A Course on Consistency of Random forests and their Aadaptiveness to Sparsity

Created for the Statistical Learning course of the Sorbonne University Master of Statistics.

Authors :

Guillhem Artis, Thibaud Hadamczik

12/01/2024

école normale supérieure rennes

HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

Contents

1 Introduction

For this course we place ourselves in a nonparametric regression setting, in which the objective is to estimate a regression function :

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}],$$

where $\mathbf{x}$ is a realization of the random variable $\mathbf{X}$ whose support is assumed to be contained in $[0,1]^d$ and $Y \in L^2(\mathbb{R})$ is a square-integrable random variable. We aim to reconstruct r from an i.i.d. sample of n observations $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$, where each pair $(\mathbf{X}_i, Y_i)$ is distributed as $(\mathbf{X}, Y)$. A popular method to reconstruct r are random forests introduced by Breiman [Bre01], which are one example of ensemble estimators since they combine the predictions of multiple base decision trees fitted on the same data. Due to their good performance in practical settings, they have been undergoing a lot of research since their introduction and many different versions now exist varying in how the underlying decision trees are constructed.

We will introduce the main notions of random forests in Section 2 but refer to [BS15] for a more extensive presentation and an overview of recent advances. Subsequently we will see under which conditions consistency of a simplified version of the original forests can be proved as done in [Bia12], extend this result to a a slightly adapted version of these simplified forests and then discuss a result of [SBV15] which shows the consistency of Breiman's forests for a specific form of the regression function.

In Section 4 we will consider a sparse setting, where the regression function only depends on a subset of the features. We will discuss a result of [Bia12] which shows that for a specific construction of the underlying decision trees the rate of convergence is not affected by sparsity and present a result of [SBV15] which shows that Breiman's forests adapt to sparsity.

2 Random Forests

The goal of this section is to introduce the random forest estimator, the different notions that come with and to identify random forests within the more general framework of local averaging estimators. To construct a random forest estimator we depart from base regression trees. A base regression tree yields a random partition of the feature space $[0,1]^d$ into a number of cells and then estimates the target value of a new point $\mathbf{X}$ as the average of the target values Y_i whose corresponding features $\mathbf{X}_i$ fall into the same cell of the partition as $\mathbf{X}$.

The general procedure for constructing the base regression trees is as follows: the root of the tree corresponds to the whole feature space $[0,1]^d$. For the child nodes we

1. randomly choose a coordinate $j \in \{1,\dots,d\}$ with probability $p_{nj} \in (0,1)$ and
2. split the cell corresponding to the parent node at a point according to a predefined rule (potentially random).

This procedure is repeated $\lceil \log_2 k_n \rceil$ times resulting in a tree with $2^{\lceil \log_2 k_n \rceil} \approx k_n$ terminal nodes, each one corresponding to a cell of Lebesgue measure $2^{-\lceil \log_2 k_n \rceil}$ such that the cells corresponding to terminal nodes constitute a partition of $[0,1]^d$. The parameter k_n is set manually, usually depends on n and determines the size of the tree, i.e. the granularity of the partition.

If we denote by $A_n(\mathbf{X},\Theta)$ the cell of the partition that $\mathbf{X}$ falls into, we can write the estimate of a base regression tree as

$$r_n(\mathbf{X},\Theta) = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{[\mathbf{X}_i \in A_n(\mathbf{X},\Theta)]}}{\sum_{i=1}^n \mathbb{1}_{[\mathbf{X}_i \in A_n(\mathbf{X},\Theta)]}} \mathbb{1}_{\mathcal{E}_n(\mathbf{X},\Theta)},$$

where $\mathcal{E}_n$ is the event that at least one $\mathbf{X}_i$ falls into the same cell as $\mathbf{X}$ (output is 0 if $\mathbf{X}$ falls into an empty cell). Θ is a random variable which is independent of $\mathbf{X}$ and describes the random mechanism according to which a base regression tree (and equivalently the corresponding random partition of $[0,1]^d$) is constructed. In particular, it assigns probabilities p_{nj}, with which a cell is split at side j at each node, to each coordinate j.

A random forest estimator combines the estimates of M (usually large) base decision trees $\{r_n(\mathbf{X},\Theta_m,\mathcal{D}_n), 1 \le m \le M\}$, where $(\Theta_1,\dots,\Theta_M)$ are i.i.d. realizations of Θ, by averaging them. The (finite) random forest regressor estimate can be defined as

$$r_{M,n}(\mathbf{X},(\Theta_1,\dots,\Theta_M),\mathcal{D}_n) = \frac{1}{M}\sum_{m=1}^M r_n(\mathbf{X},\Theta_m,\mathcal{D}_n).$$

To ease notation, we denote the number of points falling into the same cell as $\mathbf{X}$ as

$$N_n(\mathbf{X},\Theta) = \sum_{i=1}^n \mathbb{1}_{[\mathbf{X}_i \in A_n(\mathbf{X},\Theta)]},$$

which allows us to rewrite the finite random forest estimate as

$$\begin{aligned} r_{M,n}(\mathbf{X},(\Theta_1,\dots,\Theta_M),\mathcal{D}_n) &= \frac{1}{M}\sum_{m=1}^M \frac{\sum_{i=1}^n Y_i \mathbb{1}_{[\mathbf{X}_i \in A_n(\mathbf{X},\Theta_m)]}}{N_n(\mathbf{X},\Theta_m)} \mathbb{1}_{\mathcal{E}_n(\mathbf{X},\Theta_m)} && (1) \\ &= \sum_{i=1}^n W_{M,ni}(\mathbf{X}) Y_i && (2) \end{aligned}$$

with weights :

$$W_{M,ni} = W_{M,ni}(\mathbf{X}) = W_{M,ni}(\mathbf{X}, (\Theta_1, \ldots, \Theta_M)) = \frac{1}{M} \sum_{m=1}^{M} \frac{\mathbb{1}_{[\mathbf{X}_i \in A_n(\mathbf{X}, \Theta_m)]}}{N_n(\mathbf{X}, \Theta_m)} \mathbb{1}_{\mathcal{E}_n(\mathbf{X}, \Theta_m)}.$$

From the definition of $W_{M,ni}$ it follows that $\sum_{i=1}^{n} W_{M,ni} = 1$ and $W_{M,ni} \geq 0$ for $1 \leq i \leq n$. In the case where Θ is independent of $\mathbf{X}$, random forests can therefore be interpreted as local averaging estimators (cf. Section 4.2 in[Gyö+02]). From (2) it is indeed obvious that the random forest estimate is a convex combination of target values Y_i whose features $\mathbf{X}_i$ fall into the same cell as $\mathbf{X}$ for each base decision tree. Note that the higher the number of decision trees, for which $\mathbf{X}_i$ shares a cell with $\mathbf{X}$, the higher the weight of Y_i. To simplify notation we will often write $r_{M,n}(\mathbf{X})$ instead of $r_{M,n}(\mathbf{X}, (\Theta_1, \ldots, \Theta_M), \mathcal{D}_n)$. For $M = 1$ the random forest estimate is just the estimate of one base decision tree and we denote its weights as $W_{ni} := W_{1,ni}$.

Having presented the broad idea of random forests, we want to highlight that the random mechanism constructing the partition, encoded in the random variable Θ, has a major influence on the performance of the resulting estimator. The numerous types of random forests can be grouped into three categories according to the the level of independence between Θ and $\mathcal{D}_n$.

1. Purely random forests

The random mechanism of purely random forests is completely independent of the data sample $\mathcal{D}_n$. First for each node a coordinate $j \in 1, \ldots, d$ is randomly chosen independently of $\mathcal{D}_n$ and then a point of the sampled side j of the cell corresponding to the parent node is sampled independently from $\mathcal{D}_n$. The cell is then cut at the sampled point of the selected side j.

Examples include centered forests and uniform forests. During the construction of a centered forest each coordinate j has the same probability of being selected, $p_{nj} = \frac{1}{d}$, and the split is then performed at the midpoint of the selected side. Similarly, for uniform forests a coordinate is sampled in the same manner but then instead of splitting the cell at the midpoint of the selected side, the splitting point is sampled from a uniform distribution on the selected side.

2. Random forests satisfying the X-property

The second category comprises forests which make use of the features $\mathbf{X}$ to construct the partition. We say that a random forest estimator satisfies the **X-property** if the construction of the underlying base regression trees depends only on $\mathbf{X}$. Note that purely random forests are a subset of random forests satisfying the X-property. Another example of such forests are quantile forests where, for $q \in (0, 1)$, after random selection of a coordinate j the splitting point is chosen as the (empirical) q-quantile of the $(\mathbf{X}_i^{(j)})$ present in the cell. The procedure can potentially involve subsampling of $\mathcal{D}_n$ such that for each split only a subset of the sample is used for determining the splitting point.

3. Random forests not satisfying the X-property

The third and last category are random forests whose random mechanism depends both on the features $\mathbf{X}$ and the targets Y of the data sample and is thus the most involved one among the three categories presented here. The most famous examples are Breiman's CART-forests. They are constructed by selecting at each node a coordinate j from a subset of $\{1, \ldots, d\}$ and a splitting point such that an error measure (which depends on both $\mathbf{X}$ and Y) is minimized over (a subset of) $\mathcal{D}_n$ (Gini impurity for classification/MSE for regression) when splitting at coordinate j. In practical settings these forests are the most powerful ones, however, as we will see in section 3, the fact that the X-property is not satisfied anymore for this group of forests will inhibit us from proving their consistency with the same tools as for the two other groups. Also for this group the procedure may involve subsampling.

In the following, we focus on the consistency of random forests. We say that a regression function estimate r_n is consistent for r if the L^2-risk vanishes as the sample size goes to infinity,

$$\lim_{n\to\infty} \mathcal{R}(r_n) \coloneqq \lim_{n\to\infty} \mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2 = 0, \tag{3}$$

where the expectation is taken with respect to $\mathbf{X}$ and $\mathcal{D}_n$. If we let the number of base regression trees M in a random forest grow to infinity, by the law of large number we obtain the infinite random forest estimate :

$$\bar{r}_n(\mathbf{X}, \mathcal{D}_n) \coloneqq \lim_{M\to\infty} r_{M,n}(\mathbf{X}) = \mathbb{E}_\Theta[r_n(\mathbf{X}, \Theta, \mathcal{D}_n)]$$

where the expectation is taken with respect to Θ and conditionally on $\mathcal{D}_n$. In practice the infinite random forest estimate is approximated by sampling a large number of base regression trees and averaging their outputs. In Section 3 we will show that under certain conditions it suffices to show consistency of the infinite random forest estimate in order to establish consistency of the underlying finite random forest estimate. We will need the notion of the diameter of a set which is defined as

$$\operatorname{diam}(A_n) \coloneqq \sup_{x,y\in A_n} \|x - y\|_2 \tag{4}$$

for a cell $A_n \subset [0,1]^d$.

3 Consistency of random forests

Having established random forests as local averaging estimators in Section 2, one might want to use Stone's theorem which provides sufficient conditions for consistency for this class of estimators. Instead of doing that, we use the following theorem (adaptation of Theorem 4.2 in [Gyö+02]), which is a consequence of Stone's theorem. The result provides sufficient conditions for the consistency of estimators which, like random forests, partition the feature space and estimate the value for a new point $\mathbf{X}$ as the average of target points whose features fall into the same cell as $\mathbf{X}$ (partitioning regression function estimators).

Theorem 3.1. *Let r_n be a random forest estimate with the X-property. If*

(i) $\operatorname{diam}(A_n(\mathbf{X}, \Theta)) \to 0$ *in probability, and*

(ii) $N_n(\mathbf{X}, \Theta) \to \infty$ *in probability,*

then r_n is consistent in the sense of (3).

Intuitively the theorem states that a random forest estimator is consistent if its random mechanism Θ ensures that, when n goes to infinity, the cell $A_n(\mathbf{X}, \Theta)$ containing $\mathbf{X}$ becomes infinitely small while the number of sample points in this cell tends to ∞. Requiring r_n to satisfy the X-property means in particular that this theorem cannot be used to prove consistency of random forests whose splits depend also on the target values Y_i, e.g. CART-forests.

We will now show consistency of the infinite random forests estimators whose underlying base decision trees are grown by splitting at the midpoints. An assumption on the tree size, parameterized by k_n, and the probabilities p_{nj} of selecting a dimension j to split on, will be required.

Theorem 3.2. *Assume that the distribution of* $\mathbf{X}$ *has support on* $[0,1]^d$*. Then the random forests estimate* $\bar{r}_n$ *whose trees split the cell at the midpoint of the selected dimension is consistent whenever* $p_{nj}\log(k_n) \to \infty$ *for all* $j = 1, \dots, d$ *and* $\frac{k_n}{n} \to 0$ *as* $n \to \infty$*.*

Proof. In order to prove consistency of the infinite random forest estimate, we show that a single base decision tree of the random forest is consistent. This procedure is justified by Jensen's inequality,

$$\begin{aligned}\mathbb{E}[\bar{r}_n(\mathbf{X}) - r(\mathbf{X})]^2 &= \mathbb{E}[\mathbb{E}_\Theta[r_n(\mathbf{X}) - r(\mathbf{X})]]^2 \\ &\le \mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^2.\end{aligned} \tag{5}$$

We recall that, by our assumptions on the underlying base decision trees, they are constructed by first selecting a coordinate $j \in \{1, \dots, d\}$ with probability p_{nj} and then splitting the cell corresponding to the parent node at the midpoint of the selected dimension. Repeating this procedure $\lceil \log_2 k_n \rceil$ times will result in a tree with $2^{\lceil \log_2 k_n \rceil}$ terminal nodes. We will denote the terminal cells by $A_1, \dots, A_{2^{\lceil \log k_n \rceil}}$ and the number of observations among $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ falling into each of these cells by $N_1, \dots, N_{2^{\lceil \log k_n \rceil}}$. Further we denote by $\mathcal{C} = \{\mathbf{X}, \mathbf{X_1}, \dots, \mathbf{X_n}\}$ the set of positions of these $n+1$ points. To show that $N_n(\mathbf{X}, \Theta) \to \infty$ in probability, one can notice that, since $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d., conditioning on $\mathcal{C}$ and Θ, the probability that $\mathbf{X}$ falls in the ℓ-th cell equals $\frac{N_\ell}{n+1}$. Thus, for all $M > 0$,

$$\begin{aligned}\mathbb{P}(N_n(\mathbf{X}, \Theta) < M) &= \mathbb{E}[\mathbb{P}(N_n(\mathbf{X}, \Theta) < M) | \mathbf{X}, \Theta] \\ &= \mathbb{E}\left[\sum_{\ell=1,\dots,2^{\lceil \log_2 k_n \rceil}: N_\ell < M} \frac{N_\ell}{n+1}\right] \\ &\le \sum_{\ell=1,\dots,2^{\lceil \log_2 k_n \rceil}} \frac{M}{n+1} \\ &= \frac{2^{\lceil \log_2 k_n \rceil} M}{n+1} \\ &\le \frac{2k_n M}{n+1},\end{aligned}$$

which goes to 0 since $\frac{k_n}{n} \to 0$. This shows that $N_n(\mathbf{X}, \Theta) \to \infty$ in probability.

We now show that $\operatorname{diam}(A_n(\mathbf{X}, \Theta)) \to 0$ by showing that the length $V_{nj}(\mathbf{X}, \Theta)$ of the side corresponding to dimension j of the cell containing $\mathbf{X}$ goes to 0 in probability for all $j \in \{1, \dots, d\}$, which can be justified by a union bound. Since the length of side j is halved every time j is selected, denoting by $K_{nj}(\mathbf{X}, \Theta)$ the number of times j has been selected before getting $A_n(\mathbf{X}, \Theta)$, we obtain

$$V_{nj}(\mathbf{X}, \Theta) \overset{\mathfrak{D}}{=} 2^{-K_{nj}(\mathbf{X}, \Theta)}. \tag{6}$$

Now, by independence of $\mathbf{X}$ and Θ, and since each cell is the result of $\lceil \log_2 k_n \rceil$ splits with probability p_{nj} to perform it on the dimension j, conditionally on $\mathbf{X}$, $K_{nj}(\mathbf{X}, \Theta)$ follows a binomial distribution,

$$K_{nj}(\mathbf{X}, \Theta) | \mathbf{X} \sim \mathcal{B}(\lceil \log_2 k_n \rceil, p_{nj}).$$

Thus, we have for all $\varepsilon > 0$

$$\mathbb{P}(V_{nj}(\mathbf{X}, \Theta) \geq \varepsilon) \leq \frac{\mathbb{E}[V_{nj}(\mathbf{X}, \Theta)]}{\varepsilon}$$

and

$$\begin{aligned}
\mathbb{E}[V_{nj}(\mathbf{X}, \Theta)] &= \mathbb{E}\left[2^{-K_{nj}(\mathbf{X}, \Theta)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[2^{-K_{nj}(\mathbf{X}, \Theta)} | \mathbf{X}\right]\right] \\
&\overset{(6)}{=} \mathbb{E}[2^{-\mathcal{B}(\lceil \log_2 k_n \rceil, p_{nj})}] \\
&= \left(1 - \frac{p_{nj}}{2}\right)^{\lceil \log_2 k_n \rceil},
\end{aligned}$$

where the last equality follows from the Binomial theorem. To show that the above expectation tends to 0, it suffices to use the concavity of the logarithm,

$$\begin{aligned}
\left(1 - \frac{p_{nj}}{2}\right)^{\lceil \log_2 k_n \rceil} &\leq \exp\left((\log_2 k_n)\left(\log(1 - \frac{p_{nj}}{2}\right)\right) \\
&\leq \exp\left((\log_2 k_n)\left(-\frac{p_{nj}}{2}\right)\right) \\
&= \exp\left(-\frac{1}{2} p_{nj} \log_2(k_n))\right)
\end{aligned}$$

and notice that the final expression indeed goes to 0 since $p_{nj} \log k_n \to \infty$. □

The proof shows that in the setting of the theorem, consistency of a single tree implies consistency of the infinite random forest estimate. By replacing $\bar{r}_n$ in (5) by $r_{M,n}$ the same holds true for the finite random forest estimator. We will later derive sufficient conditions under which consistency of the infinite random forest estimator implies consistency of the finite one. Furthermore, we note that the assumptions of the theorem are satisfied for $p_{nj} = \frac{1}{d}$ as long as $k_n \to \infty$ and $\frac{k_n}{n} \to 0$, therefore implying consistency of centered forests e.g. for a choice of $k_n = \log n$. What if instead of splitting at the midpoint of the cell, after selecting the coordinate j to split on, the splitting point is sampled from a uniform distribution on the selected side? First note that in this case, $N_n(\mathbf{X}, \Theta) \to \infty$ in probability as in the proof of the theorem, since we still have independence between $\mathbf{X}$ and Θ. Next, the distribution of the length $V_{nj}(\mathbf{X}, \Theta)$ of the side of the cell that $\mathbf{X}$ falls into, conditionally on the number of times $K_{nj}(\mathbf{X}, \Theta)$ that the cell has been split at dimension j, satisfies

$$V_{nj}(\mathbf{X}, \Theta) | K_{nj}(\mathbf{X}, \Theta) \preceq \prod_{k=1}^{K_{nj}(\mathbf{X}, \Theta)} \max(U_k, 1 - U_k)$$

with $U_1, \ldots, U_{K_{nj}(\mathbf{X}, \Theta)}$ i.i.d. uniformly on $[0, 1]$ and $A \preceq B$ denoting stochastic dominance of B over A for A, B two random variables. This follows from the fact that we depart from a side of length 1 and progressively scale the length by a factor of either U or $1-U$ with U uniformly distributed on $[0,1]$ each time the cell is split along the respective dimension. More explicitly, each time we split at a point given by U on the j-th coordinateand $\mathbf{X}$ is either on the side of length U or on the side of length $1-U$. We can therefore proceed analogously to the second part of the proof and write

$$\begin{aligned}\mathbb{E}[V_{nj}(\mathbf{X},\Theta] &\leq \mathbb{E}\left[\mathbb{E}\left[\prod_{k=1}^{K_{nj}(\mathbf{X},\Theta)} \max(U_k, 1-I_k)\middle| K_{nj}(\mathbf{X},\Theta)\right]\right] \\ &= \mathbb{E}\left[\left(\mathbb{E}\left[\max(U_1,1-U_1)\right]\right)^{K_{nj}(\mathbf{X},\Theta)}\right] \\ &\overset{(6)}{=} \mathbb{E}\left[\left(\frac{3}{4}\right)^{-\mathcal{B}(\lceil\log_2 k_n\rceil, p_{nj})}\right] \\ &= \left(1-\frac{p_{nj}}{4}\right)^{\lceil\log_2 k_n\rceil},\end{aligned}$$

which tends to 0 under the conditions of Theorem 3.2 with the same reasoning as in the proof. The conditions are in particular satisfied if $p_{nj} = \frac{1}{d}$ for all $j \in \{1,\ldots,d\}$ and $k_n = \log n$. We have therefore, by an adaptation of the proof of Corollary 1 in [Sco16], shown that uniform forests are consistent for this choice of k_n.

This proof strategy can be used to show consistency of other random forest estimators satisfying the X-property like median forests, or, more generally, quantile forests. For CART-trees however, the X-property is not satisfied anymore, which makes proving their consistency a real challenge. So far, consistency for CART-forests has not been shown yet, but we will present a result of [SBV15] proving consistency of CART-forests in an additive regression framework. In this framework we assume that the regression function can be written as a finite sum of functions of the coordinates of $\mathbf{X}$,

$$r(\mathbf{X}) = \sum_{j=1}^{d} r_j(\mathbf{X}^{(j)}),$$

Additionally, the components r_j are supposed to be continuous and $\mathbf{X}$ uniformly distributed uniformly $[0,1]^d$. Furthermore the assumption of an independent Gaussian noise is made, such that

$$Y - r(\mathbf{X}) \sim \mathcal{N}(0,\sigma^2), \quad \sigma^2 < \infty.$$

Denoting these hypotheses by $(\mathcal{H})$, we cite the following result (Theorem 4.1 in [SBV15])

Theorem 3.3. *Assume that $(\mathcal{H})$ is satisfied and let r_n be a CART-forest, where, at each iteration, a_n sample points are subsampled uniformly with replacement from $\mathcal{D}_n$ and the splits are performed minimizing the CART-criterion until the tree has $t_n \in \{1,\ldots,a_n\}$ leaves. Then, provided $a_n \to \infty$, $t_n \to \infty$ and $t_n(\log a_n)^9/a_n \to 0$, r_n is consistent.*

This result shows that, by constraining the shape of the regression function, it is possible to obtain consistency results even for trees whose constructions involve several sources of randomness interacting with each other.

The next result enables us to derive conditions under which consistency of the infinite forest implies consistency of the random forest.

Lemma 3.4. *If* $\mathbf{X}$ *is uniformly distributed on* $[0,1]^d$, $Y \in L^2$ *and* $r_{M,n}$ *is a purely random forest,*

$$R(r_{M,n}) - R(\bar{r}_n) = \mathcal{O}\left(\frac{k_n}{M}\right).$$

Proof. Using an equality derived in the proof of Theorem 3.3. in [Sco16] and the definition of the base regression trees, we obtain

$$\begin{aligned} R(r_{M,n}) - R(\bar{r}_n) &= \frac{1}{M}\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\left[V_\Theta[r_n(\mathbf{X},\Theta)]\right] \\ &= \frac{1}{M}\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\left[V_\Theta\left[\sum_{i=1}^n W_{ni}Y_i\right]\right] \\ &\leq \frac{1}{M}\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\left[\mathbb{E}_\Theta\left[\left(\sum_{i=1}^n W_{ni}Y_i\right)^2\right]\right] \end{aligned}$$

Since $\sum_{i=1}^n W_{ni} = 1$, by convexity of the square function we can bound

$$\begin{aligned} R(r_{M,n}) - R(\bar{r}_n) &\leq \frac{1}{M}\mathbb{E}_{\mathbf{X},\mathcal{D}_n}\left[\mathbb{E}_\Theta\left[\sum_{i=1}^n W_{ni}Y_i^2\right]\right] \\ &= \frac{1}{M}\mathbb{E}_{X,\mathcal{D}_n,\Theta}\left[\sum_{i=1}^n W_{ni}Y_i^2\right] \\ &= \frac{1}{M}\sum_{i=1}^n \mathbb{E}_{\Theta,X,(X_i,Y_i)}\left[\mathbb{E}_{\mathcal{D}_n^{-i}}\left[W_{ni}Y_i^2|\Theta,X,(X_i,Y_i)\right]\right], \\ &= \frac{1}{M}\sum_{i=1}^n \mathbb{E}_{\Theta,\mathbf{X},(X_i,Y_i)}\left[Y_i^2\mathbb{E}_{\mathcal{D}_n^{-i}}\left[W_{ni}|\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)\right]\right] \end{aligned}$$

where

$$\mathcal{D}_n^{-i} := ((\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_{i-1},Y_{i-1}),(\mathbf{X}_{i+1},Y_{i+1}),\ldots,(\mathbf{X}_n,Y_n)).$$

Since $Y_i^2\mathbb{1}_{\{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)\}}$ is $\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)$-measurable,

$$\begin{aligned} &R(r_{M,n}) - R(\bar{r}_n) \\ &\leq \frac{1}{M}\sum_{i=1}^n \mathbb{E}_{\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)}\left[Y_i^2\mathbb{1}_{\{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)\}}\mathbb{E}_{\mathcal{D}_n^{-i}}\left[\frac{\mathbb{1}_{\mathcal{E}_n(\mathbf{X},\Theta)}}{\sum_{j=i}^n \mathbb{1}_{\{\mathbf{X}_j\in A_n(\mathbf{X},\Theta)\}}}\middle|\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)\right]\right] \\ &\leq \frac{1}{M}\sum_{i=1}^n \mathbb{E}_{\Theta,X,(\mathbf{X}_i,Y_i)}\left[Y_i^2\mathbb{1}_{\{\mathbf{X}_i\in A_n(\mathbf{X},\Theta)\}}\mathbb{E}_{\mathcal{D}_n^{-i}}\left[\frac{\mathbb{1}_{\mathcal{E}_n(\mathbf{X},\Theta)}}{\sum_{j\neq i} \mathbb{1}_{\{X_j\in A_n(\mathbf{X},\Theta)\}}}\middle|\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)\right]\right] \end{aligned}$$

Since $\mathbf{X},\mathbf{X}_1,\ldots,\mathbf{X}_n$ are independently, uniformly distributed on $[0,1]^d$, conditionally to $\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)$ the average number of $(\mathbf{X}_j)_{1\leq j\leq n, j\neq i}$ falling into the same cell as $\mathbf{X}$ is $\frac{n-1}{k_n}$.

$$
\begin{aligned}
R(r_{M,n}) - R(\bar{r}_n) &\leq \frac{1}{M}\sum_{i=1}^{n} \mathbb{E}_{\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)} \left[Y_i^2 \mathbb{1}_{\{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)\}} \frac{k_n}{n-1} \right] \\
&= \frac{k_n}{M(n-1)} \sum_{i=1}^{n} \mathbb{E}_{\Theta,\mathbf{X},(\mathbf{X}_i,Y_i)} \left[Y_i^2 \mathbb{1}_{\{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)\}} \right] \\
&= \frac{k_n}{M(n-1)} \mathbb{E}_{\Theta,X} \left[\sum_{i=1}^{n} \mathbb{E}_{(\mathbf{X}_i,Y_i)} \left[Y_i^2 \mathbb{1}_{\{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)} \middle| \Theta, X \right] \right] \\
&\leq \frac{k_n}{M(n-1)} \mathbb{E}_{\Theta,\mathbf{X}} \left[\sum_{i=1}^{n} \mathbb{E}_{(\mathbf{X}_i,Y_i)} \left[Y_i^2 \middle| \Theta, \mathbf{X} \right] \right] \\
&= \frac{n k_n}{M(n-1)} \mathbb{E}\left[Y^2 \right] \\
&\approx \frac{k_n}{M} \mathbb{E}\left[Y^2 \right]
\end{aligned}
$$

□

This lemma, coupled with Biau's consistency theorem of purely random forests of infinite trees [Bia12] enable us to derive sufficient conditions for purely random forests with finitely many trees' consistency :

Proposition 3.5. *Assume that the distribution of X is uniformly distributes on $[0,1]^d$. If for $\frac{M}{k_n} \to \infty$, the consistency of purely random infinite random forests regressor estimate implies the consistency of purely finite random forests.*

Remark 3.6. The proposition above provides a one way relationship between random forests with infinitely many trees and random forests with finitely many trees. In the setting of theorem 3.2, and provided that $\frac{M}{k_n} \to \infty$, the consistency of the infinite random forests implies the consitency of the finite random forests (with M trees). This result is interesting in the sense that infinite forests are imposssible to get in practice, but theoritical results on such estimates can be from then on transfer to the finite ones.

4 Rate of convergence and adaptiveness to sparsity

In this section we will derive the rate of convergence for the L^2-risk of random forests which perform splits at the midpoint of the selected side and see to what extent they adapt to sparsity. Adaptiveness to sparsity is a desirable feature of machine learning models. In practical settings we often have high-dimensional data while the regression function only depends on a small fraction of the available features. An exemplary use case is the domain of biostatistics, where one might be interested in knowing the genes that are responsible for a given genetic disease. The number of genes whose expression is considered having an impact on the disease may be very high-dimensional, whilst the number of genes that actually do is probably very small. This motivates the requirement for machine learning algorithms that adapt to sparsity and leads to the following question: In how far do random forest estimators adapt to sparsity? This question is answered in [Bia12] for the simplified version of random forests that split the cell at the midpoint of the selected dimension.

We will firstly define the considered sparsity setting formally before giving convergence results for the bias and variance of the random forest estimator. Those results will then be combined to obtain the convergence rate of the L^2-error. We will see that, if we have a method to tune the probabilities $(p_{nj})_{1\leq j\leq d}$ of the splits, the L^2-risk will decrease with the number of features used in the regression function.

For the sparse regression setting we assume that the dimension d is possibly large but r, the regression function, depends on only S coordinates of $\mathbf{X}$, called strong features. This allows to express the regression function as

$$r(\mathbf{x}) = r^*(\mathbf{x}_{\mathcal{S}}),$$

where $x_{\mathcal{S}}$ is the projection of $\mathbf{x}$ on the domain of the strong features which can be identified as $[0,1]^S$. To make this framework interesting, we assume that $2 \leq S << d$. We denote by $\mathcal{S} \subset \{1,\ldots,d\}$ the set of strong features and by $\mathcal{W} = \{1,\ldots,d\} \setminus \mathcal{S}$ the set of weak features, meaning $S = |\mathcal{S}|$. The goal is to study the L^2-risk of $\bar{r}_n$ as a function of S. To this end, we assume that at each step, we choose the j-th coordinate with probability $p_{nj} = \frac{1}{S}(1+\xi_{nj})$ if $j \in \mathcal{S}$ or $p_{nj} = \xi_{nj}$ otherwise (i.e. we assume to possess a method to tune this probability). The strategy of the proof, is to control both terms of the bias-variance decomposition

$$\mathbb{E}[\bar{r}_n(\mathbf{X}) - r(\mathbf{X})]^2 = \mathbb{E}[\bar{r}_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})]^2 + \mathbb{E}[\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})]^2,$$

where

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

The following proposition gives a bound on the variance term.

Proposition 4.1. *Assume that $\mathbf{X}$ is uniformly distributed on $[0,1]^d$ and for all $\mathbf{x} \in \mathbb{R}^d$*

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

for some positive constant σ^2. Then if $p_{nj} = \frac{1}{S}(1+\xi_{nj})$ for $j \in S$

$$\mathbb{E}[\bar{r}_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})]^2 \leq C\sigma^2 \left(\frac{S^2}{S-1}\right)^{\frac{S}{2d}} (1+\xi_n) \frac{k_n}{n(\log k_n)^{\frac{S}{2d}}}$$

where

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

and the sequence (ξ_n) depends on the sequences $\{(\xi_{nj}) : j \in \mathcal{S}\}$ only and tends to 0 as n tends to infinity.

Proof sketch. The proof of the proposition is tedious and technical, which is why we will only explain the main ideas. In the following we will shorten notation and write W_{ni} instead of $W_{ni}(\mathbf{X},\Theta)$. First, using the definition of the regression function and the assumption of a uniformly bounded conditional variance, we get

$$\mathbb{E}[\bar{r}_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})]^2 \leq n\sigma^2 \mathbb{E}\left[\mathbb{E}_\Theta^2[W_{n1}]\right]. \tag{7}$$

By using a symmetry trick,

$$\mathbb{E}\left[\mathbb{E}_{\Theta}^{2}[W_{n1}]\right] = \mathbb{E}\left[\mathbb{E}_{\Theta}[W_{n1}]\right]\mathbb{E}\left[\mathbb{E}_{\Theta'}[W_{n1}]\right],$$

with Θ' independent of Θ. It now suffices to replace W_{n1} by its expression and use (7) to obtain

$$\mathbb{E}[\bar{r}_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})]^2 \leq n\sigma^2 \mathbb{E}\left[\frac{\mathbb{1}_{[\mathbf{X}_1 \in A_n(\mathbf{X},\Theta) \cap A_n(\mathbf{X},\Theta')]}}{N_n(\mathbf{X},\Theta) N_n(\mathbf{X},\Theta')} \mathbb{1}_{[\mathcal{E}_n(\mathbf{X},\Theta)]} \mathbb{1}_{[\mathcal{E}_n(\mathbf{X},\Theta')]}\right].$$

Then, by conditioning on the independent variables $\mathbf{X}, \mathbf{X}_1, \Theta, \Theta'$ and noticing that

$$\sum_{i=1}^{n} \mathbb{1}_{[\mathbf{X}_1 \in A_n(\Theta,\mathbf{X})]} \geq 1 + \sum_{i=2}^{n} \mathbb{1}_{[\mathbf{X}_1 \in A_n(\Theta,\mathbf{X})]},$$

we bound

$$\begin{aligned} &\mathbb{E}[\bar{r}_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})]^2 \\ \leq\ & n\sigma^2 \mathbb{E}\left[\mathbb{1}_{[X_1 \in A_n(\mathbf{X},\Theta) \cap A_n(\mathbf{X},\Theta')]} \mathbb{E}\left[\frac{1}{1 + \sum_{i=2}^{n} \mathbb{1}_{[\mathbf{X}_1 \in A_n(\Theta,\mathbf{X})]}} \frac{1}{1 + \sum_{i=2}^{n} \mathbb{1}_{[\mathbf{X}_1 \in A_n(\Theta',\mathbf{X})]}} \middle| \mathbf{X}, \Theta, \Theta'\right]\right]. \end{aligned}$$

Since $\mathbf{X}, \Theta$ and Θ' are independent, the conditional expectation occurring in the last inequality can be bounded by using the Cauchy-Schwarz inequality and a technical result providing an upper bound on $\mathbb{E}[1/(1+Z^2)]$ when Z follows a binomial distribution (note that, because Z is positive, $\mathbb{E}[1/(1+Z^2)] \geq \mathbb{E}[1/(1+Z)^2]$, resulting in

$$\sqrt{\mathbb{E}\left[\frac{1}{(1 + \sum_{i=2}^{n} \mathbb{1}_{[\mathbf{X}_1 \in A_n(\Theta,\mathbf{X})]})^2} \middle| \mathbf{X}, \Theta\right] \mathbb{E}\left[\frac{1}{(1 + \sum_{i=2}^{n} \mathbb{1}_{[\mathbf{X}_1 \in A_n(\Theta',\mathbf{X})]})^2} \middle| \mathbf{X}, \Theta'\right]} \leq \frac{12k_n^2}{n^2}.$$

Now, combining the two previous inequalities yields

$$\mathbb{E}[\bar{r}_n(\mathbf{X}) - \tilde{r}_n(\mathbf{X})]^2 \leq \frac{24\sigma^2 k_n^2}{n} \mathbb{E}\left[\mathbb{P}_{\mathbf{X}_1}(\mathbb{1}_{[\mathbf{X}_1 \in A_n(\mathbf{X},\Theta) \cap A_n(\mathbf{X},\Theta')]})\right].$$

After several computations leading to an upper bound of the above expectation the Lebesgue measure of the set $A_n(\mathbf{X},\Theta) \cap A_n(\mathbf{X},\Theta')$ appears. Recall that structure of the cell $A_n(\mathbf{X},\Theta) = \Pi_{j=1}^{d} A_{nj}(X,\Theta)$ depends on the split directions and thus on $(p_{nj})_{1 \leq j \leq d}$ meaning on the strong features. We do not give supplementary details of this proof, our interest was to explain the steps until sparsity comes into play. The interested reader might take a look at the end of the proof of Proposition 2 in [Bia12]. □

Proposition 4.1 indicates that the variance is of order $\mathcal{O}\left(\frac{k_n}{n(\log k_n)^{\frac{S}{2d}}}\right)$. Comparing this rate to the order of the variance of individual trees, which has been established by previous research, trees achieve consistency as the number of sample points in each terminal node increases, typically resulting in a variance of the order $\frac{k_n}{n}$. From this perspective, setting $k_n = n$, which equates to roughly one observation per terminal node, is not advisable for individual trees. This approach can lead to substantial overfitting and a dramatic increase in variance. However, when considering random forests, the situation changes. Using $k_n = n$ for forests results in a variance of the order $\frac{1}{(\log n)^{\frac{S}{2d}}}$, a value that diminishes as n increases. The proof of Proposition 4.1 highlights that the logarithmic term in the variance formula is a direct consequence of the Θ-averaging process. It emerges when accounting for the correlation between trees. This offers an explanation for the effectiveness of random forests, even though the individual trees are not pruned.

We will now quote a proposition that yields an upper bound on the bias term.

Proposition 4.2. *Assume that* $\mathbf{X}$ *is uniformly distributed on* $[0,1]^d$ *and* r^* *is L-Lipschitz on* $[0,1]^S$*. Then if* $p_{nj} = \frac{1}{S}(1+\xi_{nj})$ *for* $j \in S$

$$E[\tilde{r}_n(\mathbf{X}) - r(\mathbf{X})]^2 \leq \frac{2SL^2}{k^{0.75}S\log^2(1+\gamma_n)n} + \sup_{\mathbf{x}\in[0,1]^d} r^2(x)e^{-\frac{n}{2k_n}}$$

Theorem 4.3. *Assume that* $\mathbf{X}$ *is uniformly distributed on* $[0,1]^d$*,* r^* *is L-Lipschitz on* $[0,1]^S$*, and for all* $\mathbf{x} \in \mathbb{R}^d$*,*

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

for some positive constant σ^2*. Then, if* $p_{nj} = \frac{1}{S}(1+\xi_{nj})$ *for* $j \in S$*, letting* $\gamma_n = \min_{j\in S}\xi_{nj}$*, we have*

$$\mathbb{E}[\bar{r}_n(\mathbf{X}) - r(\mathbf{X})]^2 \leq \Xi_n \frac{k_n}{n} + \frac{2SL^2}{k_n^{\frac{0.75}{S\log 2}(1+\gamma_n)}},$$

where

$$\Xi_n = C\sigma^2\left(\frac{S^2}{S-1}\right)^{S/2d}(1+\xi_n) + 2e^{-1}\left[\sup_{\mathbf{x}\in[0,1]^d} r^2(\mathbf{x})\right]$$

and

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

The sequence (ξ_n) *depends on the sequences* $\{(\xi_{nj}) : j \in S\}$ *only and tends to* 0 *as* n *tends to infinity.*

Corollary 4.4. *Let*

$$\Xi = C\sigma^2\left(\frac{S^2}{S-1}\right)^{S/2d} + 2e^{-1}\left[\sup_{\mathbf{x}\in[0,1]^d} r^2(\mathbf{x})\right]$$

and

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

Then, if $p_{nj} = (1/S)(1+\xi_{nj})$ *for* $j \in S$*, with* $\xi_{nj}\log n \to 0$ *as* $n \to \infty$*, for the choice*

$$k_n \propto \left(\frac{L^2}{\Xi}\right)^{1/(1+\frac{0.75}{S\log 2})} n^{1/(1+\frac{0.75}{S\log 2})},$$

and by denoting $\mathcal{F}_S$ the class of (L, σ^2)-smooth distributions $(\boldsymbol{X}, Y)$ such that $\boldsymbol{X}$ has a uniform distribution on $[0,1]^d$, the regression function $r^\star$ is Lipschitz with constant L on $[0,1]^S$ and, for all $x \in \mathbb{R}^d$, $\sigma^2(x) = \mathbb{V}[Y|\boldsymbol{X} = x] \leq \sigma^2$, we have,

$$\limsup_{n\to\infty} \sup_{(\mathbf{X},Y)\in\mathcal{F}_S} \frac{\mathbb{E}[\bar{r}_n(\mathbf{X}) - r(\mathbf{X})]^2}{\left(\Xi L^{\frac{2S\log 2}{0.75}}\right)^{\frac{0.75}{S\log 2+0.75}} n^{\frac{-0.75}{S\log 2+0.75}}} \leq \Lambda,$$

where Λ is a positive constant independent of r, L and σ^2.

This result leads to the pinnacle of the article [Bia12]. Within the sparsity assumption frameworks (and all the above hypothesis),

$$\mathbb{E}[\bar{r_n}(X) - r(X)]^2 = \mathcal{O}\left(n^{\frac{-0.75}{S\log(2)+0.75}}\right).$$

Strikingly, this bound does only depend on the number of strong variables S (for the rigorous reader, the more precise bound given in the corollary above does effectively depend on d in the term $\frac{S}{2d}$, but it does not effect the consistency result). The main message of this part is that, provided that one can tune correctly the probabilities $(p_{nj})_{1\leq j\leq d}$ in a way that favours choosing the strong variables for spliting, then random forest will adapt to sparsity. Indeed, we have seen that adaptiveness only arises from the $(p_{nj})_{1\leq j\leq d}$ expressions. How to tune the split probabilities in practice? First, in the ideal scenario where we already know the strong features we can mimic the random mechanism by following the procedure detailed below. At each split

1. select at random, with replacement, M_n candidate coordinates to split on,
2. if the selection is all weak, then choose one at random to split on. If there is more than one strong variable elected, choose one at random and cut.

With this procedure, the (ideal) probability p_n^* to split on a strong variable is :

$$p_n^{*S} = \frac{1}{S}\left[1 - \left(1 - \frac{S}{d}\right)^{M_n}\right] \tag{8}$$

Proof. First, we introduce, or recall some notations:

- $\mathcal{M}_n$ is the subset composed of the M_n variables that have been chosen with replacement.
- Z denotes the random variable that takes the value $j \in \{1, \dots, d\}$ if the j-coordinate has been chosen.
- $\mathcal{S}$ denotes the subset of strong variables. Its cardinality is S.

Now we can compute

$$\mathbb{P}(Z \in \mathcal{S}) = 1 - \mathbb{P}(Z \notin \mathcal{S})$$

The event $\{Z \notin \mathcal{S}\}$ signifies that, we choose M_n times (independently) a weak feature. Mathematically it represents M_n successive failures in a Bernoulli setting. Since there are $d-S$ weak features, and a uniform probability to choose one feature, the probability to choose a weak feature is $\frac{d-S}{d} = 1 - \frac{S}{d}$. By independence,

$$\mathbb{P}(Z \notin \mathcal{S}) = \left(1 - \frac{S}{d}\right)^{M_n}.$$

Hence,

$$\mathbb{P}(Z \in \mathcal{S}) = 1 - \left(1 - \frac{S}{d}\right)^{M_n}.$$

And since, when $\mathcal{M}_n \cap \mathcal{S} \neq \emptyset$ and we choose uniformly at random a strong feature, when $\mathcal{M}_n \cap \mathcal{S} = \emptyset$ it is impossible to split on a strong variable :

$$p_n^{*\mathcal{S}} = \frac{1}{S}\left(1 - \left(1 - \frac{S}{d}\right)^{M_n}\right).$$

□

Because the sum of all splitting probabilities equals one, the probability to split on a weak feature is ideally equal to $p_n^{*\mathcal{W}} = \frac{1}{d}\left(1 - \frac{S}{d}\right)^{M_n}$. To track the strong features and make the random forest adaptive to sparsity, $p_n^{*\mathcal{W}}$ has to tend to 0 and therefore $M_n \to \infty$. Now, if use the condition of Corollary 4.4, i.e. $p_n^{*\mathcal{W}} log(n) \to 0$, then the condition on M_n reads as :

$$\frac{M_n}{\log(n)} \to \infty$$

However, we do not know the strong features beforehand. A solution proposed by [Bia12] is to rely on an independent sample $\mathcal{D}'_n$ with the sample points distributed as $(\mathbf{X}, Y)$ and independent of $(\mathbf{X}, Y)$ and $\mathcal{D}_n$ (which can be done by splitting the training set). In the linear regression framework,

$$Y = \sum_{j \in S} a_j X^{(j)} + \varepsilon,$$

(where the a_j are non-zero real numbers, and ε is a zero-mean random noise, which is assumed to be independent of $\mathbf{X}$ and with finite variance) [Bia12] derives a method to tune these probabilities based on the following result: the split on the j-th side which most (and strictly) decreases the weighted conditional variance $\mathbb{V}[Y|\mathbf{X}^{(j)} \in A_j]\mathbb{P}(\mathbf{X}^{(j)} \in A_j)$ (where $A = \prod_{j=1}^{d} A_j$ is a terminal node) is at the midpoint of the node A if $j \in \mathcal{S}$. On the contrary, such a point does not exist when $j \in \mathcal{W}$: every location of the cut on the j-th dimension reduces the variance equally. By using the following procedure, at each node of the tree:

1. Select at random, with replacement, M_n candidate coordinates to split on.

2. For each of the M_n elected coordinates, calculate the best split, that is, the split which most decreases the within-nodesum of squares on the second sample $\mathcal{D}_n$.

3. Select one variable at random among the coordinates which output the best within-node sum of squares decreases, and cut.

a probability of choosing the strong feature is obtained:

$$p_{nj} = \frac{1}{S}\left(1 - \left(1 - \frac{S}{d}\right)^{M_n}\right)(1 + \xi_{nj})$$

with $\xi_{nj} \to 0$, hence the expression (8) used before.

The interested reader might at this point want to know whether other types of random forest estimators posses this adaptive capacity. And yes, the most popular random forests method the CART-forests can adapt to sparsity. In [SBV15], the following result has been proved.

Theorem 4.5. *Assume that the response Y satisfies the assumption $\mathcal{H}$ of an additive regression framework seen above. Let $\xi > 0$. If there is no interval $[a, b]$ and no $j \in \{1, \dots, S\}$ such that r_j is constant on $[a, b]$. Then, with probability $1 - \xi$, for all n large enough, each split is performed on a strong feature.*

5 Conclusion

The aim of this article was to give hindsights to M2 students on some aspects of random forests. This course was mostly based [Bia12]. It provided some additional results such as a link between the risk of random forests estimator with infinitely many trees and finitely many trees, and generalisation/adaptation to other type of random forests estimators (such as uniform random forests). The main message is that those estimators are consistent and can adapt to sparsity under certain assumptions is the main message of this course.

From 2001, with their introduction by Leo Breiman, until today, random forests have been widely used in practice and achieve state-of-the-art performances. However, as for large neural networks, several undelying mechanisms enabling such results remain to be discovered.

References

[Bia12] Gerard Biau. "Analysis of a Random Forests Model". In: *Journal of Machine Learning Research* 13 (2012), pp. 1063–1095.

[Bre01] L Breiman. "Random Forests". In: *Machine Learning* 45 (Oct. 2001), pp. 5–32. DOI: 10.1023/A:1010950718922.

[BS15] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: *TEST* 25 (2015), pp. 197–227. URL: https://api.semanticscholar.org/CorpusID:14518730.

[Gyö+02] László Györfi et al. *A Distribution-Free Theory of Nonparametric Regression.* Springer series in statistics. Springer, 2002. ISBN: 978-0-387-95441-7.

[SBV15] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. "Consistency of random forests". In: *The Annals of Statistics* 43.4 (Aug. 2015). ISSN: 0090-5364. DOI: 10.1214/15-aos1321. URL: http://dx.doi.org/10.1214/15-AOS1321.

[Sco16] Erwan Scornet. "On the asymptotics of random forests". In: *Journal of Multivariate Analysis* 146 (2016), pp. 72–83. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2015.06.009.