Random forests for time-dependent processes

Benjamin Goehry

RANDOM FORESTS FOR TIME-DEPENDENT PROCESSES

Benjamin Goehry [1]

Abstract. Random forests were introduced by Breiman in 2001. We are interested in the theoretical study of both the random forest-random input and a simplified version of the random forest: the centred random forest. Let $((X_1, Y_1), \dots, (X_n, Y_n))$ be random variables. Under the independent and identically distributed hypothesis, Biau studied the simplified version and got rate of convergence in the sparse case. Biau, Scornet and Vert proved the consistency of the original algorithm when the regression model follows an additive model and that $X \sim \text{Unif}(0,1)^p$. However we are commonly faced to applications where the i.i.d hypothesis is not satisfied for example when dealing with time series. We extend the previous results to the case where observations are weakly dependent.

.

1. Introduction

Random forests were introduced in 2001 by Breiman in [4] 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 parameters tuning. We can cite as successful applications: chemo-informatics [19], ecology [6, 14], 3D object recognition [17], time series prediction [8, 10, 13]. Often used as benchmark because of efficient and fast results, random forests have become a must-have tool.

Basically a random forest is a collection of random trees which are constructed independently of each others. In order to construct a random forest we then have only to explain how to grow one random tree. One tree is constructed recursively based on some criterion.

There are different ways to introduce this randomness. In the variant of random forests which is the most commonly used in practice: random forest-random input (RF-RI), the first step is to choose α_n points among the n points we have, with or without replacement. On these α_n selected points we then construct a tree using the CART criterion: at each node we look for the best split (the direction and the location on that direction) which minimises the variance but instead of considering the criterion on all the directions, we restrict the minimisation at each node to a random subset of size m_{try}.

Suppose we have a stationary random sequence $(X_t, Y_t)_{t\in\mathbb{Z}} \in [0,1]^p \times \mathbb{R}$ such that

$$Y_t = f(X_t) + \epsilon_t$$

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

$$\forall x \in [0,1]^p, f(x) = \mathbb{E}[Y_t | X_t = x].$$

In the statistical context we only observe a training sample $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$.

The RF-RI is highly difficult to analyse due to the high-dependency between data and construction. The only theoretical result we are aware of is the establishment of consistency when the $(X_i, Y_i)_{1\le i\le n}$ are i.i.d by [16] *i.e.* $\mathbb{E}[f_n(X) - f(X)]^2 \to 0$

[1] LMO, Univ. Paris Sud, Orsay, France
e-mail: benjamin.goehry@math.u-psud.fr

as $n \to +\infty$ where $\hat{f}_n$ denotes the RF-RI estimator. Note that this result relies on heavy hypotheses: it is true for trees where points are selected without replacement and the regression function is an additive model where $X \sim \mathcal{U}(0,1)^p$.

Later on, many variants have been considered which are easier to analyse from a theoretical point of view: the *purely random forests*'s family. The RF-RI is based on the CART criterion which is heavily data-dependent. The purely random forests are based on criteria which are independent of the data. The variant we will analyse later on is called *centred forest* which was introduced by [5]. The first difference is that there is no re-sampling step. We then construct recursively a tree 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. Under the hypothesis that $(X_i, Y_i)_{1\leq i\leq n}$ are i.i.d [2] proved that this procedure adapts to sparsity by giving a rate of convergence which depends on the number of strong features. This kind variant and others have been preferred for statistical analysis but we shall note that even though simpler they offer new perspectives that seem really interesting in practice as well. We refer to [3] for a complete survey on random forests.

The previous results that we cited for the centred random forest as well as the result of consistency for the RF-RI of [16] are proven 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 time series.

Many algorithms were already studied in the case of weakly dependent observations. The general problem of one-step ahead predicting of time series was considered by [12] 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. The consistency of the SVM algorithm was studied by [18] under $\alpha-$mixing and not necessarily stationary processes. The consistency and the rate of convergence of boosting are studied in [11] when the observations are $\beta-$mixing.

Our contribution is the extension of [2]'s result of rate of convergence as well as the the extension of [16]'s result on the consistency of the RF-RI in the case where the observations are weakly dependent. We will first consider the centred forest, easier to analyse and will ease the introduction of weakly dependent data. We denote $\hat{f}_n$ the regression function estimator. We will compute the convergence rates for this specific random forest model that is, at which rate, depending on n, p and assumptions over the model, does the following holds:

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

This will also prove the consistency under mild hypotheses.

In the case of the RF-RI, the goal is also to prove consistency in the same form as in the simpler case but this time without the computation of the rate of convergence.

The paper is organised as follows: we will first introduce the models studied. We follow with the notion of dependence. We can then present the result we are interested in: the convergence rates and consistency under weak dependence. The proofs are given at the end for ease of readability.

2. Models

We will first introduce the RF-RI and then the simpler model to see the main differences. Before going into the specific algorithms, let us first state some notations and remarks.

A random forest (either RF-RI or simpler models) is a collection of M random trees. We denote for the j-th 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 will be defined later on depending on the variant. The $j-$th tree is defined as follows:

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

where $\mathcal{D}_n\left(\Theta_j\right)$ 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 tree j. The cell containing the point x is denoted $A_n(x, \Theta_j, \mathcal{D}_n)$ and $N_n\left(x; \Theta_j; \mathcal{D}_n\right) = \#\left\{X_i \in A_n(x, \Theta_j, \mathcal{D}_n)\right\}$.

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

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

2.1. **Random forest - random input**

We begin by introducing the variant of random forests which is the most commonly used in practice: random forest-random input (RF-RI). We denote:

- $\alpha_n \in [\![1, \ldots, n]\!]$ the number of sampled data points in each tree;
- $m_{try} \in [\![1, \ldots, p]\!]$ the pre-selected number of directions for splitting;
- $\tau_n \in [\![1, \ldots, \alpha_n]\!]$ the number of leaves in each tree.

The random forest is then computed as detailed in algorithm 1.

input : $((X_1, Y_1), \ldots, (X_n, Y_n))$
parameters: M, α_n, m_{try}
for *1 to M* **do**
  Select $\alpha_n \leq n$ points (with or without replacement);
  Construct a tree :
    - At each node, select a random subset of m_{try} directions
    - Select the best split using CART criterion
end
output : Classification: majority vote.
  Regression: empirical mean of the M trees

Algorithm 1: Random forest - random input

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

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

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).

2.2. **Centred forest**

We will now explicit the construction of centred forest introduced by [5] and more precisely, we will explicit the construction of one random tree by the definition of a random forest.

We need to keep in mind that all nodes of the trees are associated with rectangular cells such that at each step of the construction of the tree, the collection of the cells forms a partition of $[0, 1]^p$. The root of the tree is then $[0, 1]^p$ itself. The centred forest algorithm is detailed in algorithm 2.

Data: $((X_1, Y_1), \dots, (X_n, Y_n))$
initialization : $\tau_n \geq 2$;
while $i \leq \lceil \log_2 \tau_n \rceil$ *times with* $\tau_n \geq 2$ **do**
  At each node, a coordinate $j \in [\![1, \dots, d]\!]$ is chosen with probability $p_{n,j} \in (0,1)$ such that $\sum_{j=1}^{d} p_{n,j} = 1$;
  The split is done in the middle of the chosen side once the direction chosen.
end

Algorithm 2: Centred random forest

We note that $\tau_n \geq 2$ is a fixed deterministic parameter which may depend on n and that each tree has exactly $2^{\lceil \log_2 \tau_n \rceil} \approx \tau_n$ nodes.

Each random trees then output the average over all Y_i for which the corresponding X_i fall into the same cell of the random partition. Adapting the previous notations we can write the random tree as,

$$\hat{f}_{tree,n}(X, \Theta_j, \mathcal{D}_n) = \sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in A_n(X,\Theta)} Y_i}{N_n(X,\Theta)} \mathbb{1}_{E_n(X,\Theta)}$$

where $E_n(X,\Theta)$ the event defined by $\{N_n(X,\Theta) \neq 0\}$. We will use the following notation in the case of centred random forest,

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

To ease the analysis we define the random forest regression estimate by taking the expectation over Θ:

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

and will omit the dependency on $\mathcal{D}_n$ and denote $\hat{f}_n(X) := \hat{f}_n(X, \mathcal{D}_n)$.

3. Dependency

We recall the notion of weak dependence. We refer to [15] and [7] for more details about dependent processes. In this paper we will only consider the notion of $\beta-$mixing. Let $(W_t)_{t\in\mathbb{Z}} := (X_t, Y_t)_{t\in\mathbb{Z}}$.

Definition 3.1 (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 coefficient is given by

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

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 coefficient are known as the algebraic and exponential mixing defined as follows,

$$\mathbf{H(m)} \begin{cases} \text{Algebraic mixing: } \beta_m = O(m^{-r}) \text{ for } r > 0 \\ \text{Exponential mixing: } \beta_m = O(\exp(-bm^k)) \text{ for } b, k > 0. \end{cases}$$

Notice that the exponential mixing hypothesis is stronger than algebraic mixing.

The $\beta-$mixing processes arise in many examples as Markov chains. They are also interesting in the theoretical setting since we can almost work on them as if we were in an independent setting using the following lemma from [20] which link the dependent process to an independent process up to a term linear in β.

Remark: We explicit the construction of [20] but we note that [1] proved a similar coupling lemma in which they worked with a random variable X in some Polish space.

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 remainder terms. We then define for $1 \leq i \leq \mu_n$,

$$H_j = \{i : 2(j-1)a_n + 1 \leq i \leq (2j-1)a_n\}$$

and

$$T_j = \{i : (2j-1)a_n + 1 \leq i \leq 2ja_n\}.$$

We denote

$$W^{(j)} = \left\{W_i, i \in H_j\right\}$$

and

$$W'^{(j)} = \left\{W_i, i \in T_j\right\}.$$

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 $\Xi^{(j)} = \left\{\xi_i, i \in H_j\right\}$ and such that $\forall j \in [\![1, \ldots, n]\!]$,

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

We construct in the same way a sequence of T-blocks. Figure 1 illustrates this construction.

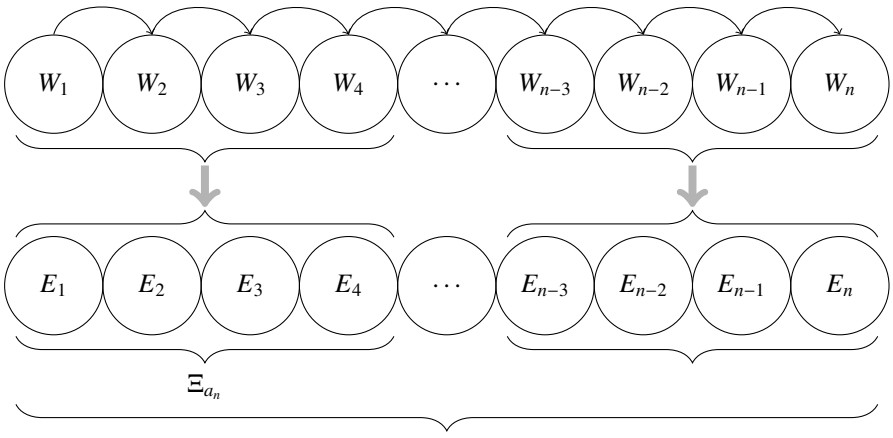

FIGURE 1. Construction of the new independent sequence Ξ.

Lemma 3.2 ([20])**.** *Let the distributions of W_{a_n} and Ξ_{a_n} be Q and $\tilde{Q}$ respectively. Then for any measurable function u on $\mathbb{R}^{a_n\mu_n}$ with bound m,*

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

We first state the result on centred forest to ease the introduction of results with dependency.

4. Results on centred random forest

We first analyse the convergence rates of the random forest model in the β-mixing context and the consistency will follow from it.

We analyse the centred random forest in a sparse framework; in most applications the true dimension is always smaller than p. We will 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,

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

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

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

We also need the following hypotheses to state the results:

- **H(d)**: Given a data set $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ with stationary $\beta-$mixing $(X_i, Y_i) \in [0,1]^p \times \mathbb{R}$;
- **H(e)**: the errors $\epsilon_i := Y_i - f(X_i)$ are bounded such that $\forall i \in [\![1, \ldots, n]\!], |\epsilon_i| \leq M$.

4.1. 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

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

The first result concerns the variance, the second the bias.

Proposition 4.1. *Under the hypotheses of stationary $\beta-$mixing data **H(d)**, bounded errors **H(e)**, assuming that X is uniformly distributed on $[0,1]^p$ and for all $x \in [0,1]^p$,*

$$\sigma^2(x) = \mathbb{V}\left[Y|X = x\right] \leq \sigma^2$$

for some positive constant σ^2. Then, if $p_{n,j} = \frac{1}{S}\left(1 + \nu_{n,j}\right)$ for $j \in \mathcal{S}$,

$$\begin{aligned}\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 \leq & C\sigma^2\left(\frac{S^2}{S-1}\right)^{S/2p}(1+\nu_n)\frac{\tau_n a_n^2}{n(\log \tau_n)^{S/2p}} \\ & + \frac{a_n^3}{2n}M^2 + 2\beta_{a_n}\mu_n\left(\sigma^2 + \left(1 + \frac{\tau_n^2 a_n^3}{4n}\right)M^2\right)\end{aligned}$$

where

$$C = \frac{288}{\pi}\left(\frac{\pi \log 2}{16}\right)^{S/2p}$$

and

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

As noticed in [2], if $a < p_{n,j} < b$ for some constants $a, n \in (0,1)$ we have

$$1 + \nu_n \leq \left(\frac{S-1}{S^2 a (1-b)} \right)^{\frac{S}{2p}}.$$

Proposition 4.2. *Under the hypotheses of stationary $\beta-$mixing data **H(d)**, assuming that X is uniformly distributed on $[0,1]^p$ and f^* is L-Lipschitz on $[0,1]^p$. Then, if $p_{n,j} = \frac{1}{S}\left(1 + \nu_{n,j}\right)$ for $j \in \mathcal{S}$,*

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

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 noticed by [2].

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 4.3. *Under the hypotheses of stationary $\beta-$mixing data **H(d)**, bounded errors **H(e)** and assuming that X is uniformly distributed on $[0,1]^p$ and f^* is L-Lipschitz on $[0,1]^p$. Moreover, for all $x \in [0,1]^p$,*

$$\sigma^2(x) = \mathbb{V}[Y|X = x] \leq \sigma^2$$

for some positive constant σ^2. Then, if $p_{n,j} = \frac{1}{S}\left(1 + \nu_{n,j}\right)$ for $j \in \mathcal{S}$,

$$\mathbb{E}\left[\hat{f}_n(X) - f(X)\right]^2 \leq \frac{2SL^2 a_n}{\tau_n^{\frac{0,75}{S \log 2}(1+\gamma_n)}} + C_{1,n}\frac{a_n^2 \tau_n}{n} + \frac{a_n^3 M^2}{2n} + C_2 \mu_n \beta_{a_n}$$

with

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

$$C_2 = 2\left(SL^2 + \sigma^2 + \left(1 + \frac{\tau_n^2 a_n^3}{4n}\right) M^2\right) + \sup_{x\in[0,1]^p} f^2(x).$$

The first remark we make is about the assumption **H(e)**. If we suppose furthermore that the errors $(\epsilon_i)_{1\leq i\leq n}$ are independent then all the terms with M in the previous rates disappear. This extended hypothesis is generally not true when $(X_i, Y_i)_{1\leq i\leq n}$ are $\beta-$mixed but it is supposed in some theoretical models as AR. 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.

Under the hypothesis of algebraic mixing and thus exponential mixing **H(m)**, the term depending on β is converging to 0 when n tends to infinity.

By the construction of the blocks, $n = 2\mu_n a_n$ where a_n is the number of variables in each block and $2\mu_n$ the total number of blocks. Let us suppose that we are in the independent case hence $\beta_m = 0, \forall m \geq 0$. We take blocks of length $a_n = 1$ and so $\mu_n = \frac{n}{2}$. Plugging these into Propositions 4.1 and 4.2, we get exactly the same upper bound for the variance as [2]. In the case of the bias, we have 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 the necessary pre-processing to use Lemma 3.2.

4.2. **Consistency**

This latter theorem also implies consistency for the centred random forest under mild hypothesis,

Theorem 4.4. *Under the hypotheses of stationary $\beta-$mixing data **H(d)**, bounded errors **H(e)**, algebraic mixing **H(m)**, assuming that X is uniformly distributed on $[0,1]^p$ and f^* is L-Lipschitz on $[0,1]^p$. Moreover, for all $x \in [0,1]^p$,*

$$\sigma^2(x) = \mathbb{V}\left[Y|X=x\right] \leq \sigma^2$$

for some positive constant σ^2. Then, if $p_{n,j} = \frac{1}{S}\left(1+\nu_{n,j}\right)$ for $j \in \mathcal{S}$, for a well chosen τ_n, μ_n, a_n, the centred random forest model is consistent i.e.

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

5. Result for the RF-RI

We introduced the algorithm in the first section and then the notion of weak dependence. To establish consistency of the RF-RI we need a tool we detail in appendix A.

We need additionally the following hypotheses to prove the consistency of the RF-RI.

H(a): the response Y follows

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

where $X = \left(X^{(1)}, \ldots, X^{(p)}\right)$ 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.

Theorem 5.1. *Given a data set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with stationary ergodic β-mixing $(X_i, Y_i) \in [0,1]^p \times \mathbb{R}$ for $i \in [\![1, \ldots n]\!]$.*

If

- ***H(a)** is satisfied;*
- $\frac{\tau_n \log(\alpha_n)^9}{\mu_n} \underset{n\to\infty}{\longrightarrow} 0$;
- $\log(\alpha_n)^4 \mu_n \beta_{a_n} \underset{n\to\infty}{\longrightarrow} 0$.

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 consider the independent case. If the $(X_i, Y_i)_{1\leq i\leq n}$ are independent, $\beta_m = 0, \forall m \geq 0$. Remember that $\alpha_n = 2\mu_n a_n$ where α_n is the number of sampled data points in each tree. Set a_n the length of each block to 1 and thus $\mu_n = \frac{\alpha_n}{2}$. We get exactly the same result as in [16].

The first two conditions are quite the same as in the original theorem. Note that the additivity part in the hypothesis **H(a)** is quite restrictive even though many applications are well simulated with such models. However we need $X \sim \mathcal{U}(0,1)^p$ which is way too restrictive in practice.

The third one is simply saying that the dependence between the data is not too long. We take the example of algebraic mixing. In this case, $\beta_{a_n} \leq C a_n^{-r}$ for some $C, r > 0$. Plugging this in the third hypothesis and using the fact that $\alpha_n = 2\mu_n a_n$ leads to

$$\log(\alpha_n)^4 \mu_n \beta_{a_n} = \frac{\alpha_n \log(\alpha_n)^4}{2a_n^{1+r}}.$$

Now let us suppose for example that $a_n = \alpha_n^{1/2}$ and we get

$$\log(\alpha_n)^4 \mu_n \beta_{a_n} = \frac{\log(\alpha_n)^4}{2\alpha_n^{\frac{r-1}{2}}} \underset{n\to\infty}{\longrightarrow} 0$$

for $r > 1$.

6. Proofs for centred forests

Proof of the variance rate, Proposition 4.1. We follow the proof given by [2]. Since the training sample is not independent, we cannot get the same lines and results but the *main* ideas are, associate with Lemma 3.2, the same.

Remember that the random forest estimator is written

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

with

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

Thus, the random forest estimator can be written

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

We define also $\tilde{f}_n(X)$,

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

We can now begin the computation,

$$\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}_\Theta\left[W_{n,i}(X, \Theta)\right](Y_i - f(X_i))\right]^2 \tag{6.1}$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}_\Theta^2\left[W_{n,i}(X, \Theta)\right](Y_i - f(X_i))^2\right] \tag{6.2}$$

$$+ \mathbb{E}\left[\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathbb{E}_\Theta\left[W_{n,i}(X, \Theta) W_{n,j}(X, \Theta)\right] \epsilon_i \epsilon_j\right]. \tag{6.3}$$

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

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}_\Theta^2\left[W_{n,i}(X, \Theta)\right] \sigma^2(X_i)\right] \leq \sigma^2 \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}_\Theta^2\left[W_{n,i}(X, \Theta)\right]\right] \text{ (by hypothesis on } \sigma(X)).$$

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 [2]. We need to rewrite the sum over n, decompose it in blocks and then use Lemma 3.2. 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:

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

where

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

with $B = H$ 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

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

and

$$\widetilde{E}_n(X,\Theta)=\left\{\sum_{j=1}^{n}\xi_j\in A_n\left(X,\Theta\right)\right\}.$$

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

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

To ease readability, we denote

$$F_{H_j}=\left\{\left\{\xi_i,i\in H_j\right\}\in A_n(X,\Theta)\cap A_n(X,\Theta')\right\}.$$

For a fixed j,

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

Let's denote

$$\text{fraction}\;=\frac{1}{\left(a_n+\sum_{k=1,\xi_k\notin H_j}^{n}\mathbb{1}_{\xi_k\in A_n(X,\Theta)}\right)\left(a_n+\sum_{k=1,\xi_k\notin H_j}^{n}\mathbb{1}_{\xi_k\in A_n(X,\Theta')}\right)}.$$

By independence of the blocks we get,

$$\mathbb{E}\Big[\text{ fraction }|X,\Theta,\Theta',F_{H_j}\Big]=\mathbb{E}\left[\text{ fraction }|X,\Theta,\Theta'\right].$$

Using now, Cauchy-Schwartz's inequality, for a fixed j,

$$\mathbb{E}\left[\text{ fraction }\Big|X,\Theta,\Theta'\right]\leq\mathbb{E}^{1/2}\left[\frac{1}{\left(a_n+\sum_{k=1,\xi_k\notin H_j}^{n}\mathbb{1}_{\xi_k\in A_n(X,\Theta)}\right)^2}\middle|X,\Theta\right]\mathbb{E}^{1/2}\left[\frac{1}{\left(a_n+\sum_{k=1,\xi_k\notin H_j}^{n}\mathbb{1}_{\xi_k\in A_n(X,\Theta')}\right)^2}\middle|X,\Theta'\right].$$

Using the fact (*cf.* [9]) that if Z is a random binomial variable of parameters (N,p) then,

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

Since each blocks are independent and $a_n\geq 1$,

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

where $\tilde{j}$ denotes one component of each block $\left(H_j\right)_{1\leq j\leq\mu_n}$ and $\left(T_j\right)_{1\leq j\leq\mu_n}$. 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,

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

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

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

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

We do the same over T.

We analyse now the second term of eq. (6.3). It is not equal to zero since $(\epsilon_i)_{1\leq n}$ are not independent. We use again the link between dependent and independent process in the $\beta-$mixing case. We rewrite this second term over the H and T as before:

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}_\Theta\left[W_{n,i}(X,\Theta)W_{n,j}(X,\Theta)\right]\epsilon_i\epsilon_j\right] = \mathbb{E}\left[w(X_{a_n}^H)\right] + \mathbb{E}\left[w(X_{a_n}^T)\right]$$

where

$$\mathbb{E}\left[w(X_{a_n}^B)\right] = \mathbb{E}_\epsilon\left[\mathbb{E}_X\left[\sum_{i=1}^{\mu_n}\sum_{j=1}^{\mu_n}\sum_{k\in B_i}\sum_{l\in B_j,l\neq k}\epsilon_i\epsilon_j\mathbb{E}_\Theta[W_{n,i}W_{n,j}]\middle|\epsilon_1,\ldots,\epsilon_n\right]\right]$$

Using the hypothesis **H(e)**, Cauchy-Schwartz and Jensen's inequality it is easy to show that

$$w(X_{a_n}) \leq M^2.$$

Using Lemma 3.2 we then have:

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}_\Theta\left[W_{n,i}(X,\Theta)W_{n,j}(X,\Theta)\right]\epsilon_i\epsilon_j\right] = 2M^2\beta_{a_n}\mu_n + \mathbb{E}\left[\sum_{i=1}^{\mu_n}\sum_{j=1}^{\mu_n}\sum_{k\in B_i}\sum_{l\in B_j,l\neq k}\tilde{\epsilon}_i\tilde{\epsilon}_j\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right].$$

It remains to analyse the last term. Since the blocks are independent of each others we have,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{\mu_n}\sum_{j=1}^{\mu_n}\sum_{k\in B_i}\sum_{l\in B_j,l\neq k}\tilde{\epsilon}_i\tilde{\epsilon}_j\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right] &= \mathbb{E}\left[\sum_{i=1}^{\mu_n}\sum_{k\in B_i}\sum_{l\in B_i,l\neq k}\tilde{\epsilon}_i\tilde{\epsilon}_j\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right] \\
&= \mu_n\mathbb{E}\left[\sum_{k\in B_1}\sum_{l\in B_1,l\neq k}\tilde{\epsilon}_i\tilde{\epsilon}_j\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right] \\
&\leq M^2\mu_n\mathbb{E}\left[\sum_{k\in B_1}\sum_{l\in B_1,l\neq k}\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right].
\end{aligned}$$

Using a similar trick as before we get:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k\in B_1}\sum_{l\in B_1,l\neq k}\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right] &\leq \mathbb{E}\left[\sum_{k\in B_1}\sum_{l\in B_1,l\neq k}\mathbb{1}_{\xi_k\in A_n(X,\Theta)}\mathbb{1}_{\xi_l\in A_n(X,\Theta)}\mathbb{E}\left[\frac{1}{1+\left(\sum_{\tilde{j}=1}^{2\mu_n-1}\mathbb{1}_{\xi_{\tilde{j}\in A_n(X,\Theta)}}\right)^2}\middle|X,\Theta,F_{B_1}\right]\right] \\
&\leq \mathbb{E}\left[\sum_{k\in B_1}\sum_{l\in B_1,l\neq k}\mathbb{1}_{\xi_k\in A_n(X,\Theta)}\mathbb{1}_{\xi_l\in A_n(X,\Theta)}\frac{2^{2\lceil\log_2\tau_n\rceil}}{4\mu_n^2}\right] \\
&\leq \frac{a_n^2 2^{2\lceil\log_2\tau_n\rceil}}{4\mu_n^2}\mathbb{P}\left(\xi_1\in A_n(X,\Theta),\xi_2\in A_n(X,\Theta)\right) \\
&\leq \frac{a_n^2 2^{2\lceil\log_2\tau_n\rceil}}{4\mu_n^2}\left(\frac{\beta_{a_n}}{2}+\mathbb{P}\left(\xi_1\in A_n(X,\Theta)\right)\mathbb{P}\left(\xi_2\in A_n(X,\Theta)\right)\right) \\
&\leq \frac{a_n^2}{4\mu_n^2}\left(1+\frac{\tau_n^2\beta_{a_n}}{2}\right).
\end{aligned}$$

Hence

$$\mathbb{E}\left[\sum_{i=1}^{\mu_n}\sum_{j=1}^{\mu_n}\sum_{k\in B_i}\sum_{l\in B_j,l\neq k}\tilde{\epsilon}_i\tilde{\epsilon}_j\widetilde{W}_{n,i}\widetilde{W}_{n,j}\right] \leq \frac{a_n^2}{4\mu_n}\left(1+\frac{\tau_n^2\beta_{a_n}}{2}\right)M^2$$

and

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}_{\Theta}\left[W_{n,i}(X,\Theta)W_{n,j}(X,\Theta)\right]\epsilon_i\epsilon_j\right] \leq \frac{a_n^2}{4\mu_n}M^2 + 2\beta_{a_n}\mu_n\left(M^2 + \frac{\tau_n^2 a_n^2}{8\mu_n}M^2\right).$$

Combining both analyses we have,

$$\mathbb{E}\left[\hat{f}_n(X) - \tilde{f}_n(X)\right]^2 \leq C\sigma^2\left(\frac{S^2}{S-1}\right)^{S/2p}(1+\nu_n)\frac{\tau_n a_n^2}{n(\log\tau_n)^{S/2p}} + \frac{a_n^2}{4\mu_n}M^2 + 2\beta_{a_n}\mu_n\left(\sigma^2 + \left(1+\frac{\tau_n^2 a_n^2}{8\mu_n}\right)M^2\right)$$

with

$$C = \frac{288}{\pi}\left(\frac{\pi\log 2}{16}\right)^{S/2p}$$

and

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

□

Proof of the bias term, Proposition 4.2. The start of the proof is the same as in [2] since it does not use the hypothesis of independence between the points:

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

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 3.2 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

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

where $B = H$ or T. We observe that,

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

Thus, using Lemma 3.2:

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

We do the same over T.

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

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

with

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

Using Lemma 3.2,

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

We get,

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

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

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

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

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

Thus,

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

We denote

$$G_{B_j} = \left\{\left\{\xi_i, i \in B_j\right\} \in A_n(X,\Theta)\right\}.$$

for $B = H$ or T.

Let us analyse the first term:

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

For a fixed j,

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

where $\tilde{k}$ denotes one component of each block $\left(H_j\right)_{1\leq j\leq\mu_n}$ and $\left(T_j\right)_{1\leq j\leq\mu_n}$. 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". Using the following inequality(*cf.* [9]),

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

This gives:

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

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

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

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

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

□

7. Proof of consistency of the RF-RI

The computation of the approximation error is the same as [16] since it does not require the independence of $(X_i, Y_i)_{1\leq i\leq n}$ but only that it is stationary and that $(\epsilon_i)_{1\leq i\leq n}$ are independent.

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

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

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

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

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 $\left\{z_1^n \cap A, A \in \mathcal{P}\right\}$ of z_1^n, for $\mathcal{P} \in \Pi_n(\Theta)$. Consequently, the partitioning number $\Gamma_n(\Pi_n(\Theta))$ is defined by

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

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 [16], $C_n = \|m\|_\infty + \sigma\sqrt{2}\log(\alpha_n)^2$, hence eq. (A.1a) is verified.

Estimation error: The computation of the estimation error is very similar to [16] but we need to use a result from [12] to introduce the mixing coefficient which follows from Lemma 3.2.

Theorem 7.1. *Let W_t be a β-mixing stationary stochastic process, with $|Y_i| \leq A_n$ and let $\mathcal{G}_n$ be a class of functions $g : \mathbb{R}^p \to \mathbb{R}$. Then, for any $d \geq 2$,*

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

where $\mathcal{N}(\nu, \mathcal{G}(\Theta), l_{1,n})$ 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 7.1 we get,

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

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 $\mathbb{E}\mathcal{N}\left(\frac{\epsilon}{128C_N}, \mathcal{G}_n(\Theta), l_{1,n}\right)$ (cf. [9]),

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

Hence

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

Going back to the probability computation

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

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

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

For n large enough,

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

with

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

We can now show that eq. (A.1c) holds:

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

We denote

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

We observe that

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

Thus for n large enough,

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

Hence with the $\beta-$mixing condition,

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

Thus, according to appendix Theorem A.1,

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

We only need to check if the non-truncated random forest estimate is consistent. This step is identical to [16].

Appendix A. Tool to establish consistency in stationary ergodic case

To get consistency results for random forests under $\beta-$mixing process, we follow [16]'s ideas. The result is stemming from a general consistency theorem from [9]. We cannot use the latter theorem since we are not in an independent frame thus we will need to adapt it first.

We introduce the operator T defined such that for all function u and real L,

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

Following the definition of the operator T we denote,

$$W_L = T_L W$$

and

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

for $W = X$ or Y and the set

$$T_L \mathcal{G}_n = \{T_L g, g \in \mathcal{G}_n\}.$$

We want to estimate the target function f as

$$\mathbb{E}\left[f(X) - Y\right]^2 = \inf_g \mathbb{E}\left[g(X) - Y\right]^2$$

where the infimum is taken over all measurable functions $g : \mathcal{X} \to \mathcal{Y}$. The solution to this problem is the regression function

$$f(x) = \mathbb{E}\left[Y | X = x\right].$$

This can obviously not be possible to compute since (X, Y) is unknown. We choose a class of functions $\mathcal{G}_n$ which can depend on the data. We then select an estimator $\hat{f}_n$ which minimises the empirical L^2 risk on this class *i.e.*

$$\hat{f}_n(\cdot) = \arg\min_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{j=1}^{n} \left| g(X_j) - Y_j \right|^2 .$$

We first introduce the general consistency theorem as known from [9] and used by [16]. From now on μ denotes the distribution of X.

Theorem A.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 $\hat{f}_n$ and f defined as above. If*

$$\lim_{n\to\infty} C_n = \infty, \tag{A.1a}$$

$$\lim_{n\to\infty} \mathbb{E}\left\{ \inf_{g \in \mathcal{G}_n, \|g\|_\infty \le C_n} \int |g(x) - f(x)|^2 \, \mu(\mathrm{d}x) \right\} = 0, \tag{A.1b}$$

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

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. Before claiming the result, we remind the definition of stationarity and ergodicity. We will give the definition under stationary assumption since we will only consider this case.

Definition A.2. The process $(W_t)_{t\in\mathbb{Z}}$ is said to be stationary if $\forall k \in \mathbb{N}, \forall (t_1, \dots, t_k) \in \mathbb{Z}^k$and $\forall \tau \in \mathbb{Z}$,

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

in distribution.

Definition A.3. The process $(W_t)_{t\in\mathbb{Z}}$ is said to be (mean-)ergodic if

$$\frac{1}{T} \int_0^T W_t \, \mathrm{d}t \xrightarrow[T\to\infty]{\mathrm{L}^2} \mathbb{E}(W_t) .$$

Proposition A.4. *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 $\hat{f}_n$ and f as before. Under eqs.* (A.1a) *to* (A.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 te same line as [9]. 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

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

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,

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

The last term can be bounded,

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

It remains to show that

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

We can lower bound this term by

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

and upper bound it by

$$\begin{aligned}&\mathbb{E}\Bigg(2\left(\mathbb{E}|Y - Y_L|^2\right)^{1/2} + 2\left(\frac{1}{n}\sum_{j=1}^n |Y_i - Y_{i,L}|^2\right)^{1/2} \\ &+ 2\sup_{g\in T_{\beta_n}\mathcal{G}_n} \left|\left(\frac{1}{n}\sum_{j=1}^n |g(X_i) - Y_{i,L}|^2\right)^{1/2} - \left(\mathbb{E}|g(X) - Y|^2\right)^{1/2}\right|\Bigg)^2. \qquad \text{(A.2)}\end{aligned}$$

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

$$\begin{aligned}
\textit{eq.}\ (\text{A.2}) \leq\ & \mathbb{E}\left[\inf_{G\in\mathcal{G}_n, \|g\|\leq\beta_n} \int_{\mathbb{R}^p} |g(x)-f(x)|^2 \mu(\,\mathrm{d}x)\right] \\
& + 6\mathbb{E}\left[\sup_{g\in T_{\beta_n}\mathcal{G}_n} \left|\frac{1}{n}\sum_{j=1}^{n} |g(X_i)-Y_{i,L}|^2 - \mathbb{E}|g(X)-Y|^2\right|\right] \\
& + 6\mathbb{E}|Y-Y_L|^2 + 6\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n}|Y_i-Y_{i,L}|^2\right) \\
& \underset{n\to\infty}{\longrightarrow} 12\mathbb{E}|Y-Y_L|^2.
\end{aligned}$$

The last line using eqs. (A.1b) and (A.1c) and the strong law for stationary ergodic process.

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

□

References

[1] Henry CP Berbee. Random walks with stationary increments and renewal theory. *MC Tracts*, 112:1–223, 1979.

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

[3] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Leo Breiman. Consistency for a simple model of random forests. 2004.

[6] D Richard Cutler, Thomas C Edwards, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[7] Jérôme Dedecker, Paul Doukhan, Gabriel Lang, León R José Rafael, Sana Louhichi, and Clémentine Prieur. Weak dependence. In *Weak Dependence: With Examples and Applications*, pages 9–20. Springer, 2007.

[8] Aurélie Fischer, Lucie Montuelle, Mathilde Mougeot, and Dominique Picard. Real-time wind power forecast. *arXiv preprint arXiv:1610.01000*, 2016.

[9] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[10] Michael J Kane, Natalie Price, Matthew Scotch, and Peter Rabinowitz. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, 15(1):276, 2014.

[11] Aurelie C Lozano, Sanjeev R Kulkarni, and Robert E Schapire. Convergence and consistency of regularized boosting with weakly dependent observations. *IEEE Transactions on Information Theory*, 60(1):651–660, 2014.

[12] Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39(1):5–34, 2000.

[13] Dongxiao Niu, Di Pu, Shuyu Dai, et al. Ultra-short-term wind-power forecasting based on the weighted random forest optimized by the niche immune lion algorithm. *Energies*, 11(5):1–21, 2018.

[14] Anantha M Prasad, Louis R Iverson, and Andy Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, 2006.

[15] Emmanuel Rio. Inequalities and limit theorems for weakly dependent sequences. 2013.

[16] Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

[17] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[18] Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

[19] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[20] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.