Consistency of Random Forests

Erwan Scornet, Gérard Biau, Jean-Philippe Vert

Consistency of Random Forests

Erwan Scornet
Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France
erwan.scornet@upmc.fr

Gérard Biau[1]
Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France
& Institut universitaire de France
gerard.biau@upmc.fr

Jean-Philippe Vert
Centre for Computational Biology, Mines ParisTech, Fontainebleau, F-77300, France
& Institut Curie, Paris, F-75248, France
& U900, INSERM, Paris, F-75248, France
jean-philippe.vert@mines-paristech.fr

Abstract

Random forests are a learning algorithm proposed by Breiman (2001) which combines several randomized decision trees and aggregates their predictions by averaging. Despite its wide usage and outstanding practical performance, little is known about the mathematical properties of the procedure. This disparity between theory and practice originates in the difficulty to simultaneously analyze both the randomization process and the highly data-dependent tree structure. In the present paper, we take a step forward in forest exploration by proving a consistency result for Breiman's (2001) original algorithm in the context of additive regression models. Our analysis also sheds an interesting light on how random forests can nicely adapt to sparsity in high-dimensional settings.

Index Terms — Random forests, randomization, consistency, additive model, sparsity, dimension reduction.

2010 Mathematics Subject Classification: 62G05, 62G20.

[1]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

1 Introduction

Random forests are an ensemble learning method for classification and regression that constructs a number of randomized decision trees during the training phase and predicts by averaging the results. Since its publication in the seminal paper of Breiman (2001), the procedure has become a major data analysis tool, that performs well in practice in comparison with many standard methods. What has greatly contributed to the popularity of forests is the fact that they can be applied to a wide range of prediction problems and have few parameters to tune. Aside from being simple to use, the method is generally recognized for its accuracy and its ability to deal with small sample sizes, high-dimensional feature spaces, and complex data structures. The random forest methodology has been successfully involved in many practical problems, including air quality prediction (winning code of the EMC data science global hackathon in 2012, see `http://www.kaggle.com/c/dsg-hackathon`), chemoinformatics (Svetnik et al., 2003), ecology (Prasad et al., 2006; Cutler et al., 2007), 3D object recognition (Shotton et al., 2011), and bioinformatics (Díaz-Uriarte and de Andrés, 2006), just to name a few. In addition, many variations on the original algorithm have been proposed to improve the calculation time while maintaining good prediction accuracy (see, e.g., Geurts et al., 2006; Amaratunga et al., 2008). Breiman's forests have also been extended to quantile estimation (Meinshausen, 2006), survival analysis (Ishwaran et al., 2008), and ranking prediction (Clémençon et al., 2013).

On the theoretical side, the story is less conclusive and, regardless of their extensive use in practical settings, little is known about the mathematical properties of random forests. To date, most studies have concentrated on isolated parts or simplified versions of the procedure. The most celebrated theoretical result is that of Breiman (2001), which offers an upper bound on the generalization error of forests in terms of correlation and strength of the individual trees. This was followed by a technical note (Breiman, 2004), that focuses on a stylized version of the original algorithm. A critical step was subsequently taken by Lin and Jeon (2006), who established an interesting connection between random forests and a particular class of nearest neighbor predictors, further explored by Biau and Devroye (2010). In recent years, various theoretical studies (e.g., Biau et al., 2008; Ishwaran and Kogalur, 2010; Biau, 2012; Genuer, 2012; Zhu et al., 2012) have been performed, analyzing consistency of simplified models, and moving ever closer to practice. A recent attempt towards narrowing the gap between theory and practice is by Denil et al. (2013), where the first consistency result for

online random forests is presented.

The difficulty to properly analyze random forests can be explained by the black-box nature of the procedure, which is actually a subtle combination of different components it is illusory to analyze separately. Among the forest essential ingredients, both bagging (Breiman, 1996) and the Classification And Regression Trees (CART)-split criterion (Breiman et al., 1984) play a critical role. Bagging (a contraction of bootstrap-aggregating) is a general aggregation scheme which proceeds by generating subsamples from the original data set, constructing a predictor from each resample, and deciding by averaging. It is one of the most effective computationally intensive procedures to improve on unstable estimates, especially for large, high-dimensional data sets where finding a good model in one step is impossible because of the complexity and scale of the problem (Bühlmann and Yu, 2002; Kleiner et al., 2012; Wager et al., 2013). On the other hand, the CART-split selection, originated from the most influential CART algorithm of Breiman et al. (1984), is used in the construction of the individual trees to choose the best cuts perpendicular to the axes. At each node of each tree, the best cut is selected by optimizing the CART-split criterion, based on the notion of Gini impurity (classification) and prediction squared error (regression).

Yet, while bagging and the CART-splitting scheme play a key role in the random forest mechanism, both are difficult to analyze, thereby explaining why theoretical studies have considered so far simplified versions of the original procedure. This is often done by simply ignoring the bagging step and by replacing the CART-split selection by a more elementary cut protocol. Besides, in Breiman's forests, each leaf (that is, a terminal node) of the individual trees contains a fixed pre-specified number of observations (usually small). Since this feature is hardly amenable to a rigorous mathematical investigation, most authors focus on a simplified, data-independent, stopping criterion. All in all, in these toy models, the forest construction is independent of the data, thus creating a gap between theory and practice.

Motivated by the above discussion, we study in the present paper some asymptotic properties of Breiman's (2001) algorithm in the context of additive regression models. We prove the $\mathbb{L}^2$ consistency of random forests, which gives a first basic theoretical guarantee of efficiency for this algorithm. Up to our knowledge, this is the first consistency result for Breiman's (2001) original procedure, since most of the previous studies focused on data-independent splitting criteria and cells containing a number of points growing to infinity with the sample size. Our approach rests upon a detailed analysis of the behavior of the cells generated by CART-split selection as the sample size

grows. In fact, a good control of the regression function variation inside each cell, together with a proper choice of the resampling rate in bagging are sufficient to ensure the forest consistency in a $\mathbb{L}^2$ sense. It also turns out that our analysis has interesting consequences for the understanding of the forest behavior in a sparse framework, that is, when the ambient dimension p is large but only a smaller number of coordinates carry out information.

The paper is organized as follows. In Section 2, we introduce some notation and describe the random forest method. The main asymptotic results are presented in Section 3 and further discussed in Section 4. Section 5 is devoted to the main proofs, and technical results are postponed to Section 6.

2 Random forests

The general framework is that of $\mathbb{L}^2$ regression estimation, in which an input random vector $\mathbf{X} \in [0,1]^p$ is observed, and the goal is to predict the square integrable random response $Y \in \mathbb{R}$ by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. To this aim, we assume given a training sample $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $[0,1]^p \times \mathbb{R}$-valued independent random variables distributed as the prototype pair $(\mathbf{X}, Y)$. The objective is to use the data set $\mathcal{D}_n$ to construct an estimate $m_n : [0,1]^p \to \mathbb{R}$ of the function m. In this respect, we say that a regression function estimate m_n is $\mathbb{L}^2$ consistent if $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2 \to 0$ as $n \to \infty$ (where the expectation is over $\mathbf{X}$ and $\mathcal{D}_n$).

A random forest is a predictor consisting of a collection of M randomized regression trees. For the m-th tree in the family, the predicted value at the query point $\mathbf{x}$ is denoted by $m_n(\mathbf{x}; \Theta_m, \mathcal{D}_n)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed as a generic random variable Θ and independent of the sample $\mathcal{D}_n$. In practice, this variable is used to subsample the training set prior to the growing of individual trees and to select the successive candidate directions for splitting. The trees are combined to form the (finite) forest estimate

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}; \Theta_m, \mathcal{D}_n). \tag{1}$$

Since in practice we can choose M as large as possible, we study in this paper the property of the infinite forest estimate obtained as the limit of the finite forest estimate when the number of trees M grows to infinity:

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

where $\mathbb{E}_\Theta$ denotes expectation with respect to the random parameter Θ, conditionally on $\mathcal{D}_n$. This operation is justified by the law of large numbers, which asserts that, almost surely, conditionally on $\mathcal{D}_n$,

$$\lim_{M\to\infty} m_{n,M}(\mathbf{x};\Theta_1,\ldots,\Theta_M,\mathcal{D}_n) = m_n(\mathbf{x};\mathcal{D}_n),$$

(see the appendix in Breiman, 2001, for details). In the sequel, to lighten notation, we simply write $m_n(\mathbf{x})$ instead of $m_n(\mathbf{x};\,\mathcal{D}_n)$.

In Breiman's (2001) original forests, each node of a single tree is associated with a hyper-rectangular cell. At each step of the tree construction, the collection of cells forms a partition of $[0,1]^p$. The root of the tree is $[0,1]^p$ itself, and each tree is grown as follows:

Algorithm 1: Breiman random forest predicted value at $\mathbf{x}$.

Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1,\ldots,n\}$, $m_{\text{try}} \in \{1,\ldots,p\}$, and $\mathbf{x} \in [0,1]^p$.

Output: Prediction of the random forest at $\mathbf{x}$

for $j = 1,\ldots,M$ **do**
  Select a_n points, without replacement, uniformly in $\mathcal{D}_n$.
  Set $\mathcal{P} = \{[0,1]^p\}$ the partition associated with the root of the tree.
  while *there exists* $A \in \mathcal{P}$ *that contains strictly more than one point* **do**
    Select uniformly, without replacement, a subset $\mathcal{M}_{\text{try}} \subset \{1,\ldots,p\}$ of cardinality m_{try}
    Select the best split in A by optimizing the CART-split criterion along the coordinates in $\mathcal{M}_{\text{try}}$ *(see details below)*.
    Cut the cell A according to the best split. Call A_L and A_R the two resulting cell.
    Set $\mathcal{P} \leftarrow (\mathcal{P}\backslash\{A\}) \cup \{A_L\} \cup \{A_R\}$.
  end
  Compute the predicted value $m_n(\mathbf{x};\Theta_j,\mathcal{D}_n)$ at $\mathbf{x}$ equal to the only Y_i falling in the cell of $\mathbf{x}$ in partition $\mathcal{P}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x};\Theta_1,\ldots,\Theta_M,\mathcal{D}_n)$ at the query point $\mathbf{x}$ according to (1).

So far, we have not made explicit the CART-split criterion used in **Algorithm 1**. To properly define it, we let A be a generic cell and $N_n(A)$ be the number of data points falling in A. A cut in A is a pair (j,z), where j is a dimension in $\{1,\ldots,p\}$ and z is the position of the cut along the j-th coordinate, within the limits of A. We let $\mathcal{C}_A$ be the set of all such possible

cuts in A. Then, with the notation $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \ldots, \mathbf{X}_i^{(p)})$, for any $(j, z) \in \mathcal{C}_A$, the CART-split criterion (Breiman et al., 1984) takes the form

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

where $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$ 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), with the convention $0/0 = 0$. At each cell A, the best cut $(j_n^\star, z_n^\star)$ is finally selected by maximizing $L_n(j, z)$ over $\mathcal{M}_{\text{try}}$ and $\mathcal{C}_A$, that is

$$(j_n^\star, z_n^\star) \in \underset{\substack{j \in \mathcal{M}_{\text{try}} \\ (j,z) \in \mathcal{C}_A}}{\arg\max} L_n(j, z).$$

To remove ties, the best cut is always performed along the best cut direction $j_n^\star$, at the middle of two consecutive data points. It should be noted that, in the original algorithm, the resampling step in **Algorithm 1** (line 2) is done by bootstrapping, that is by choosing n out of n points with replacement. Here, we consider a slightly different version where resampling is done by choosing a_n out of n points without replacement, where $a_n < n$. Picking data points without replacement is just a convenient hypothesis for the proofs. Considering resampling instead of bootstrapping is an easy way to ensure that the estimation error of the forest can be controlled.

3 Main results

We consider an additive regression model satisfying the following properties:

Assumption 1. (H1) *The response Y follows*

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

where $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ is uniformly distributed over $[0,1]^p$, ε is an independent centered Gaussian noise with finite variance $\sigma^2 > 0$, and each component m_j is assumed to be continuous.

Additive regression models, which extend linear models, were popularized by Stone (1985) and Hastie and Tibshirani (1986). These models, which decompose the regression function as a sum of univariate functions, are flexible and easy to interpret. They are acknowledged for providing a good trade-off between model complexity and calculation time, and were accordingly extensively studied for the last thirty years. Additive models also play an important role in the context of high-dimensional data analysis and sparse modelling, where they are successfully involved in procedures such as the Lasso and various aggregation schemes (for an overview, see, e.g., Hastie et al., 2009).

Throughout the document, $\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}$ stands for the indicator that $\mathbf{X}_i$ falls in the same cell as $\mathbf{X}$ in the random tree designed with $\mathcal{D}_n$ and the random parameter Θ. We denote by $\mathcal{P}_\Theta$ the partition corresponding to such a random tree. The following assumption will be needed for our analysis:

Assumption 2. (H2) *There exists a sequence $(\gamma_n)_n \to 0$ such that, for all $n \in \mathbb{N}^\star$, and for all $1 \leq i, j \leq n$ with $j \neq i$, almost surely,*

$$\left| \mathbb{E}\left[Y_i - m(\mathbf{X}_i) \middle| \mathbf{X}_i, \mathbf{X}_j, Y_j, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \right] \right| \leq \gamma_n,$$

where Θ' is an independent copy of Θ. Besides, there exists a constant $\sigma'^2 > 0$ such that, for all $1 \leq i \leq n$, almost surely,

$$\mathbb{E}\left[(Y_i - m(\mathbf{X}_i))^2 \middle| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \right] \leq \sigma'^2.$$

It is stressed that **(H2)** is satisfied in the ideal case where partitions are independent of the labels Y_i's. Indeed, in that case, the indicators $\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}$ and the sample $\mathcal{D}_n$ are independent, which implies that, almost surely, for all $1 \leq i, j \leq n$ such that $i \neq j$,

$$\mathbb{E}\left[Y_i - m(\mathbf{X}_i) \middle| \mathbf{X}_i, \mathbf{X}_j, Y_j, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \right] = 0,$$

and, for all $1 \leq i \leq n$,

$$\mathbb{E}\left[(Y_i - m(\mathbf{X}_i))^2 \middle| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \right] = \sigma^2.$$

However, assuming such an independence in the Breiman's forests is by far unrealistic, since CART-splits make an extensive use of the whole sample $\mathcal{D}_n$ to grow the trees. Thus, assumption **(H2)** states that the dependency

between partitions and a fixed data point is weak. Despite its technical aspect, the first part of **(H2)** can be easily interpreted by saying that given two indicators $\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}$ and $\mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}$ (which contain information about partitions $\mathcal{P}_\Theta$ and $\mathcal{P}_{\Theta'}$), the pair $(\mathbf{X}_j, Y_j)$ does not much influence the value Y_i of another data point. The two random variables Y_i and Y_j are independent but, since partitions depend upon the whole training sample, Y_i is **not** independent of $(\mathbf{X}_i, \mathbf{X}_j, Y_j, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j})$. Thus, **(H2)** requires this independence to be weak, in the sense that $\mathbb{E}[Y_i | \mathbf{X}_i, \mathbf{X}_j, Y_j, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}]$ is close to $m(\mathbf{X}_i) = \mathbb{E}[Y_i | \mathbf{X}_i]$. Similarly, the second part of **(H2)** can be understood by saying that the indicator $\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}$ does not alter too much the conditional law of Y_i knowing X_i, in the sense that $\mathbb{E}\left[(Y_i - m(\mathbf{X}_i))^2 \middle| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right]$ should be close to $\sigma^2 = \mathbb{E}\left[(Y_i - m(\mathbf{X}_i))^2 \middle| \mathbf{X}_i\right]$. In that case, this would ensure that the first quantity is bounded above. It is our belief that requirements in **(H2)** are mild, since partitions based on CART-split criterion are not strongly influenced by single observations. Indeed, splits are decided via quantities involving empirical mean and variance, which are just averages over sets of data points and, as such, do not too much depend upon single values. Thus, knowing the value of one data point does not provide much information about the value of another one.

We are now equipped to state our main result.

Theorem 3.1. *Assume that* **(H1)** *and* **(H2)** *are satisfied. Then, provided* $a_n \to \infty$ *and* $a_n \log n/n \to 0$*, random forests are consistent, i.e., we have*

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

Up to our knowledge, apart from the fact that bootstrapping is replaced by subsampling, this is the first consistency result for Breiman's (2001) forests. Indeed, models studied so far are designed independently of the sample $\mathcal{D}_n$, and this makes them much easier to analyse. However, such models are clearly an unrealistic representation of the true procedure, which uses both the positions $\mathbf{X}_i$'s and the values Y_i's to grow the trees. The resulting forest is then highly data-dependent and understanding its behavior deserves a more involved mathematical treatment. This important issue will be thoroughly discussed in Section 4.

Our analysis sheds also some interesting light on the behavior of forests when the ambient dimension p is large but the true underlying dimension of the

model is small. To see how, assume that the additive model **(H1)** satisfies a sparsity constraint of the form

$$Y = \sum_{j=1}^{S} m_j(\mathbf{X}^{(j)}) + \varepsilon,$$

where $S < p$ represents the true, but unknown, dimension of the model. Thus, among the p original features, it is assumed that only the first (without loss of generality) S variables are informative. Put differently, Y is assumed to be independent of the last $(p - S)$ variables. In this dimension reduction context, the ambient dimension p can be very large, but we believe that the representation is sparse, i.e., that few components of m are non-zero. As such, the value S characterizes the sparsity of the model: the smaller S, the sparser m.

Proposition 1 below shows that random forests nicely adapt to the sparsity setting by asymptotically performing, with high probability, splits along the S informative variables.

In this proposition, we set $m_{\text{try}} = p$ and, for all k, we denote by $j_{1,n}(\mathbf{X}), \ldots,$ $j_{k,n}(\mathbf{X})$ the first k cut directions used to construct the cell containing $\mathbf{X}$, with the convention that $j_{q,n}(\mathbf{X}) = \infty$ if the cell has been cut strictly less than q times.

Proposition 1. *Assume that* **(H1)** *is satisfied. Let* $k \in \mathbb{N}^\star$ *and* $\xi > 0$. *Assume that there is no interval* $[a, b]$ *and no* $j \in \{1, \ldots, S\}$ *such that* m_j *is constant on* $[a, b]$. *Then, with probability* $1 - \xi$, *for all* n *large enough, we have, for all* $1 \leq q \leq k$,

$$j_{q,n}(\mathbf{X}) \in \{1, \ldots, S\}.$$

This proposition provides an interesting perspective on why random forests are still able to do a good job in high-dimensional settings. Since the algorithm selects splits mostly along informative variables, everything happens as if data were projected onto the vector space generated by the S informative variables. Therefore, forests are likely to only depend upon these S variables, which, in turn, improves the performance of the method compared to other non-adaptive (i.e., whose construction is independent of the Y_i's) ones. This is in line with the results of Biau (2012), who proved that, for a simplified model, the performance of the method only depends upon the true dimension S and not on the ambient dimension p.

4 Discussion

At first, it should be mentioned that, contrary to most previous works, there is only one observation per leaf of each individual tree. This implies that the single trees are eventually not consistent, since standard conditions for tree consistency require that the number of observations in the terminal nodes tends to infinity as n grows (see, e.g., Devroye et al., 1996; Györfi et al., 2002). Thus, the random forest algorithm aggregates rough individual tree predictors to build a provably consistent general architecture.

One of the main difficulties in assessing the mathematical properties of Breiman's (2001) forests is that the construction process of the individual trees strongly depends on both the X_i's and the Y_i's. For partitions that are independent of the Y_i's, consistency can be shown by relatively simple means via Stone's (1977) theorem for local averaging estimates (see also Györfi et al., 2002, Chapter 6). However, our partitions and trees depend upon the Y-values in the data. This makes things complicated, but mathematically interesting too. Thus, logically, the proof of Theorem 3.1 starts with an adaptation of Stone's (1977) theorem tailored for random forests. More precisely, the proof relies on two main arguments, both of which stress an important feature of the random forest mechanism.

The first argument is outlined in Proposition 2 below. It states that the variation of the regression function m within a cell of a random tree is small provided n is large enough. To this aim, we define, for any cell A, the variation of m within A as

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

Furthermore, we denote by $A_n(\mathbf{X}, \Theta)$ the cell of a tree built with random parameter Θ that contains the point $\mathbf{X}$.

Proposition 2. *Assume that* **(H1)** *holds. For all* $\rho, \xi > 0$*, there exists* $N \in \mathbb{N}^\star$ *such that, for all* $n > N$*,*

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

It should be noted that one of the main requirements of Stone's theorem applied to Y-independent partitioning estimates is that the diameter of the tree cells tends to zero in probability. Instead of such a geometrical assumption, Proposition 2 ensures that the variation of m inside a cell is small, thereby forcing the approximation error of the forest to asymptotically approach zero.

The second important argument relies on the fact that the subsampling rate a_n/n is $o(1/\log n)$ in Theorem 3.1. This guarantees that every single observation $(\mathbf{X}_i, Y_i)$ is used in the tree construction with a probability that becomes small with n. It also implies that the query point $\mathbf{x}$ cannot be connected to the same data point in a high proportion of trees. If not, the predicted value at $\mathbf{x}$ would be too much influenced by one single pair $(\mathbf{X}_i, Y_i)$, making the forest inconsistent. In fact, the proof of Theorem 3.1 reveals that the estimation error of a forest estimate is small as soon as the maximum probability of connection between the query point and all observations is small. Thus, the assumption on the subsampling rate is just a convenient way to control these probabilities, by ensuring that partitions are dissimilar enough. This is the case if $\mathbf{x}$ is connected with many data points through the forest. This idea of diversity among trees was introduced by Breiman (2001), but is generally difficult to analyse. In our approach, the subsampling is the key component for imposing tree diversity.

5 Proof of Theorem 3.1

For the sake of clarity, proofs of the intermediary results are postponed to Section 6. We start with some notations.

5.1 Notations

In the sequel, to clarify the notation, we will sometimes write $d = (d^{(1)}, d^{(2)})$ to represent a cut (j, z).

Recall that, for any cell A, $\mathcal{C}_A$ is the set of all possible cuts in A. Thus, with this notation, $\mathcal{C}_{[0,1]^p}$ is just the set of all possible cuts at the root of the tree, that is, all possible choices $d = (d^{(1)}, d^{(2)})$ with $d^{(1)} \in \{1, \ldots, p\}$ and $d^{(2)} \in [0, 1]$.

More generally, for any $\mathbf{x} \in [0, 1]^p$, we call $\mathcal{A}_k(\mathbf{x})$ the collection of all possible $k \geq 1$ consecutive cuts used to build the cell containing $\mathbf{x}$. Such a cell is obtained after a sequence of cuts $\mathbf{d}_k = (d_1, \ldots, d_k)$, where the dependency of $\mathbf{d}_k$ upon $\mathbf{x}$ is understood. Accordingly, for any $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$, we let $A(\mathbf{x}, \mathbf{d}_k)$ be the cell containing $\mathbf{x}$ built with the particular k-tuple of cuts $\mathbf{d}_k$. The proximity between two elements $\mathbf{d}_k$ and $\mathbf{d}'_k$ in $\mathcal{A}_k(\mathbf{x})$ will be measured via

$$\|\mathbf{d}_k - \mathbf{d}'_k\|_\infty = \sup_{1 \leq j \leq k} \max \left(|d_j^{(1)} - d_j'^{(1)}|, |d_j^{(2)} - d_j'^{(2)}| \right).$$

Accordingly, the distance d_∞ between $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$ and any $\mathcal{A} \subset \mathcal{A}_k(\mathbf{x})$ is

$$d_\infty(\mathbf{d}_k, \mathcal{A}) = \inf_{\mathbf{z}\in\mathcal{A}} \|\mathbf{d}_k - \mathbf{z}\|_\infty.$$

Remember that $A_n(\mathbf{X}, \Theta)$ denotes the cell of a tree containing $\mathbf{X}$ and designed with random parameter Θ. Similarly, $A_{k,n}(\mathbf{X}, \Theta)$ is the same cell but where only the first k cuts are performed ($k \in \mathbb{N}^\star$ is a parameter to be chosen later). We also denote by $\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta) = (\hat{d}_{1,n}(\mathbf{X}, \Theta), \ldots, \hat{d}_{k,n}(\mathbf{X}, \Theta))$ the k cuts used to construct the cell $A_{k,n}(\mathbf{X}, \Theta)$.

Recall that, for any cell A, the empirical criterion used to split A in the random forest algorithm is defined in (2). For any cut $(j, z) \in \mathcal{C}_A$, we define the following theoretical version of $L_n(\cdot, \cdot)$:

$$\begin{aligned} L^\star(j, z) = \mathbb{V}[Y|\mathbf{X} \in A] - \mathbb{P}[\mathbf{X}^{(j)} < z \,|\, \mathbf{X} \in A]\ \mathbb{V}[Y|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\ - \mathbb{P}[\mathbf{X}^{(j)} \geq z \,|\, \mathbf{X} \in A]\ \mathbb{V}[Y|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]. \end{aligned}$$

Observe that $L^\star(\cdot, \cdot)$ does not depend upon the training set and that, by the strong law of large numbers, $L_n(j, z) \to L^\star(j, z)$ almost surely as $n \to \infty$ for all cuts $(j, z) \in \mathcal{C}_A$. Therefore, it is natural to define the best theoretical split $(j^\star, z^\star)$ of the cell A as

$$(j^\star, z^\star) \in \operatorname*{arg\,min}_{\substack{(j,z)\in\mathcal{C}_A \\ j\in\mathcal{M}_{\text{try}}}} L^\star(j, z).$$

In view of this criterion, we define the theoretical random forest as before, but with consecutive cuts performed by optimizing $L^\star(\cdot, \cdot)$ instead of $L_n(\cdot, \cdot)$. We note that this new forest does depend on Θ through $\mathcal{M}_{\text{try}}$, but not on the sample $\mathcal{D}_n$. In particular, the stopping criterion for dividing cells has to be changed in the theoretical random forest; instead of stopping when a cell has a single training point, we impose that each tree of the theoretical forest is stopped at a fixed level $k \in \mathbb{N}^\star$. We also let $A_k^\star(\mathbf{X}, \Theta)$ be a cell of the theoretical random tree at level k, containing $\mathbf{X}$, designed with randomness Θ, and resulting from the k theoretical cuts $\mathbf{d}_k^\star(\mathbf{X}, \Theta) = (d_1^\star(\mathbf{X}, \Theta), \ldots, d_k^\star(\mathbf{X}, \Theta))$. Since there can exist multiple best cuts at, at least, one node, we call $\mathcal{A}_k^\star(\mathbf{X}, \Theta)$ the set of all k-tuples $\mathbf{d}_k^\star(\mathbf{X}, \Theta)$ of best theoretical cuts used to build $A_k^\star(\mathbf{X}, \Theta)$.

We are now equipped to prove Proposition 2. For clarity reasons, the proof has been divided in three steps. Firstly, we study in Lemma 1, the theoretical random forest. Then we prove in Lemma 3 (via Lemma 2), that theoretical and empirical cuts are close to each other. Proposition 2 is finally established as a consequence of Lemma 1 and Lemma 3. Proofs of these lemmas are to be found in Section 6.

5.2 Proof of Proposition 2

We first need a lemma which states that the variation of $m(\mathbf{X})$ within the cell $A_k^\star(\mathbf{X},\Theta)$ where $\mathbf{X}$ falls, as measured by $\Delta(m, A_k^\star(\mathbf{X},\Theta))$, tends to zero.

Lemma 1. *Assume that* **(H1)** *is satisfied. Then, for all* $\mathbf{x}\in[0,1]^p$,

$$\Delta(m, A_k^\star(\mathbf{x},\Theta)) \to 0, \quad \textit{almost surely, as } k\to\infty.$$

The next step is to show that cuts in theoretical and original forests are close to each other. To this aim, for any $\mathbf{x}\in[0,1]^p$ and any k-tuple of cuts $\mathbf{d}_k \in \mathcal{A}_k(\mathbf{x})$, we define

$$\begin{aligned} L_{n,k}(\mathbf{x},\mathbf{d}_k) = \frac{1}{N_n(A(\mathbf{x},\mathbf{d}_{k-1}))}\sum_{i=1}^n (Y_i - \bar{Y}_{A(\mathbf{x},\mathbf{d}_{k-1})})^2 \mathbb{1}_{\mathbf{X}_i\in A(\mathbf{x},\mathbf{d}_{k-1})} \\ - \frac{1}{N_n(A(\mathbf{x},\mathbf{d}_{k-1}))}\sum_{i=1}^n \Big(Y_i - \bar{Y}_{A_L(\mathbf{x},\mathbf{d}_{k-1})}\mathbb{1}_{\mathbf{X}_i^{(d_k^{(1)})}<d_k^{(2)}} \\ - \bar{Y}_{A_R(\mathbf{x},\mathbf{d}_{k-1})}\mathbb{1}_{\mathbf{X}_i^{(d_k^{(1)})}\geq d_k^{(2)}}\Big)^2 \mathbb{1}_{\mathbf{X}_i\in A(\mathbf{x},\mathbf{d}_{k-1})}, \end{aligned}$$

where $A_L(\mathbf{x},\mathbf{d}_{k-1}) = A(\mathbf{x},\mathbf{d}_{k-1})\cap\{\mathbf{z} : \mathbf{z}^{(d_k^{(1)})} < d_k^{(2)}\}$ and $A_R(\mathbf{x},\mathbf{d}_{k-1}) = A(\mathbf{x},\mathbf{d}_{k-1})\cap\{\mathbf{z} : \mathbf{z}^{(d_k^{(1)})} \geq d_k^{(2)}\}$, and where we use the convention $0/0=0$ when $A(\mathbf{x},\mathbf{d}_{k-1})$ is empty. Besides, we let $A(\mathbf{x},\mathbf{d}_0)=[0,1]^p$ in the previous equation. The quantity $L_{n,k}(\mathbf{x},\mathbf{d}_k)$ is nothing but the criterion to maximize in d_k to find the best k-th cut in the cell $A(\mathbf{x},\mathbf{d}_{k-1})$. Lemma 2 below ensures that $L_{n,k}(\mathbf{x},\cdot)$ is stochastically equicontinuous, for all $\mathbf{x}\in[0,1]^p$. To this aim, for all $\xi>0$, and for all $\mathbf{x}\in[0,1]^p$, we denote by $\mathcal{A}_{k-1}^\xi(\mathbf{x})\subset\mathcal{A}_{k-1}(\mathbf{x})$ the set of all $(k-1)$-tuples $\mathbf{d}_{k-1}$ such that the cell $A(\mathbf{x},\mathbf{d}_{k-1})$ contains a hypercube of edge length ξ. Moreover, we let $\bar{\mathcal{A}}_k^\xi(\mathbf{x}) = \{\mathbf{d}_k : \mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^\xi(\mathbf{x})\}$ equipped with the norm $\|\mathbf{d}_k\|_\infty$.

Lemma 2. *Assume that* **(H1)** *is satisfied. Fix* $\mathbf{x}\in[0,1]^p$, $k\in\mathbb{N}^\star$ *and let* $\xi>0$. *Then* $L_{n,k}(\mathbf{x},\cdot)$ *is stochastically equicontinuous on* $\bar{\mathcal{A}}_k^\xi(\mathbf{x})$, *that is, for all* $\alpha,\rho>0$, *there exists* $\delta>0$ *such that*

$$\lim_{n\to\infty}\mathbb{P}\left[\sup_{\substack{\|\mathbf{d}_k-\mathbf{d}_k'\|_\infty\leq\delta \\ \mathbf{d}_k,\mathbf{d}_k'\in\bar{\mathcal{A}}_k^\xi(\mathbf{x})}} |L_{n,k}(\mathbf{x},\mathbf{d}_k) - L_{n,k}(\mathbf{x},\mathbf{d}_k')| > \alpha\right] \leq \rho.$$

Lemma 2 is then used in Lemma 3 to assess the distance between theoretical and empirical cuts.

Lemma 3. *Assume that* **(H1)** *is satisfied. Fix* $\xi, \rho > 0$ *and* $k \in \mathbb{N}^\star$. *Then there exists* $N \in \mathbb{N}^\star$ *such that, for all* $n \geq N$,

$$\mathbb{P}\left[d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho.$$

We are now ready to prove Proposition 2. Fix $\rho, \xi > 0$. Since almost sure convergence implies convergence in probability, according to Lemma 1, there exists $k_0 \in \mathbb{N}^\star$ such that

$$\mathbb{P}\left[\Delta(m, A_{k_0}^\star(\mathbf{X}, \Theta)) \leq \xi\right] \geq 1 - \rho. \tag{3}$$

By Lemma 3, for all $\xi_1 > 0$, there exists $N \in \mathbb{N}^\star$ such that, for all $n \geq N$,

$$\mathbb{P}\left[d_\infty(\hat{\mathbf{d}}_{k_0,n}(\mathbf{X}, \Theta), \mathcal{A}_{k_0}^\star(\mathbf{X}, \Theta)) \leq \xi_1\right] \geq 1 - \rho. \tag{4}$$

Since m is uniformly continuous, we can choose ξ_1 sufficiently small such that, for all $\mathbf{x} \in [0,1]^p$, for all $\mathbf{d}_{k_0}, \mathbf{d}'_{k_0}$ satisfying $d_\infty(\mathbf{d}_{k_0}, \mathbf{d}'_{k_0}) \leq \xi_1$, we have

$$\left|\Delta(m, A(\mathbf{x}, \mathbf{d}_{k_0})) - \Delta(m, A(\mathbf{x}, \mathbf{d}'_{k_0}))\right| \leq \xi. \tag{5}$$

Thus, combining inequalities (4) and (5), we obtain

$$\mathbb{P}\Big[\left|\Delta(m, A_{k_0,n}(\mathbf{X}, \Theta)) - \Delta(m, A_{k_0}^\star(\mathbf{X}, \Theta))\right| \leq \xi\Big] \geq 1 - \rho. \tag{6}$$

Using the fact that $\Delta(m, A) \leq \Delta(m, A')$ whenever $A \subset A'$, we deduce from (3) and (6) that, for all $n \geq N$,

$$\mathbb{P}\left[\Delta(m, A_n(\mathbf{X}, \Theta)) \leq 2\xi\right] \geq 1 - 2\rho.$$

This concludes the proof of Proposition 2.

5.3 Proof of Theorem 3.1

Recall that each cell contains exactly one data point. Thus, letting

$$W_{ni}(\mathbf{X}) = \mathbb{E}_\Theta\left[\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}, \Theta)}\right],$$

the random forest estimate m_n may be rewritten as

$$m_n(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}(\mathbf{X}) Y_i.$$

We have in particular that $\sum_{i=1}^{n} W_{ni}(\mathbf{X}) = 1$. Thus,

$$\begin{aligned}
\mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 &\leq 2\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X})(Y_i - m(\mathbf{X}_i))\right]^2 \\
&\quad + 2\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(\mathbf{X})(m(\mathbf{X}_i) - m(\mathbf{X}))\right]^2 \\
&\stackrel{\text{def}}{=} 2I_n + 2J_n.
\end{aligned}$$

Fix $\alpha > 0$ and let $\|m\|_\infty = \sup_{\mathbf{x}\in[0,1]^p} |m(\mathbf{x})|$. To upper bound J_n, note that by Jensen's inequality,

$$\begin{aligned}
J_n &\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)} (m(\mathbf{X}_i) - m(\mathbf{X}))^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{X},\Theta)} \Delta^2(m, A_n(\mathbf{X},\Theta))\right] \\
&\leq \mathbb{E}\left[\Delta^2(m, A_n(\mathbf{X},\Theta))\right].
\end{aligned}$$

So, by definition of $\Delta(m, A_n(\mathbf{X},\Theta))^2$,

$$\begin{aligned}
J_n &\leq 4\|m\|_\infty^2 \, \mathbb{E}[\mathbb{1}_{\Delta^2(m,A_n(\mathbf{X},\Theta))\geq\alpha}] + \alpha \\
&\leq \alpha(4\|m\|_\infty^2 + 1),
\end{aligned}$$

for all n large enough, according to Proposition 2.

To upper bound I_n, we note that

$$I_n = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ni}(\mathbf{X}) W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right].$$

Hereafter, to simplify notation, we write $\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}$ instead of $\mathbb{1}_{\mathbf{X}\in A_n(\mathbf{X}_i,\Theta)}$, keeping in mind the fact that the indicator $\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}$ depends upon the whole sample $\mathcal{D}_n$. Thus, recalling that $W_{ni}(\mathbf{X}) = \mathbb{E}_\Theta[\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}]$, we have, for any fixed i, and

for all $j \neq i$,

$$\begin{aligned}
&\mathbb{E}\left[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right] \\
&= \mathbb{E}\Bigg[\mathbb{E}\Bigg[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} (Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j)) \\
&\qquad\qquad \Bigg| \mathbf{X}_i, \mathbf{X}_j, Y_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j}\Bigg]\Bigg] \\
&= \mathbb{E}\Bigg[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} (Y_i - m(\mathbf{X}_i)) \\
&\qquad\qquad \times \mathbb{E}\left[Y_j - m(\mathbf{X}_j) | \mathbf{X}_i, \mathbf{X}_j, Y_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}, \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_i}\right]\Bigg].
\end{aligned}$$

Therefore, by assumption **(H2)**,

$$\begin{aligned}
&\left|\mathbb{E}\big[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\big]\right| \\
&\quad \leq \gamma_n \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} |Y_i - m(\mathbf{X}_i)|\right].
\end{aligned}$$

Taking the sum over (i, j) for $i \neq j$, we obtain

$$\begin{aligned}
&\left|\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbb{E}\left[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right]\right| \\
&\quad \leq \gamma_n \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} |Y_i - m(\mathbf{X}_i)|\right] \\
&\quad \leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} |Y_i - m(\mathbf{X}_i)|\right] \\
&\quad \leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{E}\left[|Y_i - m(\mathbf{X}_i)| \Big| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right]\right] \\
&\quad \leq \gamma_n \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{E}^{1/2}\left[|Y_i - m(\mathbf{X}_i)|^2 \Big| \mathbf{X}_i, \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}\right]\right] \\
&\quad \leq \gamma_n \sigma'.
\end{aligned}$$

We conclude that, for all n large enough,

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathbb{E}\left[W_{ni}(\mathbf{X}, \Theta) W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\right] \leq \alpha.$$

Consequently, recalling that $\varepsilon_i = Y_i - m(\mathbf{X}_i)$, we have, for all n large enough,

$$\begin{aligned} I_n &\leq \alpha + \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}^2(\mathbf{X})(Y_i - m(\mathbf{X}_i))^2\right] \\ &\leq \alpha + \mathbb{E}\left[\max_{1 \leq \ell \leq n} W_{n\ell}(\mathbf{X}) \sum_{i=1}^{n} W_{ni}(\mathbf{X}) \varepsilon_i^2\right] \\ &\leq \alpha + \mathbb{E}\left[\max_{1 \leq \ell \leq n} W_{n\ell}(\mathbf{X}) \max_{1 \leq i \leq n} \varepsilon_i^2\right]. \end{aligned} \tag{7}$$

Now, observe that in the subsampling step, there are exactly $\binom{a_n - 1}{n-1}$ choices to pick a fixed observation $\mathbf{X}_i$. Since $\mathbf{x}$ and $\mathbf{X}_i$ belong to the same cell only if $\mathbf{X}_i$ is selected in the subsampling step, we see that

$$\mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{\binom{a_n - 1}{n-1}}{\binom{a_n}{n}} = \frac{a_n}{n},$$

where $\mathbb{P}_\Theta$ denotes the probability with respect to Θ, conditionally on $\mathbf{X}$ and $\mathcal{D}_n$. So,

$$\max_{1 \leq i \leq n} W_{ni}(\mathbf{X}) \leq \max_{1 \leq i \leq n} \mathbb{P}_\Theta\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right] \leq \frac{a_n}{n}. \tag{8}$$

Thus, combining inequalities (7) and (8), for all n large enough,

$$I_n \leq \alpha + \frac{a_n}{n} \mathbb{E}\left[\max_{1 \leq i \leq n} \varepsilon_i^2\right].$$

The term inside the brackets is the maximum of n χ^2-squared distributed random variables. Thus, for some positive constant C,

$$\mathbb{E}\left[\max_{1 \leq i \leq n} \varepsilon_i^2\right] \leq C \log n,$$

(see, e.g., Chapter 1 in Boucheron et al., 2013). We conclude that, for all n large enough,

$$I_n \leq \alpha + C \frac{a_n \log n}{n} \leq 2\alpha.$$

Since α was arbitrary, the proof is complete.

6 Technical results

6.1 Proof of Lemma 1

Technical Lemma 1. *Assume that* **(H1)** *is satisfied and that* $L^\star \equiv 0$ *for all cuts in some given cell* A*. Then the regression function* m *is constant on* A*.*

Proof of Technical Lemma 1. We start by proving the result in dimension $p = 1$. Letting $A = [a, b]$ $(0 \leq a < b \leq 1)$, and recalling that $Y = m(\mathbf{X}) + \varepsilon$, one has

$$\begin{aligned} L^\star(1, z) &= \mathbb{V}\left[Y | \mathbf{X} \in A\right] - \mathbb{P}\left[a \leq \mathbf{X} \leq z \,|\, \mathbf{X} \in A\right] \mathbb{V}\left[Y | a \leq \mathbf{X} \leq z\right] \\ &\quad - \mathbb{P}\left[z \leq \mathbf{X} \leq b \,|\, \mathbf{X} \in A\right] \mathbb{V}\left[Y | z < \mathbf{X} \leq b\right] \\ &= -\frac{1}{(b-a)^2} \left(\int_a^b m(t) \mathrm{d}t \right)^2 + \frac{1}{(b-a)(z-a)} \left(\int_a^z m(t) \mathrm{d}t \right)^2 \\ &\quad + \frac{1}{(b-a)(b-z)} \left(\int_z^b m(t) \mathrm{d}t \right)^2 . \end{aligned}$$

Let $C = \int_a^b m(t) \mathrm{d}t$ and $M(z) = \int_a^z m(t) \mathrm{d}t$. Simple calculations show that

$$L^\star(1, z) = \frac{1}{(z-a)(b-z)} \left(M(z) - C \frac{z-a}{b-a} \right)^2 .$$

Therefore, since $L^\star \equiv 0$ on $\mathcal{C}_A$ by assumption, we obtain,

$$M(z) = C \, \frac{z-a}{b-a} \, .$$

This proves that $M(z)$ is linear in z, and that m is therefore constant on $[a, b]$.

Let us now examine the general multivariate case, where $A = \Pi_{j=1}^p [a_j, b_j] \subset [0, 1]^p$. From the univariate analysis, we know that, for all $1 \leq j \leq p$, there exists a constant C_j such that

$$\int_{a_1}^{b_1} \ldots \int_{a_p}^{b_p} m(\mathbf{x}) \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \ldots \mathrm{d}x_p = C_j \, .$$

Since m is additive this implies that, for all j and for all x_j,

$$m_j(x_j) = C_j - \int_{a_1}^{b_1} \dots \int_{a_p}^{b_p} \sum_{\ell \neq j} m_\ell(x_\ell) \mathrm{d}x_1 \dots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \dots \mathrm{d}x_p \,,$$

which does not depend upon x_i. This shows that m is constant on A. □

Proof of Lemma 1. Take $\xi > 0$ and $\mathbf{x} \in [0,1]^p$. Let θ be a realization of the random variable Θ. Since m is uniformly continuous, the result is clear if $\mathrm{diam}(A_k^\star(\mathbf{x},\theta))$ tends to zero as k tends to infinity. Thus, in the sequel, it is assumed that $\mathrm{diam}(A_k^\star(\mathbf{x},\theta))$ does not tend to zero. In that case, since $(A_k^\star(\mathbf{x},\theta))_k$ is a decreasing sequence of compact sets, there exist $\mathbf{a}_\infty(\mathbf{x},\theta) = (\mathbf{a}_\infty^{(1)}(\mathbf{x},\theta), \dots, \mathbf{a}_\infty^{(p)}(\mathbf{x},\theta)) \in [0,1]^p$ and $\mathbf{b}_\infty(\mathbf{x},\theta) = (\mathbf{b}_\infty^{(1)}(\mathbf{x},\theta), \dots, \mathbf{b}_\infty^{(p)}(\mathbf{x},\theta)) \in [0,1]^p$ such that

$$\begin{aligned} \bigcap_{k=1}^{\infty} A_k^\star(\mathbf{x},\theta) &= \prod_{j=1}^{p} [\mathbf{a}_\infty^{(j)}(\mathbf{x},\theta), \mathbf{b}_\infty^{(j)}(\mathbf{x},\theta)] \\ &\stackrel{\text{def}}{=} A_\infty^\star(\mathbf{x},\theta). \end{aligned}$$

Since $\mathrm{diam}(A_k^\star(\mathbf{x},\theta))$ does not tend to zero, there exists an index j' such that $\mathbf{a}_\infty^{(j')}(\mathbf{x},\theta) < \mathbf{b}_\infty^{(j')}(\mathbf{x},\theta)$ (i.e., the cell $A_\infty^\star(\mathbf{x},\theta)$ is not reduced to a point). Let $A_k^\star(\mathbf{x},\theta) \stackrel{\text{def}}{=} \prod_{j=1}^p [\mathbf{a}_k^{(j)}(\mathbf{x},\theta), \mathbf{b}_k^{(j)}(\mathbf{x},\theta)]$ be the cell containing $\mathbf{x}$ at level k. If the criterion $L^\star$ is identically zero for all cuts in $A_\infty^\star(\mathbf{x},\theta)$ then m is constant on $A_\infty^\star(\mathbf{x},\theta)$ according to Lemma 1. This implies that $\Delta(m, A_\infty^\star(\mathbf{x},\theta)) = 0$. Thus, in that case, since m is uniformly continuous,

$$\lim_{k\to\infty} \Delta(m, A_k^\star(\mathbf{x},\theta)) = \Delta(m, A_\infty^\star(\mathbf{x},\theta)) = 0.$$

Let us now show by contradiction that $L^\star$ is a.s. necessarily null on the cuts of $A_\infty^\star(\mathbf{x},\theta)$. In the rest of the proof, for all $k \in \mathbb{N}^\star$, we let $L_k^\star$ as the criterion $L^\star$ used in the cell $A_k^\star(\mathbf{x},\theta)$, that is

$$\begin{aligned} L_k^\star(d) = \mathbb{V}[Y|\mathbf{X} \in A_k^\star(\mathbf{x},\theta)] \\ - \mathbb{P}[\mathbf{X}^{(j)} < z \,|\, \mathbf{X} \in A_k^\star(\mathbf{x},\theta)]\ \mathbb{V}[Y|\mathbf{X}^{(j)} < z, \mathbf{X} \in A_k^\star(\mathbf{x},\theta)] \\ - \mathbb{P}[\mathbf{X}^{(j)} \geq z \,|\, \mathbf{X} \in A_k^\star(\mathbf{x},\theta)]\ \mathbb{V}[Y|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A_k^\star(\mathbf{x},\theta)], \end{aligned}$$

for all $d = (j,z) \in \mathcal{C}_{A_k^\star(\mathbf{x},\theta)}$. If $L_\infty^\star$ it is not identically zero, then there exists a cut $d_\infty(\mathbf{x},\theta)$ in $\mathcal{C}_{A_\infty^\star(\mathbf{x},\theta)}$ such that $L^\star(d_\infty(\mathbf{x},\theta)) = c > 0$. Fix $\xi > 0$. By the uniform continuity of m, there exists $\delta_1 > 0$ such that

$$\sup_{\|\mathbf{w}-\mathbf{w}'\|_\infty \leq \delta_1} |m(\mathbf{w}) - m(\mathbf{w}')| \leq \xi \,.$$

Since $A_k^\star(\mathbf{x},\theta) \downarrow A_\infty^\star(\mathbf{x},\theta)$, there exists k_0 such that, for all $k \geq k_0$,

$$\max\left(\|\mathbf{a}_k(\mathbf{x},\theta) - \mathbf{a}_\infty(\mathbf{x},\theta)\|_\infty, \|\mathbf{b}_k(\mathbf{x},\theta) - \mathbf{b}_\infty(\mathbf{x},\theta)\|_\infty\right) \leq \delta_1\,. \tag{9}$$

Observe that for all $k \in \mathbb{N}^\star$, $\mathbb{V}[Y|\mathbf{X} \in A_{k+1}^\star(\mathbf{x},\theta)] < \mathbb{V}[Y|\mathbf{X} \in A_k^\star(\mathbf{x},\theta)]$. Thus,

$$\underline{L}_k^\star := \sup_{\substack{d\in\mathcal{C}_{A_k(\mathbf{x},\theta)}\\ d^{(1)}\in\mathcal{M}_{\text{try}}}} L_k^\star(d) \leq \xi\,. \tag{10}$$

From inequality (9), we deduce that

$$\left|\mathbb{E}[m(\mathbf{X})|\mathbf{X} \in A_k^\star(\mathbf{x},\theta)] - \mathbb{E}[m(\mathbf{X})|\mathbf{X} \in A_\infty^\star(\mathbf{x},\theta)]\right| \leq \xi.$$

Consequently, there exists a constant $C > 0$, such that, for all $k \geq k_0$, for all cuts $d \in \mathcal{C}_{A_\infty^\star(\mathbf{x},\theta)}$,

$$|L_k^\star(d) - L_\infty^\star(d)| \leq C\xi^2. \tag{11}$$

Let $k_1 \geq k_0$ be the first level after k_0 at which the direction $d_\infty^{(1)}(\mathbf{x},\theta)$ is amongst the m_{try} selected coordinates. With probability 1, $k_1 < \infty$. Thus, by the definition of $d_\infty(\mathbf{x},\theta)$, and inequality (11),

$$c - C\xi^2 \leq L_\infty^\star(d_\infty(\mathbf{x},\theta)) - C\xi^2 \leq L_k^\star(d_\infty(\mathbf{x},\theta)),$$

which implies that $c - C\xi^2 \leq \underline{L}_k^\star$. Hence, using inequality (10), we have

$$c - C\xi^2 \leq \underline{L}_k^\star \leq \xi,$$

which is absurd, since $c > 0$ is fixed and ξ is arbitrarily small. Thus, by Lemma 1, m is constant on $A_\infty^\star(\mathbf{x},\theta)$. This implies that $\Delta(m, A_k^\star(\mathbf{x},\Theta)) \to 0$ as $k \to \infty$.

6.2 Proof of Lemma 2

We start by proving Lemma 2 in the case $k = 1$, i.e., when we perform the first cut at the root of a tree. Since in that case $L_{n,1}(\mathbf{x},\cdot)$ does not depend on $\mathbf{x}$, we simply write $L_{n,1}(\cdot)$ instead of $L_{n,1}(\mathbf{x},\cdot)$.

Proof of Lemma 2 in the case $k = 1$. Fix $\alpha, \rho > 0$. Observe that if two cuts d_1, d_2 satisfy $\|d_1 - d_2\|_\infty < 1$, the cut directions are the same, i.e. $d_1^{(1)} = d_2^{(1)}$. Using this fact and for reasons of symmetry, we just need to prove Lemma 2

when the cuts are performed along the first dimension. In other words, we only need to prove that

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{|x_1-x_2|\leq\delta} |L_{n,1}(1,x_1) - L_{n,1}(1,x_2)| > \alpha\right] \leq \rho/p. \tag{12}$$

Recall that, for all i, $Y_i = m(\mathbf{X}_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$. Letting $Z_i = \max_{1\leq i\leq n} |\varepsilon_i|$, simple calculations show that

$$\mathbb{P}\left[Z_i \geq t\right] = 1 - \exp\left(n \ln\left(1 - 2\mathbb{P}\left[\varepsilon_1 \geq t\right]\right)\right).$$

The last probability can be upper bounded by using the following standard inequality on Gaussian tail:

$$\mathbb{P}\left[\varepsilon_1 \geq t\right] \leq \frac{\sigma}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Consequently, there exists a constant $C_\rho > 0$ and $N_1 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $n > N_1$,

$$\max_{1\leq i\leq n} |\varepsilon_i| \leq C_\rho \sqrt{\log n}\,. \tag{13}$$

Besides, by simple calculations on Gaussian tail, for all $n \in \mathbb{N}^\star$, we have,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\right| \geq \alpha\right] \leq \frac{\sigma}{\alpha\sqrt{n}} \exp\left(-\frac{\alpha^2 n}{2\sigma^2}\right).$$

Since there are, at most, n^2 sets of the form $\{i : X_i \in [a_n, b_n]\}$ for $0 \leq a_n < b_n \leq 1$, we deduce that, from the last inequality and using an union bound, there exists $N_2 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $n > N_2$, for all $0 \leq a_n < b_n \leq 1$ satisfying $N_n([a_n, b_n] \times [0,1]^{p-1}) > \sqrt{n}$, we have

$$\left|\frac{1}{N_n([a_n, b_n] \times [0,1]^{p-1})} \sum_{\substack{i:X_i\in[a_n,b_n]\\ \times[0,1]^{p-1}}} \varepsilon_i\right| \leq \alpha. \tag{14}$$

By the Glivenko-Cantelli theorem, there exists $N_3 \in \mathbb{N}^\star$ such that, with probability $1-\rho$, for all $0 \leq a < b \leq 1$, for all $n > N_3$,

$$(b - a - \delta^2)n \leq N_n([a,b] \times [0,1]^{p-1}) \leq (b - a + \delta^2)n. \tag{15}$$

Throughout the proof, we assume to be on the event where assertions (13)-(15) hold, which occurs with probability $1-3\rho$, for all $n > N$, where $N = \max(N_1, N_2, N_3)$.

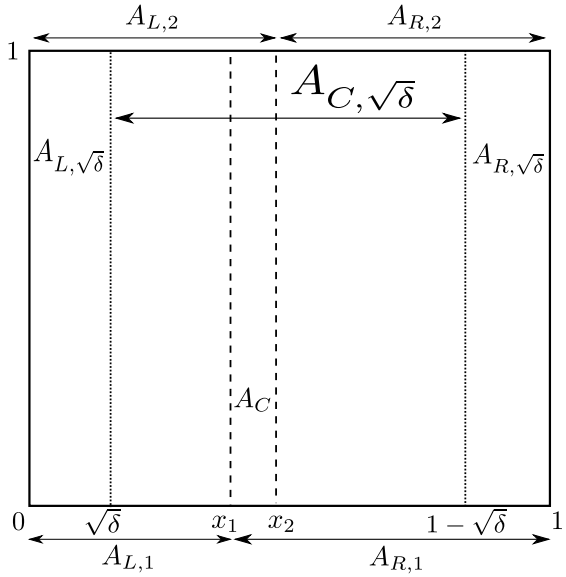

Figure 1: Illustration of the notations in dimension $p = 2$.

Take $x_1, x_2 \in [0,1]$ such that $|x_1 - x_2| \leq \delta$ and assume, without loss of generality, that $x_1 < x_2$. In the remainder of the proof, we will need the following quantities (see Figure 1 for an illustration in dimension two):

$$\begin{cases} A_{L,\sqrt{\delta}} = [0, \sqrt{\delta}] \times [0,1]^{p-1} \\ A_{R,\sqrt{\delta}} = [1-\sqrt{\delta}, 1] \times [0,1]^{p-1} \\ A_{C,\sqrt{\delta}} = [\sqrt{\delta}, 1-\sqrt{\delta}] \times [0,1]^{p-1}. \end{cases}$$

Similarly, we define

$$\begin{cases} A_{L,1} = [0, x_1] \times [0,1]^{p-1} \\ A_{R,1} = [x_1, 1] \times [0,1]^{p-1} \\ A_{L,2} = [0, x_2] \times [0,1]^{p-1} \\ A_{R,2} = [x_2, 1] \times [0,1]^{p-1} \\ A_C = [x_1, x_2] \times [0,1]^{p-1}. \end{cases}$$

Recall that, for any cell A, $\bar{Y}_A$ is the mean of the Y_i's falling in A and $N_n(A)$ is the number of data points in A. To prove (12), five cases are to be considered, depending upon the positions of x_1 and x_2. We repeatedly use the decomposition

$$L_{n,1}(1, x_1) - L_{n,1}(1, x_2) = J_1 + J_2 + J_3,$$

where

$$J_1 = \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}(Y_i - \bar{Y}_{A_{L,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}<x_1}(Y_i - \bar{Y}_{A_{L,2}})^2,$$

$$J_2 = \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}(Y_i - \bar{Y}_{A_{R,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}(Y_i - \bar{Y}_{A_{L,2}})^2,$$

$$\text{and} \quad J_3 = \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\geq x_2}(Y_i - \bar{Y}_{A_{R,1}})^2 - \frac{1}{n}\sum_{i:\mathbf{X}_i^{(1)}\geq x_2}(Y_i - \bar{Y}_{A_{R,2}})^2.$$

First case: $x_1, x_2 \in A_{C,\sqrt{\delta}}$. Since $N_n(A_{L,2}) > N_n(A_{L,\sqrt{\delta}}) > \sqrt{n}$ for all $n > N$, we have, according to inequalities (14),

$$|\bar{Y}_{A_{L,2}}| \leq \|m\|_\infty + \alpha \quad \text{and} \quad |\bar{Y}_{A_{R,1}}| \leq \|m\|_\infty + \alpha.$$

Therefore

$$\begin{aligned}
|J_2| &= 2\left|\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{R,1}}\right| \times \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\left(Y_i - \frac{\bar{Y}_{A_{L,2}} + \bar{Y}_{A_{R,1}}}{2}\right)\right| \\
&\leq 4(\|m\|_\infty + \alpha)\left(\frac{(\|m\|_\infty + \alpha)N_n(A_C)}{n} + \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} m(\mathbf{X}_i)\right|\right. \\
&\qquad\qquad \left. + \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\right|\right) \\
&\leq 4(\|m\|_\infty + \alpha)\left((\delta+\delta^2)(\|m\|_\infty + \alpha) + \|m\|_\infty(\delta+\delta^2)\right. \\
&\qquad\qquad \left. + \frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\right|\right).
\end{aligned}$$

If $N_n(A_C) \geq \sqrt{n}$, we have

$$\frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\right| \leq \frac{1}{N_n(A_C)}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\right| \leq \alpha \quad \big(\text{according to (14)}\big)$$

or, if $N_n(A_C) < \sqrt{n}$, we have

$$\frac{1}{n}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]}\varepsilon_i\right| \leq \frac{C_\rho\sqrt{\log n}}{\sqrt{n}} \quad \big(\text{according to (13)}\big).$$

Thus, for all n large enough,

$$|J_2| \leq 4(\|m\|_\infty + \alpha)\bigg((\delta + \delta^2)(2\|m\|_\infty + \alpha) + \alpha\bigg). \tag{16}$$

With respect to J_1, observe that

$$\begin{aligned}
|\bar{Y}_{A_{L,1}} - \bar{Y}_{A_{L,2}}| &= \left|\frac{1}{N_n(A_{L,1})}\sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i - \frac{1}{N_n(A_{L,2})}\sum_{i:\mathbf{X}_i^{(1)}<x_2} Y_i\right| \\
&\leq \left|\frac{1}{N_n(A_{L,1})}\sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i - \frac{1}{N_n(A_{L,2})}\sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i\right| \\
&\quad + \left|\frac{1}{N_n(A_{L,2})}\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} Y_i\right| \\
&\leq \left|1 - \frac{N_n(A_{L,1})}{N_n(A_{L,2})}\right| \times \frac{1}{N_n(A_{L,1})} \times \left|\sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i\right| \\
&\quad + \frac{1}{N_n(A_{L,2})}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} Y_i\right|.
\end{aligned}$$

Since $N_n(A_{L,2}) - N_n(A_{L,1}) \leq n(\delta + \delta^2)$, we obtain

$$1 - \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \leq \frac{n(\delta+\delta^2)}{N_n(A_{L,2})} \leq \frac{\delta+\delta^2}{\sqrt{\delta}-\delta^2} \leq 4\sqrt{\delta},$$

for all δ small enough, which implies that

$$\begin{aligned}
|\bar{Y}_{A_{L,1}} - \bar{Y}_{A_{L,2}}| &\leq \frac{4\sqrt{\delta}}{N_n(A_{L,1})}\left|\sum_{i:\mathbf{X}_i^{(1)}<x_1} Y_i\right| \\
&\quad + \frac{N_n(A_{L,1})}{N_n(A_{L,2})} \times \frac{1}{N_n(A_{L,1})}\left|\sum_{i:\mathbf{X}_i^{(1)}\in[x_1,x_2]} Y_i\right| \\
&\leq 4\sqrt{\delta}(\|m\|_\infty + \alpha) + \frac{N_n(A_{L,1})}{N_n(A_{L,2})}(\|m\|_\infty\delta + \alpha) \\
&\leq 5(\|m\|_\infty\sqrt{\delta} + \alpha).
\end{aligned}$$

Thus,

$$
\begin{aligned}
|J_1| &= \left| \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} (Y_i - \bar{Y}_{A_{L,1}})^2 - \frac{1}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} (Y_i - \bar{Y}_{A_{L,2}})^2 \right| \\
&= \left| (\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{L,1}}) \times \frac{2}{n} \sum_{i:\mathbf{X}_i^{(1)} < x_1} \left(Y_i - \frac{\bar{Y}_{A_{L,1}} + \bar{Y}_{A_{L,2}}}{2} \right) \right| \\
&\leq |\bar{Y}_{A_{L,2}} - \bar{Y}_{A_{L,1}}|^2 \\
&\leq 25(\|m\|_\infty \sqrt{\delta} + \alpha)^2. \qquad (17)
\end{aligned}
$$

The term J_3 can be bounded with the same arguments.

Finally, by (16) and (17), for all $n > N$, for all δ small enough, we conclude that

$$
\begin{aligned}
|L_n(1, x_1) - L_n(1, x_2)| \leq 4(\|m\|_\infty + \alpha) &\Big((\delta + \delta^2)(2\|m\|_\infty + \alpha) + \alpha \Big) \\
&+ 25(\|m\|_\infty \sqrt{\delta} + \alpha)^2 \\
&\leq \alpha.
\end{aligned}
$$

Second case: $x_1, x_2 \in A_{L,\sqrt{\delta}}$. With the same arguments as above, one proves that

$$
\begin{aligned}
|J_1| &\leq \max\left(4(\sqrt{\delta} + \delta^2)(\|m\|_\infty + \alpha)^2, \alpha \right), \\
|J_2| &\leq \max(4(\|m\|_\infty + \alpha)(2\delta\|m\|_\infty + 2\alpha), \alpha), \\
|J_3| &\leq 25(\|m\|_\infty \sqrt{\delta} + \alpha)^2.
\end{aligned}
$$

Consequently, for all n large enough,

$$
|L_n(1, x_1) - L_n(1, x_2)| = J_1 + J_2 + J_3 \leq 3\alpha.
$$

The other cases $\{x_1, x_2 \in A_{R,\sqrt{\delta}}\}$, $\{x_1, x_2 \in A_{L,\sqrt{\delta}} \times A_{C,\sqrt{\delta}}\}$, and $\{x_1, x_2 \in A_{C,\sqrt{\delta}} \times A_{R,\sqrt{\delta}}\}$ can be treated in the same way. Details are omitted. □

Proof of Lemma 2. We proceed similarly as in the proof for the case $k = 1$. Here, we establish the result for $k = 2$ and $p = 2$ only. Extensions are easy and left to the reader. Fix $\rho > 0$. At first, it should be noted that, there exists $N_1 \in \mathbb{N}^\star$ such that, with probability $1 - \rho$, for all $n > N_0$, $A_n \equiv [a_n^{(1)}, b_n^{(1)}] \times [a_n^{(2)}, b_n^{(2)}] \subset [0, 1]^2$ such that $N_n(A_n) > \sqrt{n}$, we have

$$
\left| \frac{1}{N_n(A_n)} \sum_{i:X_i \in A_n} \varepsilon_i \right| \leq \alpha, \qquad (18)
$$

and,

$$\frac{1}{N_n(A_n)} \sum_{i:X_i \in A_n} \varepsilon_i^2 \leq \tilde{\sigma}^2, \tag{19}$$

where $\tilde{\sigma}^2$ is a positive constant, depending only on ρ. Inequality (19) is a straightforward consequence of the following inequality (see e.g. Laurent and Massart, 2000), which is valid for all $n \in \mathbb{N}^\star$:

$$\mathbb{P}\left[\chi^2(n) \geq 5n\right] \leq \exp(-n).$$

Throughout the proof, we assume to be on the event where assertions (13), (15), (18)-(19) hold, which occurs with probability $1 - 3\rho$, for all n large enough. We also assume that $d_1 = (1, x_1)$ and $d_2 = (2, x_2)$ (see Figure 2). The other cases can be treated similarly.

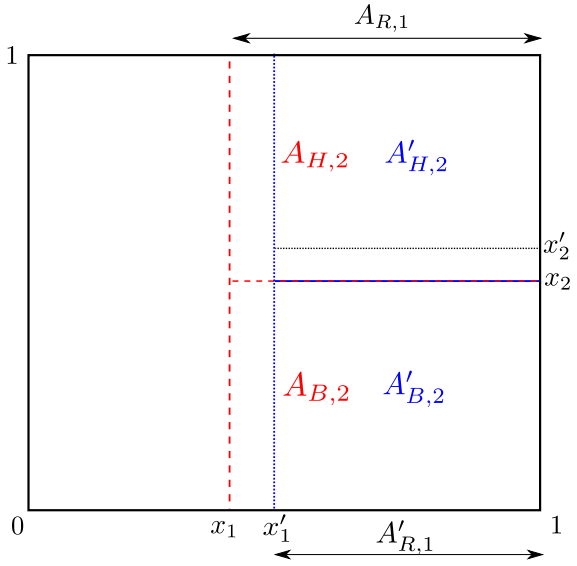

Figure 2: An example of cells in dimension $p = 2$.

Let $d'_1 = (1, x'_1)$ and $d'_2 = (2, x'_2)$ be such that

$$|x_1 - x'_1| < \delta \quad \text{and} \quad |x_2 - x'_2| < \delta.$$

Then the CART-split criterion $L_{n,2}$ writes

$$\begin{aligned} L_n(d_1, d_2) = \frac{1}{N_n(A_{R,1})} \sum_i (Y_i - \bar{Y}_{A_{R,1}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ - \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ - \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1}. \end{aligned}$$

Clearly,

$$L_n(d_1, d_2) - L_n(d_1', d_2') = L_n(d_1, d_2) - L_n(d_1', d_2) + L_n(d_1', d_2) - L_n(d_1', d_2').$$

We have (Figure 2):

$$\begin{aligned} L_n(d_1, d_2) - L_n(d_1', d_2) = \Bigg[\frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ - \frac{1}{N_n(A'_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'} \Bigg] \\ + \Bigg[\frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1} \\ - \frac{1}{N_n(A'_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} \leq x_2} (Y_i - \bar{Y}_{A'_{B,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'} \Bigg] \\ \stackrel{\text{def}}{=} A_1 + B_1. \end{aligned}$$

The term A_1 can be rewritten as $A_1 = A_{1,1} + A_{1,2} + A_{1,3}$, where

$$\begin{aligned} A_{1,1} = \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'} \\ - \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'}, \\ A_{1,2} = \frac{1}{N_n(A_{R,1})} \sum_{i: \mathbf{X}_i^{(2)} > x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)} > x_1'} \end{aligned}$$

$$- \frac{1}{N_n(A'_{R,1})} \sum_{i:\mathbf{X}_i^{(2)}>x_2} (Y_i - \bar{Y}_{A'_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)}>x'_1},$$

$$\text{and} \quad A_{1,3} = \frac{1}{N_n(A_{R,1})} \sum_{i:\mathbf{X}_i^{(2)}>x_2} (Y_i - \bar{Y}_{A_{H,2}})^2 \mathbb{1}_{\mathbf{X}_i^{(1)}\in[x_1,x'_1]}.$$

Calculations show that

$$A_{1,1} = \frac{N_n(A'_{H,2})}{N_n(A_{R,1})} (\bar{Y}_{A'_{H,2}} - \bar{Y}_{A_{H,2}})^2,$$

which implies, with the same arguments as in the proof for $k = 1$, that $A_{1,1} \to 0$ as n tends to infinity. With respect to $A_{1,2}$ and $A_{1,3}$, we write

$$\max(A_{1,2}, A_{1,3}) \leq \max(C_\rho \frac{\log n}{\sqrt{n}}, 2(\tilde{\sigma}^2 + 4\|m\|_\infty^2 + \alpha^2)\frac{\sqrt{\delta}}{\xi}).$$

Thus, $A_{1,2} \to 0$ and $A_{1,3} \to 0$ as $n \to \infty$. Collecting bounds, we conclude that $A_1 \to 0$. One proves with similar arguments that $B_1 \to 0$ and, consequently, that

$$L_n(d'_1, d_2) - L_n(d'_1, d'_2) \to 0.$$

□

6.3 Proof of Lemma 3

We prove by induction that, for all k, with probability $1 - \rho$, for all $\xi > 0$ and for all n large enough,

$$d_\infty(\hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta), \mathcal{A}_k^\star(\mathbf{X}, \Theta)) \leq \xi.$$

Call this property H_k. Fix $k > 1$ and assume that H_{k-1} is true. We momentarily keep $\mathbf{X}$ fixed. For all $\mathbf{d}_{k-1} \in \mathcal{A}_{k-1}(\mathbf{X})$, let

$$\hat{d}_{k,n}(\mathbf{d}_{k-1}) \in \arg\min_{d_k} L_n(\mathbf{X}, \mathbf{d}_{k-1}, d_k),$$

and

$$d_k^\star(\mathbf{d}_{k-1}) \in \arg\min_{d_k} L^\star(\mathbf{X}, \mathbf{d}_{k-1}, d_k),$$

where the minimum is evaluated, as usual, over $\{d_k \in \mathcal{C}_{A(\mathbf{X},\mathbf{d}_{k-1})} : d_k^{(1)} \in \mathcal{M}_{\text{try}}\}$.

Fix $\rho > 0$. Observe that the volume of a cell which does not contain a hypercube of edge length ξ is necessarily less than ξ. Thus, the probability to fall in such a cell at level k is, at most, $2^k\xi$. Since k is fixed, letting $\xi = \rho/2^k$, conditionally on $\mathcal{D}_n$, with probability $1-\rho$, the cell $A_{k,n}(\mathbf{X},\Theta)$ contains a hypercube of edge length ξ. In the rest of the proof, we assume $\mathbf{X}$ and Θ to be fixed and that $A_{k,n}(\mathbf{X},\Theta)$ satisfies this geometrical property. Moreover, since $\mathbf{X}$ and Θ are fixed, we omit the dependence on $\mathbf{X}$ and Θ.

Note that, for all $\mathbf{d}_{k-1}$,

$$\begin{aligned}
&L_n(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) \\
&\quad \le L_n(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) \\
&\qquad \text{(by definition of } d_k^\star(\mathbf{d}_{k-1})) \\
&\quad \le L_n(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) \\
&\qquad \text{(by definition of } \hat{d}_{k,n}(\mathbf{d}_{k-1})).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left|L_n(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\right| \\
&\quad \le \max\Big(\left|L_n(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))\right|, \\
&\qquad\qquad \left|L_n(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))\right|\Big) \\
&\quad \le \sup_{d_k} \left|L_n(\mathbf{d}_{k-1}, d_k) - L^\star(\mathbf{d}_{k-1}, d_k)\right|.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&|L^\star(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))| \\
&\quad \le |L^\star(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L_n(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1}))| \\
&\qquad + |L_n(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^\star(\mathbf{d}_{k-1}, d_k^\star(\mathbf{d}_{k-1}))| \\
&\quad \le 2 \sup_{d_k} |L_n(\mathbf{d}_{k-1}, d_k) - L^\star(\mathbf{d}_{k-1}, d_k)| \\
&\quad = 2 \sup_{d_k} |L_n(\mathbf{d}_k) - L^\star(\mathbf{d}_k)|. \qquad (20)
\end{aligned}$$

Let $\bar{\mathcal{A}}_k^\xi = \left\{\mathbf{d}_k : \mathbf{d}_{k-1} \in \mathcal{A}_{k-1}^\xi\right\}$. So, taking the supremum on both sides of

(20) leads to

$$\sup_{\mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^{\xi}} |L^{\star}(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^{\star}(\mathbf{d}_{k-1}, d_k^{\star}(\mathbf{d}_{k-1}))|$$
$$\leq 2 \sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^{\xi}} |L_n(\mathbf{d}_k) - L^{\star}(\mathbf{d}_k)|. \tag{21}$$

By Lemma 2, for all $\xi' > 0$, one can find $\delta > 0$ such that, for all n large enough,

$$\mathbb{P}\left[\sup_{\substack{\|\mathbf{d}_k-\mathbf{d}'_k\|_\infty\leq\delta \\ \mathbf{d}_k,\mathbf{d}'_k\in\bar{\mathcal{A}}_k^{\xi}}} |L_n(\mathbf{d}_k) - L_n(\mathbf{d}'_k)| \leq \xi'\right] \geq 1-\rho. \tag{22}$$

Since $\bar{\mathcal{A}}_k^{\xi}$ is compact, there exists a finite subset $\mathbf{X}_\delta = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\} \subset \bar{\mathcal{A}}_k^{\xi}$ such that, for all $\mathbf{d}_k \in \bar{\mathcal{A}}_k^{\xi}$, $d_\infty(\mathbf{d}_k, \mathbf{X}_\delta) \leq \delta$. Hence, for all $\mathbf{d}_k \in \bar{\mathcal{A}}_k^{\xi}$, there is an index j such that, with probability $1-\rho$, for all n large enough,

$$\begin{aligned}
|L_n(\mathbf{d}_k) - L^{\star}(\mathbf{d}_k)| &\leq |L_n(\mathbf{d}_k) - L_n(\mathbf{x}_j)| + |L_n(\mathbf{x}_j) - L^{\star}(\mathbf{x}_j)| \\
&\quad + |L^{\star}(\mathbf{x}_j) - L^{\star}(\mathbf{d}_k)| \\
&\leq 2\xi' + |L_n(\mathbf{x}_j) - L^{\star}(\mathbf{x}_j)| \\
&\quad \text{(by inequality (22))} \\
&\leq 3\xi',
\end{aligned}$$

since, for all $\mathbf{x}_j$, $L_n(\mathbf{x}_j) \to L^{\star}(\mathbf{x}_j)$ almost surely, as n tends to infinity. In consequence, with probability $1-\rho$, for all n large enough,

$$\sup_{\mathbf{d}_k\in\bar{\mathcal{A}}_k^{\xi}} |L_n(\mathbf{d}_k) - L^{\star}(\mathbf{d}_k)| \leq 3\xi',$$

and, using inequality (21), we finally obtain that with probability $1-\rho$, for all n large enough,

$$\sup_{\mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^{\xi}} |L^{\star}(\mathbf{d}_{k-1}, \hat{d}_{k,n}(\mathbf{d}_{k-1})) - L^{\star}(\mathbf{d}_{k-1}, d_k^{\star}(\mathbf{d}_{k-1}))| \leq 6\xi'. \tag{23}$$

Hereafter, to simplify, we assume that, for any given $(k-1)$-tuple of theoretical cuts, there is only one theoretical cut at level k, and leave the general case as an easy adaptation. Thus, we can define unambiguously

$$d_k^{\star}(\mathbf{d}_{k-1}) = \arg\min_{d_k} L^{\star}(\mathbf{d}_{k-1}, d_k).$$

Fix $\xi'' > 0$. From inequality (23), by evoking the equicontinuity of L_n and the compactness of $\mathcal{A}_{k-1}^{\xi}$, we deduce that, with probability $1-\rho$, for all n large enough,

$$\sup_{\mathbf{d}_{k-1}\in\mathcal{A}_{k-1}^{\xi}} d_\infty\Big(\hat{d}_{k,n}(\mathbf{d}_{k-1}), d_k^\star(\mathbf{d}_{k-1})\Big) \leq \xi''.$$

In particular, with probability $1-\rho$, for all n large enough,

$$d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\hat{\mathbf{d}}_{k-1,n})\Big) \to 0.$$

Now, using triangle inequality,

$$\begin{aligned} d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), \overline{d_k^\star(\mathbf{d}_{k-1}^\star)}\Big) \leq\ & d_\infty\Big(\hat{d}_{k,n}(\hat{\mathbf{d}}_{k-1,n}), \overline{d_k^\star(\hat{\mathbf{d}}_{k-1,n})}\Big) \\ & + d_\infty\Big(\overline{d_k^\star(\hat{\mathbf{d}}_{k-1,n})}, \overline{d_k^\star(\mathbf{d}_{k-1}^\star)}\Big). \end{aligned}$$

Thus, we just have to show that $d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), \mathcal{A}_k^\star) \to 0$ in probability, as $n \to \infty$, and the proof will be complete.

To avoid confusion, we let

$$\overline{\mathbf{d}_{k-1}^\star} = \{\mathbf{d}_{k-1}^{\star,i} : i \in \mathcal{I}\}$$

be the set of best first $(k-1)$-th theoretical cuts (which can be either countable or not). With this notation, $d_k^\star(\mathbf{d}_{k-1}^{\star,i})$ is the k-th theoretical cuts given that the $(k-1)$ previous ones are $\mathbf{d}_{k-1}^{\star,i}$. For simplicity, let

$$L^{i,\star}(d_k) = L_k^\star(\mathbf{d}_{k-1}^{\star,i}, d_k) \quad \text{and} \quad \hat{L}^\star(d_k) = L_k^\star(\hat{\mathbf{d}}_{k-1,n}, d_k).$$

As before,

$$d_k^\star(\mathbf{d}_{k-1}^{\star,i}) \in \arg\min_{d_k} L^{i,\star}(d_k) \quad \text{and} \quad d_k^\star(\hat{\mathbf{d}}_{k-1,n}) \in \arg\min_{d_k} \hat{L}^\star(d_k).$$

Clearly, the result will be proved if we establish that,

$$\inf_{i\in\mathcal{I}} d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \to 0 \quad \text{in probability,} \quad \text{as } n \to \infty.$$

Since d_k belongs to a compact set, we can find a finite set $\mathbf{X} = \{x_1, \ldots, x_m\} \subset \mathcal{C}_{[0,1]^p}$ satisfying $d_\infty(d_k, \mathbf{X}) \leq \xi$. Thus, there is an index j such that, with probability $1-\rho$, for all n large enough,

$$\begin{aligned} |\hat{L}^\star(d_k) - L^{i,\star}(d_k)| &\leq |\hat{L}^\star(d_k) - \hat{L}^\star(\mathbf{x}_j)| + |\hat{L}^\star(\mathbf{x}_j) - L^{i,\star}(\mathbf{x}_j)| \\ &\quad + |L^{i,\star}(\mathbf{x}_j) - L^{i,\star}(d_k)| \\ &\leq 2\xi' + |\hat{L}^\star(\mathbf{x}_j) - L^{i,\star}(\mathbf{x}_j)|. \end{aligned}$$

Therefore, as in inequality (21), with probability $1-\rho$, for all i, and for all n large enough,

$$\begin{aligned}|L^{i,\star}(d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| &\leq 2\sup_{d_k}|\hat{L}^\star(d_k) - L^{i,\star}(d_k)| \\ &\leq 4\xi' + 2\max_j |\hat{L}^\star(\mathbf{x}_j) - L^{i,\star}(\mathbf{x}_j)|.\end{aligned}$$

Taking the infimum over all i, we obtain

$$\inf_i |L^{i,\star}(d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \leq 4\xi' + 2\inf_i \max_j |\hat{L}^\star(x_j) - L^{i,\star}(x_j)|. \tag{24}$$

Introduce ω, the modulus of continuity of $L_k^\star$:

$$\omega(\delta) = \sup_{\|x-y\|_\infty \leq \delta} |L_k^\star(x) - L_k^\star(y)|.$$

Observe that, since $L_k^\star(\cdot)$ is uniformly continuous, $\omega(\delta) \to 0$ as $\delta \to 0$. Hence, for all n large enough,

$$\begin{aligned}\inf_i \max_j |\hat{L}^\star(x_j) - L^{i,\star}(x_j)| &= \inf_i \max_j |L_k^\star(\hat{\mathbf{d}}_{k-1,n}, x_j) - L_k^\star(\mathbf{d}_{k-1}^{\star,i}, x_j)| \\ &\leq \inf_i \omega(\|\hat{\mathbf{d}}_{k-1,n} - \mathbf{d}_{k-1}^{\star,i}\|_\infty) \\ &\leq \xi', \end{aligned} \tag{25}$$

since, by assumption H_{k-1}, $\inf_i \|\hat{\mathbf{d}}_{k-1,n} - \mathbf{d}_{k-1}^{\star,i}\|_\infty \to 0$. Therefore, combining (24) and (25), with probability $1-\rho$, for all n large enough,

$$\inf_i |L^{i,\star}(d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \leq 6\xi.$$

Finally, by Lemma 4 below, H_k is true. Property H_1 can be proved in the same way.

Lemma 4. *For all $\delta > 0$, there exists $\xi > 0$ such that*

$$\inf_i d_\infty(d_k^\star(\hat{\mathbf{d}}_{k-1,n}), d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \leq \delta \tag{26}$$

whenever

$$\inf_i |L^{i,\star}(d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \leq \xi.$$

Proof of Lemma 4. Note that, for all $\delta > 0$, there exists $\xi > 0$ such that

$$\inf_i \inf_{y: d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \geq \delta} |L_k^\star(\mathbf{d}_{k-1}^{\star,i}, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) - L_k^\star(\mathbf{d}_{k-1}^{\star,i}, y)| \geq \xi.$$

To see this, assume that one can find $\delta > 0$ such that, for all $\xi > 0$, there exist i, y satisfying

$$|L_k^\star(\mathbf{d}_{k-1}^{\star,i}, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) - L_k^\star(\mathbf{d}_{k-1}^{\star,i}, y)| \leq \xi,$$

with $d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \geq \delta$. Letting $\xi_p = 1/p$ and recalling that $\{\mathbf{d}_{k-1}^{\star,i} : i \in \mathbb{N}\}$, $\{d_k^\star(\mathbf{d}_{k-1}^{\star,i}) : i \in \mathbb{N}\}$, and $\{y_i : i \in \mathbb{N}\}$ are compact, we can extract three sequences $\mathbf{d}_{k-1}^{\star,p} \to \mathbf{d}_{k-1}$, $d_k^\star(\mathbf{d}_{k-1}^{\star,p}) \to d_k$ and $y_p \to \tilde{y}$ as $p \to \infty$. Therefore,

$$L_k^\star(\mathbf{d}_{k-1}, d_k) = L_k^\star(\mathbf{d}_{k-1}, y),$$

with $d_\infty(y, \{d : L_k^\star(\mathbf{d}_{k-1}, d_k) = L_k^\star(\mathbf{d}_{k-1}, d)\}) \geq \delta$, which is absurd.

Assume now that (26) is not true. If this is the case, then there exists $\delta > 0$ such that

$$\inf_i d_\infty(d_k^\star(\mathbf{d}_{k-1}^{\star,i}), d_k^\star(\hat{\mathbf{d}}_{k-1,n})) \geq \delta.$$

Thus,

$$\begin{aligned}
&\inf_i |L^{i,\star}(d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L^{i,\star}(d_k^\star(\mathbf{d}_{k-1}^{\star,i}))| \\
&\quad = \inf_i |L_k^\star(\mathbf{d}_{k-1}^{\star,i}, d_k^\star(\hat{\mathbf{d}}_{k-1,n})) - L_k^\star(\mathbf{d}_{k-1}^{\star,i}, d_k^\star)| \\
&\quad \geq \inf_i \inf_{d_\infty(y, d_k^\star(\mathbf{d}_{k-1}^{\star,i})) \geq \delta} |L_k^\star(\mathbf{d}_{k-1}^{\star,i}, y) - L_k^\star(\mathbf{d}_{k-1}^{\star,i}, d_k^\star)| \\
&\quad \geq \xi,
\end{aligned}$$

which concludes the proof. □

Proof of Proposition 1. Fix $k \in \mathbb{N}^\star$ and $\rho, \xi > 0$. According to Lemma 3, with probability $1 - \rho$, for all n large enough, there exists a sequence of theoretical first k cuts $\mathbf{d}_k^\star(\mathbf{X}, \Theta)$ such that

$$d_\infty(\mathbf{d}_k^\star(\mathbf{X}, \Theta), \hat{\mathbf{d}}_{k,n}(\mathbf{X}, \Theta)) \leq \xi. \tag{27}$$

This implies that, with probability $1 - \rho$, for all n large enough and for all $1 \leq j \leq k$, the j-th empirical cut $\hat{d}_{j,n}(\mathbf{X}, \Theta)$ is performed along the same coordinate as $d_j^\star(\mathbf{X}, \Theta)$.

Now, for any cell A, since the regression function is not constant on A, one can find a theoretical cut $d_A^\star$ on A such that $L^\star(d_A^\star) > 0$. Thus, the cut $d_A^\star$ is

made along an informative variable, in the sense that it is performed along one of the first S variables. Consequently, for all $\mathbf{X}, \Theta$ and for all $1 \leq j \leq k$, each theoretical cut $d_j^\star(\mathbf{X}, \Theta)$ is made along one of the first S coordinate. The proof is then a consequence of inequality (27). □

Acknowledgements

This work was supported by the European Research Council [SMAC-ERC-280032].

References

D. Amaratunga, J. Cabrera, and Y.-S. Lee. Enriched random forests. *Bioinformatics*, 24:2010–2014, 2008.

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

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman. *Consistency for a simple model of random forests*. Technical Report 670, UC Berkeley, 2004.

L. Breiman, J. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30: 927–961, 2002.

S. Clémençon, M. Depecker, and N. Vayatis. Ranking forests. *Journal of Machine Learning Research*, 14:39–73, 2013.

D.R. Cutler, T.C. Edwards Jr, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, and J.J. Lawler. Random forests for classification in ecology. *Ecology*, 88: 2783–2792, 2007.

M. Denil, D. Matheson, and N. de Freitas. *Consistency of online random forests*. arXiv:1302.4853, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:1–13, 2006.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24:543–562, 2012.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–310, 1986.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Second Edition*. Springer, New York, 2009.

H. Ishwaran and U.B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80:1056–1064, 2010.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forest. *The Annals of Applied Statistics*, 2:841–860, 2008.

A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan. A scalable bootstrap for massive data. arXiv:1112.5016, 2012.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1302–1338, 2000.

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

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

A.M. Prasad, L.R. Iverson, and A. Liaw. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9:181–199, 2006.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

C.J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, pages 689–705, 1985.

V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43:1947–1958, 2003.

S. Wager, T. Hastie, and B. Efron. Standard errors for bagged predictors and random forests. arXiv:1311.4555, 2013.

R. Zhu, D. Zeng, and M.R. Kosorok. *Reinforcement learning trees*. Technical Report, University of North Carolina, 2012.