Consistency of Random Forests

Project Report

Benjamin Moser

Abstract

Random Forest methods are some of the most widely used methods for regression and classification learning tasks, and their effectiveness has been proven empirically. However, the exact forces determining if, when, and why a Random Forest performs well are not yet completely clear. From a statistical point of view, a Random Forest is an estimator of some assumed true regression or classification function. A basic notion in statistics is that, as we observe more and more data samples, an estimate converges indeed to the estimated quantity. This property is called *consistency*. In this work, we review proofs of the consistency of Random Forests under different conditions as published in **Scornet2015**.

1 Overview

In this report, we review the main contributions of **Scornet2015**. We begin by recapitulating some basic definitions and introducing the regression learning task. We then proceed to define the Random Forest estimator and what it means for an estimator to be consistent. We review two theorems showing the consistency of Random Forests under different conditions. One proof is treated in full and repeated in a narrative manner, with a focus on providing intuition and motivating the imposed assumptions. For the other theorem we illustrate the overall proof structure and point out the basic ingredients.

2 Preliminaries

2.1 Probability & Statistics

Since we want to reason about an estimator independent of concrete data, we make statistical statements based on the in practice unknown probabilistic distribution of data. In this section, we fixate some basic language and notation to do that. The definitions are based on Wasserman 2010 and Györfi et al. 2002.

Definition 2.1. Probability Space A *probability space* is a triple $(\Omega, \Sigma, \mathbb{P})$ where

- Ω is an arbitrary set modelling the *sample space* i.e. the set of all possible outcomes.
- Σ is a σ-algebra over Ω, modelling the set of *events*.

- $\mathbb{P}$ is a function $\Sigma \to [0,1]$ such that $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for a countable collection of (pairwise disjoint) sets in Σ, and models the *probability measure*.

In the following we assume an underyling probability space implicitly.

Definition 2.2 (Random Variable)**.** A *random variable* is a quantity that depends on a random event, i.e. a function $\Omega \to M$ (commonly, we have $M = \mathbb{R}$).

Random variables are commonly denoted with uppercase letters. Random variables extend naturally to random vectors. Any function of a random variable is in turn a random variable.

Definition 2.3 (Convergence of Random Variables)**.** A sequence of random variables $X_n := \{X_1, \dots, X_n\}$ is said to *converge almost surely* (a.s.) towards random variable X iff

$$\mathbb{P}\left(\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\right) = 1$$

Since we want to make statements independent of concrete outcomes but rather about the general behaviour of random variables, we consider their expected value. For random variables with continuous range, this is based the probability density function.

Definition 2.4 (Probability Density Function)**.** The *probability density function f_X* of a random variable X is a nonnegative function such that

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x) dx$$

Definition 2.5 (Expected Value)**.** The *expected value* (expectation) of random variable X is defined as

$$\mathbb{E}[X] := \int x \cdot f_X(x)\, dx$$

where the integral is over the support of X. Note that a function $g(X)$ of a random variable X is a random variable itself and $\mathbb{E}[g(X)] = \int g(x) f_X(x)\, dx$. To emphasize the random variable the function is dependent on, we sometimes mention it in the subscript, e.g. $\mathbb{E}_X[g(X)]$.

Analogous to probabilities, probability densities and expectations can be conditioned on the outcome of another random variable. Note that $\mathbb{E}[X|Y = y]$ depends on a concrete observation to determine a value for Y, thus it is a random variable itself. We denote it the conditional expecation by $\mathbb{E}[X|Y]$ as a function of Y.

Let's collect some facts about expectations that will be useful to us.

Lemma For random variables X and Y:

- The expectation of the indicator function of an event and consequently of a random variable taking on some value is just the probability of that happening: $\mathbb{E}[\mathbb{1}_{X=x}] = \mathbb{P}(X = x)$
- We can relate a bounded random variable to its cumulative distribution function. For a random variable X with $X(\omega) < u \; \forall \omega \in \Omega$ and $\xi \in \mathbb{R}$, we have $\mathbb{E}[X] \leq \xi + u\mathbb{P}(X > \xi)$
- The rule of iterated expectations states that, for any function r: $\mathbb{E}[r(X,Y)] = \mathbb{E}\left[\mathbb{E}[r(X,Y)|X]\right]$ and, in particular, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$
- Jensen's Inequality: Let X be a random variable such that $\mathbb{P}(a \leq X \leq b) = 1$. If $g : \mathbb{R} \to \mathbb{R}$ is convex on $[a,b]$, then $g(\mathbb{E}\,[X]) \leq \mathbb{E}\,[g(X)]$.

- If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

The following quantities consider interactions between random variables.

Definition 2.6 (Variance, Covariance, Correlation)**.**

- *Variance*: $\mathrm{Var}(X) := \mathbb{E}[X - \mathbb{E}[X]]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- *Covariance*: $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- *Correlation*: $\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X,Y)}{(\mathrm{Var}(X)\mathrm{Var}(Y))^{1/2}}$

The definitions extend naturally to the conditional case.

2.2 Learning

Regression We consider the *regression* task, in which, given some input vector (*point*) $x \in [0, 1]^p$, we want to predict a response $y \in \mathbb{R}$. Generally, we consider input and response to be random variables X and Y. We want to find a function f such that $f(X)$ is an optimal approximation of Y. Note that $|f(X) - Y|$ is a random variable, thus we are interested in its expectation. To assess the quality of f, we use the L_2*-risk* of f given by $\mathbb{E}\left[|f(X) - Y|^2\right]$, where the square is motivated by mathematical and computational convenience. In fact, we can derive a L_2-optimal function explicitly. Let $m(x) := \mathbb{E}[Y|X = x]$ be the *regression function*. We often also refer to it as a random variable over the distribution of X, that is $m(X) := \mathbb{E}[Y|X]$. It can be shown that, for some candidate estimator f (see Györfi et al. 2002 sec. 1.1):

$$\mathbb{E}_{(X,Y)}\left[|f(X) - Y|^2\right] = \mathbb{E}_X\left[|f(X) - m(X)|^2\right] + \mathbb{E}_{(X,Y)}\left[|m(X) - Y|^2\right] \tag{1}$$

One can see that f is optimal if $f(x) = m(x)$.

In practice, we do not know the distributions of X and Y and hence cannot use m directly. However, we do have access to samples of their distributions, commonly also referred to as the *dataset* $D_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. We will consider D_n to be a random variable. Based on D_n, we may construct an *estimator of m*, which we will refer to as m_n. Note that $m_n(X)$ is a function of random variables X and D_n and in itself a random variable. We assume the samples in D_n to be independently and identically distributed and X and Y to be *i.i.d.* to any (X_i, Y_i) in D_n.

Similarly to (1), the L_2*-risk of the estimate* can be decomposed as

$$\mathbb{E}_{D_n}\left[\mathbb{E}_{(X,Y)}\left[|m_n(X) - Y|^2\right]\right] = \mathbb{E}_{X,D_n}\left[|m_n(X) - m(X)|^2\right] + \mathbb{E}_{(X,Y)}\left[|m(X) - Y|^2\right] \tag{2}$$

Therefore, we are interested in $\mathbb{E}\left[|m_n(X) - m(X)|^2\right]$ as a central quality measure that is independent of the concrete realisation of D_n (but note that it is not independent of n). This notion of error will be the basis for our definition the consistency of an estimator later.

Note that we can not directly compute such risks since they depend on the distributions of X and Y. As an approximation we consider the *empirical* L_2*-risk* of some function f defined as

$$\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 \approx \mathbb{E}_{(X,Y)}\left[|f(X) - Y|^2\right]$$

Minimising over the set of all functions will likely fit D_n perfectly and result in an estimator that is unlikely to generalise to the entire distribution of (X, Y). Due to this, we restrict ourselves to a *hypothesis class* of functions $\mathcal{F}_n$ and consider

$$m_n := \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2$$

Indeed, any learning method — such as Random Forests — characterise a hypothesis class that they are algorithmically able to cover. For Random Forests in particular, $\mathcal{F}_n$ will depend on n as well as the randomness that goes into constructing individual trees.

Estimation and approximation error The error of an estimator can be decomposed into different aspects. First, we may not have found the best possible choice in $\mathcal{F}_n$. Second, $\mathcal{F}_n$ itself may be too restrictive to allow a close estimate to the regression function m. Starting from eq. (2), one can write

$$\begin{aligned} \mathbb{E}\left[|m_n(X) - m(X)|^2\right] &= \mathbb{E}\left[|m_n(X) - Y|^2 \mid D_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] \\ &= \left(\mathbb{E}\left[|m_n(X) - Y|^2 \mid D_n\right] - \inf_{f \in \mathcal{F}_n} \mathbb{E}\left[|f(X) - Y|^2\right]\right) \\ &\quad + \left(\inf_{f \in \mathcal{F}_n} \mathbb{E}\left[|f(X) - Y|^2\right] - \mathbb{E}\left[|m(x) - Y|^2\right]\right) \end{aligned}$$

The first term is referred to as the *estimation error* and quantifies, in terms of error, the distance from the given estimator to the best possible estimator. The second term is the *approximation error* and quantifies the distance of the best possible estimator to the (true) regression function. We will be seeing very similar notions later on, and often such decompositions will be the first step towards showing consistency.

Bias-Variance-Covariance decomposition Alternatively, the risk can be divided into components of bias and variance (see Ueda and Nakano 1996). The *bias* measures how well the average estimate of the learning algorithm approximates the true function. The *variance* measures the variation of the estimate across different training sets of the same size. Assume there is a functional relationship between X_i and Y_i in the form of $Y_i = g(X_i) + \varepsilon$ where ε is noise with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 < \infty$. Then

$$\begin{aligned} \mathbb{E}\left[|m_n(X) - Y|^2\right] &= \mathbb{E}_X\left[\mathrm{Var}(m_n|X) + \mathrm{Bias}(m_n|X)^2\right] + \sigma^2 \\ \text{where} \qquad \mathrm{Var}(m_n|X) &:= \mathbb{E}_{D_n}\left[(m_n(X) - \mathbb{E}_{D_n}[m_n(X)])^2\right] \\ \mathrm{Bias}(m_n|X) &:= \mathbb{E}_{D_n}[m_n(X)] - g(X) \end{aligned}$$

Recall that Random Forests are ensembles of trees. When is an ensemble better than a single tree? Certainly, if the individual tree estimators are very similar, we can not expect to gain much over using just a single tree. The individual trees ought to be *diverse*. For ensemble methods that combine individual estimators $m_n^{(i)}$ for $i \in [M]$ into an ensemble estimator m_n by means of $m_n(X) = \sum_{i=1}^{M} w_i m_n^{(i)}(X)$ where the weights w_i are nonnegative and sum to one, the variance term can be decomposed further into the variance of individual estimators and the pairwise covariance between estimators.

$$\mathbb{E}\left[|m_n(X) - Y|^2\right] = \mathbb{E}_X\left[\frac{1}{M}\overline{\mathrm{Var}}(X) + \left(1 - \frac{1}{M}\right)\overline{\mathrm{Cov}}(X) + \overline{\mathrm{Bias}}(X)^2\right] + \sigma^2$$

$$\begin{aligned}
\text{where} \qquad \overline{\mathrm{Var}}(X) &:= \frac{1}{M}\sum_{i=1}^{M}\mathrm{Var}(m_n^{(i)}|X) \\
\overline{\mathrm{Cov}}(X) &:= \frac{1}{M(M-1)}\sum_{i\neq j}\mathrm{Cov}(m_n^{(i)}, m_n^{(j)}|X) \\
\mathrm{Cov}(m_n^{(i)}, m_n^{(i)}|X) &:= \mathbb{E}_{D_n}\left[\left(m_n^{(i)}(X) - \mathbb{E}_{D_n}\left[m_n^{(i)}(X)\right]\right)\left(m_n^{(j)}(X) - \mathbb{E}_{D_n}\left[m_n^{(j)}(X)\right]\right)\right] \\
\overline{\mathrm{Bias}}(X) &:= \frac{1}{M}\sum_{i=1}^{M}\mathrm{Bias}(m_n^{(i)}|X)
\end{aligned}$$

This tells us that while combining multiple estimators may mitigate individual variances, a good ensemble must be also be one in which the individual estimators are not too similar. In particular, their errors should exhibit small (absolute) covariance. Random Forests establish diversity implicitly via the randomness that goes into constructing individual trees, namely the subsampling prior to the tree construction and the random selection of candidate split dimensions.

2.3 Random Forests

A Random Forest is an ensemble of M randomized regression trees. The randomness that goes into constructing each tree is captured in identically and independently distributed random variables $\Theta_1, ..., \Theta_M$. The k-th tree is constructed as follows.

1. *Initialisation:* Sample a_n points from the dataset D_n at random based on Θ_k without replacement. These samples define the root node of the tree.
2. *Recursion:* Split a node A, dividing the points in A into two child nodes. A *cut* is a pair (j, z) where $j \in [p]$ indexes the dimension in which to perform the split and z the threshold at which to split. We will split at $(j^\star, z^\star) = \arg\max_{j\in\mathcal{M}_{\mathrm{try}},(j,z)\in C_A} \mathrm{L}(j, z)$ where L is the CART split criterion (see def. 2.7).
3. *Termination:* Stop as soon as the total number of leaf nodes reaches some given t_n (i.e. in case of $t_n = a_n$, each leaf node contains exactly one point).

Definition 2.7 (CART split criterion)**.** Let C_A be the set of possible cuts in A. Let $\mathcal{M}_{try}$ be a set of possible dimensions, chosen at random based on Θ_j. The *CART Criterion* measures the difference in the (renormalised variance) before and after a given split.

$$\begin{aligned}
L_n(j, z) &:= \mathrm{Var}_{\mathrm{before}} - \mathrm{Var}_{\mathrm{after}} \\
\text{for} \quad \mathrm{Var}_{\mathrm{before}} &= \frac{1}{N_n(A)}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i\in A} \\
\mathrm{Var}_{\mathrm{after}} &= \frac{1}{N_n(A)}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_{A_L}\mathbb{1}_{\mathbf{x}_i^{(j)}<z} - \bar{Y}_{A_R}\mathbb{1}_{\mathbf{x}_i^{(j)}\geq z}\right)^2 \mathbb{1}_{\mathbf{X}_i\in A}
\end{aligned}$$

We refer to Biau and Scornet 2016 for a comprehensive formulation of the algorithm. Each level of the tree induced by Θ_j corresponds exactly to a partition of the entire data space. Each leaf of the tree corresponds to a cell in the partition. Let $A_n(x, \Theta_j)$ denote the cell of x in the tree generated by Θ_j. We say that two points are *connected* in a tree if they are in the same cell.

Definition 2.8 (Tree and Forest estimates)**.** The individual tree estimates are aggregated to form the final Random Forest estimate.

- Given a query point x, the *estimate of the j-th tree*, denoted by $m_n(x, \Theta_j, D_n)$ is the mean response of the cell of x.
- The *finite forest estimate* is the average over all individual tree estimates, that is

$$m_{M,n}(x, \Theta_{1..M}, D_n) := \frac{1}{M} \sum_{j=1}^{M} m_n(x, \Theta_j, D_n)$$

- In practice, M can be chosen arbitrarily large and indeed the finite forest estimate converges almost surely to its expectation as the number of trees M grows (see Biau and Scornet 2016). Due to this, in the following, we will consider the *infinite forest estimate* $m_n(x) = \mathbb{E}_\Theta\left[m_n(x, \Theta, D_n)\right]$.

2.4 Consistency

As established earlier, a good estimate is one with small L_2-error. In many methods, the quality of the estimate m_n depends on the dataset, in particular its size n. As such, we are interested in the question whether, as n grows, the estimate converges indeed to the optimal regression function m. This is equivalent to saying that the L_2-error of the estimate vanishes as n grows to infinity. This property is called *consistency*.

Definition 2.9 (Consistency)**.** A sequence of regression function estimates $\{m_n\}_n$ is called (strongly universally) *consistent* if, for all possible distributions of X and Y, the L_2-error converges almost surely to 0, that is

$$\mathbb{E}[m_n(X) - m(X)]^2 \underset{a.s.}{\rightarrow} 0$$

where the expectation is with respect to the query point X and the samples D_n.

3 Consistency of Random Forests

We consider two different settings. In the setting of Theorem 1, cells may contain more than one point. Theorem 2 works in the setting of fully-grown trees, that is, each cell contains exactly one point.

Hypothesis 1 Throughout, we will assume that the response follows an *additive model*, i.e.

$$Y = \sum_{j=1}^{p} m_j\left(\mathbf{X}^{(j)}\right) + \varepsilon$$

where $\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)}\right)$ is uniformly distributed over $[0,1]^p$ and each m_j is continuous. ε is an independent, centered Gaussian noise.

Hypothesis 1 (H1) will be used mostly to relate the estimate to the true response via a noise term and consequently exploit the fact that the noise ε is assumed to satisfy $\mathbb{E}[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma^2$. We could, in fact, also simply work with bounded noise with zero mean.

An essential tool for both theorems is Proposition 2, which states that the variation of m within a cell goes to 0 as n grows. This will be used to address the estimation error.

Definition 3.1 (Variation)**.** The *variation of m within cell A* is given as

$$\Delta(m, A) := \sup_{\mathbf{x},\mathbf{x}' \in A} \left| m(\mathbf{x}) - m\left(\mathbf{x}'\right) \right|$$

Proposition 2 (Scornet, Biau, and Vert 2015) Let $A_n(X, \Theta)$ be the cell that contains X in a tree generated via Θ. Assume that (H1) holds. Then, for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}$ such that, for all $n > N$,

$$\mathbb{P}\left[\Delta\left(m, A_n(X, \Theta)\right) \leq \xi\right] \geq 1 - \rho$$

Intuitively, this means that the variation of m is small, provided that n is large enough.

3.1 Fully-grown trees (Theorem 2)

Since, by definition, the tree estimate for a query point x is the mean response of the cell of x, We can express the forest estimate aswell by a local averaging formulation. Starting from the finite forest estimate, we have (see also Biau and Scornet 2016)

$$\begin{aligned} m_{M,n}(x, \Theta_{1..M}) &= \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{n} \frac{\mathbb{1}_{x_i \in A_n(x, \Theta_j)}}{||A_n(x, \Theta_j)||} Y_i \\ &= \sum_{i=1}^{n} W_{M,ni}(x) Y_i \qquad \text{for } W_{M,ni}(x) := \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{x_i \in A_n(x, \Theta_j)}}{||A_n(x, \Theta_j)||} \end{aligned}$$

In Theorem 2, we assume that each cell contains exactly one point. Thus, motivated by the above discussion, the infinite forest estimate can be written as

$$m_n(X) = \sum_{i=1}^{n} W_{ni}(X) Y_i \qquad \text{for} \quad W_{ni}(X) := \mathbb{E}_\Theta \left[\mathbb{1}_{X_i \in A_n(X, \Theta)}\right]$$

In both cases, the weights sum to one. Note that W_{ni} is the expectation over tree randomness Θ and nothing else but the probability that x_i and x are connected in a tree generated with Θ. The error of the estimate can be bounded from above by

$$\mathbb{E}\left[m_n(X) - m(X)\right]^2 \leq 2\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(Y_i - m(X_i))\right]^2 + 2\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(m(X_i) - m(X))\right]^2 \quad (3)$$

The first term on the right-hand-side can be considered a kind of approximation error and the second term a kind of estimation error. (since $\sum_{i=1}^{n} W_{ni}(X) m(X)$ is the best possible estimate) To show consistency, we will show the convergence of each of the two terms separately.

Approximation error Let us first turn towards the second term. Note that it depends on the difference between the responses for two points. Indeed, we only need to consider situations in which the two points are in the same cell, enabling us to apply Proposition 2, which bounds the variation of m within a cell. Using Jensen's inequality and the definition of Δ, it holds that

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(m(X_i) - m(X))\right]^2 \leq \mathbb{E}\left[\mathbb{1}_{X_i \in A_n(x,\Theta)}(m(X_i) - m(X))^2\right] \leq \mathbb{E}\left[\Delta^2(m, A_n(x,\Theta)\right]$$

which, according to Proposition 2 is bounded from above by $\alpha(4||m||_\infty + 1)$ for sufficiently large n, where $\alpha > 0$ is arbitrary.

Estimation error Let us now rest our attention on the first term in (3). We denote the noise as $\varepsilon_i = Y_i - m(X_i)$ and the indicator that points are connected to the query point in some tree as $Z_i := \mathbb{1}_{X_i \in A_n(X,\Theta)}$ (resp. $Z'_j := \mathbb{1}_{X_i \in A_n(X,\Theta')}$). It holds that

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\varepsilon_i\right]^2 = \mathbb{E}\left[\sum_{i,j=1}^{n} W_{ni}(X)W_{nj}(X)\varepsilon_i\varepsilon_j\right] = \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)^2\varepsilon_i^2\right] + \mathbb{E}\left[\sum_{i \neq j} Z_i Z'_j \varepsilon_i \varepsilon_j\right]$$

Note that the left term considers single trees, whereas the right term considers interactions between two trees generated with Θ and Θ', respectively.

Single Trees To find a vanishing upper bound for the left term, notice that we have assumed the noise ε_i to be bounded in (H1). We can bound the probability of connection $W_{ni}(X)$ with a simple counting argument. Informally, since

$$\begin{aligned} W_{ni}(X) = \mathbb{E}\left[Z_i\right] &= \mathbb{P}(X, X_i \text{ selected in subsampling step} \wedge X, X_i \text{ connected in tree generated via } \Theta) \\ &\leq \mathbb{P}(X, X_i \text{ selected in subsampling step}) \\ &= \frac{\binom{a_n - 1}{n_1}}{\binom{a_n}{n}} = \frac{a_n}{n} \end{aligned}$$

Putting this together, assuming that $\frac{a_n}{n} \to 0$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\varepsilon_i^2\right] \leq \mathbb{E}\left[\max_{l \in [n]} W_{nl}(X) \max_{i \in [n]} \varepsilon_i^2\right] \leq \frac{a_n}{n} C \to 0$$

Tree interactions There are two different assumptions (H2.1) and (H2.2) that we can make that will allow us to bound the tree interaction term. Here, we wil only treat the proof based on (H2.2). By the rule of iterated expectation, we have

$$\mathbb{E}\left[\sum_{i \neq j} Z_i Z'_j \varepsilon_i \varepsilon_j\right] = \mathbb{E}\left[\sum_{i \neq j} Z_i Z'_j \varepsilon_i \mathbb{E}\left[\varepsilon_j \mid X_i, X_j, Z_i, Z'_j, Y_i\right]\right] \tag{4}$$

We will handle the inner and outer expectations separately. To assert their boundedness, we will have to impose two assumptions (H2.2a and H2.2b).

Inner expectation In general, for random variables A, B, C (where B is discrete), it holds that

$$\mathbb{E}\left[A \mid B, C\right] = \sum_{b \in \text{range}(B)} \frac{\mathbb{E}\left[A \mathbb{1}_{B=b} \mid C\right]}{\mathbb{P}(B = b \mid C)} \mathbb{1}_{B=b}$$

Applied to eq. (4), this yields

$$\mathbb{E}\left[\varepsilon_j \mid X_i, X_j, Z_i, Z'_j, Y_i\right] = \sum_{k \in \{0,1\}^2} \frac{\mathbb{E}\left[\varepsilon_i \mathbb{1}_{(Z_i, Z'_j)=k} \mid X_i, X_j, Y_i\right]}{\mathbb{P}((Z_i, Z'_j) = k \mid X_i, X_j, Y_i)} \mathbb{1}_{(Z_i, Z'_j)=k} \tag{5}$$

This is useful since, in fact, $\mathbb{E}\left[\varepsilon_i \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_i\right] = \text{Cov}(\varepsilon_1, \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j)$ due to that $\mathbb{E}\left[\varepsilon_i \mid X_i, X_j, Y_j\right] = 0$ (by H1). Note that, if ε_i and (Z_i, Z'_j) are independent, this term and any covariance or correlation based on it will be zero. However, in Random Forests, the partitions – and thus (Z_i, Z'_j) – depend on the responses Y_i due to optimising for purity in the CART Criterion during tree construction. Consequently, we will formulate an assumption that will express the boundedness of this correlation, and consequently the expectation. Based on the definition of correlation, we obtain

$$\text{Corr}(\varepsilon_i, \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j) = \frac{\text{Cov}(\varepsilon_i, \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j)}{\text{Var}^{1/2}(\varepsilon_i \mid X_i, X_j, Y_j)\text{Var}^{1/2}(\mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j)}$$

where $\text{Var}^{1/2}(\varepsilon_i \mid X_i, X_j, Y_j) = \sigma$ and $\text{Var}^{1/2}(\mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j) \leq \mathbb{P}((Z_i, Z_j) = k)$.

Consequently,

$$\mathbb{E}\left[\varepsilon_i \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j\right] \leq \text{Corr}(\varepsilon_i, \mathbb{1}_{(Z_i,Z'_j)=k \mid X_i, X_j, Y_j})\mathbb{P}^{1/2}((Z_i, Z_j) = k)\sigma$$

which, applied to (5), yields

$$\mathbb{E}\left[\varepsilon_j \mid X_i, X_j, Z_i, Z'_j, Y_i\right] \leq \sigma \cdot 4 \max_{k \in \{0,1\}^2} \frac{|\text{Corr}(\varepsilon_i, \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j)|}{\mathbb{P}^{1/2}((Z_i, Z'_j) = k)}$$

This motivates assumption (H2.2a), stated as

$$\max_{k \in \{0,1\}^2} \frac{|\text{Corr}(\varepsilon_i, \mathbb{1}_{(Z_i,Z'_j)=k} \mid X_i, X_j, Y_j)|}{\mathbb{P}^{1/2}((Z_i, Z'_j) = k)} \leq \gamma_n$$

for sufficiently large n and a sequence $\{\gamma_n\}_n \to 0$. In summary, this means that (H2.2a) implies that, almost surely

$$\mathbb{E}\left[\varepsilon_i \mid Z_i, Z'_j, X_i, X_j, Y_j\right] \leq 4\sigma\gamma_n$$

Outer Expectation If (H2.2a) is given, we have, by linearity of expectation

$$\mathbb{E}\left[\sum_{i \neq j} Z_i Z'_j \varepsilon_i \varepsilon_j\right] = \mathbb{E}\left[\sum_{i \neq j} Z_i Z'_j \varepsilon_i \mathbb{E}\left[\varepsilon_j \mid X_i, X_j, Z_i, Z'_j, Y_i\right]\right] \leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[Z_i \varepsilon_i\right] \tag{6}$$

This is where we extract another assumption. We use the law of iterated expectation to factor out an expection of the noise term:

$$\cdots \leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[Z_i \mathbb{E}^{1/2}\left[|\varepsilon_i|^2 \mid X_i, Z_i\right]\right]$$

Similar to previously, one can then derive assumption (H2.2b) and show that for some constant $C > 0$

$$\max_{k\in\{0,1\}} \frac{|\mathrm{Corr}(\varepsilon_i^2, \mathbb{1}_{Z_i=k} \mid X_i)|}{\mathbb{P}^{1/2}(Z_i = k \mid X_i)} \leq C \quad \Rightarrow \qquad \mathbb{E}\left[|\varepsilon_i|^2 \mid X_i, Z_i\right] \leq 4C\sigma^2$$

Assuming this holds, we obtain, continuing from eq. (6)

$$\gamma_n \sum_{i=1}^{n} \mathbb{E}\left[Z_i \mathbb{E}^{1/2}\left[|\varepsilon_i|^2 \mid X_i, Z_i\right]\right] \leq \gamma_n 2C^{1/2}\sigma \sum_{i=1}^{n} \mathbb{E}\left[Z_i\right] \leq \gamma_n 2C^{1/2}\sigma$$

Which concludes the proof of (H2.2) ⇒ Theorem 2. In summary, the full statement is given below.

Hypothesis 2.2 There exist a constant $C > 0$ and a sequence $(\gamma_n)_n \to 0$ such that, almost surely, the following two statements hold:

$$\max_{\ell_1,\ell_2=0,1} \frac{\left|\mathrm{Corr}\left(Y_i - m\left(\mathbf{X}_i\right), \mathbb{1}_{Z_{i,j}=(\ell_1,\ell_2)} \mid \mathbf{X}_i, \mathbf{X}_j, Y_j\right)\right|}{\mathbb{P}^{1/2}\left[Z_{i,j} = (\ell_1, \ell_2) \mid \mathbf{X}_i, \mathbf{X}_j, Y_j\right]} \leq \gamma_n \tag{H2.2a}$$

$$\max_{\ell_1=0,1} \frac{\left|\mathrm{Corr}\left(\left(Y_i - m\left(\mathbf{X}_i\right)\right)^2, \mathbb{1}_{Z_i=\ell_1} \mid \mathbf{X}_i\right)\right|}{\mathbb{P}^{1/2}\left[Z_i = \ell_1 \mid \mathbf{X}_i\right]} \leq C. \tag{H2.2b}$$

Theorem 2 (Scornet, Biau, and Vert 2015) Assume that (H1) and (H2.2) are satisfied, and let $t_n = a_n$. Then, provided $a_n \to \infty$, $t_n \to \infty$ and $a_n \log n/n \to 0$, random forests are consistent, that is,

$$\lim_{n\to\infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0$$

Let us briefly discuss an alternate set of assumptions. Again, it comes into action when handling the tree interaction term $\mathbb{E}\left[\sum_{i\neq j} Z_i Z_j' \varepsilon_i \varepsilon_j\right]$. Recall that $Z_i = 1$ if and only if X_i is connected to the query point X in a tree generated via Θ. We bring in the additional observation that, since each cell contains exactly one point, x_i can only be connected to the query point x if there are no other points in the hyperrectangle defined by x_i and x. In this case, we call x_i and x *layered nearest neighbours* (LNNs). Previous work gives us a bound on the expected number of LNNs of a random query point, denoted by $l_{a_n}(X)$ (see Scornet, Biau, and Vert 2015). Let L_i be the indicator of the event that X_i is an LNN to X in a tree generated via Θ and likewise for L_j'. Let $\psi_{i,j} = \mathbb{E}\left[Z_i Z_j' \mid \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n\right]$ and $\psi_{i,j}\left(Y_i, Y_j\right) = \mathbb{E}\left[Z_i Z_j' \mid \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i, Y_j\right]$. One can show that

$$\begin{aligned}\mathbb{E}\left[\sum_{i\neq j} Z_i, Z_j' \varepsilon_i \varepsilon_j\right] &\leq \mathbb{E}\left[\sum_{i\neq j} |\varepsilon_i|\, |\varepsilon_j|\, L_i L_j' \,|\psi_{ij}(Y_i, Y_j) - \psi_{ij}|\right] \\ &\leq \mathbb{E}\left[\max_{i\in[n]} |\varepsilon_i|^2 \max_{i\neq j} |\psi_{ij}(Y_i, Y_j) - \psi_{ij}| \sum_{i\neq j} L_i L_j'\right]\end{aligned}$$

The first factor vanishes since by (H1) noise is assumed to be bounded. The third factor is bounded by the total number of LNNs l_{a_n}, seeing that $\sum_{i\neq j} L_i, L_j' \leq l_{a_n}^2(X)$. The second factor is assumed to vanish by means of (H2.1).

Hypothesis 2.1 If ε is a bounded random variable, (H2.1) can be stated as

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{\substack{i,j \\ i\neq j}} \left|\psi_{i,j}\left(Y_i, Y_j\right) - \psi_{i,j}\right|\right]^2 = 0$$

If we assume Gaussian noise, the above term requires an extra factor of $(loga_n)^{2p-2}(\log n)^2$.

3.2 Trees not fully grown (Theorem 1)

Theorem 1 acts in the setting of leaves containing more than one point.

Theorem 1 (Scornet, Biau, and Vert 2015) Assume that (H1) is satisfied. Then, provided $a_n \to \infty, t_n \to \infty$ and $t_n \left(\log a_n\right)^9 / a_n \to 0$, random forests are consistent, that is,

$$\lim_{n\to\infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0$$

Figure 1 illustrates the individual statements that yield Theorem 1. The *theoretical Random Forest* is one built by optimizing a CART criterion that is not based on empirical variance but instead on the general definition of variance in the limit. Trees in a theoretical Random Forest do not depend on a given dataset D_n, but are still random due to the random selection of candidate split dimensions. A theoretical random forest is assumed to be of fixed depth.

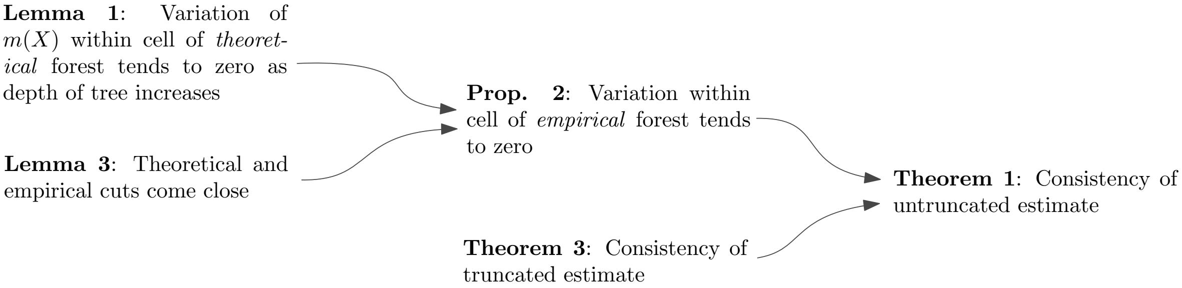
Lemma 1: Variation of $m(X)$ within cell of *theoretical* forest tends to zero as depth of tree increases
Lemma 3: Theoretical and empirical cuts come close
Prop. 2: Variation within cell of *empirical* forest tends to zero
Theorem 3: Consistency of truncated estimate
Theorem 1: Consistency of untruncated estimate

Figure 1: Overview of the proof of Theorem 1 (omitting Technical Lemma 1 and Lemma 2).

We refrain from illustrating the entire proof of Theorem 1 but nevertheless attempt to shed light on the overall strategy and some key ideas. Recall that in finding a regression function estimate, we restrict ourselves to a hypothesis class $\mathcal{F}_n$. The following Lemma gives a bound on the L_2-risk of the estimate in terms of estimation and approximation error with respect to $\mathcal{F}_n$.

Lemma 10.1 (Györfi et al. 2002) Let $\mathcal{F}_n$ be a class of functions $f : \mathbb{R}^p \to \mathbb{R}$ depending on the data D_n. Then,

$$\begin{aligned} \mathbb{E}\left[\left|m_n(x) - m(x)\right|^2\right] &\leq 2 \sup_{f\in\mathcal{F}_n} \left| \frac{1}{n}\sum_{j=1}^{n} \left|f\left(X_j\right) - Y_j\right|^2 - \mathbb{E}\left[(f(X) - Y)\right]^2 \right| \\ &+ \inf_{f\in\mathcal{F}_n} \mathbb{E}\left[\left|f(x) - m(x)\right|^2\right] \end{aligned}$$

Estimation error The first term quantifies the difference between the empirical L_2-risk and the true L_2-risk. This is reminiscent of the notion of estimation error. The bound for this term rests on theory developed in Györfi et al. 2002. To enable this, we have to take a detour via first considering the truncated estimate. Luckily, it turns out that we can conclude the consistency of the untruncated estimate from there. To bound the estimation error, we relate it to characteristics of the induced partition (Györfi et al. 2002 Thm 9.1). These characteristics are indirectly controlled by the rates required by the theorem.

Approximation error The second term is the classical approximation error. To handle this term, the basic idea is that any Random Forest regression function will be cell-wise constant. So, the approximation error will compare the difference of the cell-wise estimate with the true estimate at a given point. This is bounded by the variation of m, enabling the application of Proposition 2.

References

Wasserman, Larry (2010). *All of Statistics: A Concise Course in Statistical Inference*. Corr. 2. print., [repr.] Springer Texts in Statistics. New York Berlin Heidelberg: Springer. 442 pp. ISBN: 978-0-387-21736-9 978-1-4419-2322-6.

Györfi, László et al. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-95441-7 978-0-387-22442-8. DOI: `10.1007/b97848`.

Ueda, N. and R. Nakano (1996). "Generalization Error of Ensemble Estimators". In: *Proceedings of International Conference on Neural Networks (ICNN'96)*. International Conference on Neural Networks (ICNN'96). Vol. 1. Washington, DC, USA: IEEE, pp. 90–95. ISBN: 978-0-7803-3210-2. DOI: `10.1109/ICNN.1996.548872`.

Biau, Gérard and Erwan Scornet (June 1, 2016). "A Random Forest Guided Tour". In: *TEST* 25.2, pp. 197–227. ISSN: 1863-8260. DOI: `10.1007/s11749-016-0481-7`.

Scornet, Erwan, Gérard Biau, and Jean-Philippe Vert (Aug. 1, 2015). "Consistency of Random Forests". In: *The Annals of Statistics* 43.4. ISSN: 0090-5364. DOI: `10.1214/15-AOS1321`.