



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-02170687v1

HAL Id: hal-02170687

<https://hal.science/hal-02170687v1>

Preprint submitted on 3 Jul 2019 (v1), 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

In this paper, we introduce a novel method to generate *interpretable* regression function estimators. The idea is based on called *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 $\mathbf{D}_n = ((\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n))$, 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 $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is measured by its quadratic risk, defined as

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

Thanks to Hypothesis (H1), we have

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

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,

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

The aim of this paper is to provide interpretable learning algorithms (see Section 1.2 for a discussion on the notion of interpretability) that generate \mathcal{G}_n so that the associated empirical risk minimizer g_n is consistent, i.e. g_n converges to g^* as $n \rightarrow \infty$. More precisely, we show the weak consistency of the estimator g_n , i.e. its excess of risk

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

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:

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

where $\mathbf{X}[i]$ is the i^{th} 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

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

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 (see Section 1.2).

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 fulfil *coverage* and *significance* conditions (see Definition 2.1). 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:

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

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

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

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

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

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

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

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

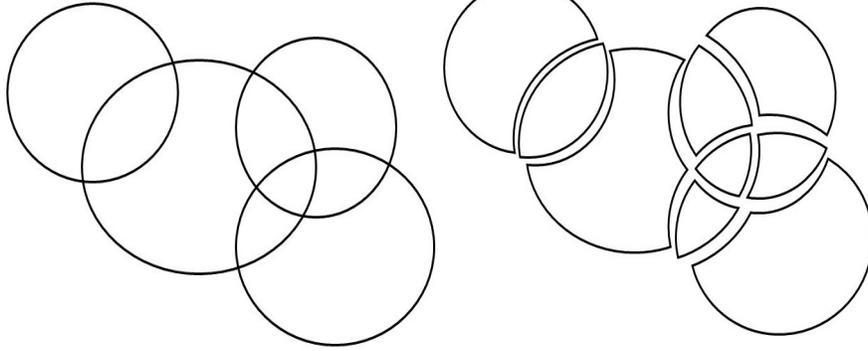


Figure 1: The four elements of \mathcal{C} .

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 $\text{Im}(\varphi_{\mathcal{C}})$.

For each element \mathbf{r} of \mathcal{C} , the cells of the partition generated by \mathcal{C} that are included in \mathbf{r} are gathered in

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

We also introduce the maximal (resp. minimal) redundancy of \mathcal{C} on a subset $\mathbf{r} \in \mathcal{C}$:

$$M(\mathcal{C}, \mathbf{r}) := \max_{x \in \mathbf{r}} \#\varphi_{\mathcal{C}}(x)$$

$$m(\mathcal{C}, \mathbf{r}) := \min_{x \in \mathbf{r}} \#\varphi_{\mathcal{C}}(x).$$

We shorten $M(\mathcal{C}, \mathbf{c})$ in $M(\mathcal{C})$ and $m(\mathcal{C}, \mathbf{c})$ in $m(\mathcal{C})$.

Remark 2. If \mathcal{C} is a partition then for any $\mathbf{r} \in \mathcal{C}$, we have $\mathcal{P}(\mathbf{r}) = \{\mathbf{r}\}$ and $M(\mathcal{C}, \mathbf{r}) = m(\mathcal{C}, \mathbf{r}) = 1$.

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 (see Section 1.2). 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).

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

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

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

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

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

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

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 [2016], there are several meanings of the term 'interpretable' and no rigorous mathematical foundation of the concept. In this paper, *interpretability* correspond to parsimonious characterisation of the estimators of g^* generated by a given algorithm, *i.e.* the facility to describe the generated model in human words. Nowadays, the most popular

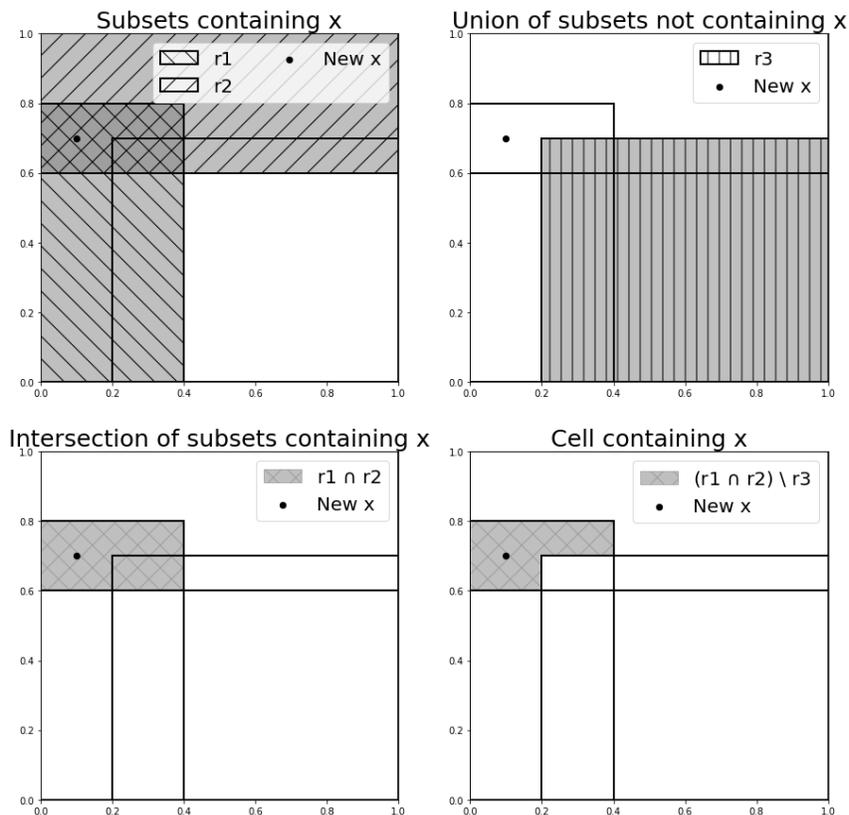


Figure 3: Evaluation steps of the cell containing $\mathbf{x} = (0.1, 0.7)$ of the partition generated from the covering of $[0, 1]^2$, $\mathcal{C} = \{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3\}$. Using partition from a covering allows to generate complex cells with a simple interpretation (\mathbf{r}_1 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 .

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 summarise them. 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 im-

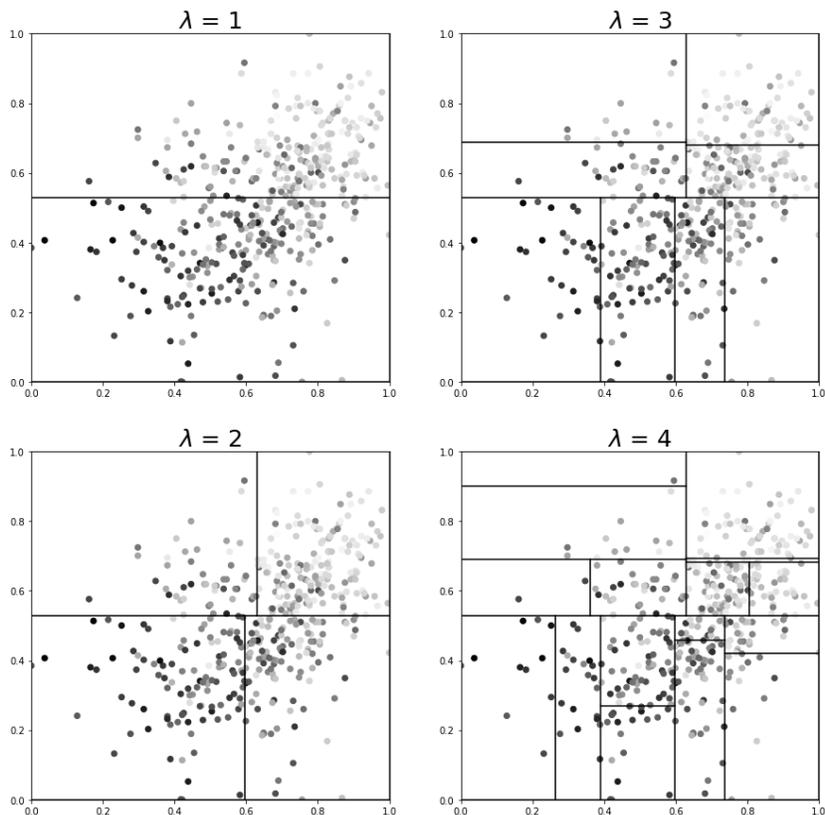


Figure 4: Partitions generated by fully deployed decision tree algorithm, it means without pruning, for a maximal depth $\lambda \in \{1, 2, 3, 4\}$.

portance 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 algorithms that only generate interpretable models, such as rule-based algorithms.

The interpretability of the rule-based algorithms of type (3) is achieved when the length k of each rule is small. But 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

1 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 (10) and (11) 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 $\mathbf{X}_1, \dots, \mathbf{X}_n$. For any $\mathbf{r} \subseteq \mathbb{R}^d$ such that $\mathbb{P}_n(\mathbf{r}) > 0$, we also denote

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

and

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

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)

$$\exists \alpha \in [0, 1/2), \forall \mathbf{r} \in \mathcal{C}_n, \mathbb{P}_n(\mathbf{r}) > n^{-\alpha} \quad a.s., \quad (8)$$

for n sufficiently large;

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

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

$$\mathcal{C}_n = \mathcal{C}_n^s \cup \mathcal{C}_n^i \quad \text{a.s.}, \quad (9)$$

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

$$\mathcal{C}_n^s := \left\{ \mathbf{r} \in \mathcal{C}_n : \beta_n |\mathbb{E}_n[Y|\mathbf{X} \in \mathbf{r}] - \mathbb{E}_n[Y]| \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma^2)_+} \right\}, \quad (10)$$

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

$$\mathcal{C}_n^i := \left\{ \mathbf{r} \in \mathcal{C}_n \setminus \mathcal{C}_n^s : \varepsilon_n \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma^2)_+} \right\}, \quad (11)$$

and their redundancies satisfy

$$\frac{M(\mathcal{C}_n^s)}{m(\mathcal{C}_n^s)} = o_{\mathbb{P}}(\beta_n^{-2} \wedge n^{1/2-\alpha}) \quad (12)$$

and

$$\frac{M(\mathcal{C}_n^i)}{m(\mathcal{C}_n^i)} = o_{\mathbb{P}}(\varepsilon_n^{-2} \wedge n^{1/2-\alpha}) \quad . \quad (13)$$

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 fulfils the significance condition (10) 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 (11) 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 (12) and (13) 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

$$\mathbb{P}_n \left(\mathbf{r}_i \cap \left\{ \bigcup_{1 \leq j \leq i-1} \mathbf{r}_j \right\} \right) \leq \gamma \mathbb{P}_n(\mathbf{r}_i), \quad 1 \leq i \leq \#\mathcal{C}_n .$$

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

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

Thus (12) and (13) 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 (10) 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 (10) 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 $n \rightarrow \infty$. On the one hand, we obtain thanks to several applications of the Central Limit Theorem

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

On the other hand, we obtain

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

Then

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

Thus (10) holds for $\mathbf{r} = A$ with high probability for n sufficiently large. Note that for similar reasons (10) also holds with high probability for $\mathbf{r} = A^c$, $n^{-1/4} = o(\beta_n)$ and n sufficiently large. Finally, conditions (8), (12) and (13) 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 (10) does not follow from a condition on the diameter of the subset. On the opposite, the insignificant condition (11) 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 $\mathbf{x}_1^n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in (\mathbb{R}^d)^n$, let

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

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:

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

$$\begin{aligned} \text{where } \Pi_n &:= \{ \mathcal{P}(\mathcal{C}_n(d_n)) : d_n \in \mathcal{S}^n \} \text{ for any } n \in \mathbb{N}^*; \\ &\forall n \in \mathbb{N}^*, \{ \mathbf{c} \times \mathbb{R}, \mathbf{c} \in \mathcal{C}_n \} \subseteq \mathcal{B}, \end{aligned} \tag{H6}$$

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

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

$$\ell(g^*, g_n) = o_{\mathbb{P}}(1). \tag{14}$$

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 $f : \mathcal{S} \rightarrow \mathbb{R}$ 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 $v : \mathcal{F} \rightarrow \mathbb{R}$, $\|v\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |v(f)|$.

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

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 $Z_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} 0$.

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

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

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

3.1 Empirical estimation of conditional expectations

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

Proposition 3.2. Let $\mathcal{B} \subseteq \mathcal{B}_S$ 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 $\mathcal{B}_n := \{A \in \mathcal{B}, \mathbb{Q}_n(A) \geq n^{-\alpha}\}$,

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

Proof of Proposition 3.2. Let $\varepsilon > 0$. First, for any $f \in \mathcal{F}$ and $A \in \mathcal{B}_n$, since $\mathbb{Q}_n(A) > 0$ and then $\mathbb{Q}(A) > 0$,

$$\begin{aligned} & |\mathbb{E}_n[f | A] - \mathbb{E}[f | A]| \\ &= \left| \frac{\int_A f d\mathbb{Q}_n}{\mathbb{Q}_n(A)} - \frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)} \right| \\ &= \left| \frac{\mathbb{Q}(A) (\int_A f d\mathbb{Q}_n - \int_A f d\mathbb{Q}) + (\mathbb{Q}(A) - \mathbb{Q}_n(A)) \int_A f d\mathbb{Q}}{\mathbb{Q}(A)\mathbb{Q}_n(A)} \right| \\ &\leq \left| \frac{\int_A f d\mathbb{Q}_n - \int_A f d\mathbb{Q}}{\mathbb{Q}_n(A)} \right| + \left| (\mathbb{Q}(A) - \mathbb{Q}_n(A)) \frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)\mathbb{Q}_n(A)} \right|. \end{aligned} \quad (15)$$

Now, according to Proposition 3.1,

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

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,

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

and

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

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

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

Then (15) 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 ,

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

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

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

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

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

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

which, together with Remark 8 again, proves the proposition. \square

Corollary 3.1. *Let $\mathcal{B} \subseteq \mathcal{B}_S$ 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*

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

and

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

Proof of Corollary 3.1 (16). Let $L = \text{ess sup } Y$, $i \in \mathbb{N}$, and $f_i \in \mathcal{L}^1(\mathbb{Q})$ be defined by

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

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

Proof of Corollary 3.1 (17). This part follows from (17) since Y is bounded and

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

□

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 $\text{Diam}(\mathbf{r}) = \sup_{x \in \mathbf{r}, x' \in \mathbf{r}} \|x - x'\|$, 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 (8) 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*

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

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

Proof. Let us show that each cell is significant or insignificant. Thanks to Condition (8), Corollary 3.1 Eq. (17) and Remark 7,

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

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 (13) is automatically satisfied for cells of a partition, the desired result will follow if we check that

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

converges to 0.

From (19) we remark that

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

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

$$\begin{aligned} \mathbb{V}(g^*(\mathbf{X}) \mid \mathbf{X} \in \mathbf{r}) &= \mathbb{V}(g^*(\mathbf{X}_{\mathbf{r}})) \\ &= \frac{1}{2} \mathbb{V}(g^*(\mathbf{X}_{\mathbf{r}}) - g^*(\mathbf{X}'_{\mathbf{r}})) \\ &\leq \frac{1}{2} \mathbb{E}((g^*(\mathbf{X}_{\mathbf{r}}) - g^*(\mathbf{X}'_{\mathbf{r}}))^2). \end{aligned}$$

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

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

By uniform continuity, the condition (18) implies that

$$\varepsilon_n \leq 2^{-1/2} \max_{\mathbf{r} \in \mathcal{P}_n} \left(w(\text{Diam}(\mathbf{r})) \right) + O_{\mathbb{P}}(n^{\alpha/2-1/4}) = o_{\mathbb{P}}(1).$$

Thus, from (11), each cell which is not significant is insignificant and the corresponding covering sequence is suitable. \square

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 $g : \mathbb{R}^d \rightarrow [-L, L]$ depending on the data $D_n = ((\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n))$. If g_n satisfies (2) then*

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

Hence, to prove (14) it is sufficient to prove that:

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

and

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

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

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

From (7),

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

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:

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

where

$$\mathbb{B}_n := \sum_{A \in \mathcal{P}(\mathcal{C}_n)} (\mathbb{E}[Y | \mathbf{X} \in A] - \mathbb{E}[Y])^2 \mathbb{P}(\mathbf{X} \in A). \tag{23}$$

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 $M(\mathcal{C}_n, \mathbf{r})$,

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

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

$$\begin{aligned}
\mathbb{B}_n &\geq \sum_{\mathbf{r} \in \mathcal{C}_n^s} \frac{1}{M(\mathcal{C}_n, \mathbf{r})} (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - \mathbb{E}[Y])^2 \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - \mathbb{E}[Y])^2 \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} U_{\mathbf{r}}^2 \mathbb{P}(\mathbf{X} \in \mathbf{r}) \tag{24}
\end{aligned}$$

where

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

Let's define

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

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 :

$$\begin{aligned}\Delta_n &= \sup_{\mathbf{r} \in \mathcal{C}_n} \{(V_{n,\mathbf{r}} - U_{\mathbf{r}})(V_{n,\mathbf{r}} + U_{\mathbf{r}})\} \\ &= O_{\mathbb{P}^*}(n^{\alpha-1/2}).\end{aligned}\tag{25}$$

Continuing (24),

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

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

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

Thus

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

Let

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

Using again Corollary 3.1 leads to

$$\Delta'_n = O_{\mathbb{P}^*}(n^{\alpha-1/2})\tag{26}$$

Thus,

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

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

$$\begin{aligned}\mathbb{V}(Y|\mathbf{X} \in \mathbf{r}) - \sigma^2 &= \mathbb{E} \left[(Y - \mathbb{E}[Y | \mathbf{X} \in \mathbf{r}])^2 | \mathbf{X} \in \mathbf{r} \right] - \sigma^2 \\ &= \mathbb{E} \left[(\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X}))^2 | \mathbf{X} \in \mathbf{r} \right].\end{aligned}\tag{27}$$

Hence we have

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

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 $m(\mathcal{C}_n^s)$,

$$\begin{aligned}
\mathbb{B}_n &\geq \beta_n^{-2} \frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)} \times \\
&\quad \sum_{A \in \mathcal{P}(\mathcal{C}_n^s)} \left(\mathbb{E} \left[(\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X}))^2 \mid \mathbf{X} \in A \right] \right) \mathbb{P}(\mathbf{X} \in A) \\
&\quad - (\beta_n^{-2} \Delta'_n + \Delta_n) \tag{28}
\end{aligned}$$

We remark that

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

Let us define

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

Since $\mathcal{C}_n = \mathcal{C}_n^s \cup \mathcal{C}_n^i$,

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

Continuing (28),

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

From (22) and (29), we conclude:

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

using Remark 7, Equations (25) and (26) and (H4). Regarding the insignificant part of the within group variance and assuming that \mathcal{C}_n^i is not empty, we have

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

Using (27) we have

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

Then, (H4)

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

Hence, (21) is proved.

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 $\mathbb{R}^d \rightarrow [-L, L]$ and

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

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,

$$\begin{aligned} & \mathbb{P} \left\{ \sup_{g \in \mathcal{G}_c \circ \Pi_n} |\mathcal{L}_n(g) - \mathcal{L}(g)| > \varepsilon \right\} \\ & \leq 8\mathbb{E} \left[\mathcal{N}_1 \left(\frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, \mathbf{X}_1^n \right) \right] \exp \left\{ \frac{-n\varepsilon^2}{128.(4L^2)^2} \right\}, \quad (30) \end{aligned}$$

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

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 $g_1, \dots, g_N : \mathbb{R}^d \rightarrow [-L, L]$ with the property that for every $g \in \mathcal{G}_c \circ \Pi_n$ there is a $j \in \{1, \dots, N\}$ 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

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

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 $g_1, \dots, g_N \in \mathcal{G}$ 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 ,

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

Finally,

$$\sup_{z_1, \dots, z_m \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}, m \leq n} \mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, z_1^m\right) \leq \Delta(\Pi_n) \left\lceil \frac{64L^2}{\varepsilon} \right\rceil^{\mathcal{M}(\Pi_n)}. \quad (31)$$

According to (30) and (31) we have:

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

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 Application

In this section we propose a simple algorithm to generate a suitable sequence of data-dependent coverings using the Random Forests algorithm (RF) Breiman [2001] as rule generator. The interest is double; first, it shows that there exists a sequence of suitable data-dependent coverings in practice. Second, it could prove the consistency of an estimator generated from RF. For 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 covering \mathcal{C}_n such that $\mathcal{C}_n \subseteq \mathcal{C}$ satisfies (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 Algorithm

The proposed algorithm is an easy way to generate an estimator using data-dependent coverings. It can be decomposed into four steps.

1. The generation of RF with m_{tree} trees fully deployed (without pruning).
2. For a chosen $\alpha \in (0, 1/4)$, $\beta_n = n^{\alpha/2-1/4}$ and $\varepsilon_n = \beta_n \mathbb{V}_n(Y)$, it extracts all significant and insignificant rules according to (H3) and (H4) from all nodes and leaves of all trees generated by RF.
3. The selection of a minimal set of rules using Algorithm 1. The redundancy is controlled recursively as described in Remark 4; 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.
4. If the selected set of rules does not form a covering, generation of a unique *no-rule* that is one of the smallest hyperrectangle satisfying (H3) containing the remaining points.

Remark 10. The sequence (ε_n) of the no-rule condition (11) is not controlled by this algorithm. Indeed, the no-rule is added to ensure a covering without any control on its variance.

4.2 Simulation

We generate artificial $n = 2000$ data following the regression setting

$$Y = \eta_1 X_1 + \eta_2 X_2 + Z,$$

where $Z \sim \mathcal{N}(0, 1)$, $X_1, X_2 \sim \mathcal{U}(-1, 1)$, $\eta_1 = 3.5$ and $\eta_2 = -2.5$.

We $\alpha = 1/4 - 1/100$. The data are randomly split into a training set and a test set, with a ratio of 80% / 20%, respectively. We use RF with $m_{tree} = 50$ generating 100808 rules among which 73 are significant according to (8) and (10) and 18 are insignificant according to (8) and (11). Then, the selection process, with $\gamma = 0.95$, extracts a set of 5 significant rules which cover 57% of the training data and add 5 insignificant rules to generate a covering (see Table 1). It is no necessary to add a no-rule.

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 ;

```
1  $\mathcal{C}_n \leftarrow \arg \max_{\mathbf{r} \in S} \mathbb{P}_n(\mathbf{r})$ ;  
2  $S \leftarrow S \setminus \mathcal{C}_n$ ;  
3 while  $\sum_{\mathbf{r} \in \mathcal{C}_n} \mathbb{P}_n(\mathbf{r}) < 1$  do  
4    $\mathbf{r}^* \leftarrow \arg \max_{\mathbf{r} \in S} \mathbb{P}_n(\mathbf{r})$ ;  
5   if  $\mathbb{P}_n(\mathbf{r}^* \cap \{\cup_{\mathbf{r} \in \mathcal{C}_n} \mathbf{r}\}) \leq \gamma \mathbb{P}_n(\mathbf{r}^*)$  then  
6      $\mathcal{C}_n \leftarrow \mathcal{C}_n \cup \mathbf{r}^*$ ;  
7    $S \leftarrow S \setminus \mathbf{r}^*$ ;  
8   if  $\#S = 0$  then  
9     Break ;  
10 end  
11 if  $\sum_{\mathbf{r} \in \mathcal{C}_n} \mathbb{P}_n(\mathbf{r}) < 1$  then  
12   while  $\sum_{\mathbf{r} \in \mathcal{C}_n} \mathbb{P}_n(\mathbf{r}) < 1$  do  
13      $\mathbf{r}^* \leftarrow \arg \min_{\mathbf{r} \in I} \mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r})$ ;  
14     if  $\mathbb{P}_n(\mathbf{r}^* \cap \{\cup_{\mathbf{r} \in \mathcal{C}_n} \mathbf{r}\}) \leq \gamma \mathbb{P}_n(\mathbf{r}^*)$  then  
15        $\mathcal{C}_n \leftarrow \mathcal{C}_n \cup \mathbf{r}^*$ ;  
16        $I \leftarrow I \setminus \mathbf{r}^*$ ;  
17       if  $\#I = 0$  then  
18         Break ;  
19     end  
20 return  $\mathcal{C}_n$ ;
```

Rule	Conditions	Coverage	Prediction	Var	MSE
R1	$X0 \in [0.09, 1.0]$	0.26	2.99	2.35	4.07
	$X1 \in [-1.0, 0.13]$				
R2	$X0 \in [-0.06, 1.0]$	0.26	2.98	2.39	4.13
	$X1 \in [-1.0, -0.01]$				
R3	$X0 \in [-1.0, -0.03]$	0.21	-3.33	2.51	4.57
	$X1 \in [0.13, 1.0]$				
R4	$X0 \in [-1.0, -0.32]$	0.22	-3.46	2.21	4.33
	$X1 \in [-0.18, 1.0]$				
R5	$X0 \in [-1.0, 0.05]$	0.19	-3.47	2.42	4.84
	$X1 \in [0.31, 1.0]$				
R6	$X0 \in [-0.03, 0.52]$	0.17	-0.08	1.76	7.42
	$X1 \in [-0.24, 1.0]$				
R7	$X0 \in [-1.0, -0.06]$	0.18	-2.4	1.95	6.23
	$X1 \in [-0.18, 0.61]$				
R8	$X0 \in [-0.59, -0.01]$	0.18	-0.27	2.02	7.42
	$X1 \in [-1.0, 0.28]$				
R9	$X0 \in [0.09, 1.0]$	0.19	0.44	2.02	7.35
	$X1 \in [0.13, 1.0]$				
R10	$X0 \in [-1.0, -0.37]$	0.2	-1.41	2.04	6.99
	$X1 \in [-1.0, 0.13]$				

Table 1: Summary of selected rules.

4.3 Results

Regarding the accuracy on the test set, RF has a MSE score of 1.32 and the data-dependent covering estimator has a MSE score of 1.57 (see Fig 5). This loss of accuracy is the price of turning a black box model into an interpretable one. The Figure 5 is composed of four graphics: the dataset (upper left), the model generated by RF (upper right), the model generated by the selected set of rules (lower left) and the 10 selected rules (lower right). It is interesting to note that the cells of the partition generated by the 10 set covering are represented in the bottom left graphic.

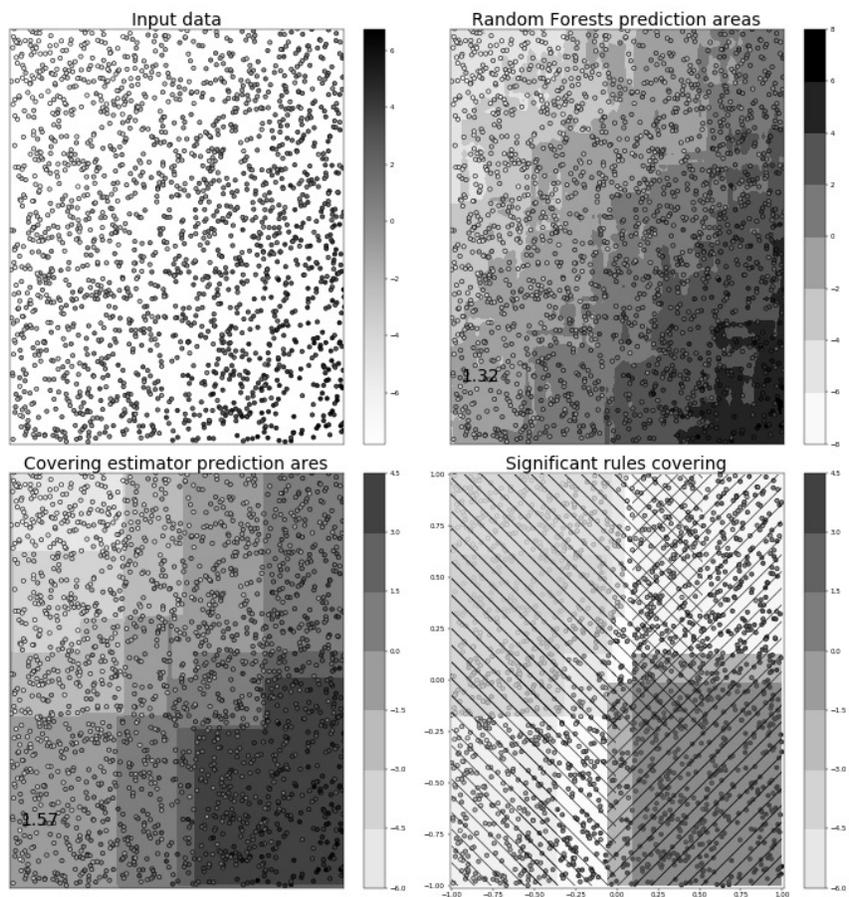


Figure 5: Top-left: the dataset, the darker the color the higher the value of Y . Top-right: the model inferred by RF. Bottom-left: the model inferred by the minimal data-dependent covering. Bottom-right: the model inferred by the 10 selected rules.

4.4 Comments

The presented algorithm has no control over the sequence (ε_n) of the no-rule. So, it cannot guarantee that the generated sequence of data-dependent coverings \mathcal{C}_n is suitable. However this application emphasises that data-dependent coverings are very efficient to generate an interpretable estimator without a significant loss of accuracy. With 0.01% of the rules from RF, it constructs an estimator with 1.19 *MSE* ratio accuracy compared with the RF one.

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 set, 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 an algorithm that extracts from any rule generator a suitable data-dependent covering. We apply it to Random Forest.

This very effective approach appeals for an algorithm that generates significant and insignificant rules and a suitable sequence of data-dependent coverings on its own. Generating insignificant rules with shrinking diameters as in Proposition 3.3, a control on the sequence (ε_n) of the insignificant condition (11) seems possible. 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. Finally, the scope could be broaden from the regression setting to the classification one by adapting the significant condition.

Appendix

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 $\|u - l\|_{L_2(\mathbb{Q})} < \varepsilon$. 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 $J_{[]}(\delta, \mathcal{F}, \mathcal{L}_2(\mathbb{Q})) < \infty$ is \mathbb{Q} -Donsker.*

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

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

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

$$\begin{aligned} \mathbb{Q}(\mathbb{R}^{k-1} \times]t_{i-1,k}, t_{i,k}[\times \mathbb{R}^{d-k} \times [-L, L]) &\leq \varepsilon/2d \\ \mathbb{Q}(\mathbb{R}^{k-1} \times]t_{i-1,k}, t_{i,k}[\times \mathbb{R}^{d-k} \times [-L, L]) &\geq \varepsilon/2d \\ \mathbb{Q}(\mathbb{R}^{k-1} \times]t_{0,k}, t_{i,k}[\times \mathbb{R}^{d-k} \times [-L, L]) &\geq i\varepsilon/2d \\ \mathbb{Q}(\mathbb{R}^{k-1} \times]t_{0,k}, t_{\lceil 2d/\varepsilon \rceil, k}[\times \mathbb{R}^{d-k} \times [-L, L]) &\geq 1 - \varepsilon/2d \text{ (since } \lceil 2d/\varepsilon \rceil \geq 2d/\varepsilon - 1) \\ \mathbb{Q}(\mathbb{R}^{k-1} \times]t_{\lceil 2d/\varepsilon \rceil, k}, +\infty[\times \mathbb{R}^{d-k} \times [-L, L]) &\leq \varepsilon/2d \end{aligned}$$

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

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

Consider now the set of brackets defined by (see Fig 6)

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

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

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

The term after the first equality corresponds to the integration of the hatched area in Figure 6 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 $k \in \{1, \dots, d\}$,

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

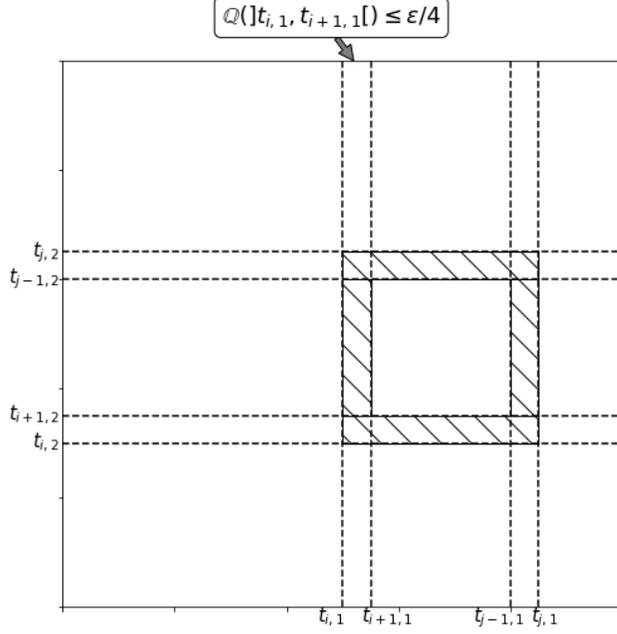


Figure 6: Example of bracket for $d = 2$. With $l = \mathbf{1}_{[t_{i+1,1}; t_{j-1,1}] \times [t_{i+1,2}; t_{j-1,2}]}$ and $u = \mathbf{1}_{[t_{i,1}; t_{j,1}] \times [t_{i,2}; t_{j,2}]}$, for any rectangle A , $\mathbf{1}_A \in [l, u]$ if and only if its boundary $\bar{A} \setminus \mathring{A}$ is included in the hatched area.

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,

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

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

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

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

References

- L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 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. 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.
- 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*, 2016.
- 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.
- Q. Zhao and S. S. Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 2003.