \in A_n(x, \Theta_j, \mathcal{D}_n)\right\}$$

and $E_n(x, \Theta_j)$ the event defined by $\{N_n(x, \Theta_j) \neq 0\}$. This means that each random tree outputs for a new point x the average value over all Y_i for which the corresponding X_i falls into the cell $A_n(x, \Theta_j, \mathcal{D}_n)$ of the random partition.

In the regression case, we aggregate the predictions by taking the average in the following way to get the random forest estimator

$$\hat{f}_{M,n}(x; \Theta_1, \dots, \Theta_M; \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} \hat{f}_n(x, \Theta_j, \mathcal{D}_n). \tag{2.1}$$

Since M can be chosen as large as possible in practice, we study the properties of the infinite random forest estimate which is obtained as the limit of equation (2.1) when the number of trees M grows to infinity. The law of large numbers then justifies using

$$\hat{f}_n(x, \mathcal{D}_n) = \mathbb{E}_\Theta\left[\hat{f}_n(x, \Theta, \mathcal{D}_n)\right]$$

instead of $\hat{f}_{M,n}(x; \Theta_1, \dots, \Theta_M; \mathcal{D}_n)$, where $\mathbb{E}_\Theta$ denotes expectation with respect to Θ conditionally to $\mathcal{D}_n$. In the following, to ease legibility we omit the dependency on $\mathcal{D}_n$ and denote simply $\hat{f}_n(x) := \hat{f}_n(x, \mathcal{D}_n)$.

2.1. Random forest - random input

We begin by recalling the variant of random forest which is the most commonly used in practice, the random forest-random input. We denote:

- $\alpha_n \in \{1, \dots, n\}$ the number of sampled data points in each tree;
- $m_{try} \in \{1, \dots, p\}$ the preselected number of variables for splitting;
- $\tau_n \in \{1, \dots, \alpha_n\}$ the number of leaves in each tree.

Here we consider the stopping criterion where the number of leaves must not exceed the given parameter τ_n. The random forest is then computed as detailed in Algorithm 1. We shall make a remark regarding the selection of the nodes. They are not chosen uniformly among all the childless nodes, otherwise, there could exist tree branches far more developed than other only because of randomness. Usually, all nodes of a given level are split

(if permitted) then the algorithm considers the nodes of the next level and so on. This remark holds also for the centred forest in Section 2.2.

Algorithm 1: Random forest - random input
input: Training set $((X_1, Y_1), \ldots, (X_n, Y_n))$
parameters: number of trees M, number of observations per tree α_n, size of preselected variables for splitting m_{try}, number of leaves τ_n
for $j \leftarrow 1$ *to* M **do**
  Construct the jth tree:
  - Draw uniformly $\alpha_n \leq n$ points without replacement.
  - Set $n_{nodes} = 1$.
  - **while** $n_{nodes} < \tau_n$ **do**
    - Choose a childless node A, containing more than one observation.
    - Select uniformly (without replacement), the set $Mtry \subset \{1, \ldots, p\}$ such that $|Mtry| = m_{try}$.
    - Choose the best split in the cell A maximising the CART criterion, defined in equation (2.2), on $Mtry$.
    - Cut the cell A according to the best split. Let A_R and A_L be the cells we obtain.
    - $n_{nodes} = n_{nodes} + 1$.
  end
end
output for a new observation x: mean of the M predictions given by the trees for x.

The CART criterion is defined as follows. Let $\mathcal{C}_A$ be the set of all possible cuts in the cell A. For any $(j, z) \in \mathcal{C}_A$, the CART-split criterion takes the form

$$L_n(j,z) = \frac{1}{N_n(A)} \sum_{i=1, X_i \in A}^{n} (Y_i - \bar{Y}_A)^2 - \frac{1}{N_n(A)} \sum_{i=1, X_i \in A}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{X_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{X_i^{(j)} \geq z})^2, \tag{2.2}$$

with $X_i = \left(X_i^{(1)}, \ldots, X_i^{(p)}\right), A_L = \left\{x \in A, x^{(j)} < z\right\}, A_R = \left\{x \in A, x^{(j)} \geq z\right\}$ and $\bar{Y}_A$ (resp. $\bar{Y}_{A_L}, \bar{Y}_{A_R}$) is the average of the Y_i's belonging to A (resp. A_L, A_R).

Let us suppose that the observations $(X_i, Y_i)_{1 \leq i \leq n}$ are independent and identically distributed as (X, Y). A link between the error of the finite and infinite forest is established in [21] and shows that the error of the finite forest can be made arbitrary close to the infinite one provided that the number of trees is large enough,

$$\mathbb{E}\left[\hat{f}_{M,n}(X) - f(X)\right]^2 - \mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 \leq \frac{8\left(\|f\|_\infty + \sigma^2\left(1 + 4\log n\right)\right)}{M}$$

when ϵ is a centred Gaussian noise with finite variance $\sigma^2 > 0$ and independent of X. Another consequence of this result is that as soon as infinite random forests are consistent then the finite random forests are consistent provided that $\frac{\log n}{M} \underset{n\to\infty}{\longrightarrow} 0$. Asymptotic normality of random forests based on subsampling was proven in [18] when the subsample size α_n grows slower than $\sqrt{n}$, *i.e.* that $\frac{\alpha_n}{\sqrt{n}} \underset{n\to\infty}{\longrightarrow} 0$ and that the number of trees M varies with n, *i.e.* that $\frac{n}{M} \underset{n\to\infty}{\longrightarrow} C$ for some constant $C > 0$. However, this does not necessarily imply that random forests are asymptotically unbiased. This gap was filled in [25] and also established that the infinitesimal jackknife consistently estimates the forest variance under the less restrictive condition that the subsample size grows such that $\frac{\alpha_n \log n^p}{n} \underset{n\to\infty}{\longrightarrow} 0$.

2.2. Centred forest

We now recall the construction of the centred random forest introduced in [7], detailed in Algorithm 2.

Algorithm 2: Centred random forest
Data:$((X_1, Y_1), \dots, (X_n, Y_n))$
Initialisation: τ_n
Repeat recursively $\lceil \log_2 \tau_n \rceil$ times:
- At each node, select a coordinate $j \in \{1, \dots, d\}$ with probability $p_{n,j} \in (0, 1)$ where $\sum_{j=1}^{d} p_{n,j} = 1$;
- The split is performed at the centre of the cell along the selected variable.

We note that $\tau_n \geq 2$ is a fixed deterministic parameter which may depend on n but not on $\mathcal{D}_n$ and that each tree has exactly $2^{\lceil \log_2 \tau_n \rceil} \approx \tau_n$ nodes. However, there is no re-sampling step in the centred random forest algorithm and so $\mathcal{D}_n(\Theta) = ((X_1, Y_1), \dots, (X_n, Y_n))$.

3. Statistical framework

Let us denote $(W_t)_{t\in\mathbb{Z}} := (X_t, Y_t)_{t\in\mathbb{Z}}$ where (X_t, Y_t) are defined in equation (1.1). The first assumption throughout this paper is that the random sequence $(W_t)_{t\in\mathbb{Z}}$ is stationary. More precisely, we assume that $(W_t)_{t\in\mathbb{Z}}$ is a strongly stationary process as defined in Definition 3.1.

Definition 3.1. The process $(W_t)_{t\in\mathbb{Z}}$ is said to be (strongly) stationary if $\forall k \in \mathbb{N}, \forall (t_1, \dots, t_k) \in \mathbb{Z}^k$ and for all $\tau \in \mathbb{Z}$,

$$(W_{t_1+\tau}, \dots W_{t_k+\tau}) = (W_{t_1}, \dots W_{t_k})$$

in distribution.

In order to prove the consistency of the RF-RI we also need to assume that $(W_t)_{t\in\mathbb{Z}}$ is an ergodic process as defined in Definition 3.2.

Definition 3.2. The process $(W_t)_{t\in\mathbb{Z}}$ is said to be (mean-)ergodic if

$$\frac{1}{2K} \sum_{t=-K}^{K} W_t \xrightarrow[K\to\infty]{\mathrm{L}^2} \mathbb{E}(W_t).$$

Let $(C_n)_n$ be a positive sequence and define the truncation operator T_{C_n} by

$$T_{C_n} u = \begin{cases} u & \text{when } |u| \leq C_n \\ C_n & \text{when } |u| > C_n. \end{cases}$$

and the set

$$T_{C_n} \mathcal{G}_n = \{T_{C_n} g, g \in \mathcal{G}_n\}$$

where $\mathcal{G}_n = \mathcal{G}(\mathcal{D}_n)$ denotes a class of functions $g : \mathcal{X} \to \mathcal{Y}$. Following the definition of the truncation operator T we denote

$$W_L = T_L W$$

and

$$W_{i,L} = T_L W_i$$

for $W = X$ or Y.

The consistency proof in [22] relies on the general consistency theorem found in [13]. In order to extend the consistency result to the dependent case, we use the extension of the general consistency theorem to the stationary ergodic setting as stated in Proposition 3.3. We postpone the proof in Appendix B for ease of readability.

Proposition 3.3. *Let $(W_t)_{t\in\mathbb{Z}}$ be a stationary ergodic process and $\mathcal{D}_n$ a data set. Let $\mathcal{G}_n = \mathcal{G}(\mathcal{D}_n)$ be a class of functions $g : \mathcal{X} \to \mathcal{Y}$, $(C_n)_n$ a positive sequence, f the regression function in equation (1.1) and $\hat{f}_n$ an estimator which minimises the empirical L^2 risk on $\mathcal{G}_n$. If*

$$\lim_{n\to\infty} C_n = \infty, \tag{3.1a}$$

$$\lim_{n\to\infty} \mathbb{E}\left\{ \inf_{g\in\mathcal{G}_n, \|g\|_\infty \le C_n} \int |g(x) - f(x)|^2 \, \mu(\mathrm{d}x) \right\} = 0, \tag{3.1b}$$

$$\lim_{n\to\infty} \mathbb{E}\left\{ \sup_{g\in T_{C_n}\mathcal{G}_n} \left| \frac{1}{n}\sum_{i=1}^{n} |g(X_i) - Y_{i,L}|^2 - \mathbb{E}\left[g(X) - Y_L\right]^2 \right| \right\} = 0 \ \forall L > 0 \tag{3.1c}$$

then

$$\lim_{n\to\infty} \mathbb{E}\left\{ \int |\hat{f}_n(x) - f(x)|^2 \mu(\mathrm{d}x) \right\} = 0.$$

We recall the notion of weak dependence, more precisely the $\beta-$mixing case in which we establish the results.

Definition 3.4 (β-mixing process)**.** Let $\sigma_l = \sigma(W_1^l)$ and $\sigma'_{l+m} = \sigma(W_{l+m}^\infty)$ be the sigma-algebras of events generated by the random variables $W_1^l = (W_1, \dots W_l)$ and $W_{l+m}^\infty = (W_{l+m}, W_{l+m+1}, \dots)$. The β-mixing coefficients is given by

$$\beta_m = \sup_{l\ge 1} \mathbb{E}\left[\sup_{B\in\sigma'_{l+m}} |\mathbb{P}(B|\sigma_l) - \mathbb{P}(B)| \right]$$

where the expectation is taken with respect to σ_l.

A stochastic process is said to be absolutely regular, or β-mixed, if

$$\lim_{m\to\infty} \beta_m = 0.$$

The most common $\beta-$mixing coefficients are known as the algebraic and exponential mixing defined as follows,

1. Algebraic mixing: $\beta_m = \mathcal{O}(m^{-r_\beta})$ for $r_\beta > 0$.
2. Exponential mixing: $\beta_m = \mathcal{O}(\exp(-bm^{k_\beta}))$ for $b, k_\beta > 0$.

The exponential mixing hypothesis is stronger than algebraic mixing. The values r_β and k_β are called the mixing exponents and the i.i.d process can be recovered by taking either the limit $r_\beta \to +\infty$ for the algebraic mixing or $k_\beta \to +\infty$ for the exponential mixing.

The $\beta-$mixing property is appealing in the theoretical setting since many statistical properties are preserved under this condition and are easy to manipulate. One method to manipulate $\beta-$mixing sequences is by using a lemma established in [26], recalled in Lemma A.1. Using this lemma, the dependent process is approximated with independent blocks of observations plus some linear function in β.

4. Result for the RF-RI

We recalled the studied models and the notion of weak dependence. We need the following hypotheses to establish the consistency of the RF-RI when the observations are weakly dependent:

- **H1a:** the data set $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ is composed of stationary ergodic $\beta-$mixing $(X_i, Y_i) \in [0,1]^p \times \mathbb{R}$;
- **H2a:** the errors $(\epsilon_i)_{1 \leq i \leq n}$ are independent;
- **H3a:** the response Y follows the additive model

$$Y = \underbrace{\sum_{j=1}^{p} f_j(X^{(j)})}_{f(X)} + \epsilon$$

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

We can now state the result of consistency of random forests when the observations are weakly dependent under the regime $\tau_n < \alpha_n$ (*i.e.* the trees are not fully grown).

Theorem 4.1. *Assume the hypothesis of stationary ergodic $\beta-$mixing data **H1a**, the independent errors hypothesis **H2a** and that the additive model hypothesis **H3a**. If there exists a sequence a_n verifying $1 \leq a_n \leq n$ such that $\frac{\tau_n \log(\alpha_n)^9 a_n}{\alpha_n} \underset{n\to\infty}{\longrightarrow} 0$ and $\frac{\log(\alpha_n)^4 \beta_{a_n} \alpha_n}{a_n} \underset{n\to\infty}{\longrightarrow} 0$, then RF-RI are consistent, i.e.*

$$\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 \underset{n\to\infty}{\longrightarrow} 0.$$

Let us first verify if we recover the result in the independent case. If the observations $(X_i, Y_i)_{1 \leq i \leq n}$ are independent, $\beta_m = 0$ for all $m \geq 0$. We then get exactly the same hypotheses and result as in [22] by setting a_n equal to 1.

The hypotheses **H2a** and **H3a** are the same as in [22]. Note however that in the context of $\beta-$ mixing processes, the independent errors hypothesis **H2a** is not necessarily true but is assumed in some theoretical models as in the autoregressive model. We refer to [4] for a complete survey of processes verifying the $\beta-$mixing condition. An interesting perspective would be to extend the result to the case where the errors are not assumed to be i.i.d.

The condition $\frac{\tau_n \log(\alpha_n)^9 a_n}{\alpha_n} \longrightarrow 0$ as n tends to infinity is also highly similar to the last hypothesis in their theorem and recover it by setting a_n equal to 1. The last one is simply saying that the dependence between the data must not be too long in order to have consistency of the forest. Let us see how the dependence influences the number of leaves parameter τ_n. Let us suppose, in the following analysis, that r_β (or k_β in the exponential mixing case) is known. Let us consider the algebraic mixing case and suppose that $a_n = \alpha_n^a$ with $\frac{1}{1+r_\beta} < a < 1$. The last condition is then verified and the greatest value of τ_n must verify the following in order to obtain consistency:

$$\frac{\tau_n \log(\alpha_n)^9}{\alpha_n^{\frac{r_\beta}{1+r_\beta}}} \underset{n\to\infty}{\longrightarrow} 0.$$

In the exponential mixing case, suppose that $a_n = \frac{c}{b}\log(\alpha_n)^{\frac{1}{k_\beta}}$ with $c > 1$. The last condition is then equal to $\frac{\log(\alpha_n)^{4-\frac{1}{k_\beta}}}{\alpha_n^{c-1}}$ which tends to 0 as n tends to infinity. The penultimate condition can then be rewritten, implying that τ_n cannot be greater than the following condition is true,

$$\frac{\tau_n \log(\alpha_n)^{9+\frac{1}{k_\beta}}}{\alpha_n} \underset{n\to\infty}{\longrightarrow} 0.$$

This analysis leads to the following conclusion. The nature of the hypothesis appears in the choice of the parameter τ_n, influenced by r_β (or k_β): the stronger the dependence between the observations, meaning that r_β (or k_β) is small, the shallower the trees need to be compared to the trees constructed based on i.i.d observations, in order to guarantee convergence.

5. Results on centred random forest

We analyse now the convergence rates of the centred random forest model when the observations are stationary β-mixing. The space $[0,1]^p$ is equipped with the standard Euclidean metric. We analyse the centred random forest in a sparse framework; this arises from the fact that in many applications the true dimension is always smaller than p. We assume that the regression function only depends on a nonempty subset $\mathcal{S}$ of the p features. We use the letter S to denote the cardinal of $\mathcal{S}$. Based on this assumption we have

$$f(X) = \mathbb{E}[Y|X_{\mathcal{S}}]$$

where $X_{\mathcal{S}} = \left\{X^{(i)}, i \in \mathcal{S}\right\}$. Let us introduce $f^* : [0,1]^S \to \mathbb{R}$ that is the section of f corresponding to $\mathcal{S}$. We then have

$$f(X) = f^*(X_{\mathcal{S}}).$$

We also need the following hypotheses to establish the results:

- **H1b:** the data set $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is composed of stationary $\beta-$mixing $(X_i, Y_i) \in [0,1]^p \times \mathbb{R}$;
- **H2b:** the errors $\epsilon_i := Y_i - f(X_i)$ are independent of finite variance $\sigma^2 > 0$.

5.1. Convergence rates

We first decompose $\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2$ with the variance/bias decomposition:

$$\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 = \underbrace{\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2}_{\text{Variance}} + \underbrace{\mathbb{E}\left[\tilde{f}_n(X) - f(X)\right]^2}_{\text{Bias}}$$

where

$$\tilde{f}_n(X) = \sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{n,i}(X,\Theta)\right] f(X_i). \tag{5.1}$$

with

$$W_{n,i}(X,\Theta) = \frac{\mathbb{1}_{X_i \in A_n(X,\Theta)}}{\sum_{k=1}^{n} \mathbb{1}_{X_k \in A_n(X,\Theta)}} \mathbb{1}_{E_n(X,\Theta)} \quad \forall i \in \{1, \ldots n\}.$$

We assume throughout that the coordinate-sampling probabilities are such that $p_{n,j} = \frac{1}{S}(1+\nu_{n,j})$ for $j \in \mathcal{S}$ and $p_{n,j} = \nu_{n,j}$ otherwise where each $\nu_{n,j}$ tends to 0 as n tends to infinity.

The first result concerns the variance term and the second the bias term.

Proposition 5.1. *Assume the hypotheses of stationary $\beta-$mixing data **H1b**, independent errors **H2b** and that X is uniformly distributed on $[0,1]^p$ .Assuming that the coordinate-sampling probabilities are such that $p_{n,j} = \frac{1}{S}(1+\nu_{n,j})$ for $j \in \mathcal{S}$ and if there exists a sequence a_n verifying $1 \leq a_n \leq n$ then*

$$\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 \leq C\sigma^2 \left(\frac{S^2}{S-1}\right)^{S/2p} (1+\nu_n) \frac{\tau_n a_n^2}{n(\log \tau_n)^{S/2p}} + \sigma^2 \frac{\beta_{a_n} n}{a_n}$$

where

$$C = \frac{576}{\pi}\left(\frac{\pi \log 2}{16}\right)^{S/2p}$$

and

$$1+\nu_n = \prod_{j\in\mathcal{S}} \left[(1+\nu_{n,j})^{-1}\left(1-\frac{\nu_{n,j}}{S-1}\right)^{-1}\right]^{1/2p}.$$

As noted in [2], if $p_{\text{lower}} < p_{n,j} < p_{\text{upper}}$ for some constants $p_{\text{lower}}, p_{\text{upper}} \in (0,1)$ we have

$$1+\nu_n \leq \left(\frac{S-1}{S^2 p_{\text{lower}}(1-p_{\text{upper}})}\right)^{\frac{S}{2p}}.$$

Proposition 5.2. *Assume the hypotheses of stationary $\beta-$mixing data **H1b**, X is uniformly distributed on $[0,1]^p$ and f^* is L-Lipschitz on $[0,1]^S$. Assuming that the coordinate-sampling probabilities are such that $p_{n,j} = \frac{1}{S}(1+\nu_{n,j})$ for $j \in \mathcal{S}$ and if there exists a sequence a_n verifying $1 \leq a_n \leq n$ then*

$$\mathbb{E}\left[\tilde{f}_n(X) - f(X)\right]^2 \leq \frac{2SL^2 a_n}{\tau_n^{\frac{0,75}{S\log 2}(1+\gamma_n)}} + \exp\left(-\frac{\mu_n}{2\tau_n}\right) \sup_{x\in[0,1]^p} f^2(x) + \frac{\beta_{a_n} n}{a_n}\left[SL^2 + \sup_{x\in[0,1]^p} f^2(x)\right]$$

where $\gamma_n = \min_j \nu_{n,j}$.

The bias in the weakly dependent case only depends on the true dimension and not p which confirms the intuition and the result in the independent case as noted in [2]. However, we should keep in mind, whether in the dependent or independent setting, that the result relies on the assumption that the splits concentrate on the relevant variables.

Using the inequality $z\exp(-nz) \leq \frac{1}{en}$ for $z \in (0,1]$ and combining both previous convergence rates we get the following result.

Theorem 5.3. *Assume the hypotheses of stationary $\beta-$mixing data **H1b**, independent errors **H2b**, X is uniformly distributed on $[0,1]^p$ and f^* is L-Lipschitz on $[0,1]^S$. Assuming that the coordinate-sampling probabilities*

are such that $p_{n,j} = \frac{1}{S}(1+\nu_{n,j})$ *for* $j \in \mathcal{S}$ *and if there exists a sequence* a_n *verifying* $1 \leq a_n \leq n$ *then*

$$\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 \leq \frac{2SL^2 a_n}{\tau_n^{\frac{0,75}{S\log 2}(1+\gamma_n)}} + \mathcal{C}_{1,n}\frac{\tau_n a_n^2}{n} + \mathcal{C}_2\frac{\beta_{a_n} n}{a_n}$$

with

$$\mathcal{C}_{1,n} = 4e^{-1}\sup_{x\in[0,1]^p} f^2(x) + C\sigma^2\left(\frac{S^2}{S-1}\right)^{S/2p}(1+\nu_n),$$

$$\mathcal{C}_2 = SL^2 + \sigma^2 + \sup_{x\in[0,1]^p} f^2(x).$$

The remark done previously on the independent error hypothesis **H2a** holds obviously for **H2b** as well. The hypothesis $X \sim \mathcal{U}(0,1)^p$ is only a convenience and can be easily extended to the case where X admits a Lebesgue density which is lower and upper bounded.

We also recover the convergence rate in the independent setting given in [2] up to a constant factor. Let us suppose that we are in the independent case hence $\beta_m = 0$ for all $m \geq 1$. Setting a_n equal to 1 and plugging into Propositions 5.1 and 5.2, we get exactly the same upper bound for the variance as in [2] back. However, regarding the bias term, we get a term with $\exp\left(-\frac{n}{4\tau_n}\right)$ instead of $\exp\left(-\frac{n}{2\tau_n}\right)$ which is due to a necessary pre-processing needed in order to work with $\beta-$mixing sequences.

Under the hypothesis of algebraic mixing and thus exponential mixing, the term depending on β is converging to 0 when n tends to infinity. The last term shows the price we must pay when dealing with $\beta-$mixing sequences instead of independent observations. More precisely, under algebraic mixing the penalty is of the form $\mathcal{O}\left(n^{1-a(r_\beta+1)}\right)$ with $a \in [0,1)$ which is the same penalty as in the convergence rate of boosting established in [16]. The following corollary precises, under algebraic and exponential mixing conditions, the choices of τ_n with the associated upper bound on the rate of consistency.

Corollary 5.4. *Suppose that* r_β *and* k_β *are known.*

1. *Under algebraic mixing condition; choosing*

$$a_n \propto n^{\frac{1.5+S\log 2}{2.25+2S\log 2+r_\beta(0.75+S\log 2)}}.$$

This implies that the parameter τ_n *is of the form*

$$\tau_n \propto n^{\frac{(1+r_\beta)S\log 2}{2.25+2S\log 2+r_\beta(0.75+S\log 2)}}$$

and achieves the following convergence rate:

$$\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 = \mathcal{O}\left(n^{\frac{-0.75r_\beta+0.75+S\log 2}{r_\beta(0.75+S\log 2)+2.25+2S\log 2}}\right).$$

2. *Under exponential mixing; taking*

$$a_n \propto \log n^{\frac{1}{k_\beta}}$$

gives

$$\tau_n \propto \left(\frac{n}{\log n^{\frac{1}{k_\beta}}} \right)^{\frac{S \log 2}{0.75 + S \log 2}}.$$

Plugging into equation (**??**) *we get*

$$\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 = \mathcal{O}\left(\left(\frac{\log n^{\frac{1.5 + S \log 2}{k_\beta}}}{n^{0.75}} \right)^{\frac{1}{0.75 + S \log 2}} \right).$$

The form of the convergence rate under algebraic mixing condition implies that in order to have consistency, we need the couple (r_β, S) to satisfy the inequality $0.75 + S \log 2 < 0.75 r_\beta$. It also implies that this result only treats the case where $r_\beta \geq 1.41$. We note that we recover the same optimal parameter and convergence rate as in [2] by letting r_β go to infinity. Under exponential mixing condition, the chosen τ_n is, up to a logarithmic factor in the denominator, the optimal parameter in the i.i.d case and gives the same convergence rate up to a logarithmic term depending on the inverse of k_β.

The previous analysis leads to the following conclusion. The choice of the parameter τ_n is determined by the nature of the hypothesis; the stronger the dependence between the observations, meaning that r_β (or k_β) is small, the shallower the trees need to be compared to the trees constructed based on i.i.d observations, in order to guarantee convergence.

6. Conclusion

The results for either the random forest-random input or the centred forest lead to the same conclusion: the more the dependence between the observations is long, the shallower the trees need to be compared to the trees constructed based on independent and identically distributed observations.

Theses results may also lead to new variants of random forests. The proofs of the results are based on a decomposition in blocks of the random process and the blocks are close to being independent. An analogy can be drawn between this decomposition and the so-called *block bootstraps* commonly used in time series estimation. Instead of considering the observations one by one, the algorithm is fed with blocks of observations and lead to better estimations. It could be interesting to modify the random forest algorithm in the same way to get a random forest adapted to time series.

Appendix A. Proofs

The proofs are based on the construction and lemma given in [26], also recalled below, but we note that a similar coupling lemma is proved in [1].

We divide the sequence $(W_i)_{1 \leq i \leq n}$ into $2\mu_n$ blocks each of size a_n. We assume that $n = 2\mu_n a_n$ and so consider that there is no remaining terms. We then define for $1 \leq i \leq \mu_n$,

$$\begin{aligned} H_j &= \{i : 2(j-1)a_n + 1 \leq i \leq (2j-1)a_n\} \\ T_j &= \{i : (2j-1)a_n + 1 \leq i \leq 2ja_n\}. \end{aligned}$$

and we denote

$$\begin{aligned} W^{(j)} &= \{W_i, i \in H_j\} \\ W'^{(j)} &= \{W_i, i \in T_j\}. \end{aligned}$$

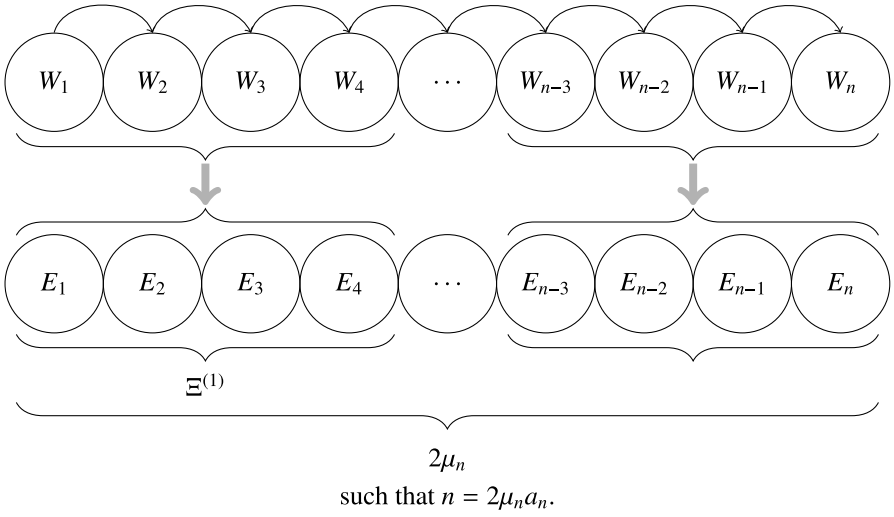

FIGURE A.1. Construction of the new independent sequence Ξ.

We then denote the sequence of H-blocks $W_{a_n} = \left(W^{(j)}\right)_{1\leq j\leq \mu_n}$. We construct a sequence of independently distributed blocks $\Xi_{a_n} = \left(\Xi^{(j)}\right)_{1\leq j\leq \mu_n}$ where $\Xi^{(j)} = \{\xi_i, i \in H_j\}$ and such that for all $j \in \{1, \ldots, n\}$,

$$W^{(j)} \overset{(d)}{=} \Xi^{(j)}.$$

We construct in the same way a sequence of T-blocks. An illustration of this construction is given in Figure A.1.

Lemma A.1 ([26]). *Let the distributions of W_{a_n} and Ξ_{a_n} be $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ respectively. Then for any measurable function u on $\mathbb{R}^{a_n\mu_n}$ with bound m,*

$$|\mathbb{E}_{\mathcal{Q}} u(W_{a_n}) - \mathbb{E}_{\tilde{\mathcal{Q}}} u(\Xi_{a_n})| \leq m\mu_n\beta_{a_n}.$$

A.1 Proof of theorem 4.1

The proof consists in applying Proposition 3.3. The computation of the approximation error is the same as in [22] since it does not require the independence of $(X_i, Y_i)_{1\leq i\leq n}$ but only stationarity and that the errors $(\epsilon_i)_{1\leq i\leq n}$ are independent. This verifies equation (3.1b).

The partition obtained with the random variable Θ and the data set $\mathcal{D}_n$ is denoted by $\mathcal{P}_n(\mathcal{D}_n, \Theta)$. We let

$$\Pi_n(\Theta) = \{\mathcal{P}\left((x_1, y_1), \ldots, (x_n, y_n), \Theta\right), (x_i, y_i) \in [0,1]^p \times [0,1]\}$$

be the family of all achievable partitions with random parameter Θ. We let

$$M\left(\Pi_n(\Theta)\right) = \max\left\{\text{Card}(\mathcal{P}, \mathcal{P} \in \Pi_n(\Theta)\right\}$$

be the maximal number of terminal nodes among all partitions in $\Pi_n(\Theta)$.

Given a set $z_1^n = \{z_1, \ldots, z_n\} \subset [0,1]^p$, $\Gamma_n\left(z_1^n, \Pi_n(\Theta)\right)$ denotes the number of distinct partitions of z_1^n induced by elements of $\Pi_n(\Theta)$, that is, the number of different partitions $\{z_1^n \cap A, A \in \mathcal{P}\}$ of z_1^n, for $\mathcal{P} \in \Pi_n(\Theta)$.

Consequently, the partitioning number $\Gamma_n\left(\Pi_n(\Theta)\right)$ is defined by

$$\Gamma_n\left(\Pi_n(\Theta)\right) = \max\left\{\Gamma\left(z_1^n, \Pi_n(\Theta)\right), z_1, \dots, z_n \in [0,1]^p\right\}.$$

Let $\mathcal{G}_n(\Theta)$ be the set of all functions $g : [0,1]^p \to \mathbb{R}$ piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$. We define as in [22], $C_n = \|f\|_\infty + \sigma\sqrt{2}\log\left(\alpha_n\right)^2$, hence equation (3.1a) is verified.

Regarding the estimation error, it is very similar to the computation done in [22] but we need to use a result established in [17] to introduce the $\beta-$mixing coefficient. This will prove equation (3.1c).

Theorem A.2. *Let $(W_t)_{t\in\mathbb{Z}}$ be a β-mixing stationary stochastic process, with $|Y_i| \le A_n$ and let $\mathcal{G}_n$ be a class of functions $g : \mathbb{R}^p \to \mathbb{R}$. Then, for any $d \ge 2$,*

$$\begin{aligned}&\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}_n(\Theta)\\ \|g\|\le A_n}}\left|\frac{1}{n}\sum_{j=1}^{n}|Y_j - g(X_j)|^d - \mathbb{E}\left[Y - g(X)\right]^d\right| > \delta\right)\\ &\le 8\mathbb{E}\mathcal{N}\left(\frac{\delta}{32d(2A_n)^{d-1}}, \mathcal{G}_n(\Theta), l_{1,n}\right)\exp\left(-\frac{\mu_n\delta^2}{128(2A_n)^{2d}}\right) + 2\mu_n\beta_{a_n}\end{aligned}$$

where $\mathcal{N}\left(\nu, \mathcal{G}(\Theta), l_{1,n}\right)$ is the ν-covering number of $\mathcal{G}_n(\Theta)$ w.r.t $l_{1,n} := \frac{1}{n}\sum_{i=1}^{n}|f(X_i) - g(X_i)|$.

Using Theorem A.2 we get,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}_n(\Theta)\\ \|g\|\le C_n}}\left|\frac{1}{\alpha_n}\sum_{i=1}^{\alpha_n}|g(X_i) - Y_{i,L}|^2 - E|g(X) - Y_L|^2\right| > \delta\right)\\ &\le 8\mathbb{E}\mathcal{N}\left(\frac{\delta}{128C_n}, \mathcal{G}_n(\Theta), l_{1,n}\right)\exp\left(-\frac{\mu_n\delta^2}{128(2C_n)^4}\right) + 2\mu_n\beta_{a_n}\end{aligned}$$

where $\alpha_n = 2\mu_{\alpha_n}a_{\alpha_n}$. For simplicity's sake, we denote $\mu_n = \mu_{\alpha_n}$ and $a_n = a_{\alpha_n}$.
Let us compute $\mathbb{E}\mathcal{N}\left(\frac{\delta}{128C_N}, \mathcal{G}_n(\Theta), l_{1,n}\right)$ (*cf.* [13]),

$$\begin{aligned}\mathcal{N}\left(\frac{\delta}{128C_n}, \mathcal{G}_n(\Theta), l_{1,n}\right) &\le \Gamma_n(\Pi_n(\Theta))\left[3\left(\frac{3e(2C_n)}{\frac{\delta}{128C_n}}\right)^2\right]^{M(\Pi_n(\Theta))}\\ &\le \Gamma_n(\Pi_n(\Theta))\left[3\left(\frac{768eC_n^2}{\delta}\right)^2\right]^{M(\Pi_n(\Theta))}\\ &\le \Gamma_n(\Pi_n(\Theta))\left[\frac{1331eC_n^2}{\delta}\right]^{2M(\Pi_n(\Theta))}.\end{aligned}$$

Hence

$$\mathbb{E}\mathcal{N}\left(\frac{\delta}{128C_n}, \mathcal{G}_n(\Theta), l_{1,n}\right) \le \mathbb{E}\left(\Gamma_n(\Pi_n(\Theta))\left[\frac{1331eC_n^2}{\delta}\right]^{2M(\Pi_n(\Theta))}\right).$$

Going back to the probability computation

$$\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}_n(\Theta)\\ \|g\|\le C_n}}\left|\frac{1}{\alpha_n}\sum_{i=1}^{\alpha_n}|g(X_i)-Y_i|^2-E|g(X)-Y|^2\right|>\delta\right)$$
$$\le 2\mu_n\beta_{a_n}+8\exp\left(-\frac{\mu_n\delta^2}{2048C_n^4}\right)\mathbb{E}\left(\exp\left(2M(\Pi_n(\Theta))\log\left(\frac{1331eC_n^2}{\delta}\right)\right)\exp\left(\log(\Gamma_n(\Pi_n(\Theta)))\right)\right).$$

Since $M(\Pi_n(\Theta))\le\tau_n$ and $\Gamma_n(\Pi_n(\Theta))\le(d\alpha_n)^{\tau_n}$,

$$\begin{aligned}
&2\mu_n\beta_{a_n}+8\exp\left(-\frac{\mu_N\delta^2}{2048C_N^4}\right)\mathbb{E}\left(\exp\left(2M(\Pi_N(\Theta))\log\left(\frac{1331eC_n^2}{\delta}\right)\right)\exp\left(\log(\Gamma_n(\Pi_n(\Theta)))\right)\right.\\
&\le 2\mu_n\beta_{a_n}+8\exp\left(-\frac{\mu_n\delta^2}{2048C_n^4}+2\tau_n\log\left(\frac{1331eC_n^2}{\delta}\right)+\tau_n\log\left(d\alpha_n\right)\right)\\
&\le 2\mu_n\beta_{a_n}+8\exp\left(-\frac{\mu_n}{C_n^4}\left[\frac{\delta^2}{2048}-\frac{2\tau_nC_n^4}{\mu_n}\log\left(\frac{1331eC_n^2}{\delta}\right)-\frac{\tau_nC_n^4}{\mu_n}\log\left(d\alpha_n\right)\right]\right).
\end{aligned}$$

For n large enough,

$$\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G}_n(\Theta)\\ \|g\|\le C_n}}\left|\frac{1}{\alpha_n}\sum_{i=1}^{\alpha_n}|g(X_i)-Y_i|^2-E|g(X)-Y|^2\right|>\delta\right)\le 2\mu_n\beta_{a_n}+8\exp\left(-\frac{\mu_n}{C_n^4}\eta_{\delta,n}\right)$$

with

$$\begin{aligned}
\eta_{\delta,n}&=\frac{\delta^2}{2048}-\frac{8\sigma^4\tau_n\log\left(\alpha_n\right)^8\log\left(\frac{2662e\sigma^2\log(\alpha_n)^4}{\delta}\right)}{\mu_n}-\frac{4\sigma^4\tau_n\log\left(\alpha_n\right)^8\log\left(d\alpha_n\right)}{\mu_n}\\
&\le\frac{\delta^2}{2048}-\frac{8\sigma^4\tau_n\log\left(\alpha_n\right)^8\log\left(\frac{2662e\sigma^2\log(\alpha_n)^4}{\delta}\right)}{\mu_n}-\frac{4\sigma^4\tau_n\log\left(d\alpha_n\right)^9}{\mu_n}.
\end{aligned}$$

We can now show that equation (3.1c) holds:

$$\lim_{n\to\infty}\mathbb{E}\left\{\sup_{g\in T_{C_n}\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|g(X_i)-Y_{i,L}|^2-\mathbb{E}\left[g(X)-Y_L\right]^2\right|\right\}=0\,\forall L>0.$$

We denote

$$I=\sup_{g\in T_{C_n}\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|g(X_i)-Y_{i,L}|^2-\mathbb{E}\left[g(X)-Y_L\right]^2\right|.$$

and observe that

$$I\le 2(C_n+L)^2.$$

Thus for n large enough

$$\begin{aligned}\mathbb{E}\{I\} &\leq \mathbb{E}\{I\mathbb{1}_{I>\delta} + I\mathbb{1}_{I\leq\delta}\}\\ &\leq \delta + 2(C_n+L)^2\left(2\mu_n\beta_{a_n} + 8\exp\left(-\frac{\mu_n}{C_n^4}\eta_{\delta,n}\right)\right)\\ &= \delta + 16(C_n+L)^2\exp\left(-\frac{\mu_n}{C_n^4}\eta_{\delta,n}\right) + 4(C_n+L)^2\mu_n\beta_{a_n}.\end{aligned}$$

Hence with the $\beta-$mixing condition

$$\lim_{n\to\infty}\mathbb{E}\left\{\sup_{g\in T_{C_n}\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}|g(X_i)-Y_{i,L}|^2 - \mathbb{E}\left[g(X)-Y_L\right]^2\right|\right\} = 0\,\forall L>0.$$

Thus, according to Proposition 3.3,

$$\lim_{n\to\infty}\mathbb{E}\left(T_{C_n}\hat{f}_n(X,\Theta) - f(X)\right)^2 = 0.$$

We only need to check if the non-truncated random forest estimate is consistent, this step is identical to [22]. □

A.2 Proofs for centred forests

Proof of the variance rate, Proposition 5.1. We follow the proof given in [2]. Since the training sample is not independent, we cannot get the same lines and results but the *main* ideas are, associated with Lemma A.1, the same.

Remember that the random forest estimator is written

$$\hat{f}_n(X,\mathcal{D}_n) = \mathbb{E}_\Theta\left[\hat{f}_n(X,\Theta,\mathcal{D}_n)\right]$$

with

$$\hat{f}_n(X,\Theta,\mathcal{D}_n) = \sum_{i=1}^{n} W_{n,i}(X,\Theta)Y_i$$

Thus, omitting the dependence in $\mathcal{D}_n$, the random forest estimator can be written

$$\hat{f}_n(X) = \sum_{i=1}^{n}\mathbb{E}_\Theta\left[W_{n,i}(X,\Theta)\right]Y_i.$$

We can now begin the computation,

$$\begin{aligned}\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 &= \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}_\Theta\left[W_{n,i}(X,\Theta)\right](Y_i - f(X_i))\right]^2\\ &= \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}_\Theta^2\left[W_{n,i}(X,\Theta)\right](Y_i - f(X_i))^2\right]\\ &\quad + \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}_\Theta\left[W_{n,i}(X,\Theta)W_{n,j}(X,\Theta)\right]\epsilon_i\epsilon_j\right].\end{aligned} \tag{A.1}$$

The second term of equation (A.1) is equal to zero since the errors $(\epsilon_i)_{1\leq n}$ are independent by hypothesis **H2b**.

We next analyse the first term. We can upper-bound

$$\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]\epsilon_i^2\right] \leq \sigma^2\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]\right] \text{ (by hypothesis on the variance of the errors } \mathbf{H2b}).$$

The next step is to analyse the expectation of $W_{n,i}$. Since the data is not independent we cannot do exactly the same as in [2]. We need to rewrite the sum over n, decompose it in blocks and then use Lemma A.1. We can then use a similar argument as [2] which is, by introducing another random variable, to reveal a random binomial variable in the denominator. Let us first decompose the previous term

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]\right] &= \mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in H_j}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]\right] + \mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in T_j}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]\right]\\ &= \mathbb{E}\left[u(X_{a_n}^{H})\right] + \mathbb{E}\left[u(X_{a_n}^{T})\right]\end{aligned}$$

where

$$u(X_{a_n}^{B}) = \sum_{j=1}^{\mu_n}\sum_{i\in B_j}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]$$

for $B = H$ or T. We easily observe that $u \leq 1$ by definition of $W_{n,i}$.

Let us begin with the first part of the right hand:

$$\mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in H_j}\mathbb{E}_{\Theta}^{2}\left[W_{n,i}(X,\Theta)\right]\right] \leq \mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in H_j}\mathbb{E}_{\Theta}^{2}\left[\widetilde{W}_{n,i}(X,\Theta)\right]\right] + \mu_n\beta_{a_n}$$

with

$$\widetilde{W}_{n,i}(X,\Theta) = \frac{\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}}{\sum_{k=1}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}}\mathbb{1}_{\widetilde{E}_n(X,\Theta)}$$

where $\left(\xi_i^1\right)_{1\leq i\leq n}$ denotes the first coordinate of the sequence $(\xi_i)_{1\leq i\leq n}$ and

$$\widetilde{E}_n(X,\Theta) = \left\{\sum_{i=1}^{n}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)} \neq 0\right\}.$$

We introduce Θ' independent of Θ but with same distribution,

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in H_j}\mathbb{E}_\Theta^2\left[\widetilde{W}_{n,i}(X,\Theta)\right]\right] &= \sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in H_j}\mathbb{E}_\Theta\left[\widetilde{W}_{n,i}(X,\Theta)\right]\mathbb{E}_{\Theta'}\left[\widetilde{W}_{n,i}(X,\Theta')\right]\right]\\ &= \sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in H_j}\mathbb{E}_{\Theta,\Theta'}\left[\widetilde{W}_{n,i}(X,\Theta)\widetilde{W}_{n,i}(X,\Theta')\right]\right]\\ &= \sum_{j=1}^{\mu_n}\mathbb{E}_{X,\Theta,\Theta'}\left[\sum_{i\in H_j}\frac{\mathbb{1}_{\xi_i\in A_n(X,\Theta)\cap A_n(X,\Theta')}}{\left(\sum_{k=1}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(\sum_{k=1}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\mathbb{1}_{\tilde{E}_n(X,\Theta)}\mathbb{1}_{\tilde{E}_n(X,\Theta')}\right].\end{aligned}$$

For a fixed j,

$$\begin{aligned}&\mathbb{E}_{X,\Theta,\Theta'}\left[\sum_{i\in H_j}\frac{\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)\cap A_n(X,\Theta')}}{\left(\sum_{k=1}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(\sum_{k=1}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\mathbb{1}_{\tilde{E}_n(X,\Theta)}\mathbb{1}_{\tilde{E}_n(X,\Theta')}\right]\\ &\leq \mathbb{E}_{X,\Theta,\Theta'}\left[\sum_{i\in H_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)\cap A_n(X,\Theta')}\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\middle|X,\xi_i^1,\Theta,\Theta'\right]\right].\end{aligned}$$

By independence of the blocks we can remove the conditioning to ξ_i^1,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\middle|X,\xi_i^1,\Theta,\Theta'\right]\\ &\quad=\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\middle|X,\Theta,\Theta'\right].\end{aligned}$$

Using Cauchy-Schwarz's inequality, for a fixed j,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\middle|X,\Theta,\Theta'\right]\\ &\quad\leq\mathbb{E}^{1/2}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)^2}\middle|X,\Theta\right]\\ &\quad\quad\times\mathbb{E}^{1/2}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^n\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)^2}\middle|X,\Theta'\right].\end{aligned}$$

Using the following fact (*cf.*[13]) that

$$\mathbb{E}\left[\frac{1}{1+Bin(N,p)^2}\right]\leq\frac{3}{(N+1)(N+2)p^2}.$$

and since each blocks are independent

$$\mathbb{E}^{1/2}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)^2}\middle|X,\Theta\right]\leq\mathbb{E}^{1/2}\left[\frac{1}{1+\left(\sum_{\tilde{j}=1}^{2\mu_n-1}\mathbb{1}_{\xi_{\tilde{j}}^1\in A_n(X,\Theta)}\right)^2}\middle|X,\Theta\right]$$

where $\tilde{j}$ denotes one component of each block $(H_j)_{1\leq j\leq\mu_n}$ and $(T_j)_{1\leq j\leq\mu_n}$. By independence of the blocks we get

$$\sum_{\tilde{j}=1}^{2\mu_n-1}\mathbb{1}_{\xi_{\tilde{j}\in A_n(X,\Theta)}}\sim Bin\left(2\mu_n-1,\mathbb{P}(X\in A_n(X,\Theta)|X,\Theta)\right).$$

Since we suppose that the law is uniform on $[0,1]^p$ and by the construction of the tree we get

$$\mathbb{P}\left(X\in A_n(X,\Theta)|X,\Theta\right)=2^{-\lceil\log_2\tau_n\rceil}.$$

The same is done for the conditional expectation with respect to (X,Θ'). Thus

$$\begin{aligned}
&\mathbb{E}_{X,\Theta,\Theta'}\left[\sum_{i\in H_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)\cap A_n(X,\Theta')}\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)\left(1+\sum_{k=1,k\notin H_j}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta')}\right)}\middle|X,\xi_i^1,\Theta,\Theta'\right]\right]\\
&\quad\leq\frac{3\times2^{2\lceil\log_2\tau_n\rceil}}{4\mu_n^2}\mathbb{E}_{X,\Theta,\Theta'}\left[\sum_{i\in H_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)\cap A_n(X,\Theta')}\right]\\
&\quad\leq\frac{12\tau_n^2}{4\mu_n^2}\mathbb{E}_{X,\Theta,\Theta'}\left[\sum_{i\in H_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)\cap A_n(X,\Theta')}\right]\\
&\quad\leq\frac{3\tau_n^2}{\mu_n^2}a_n\mathbb{P}\left(\xi_1^1\in\widetilde{A}_n\left(X,\Theta\right)\cap\widetilde{A}_n\left(X,\Theta'\right)\right).
\end{aligned}$$

The last inequality using the fact that even though dependent, they have the same distribution.

The rest is the same as in [2]. After the computations over H, we get

$$\mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in H_j}\mathbb{E}_\Theta^2\left[\widetilde{W}_{n,i}(X,\Theta)\right]\right]\leq\widetilde{C}\left(\frac{S^2}{S-1}\right)^{S/2p}(1+\nu_n)\frac{\tau_n a_n\mu_n}{\mu_n^2(\log\tau_n)^{S/2p}}$$

with

$$\widetilde{C}=\frac{144}{\pi}\left(\frac{\pi\log2}{16}\right)^{S/2p}$$

and

$$1+\nu_n=\prod_{j\in\mathcal{S}}\left[(1+\nu_{n,j})^{-1}\left(1-\frac{\nu_{n,j}}{S-1}\right)^{-1}\right]^{1/2p}.$$

We do the same over T.

Combining both analyses we have

$$\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 \leq 2\widetilde{C}\sigma^2 \left(\frac{S^2}{S-1}\right)^{S/2p} (1+\nu_n)\frac{\tau_n a_n}{\mu_n (\log \tau_n)^{S/2p}} + 2\sigma^2 \beta_{a_n}\mu_n.$$

By construction of the blocs $\mu_n = \frac{n}{2a_n}$, plugging in the previous expression we have

$$\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 \leq C\sigma^2 \left(\frac{S^2}{S-1}\right)^{S/2p} (1+\nu_n)\frac{\tau_n a_n^2}{n (\log \tau_n)^{S/2p}} + \frac{\sigma^2 \beta_{a_n} n}{a_n}$$

with

$$C = \frac{576}{\pi}\left(\frac{\pi \log 2}{16}\right)^{S/2p}$$

and

$$1+\nu_n = \prod_{j\in\mathcal{S}} \left[(1+\nu_{n,j})^{-1}\left(1 - \frac{\nu_{n,j}}{S-1}\right)^{-1}\right]^{1/2p}.$$

□

Proof of the bias term, Proposition 5.2. The start of the proof is the same as in [2] since it does not use the hypothesis of independence between the observations:

$$\begin{aligned}
\mathbb{E}\left[\tilde{f}_n(X) - f(X)\right]^2 &\leq \mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X,\Theta)\left(f(X_i) - f(X)\right)\right]^2 + \sup_{x\in[0,1]^d} f^2(x)\mathbb{P}\left(E_n^c(X,\Theta)\right) \\
&\leq \mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X,\Theta)\left(f^*(X_{i,\mathcal{S}}) - f^*(X_{\mathcal{S}})\right)^2\right] + \sup_{x\in[0,1]^d} f^2(x)\mathbb{P}\left(E_n^c(X,\Theta)\right) \ (cf.\ [2]) \\
&\leq L^2\mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X,\Theta)\|X_i - X\|_{\mathcal{S}}^2\right] + \sup_{x\in[0,1]^d} f^2(x)\mathbb{P}\left(E_n^c(X,\Theta)\right)
\end{aligned}$$

where we get the last inequality using the hypothesis that f^* is $L-$Lipschitz. To go further in the analysis we have to use Lemma A.1 to get independent variables. We proceed similarly to the first proof,

$$\mathbb{E}\left[\sum_{i=1}^n W_{n,i}(X,\Theta)\|X_i - X\|_{\mathcal{S}}^2\right] = \mathbb{E}\left[v\left(X_{a_n}^H\right)\right] + \mathbb{E}\left[v\left(X_{a_n}^T\right)\right]$$

with

$$v\left(X_{a_n}^B\right) = \sum_{j=1}^{\mu_n}\sum_{i\in H_j} W_{n,i}(X,\Theta)\|X_i - X\|_{\mathcal{S}}^2$$

for $B = H$ or T. We observe that

$$v \leq \sup_{(x,y)\in[0,1]^{\mathcal{S}}\times[0,1]^{\mathcal{S}}} \|x-y\|_{\mathcal{S}}^2 \leq S.$$

Thus, using Lemma A.1,

$$\mathbb{E}\left[v\left(X_{a_n}^H\right)\right] \leq \mathbb{E}\left[\sum_{j=1}^{\mu_n}\sum_{i\in H_j}\widetilde{W}_{n,i}(X,\Theta)\|\xi_i^1 - X\|_{\mathcal{S}}^2\right] + S\mu_n\beta_{a_n}.$$

We do the same over T.

We need to do a similar operation to compute the probability $\mathbb{P}\left(E_n^c(X,\Theta)\right)$. We recall that $E_n := \left\{\sum_{i=1}^n \mathbb{1}_{X_i\in A_n(X,\Theta)} \neq 0\right\}$:

$$\begin{aligned}\mathbb{P}\left(E_n^c(X,\Theta)\right) &= \mathbb{E}\left[\mathbb{1}_{\sum_{i=1}^n \mathbb{1}_{X_i\in A_n(X,\Theta)}=0}\right]\\ &= \mathbb{E}\left[\mathbb{1}_{X_1\notin A_n(X,\Theta)}\cdots\mathbb{1}_{X_n\notin A_n(X,\Theta)}\right]\\ &\leq E\left[w(X_{a_n}^H)\right]\end{aligned}$$

where

$$w(X_{a_n}^H) = \prod_{j=1}^{\mu_n}\prod_{i\in H_j}\mathbb{1}_{X_i\notin A_n(X,\Theta)} \Rightarrow w \leq 1.$$

Using Lemma A.1,

$$\mathbb{E}\left[w(X_{a_n}^H)\right] \leq \mathbb{P}\left[\forall\, 1\leq j\leq \mu_n, \forall i\in H_j, \xi_i^1\notin A_n(X,\Theta)\right] + \mu_n\beta_{a_n}.$$

We get

$$\begin{aligned}\mathbb{E}\left[\tilde{f}_n(X) - f(X)\right]^2 \leq L^2\mathbb{E}\left[\sum_{i=1}^n \widetilde{W}_{n,i}(X,\Theta)\|\xi_i^1 - X\|_{\mathcal{S}}^2\right] &+ \sup_{x\in[0,1]^p} f^2(x)\,\mathbb{P}\left[\forall\, 1\leq j\leq\mu_n, \forall i\in H_j, \xi_i^1\notin A_n(X,\Theta)\right]\\ &+ \mu_n\beta_{a_n}\left[2SL^2 + \sup_{x\in[0,1]^p} f^2(x)\right]. \qquad \text{(A.2)}\end{aligned}$$

We first analyse the term $\mathbb{P}\left[\forall\, 1\leq j\leq\mu_n, \forall i\in H_j, \xi_i^1\notin A_n(X,\Theta)\right]$,

$$\mathbb{P}\left[\forall\, 1\leq j\leq\mu_n, \forall i\in H_j, \xi_i^1\notin A_n(X,\Theta)\right] \leq \mathbb{P}\left[\forall\, 1\leq j\leq\mu_n,\ \text{pick } \tilde{i}\in H_j, \xi_{\tilde{i}}\notin A_n(X,\Theta)\right]$$

where $\tilde{i}$ is an arbitrary index chosen in $\{1,\ldots,a_n\}$. Since the blocks are independent, the terms in the probability are independent. Furthermore, they have the same distribution. Thus

$$\begin{aligned}\mathbb{P}\left[\forall 1\leq j\leq\mu_n, \forall i\in H_j, \xi_i^1\notin A_n(X,\Theta)\right] &\leq \mathbb{P}^{\mu_n}\left[\xi_1\notin A_n(X,\Theta)\right]\\ &= \left(1 - 2^{-\lceil\log_2\tau_n\rceil}\right)^{\mu_n} \text{ (by construction of the tree)}\\ &\leq \exp\left(-\frac{\mu_n}{2\tau_n}\right).\end{aligned}$$

Plugging in equation (A.2) we have

$$\mathbb{E}\left[\tilde{f}_n(X) - f(X)\right]^2 \leq L^2\mathbb{E}\left[\sum_{i=1}^n \widetilde{W}_{n,i}(X,\Theta)\|\xi_i^1 - X\|_{\mathcal{S}}^2\right] + \exp\left(-\frac{\mu_n}{2\tau_n}\right)\sup_{x\in[0,1]^d} f^2(x) + \mu_n\beta_{a_n}\left[2SL^2 + \sup_{x\in[0,1]^d} f^2(x)\right].$$

Let us analyse the first term:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n}\widetilde{W}_{n,i}(X,\Theta)\|\xi_i^1-X\|_{\mathcal{S}}^2\right] &= \sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in H_j}\frac{\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}}{\widetilde{N}_n(X,\Theta)}\mathbb{1}_{\widetilde{E}_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\right]\\
&\quad+\sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in T_j}\frac{\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}}{\widetilde{N}_n(X,\Theta)}\mathbb{1}_{\widetilde{E}_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\right]\\
&\leq \sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in H_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\,\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)}\middle|X,\xi_i^1,\Theta\right]\right]\\
&\quad+\sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in T_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\,\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin T_j}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)}\middle|X,\xi_i^1,\Theta\right]\right].
\end{aligned}$$

For a fixed j

$$\mathbb{E}\left[\frac{1}{\left(1+\sum_{k=1,k\notin H_j}^{n}\mathbb{1}_{\xi_k^1\in A_n(X,\Theta)}\right)}\middle|X,\xi_i^1,\Theta\right]\leq\mathbb{E}\left[\frac{1}{\left(1+\sum_{\tilde{k}=1}^{2\mu_n-1}\mathbb{1}_{\xi_{\tilde{k}}^1\in A_n(X,\Theta)}\right)}\middle|X,\Theta\right]$$

where $\tilde{k}$ denotes one component of each block $(H_j)_{1\leq j\leq\mu_n}$ and $(T_j)_{1\leq j\leq\mu_n}$. By independence of the blocks we have

$$\sum_{\tilde{k}=1}^{2\mu_n-1}\mathbb{1}_{\xi_{\tilde{k}}\in A_n(X,\Theta)}\sim Bin(2\mu_n-1,2^{-\lceil\log_2\tau_n\rceil})$$

using the same argument as in the proof "convergence rate for the variance". The following inequality (*cf.*[13]),

$$\mathbb{E}\left[\frac{1}{1+Bin(N,p)}\right]\leq\frac{1}{(N+1)p},$$

gives

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n}\widetilde{W}_{n,i}(X,\Theta)\|\xi_i^1-X\|_{\mathcal{S}}^2\right] &\leq \sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in H_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\frac{2^{\lceil\log_2\tau_n\rceil}}{2\mu_n}\right]\\
&\quad+\sum_{j=1}^{\mu_n}\mathbb{E}\left[\sum_{i\in T_j}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\frac{2^{\lceil\log_2\tau_n\rceil}}{2\mu_n}\right]\\
&\leq \tau_n\mathbb{E}\left[\sum_{i\in H_1}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\right]+\tau_n\mathbb{E}\left[\sum_{i\in T_1}\mathbb{1}_{\xi_i^1\in A_n(X,\Theta)}\|\xi_i^1-X\|_{\mathcal{S}}^2\right]\\
&\leq 2a_n\tau_n\mathbb{E}\left[\mathbb{1}_{\xi_1^1\in A_n(X,\Theta)}\|\xi_1^1-X\|_{\mathcal{S}}^2\right]\text{ by stationarity.}
\end{aligned}$$

The rest is the same as in [2]. We get

$$\mathbb{E}\left[\sum_{i=1}^{n}\widetilde{W}_{n,i}(X,\Theta)\|\xi_i^1 - X\|_{\mathcal{S}}^2\right] \leq \frac{2a_nS}{\tau_n^{\frac{0.75}{S\log 2}(1+\gamma_n)}}$$

with $\gamma_n = \min_j \nu_{n,j}$. We conclude

$$E\left[\tilde{f}_n(X) - f(X)\right]^2 \leq \frac{2SL^2a_n}{\tau_n^{\frac{0.75}{S\log 2}(1+\gamma_n)}} + \exp\left(-\frac{\mu_n}{2\tau_n}\right)\sup_{x\in[0,1]^d} f^2(x) + \mu_n\beta_{a_n}\left[2SL^2 + \sup_{x\in[0,1]^d} f^2(x)\right].$$

Replacing using that $\mu_n = \frac{n}{2a_n}$ we have

$$E\left[\tilde{f}_n(X) - f(X)\right]^2 \leq \frac{2SL^2a_n}{\tau_n^{\frac{0.75}{S\log 2}(1+\gamma_n)}} + \exp\left(-\frac{n}{4a_n\tau_n}\right)\sup_{x\in[0,1]^d} f^2(x) + \frac{\beta_{a_n}n\left[SL^2 + \sup_{x\in[0,1]^d} f^2(x)\right]}{a_n}.$$

□

Appendix B. Tool to establish consistency in stationary ergodic case

We first introduce the general consistency theorem as known from [13] and used in [22]. From now on μ denotes the distribution of X.

Theorem B.1. *Let $\mathcal{D}_n = \{(X_1, Y_1), \dots (X_n, Y_n)\}$ i.i.d. Let $\mathcal{G}_n = \mathcal{G}(\mathcal{D}_n)$ be a class of functions $g : \mathcal{X} \to \mathcal{Y}$,the estimator $\hat{f}_n$ which minimises the empirical L^2 risk on $\mathcal{G}_n$ and f the regression function. If*

$$\lim_{n\to\infty} C_n = \infty,$$

$$\lim_{n\to\infty} \mathbb{E}\left\{\inf_{g\in\mathcal{G}_n, \|g\|_\infty \leq C_n} \int |g(x) - f(x)|^2 \mu(\mathrm{d}x)\right\} = 0,$$

$$\lim_{n\to\infty} \mathbb{E}\left\{\sup_{g\in T_{C_n}\mathcal{G}_n} \left|\frac{1}{n}\sum_{i=1}^{n} |g(X_i) - Y_{i,L}|^2 - \mathbb{E}\left[g(X) - Y_L\right]^2\right|\right\} = 0 \ \forall L > 0$$

then

$$\lim_{n\to\infty} \mathbb{E}\left\{\int |\hat{f}_n(x) - f(x)|^2 \mu(\mathrm{d}x)\right\} = 0.$$

We extend this theorem to dependent process. The only assumption we actually need is that the stochastic process is stationary and ergodic.

Proposition B.2. *Let $(X_t, Y_t)_{t\in\mathbb{Z}}$ be a stationary ergodic process and a data set $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Let $\mathcal{G}_n = \mathcal{G}(\mathcal{D}_n)$ be a class of functions $g : \mathcal{X} \to \mathcal{Y}$, the estimator $\hat{f}_n$ which minimises the empirical L^2 risk on $\mathcal{G}_n$ and f the regression function. Under equations (3.1a–3.1c),*

$$\lim_{n\to\infty} \mathbb{E}\left\{\int |\hat{f}_n(x) - f(x)|^2 \mu(\mathrm{d}x)\right\} = 0.$$

Proof. To prove this result, we follow the same line as in [13]. Instead of using the large of law numbers for i.i.d variables we use the law of large numbers for stationary ergodic processes.

We write

$$\begin{aligned}\int_{\mathbb{R}^p} |\hat{f}_n(x) - f(x)|^2 \mu(\mathrm{d}x) &= \mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right] - \mathbb{E}|f(X) - Y|^2 \\ &= \left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right) \\ &\quad \times \left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} + \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right) \\ &= \left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right)^2 \\ &\quad + 2\left(\mathbb{E}|f(X) - Y|^2\right)^{1/2} \left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right)^2.\end{aligned}$$

It suffices to show

$$\mathbb{E}\left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right)^2 \underset{n\to\infty}{\longrightarrow} 0.$$

We rewrite this term,

$$\begin{aligned}&\mathbb{E}\left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right)^2 \\ &\leq 2\mathbb{E}\left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \inf_{g\in\mathcal{G}_n, \|g\|\leq C_n} \left(\mathbb{E}|g(X) - Y|^2\right)^{1/2}\right)^2 \\ &\quad + 2\mathbb{E}\left(\inf_{g\in\mathcal{G}_n, \|g\|_\infty\leq C_n} \left(\mathbb{E}|g(X) - Y|^2\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right)^2.\end{aligned}$$

The last term can be bounded using the reverse triangle inequality,

$$\begin{aligned}&2\mathbb{E}\left(\inf_{g\in\mathcal{G}_n, \|g\|_\infty\leq C_n} \left(\mathbb{E}|g(X) - Y|^2\right)^{1/2} - \left(\mathbb{E}|f(X) - Y|^2\right)^{1/2}\right)^2 \\ &\leq 2\mathbb{E}\left(\inf_{g\in\mathcal{G}_n, \|g\|_\infty\leq C_n} \left(\mathbb{E}|g(X) - f(X)|^2\right)^{1/2}\right)^2 \\ &\leq 2\mathbb{E}\left(\inf_{g\in\mathcal{G}_n, \|g\|_\infty\leq C_n} \mathbb{E}|g(X) - f(X)|^2\right) \underset{n\to\infty}{\longrightarrow} 0 \text{ by equation (3.1b).}\end{aligned}$$

It remains to show that

$$2\mathbb{E}\left(\left(\mathbb{E}\left[|\hat{f}_n(X) - Y|^2 \middle| \mathcal{D}_n\right]\right)^{1/2} - \inf_{g\in\mathcal{G}_n, \|g\|_\infty\leq C_n} \left(\mathbb{E}|g(X) - Y|^2\right)^{1/2}\right)^2 \underset{n\to\infty}{\longrightarrow} 0.$$

We can lower bound this term by

$$-2\mathbb{E}\left[\inf_{g\in\mathcal{G}_n,\|g\|_\infty\leq C_n}\left(\int_{\mathbb{R}^p}|g(x)-f(x)|^2\mu(\,\mathrm{d}x)\right)^{1/2}\right]^2$$

and upper bound it by

$$\mathbb{E}\Bigg(2\left(\mathbb{E}|Y-Y_L|^2\right)^{1/2}+2\left(\frac{1}{n}\sum_{j=1}^n|Y_i-Y_{i,L}|^2\right)^{1/2}$$
$$+2\sup_{g\in T_{C_n}\mathcal{G}_n}\left|\left(\frac{1}{n}\sum_{j=1}^n|g(X_i)-Y_{i,L}|^2\right)^{1/2}-\left(\mathbb{E}|g(X)-Y|^2\right)^{1/2}\right|\Bigg)^2. \tag{B.2}$$

Using the inequality: $(a+b+c)^2\leq 3a^2+3b^2+3c^2\ \forall(a,b,c)\in\mathbb{R}^3$ and $(\sqrt{a}-\sqrt{b})^2\leq|a-b|$ we have

$$\begin{aligned}\textit{equation (B.2)}&\leq 6\mathbb{E}\left[\sup_{g\in T_{C_n}\mathcal{G}_n}\left|\frac{1}{n}\sum_{j=1}^n|g(X_i)-Y_{i,L}|^2-\mathbb{E}|g(X)-Y|^2\right|\right]\\&\quad+6\mathbb{E}|Y-Y_L|^2+6\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^n|Y_i-Y_{i,L}|^2\right)\\&\underset{n\to\infty}{\longrightarrow}12\mathbb{E}|Y-Y_L|^2.\end{aligned}$$

The last line uses equations (3.1a–3.1c) and the strong law for stationary ergodic process.

We get the result letting $L\to\infty$.

□

References

[1] H.C.P. Berbee, Random walks with stationary increments and renewal theory. *MC Tracts* **112** (1979) 1–223.
[2] G. Biau, Analysis of a random forests model. *J. Mach. Learn. Res.* **13** (2012) 1063–1095.
[3] G. Biau and E. Scornet, A random forest guided tour. *TEST* **25** (2016) 197–227.
[4] R.C. Bradley, Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surv.* **2** (2005) 107–144.
[5] L. Breiman, Bagging predictors. *Mach. Learn.* **24** (1996) 123–140.
[6] L. Breiman, Random forests. *Mach. Learn.* **45** (2001) 5–32.
[7] L. Breiman, Consistency for a simple model of random forests. Technical report (2004).
[8] L. Breiman, J. Friedman, C.J. Stone and R.A. Olshen, Classification and Regression Trees. *The Wadsworth and Brooks-Cole statistics-probability series*. Taylor & Francis, Oxford (1984).
[9] D.R. Cutler, T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson and J.J. Lawler, Random forests for classification in ecology. *Ecology* **88** (2007) 2783–2792.
[10] J. Dedecker, P. Doukhan, G. Lang, L.R.J. Rafael, S. Louhichi and C. Prieur, Weak dependence, in *Weak Dependence: With Examples and Applications*. Springer, Berlin (2007) 9–20.
[11] G. Dudek, Short-term load forecasting using random forests, in *Intelligent Systems'2014*. Springer International Publishing, Cham (2015) 821–828.
[12] A. Fischer, L. Montuelle, M. Mougeot and D. Picard, Statistical learning for wind power: A modeling and stability study towards forecasting. *Wind Energy* **20** (2017) 2037–2047.
[13] L. Györfi, M. Kohler, A. Krzyzak and H. Walk, A distribution-free theory of nonparametric regression. Springer Science & Business Media, Berlin (2006).
[14] M.J. Kane, N. Price, M. Scotch and P. Rabinowitz, Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinform.* **15** (2014) 276.

[15] A. Lahouar and J. Ben Hadj Slama, Random forests model for one day ahead load forecasting, in *IREC2015 The Sixth International Renewable Energy Congress* (2015) 1–6.
[16] A.C. Lozano, S.R. Kulkarni and R.E. Schapire, Convergence and consistency of regularized boosting with weakly dependent observations. *IEEE Trans. Inf. Theory* **60** (2014) 651–660.
[17] R. Meir, Nonparametric time series prediction through adaptive model selection. *Mach. Learn.* **39** (2000) 5–34.
[18] L. Mentch and G. Hooker, Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** (2016) 1–41.
[19] A.M. Prasad, L.R. Iverson and A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9** (2006) 181–199.
[20] E. Rio, Inequalities and limit theorems for weakly dependent sequences. Lecture (2013).
[21] E. Scornet, On the asymptotics of random forests. *J. Multivar. Anal.* **146** (2016) 72–83.
[22] E. Scornet, G. Biau and J.-P. Vert, Consistency of random forests. *Ann. Stat.* **43** (2015) 1716–1741.
[23] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook and R. Moore, Real-time human pose recognition in parts from single depth images. *Commun. ACM* **56** (2013) 116–124.
[24] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan and B.P. Feuston, Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Comput. Sci.* **43** (2003) 1947–1958.
[25] S. Wager and S. Athey, Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113** (2018) 1228–1242.
[26] B. Yu, Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Prob.* **22** (1994) 94–116.