HAL open science

Consistent Regression using Data-Dependent Coverings

Vincent Margot, Jean-Patrick Baudry, Frédéric Guilloux, Olivier Wintenberger

To cite this version:

Vincent Margot, Jean-Patrick Baudry, Frédéric Guilloux, Olivier Wintenberger. Consistent Regression using Data-Dependent Coverings. 2019. hal-02170687v3

HAL Id: hal-02170687

https://hal.science/hal-02170687v3

Preprint submitted on 22 Oct 2019 (v3), last revised 22 Jan 2021 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Consistent Regression using Data-Dependent Coverings

Vincent Margot, Jean-Patrick Baudry, Frederic Guilloux, Olivier Wintenberger

Sorbonne Université, CNRS, LPSM, F-75005 Paris, France

Abstract

We introduce a novel method to generate *interpretable* regression function estimators. The idea is based on *data-dependent coverings*. The aim is to extract from the data a covering of the feature space instead of a partition. The estimator predicts the empirical conditional expectation over the cells of the partitions generated from the coverings. Thus, such estimator has the same form as those issued from *data-dependent partitioning* algorithms. We give sufficient conditions to ensure the consistency, avoiding the sufficient condition of shrinkage of the cells that appears in the former literature. Doing so, we reduce the number of covering elements. We show that such coverings are *interpretable* and each element of the covering is tagged as *significant* or *insignificant*.

The proof of the consistency is based on a control of the error of the empirical estimation of conditional expectations which is interesting on its own.

Keywords: Consistency, Nonparametric regression, Rule-based algorithm, Data-dependent covering, Interpretable learning.

1 Introduction

We consider the following regression setting: $(\mathbf{X}, Y)$ is a couple of random variables in $\mathbb{R}^d \times \mathbb{R}$ of unknown distribution $\mathbb{Q}$ such that

$$Y = g^*(\mathbf{X}) + Z,$$

where $\mathbb{E}[Z] = 0$, $\mathbb{V}(Z) = \sigma^2$ and g^* is a measurable function from $\mathbb{R}^d$ to $\mathbb{R}$. We make the following common assumptions:

- Z is independent of $\mathbf{X}$ and $\sigma^2 \geq 0$ is known; (H1)
- Y is bounded: $\mathbb{Q}(\mathcal{S}) = 1$ with $\mathcal{S} = \mathbb{R}^d \times [-L, L]$, for some $L > 0$ (unknown). (H2)

Given a sample $\mathbf{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$, we aim at predicting Y conditionally on $\mathbf{X}$. The observations $(\mathbf{X}_i, Y_i)$ are independent and identically distributed (i.i.d.) from the distribution $\mathbb{Q}$. The accuracy of a regression function $g : \mathbb{R}^d \to \mathbb{R}$ is measured by its quadratic risk, defined as

$$\mathcal{L}(g) = \mathbb{E}_{\mathbb{Q}}\left[(g(\mathbf{X}) - Y)^2\right] .$$

Thanks to Hypothesis (H1), we have

$$g^*(\mathbf{X}) = \mathbb{E}\left[Y|\mathbf{X}\right] = \arg\min_g \mathcal{L}(g) \text{ a.s,} \tag{1}$$

where the arg min is taken over the class of all measurable regression functions.

The regression functions generated from the data $\mathbf{D}_n$ by a learning algorithm are called estimators of g^*. We consider a set of regression functions $\mathcal{G}_n$ that contains all such estimators. Let $\mathbb{Q}_n$ be the empirical distribution of the sample $\mathbf{D}_n$. We define the empirical risk, the empirical risk minimizer and the minimizer of the risk over $\mathcal{G}_n$ as, respectively,

$$\mathcal{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n}\left(g(\mathbf{X}_i) - Y_i\right)^2 ,\ g_n = \arg\min_{g\in\mathcal{G}_n} \mathcal{L}_n(g) \text{ and } \tilde{g}_n = \arg\min_{g\in\mathcal{G}_n} \mathcal{L}(g). \tag{2}$$

The aim of this paper is to provide *interpretable* learning algorithms that generate $\mathcal{G}_n$ so that the associated empirical risk minimizer g_n is consistent, i.e. g_n converges to g^* as $n \to \infty$. More precisely, we show the weak consistency of the estimator g_n, i.e. its excess of risk

$$\ell(g^*, g_n) = \mathcal{L}(g_n) - \mathcal{L}(g^*) = \mathbb{E}[(g_n(\mathbf{X}) - g^*(\mathbf{X}))^2] = o_{\mathbb{P}}(1) .$$

But, as explain in Lipton [2017], they are several meanings of *interpretability* depending on the users desiderata and the expected properties of algorithms. In this paper, we use the definition of model interpretability from Biran and Cotton [2017]: **Interpretability is the degree to which an observer can understand the cause of a decision.**

1.1 Rule-based algorithms using partitions and coverings

In this paper we consider algorithms generating interpretable models that are rule-based, such as *CART* [Breiman et al., 1984], *ID3* [Quinlan, 1986], *C4.5* [Quinlan, 1993], *FORS* [Karalič and Bratko, 1997], *M5 Rules* [Holmes et al., 1999]. In these models, the regression function is explained by the realization of a simple condition, an *If-Then* statement of the form:

$$\begin{array}{rl} \text{IF} & (\mathbf{X}[i_1] \in c_1) \text{ And } (\mathbf{X}[i_2] \in c_2) \text{ And } \ldots \text{ And } (\mathbf{X}[i_k] \in c_k) \\ \text{THEN} & g_n(\mathbf{X}) = p \end{array} \tag{3}$$

where $\mathbf{X}[i]$ is the ith coordinate of $\mathbf{X}$ and $c_i \subseteq \mathbb{R}$.

The *If* part, called the *condition* of the rule, or simply the rule, is composed of the conjunction of $k \leq d$ tests, each of which checking whether a feature (a coordinate of $\mathbf{X}$) satisfies a specified property or not and k is called the *length* of the rule. The *Then* part, called the *conclusion* of the rule, is the estimated value when the rule is *activated*, i.e. when the condition in the *If* part is satisfied. The rules are easy to understand and allow an interpretable decision process when k is small. For a review of the best-known algorithms for descriptive and predictive rule learning, see Zhao and Bhowmick [2003] and Fürnkranz and Kliegr [2015].

Formally, the models generated by such algorithms are defined by a corresponding *data-dependent partition* $\mathcal{P}_n$ of $\mathbb{R}^d$. Each element of the partition is named a *cell* and the empirical risk minimizer associated to $\mathcal{P}_n$ satisfies

$$g_n(\mathbf{x}) = \sum_{A \in \mathcal{P}_n} \frac{\sum_{i=1}^n Y_i \mathbf{1}_{X_i \in A}}{\sum_{i=1}^n \mathbf{1}_{X_i \in A}} \mathbf{1}_{\mathbf{x} \in A} \,, \qquad \mathbf{x} \in \mathbb{R}^d \,. \tag{4}$$

Those algorithms use the dataset $\mathbf{D}_n$ twice; first, the partition $\mathcal{P}_n = \mathcal{P}_n(\mathbf{D}_n)$ is chosen according to the dataset, second, this partition and the data are used to compute $g_n(\mathbf{x})$ as in (4). Note that g_n is the empirical risk minimizer among the class of all piecewise constant functions over $\mathcal{P}_n$ denoted $\mathcal{G}_c \circ \mathcal{P}_n$. The major issue for these algorithms is the model interpretability, which requires a small value for the length k of the rule, whereas the consistency of the estimator is usually proved for conditions implying that $k = d$, *i.e.* a high model complexity.

In order to reduce the complexity of the model, we present a novel method of generating a partition. The idea is to generate a *data-dependent covering* $\mathcal{C}_n = \mathcal{C}_n(\mathbf{D}_n)$ of $\mathbb{R}^d$ rather than a partition. To do so, the dataset $\mathbf{D}_n$ is used to identify subsets of $\mathbb{R}^d$ that fulfill some specific conditions (we detail them in the next section). As elements of coverings can overlap, the construction of the subsets fulfilling these conditions can be done *separately*, which is not doable for the cells of partitions. Using a covering instead of a partition we ensure consistency without a condition on shrinkage of the cells. Moreover, each subset of the covering defines a rule with a small length k. Thus, we obtain a regression function described by a covering formed by simple rules rather than a partition formed by complex rules:

$$\begin{array}{rl} \text{IF} & (\mathbf{X} \in \mathbf{r}_1) \text{ And } (\mathbf{X} \in \mathbf{r}_2) \text{ And } \ldots \text{ And } (\mathbf{X} \in \mathbf{r}_l) \\ \text{THEN} & g_n(\mathbf{X}) = p \end{array}$$

where, for $j = 1, \ldots, l$

$$\mathbf{r}_j := \left\{ \mathbf{x} : (\mathbf{x}[i_{j,1}] \in c_{j,1}) \text{ And } (\mathbf{x}[i_{j,2}] \in c_{j,2}) \text{ And } \ldots \text{ And } (\mathbf{x}[i_{j,k_j}] \in c_{j,k_j}) \right\},$$

with $k_j \ll d$.

To estimate the value p, a partition $\mathcal{P}(\mathcal{C}_n)$ is generated from the covering $\mathcal{C}_n$ as an intermediate calculation. Formally, we define the partition generated

from any collection of subsets $\mathcal{C}$ using the power set $2^{\mathcal{C}}$ gathering all subsets of $\mathcal{C}$:

Definition 1.1. *Let $\mathcal{C}$ be a finite collection of subsets of $\mathbb{R}^d$ and let $\mathbf{c} = \bigcup_{\mathbf{r}\in\mathcal{C}} \mathbf{r}$. We define the activation function as*

$$\varphi_{\mathcal{C}} : \mathbb{R}^d \mapsto 2^{\mathcal{C}}; \qquad \varphi_{\mathcal{C}}(\mathbf{x}) = \{\mathbf{r} \in \mathcal{C} : \mathbf{x} \in \mathbf{r}\}.$$

Then $\mathcal{P}(\mathcal{C})$, the partition of $\mathbf{c}$ generated from $\mathcal{C}$, is defined as

$$\mathcal{P}(\mathcal{C}) := \varphi_{\mathcal{C}}^{-1}(Im(\varphi_{\mathcal{C}})).$$

We illustrate this transformation $\mathcal{P}$ on an example of four elements in Figures 1 and 2.

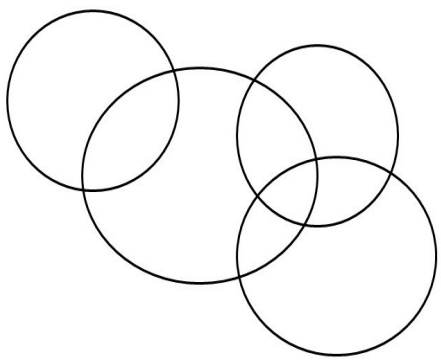

Figure 1: The four elements of $\mathcal{C}$.

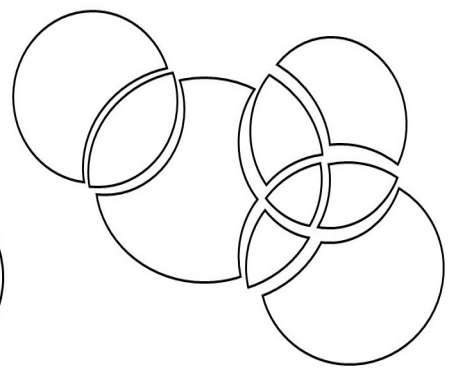

Figure 2: The 9 cells of the partition $\mathcal{P}(\mathcal{C})$.

Remark 1. If $\mathcal{C}$ is a covering of $\mathbb{R}^d$, then $\mathcal{P}(\mathcal{C})$ is a partition of $\mathbb{R}^d$. The relation $\mathcal{C} = \mathcal{P}(\mathcal{C})$ holds if and only if $\mathcal{C}$ is a partition of $\mathbf{c} = Im(\varphi_{\mathcal{C}})$.

Definition 1.2. *Let $\mathcal{C}$ be a covering of $\mathbb{R}^d$ and $\mathcal{C}' \subseteq \mathcal{C}$. The cells of $\mathcal{P}(\mathcal{C})$ wich are included in an element of $\mathcal{C}'$ are gathered in $\mathcal{P}_{\mathcal{C}}(\mathcal{C}')$:*

$$\mathcal{P}_{\mathcal{C}}(\mathcal{C}') := \{A \in \mathcal{P}(\mathcal{C}) : \exists \mathbf{r} \in \mathcal{C}',\, A \subseteq \mathbf{r}\}.$$

For $\mathbf{r} \in \mathcal{C}$, we shorten $\mathcal{P}_{\mathcal{C}}(\{\mathbf{r}\})$ in $\mathcal{P}(\mathbf{r})$.

Remark 2. By definition, we have the identity $\mathcal{P}_{\mathcal{C}}(\mathcal{C}) = \mathcal{P}(\mathcal{C})$.

By using this transformation on a data-dependent covering, $\mathcal{C}_n$, we get the partition $\mathcal{P}(\mathcal{C}_n)$ and the associated estimator (4). The major difference compared to an estimator defined on a data-dependent partition is its interpretability. Moreover, using a partition from a data-dependent covering in place of a data-dependent partition generates a more complex partition where cells are not necessarily conjunctions of tests as in (3). We illustrate it in Figure 3.

As the construction of a partition from a covering is time consuming, it is important to note that the partition $\mathcal{P}(\mathcal{C}_n)$ does not need to be constructed. The trick is to identify the unique cell of $\mathcal{P}(\mathcal{C}_n)$ which contains some $\mathbf{x} \in \mathbb{R}^d$ used for calculating the prediction at $\mathbf{x}$. By creating binary vectors of size $\#\mathcal{C}_n$, whose value is 1 if $\mathbf{x}$ fulfilled the rule's condition and 0 otherwise, this cell identification becomes a simple sequence of vectorial operations. Figure 3 is an illustration of this process (cf Margot et al. [2018] for more details).

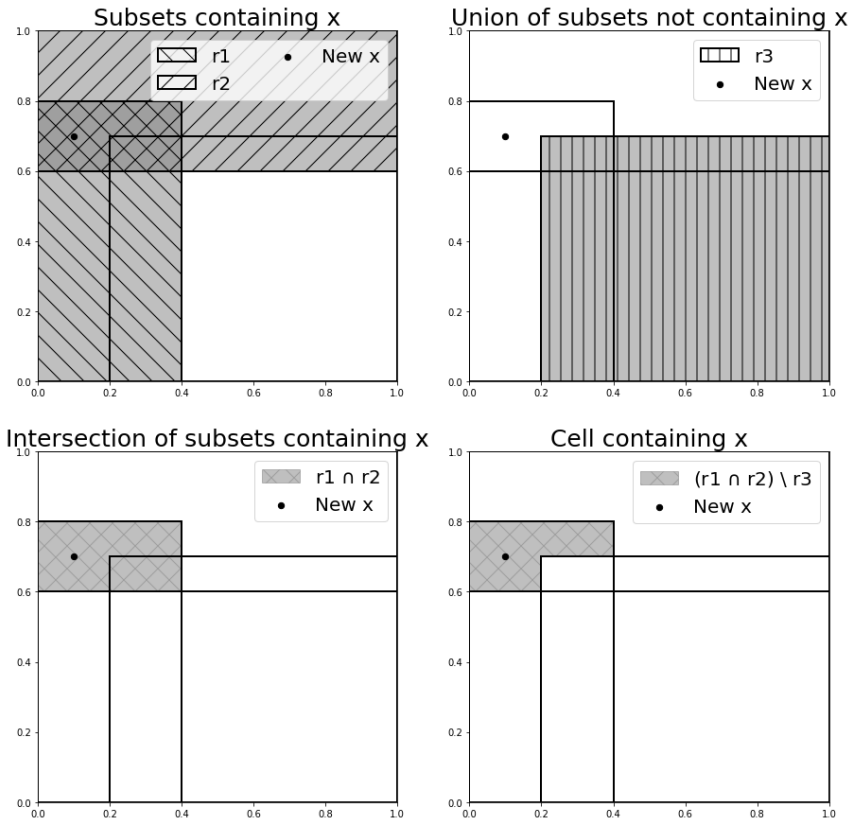

Figure 3: Evaluation steps of the cell containing $\mathbf{x} = (0.1, 0.7)$ of the partition generated from the covering of $[0,1]^2$, $\mathcal{C} = \{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3\}$. Using partition from a covering allows to generate complex cells with a simple interpretation ($\mathbf{r}_1$ *And* $\mathbf{r}_2$), where a classical partitioning algorithm cannot. Note that the condition $\mathbf{x}$ satisfies ($\mathbf{r}_1$ *And* $\mathbf{r}_2$) implicitly implies that $\mathbf{x}$ does not satisfy $\mathbf{r}_3$.

All the estimators generated by the data-dependent covering algorithm belong to the class

$$\mathcal{G}_n := \mathcal{G}_c \circ \mathcal{P}(\mathcal{C}_n) \tag{5}$$

of piecewise constant functions on the partition $\mathcal{P}(\mathcal{C}_n)$ such that $\forall g \in \mathcal{G}_n, \forall x \in \mathbb{R}^d, |g(x)| \leq L$.

Hence, from definitions (2) we have

$$g_n(\mathbf{x}) = \sum_{A \in \mathcal{P}(\mathcal{C}_n)} \frac{\sum_{i=1}^n Y_i \mathbf{1}_{X_i \in A}}{\sum_{i=1}^n \mathbf{1}_{X_i \in A}} \mathbf{1}_{\mathbf{x} \in A}\,, \qquad \mathbf{x} \in \mathbb{R}^d\,, \tag{6}$$

and the risk minimizer over $\mathcal{G}_n$ is

$$\tilde{g}_n(\mathbf{x}) = \sum_{A \in \mathcal{P}(\mathcal{C}_n)} \frac{\mathbb{E}[Y \mathbf{1}_{\mathbf{X} \in A}]}{\mathbb{P}(\mathbf{X} \in A)} \mathbf{1}_{\mathbf{x} \in A}\,, \qquad \mathbf{x} \in \mathbb{R}^d\,. \tag{7}$$

The functions g_n and $\tilde{g}_n$ are indeed both in $\mathcal{G}_n$, although the later is not computable from the data only.

Remark 3. The definition (6) of g_n guarantees that $\forall \mathbf{x} \in \mathbb{R}^d, |g_n(\mathbf{x})| \leq L$ so that L doesn't need to be known.

In the following Subsection we discuss about the important notion of *interpretability*.

1.2 Interpretability

In many fields, such as healthcare, marketing or asset management, decisions makers prefer an interpretable models rather than models with better accuracy but uninterpretable. As mentioned in Lipton [2017], there is no rigorous mathematical foundation of the concept. In this paper, *interpretability* correspond to parsimonious characterization of the estimators of g^* generated by a given algorithm, *i.e.* the facility to to describe the generated model in human words. Nowadays, the most popular and efficient algorithms for regression, such as Support Vector Machines, Neural networks, Random Forests,... are uninterpretable. The lack of interpretability comes from the complexity of the models they generate. We refer to them as black box models. Usually, these black box models have an optimal accuracy. We assert that the novel family of *covering algorithms* described here, can achieve a better Interpretability-Accuracy trade-off by reducing the complexity of the generated models keeping Accuracy guarantees, i.e. weak consistency.

There exist two ways of constructing interpretable models. The first one is to create black-box models and then to summarize them to create a so-called *post-hoc* interpretable algorithm. For example, recent researches propose to use explanation models, such as *LIME* [Ribeiro et al., 2016], *DeepLIFT* [Shrikumar et al., 2017] or *SHAP* [Lundberg and Lee, 2017], to interpret black-box models. These explanation models try to measure the importance of a feature (a coordinate of $\mathbf{X}$) on the prediction process (see Guidotti et al. [2018] for a survey of existing methods). The second way to interpretability is to use *intrinsic* interpretable algorithms, it means algorithms that only generate interpretable models, such as rule-based algorithms.

The interpretability of the rule-based estimator is achieved when the length k of each rule is small. Considering that an estimator with one rule of length

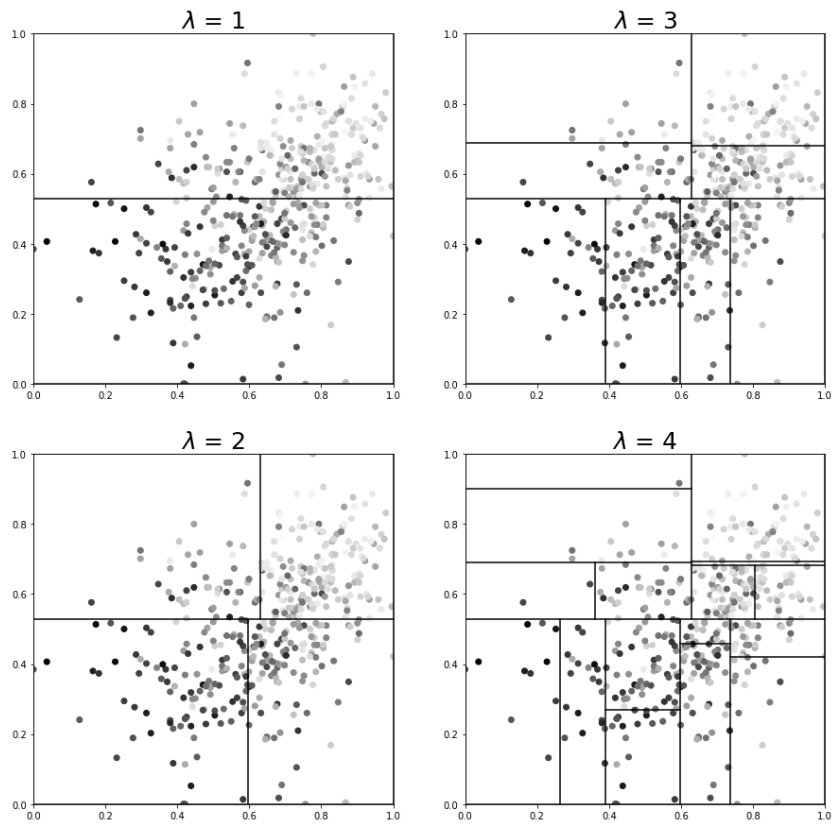

Figure 4: Partitions generated by fully deployed decision tree algorithm, it means without pruning, for a maximal depth $\lambda \in \{1, 2, 3, 4\}$.

k is as interpretable as the one with the corresponding k rules of length 1, the interpretability gets naturally the additivity property. With this in mind, we are able to quantify the interpretability of a estimator g_n generated by a set of rules C_n by the *interpretability score* defined by

$$Int(g_n) := \sum_{\mathbf{r} \in \mathcal{C}_n} \text{length}(\mathbf{r}). \tag{8}$$

However in order to prove the consistency of the estimator g_n, one usually applies results such as Theorem 13.1 in Györfi et al. [2006] under the condition of shrinkage of the cells (Condition 13.10 in Györfi et al. [2006]). Each rule (3) must have a length $k = d$ in order to fulfil this sufficient condition without extra condition on the feature space. Then, for large d, the condition becomes uninterpretable. Moreover, as illustrated in Figure 4, the number of cells necessary to have an accurate model is very large as the more precise the partition, the more complex the model.

For an estimator defined on a data-dependent covering, each prediction is explained by a small set of fulfilled rules which are easy to understand, see Table 3 in Section 4 for an example. Even if the partition generated may be finer and more complex than a classical data-dependent partition, the explanation of the prediction is given by the covering and not the partition, and it remains understandable by humans, as illustrated in Figure 3.

Despite the fact that the parsimony of the selected set of rules is not theoretically guaranteed, the redundancy conditions (11) and (12) described below are heading in the right direction.

We obtain a consistent estimator g_n by carefully constructing the covering elements. We can apply none of the classical approaches based on Stone's theorem Stone [1977] because the covering is data-dependent nor based on Theorem 13.1 in Györfi et al. [2006] as Condition 13.10 in Györfi et al. [2006] forces rules to be complex ($k = d$). The key notion of this paper is the notion of suitable data-dependent covering introduced in Section 2. Proposition 3.2 provides the main tool to prove the weak consistency of suitable data-dependent covering estimators stated in Theorem 2.1. This result of independent interest is given in Section 3. Finally we apply our approach on covering elements using Random Forest as rule generator in Section 4. Supplementary material gathers the proof of Proposition 3.2.

2 Main result

We denote $\mathbb{P}_n$ the empirical distribution associated to the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$. For any $\mathbf{r} \subseteq \mathbb{R}^d$ such that $\mathbb{P}_n(\mathbf{r}) > 0$, we also denote

$$\mathbb{E}_n[Y \mid \mathbf{X} \in \mathbf{r}] := \frac{\frac{1}{n}\sum_{i=1}^n Y_i \mathbf{1}_{\mathbf{X}_i \in \mathbf{r}}}{\mathbb{P}_n(\mathbf{r})}$$

and

$$\mathbb{V}_n(Y \mid \mathbf{X} \in \mathbf{r}) := \mathbb{E}_n\left[Y^2 \mid \mathbf{X} \in \mathbf{r}\right] - \mathbb{E}_n\left[Y \mid \mathbf{X} \in \mathbf{r}\right]^2.$$

In the same way, we define $\mathbb{E}[Y|\mathbf{X} \in \mathbf{r}] := \frac{\mathbb{E}[Y\mathbf{1}_{\mathbf{X}\in\mathbf{r}}]}{\mathbb{P}(\mathbf{X}\in\mathbf{r})}$ and $\mathbb{V}[Y|\mathbf{X} \in \mathbf{r}] := \mathbb{E}[Y^2|\mathbf{X} \in \mathbf{r}] - (\mathbb{E}[Y|\mathbf{X} \in \mathbf{r}])^2$.

2.1 Significance and coverage conditions

We introduce some conditions on each element of the covering. We use the classical notation $x_+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$.

Definition 2.1. *We call a sequence $(\mathcal{C}_n)_{n\geq 1}$ of data-dependent coverings of $\mathbb{R}^d$* ***suitable*** *if it satisfies the two following conditions:*

1. *the* ***coverage condition****:* *(H3)*

$$\exists\ \alpha \in [0, 1/2),\ \forall \mathbf{r} \in \mathcal{C}_n,\ \mathbb{P}_n(\mathbf{r}) >\ n^{-\alpha} \quad a.s., \tag{9}$$

for n sufficiently large;

2. *the* ***significance condition****:* *(H4)*

there exists two sequences $\beta_n \to 0$ *and* $\varepsilon_n \to 0$ *such that:*

$$\mathcal{C}_n = \mathcal{C}_n^s \cup \mathcal{C}_n^i \quad a.s., \tag{10}$$

for n *sufficiently large, where the significant subsets* $\mathcal{C}_n^s$ *are defined by*

$$\mathcal{C}_n^s := \Big\{ \mathbf{r} \in \mathcal{C}_n : \beta_n \big| \mathbb{E}_n[Y|\mathbf{X} \in \mathbf{r}] - \mathbb{E}_n[Y] \big| \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma^2)_+} \Big\}, \tag{11}$$

the insignificant subsets $\mathcal{C}_n^i$ *are defined by*

$$\mathcal{C}_n^i := \Big\{ \mathbf{r} \in \mathcal{C}_n \setminus \mathcal{C}_n^s : \varepsilon_n \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma^2)_+} \Big\}, \tag{12}$$

and their redundancies satisfy

$$\frac{M(\mathcal{C}_n^s)}{m(\mathcal{C}_n^s)} = o_{\mathbb{P}}(\beta_n^{-2} \wedge n^{1/2-\alpha}) \tag{13}$$

and

$$\frac{M(\mathcal{C}_n^i)}{m(\mathcal{C}_n^i)} = o_{\mathbb{P}}(\varepsilon_n^{-2} \wedge n^{1/2-\alpha}) \quad . \tag{14}$$

The coverage condition (H3) guarantees that the empirical within group expectation is a good estimation of the within group expectation. Up to our knowledge, the definitions of significant and insignificant elements of a covering in (H4) are new. An element fulfills the significance condition (11) if its conditional expectation is sufficiently different from the unconditional expectation. It ensures, in some sense, that the *within-group variances* of coverings with significant elements is controlled by the *between-group variances*. The insignificant condition (12) guarantees that the conditional variance of the insignificant elements shrinks to the noise variance. Both conditions (H3) and (H4) can be checked for each element of the covering separately. Thus the construction of such subsets can be parallelized which allow imagining algorithms less complex in comparison of usual ones.

Remark 4. An easy way to ensure (13) and (14) is to avoid inclusion between elements of the covering. Let $(\mathcal{C}_n)$ be a sequence of coverings that fulfills (H3). We consider $1 \leq i \leq \#\mathcal{C}_n$ any ordering of the covering. If

$$\mathbb{P}_n\Big(\mathbf{r}_i \bigcap \Big\{ \bigcup_{1 \leq j \leq i-1} \mathbf{r}_j \Big\}\Big) \leq \gamma\, \mathbb{P}_n(\mathbf{r}_i)\,, \qquad 1 \leq i \leq \#\mathcal{C}_n\,.$$

then the cardinal of $\mathcal{C}_n$ is upper bounded by $\frac{n^\alpha}{1-\gamma}$ for every n sufficiently large. Indeed, by the inclusion-exclusion principle we get

$$1 = \mathbb{P}_n(\mathcal{C}_n) = \sum_{i=1}^{\#\mathcal{C}_n} \mathbb{P}_n(\mathbf{r}_i \setminus \cup_{1 \leq j \leq i-1} \mathbf{r}_j) \geq \#\mathcal{C}_n\,(1-\gamma)\, n^{-\alpha}\,,$$

Thus (13) and (14) can be checked for any $\alpha \in [0, 1/4)$, using the fact that $M(\mathcal{C}_n^s)$ and $M(\mathcal{C}_n^i)$ are smaller than $\frac{n^\alpha}{1-\gamma}$ and setting $\beta_n = o_{\mathbb{P}}(n^{1/4-\alpha/2})$ and $\varepsilon_n = o_{\mathbb{P}}(n^{1/4-\alpha/2})$.

Example 1. The significant condition (11) can hold for a subset $\mathbf{r}$ with arbitrary diameter that does not satisfy Condition 13.10 of Györfi et al. [2006]. For instance, consider the case $g^* = \mathbf{1}_{x \in A}$ for some Borel set A such that $0 < \mathcal{P}(\mathbf{X} \in A) < 1$. Then $\mathbf{r} = A$ is a significant subset as it satisfies the condition (11) with high probability for any β_n such that $n^{-1/4} = o(\beta_n)$ and n sufficiently large. Indeed, from the Strong Law of Large Numbers $k_n := \#\{X_i \in A\} \sim n\mathcal{P}(\mathbf{X} \in A)$ a.s. as $n \to \infty$. On the one hand, we obtain thanks to several applications of the Central Limit Theorem

$$\begin{aligned}
\big|\mathbb{E}_n[Y|\mathbf{X} \in A] - \mathbb{E}_n[Y]\big| &\geq \mathbb{E}_n[Y|\mathbf{X} \in A] - \mathbb{E}_n[Y] \\
&= 1 - \frac{k_n}{n} + \frac{1}{k_n}\sum_{i=1}^n Z_i \mathbf{1}_{\mathbf{X}_i \in A} - \frac{1}{n}\sum_{i=1}^n Z_i \\
&= 1 - \mathbb{P}(\mathbf{X} \in A) + O_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

On the other hand, we obtain

$$\begin{aligned}
(\mathbb{V}_n(Y|\mathbf{X} \in A) - \sigma^2)_+ &\leq |\mathbb{V}_n(Y|\mathbf{X} \in A) - \sigma^2| \\
&= \left|\frac{1}{k_n}\sum_{i=1}^n Z_i^2 \mathbf{1}_{\mathbf{X}_i \in A} - \Big(\frac{1}{k_n}\sum_{i=1}^n Z_i \mathbf{1}_{\mathbf{X}_i \in A}\Big)^2 - \sigma^2\right| \\
&= O_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

Then

$$\begin{aligned}
\beta_n\big|\mathbb{E}_n[Y|\mathbf{X} \in A] - \mathbb{E}_n[Y]\big| - \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in A) - \sigma^2)_+} \\
\geq \beta_n\big(1 - \mathbb{P}(\mathbf{X} \in A) + O_{\mathbb{P}}(n^{-1/2}) + O_{\mathbb{P}}(\beta_n^{-1} n^{-1/4})\big)
\end{aligned}$$

Thus (11) holds for $\mathbf{r} = A$ with high probability for n sufficiently large. Note that for similar reasons (11) also holds with high probability for $\mathbf{r} = A^c$, $n^{-1/4} = o(\beta_n)$ and n sufficiently large. Finally, conditions (9), (13) and (14) are easily checked on the partitions $\mathcal{C}_n = \mathcal{P}_n = \{A, A^c\}$ that constitute a suitable coverings sequence with high probability for n large enough.

Remark 5. The significant condition (11) does not follow from a condition on the diameter of the subset. On the opposite, the insignificant condition (12) can follow from a condition on the diameter of the subset, see Proposition 3.3.

2.2 Partitioning number

To control the complexity of families of partitions, some tools introduced in [Nobel, 1996, Sec. 1.2] are recalled (see also [Györfi et al., 2006, Def 13.1]).

Definition 2.2. *Let Π be a family of partitions of $\mathbb{R}^d$.*

1. *The maximal number of cells in a partition of Π is denoted by*

$$\mathcal{M}(\Pi) := \sup \{\#\mathcal{P} : \mathcal{P} \in \Pi\}.$$

2. *For a set $\mathbf{x}_1^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in (\mathbb{R}^d)^n$, let*

$$\Delta(\mathbf{x}_1^n, \Pi) := \#\big\{\{\mathbf{x}_1^n \cap A : A \in \mathcal{P}\} : \mathcal{P} \in \Pi\big\}$$

   be the number of distinct partitions of $\mathbf{x}_1^n$ induced by elements of Π.

3. *The* partitioning number $\Delta_n(\Pi)$ *of Π is defined by:*

$$\Delta_n(\Pi) := \max_{\mathbf{x}_1^n \in (\mathbb{R}^d)^n} \Delta(\mathbf{x}_1^n, \Pi).$$

The partitioning number is the maximal number of different partitions of any n points set that can be induced by elements of Π.

2.3 Consistency of data-dependent covering algorithms

In the following, we use the classical notion of Donsker class that is discussed in details in Section 3.

Theorem 2.1. *Assume that $\mathbb{Q}$ satisfies (H1) and (H2). Let $(\mathcal{C}_n)$ be a suitable data-dependent covering sequence (i.e. it satisfies (H3) and (H4)) fulfilling the two following conditions:*

$$\mathcal{M}(\Pi_n) \vee \log(\Delta_n(\Pi_n)) = o(n), \tag{H5}$$

where $\Pi_n := \{\mathcal{P}(\mathcal{C}_n(d_n)) : d_n \in \mathcal{S}^n\}$ for any $n \in \mathbb{N}^*$;

$$\forall n \in \mathbb{N}^*, \{\mathbf{c} \times \mathbb{R}, \mathbf{c} \in \mathcal{C}_n\} \subseteq \mathcal{B}, \tag{H6}$$

where $\mathcal{B}$ is a $\mathbb{Q}$-Donsker class.

Then the estimator g_n defined by (6) *is weakly consistent:*

$$\ell(g^*, g_n) = o_{\mathbb{P}}(1). \tag{15}$$

The proof of this theorem is postponed to Section 3.

This theorem gives us conditions on data-dependent covering algorithms to ensure that the generated empirical risk minimizer g_n converges in probability to the regression function g^* defined in (1). The condition (H5) is a classical one (*e.g.* [Györfi et al., 2006, Conditions (13.7) and (13.8)]) used to ensure that the family of partitions Π_n is not too "complex". It means that the maximal number of cells in a partition, and the logarithm of the partitioning number, are small compared to the sample size. This condition guarantees that the estimation error tends to 0. The conditions (H3), (H4) and (H6) guarantee that the approximation error tends to 0 without any condition on the diameter of the cells.

3 Proof of Theorem 2.1

In order to prove the main theorem, we need some preliminary results based on notions of $\mathbb{Q}$-Donsker class and outer probability.

The outer probability, defined for $A \subseteq \Omega$ by $\mathbb{P}^*(A) := \inf\left\{\mathbb{P}(\tilde{A}) : A \subset \tilde{A}, \tilde{A} \in \mathcal{A}\right\}$ is introduced to handle functions which are not necessarily measurable. The notation $O_{\mathbb{P}^*}(1)$ stands for *asymptotically tight* instead of the usual $O_{\mathbb{P}}(1)$ (*bounded in probability*). See [Van der Vaart, 2000, Chapter 18].

Let us define for any $f : \mathcal{S} \to \mathbb{R}$ in $\mathcal{L}^1(\mathbb{Q})$, $v_n f := \sqrt{n}(\mathbb{Q}_n f - \mathbb{Q} f)$ and consider the empirical process indexed by a set $\mathcal{F}$ of such functions: $\{v_n f : f \in \mathcal{F}\}$.

Definition 3.1. *[Van der Vaart, 2000, Section 19.2] $\mathcal{F}$ is called $\mathbb{Q}$-Donsker if the sequence of processes $\{v_n f : f \in \mathcal{F}\}$ converges in distribution to a tight limit process in the space $\ell^\infty(\mathcal{F})$.*

The limit process is then a $\mathbb{Q}$-Brownian bridge.

Definition 3.2. *A class of sets $\mathcal{B} \subseteq \mathcal{B}_{\mathcal{S}}$ is called $\mathbb{Q}$-Donsker if $\mathcal{I}_{\mathcal{B}} := \{\mathbf{1}_A : A \in \mathcal{B}\}$ is a $\mathbb{Q}$-Donsker class of functions.*

Now, with $\mathbb{Q}f := \int f d\mathbb{Q}$ and $\mathbb{Q}_n f := \int f d\mathbb{Q}_n$, for any $f \in \mathcal{L}^1(\mathbb{Q})$, if $\mathcal{F}$ is a $\mathbb{Q}$-Donsker class of functions, then the empirical process $((\sqrt{n}(\mathbb{Q}_n f - \mathbb{Q} f))_{f\in\mathcal{F}})_{n\in\mathbb{N}}$ is asymptotically tight as a sequence of maps with values in $\ell^\infty(\mathcal{F})$ (this is a consequence of Prohorov's Theorem adapted to this framework – see Theorem 18.12 in Van der Vaart [2000]). Keeping in mind that a compact set in $\ell^\infty(\mathcal{F})$ is bounded, we have:

Proposition 3.1. *Let $\mathcal{F}$ be a $\mathbb{Q}$-Donsker class of functions. Then*

$$\|\mathbb{Q}_n - \mathbb{Q}\|_{\mathcal{F}} = O_{\mathbb{P}^*}(n^{-1/2}),$$

where for any $v : \mathcal{F} \to \mathbb{R}$, $\|v\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |v(f)|$.

Remark 6. If $\mathcal{B} \subseteq \mathcal{B}_{\mathcal{S}}$ is a $\mathbb{Q}$-Donsker class of sets, where $\mathcal{B}_{\mathcal{S}}$ is the Borel set on $\mathcal{S}$, then

$$\|\mathbb{Q}_n - \mathbb{Q}\|_{\mathcal{B}} = O_{\mathbb{P}^*}(n^{-1/2}),$$

where for any $v : \mathcal{B} \to \mathbb{R}$, $\|v\|_{\mathcal{B}} = \sup_{A\in\mathcal{B}} |v(A)|$.

Remark 7. It can be checked that if $(Z_n)_{n\in\mathbb{N}}$ is a sequence of non-negative random variables, $(a_n)_{n\in\mathbb{N}} \in (\mathbb{R}^+)^{\mathbb{N}}$ such that $a_n = o_{\mathbb{P}}(1)$ and $(M_n)_{n\in\mathbb{N}}$ is a sequence of maps such that $M_n = O_{\mathbb{P}^*}(1)$ and $Z_n \leq a_n M_n$ for any n, then $Z_n \xrightarrow[n\to+\infty]{\mathbb{P}} 0$.

The usual notion of boundedness in probability for sequences of random variables need be generalized because sequences of maps are to be considered, with values in metric spaces which are not Euclidean spaces (thus bounded and closed sets need not be compact) and which are not guaranteed to be measurable. We need involve the outer probability $\mathbb{P}^*$.

Definition 3.3. *[Van der Vaart, 2000, Chapter 18] A sequence $(M_n)_{n\in\mathbb{N}}$ of maps defined on Ω and with values in a metric space $(\mathbb{D}, d)$ is said to be asymptotically tight if*

$$\forall \varepsilon > 0, \exists K \subset \mathbb{D} \text{ compact}/\forall \delta > 0, \limsup_{n\to\infty} \mathbb{P}^*(M_n \notin K^\delta) < \varepsilon,$$

with $K^\delta = \{y \in \mathbb{D} : d(y, K) < \delta\}$ for any $K \subset \mathbb{D}$ and $\delta > 0$.

Remark 8. If $\mathbb{D} = \mathbb{R}$, (M_n) is asymptotically tight if and only if

$$\forall \varepsilon > 0, \exists M > 0/ \limsup_{n\to\infty} \mathbb{P}^*(|M_n| > M) < \varepsilon.$$

3.1 Empirical estimation of conditional expectations

We shall also use the following proposition, which is inspired by the work of Grunewalder [2018] (Proposition 3.2).

Proposition 3.2. *Let $\mathcal{B} \subseteq \mathcal{B}_\mathcal{S}$ and let $\mathcal{F}_\mathcal{B} := \{f\mathbf{1}_A : f \in \mathcal{F}, A \in \mathcal{B}\}$ where $\mathcal{F}$ is a set of functions in $\mathcal{L}^1(\mathbb{Q})$ uniformly bounded. If $\mathcal{B}$ and $\mathcal{F}_\mathcal{B}$ are $\mathbb{Q}$-Donsker classes then for any $\alpha \in [0, 1/2)$ and with $\mathcal{B}_n := \{A \in \mathcal{B}, \mathbb{Q}_n(A) \geq n^{-\alpha}\}$,*

$$\sup_{f\in\mathcal{F}} \sup_{A\in\mathcal{B}_n} \left|\mathbb{E}_n\left[f \mid A\right] - \mathbb{E}\left[f \mid A\right]\right| = O_{\mathbb{P}^*}(n^{\alpha-1/2}).$$

Corollary 3.1. *Let $\mathcal{B} \subseteq \mathcal{B}_\mathcal{S}$ be a $\mathbb{Q}$-Donsker class. If Y is bounded then for any $i \in \mathbb{N}$ and any $\alpha \in [0, 1/2)$, with $\mathcal{B}_n := \{A \in \mathcal{B}, \mathbb{Q}_n(A) \geq n^{-\alpha}\}$ we have*

$$\sup_{A\in\mathcal{B}_n} \left|\mathbb{E}_n\left[Y^i \mid (\mathbf{X}, Y) \in A\right] - \mathbb{E}\left[Y^i \mid (\mathbf{X}, Y) \in A\right]\right| = O_{\mathbb{P}^*}(n^{\alpha-1/2}), \tag{16}$$

and

$$\sup_{A\in\mathcal{B}_n} \left|\mathbb{V}_n\left[Y \mid (\mathbf{X}, Y) \in A\right] - \mathbb{V}\left[Y \mid (\mathbf{X}, Y) \in A\right]\right| = O_{\mathbb{P}^*}(n^{\alpha-1/2}). \tag{17}$$

Proofs of these results are in Appendix.

It seems that the result of Corollary 3.1, which is of independent interest, does not appear as such in the existing literature. As a first application of Corollary 3.1, we show that any partition with shrinking cells diameters is a suitable covering. We define the diameter of a cell $\mathbf{r}$ as $\mathrm{Diam}(\mathbf{r}) = \sup_{x\in\mathbf{r},\, x'\in\mathbf{r}} \|x - x'\|$, where $\|\cdot\|$ is any norm of $\mathbb{R}^d$.

Proposition 3.3. *Consider a sequence $(\mathcal{P}_n)_{n\in\mathbb{N}}$ of data-dependent partitions, that satisfies the coverage condition* (9) *and such that*

$$\bigcup_{n\in\mathbb{N}^*} \bigcup_{\mathbf{r}\in\mathcal{P}_n} (\mathbf{r} \times \mathbb{R})$$

is a.s. a $\mathbb{Q}$-Donsker class. If g^* is uniformly continuous and if

$$\max_{\mathbf{r}\in\mathcal{P}_n} Diam(\mathbf{r}) = o_\mathbb{P}(1) \tag{18}$$

then the sequence $(\mathcal{P}_n)$ is suitable.

Proof. Let us show that each cell is significant or insignificant. Thanks to Condition (9), Corollary 3.1 Eq. (17) and Remark 7,

$$\max_{\mathbf{r}\in\mathcal{P}_n} |\mathbb{V}_n(Y|\mathbf{X}\in\mathbf{r}) - \mathbb{V}(Y \mid \mathbf{X}\in\mathbf{r})| = O_{\mathbb{P}}(n^{\alpha-1/2}). \tag{19}$$

Moreover $\mathbb{V}(Y \mid \mathbf{X}\in\mathbf{r}) = \mathbb{V}(g^*(\mathbf{X}) \mid \mathbf{X}\in\mathbf{r}) + \sigma^2$. Thus, as the redundancy condition (14) is automatically satisfied for cells of a partition, the desired result will follow if we check that

$$\varepsilon_n := \max_{\mathbf{r}\in\mathcal{P}_n} \sqrt{(\mathbb{V}_n(Y|\mathbf{X}\in\mathbf{r}) - \sigma^2)_+}$$

converges to 0.

From (19) we remark that

$$\varepsilon_n \le \max_{\mathbf{r}\in\mathcal{P}_n} \sqrt{\mathbb{V}(g^*(\mathbf{X}) \mid \mathbf{X}\in\mathbf{r})} + O_{\mathbb{P}}(n^{\alpha/2-1/4}) \,.$$

For all n, if $\mathbf{r}\in\mathcal{P}_n$, then $\mathbf{r}\times\mathbb{R}\in\mathcal{B}_{\mathcal{S}}$. We denote $\mathbf{X}_\mathbf{r}$ and $\mathbf{X}'_\mathbf{r}$ two independent variables distributed as $\mathbf{X}$ given that $\mathbf{X}\in\mathbf{r}$. We obtain

$$\begin{aligned}\mathbb{V}(g^*(\mathbf{X}) \mid \mathbf{X}\in\mathbf{r}) &= \mathbb{V}(g^*(\mathbf{X}_\mathbf{r})) \\ &= \tfrac{1}{2}\mathbb{V}\big(g^*(\mathbf{X}_\mathbf{r}) - g^*(\mathbf{X}'_\mathbf{r})\big) \\ &\le \tfrac{1}{2}\mathbb{E}\left[(g^*(\mathbf{X}_\mathbf{r}) - g^*(\mathbf{X}'_\mathbf{r}))^2\right] \,.\end{aligned}$$

Thus, if we denote w the modulus of continuity of g^*, we get

$$\sqrt{\mathbb{V}(g^*(\mathbf{X}) \mid \mathbf{X}\in\mathbf{r})} \le 2^{-1/2} w(\mathrm{Diam}(\mathbf{r})).$$

By uniform continuity, the condition (18) implies that

$$\varepsilon_n \le 2^{-1/2} \max_{\mathbf{r}\in\mathcal{P}_n} \Big(w(\mathrm{Diam}(\mathbf{r})) \Big) + O_{\mathbb{P}}(n^{\alpha/2-1/4}) = o_{\mathbb{P}}(1) \,.$$

Thus, from (12), each cell which is not significant is insignificant and the corresponding covering sequence is suitable. □

Remark 9. The condition of uniform continuity of g^* in Proposition 3.3 may be simply raised. Indeed, from [Györfi et al., 2006, Corollary A.1], g^* can be approximated arbitrarily closely in $\mathcal{L}_2(\mathbb{Q}_\mathbf{X})$ by functions of $C_0^\infty(\mathbb{R}^d)$ where $\mathbb{Q}_\mathbf{X}$ is the marginal distribution of $\mathbf{X}$.

3.2 Estimation-approximation decomposition

The excess risk (1) can be decomposed into two terms using the following lemma:

Lemma 3.1 (Lemma 10.1 of Györfi et al. [2006])**.** *Let $\mathcal{G}_n$ be a class of functions $g : \mathbb{R}^d \to [-L, L]$ depending on the data $D_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$. If g_n satisfies* (2) *then*

$$\ell\left(g^*, g_n\right) \le 2 \sup_{g\in\mathcal{G}_n} \left|\mathcal{L}_n\left(g\right) - \mathcal{L}\left(g\right)\right| + \inf_{g\in\mathcal{G}_n} \mathbb{E}\left[\left(g(\mathbf{X}) - g^*(\mathbf{X})\right)^2\right].$$

Hence, to prove (15) it is sufficient to prove that:

$$\sup_{g \in \mathcal{G}_n} |\mathcal{L}_n(g) - \mathcal{L}(g)| = o_{\mathbb{P}}(1). \tag{20}$$

and

$$\inf_{g \in \mathcal{G}_n} \mathbb{E}\left[\left(g(\mathbf{X}) - g^*(\mathbf{X})\right)^2\right] = o_{\mathbb{P}}(1). \tag{21}$$

The *estimation error* (20) controls the distance between the best function in $\mathcal{G}_n$ and g_n. The *approximation error* (21) is the smallest error for a function of $\mathcal{G}_n$.

The two terms have opposite behaviors. Indeed, if $\mathcal{G}_n$ is not too complex the empirical risk will be close to the risk uniformly over $\mathcal{G}_n$. Thus, the error due to the minimization of the empirical risk instead of the risk will be small. On the other hand, the risk cannot be better than for the best function of $\mathcal{G}_n$. So, $\mathcal{G}_n$ must be complex enough. It is the classical *Bias/Variance* or *Approximation/Estimation* trade-off.

3.3 Approximation Error

In this subsection, we prove (21) using hypotheses (H1), (H2), (H3), (H4) and (H6).

The function $\tilde{g}_n$ is in $\mathcal{G}_n$, thus to prove (21), it suffices to show that $\mathbb{W}_n = o_{\mathbb{P}}(1)$ where

$$\mathbb{W}_n := \mathbb{E}\left[\left(\tilde{g}_n(\mathbf{X}) - g^*(\mathbf{X})\right)^2\right]$$

From (7),

$$\mathbb{W}_n = \mathbb{E}\left[\left(\sum_{A \in \mathcal{P}(\mathcal{C}_n)} \mathbb{E}\left[Y \mid \mathbf{X} \in A\right] \mathbf{1}_{\mathbf{X} \in A} - g^*(\mathbf{X})\right)^2\right]$$

which shows that $\mathbb{W}_n$ is a within-group variance for the variable $g^*(\mathbf{X})$ and the groups $\mathcal{P}(\mathcal{C}_n)$.

First we use the decomposition of the total variance into the sum of the within-group and the between-group variances:

$$\begin{aligned}
\mathbb{W}_n &= \sum_{A' \in \mathcal{P}(\mathcal{C}_n)} \mathbb{E}\left[\left(\sum_{A \in \mathcal{P}(\mathcal{C}_n)} \mathbb{E}\left[Y \mid \mathbf{X} \in A\right] \mathbf{1}_{\mathbf{X} \in A} - g^*(\mathbf{X})\right)^2 \mathbf{1}_{X \in A'}\right] \\
&= \sum_{A' \in \mathcal{P}(\mathcal{C}_n)} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X} \in A'\right] - g^*(\mathbf{X})\right)^2 \mathbf{1}_{X \in A'}\right] \\
&= \sum_{A' \in \mathcal{P}(\mathcal{C}_n)} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X} \in A'\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X} \in A'\right] \mathbb{P}\left(\mathbf{X} \in A'\right) && (22) \\
&= \mathbb{V}(g^*(\mathbf{X})) - \mathbb{B}_n, && (23)
\end{aligned}$$

where

$$\mathbb{B}_n := \sum_{A \in \mathcal{P}(\mathcal{C}_n)} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in A\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in A\right). \tag{24}$$

Let's consider $\mathbb{B}_n$ and replace the summation over the partition $\mathcal{P}(\mathcal{C}_n)$ by a summation over the covering $\mathcal{C}_n$. We have, from the definition of $M(\mathcal{C}_n, \mathbf{r})$,

$$\begin{aligned}
\mathbb{B}_n &= \sum_{\mathbf{r} \in \mathcal{C}_n} \sum_{A \in \mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \frac{1}{\#\varphi_{\mathcal{C}_n}(A)} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in A\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in A\right) \\
&\geq \sum_{\mathbf{r} \in \mathcal{C}_n} \frac{1}{M(\mathcal{C}_n, \mathbf{r})} \sum_{A \in \mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in A\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in A\right) \\
&\geq \sum_{\mathbf{r} \in \mathcal{C}_n} \frac{1}{M(\mathcal{C}_n, \mathbf{r})} \times \\
&\qquad \sum_{A \in \mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in A\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in A \mid \mathbf{X} \in \mathbf{r}\right) \mathbb{P}\left(\mathbf{X} \in \mathbf{r}\right) \\
&\geq \sum_{\mathbf{r} \in \mathcal{C}_n} \frac{1}{M(\mathcal{C}_n, \mathbf{r})} \times \\
&\qquad \left(\sum_{A \in \mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \mathbb{E}\left[Y \mid \mathbf{X} \in A\right] \mathbb{P}\left(\mathbf{X} \in A \mid \mathbf{X} \in \mathbf{r}\right) - \mathbb{E}\left[Y\right] \right)^2 \mathbb{P}\left(\mathbf{X} \in \mathbf{r}\right) \\
&\geq \sum_{\mathbf{r} \in \mathcal{C}_n} \frac{1}{M(\mathcal{C}_n, \mathbf{r})} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in \mathbf{r}\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in \mathbf{r}\right)
\end{aligned}$$

where we last applied Jensen's inequality.

Now, we focus on the set $\mathcal{C}_n^s$ of significant elements of the covering. Since $\mathcal{C}_n = \mathcal{C}_n^s \cup \mathcal{C}_n^i$, we have

$$\begin{aligned}
\mathbb{B}_n &\geq \sum_{\mathbf{r} \in \mathcal{C}_n^s} \frac{1}{M(\mathcal{C}_n, \mathbf{r})} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in \mathbf{r}\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in \mathbf{r}\right) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} \left(\mathbb{E}\left[Y \mid \mathbf{X} \in \mathbf{r}\right] - \mathbb{E}\left[Y\right]\right)^2 \mathbb{P}\left(\mathbf{X} \in \mathbf{r}\right) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} U_{\mathbf{r}}^2 \, \mathbb{P}(\mathbf{X} \in \mathbf{r})
\end{aligned} \tag{25}$$

where

$$U_{\mathbf{r}} := \mathbb{E}\left[Y \mid \mathbf{X} \in \mathbf{r}\right] - \mathbb{E}\left[Y\right] \ .$$

Let's define

$$V_{n,\mathbf{r}} := \mathbb{E}_n\left[Y \mid \mathbf{X} \in \mathbf{r}\right] - \mathbb{E}_n\left[Y\right]$$

the empirical counterpart of U_n,

$$\Delta_{n,\mathbf{r}} := V_{n,\mathbf{r}}^2 - U_{\mathbf{r}}^2$$

and

$$\Delta_n := \sup_{\mathbf{r}\in\mathcal{C}_n} \{\Delta_{n,\mathbf{r}}\} \ .$$

In order to control $\mathbb{B}_n$ with its empirical counterpart, we shall make use of the outer probability $\mathbb{P}^*$ defined in Section 3. Using hypotheses (H2) and (H6) and Corollary 3.1 (with $\mathcal{B}_n = \{\mathbf{c}\times[-L,L], \mathbf{c}\in\mathcal{C}_n\}$) we have :

$$\begin{aligned}\Delta_n &= \sup_{\mathbf{r}\in\mathcal{C}_n} \{(V_{n,\mathbf{r}} - U_{\mathbf{r}})(V_{n,\mathbf{r}} + U_{\mathbf{r}})\} \\ &= O_{\mathbb{P}^*}(n^{\alpha-1/2}) \ . \end{aligned} \tag{26}$$

Continuing (25),

$$\begin{aligned}\mathbb{B}_n &\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r}\in\mathcal{C}_n^s} \left(V_{n,\mathbf{r}}^2 - \Delta_{n,\mathbf{r}}\right) \mathbb{P}(X\in\mathbf{r}) \\ &\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r}\in\mathcal{C}_n^s} \left(V_{n,\mathbf{r}}^2 \ - \Delta_n\right) \mathbb{P}(X\in\mathbf{r}) \ .\end{aligned}$$

By definition of $\mathcal{C}_n^s$, $\forall \mathbf{r}\in\mathcal{C}_n^s$,

$$V_{n,\mathbf{r}}^2 \geq \beta_n^{-2}\big(\mathbb{V}_n(Y|X\in\mathbf{r}) - \sigma^2\big)_+ \geq \beta_n^{-2}\big(\mathbb{V}_n(Y|X\in\mathbf{r}) - \sigma^2\big).$$

Thus

$$\mathbb{B}_n \geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r}\in\mathcal{C}_n^s} \left(\beta_n^{-2}(\mathbb{V}_n(Y|\mathbf{X}\in\mathbf{r}) - \sigma^2) - \Delta_n\right) \mathbb{P}(X\in\mathbf{r}) \ .$$

Let

$$\Delta_n' := \sup_{\mathbf{r}\in\mathcal{C}_n} \left\{\left|\mathbb{V}(Y|\mathbf{X}\in\mathbf{r}) - \mathbb{V}_n(Y|\mathbf{X}\in\mathbf{r})\right|\right\}$$

Using again Corollary 3.1 leads to

$$\Delta_n' = O_{\mathbb{P}^*}(n^{\alpha-1/2}) \tag{27}$$

Thus,

$$\mathbb{B}_n \geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r}\in\mathcal{C}_n^s} \left(\beta_n^{-2}(\mathbb{V}(Y|\mathbf{X}\in\mathbf{r}) - \sigma^2 - \Delta_n') - \Delta_n\right) \ \mathbb{P}(X\in\mathbf{r}) \ .$$

By independence between Z and $\mathbf{X}$, we have

$$\begin{aligned}\mathbb{V}(Y|\mathbf{X}\in\mathbf{r}) - \sigma^2 &= \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y \mid \mathbf{X}\in\mathbf{r}\right]\right)^2 \mid \mathbf{X}\in\mathbf{r}\right] - \sigma^2 \\ &= \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in\mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in\mathbf{r}\right]. \end{aligned} \tag{28}$$

Hence we have

$$\begin{aligned}
\mathbb{B}_n &\geq \frac{\beta_n^{-2}}{M(\mathcal{C}_n^s)} \times \\
&\quad \sum_{\mathbf{r}\in\mathcal{C}_n^s} \left(\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in\mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in\mathbf{r} \right] - (\Delta_n' + \beta_n^2 \Delta_n) \right) \mathbb{P}(\mathbf{X}\in\mathbf{r}) \\
&\geq \frac{\beta_n^{-2}}{M(\mathcal{C}_n^s)} \times \\
&\quad \sum_{\mathbf{r}\in\mathcal{C}_n^s} \sum_{A\in\mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \left(\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in\mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in A \right] \right) \mathbb{P}(\mathbf{X}\in A) \\
&\quad -(\beta_n^{-2}\Delta_n' + \Delta_n)
\end{aligned}$$

since $\sum_{\mathbf{r}\in\mathcal{C}_n^s} \mathbb{P}(\mathbf{X}\in\mathbf{r}) \leq M(\mathcal{C}_n^s)$. Thus, by definition of $m(\mathcal{C}_n^s)$,

$$\begin{aligned}
\mathbb{B}_n &\geq \beta_n^{-2} \frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)} \times \\
&\quad \sum_{A\in\mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^s)} \left(\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in\mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in A \right] \right) \mathbb{P}(\mathbf{X}\in A) \\
&\quad -(\beta_n^{-2}\Delta_n' + \Delta_n) \qquad (29)
\end{aligned}$$

We remark that

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in\mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in A \right] &= \mathbb{E}\left[\left(\mathbb{E}\left[g^*(\mathbf{X}) \mid \mathbf{X}\in\mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in A \right] \\
&\geq \mathbb{E}\left[\left(\mathbb{E}\left[g^*(\mathbf{X}) \mid \mathbf{X}\in A\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in A \right] \\
&\geq \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in A\right] - g^*(\mathbf{X})\right)^2 \mid \mathbf{X}\in A \right]
\end{aligned}$$

Let us define

$$\mathbb{W}_n^s := \sum_{A\in\mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^s)} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in A\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X}\in A \right] \mathbb{P}\left(\mathbf{X}\in A\right)$$

$$\text{and } \mathbb{W}_n^{i\setminus s} := \sum_{A\in\mathcal{P}(\mathcal{C}_n)\setminus\mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^s)} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in A\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X}\in A \right] \mathbb{P}\left(\mathbf{X}\in A\right).$$

From (22), we infer that

$$\mathbb{W}_n = \mathbb{W}_n^s + \mathbb{W}_n^{i\setminus s}.$$

Continuing (29), we get

$$\mathbb{B}_n \geq \beta_n^{-2} \frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)} \mathbb{W}_n^s - (\beta_n^{-2}\Delta_n' + \Delta_n) \qquad (30)$$

From (23) and (30), we conclude that

$$\begin{aligned}\mathbb{W}_n^s &\leq \frac{\mathbb{V}(g^*(\mathbf{X})) + \beta_n^{-2}\Delta_n' + \Delta_n}{1 + \beta_n^{-2}\frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)}} \\ &\xrightarrow[n\to\infty]{\mathbb{P}} 0\end{aligned}$$

using Remark 7, Equations (26) and (27) and (H4).

Since $\mathcal{C}_n = \mathcal{C}_n^s \cup \mathcal{C}_n^i$, we have : $\left(\mathcal{P}(\mathcal{C}_n) \setminus \mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^s)\right) \subseteq \mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^i)$. Thus, we obtain

$$\begin{aligned}\mathbb{W}_n^{i\setminus s} &\leq \sum_{A\in\mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^i)} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in A\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X}\in A\right]\mathbb{P}\left(\mathbf{X}\in A\right) \\ &\leq \frac{1}{m(\mathcal{C}_n^i)}\sum_{\mathbf{r}\in\mathcal{C}_n^i}\sum_{A\in\mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in A\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X}\in A\right]\mathbb{P}\left(\mathbf{X}\in A\right) \\ &\leq \frac{1}{m(\mathcal{C}_n^i)}\sum_{\mathbf{r}\in\mathcal{C}_n^i}\sum_{A\in\mathcal{P}_{\mathcal{C}_n}(\mathbf{r})} \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in \mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X}\in A\right]\mathbb{P}\left(\mathbf{X}\in A\right) \\ &\leq \frac{1}{m(\mathcal{C}_n^i)}\sum_{\mathbf{r}\in\mathcal{C}_n^i}\left(\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \mathbf{X}\in \mathbf{r}\right] - g^*(\mathbf{X})\right)^2 \middle| \mathbf{X}\in \mathbf{r}\right]\right)\mathbb{P}\left(\mathbf{X}\in \mathbf{r}\right).\end{aligned}$$

Using (28) we have

$$\begin{aligned}\mathbb{W}_n^{i\setminus s} &\leq \frac{1}{m(\mathcal{C}_n^i)}\sum_{\mathbf{r}\in\mathcal{C}_n^i}\left(\mathbb{V}(Y|\mathbf{X}\in\mathbf{r}) - \sigma^2\right)\mathbb{P}\left(\mathbf{X}\in\mathbf{r}\right) \\ &\leq \frac{1}{m(\mathcal{C}_n^i)}\sum_{\mathbf{r}\in\mathcal{C}_n^i}\left(\mathbb{V}_n(Y|\mathbf{X}\in\mathbf{r}) - \sigma^2 + \Delta_n'\right)\mathbb{P}\left(\mathbf{X}\in\mathbf{r}\right).\end{aligned}$$

Then, under (H4) we obtain

$$\begin{aligned}\mathbb{W}_n^{i\setminus s} &\leq \frac{1}{m(\mathcal{C}_n^i)}\sum_{\mathbf{r}\in\mathcal{C}_n^i}\left(\varepsilon_n^2 + \Delta_n'\right)\mathbb{P}\left(\mathbf{X}\in\mathbf{r}\right) \\ &\leq \frac{\sum_{\mathbf{r}\in\mathcal{C}_n^i}\mathbb{P}\left(\mathbf{X}\in\mathbf{r}\right)}{m(\mathcal{C}_n^i)}\left(\varepsilon_n^2 + \Delta_n'\right) \\ &\leq \frac{M(\mathcal{C}_n^i)}{m(\mathcal{C}_n^i)}\left(\varepsilon_n^2 + \Delta_n'\right) \\ &\xrightarrow[n\to\infty]{\mathbb{P}} 0.\end{aligned}$$

Hence, (21) is proved and the proof is complete.

3.4 Estimation Error

In this subsection we prove (20) using hypotheses (H1), (H2) and (H5).

Recall from (5) that $\mathcal{G}_n$ is the set of piecewise constant functions with values in $[-L, L]$ on the elements of the partition $\mathcal{P}(\mathcal{C}_n(D_n))$. Then, with the definition of Π_n in (H5) in mind,

$$\sup_{g \in \mathcal{G}_n} |\mathcal{L}_n(g) - \mathcal{L}(g)| \leq \sup_{g \in \mathcal{G}_c \circ \Pi_n} |\mathcal{L}_n(g) - \mathcal{L}(g)|,$$

where $\mathcal{G}_c$ is the set of constant functions $\mathbb{R}^d \to [-L, L]$ and

$$\mathcal{G}_c \circ \Pi_n := \left\{ g : \mathbb{R}^d \to \mathbb{R} : g = \sum_{A \in \mathcal{P}} f_A \mathbf{1}_A, \mathcal{P} \in \Pi_n, f_A \in \mathcal{G}_c \right\}.$$

The following is based on the same idea as [Györfi et al., 2006, Theorem 13.1].

According to [Györfi et al., 2006, Theorem 9.1 and Problem 10.4] we have,

$$\mathbb{P}\left\{ \sup_{g \in \mathcal{G}_c \circ \Pi_n} |\mathcal{L}_n(g) - \mathcal{L}(g)| > \varepsilon \right\} \leq 8\mathbb{E}\left[\mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, \mathbf{X}_1^n \right) \right] \exp\left\{ \frac{-n\varepsilon^2}{128.(4L^2)^2} \right\}, \quad (31)$$

where $\mathbf{X}_1^n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$.

Here $\mathcal{N}_1(\varepsilon, \mathcal{G}_c \circ \Pi_n, \mathbf{X}_1^n)$ is the random variable corresponding to the minimal number $N \in \mathbb{N}$ such that there exist functions $g_1, \ldots, g_N : \mathbb{R}^d \to [-L, L]$ with the property that for every $g \in \mathcal{G}_c \circ \Pi_n$ there is a $j \in \{1, ..., N\}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} |g(\mathbf{X}_i) - g_j(\mathbf{X}_i)| \leq \varepsilon.$$

This number is called the ε-*covering number* of $\mathcal{G}_c \circ \Pi_n$. It can be interpreted as the complexity of the class. Then using [Györfi et al., 2006, Lemma 13.1] we have

$$\mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, \mathbf{X}_1^n \right) \leq \Delta(\Pi_n) \left\{ \sup_{z_1, \ldots, z_m \in \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}, m \leq n} \mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}_c, z_1^m \right) \right\}^{\mathcal{M}(\Pi_n)},$$

According to [Györfi et al., 2006, Lemma 9.2] for any set of function $\mathcal{G}$ and any sample z_1^m we have

$$\mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}, z_1^m \right) \leq \mathcal{M}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}, z_1^m \right),$$

where $\mathcal{M}_1(\varepsilon, \mathcal{G}, z_1^m)$ is the maximal $N \in \mathbb{N}$ such that there exist functions $g_1, \ldots, g_N \in \mathcal{G}$ with

$$\frac{1}{n}\sum_{i=1}^{m} |g_j(z_i) - g_k(z_i)| \geq \varepsilon,$$

for all $1 \leq j < k \leq N$. It is called L_1 *ε-packing* of $\mathcal{G}$ on z_1^m. See [Györfi et al., 2006, Definition 9.4 (c)].

Now, from the definition of $\mathcal{G}_c$,

$$\sup_{z_1,\ldots,z_m \in \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}, m \leq n} \mathcal{M}_1(\varepsilon, \mathcal{G}_c, z_1^m) = \left\lceil \frac{2L}{\varepsilon} \right\rceil.$$

Finally,

$$\sup_{z_1,\ldots,z_m \in \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}, m \leq n} \mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, z_1^m\right) \leq \Delta(\Pi_n) \left\lceil \frac{64L^2}{\varepsilon} \right\rceil^{\mathcal{M}(\Pi_n)}. \quad (32)$$

According to (31) and (32) we have:

$$\mathbb{P}\left\{ \sup_{g \in \mathcal{G}_n} \left| \frac{1}{n}\sum_{i=1}^{n} |g(\mathbf{X}_i) - Y_i|^2 - \mathbb{E}\left[|g(\mathbf{X}) - Y|^2\right] \right| > \varepsilon \right\}$$
$$\leq 8\Delta_n(\Pi_n) \left\lceil \frac{64L^2}{\varepsilon} \right\rceil^{\mathcal{M}(\Pi_n)} \exp\left(-\frac{n\varepsilon^2}{128.(4L^2)^2}\right)$$

and since

$$\begin{aligned}
&8\Delta_n(\Pi_n) \left\lceil \frac{64L^2}{\varepsilon} \right\rceil^{\mathcal{M}(\Pi_n)} \exp\left(-\frac{n\varepsilon^2}{2048L^4}\right) \\
&\leq 8\exp\left(\log \Delta_n(\Pi_n) + \mathcal{M}(\Pi_n) \log\left(\left\lceil \frac{64L^2}{\varepsilon} \right\rceil\right) - \frac{n\varepsilon^2}{2048.L^4}\right) \\
&\leq 8\exp\left(-\frac{n}{L^4}\left(\frac{\varepsilon^2}{2048} - \frac{\log \Delta_n(\Pi_n) L^4}{n} - \frac{\mathcal{M}(\Pi_n) L^4 \log\left(\left\lceil \frac{64L^2}{\varepsilon} \right\rceil\right)}{n}\right)\right),
\end{aligned}$$

this concludes the proof of (20) and of Theorem 2.1.

4 Illustrations

In this section we propose a simple algorithm to generate a suitable sequence of data-dependent coverings using the Random Forests algorithm (RF) of Breiman [2001] as rule generator. The interest is double; first, it shows that there exists a sequence of suitable data-dependent coverings in practice as in Definition 2.1. Second, it proves the consistency of a rule-based estimator generated from RF as soon as the condition (H5) on the complexity of RF is satisfied. Until now

there are few results about the consistency of an estimator generated by RF, we may cite Denil et al. [2013], Scornet et al. [2015].

Let $\mathcal{C}$ be the set of all hyperrectangles of $\mathbb{R}^d$:

$$\mathcal{C} := \left\{ [\mathbf{a}, \mathbf{a}+\mathbf{h}] \subseteq \mathbb{R}^d : \mathbf{a} \in \mathbb{R}^d,\, \mathbf{h} \in \mathbb{R}^d_+ \right\} \cup \{\emptyset\}.$$

The following result ensures that any set of rules $\mathcal{C}_n$ such that $\mathcal{C}_n \subseteq \mathcal{C}$, which is the case of sets of rules generated by RF, fulfills (H6).

Lemma 4.1. *$\mathcal{B} := \{\mathbf{c} \times \mathbb{R} : \mathbf{c} \in \mathcal{C}\}$ is a $\mathbb{Q}$-Donsker class.*

The proof is given for completeness in Appendix.

4.1 Covering Algorithm

The proposed algorithm generates an estimator using data-dependent coverings. It is decomposed into four steps.

1. Generation of RF with m_{tree} trees with a maximal depth controlled by a fixed maximal number of rules generated by RF[1].

2. Extraction of all significant and insignificant rules according to (H3) and (H4) from all nodes and leaves of all trees generated by RF that have a length lower than or equal to $k_{max} \in \{1, \ldots, d\}$ for a chosen $\alpha \in (0, 1/2)$, $\beta_n = n^{\alpha/2-1/4}$, $\varepsilon_n = \beta_n \sigma_n(Y)$, where $\sigma_n(Y)$ is the empirical standard deviation of Y.

3. Selection of a minimal set of rules using Algorithm 1 in Appendix. A rule is added to the current set of rules if and only if it has at least a rate $\gamma \in (0, 1)$ of points not covered by the current set of rules. The smaller γ the smaller the number of rules $\#\mathcal{C}_n$[2].

4. If the selected set of rules does not form a covering, generation of a unique *no-rule* that is one of the smallest hyperrectangles satisfying (H3) containing the remaining points.

Remark 10. We do not use step 4 in any of the examples treated below as we already obtain a covering at the third step. If the no-rule had to be added to ensure a covering, one should check that the no-rule is either a significant or an insignificant rule in order to obtain a consistent Covering Algorithm.

[1] The redundancies conditions (13) and (14) are automatically controlled by the fixed maximal number of rules generated by RF.

[2] This step was already described in Remark 4 in order to control the redundancy. Here it is only used for controlling the number of rules as the redundancy is already controlled by the maximal number of rules generated by RF.

4.2 Artificial data

Here we consider the same data set as in Friedman and Popescu [2008]. We generate $n = 5000$ data following the regression setting

$$Y = g^*(\mathbf{X}) + Z,$$

where $d = 100$ (the dimension of $\mathbf{X}$) and

$$g^*(\mathbf{X}) = 9 \prod_{j=1}^{3} \exp\left(-3\left(1 - X_j\right)^2\right) - 0.8 \exp\left(-2\left(X_4 - X_5\right)\right) + 2\sin^2(\pi \cdot X_6) - 2.5\left(X_7 - X_8\right), \quad (33)$$

and $Z \sim \mathcal{N}(0, \sigma^2)$. The value of $\sigma > 0$ was chosen to produce a two-to-one signal-to-noise ratio. The variables were generated from a uniform distribution on $\{0/10, \ldots, 9/10\}$. It is important to notice that only the eight first variables are informative; the 92 others are just noise. Coefficients multiplying each of the terms in g^* were chosen to ensure that variables have approximately equal influence.

As in Friedman and Popescu [2008], we evaluate the accuracy of estimators generated by RF, Covering Algorithm and RuleFit using the average absolute error defined by

$$aae = \frac{\mathbb{E}_{\mathbb{Q}}\left[|Y - g_n(\mathbf{X})|\right]}{\mathbb{E}_{\mathbb{Q}}\left[|Y - \text{median}(Y)|\right]}.$$

In order to evaluate the error without including the error of the noise, we also consider the following criterium

$$aae^* = \frac{\mathbb{E}_{\mathbb{Q}}\left[|g^*(\mathbf{X}) - g_n(\mathbf{X})|\right]}{\mathbb{E}_{\mathbb{Q}}\left[|g^*(\mathbf{X}) - \text{median}(g^*(\mathbf{X}))|\right]}.$$

We approximate the criteria aae and aae^* using 50000 test observations sampled independently from $\mathbb{Q}$.

4.2.1 Execution

We run $M = 100$ simulations. For each simulation we set the maximal number of rules generated by the Random Forest (RF) at 20000 and the number of trees at $m_{tree} = 100$. The maximal length of a rule is fixed at $k_{max} = 3$. And we set $\alpha = 1/2 - 1/100$ and $\gamma = 0.95$.

4.2.2 Results

We calculate the average absolute errors of each algorithm and the interpretability score (8) for each experience and summarize it in Figure 5. All results are resumed in Table 1.

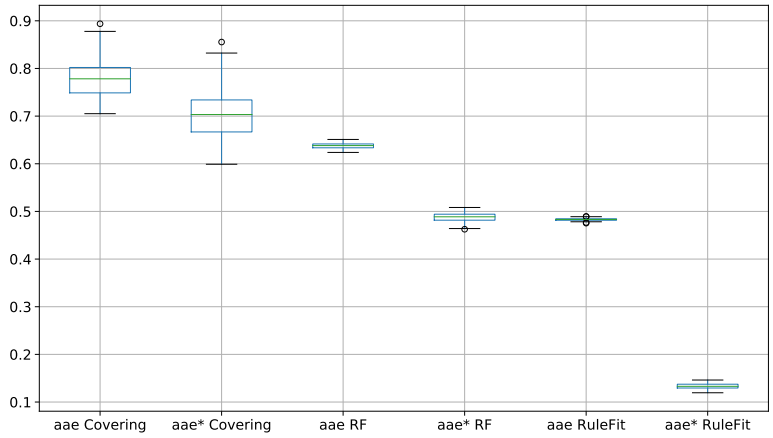

Figure 5: Box-plot of the average absolute error of each algorithm.

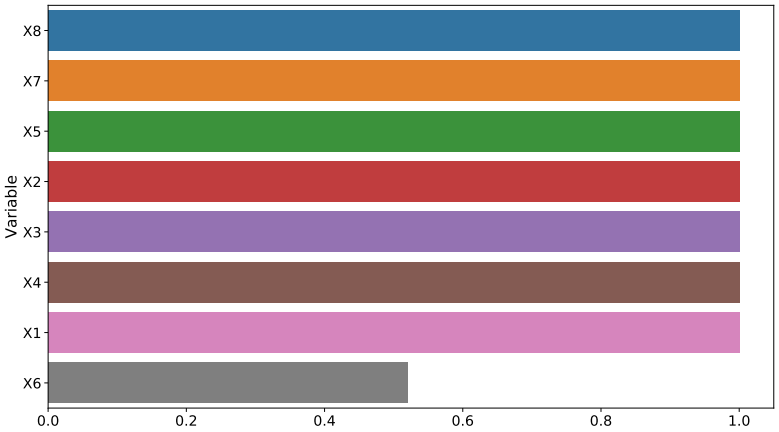

Figure 6: Empirical probability of occurrence in at least one rule of the selected set of rules generated by the Covering Algorithm.

4.2.3 Comments

Figures 6 and 7 display the frequency of occurrence of a variable in at least one rule of the selected set of rules and the average occurrence of a variable in the selected set of rules respectively. We emphasize that only the informative variables are involved in the selected rules. It means that the support of g^* is well identified. Moreover, in more than half of the experiences, all informative

Random Forest	Nb rules	Interpretability	aae	aae*
mean	12596.24	60859.48	0.64	0.48
std	3.80	363.87	0.006	0.01
min	12582	59546	0.62	0.45
25%	12596	60642	0.63	0.48
50%	12598	60965	0.64	0.49
75%	12600	61101	0.64	0.49
max	12600	61478	0.65	0.50

Covering Algorithm	Nb rules	Interpretability	aae	aae*
mean	33.59	100.27	0.77	0.69
std	3.59	10.80	0.04	0.06
min	27	81	0.68	0.56
25%	31	93	0.74	0.64
50%	33	99	0.77	0.69
75%	36	107.25	0.79	0.73
max	43	129	0.89	0.85

RuleFit	Nb Rules	Nb linear	Interpretability	aae	aae*
mean	183.11	25.02	448.82	0.48	0.13
std	18.22	6.77	49.57	0.002	0.006
min	143	12	341.0	0.47	0.11
25%	170.75	20	413.25	0.48	0.13
50%	181	25	447.50	0.48	0.13
75%	193.25	30	472.5	0.48	0.14
max	233	42	608	0.49	0.15

Table 1: Summary of the M experiences for each algorithm.

variables are identified. In the other cases, the variable $X6$ is the only rule that is not involved. One reason might be that RF is a random generator of rules that might not capture the importance of $X6$ at each run. The issue might be solved by considering a deterministic rule generator algorithm.

Nevertheless, the linear term in (33) cannot be well fitted by a finite set of rules. This setting is highly favorable to RuleFit that is the only tested algorithm including linear components. However, RuleFit involves at least 11 and up to 39 linear components (see Table 1). Its accuracy is high but many noise variables are included in its generated model. Hence the a posteriori analysis of the importances of the variables and rules of the generated model is crucial, see Section 9.1.2 of Friedman and Popescu [2008].

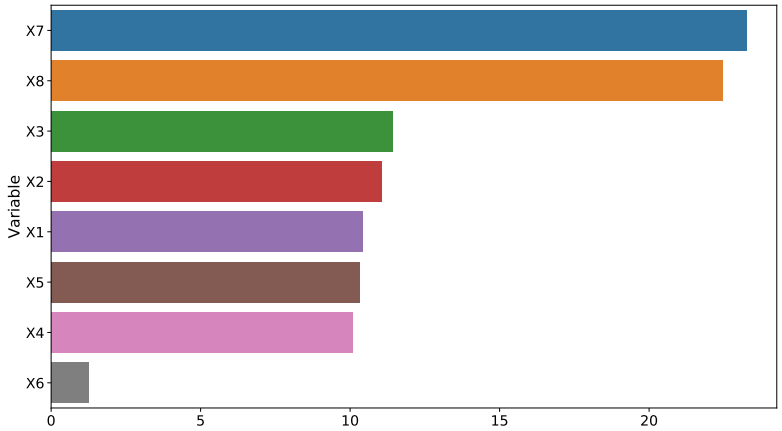

Figure 7: Average occurrence in the selected set of rules generated by the Covering Algorithm.

4.3 Real data

In this example we use the well-known *Boston housing* dataset. It consists in $n = 506$ neighborhoods in the Boston area and 14 statistics for each neighborhood describes in Table 2 (see Table IV p96-97 Harrison Jr and Rubinfeld [1978] for more details about the variables). The regression model explains the median house price Y of each neighborhood by the $d = 13$ others statistics of this neighborhood. The data are randomly split into a training set and a test set, with a ratio of 70% / 30%, respectively.

4.3.1 Execution

We set the maximal number of rules generated by the Random Forest (RF) at 20000 and the number of trees at $m_{tree} = 100$. It generates 8490 rules and the RF estimator has an interpretability score of $Int(g_n) = 29524$ according to (8). Among these rules 4066 have a length lower than or equal to $k_{max} = 3$. In this application the real value of σ^2 is unknown. We estimate it by σ_n^2 the minimal variance of the rules fulfilling the covering condition (9). We have $\sigma_n^2 = 0.815$. Considering $\sigma^2 = \sigma_n^2$ and $\alpha = 1/2 - 1/100$, the Covering Algorithm described in Section 4.1 extracts 1306 rules that are significant according to (11) and 2083 that are insignificant according to (12). Then, the selection process (Algorithm 1), with $\gamma = 0.95$, extracts a set of 6 significant rules which cover 99% of the data and adds 2 insignificant rules $R4$ and $R7$ to cover all data. The interpretability score obtained by the covering estimator is $Int(g_n) = 20$.

	Y	CRIM	ZN	INDUS	CHAS	NOX	RM
mean	22.53	3.61	11.36	11.14	0.069	0.55	6.28
std	9.19	8.60	23.32	6.86	0.25	0.11	0.70
min	5.00	0.006	0.0	0.46	0.0	0.38	3.56
25%	17.02	0.08	0.0	5.19	0.0	0.44	5.88
50%	21.20	0.25	0.0	9.69	0.0	0.53	6.20
75%	25.00	3.67	12.50	18.10	0.0	0.62	6.62
max	50	88.97	100.0	27.74	1.0	0.87	8.78

	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
mean	68.57	3.79	9.54	408.23	18.45	356.67	12.65
std	28.14	2.10	8.70	168.53	2.16	91.29	7.14
min	2.90	1.12	1.0	187.0	12.6	0.32	1.73
25%	45.02	2.10	4.0	279.0	17.40	375.37	6.95
50%	77.50	3.20	5.0	330.0	19.0	391.44	11.36
75%	94.07	5.18	24.0	666.0	20.20	396.22	16.95
max	100.0	12.12	24.00	711.0	22.0	396.90	37.97

Table 2: Description of variables.

Rule	Conditions	Coverage	Prediction	Var
$R1$	$LSTAT \in [9.62, 37.97]$	0.58	17.74	26.03
$R2$	$LSTAT \in [1.73, 7.87]$	0.31	32.31	78.57
$R3$	$RM \in [3.86, 6.10]$ $LSTAT \in [7.76, 37.97]$ $DIS \in [1.20, 12.13]$	0.37	17.62	22.52
$R4$	$RM \in [6.10, 6.54]$ $LSTAT \in [1.73, 9.98]$ $DIS \in [1.42, 12.13]$	0.12	23.39	4.62
$R5$	$RM \in [6.42, 8.78]$ $LSTAT \in [1.73, 14.40]$ $DIS \in [1.20, 12.13]$	0.33	31.88	72.88
$R6$	$RM \in [6.09, 8.78]$ $LSTAT \in [1.73, 9.98]$ $CRIM \in [0.05, 88.98]$	0.27	32.12	76.50
$R7$	$RM \in [3.86, 6.12]$ $LSTAT \in [8.13, 14.40]$ $TAX \in [187.0, 298.0]$	0.08	21.17	3.69
$R8$	$LSTAT \in [5.06, 37.97]$ $PTRATIO \in [13.90, 22.0]$ $INDUS \in [15.01, 27.74]$	0.36	16.34	35.70

Table 3: Summary of the selected rules generated by Covering Algorithm.

4.3.2 Results

Regarding the accuracy on the test set, RF has a *aae* of 0.37 and Covering Algorithm has a *aae* of 0.65. This loss of accuracy is the price of turning a black box model into an interpretable one. According to (8) we have $Int(g_n) = 20$ for the Covering Algorithm and $Int(g_n) = 29524$ for the RF, which is a huge improvement of the interpretability score.

Algorithm	Nb rules	Interpretability	*aae*
Random Forest	8490	29524	0.37
Covering Algorithm	8	20	0.63
RuleFit	602 and 1 linear component	602	0.34

Table 4: Comparison between Random Forest, Covering Algorithm and RuleFit on real data.

4.3.3 Comments

This application on real data emphasizes that data-dependent coverings are very efficient to generate an interpretable rule-based model. With 0.09% of the rules from RF and a reduction of 99.9% of the interpretability score, Covering Algorithm constructs an estimator with accuracy 76% larger than RF. The interpretability score of Covering Algorithm is also much lower than the one of RuleFit.

Using Table 2 we are able to translate in natural language the significant rules in Table 3, accordingly with the definition of interpretability of Biran and Cotton [2017].

- Rules $R1$ and $R2$ indicate that the $LSTAT$ variable, which represents the percent of lower status population, is inversely related to the median house price. This relation is easily identified by the linear component of RuleFit which corresponds to the unique variable $LSTAT$ (See Table 4).
- Rules $R3$, $R5$ and $R6$ show that for $LSTAT$ close to the median, the number of rooms RM is determining the median house price.
- Rule $R8$ can be interpreted by the fact that a very high proportion of non-retail business acres per town $INDUS$ has a negative influence on the median house price Y that cannot be offset by a percentage of the lower status $LSTAT$ and a pupil-teacher ratio $PTRATIO$ lower than the first quartile.

The variables DIS and $CRIM$ seem superfluous as the associated rules $R3$, $R5$ and $R6$ cover most of their range.

5 Conclusion and perspectives

In this paper we provide a general setting for studying the consistency of interpretable rule-based estimators. The novelty is to introduce the notion of covering composed by two kinds of sets, the significant and the insignificant ones. The significant sets are thought as interpretable sets by construction. The insignificant ones are thought as small sets which variances tend to zero. We provide Covering Algorithm that extracts from any rule generator a suitable data-dependent covering. We apply it to RF and we compare its result with the ones of RF and RuleFit (Friedman and Popescu [2008]). The loss of accuracy in the prediction is the cost to pay to have an interpretable model according to our definition of interpretability. Monte Carlo experiments in section 4.2 show that Covering Algorithm, seeking interpretability, is identifying the support of the regression function as a byproduct. The loss of accuracy of Covering Algorithm may be due to the conflict between model identification and regression estimation identified by Yang [2005]. It is worthwile to mention that despite its loss of accuracy, Covering Algorithm is still weakly consistent.

Our methodology based on coverings is very effective for generating interpretable models. The use of RF as the rule-generator of Covering Algorithm is questionable; In Section 4.2, we pointed out the possible negative effect of the randomization procedure in RF for identifying informative variables. In Section 4.3, we observed some rules generated by RF with length potentially artificially too high. We look for an algorithm that generates significant, insignificant rules and a suitable sequence of data-dependent coverings satisfying (H5) on its own. The difficulty we encounter is to ensure that the generated rules cover all the data. It is a subject of research for future works. The theoretical setting could also be refined; unbounded Y may be considered by introducing a truncation operator as in Györfi et al. [2006]; strong consistency and rates of convergence of the data-dependent covering estimators may be established under slightly stronger assumptions; the consistency of the algorithm may be extended to cases of unknown variance estimated as in Section 4.3. Finally, the scope could be broaden from the regression setting to the classification one by adapting the significant condition accordingly.

Appendix

In this Appendix, we gather some proofs and we provide the pseudo-code of Covering Algorithm selection process.

Proof of Proposition 3.2

Proof. Let $\varepsilon > 0$. First, for any $f \in \mathcal{F}$ and $A \in \mathcal{B}_n$, since $\mathbb{Q}_n(A) > 0$ and then $\mathbb{Q}(A) > 0$,

$$\begin{aligned}
&|\mathbb{E}_n[f \mid A] - \mathbb{E}[f \mid A]| \\
&= \left| \frac{\int_A f d\mathbb{Q}_n}{\mathbb{Q}_n(A)} - \frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)} \right| \\
&= \left| \frac{\mathbb{Q}(A)\left(\int_A f d\mathbb{Q}_n - \int_A f d\mathbb{Q}\right) + (\mathbb{Q}(A) - \mathbb{Q}_n(A)) \int_A f d\mathbb{Q}}{\mathbb{Q}(A)\mathbb{Q}_n(A)} \right| \\
&\leq \left| \frac{\int_A f d\mathbb{Q}_n - \int_A f d\mathbb{Q}}{\mathbb{Q}_n(A)} \right| + \left| (\mathbb{Q}(A) - \mathbb{Q}_n(A)) \frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)\mathbb{Q}_n(A)} \right|. \qquad (34)
\end{aligned}$$

Now, according to Proposition 3.1,

$$\sup_{\tilde{f} \in \mathcal{F}, \tilde{A} \in B} \left| \int_{\tilde{A}} \tilde{f} d\mathbb{Q}_n - \int_{\tilde{A}} \tilde{f} d\mathbb{Q} \right| = O_{\mathbb{P}^*}(n^{-1/2})$$

and

$$\sup_{\tilde{A} \in \mathcal{B}} \left| \mathbb{Q}_n(\tilde{A}) - \mathbb{Q}(\tilde{A}) \right| = O_{\mathbb{P}^*}(n^{-1/2}).$$

Thus, According to Remark 8, there exists $M > 0$ such that for any n large enough,

$$\mathbb{P}^* \left\{ \sup_{\tilde{f} \in \mathcal{F}, \tilde{A} \in B} \left| \int_{\tilde{A}} \tilde{f} d\mathbb{Q}_n - \int_{\tilde{A}} \tilde{f} d\mathbb{Q} \right| > M n^{-1/2} \right\} < \frac{\varepsilon}{2}$$

and

$$\mathbb{P}^* \left\{ \sup_{\tilde{A} \in \mathcal{B}} \left| \mathbb{Q}_n(\tilde{A}) - \mathbb{Q}(\tilde{A}) \right| > M n^{-1/2} \right\} < \frac{\varepsilon}{2}$$

so that $\mathbb{P}^*(\Omega_n) \geq 1 - \varepsilon$ with

$$\Omega_n := \left\{ \sup_{\tilde{f} \in \mathcal{F}, \tilde{A} \in B} \left| \int_{\tilde{A}} \tilde{f} d\mathbb{Q}_n - \int_{\tilde{A}} \tilde{f} d\mathbb{Q} \right| \leq M n^{-1/2} \right\} \bigcap \left\{ \sup_{\tilde{A} \in B} \left| \mathbb{Q}_n(\tilde{A}) - \mathbb{Q}(\tilde{A}) \right| \leq M n^{-1/2} \right\}.$$

Then (34) yields, with $c := \sup_{f \in \mathcal{F}, x \in \mathcal{S}} |f(x)| < \infty$ and since $\mathbb{Q}_n(A) \geq n^{-\alpha}$, for n large enough, in the event Ω_n,

$$\sup_{f \in \mathcal{F}, A \in \mathcal{B}_n} \left| \mathbb{E}_n[f \mid A] - \mathbb{E}[f \mid A] \right| \leq M n^{\alpha - 1/2}(1 + c),$$

since $\frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)} \leq c$.

Finally, it has been proved that $\forall \varepsilon > 0, \exists M > 0, \exists N \in \mathbb{N}^* / \forall n \geq N$,

$$\mathbb{P}^* \left\{ \sup_{f \in \mathcal{F}, A \in \mathcal{B}_n} |\mathbb{E}_n [f \mid A] - \mathbb{E} [f \mid A]| > M n^{\alpha - 1/2} \right\} < \varepsilon$$

and then $\forall \varepsilon > 0, \exists M > 0$ such that

$$\limsup_{n \to \infty} \mathbb{P}^* \left\{ \sup_{f \in \mathcal{F}, A \in \mathcal{B}_n} |\mathbb{E}_n [f \mid A] - \mathbb{E} [f \mid A]| > M n^{\alpha - 1/2} \right\} \leq \varepsilon$$

which, together with Remark 8 again, proves the proposition. □

Proof of Corollary 3.1

Proof of (16). Let $L = \operatorname{ess\,sup} Y$, $i \in \mathbb{N}$, and $f_i \in \mathcal{L}^1(\mathbb{Q})$ be defined by

$$\begin{aligned} f_i : \mathbb{R}^d \times [-L, L] &\to [-L^i, L^i] \\ (\mathbf{x}, y) &\mapsto y^i. \end{aligned}$$

f_i is bounded and $\{f_i\}$ is finite thus Donsker. The result is then a straightforward application of Proposition 3.2. □

Proof of (17). This part follows from (17) since Y is bounded and

$$\mathbb{V}_n [Y \mid (\mathbf{X}, Y) \in A] := \mathbb{E}_n \left[Y^2 \mid (\mathbf{X}, Y) \in A \right] - \mathbb{E}_n [Y \mid (\mathbf{X}, Y) \in A]^2 .$$

□

Proof of Lemma 4.1

We remind that, given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket in $L_2(\mathbb{Q})$ is a bracket $[l, u]$ with $\|u - l\|_{L_2(\mathbb{Q})} < \epsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \mathcal{L}_2(\mathbb{Q}))$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. (The bracketing functions l and u must have finite $L_2(\mathbb{Q})$-norms but need not belong to $\mathcal{F}$.) A simple condition for a class to be $\mathbb{Q}$-Donsker is that $N_{[]}(\varepsilon, \mathcal{F}, \mathcal{L}_2(\mathbb{Q}))$ don't grow too fast to infinity as ε tends to 0. The speed is measured in terms of the *bracketing integral* defined by

$$J_{[]}(\delta, \mathcal{F}, \mathcal{L}_2(\mathbb{Q})) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, \mathcal{L}_2(\mathbb{Q}))} d\varepsilon.$$

Theorem 5.1. *[Van der Vaart, 2000, Theorem 19.4] Every class $\mathcal{F}$ of measurable functions with $J_{[]}(1, \mathcal{F}, \mathcal{L}_2(\mathbb{Q})) < \infty$ is $\mathbb{Q}$-Donsker.*

Proof of Lemma 4.1. Let $\varepsilon > 0$ and consider the sequence $\{\mathbf{t}_i\}_{i \in \{0, \ldots, \lceil 2d/\varepsilon \rceil\}}$ defined as follows: for any $k \in \{1, \ldots, d\}$,

$$\begin{cases} t_{0,k} = -\infty \\ \forall i \in \{1, \ldots, \lfloor 2d/\varepsilon \rfloor\}, t_{i,k} = \sup\{t \in \mathbb{R} : \mathbb{Q}(\mathbb{R}^{k-1} \times]t_{i-1,k}, t[\times \mathbb{R}^{d-k} \times [-L, L]) < \varepsilon/2d\} \\ t_{\lceil 2d/\varepsilon \rceil, k} = +\infty \end{cases}$$

From the definition of $\mathbf{t}_i$'s we have $\forall k \in \{1,\dots,d\}$, $\forall i \in \{1,\dots,\lceil 2d/\varepsilon\rceil\}$,

$$\mathbb{Q}\left(\mathbb{R}^{k-1}\times]t_{i-1,k},t_{i,k}[\times\mathbb{R}^{d-k}\times[-L,L]\right)\le\varepsilon/2d$$
$$\mathbb{Q}\left(\mathbb{R}^{k-1}\times]t_{i-1,k},t_{i,k}]\times\mathbb{R}^{d-k}\times[-L,L]\right)\ge\varepsilon/2d$$
$$\mathbb{Q}\left(\mathbb{R}^{k-1}\times]t_{0,k},t_{i,k}]\times\mathbb{R}^{d-k}\times[-L,L]\right)\ge i\varepsilon/2d$$
$$\mathbb{Q}\left(\mathbb{R}^{k-1}\times]t_{0,k},t_{\lfloor 2d/\varepsilon\rfloor,k}]\times\mathbb{R}^{d-k}\times[-L,L]\right)\ge 1-\varepsilon/2d \text{ (since } \lfloor 2d/\varepsilon\rfloor\ge 2d/\varepsilon-1)$$
$$\mathbb{Q}\left(\mathbb{R}^{k-1}\times]t_{\lfloor 2d/\varepsilon\rfloor,k},+\infty[\times\mathbb{R}^{d-k}\times[-L,L]\right)\le\varepsilon/2d$$

Hence, $\forall i \in \{1,\dots,\lceil 2d/\varepsilon\rceil\}$ and $\forall k \in \{1,\dots,d\}$ we have that

$$\mathbb{Q}\left(\mathbb{R}^{k-1}\times]t_{i-1,k},t_{i,k}[\times\mathbb{R}^{d-k}\times[-L,L]\right)\le\varepsilon/2d.$$

Consider now the set of brackets defined by (see Fig 8)

$$A:=\left\{[\mathbf{1}_{\prod_{k=1}^d[t_{i_k+1,k},t_{j_k-1,k}]},\mathbf{1}_{\prod_{k=1}^d]t_{i_k,k},t_{j_k,k}[}],\forall k\in\{1,\dots,d\},0\le i_k<j_k\le\lceil 2d/\varepsilon\rceil\right\}.$$

We have $\#A=(\lceil 2d/\varepsilon\rceil(\lceil 2d/\varepsilon\rceil+1)/2)^d$ and

$$\begin{aligned}
&d(\mathbf{1}_{\prod_{k=1}^d[t_{i_k+1,k},t_{j_k-1,k}]\times[-L,L]},\mathbf{1}_{\prod_{k=1}^d]t_{i_k,k},t_{j_k,k}[\times[-L,L]})\\
&=\int_{\mathbb{R}^d}\left|\mathbf{1}_{\prod_{k=1}^d]t_{i_k,k},t_{j_k,k}[\times[-L,L]}-\mathbf{1}_{\prod_{k=1}^d[t_{i_k+1,k},t_{j_k-1,k}]\times[-L,L]}\right|d\mathbb{Q}\\
&\le\int_{\mathbb{R}^d}\sum_{k=1}^d\mathbf{1}_{\mathbb{R}^{k-1}\times\left(\right]t_{i_k,k},t_{i_k+1,k}\left[\cup\right]t_{j_k-1,k},t_{j_k,k}\left[\right)\times\mathbb{R}^{d-k}\times[-L,L]}d\mathbb{Q}\\
&\le\sum_{k=1}^d\mathbb{Q}\left(\mathbb{R}^{k-1}\times(]t_{i_k,k},t_{i_k+1,k}[\cup]t_{j_k-1,k},t_{j_k,k}[)\times\mathbb{R}^{d-k}\times\mathbb{R}\right)\\
&\le d\left(\varepsilon/2d+\varepsilon/2d\right)\\
&\le\varepsilon.
\end{aligned}$$

The term after the first equality corresponds to the integration of the hatched area in Figure 8 and the term after the next inequality corresponds to the integration of the area delimited by the dotted lines.

Thus, the $\mathcal{L}_1(\mathbb{Q})$-size of the brackets is not larger than ε. Since $\mathbb{Q}f^2=\mathbb{Q}f$ for every $f\in\mathcal{I}_{\mathcal{C}}=\{\mathbf{1}_A : A\in\mathcal{C}\}$, the $L_2(\mathbb{Q})$-size of the brackets is not larger than $\sqrt{\varepsilon}$.

Let $f\in\mathcal{I}_{\mathcal{C}}$. Then $\exists\,\mathbf{a}\in\mathbb{R}^d,\mathbf{h}\in\mathbb{R}^d_+$ s.t. $f=\mathbf{1}_{[\mathbf{a},\mathbf{a}+\mathbf{h}]\times[-L,L]}$. We set, for any $k\in\{1,\dots,d\}$,

$$\begin{aligned}
i_k^*&:=\max\{\iota\in\{0,\dots,\lceil 2d/\varepsilon\rceil\}:t_{\iota,k}<a_k\}\\
j_k^*&:=\min\{\iota\in\{0,\dots,\lceil 2d/\varepsilon\rceil\}:t_{\iota,k}>a_k+h_k\}.
\end{aligned}$$

There always exist j_k^* and i_k^* since the sets $\{\iota\in\{0,\dots,\lceil 2d/\varepsilon\rceil\}:t_{\iota,k}<a_k\}$ and $\{\iota\in\{0,\dots,\lceil 2d/\varepsilon\rceil\}:t_{\iota,k}>a_k+h_k\}$ are not empty (they contain respectively 0 and $\lceil 2d/\varepsilon\rceil$) and, by construction, $j_k^*>i_k^*$. Moreover,

$$\mathbf{1}_{\prod_{k=1}^d[t_{i_k^*+1},t_{j_k^*-1}]\times[-L,L]}\le\mathbf{1}_{[\mathbf{a},\mathbf{a}+\mathbf{h}]\times[-L,L]}\le\mathbf{1}_{\prod_{k=1}^d]t_{i_k^*},t_{j_k^*}[\times[-L,L]}$$

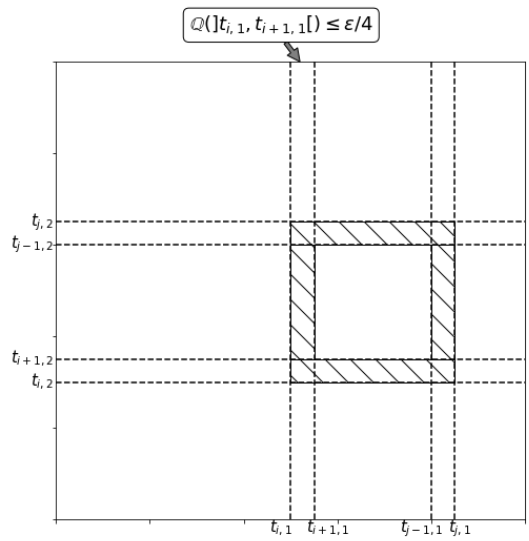

Figure 8: Example of bracket for $d = 2$. With $l = \mathbf{1}_{[t_{i+1,1};t_{j-1,1}]\times[t_{i+1,2};t_{j-1,2}]}$ and $u = \mathbf{1}_{[t_{i,1};t_{j,1}]\times[t_{i,2};t_{j,2}]}$, for any rectangle A, $\mathbf{1}_A \in [l, u]$ if and only if its boundary $\bar{A} \setminus \mathring{A}$ is included in the hatched area.

and

$$[\mathbf{1}_{\prod_{k=1}^{d}[t_{i_k^*+1},t_{j_k^*-1}]\times[-L,L]}, \mathbf{1}_{\prod_{k=1}^{d}]t_{i_k^*},t_{j_k^*}[\times[-L,L]}] \in A.$$

Thus, $\forall f \in \mathcal{I}_{\mathcal{C}}, \exists [l, u] \in A$ such that $l \leq f \leq u$.

It follows that $N_{[\,]}(\sqrt{\varepsilon}, \mathcal{I}_{\mathcal{C}}, \mathcal{L}_2(\mathbb{Q})) \leq \left(\frac{\lceil 2d/\varepsilon\rceil(\lceil 2d/\varepsilon\rceil+1)}{2}\right)^d$. Hence,

$$J_{[\,]}(1, \mathcal{I}_{\mathcal{C}}, \mathcal{L}_2(\mathbb{Q})) < \infty.$$

According to Theorem 5.1, this guarantees that $\mathcal{I}_{\mathcal{C}}$ is a $\mathbb{Q}$-Donsker class. □

Covering Algorithm selection process

References

O. Biran and C. Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, page 1, 2017.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Algorithm 1: Selection of minimal set of rules

Input:

- the rate $0 < \gamma < 1$;
- a set of significant rules S;
- a set of insignificant rules I;

Output:

- a minimal set of rules $\mathcal{C}_n$;

```
C_n ← arg max_{r∈S} P_n(r);
S ← S \ C_n;
while Σ_{r∈C_n} P_n(r) < 1 do
    r* ← arg max_{r∈S} P_n(r);
    if P_n(r* ∩ {∪_{r∈C_n} r}) ≤ γ P_n(r*) then
        C_n ← C_n ∪ r*;
    S ← S \ r*;
    if #S = 0 then
        Break ;
end
if Σ_{r∈C_n} P_n(r) < 1 then
    while Σ_{r∈C_n} P_n(r) < 1 do
        r* ← arg min_{r∈I} V_n(Y|X ∈ r);
        if P_n(r* ∩ {∪_{r∈C_n} r}) ≤ γ P_n(r*) then
            C_n ← C_n ∪ r*;
        I ← I \ r*;
        if #I = 0 then
            Break ;
    end
return C_n;
```

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. CRC press, 1984.

M. Denil, D. Matheson, and N. Freitas. Consistency of online random forests. In *International Conference on Machine Learning*, pages 1256–1264, 2013.

J. Friedman and B. Popescu. Predective learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

J. Fürnkranz and T. Kliegr. A brief overview of rule learning. In *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, pages 54–69. Springer, 2015.

S. Grunewalder. Plug-in estimators for conditional expectations and probabilities. In *International Conference on Artificial Intelligence and Statistics*, pages 1513–1521, 2018.

R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *arXiv preprint arXiv:1802.01933*, 2018.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

D. Harrison Jr and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

G. Holmes, M. Hall, and E. Prank. Generating rule sets from model trees. In *Australasian Joint Conference on Artificial Intelligence*, pages 1–12. Springer, 1999.

A. Karalič and I. Bratko. First order regression. *Machine Learning*, 26(2-3): 147–176, 1997.

Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2017.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

V. Margot, J.-P. Baudry, F. Guilloux, and O. Wintenberger. Rule induction partitioning estimator. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 288–301. Springer, 2018.

A. Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

J. R. Quinlan. C4.5: Programs for machine learning. 1993.

M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

E. Scornet, G. Biau, J.-P. Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

C. Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Y. Yang. Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

Q. Zhao and S. S. Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 2003.