



**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. 2020. hal-02170687v4

**HAL Id: hal-02170687**

**<https://hal.science/hal-02170687v4>**

Preprint submitted on 23 Sep 2020 (v4), 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 pair of random variables in  $\mathbb{R}^d \times \mathbb{R}$  of unknown distribution  $\mathbb{Q}$  such that

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

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

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

The accuracy of  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $\mathbb{E}[g(\mathbf{X})^2] < \infty$  is measured by its quadratic risk defined as

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

Then we have

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

where the arg min ranges over all the measurable functions  $g$  with  $\mathbb{E}[g(\mathbf{X})^2] < \infty$ .

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  $\mathbb{Q}$ .

Let us denote the empirical measure for any  $\mathbf{r} \subseteq \mathbb{R}^d$  and  $I \subseteq \mathbb{R}$

$$\mathbb{Q}_n(\mathbf{r} \times I) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i \in \mathbf{r}} \mathbf{1}_{Y_i \in I}.$$

For simplicity of notation, we write  $\mathbb{Q}_n(\mathbf{r})$  instead of  $\mathbb{Q}_n(\mathbf{r} \times \mathbb{R})$  as well as  $\mathbb{Q}(\mathbf{r})$  instead of  $\mathbb{Q}(\mathbf{r} \times \mathbb{R})$  for any  $\mathbf{r} \subseteq \mathbb{R}^d$ .

We consider a set of functions  $\mathcal{G}_n$ . Any  $\tilde{g}_n \in \arg \min_{g \in \mathcal{G}_n} \mathcal{L}(g)$  minimizes the risk over  $\mathcal{G}_n$  but is not an estimator since it depends on the knowledge of  $\mathbb{Q}$ . Following the Empirical Risk Minimization (ERM) principle (Vapnik, 2013, Section 1.5), we define the empirical risk and an empirical risk minimizer over  $\mathcal{G}_n$  as, respectively,

$$\mathcal{L}_n(g) := \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 \text{ and } g_n \in \arg \min_{g \in \mathcal{G}_n} \mathcal{L}_n(g). \quad (1.1)$$

The aim of this paper is to provide *interpretable*<sup>1</sup> learning algorithms that generate weakly consistent empirical risk minimizers  $g_n$ , i.e. such that their excess of risk  $\ell(g^*, g_n)$  fulfills

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

Note that  $\mathbb{P}$  refers to an underlying probability measure of reference that is unique (on the contrary to  $\mathbb{Q}$ ) and that  $\mathbb{E}$  corresponds to the expectation under  $\mathbb{P}$ .

## 1.1 Rule-based algorithms using partitions and quasi-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* (Karalić and Bratko, 1997), *M5 Rules* (Holmes et al.,

<sup>1</sup>A formal definition of interpretability convenient to our framework is proposed and discussed in Section 1.2.

1999). In these models, the estimator is explained by the realization of a simple condition, an *If-Then* statement of the form:

$$\begin{aligned} \text{IF} & \quad (\mathbf{X} \in c_1) \text{ And } (\mathbf{X} \in c_2) \text{ And } \dots \text{ And } (\mathbf{X} \in c_k) & (1.2) \\ \text{THEN} & \quad g_n(\mathbf{X}) = p \end{aligned}$$

where  $c_i \subseteq \mathbb{R}^d$  and  $p \in \mathbb{R}$ . The *If* part, called the *condition* of the rule, or simply the rule, is composed of the conjunction of  $k$  tests 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*. Let us define for any  $\mathbf{r} \subseteq \mathbb{R}^d$  such that  $\mathbb{Q}(\mathbf{r}) > 0$  and for any  $h$  measurable

$$\mathbb{E}[h(Y) \mid \mathbf{X} \in \mathbf{r}] := \frac{\mathbb{E}[h(Y)\mathbf{1}_{\mathbf{X} \in \mathbf{r}}]}{\mathbb{P}(\mathbf{X} \in \mathbf{r})}$$

and

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

and their empirical counterparts for any  $\mathbf{r} \subseteq \mathbb{R}^d$  such that  $\mathbb{Q}_n(\mathbf{r}) > 0$ :

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

and

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

Any such algorithm following the ERM principle associates to  $\mathcal{P}_n$  the estimator

$$g_n = \sum_{\substack{A \in \mathcal{P}_n \\ \mathbb{Q}_n(A) > 0}} \mathbb{E}_n[Y \mid \mathbf{X} \in A] \mathbf{1}_A \quad (1.3)$$

which minimizes the empirical risk among the class of piecewise constant functions over  $\mathcal{P}_n$ .

Those algorithms use the dataset  $\mathbf{D}_n$  twice: the partition  $\mathcal{P}_n = \mathcal{P}_n(\mathbf{D}_n)$  is constructed according to the dataset, and the dataset is used again to compute the value of  $g_n$  on each element of  $\mathcal{P}_n$  as in (1.3).

In this paper, we present a novel family of algorithms by relaxing the condition that the rules constitute a partition of  $\mathbb{R}^d$ . We consider collections of sets  $\mathcal{C}_n = \mathcal{C}_n(\mathbf{D}_n)$  which asymptotically cover  $\mathbb{R}^d$  (see (H4)). We call such a collection a *quasi-covering*, and a *quasi-partition* if the sets do not intersect, which we do not need to assume in general. Such a quasi-covering is turned into a quasi-partition  $\mathcal{P}(\mathcal{C}_n)$  using the definition below and the ERM principle

defines an estimator  $g_n$  by (1.3) associated to  $\mathcal{P}_n = \mathcal{P}(\mathcal{C}_n)$ . The main interest of quasi-covering estimators compared to partition estimators is the interpretability. This will be discussed in Section 1.2.

**Definition 1.1.** Let  $\mathcal{C}$  be a finite collection of sets 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 introduce the maximal (resp. minimal) redundancy of  $\mathcal{C}$  as

$$M(\mathcal{C}) := \max_{\mathbf{x} \in \mathbf{c}} \#\varphi_{\mathcal{C}}(\mathbf{x}), \quad m(\mathcal{C}) := \min_{\mathbf{x} \in \mathbf{c}} \#\varphi_{\mathcal{C}}(\mathbf{x}).$$

Moreover, the cells of  $\mathcal{P}(\mathcal{C})$  which are included in an element of  $\mathcal{C}' \subset \mathcal{C}$  are gathered in  $\mathcal{P}_{\mathcal{C}}(\mathcal{C}')$  :

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

We provide a more explicit characterisation of  $\mathcal{P}(\mathcal{C})$  in Proposition 1.1.

**Proposition 1.1.** Let  $\mathcal{C}$  be a finite collection of sets of  $\mathbb{R}^d$ . Then

$$\mathcal{P}(\mathcal{C}) = \left\{ \bigcap_{\mathbf{r} \in \tilde{\mathcal{C}}} \mathbf{r} \setminus \bigcup_{\mathbf{r} \in \mathcal{C} \setminus \tilde{\mathcal{C}}} \mathbf{r} : \tilde{\mathcal{C}} \subseteq \mathcal{C} \right\} \setminus \{\emptyset\}.$$

The proof of Proposition 1.1 is deferred in Appendix. Figures 1 and 2 illustrate the fact that  $\mathcal{P}(\mathcal{C})$  can contain (much) more sets than  $\mathcal{C}$  itself.

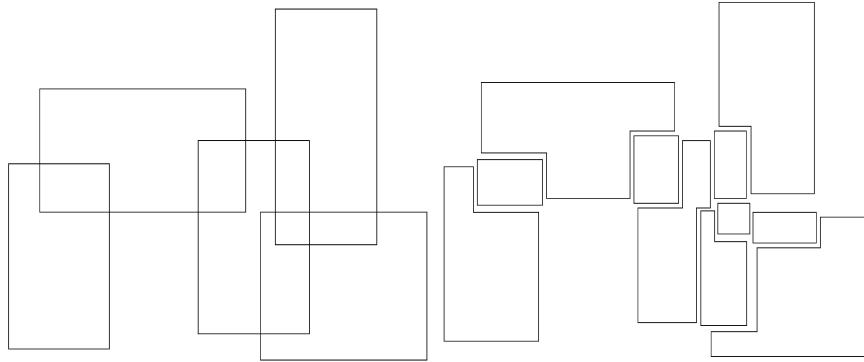


Figure 1: The 5 overlapping elements of  $\mathcal{C}$ . Figure 2: The 11 cells of the partition  $\mathcal{P}(\mathcal{C})$ .

As the construction of  $\mathcal{P}(\mathcal{C})$  can be time consuming, it is noteworthy that there is no need to fully describe  $\mathcal{P}(\mathcal{C})$  to compute the value of  $g_n(\mathbf{x})$  for some  $\mathbf{x} \in \mathbb{R}^d$ . The trick is to identify the unique cell of  $\mathcal{P}(\mathcal{C})$  which contains  $\mathbf{x}$ . By creating binary vectors of size  $\#\mathcal{C}$ , whose value is 1 if  $\mathbf{x}$  fulfills the rule’s condition and 0 otherwise, this cell identification is a simple sequence of vectorial operations. Figure 3 is an illustration of this process (cf Margot et al. (2018) for more details).

As compared to ours, the predictors designed by Random Forests (RF) and other random rule algorithms do not satisfy the ERM principle (1.1): they are based on the averaging of the activated rules at the point  $\mathbf{x}$ . This averaging is justified empirically and theoretically as it lowers the variance of the prediction thanks to the independence of the random rules as explained by Breiman (2001). The averaging does not decrease the approximation error of the prediction and there are few results about their consistency, see Denil et al. (2013); Scornet et al. (2015) for RF only. On the opposite, our estimator defined by (1.3) is the empirical conditional expectation on the cell of  $\mathcal{P}(\mathcal{C}_n)$ . Under some conditions, the approximation error of the prediction is low thanks to the ERM principle. The variance will be controlled through the significance condition (see Definition 2.1 below) as in the RIPE algorithm of Margot et al. (2018).

## 1.2 Interpretability

In many fields, such as healthcare, marketing or asset management, decision makers prefer interpretable models rather than models with better accuracy but uninterpretable. As explained in Lipton (2018), there are several meanings of *interpretability* depending on the user’s desiderata and the expected properties of algorithms. In this paper, we embrace the intuitive definition of model interpretability of Biran and Cotton (2017): **Interpretability is the degree to which an observer can understand the cause of a decision.** *Interpretable* models should provide a parsimonious characterization of an estimator of  $g^*$ . Nowadays, the most popular and accurate 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. We assert that the novel family of (quasi-) *covering algorithms* described in Section 4 can achieve another interpretability-accuracy trade-off by reducing the complexity of the generated models while keeping weak consistency guarantees.

There exist two ways of constructing interpretable models. The first one is to create black-box models and then to summarize in a so-called *post-hoc* interpretable algorithm. Recent researches propose to use explanation models such as *LIME* (Ribeiro et al., 2016), *DeepLIFT* (Shrikumar et al., 2019) or *SHAP* (Lundberg and Lee, 2017), to interpret black-box models. These explanation models measure the importance of a feature 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 that only generate interpretable models.

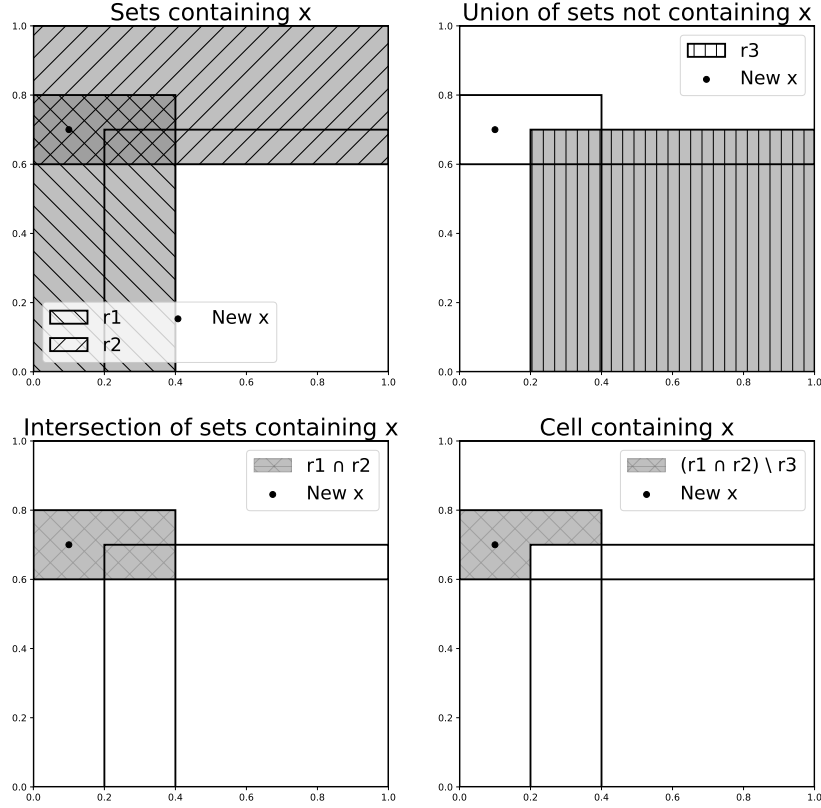


Figure 3: Evaluation steps of the cell containing  $\mathbf{x} = (0.1, 0.7)$  of the partition generated from the covering of  $[0, 1]^2$ ,  $\mathcal{C} = \{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3\}$ . Using partition from a covering allows to generate complex cells with a simple interpretation ( $\mathbf{r}_1$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

The interpretability of the rule-based estimator is achieved when the length  $k$  of each rule and the number of rules are small, i.e. when the associated collection of rules is simple. Considering that a collection  $\mathcal{C}$  with one rule of length  $k$  is as interpretable as a collection with  $k$  rules of length 1 since they consists in the same number of tests, interpretability gets some additivity property. With this in mind, we are able to quantify the interpretability index of a collection  $\mathcal{C}$ .

**Definition 1.2.** *The interpretability index of a collection of rules  $\mathcal{C}$  is defined by*

$$Int(\mathcal{C}) := \sum_{\mathbf{r} \in \mathcal{C}} length(\mathbf{r}). \quad (1.4)$$

By convention, we define the interpretability index of the estimator  $g_n$  generated by a set of rules  $\mathcal{C}_n$  as  $Int(g_n) = Int(\mathcal{C}_n)$ .

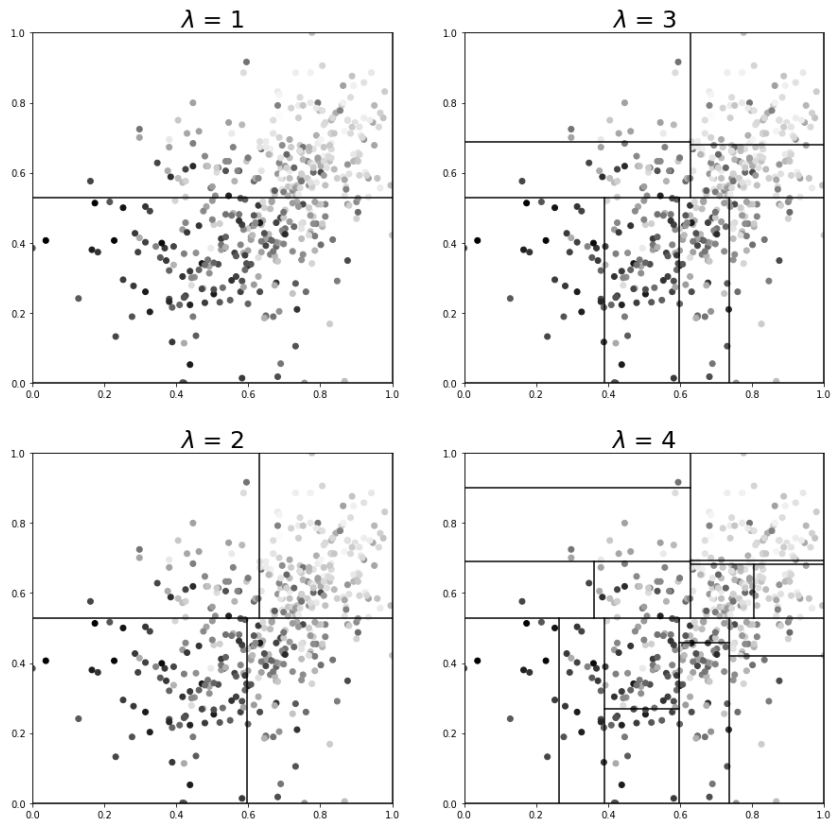


Figure 4: Partitions generated by fully deployed decision tree algorithm without pruning for a maximal depth  $\lambda \in \{1, 2, 3, 4\}$ . Here the number of rules satisfies  $2^\lambda$ ,  $\lambda \in \{1, 2, 3, 4\}$ .

*Remark 1.* The interpretability index  $Int(g_n)$  does not take into account the notion of interaction described in Friedman and Popescu (2008). Indeed, a low interpretability index does not inform whether  $g_n$  exhibits interactions between variables or not. Many interactions may exhibit from a set of  $k$  rules of length 1 whereas it is less the case for a unique rule of length  $k$ . Due to the additivity property, both cases have the same interpretability index.

We aim at designing estimators which combine consistency guarantees keeping an interpretability index as small as possible. In order to prove the consistency of a partition-based estimator  $g_n$ , one usually applies results such as Theorem 13.1 in Györfi et al. (2006), which is a generalization of Stone’s Theorem to data-dependent partitioning estimates (i.e. such that the partitions can depend on the  $Y_i$ ’s). Then, “it is required that the diameter of the cells of the data-dependent partition [...] converge in some sense to zero” (Condition (13.10) in Györfi et al., 2006). However, as illustrated in Figure 4, such



approaches based on fine partitions have a high interpretability index. A cell shrinkage condition such as Condition (13.10) in Györfi et al. (2006) implies necessarily that the number of elements in the partition  $\mathcal{P}_n$  tends to infinity and that the interpretability index tends to infinity as well.

For an estimator defined on a data-dependent quasi-covering, we can get rid of this condition and avoid the condition of shrinkage of the sets in the collection  $\mathcal{C}_n$ . Each prediction can potentially be explained by a small set of rules which are interpretable: see Table 4 in Section 4 for an example. The partition generated may be finer and more complex than a classical data-dependent partition while the explanation of the quasi-covering is well understandable by humans.

Despite the fact that we have not proved the parsimony of the selected set of rules in general yet, the use of quasi-covering rather than partition and conditions different than the usual ones reduce dramatically the number of rules in simple cases. This is discussed in detail in Section 2.1.2.

We prove the consistency of a quasi-covering based estimator  $g_n$  by carefully designing the covering elements. The key notion is the one of suitable data-dependent quasi-covering introduced in Section 2. Proposition 3.2 provides rates of convergence of empirical conditional expectations. This preliminary result of independent interest is crucial to prove the weak consistency of suitable data-dependent covering estimators stated in Theorem 2.1. Our main result Theorem 2.1 is proved in Section 3 whereas proofs of intermediate results are postponed to Appendix. We apply our quasi-covering approach on real data using different rule generators and we compare it to common existing algorithms in Section 4.

## 2 Consistent prediction based on suitable quasi-coverings

In order to ensure the consistency of quasi-covering based algorithms we have to introduce the notion of *suitable* sequence of data-dependent quasi-coverings.

### 2.1 Suitable data-dependent quasi-coverings

#### 2.1.1 Definition

We use the classical notation  $x_+ = \max\{x, 0\}$  for any  $x \in \mathbb{R}$ .

**Definition 2.1.** *Let  $\alpha \in [0, 1/2)$ . We call a sequence  $(\mathcal{C}_n)_{n \geq 1}$  of data-dependent finite collections of sets  $\mathbf{r}$  of  $\mathbb{R}^d$   $\alpha$ -suitable if it satisfies:*

1. *The set coverage condition* (H3)

$$\mathbb{Q}_n(\mathbf{r}) > n^{-\alpha}, \quad \mathbf{r} \in \mathcal{C}_n, n \geq 1.$$

2. *The collection coverage condition* (H4)

$$\mathbb{Q}_n(\mathbf{c}_n^c) = o_{\mathbb{P}}(1)$$

where  $\mathbf{c}_n = \cup_{\mathbf{r} \in \mathcal{C}_n} \mathbf{r}$ .

3. *The significance condition* (H5)

For some constant estimator  $(\sigma_n^2)$  of  $\sigma^2$  and sequences  $\beta_n = o_{\mathbb{P}}(1)$  and  $\varepsilon_n = o_{\mathbb{P}}(1)$ , it holds for all  $n \geq 1$

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

where the significant sets  $\mathcal{C}_n^s$  are defined by

$$\mathcal{C}_n^s := \left\{ \mathbf{r} \in \mathcal{C}_n : \beta_n |\mathbb{E}_n[Y|\mathbf{X} \in \mathbf{r}] - \mathbb{E}_n[Y]| \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma_n^2)_+} \right\} \quad (2.1)$$

and the insignificant sets  $\mathcal{C}_n^i$  are defined by

$$\mathcal{C}_n^i := \left\{ \mathbf{r} \in \mathcal{C}_n \setminus \mathcal{C}_n^s : \varepsilon_n \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma_n^2)_+} \right\}. \quad (2.2)$$

4. *The redundancy condition* (H6)

The redundancies of  $\mathcal{C}_n^s$  and  $\mathcal{C}_n^i$  satisfy

$$\frac{M(\mathcal{C}_n^s)}{m(\mathcal{C}_n^s)} = o_{\mathbb{P}}(\beta_n^{-2} \wedge n^{1/2-\alpha}) \quad (2.3)$$

and

$$\frac{M(\mathcal{C}_n^i)}{m(\mathcal{C}_n^i)} = o_{\mathbb{P}}(\varepsilon_n^{-2} \wedge n^{1/2-\alpha}). \quad (2.4)$$

A few remarks on the conditions of Definition 2.1 are in order. The set coverage condition (H3) controls the minimal number of observations in a set of the collection  $\mathcal{C}_n$ , ensuring that the empirical conditional expectation is close to the conditional expectation given any of the sets in  $\mathcal{C}_n$ . The collection coverage condition (H4) guarantees that  $\mathcal{C}_n$  covers a sufficient part of  $\mathbb{R}^d$ . The significance conditions (H5) control the variance of the rules in two different ways; The significant rules are predicting significantly differently than the average and may have a large variance, the insignificant rules are not significant and have a small variance. The redundancy condition (H6) controls how far from a partition the collections  $\mathcal{C}_n^s$  and  $\mathcal{C}_n^i$  can be. This condition is fulfilled if the sets in these collections are mutually disjoint.

*Remark 2.* The significant condition (2.1) is not related to a condition on the diameter of the set  $\mathbf{r}$ . It is intended to detect areas of the feature space where the conditional expectation of  $Y$  is noticeably different from its unconditional expectation over the whole feature space. See Example 1. The insignificant condition (2.2) can be proved to hold under a shrinkage condition on the diameter of the set, see Proposition 3.3. This is not a necessary condition.

*Remark 3.* We may assume in the following that the noise variance  $\sigma^2$  is known or that a good enough estimator  $\sigma_n^2$  exists. Theorem 2.1 proves the consistency of the procedure under the assumption that  $|\sigma_n^2 - \sigma^2| = O_{\mathbb{P}}(n^{\alpha-1/2})$ . One refers to Liitiäinen et al. (2009) as a reference on the difficult question of the bias-variance trade-off of some residual variance estimators. Most of the existing estimators are based on Nearest-Neighbor methods. Random forests were not used for residual variance estimation until recently, see Ramosaj and Pauly (2019). None of these results is satisfactory in our setting of rule-based procedure. We let for a future work the study of a good enough estimator  $\sigma_n^2$  based on a covering algorithm. We will use on real data in Section 4 the smallest of the empirical variances of the rules.

*Remark 4.* An easy way to ensure (H6) is to constrain the proportion of any element of  $\mathcal{C}_n$  which can be included in the union of the others to be small. Let  $(\mathcal{C}_n)$  be a sequence of collections of sets  $\mathbf{r}$  of  $\mathbb{R}^d$  that fulfills (H3) and  $\gamma \in (0, 1)$ . We consider  $(\mathbf{r}_i)_{1 \leq i \leq \#\mathcal{C}_n}$  any ordering of  $\mathcal{C}_n$ . If

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

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

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

Thus (H6) 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  $\beta_n = o_{\mathbb{P}}(n^{-\alpha/2})$  and  $\varepsilon_n = o_{\mathbb{P}}(n^{-\alpha/2})$ .

### 2.1.2 Advantages of suitable quasi-coverings

We discuss how our approach can provide estimators with low interpretability indices by considering quasi-coverings instead of partitions on the one hand and by introducing conditions on the elements of the quasi-coverings which are completely different from Condition (13.10) in Györfi et al. (2006) on the partition cells on the other hand.

We already discussed and illustrated in Figures 1 and 2 that a quasi-covering  $\mathcal{C}_n$  contains less elements than the quasi-partition  $\mathcal{P}(\mathcal{C}_n)$ . Then its interpretability index is lower.

Moreover,  $\mathcal{P}(\mathcal{C})$  contain more complex elements than those of a partition-based algorithm: the cells in  $\mathcal{P}(\mathcal{C})$  in Figure 2 are not necessarily conjunctions of simple tests on coordinates as the sets in Figure 1. Our covering-based approach generates complex cells without sacrificing interpretability since the algorithm keeps small the number of elements of  $\mathcal{C}$  which have to be designed.

Let us illustrate this point in the following example considering rules designed as hyperrectangles which is usual, see Remark 6.

**Proposition 2.1.** *Let  $C = [0, 1]^d \subset \mathbb{R}^d$ .*

1. The minimal cardinality of a partition  $\mathcal{P}$  of  $\mathbb{R}^d$  by hyperrectangles such that  $C \in \mathcal{P}$  is  $2d + 1$ .
2. The minimal cardinality of a covering  $\mathcal{C}$  of  $\mathbb{R}^d$  by hyperrectangles such that  $C \in \mathcal{P}(\mathcal{C})$  is 2.

The proof of Proposition 2.1 is deferred to Appendix and is a consequence of the fact that  $\mathbb{R}^d \setminus [0, 1]^d$  is a more complex set than an hyperrectangle.

On the other hand, our Condition (H5) suits better the requirement of interpretability than Condition (13.10) of Györfi et al. (2006) as they do not imply any shrinkage of the sets in the collection  $\mathcal{C}_n$  and the number of sets in  $\mathcal{C}_n$  does not necessarily have to grow with  $n$ . We illustrate in the following example this point when  $g^*$  itself is rule-based with a low interpretability index: in this important situation, an estimator with low interpretability index is expected.

**Example 1.** *The condition involved in (2.1) can hold for a set  $\mathbf{r}$  with arbitrary diameter that does not satisfy Condition (13.10) of Györfi et al. (2006). Consider the case  $g^* = \mathbf{1}_A$  for some Borel set  $A$  such that  $0 < \mathbb{P}(\mathbf{X} \in A) < 1$  and assume that  $\sigma^2$  is known. Then  $\mathbf{r} = A$  is a significant set as it satisfies the condition involved in (2.1) for some well-chosen  $\beta_n = o_{\mathbb{P}}(1)$ . Indeed, from the Strong Law of Large Numbers  $k_n := \#\{i : \mathbf{X}_i \in A\} \sim n\mathbb{P}(\mathbf{X} \in A)$  a.s. as  $n \rightarrow \infty$ . Then repeated application of the Central Limit Theorem yields*

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

and

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

It follows that (2.1) holds for  $\mathbf{r} = A$  with

$$\beta_n = \frac{\sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in A) - \sigma^2)_+}}{|\mathbb{E}_n[Y|\mathbf{X} \in A] - \mathbb{E}_n[Y]|} = O_{\mathbb{P}}(n^{-\frac{1}{4}}).$$

For similar reasons (2.1) also holds for  $\mathbf{r} = A^c$  with

$$\beta'_n = \beta_n \vee \frac{\sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in A^c) - \sigma^2)_+}}{|\mathbb{E}_n[Y|\mathbf{X} \in A^c] - \mathbb{E}_n[Y]|} = O_{\mathbb{P}}(n^{-\frac{1}{4}}).$$

Finally it can easily be checked that  $\mathcal{C}_n = \{\mathbf{r} \in \{A, A^c\} : \mathbb{Q}_n(\mathbf{r}) > n^{-\alpha}\}$  defines an  $\alpha$ -suitable sequence of quasi-partitions for any  $\alpha \in (0, \frac{1}{2})$ . The number of rules in  $\mathcal{C}_n$  tends to 2.

*Remark 5.* In Section 4 and for many usual procedures, only hyperrectangles are considered as rules. In Example 1, if  $A$  is assumed to be an hyperrectangle itself,  $A^c$  can be written as the union of a set of hyperrectangles as in the proof of Proposition 2.1 and similar arguments as above apply to prove that an  $\alpha$ -suitable sequence of quasi-coverings which are all subsets of a small set of hyperrectangles can thus be defined accordingly.

Note that in Proposition 2.1, Example 1, and Remark 5, the number of elements of the quasi-coverings and the interpretability index are bounded as  $d$  or  $n$  grows whereas they would need to tend to infinity for any method partition-based or relying on Condition (13.10) of Györfi et al. (2006).

Besides, another advantage is that the check of the conditions involved in (2.1) and (2.2) for elements of a quasi-covering can be done simultaneously with parallel computing on the contrary to any condition on partition cells that depend on each other.

## 2.2 Consistency of data-dependent quasi-covering algorithms

### 2.2.1 Generalization of the partitioning number

Conditions are required to control the complexity of the family of partitions  $\mathcal{P}(\mathcal{C}_n)$  that the algorithm can generate. These conditions use some concepts introduced in (Nobel, 1996, Sec. 1.2) (see also (Györfi et al., 2006, Def 13.1)). The standard definitions have to be adapted since we consider collections which do not necessarily cover  $\mathbb{R}^d$ . To discriminate with the standard definitions for partitions we denote these quantities with tildes.

**Definition 2.2.** Let  $\Pi$  be a family of finite collections of disjoint sets of  $\mathbb{R}^d$ .

1. The maximal number of sets in an element of  $\Pi$  is denoted by

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

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

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

3. The partitioning number  $\widetilde{\Delta}_n(\Pi)$  of  $\Pi$  is defined by

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

The partitioning number is the maximal number of different collections of disjoint non-empty subsets of any  $n$ -point set that can be induced by elements of  $\Pi$  and is smaller than or equal to  $(n+1)^n$ , which is an upper bound of the number of collections of disjoint subsets of  $\mathbf{x}_1^n$  as the number of maps from  $\{1, \dots, n\}$  to  $\{0, \dots, n\}$ .

Classical consistency theorems require to control the partitioning number of the family of partitions  $\Pi_n$  over which the estimators are defined (see (13.7))

and (13.8) in Györfi et al., 2006, Theorem 13.1). As illustrated in Figure 2, the complexity of the sets in  $\mathcal{P}(\mathcal{C}_n)$  may be high even if the complexity of the sets in  $\mathcal{C}_n$  is low. This fact makes the partitioning number of the family of partitions tough, if not impossible, to bound because no control on the shape of the cells is available.

Usually  $\Pi_n$  is chosen as  $\{\mathcal{P}(\mathcal{C}_n(\mathbf{d}_n)) : \mathbf{d}_n \in \mathcal{S}^n\}$ . It can be useful to take into account the building process of the family of sets  $\mathcal{C}_n(\mathbf{d}_n)$  in the evaluation of the partitioning number when the partitioning number of  $\{\mathcal{P}(\mathcal{C}_n(\mathbf{d}_n)) : \mathbf{d}_n \in \mathcal{S}^n\}$  cannot be controlled easily. This is what the definition below enables by considering  $\Pi_n$  as a refinement of  $\{\mathcal{P}(\mathcal{C}_n(\mathbf{d}_n)) : \mathbf{d}_n \in \mathcal{S}^n\}$ .

**Definition 2.3.** Let  $\mathcal{F}$  and  $\mathcal{F}'$  be two families of finite collections of disjoint sets.  $\mathcal{F}'$  is called a refinement of  $\mathcal{F}$  if

$$\forall F \in \mathcal{F}, \quad \exists F' \in \mathcal{F}', \quad \forall A \in F, \quad A = \bigcup_{A' \in F': A' \subseteq A} A'.$$

### 2.2.2 Main result

Our main result provides the consistency of estimators based on suitable quasi-covering sequences.

**Theorem 2.1.** Assume that  $g^*$  and  $\mathbb{Q}$  satisfy (H1), (H2). Let  $\alpha \in [0, \frac{1}{2})$  and  $(\mathcal{C}_n)$  be an  $\alpha$ -suitable sequence of finite collections of sets of  $\mathbb{R}^d$  such that:

$$\mathbb{Q}(\mathbf{c}_n^c) = O_{\mathbb{P}}(n^{\alpha-1/2}) \tag{H7}$$

where  $\mathbf{c}_n = \cup_{\mathbf{r} \in \mathcal{C}_n} \mathbf{r}$ ;

$$\widetilde{\mathcal{M}}(\Pi_n) = o(n) \quad \text{and} \quad \log(\widetilde{\Delta}_n(\Pi_n)) = o(n) \tag{H8}$$

where  $\Pi_n$  is a refinement of  $\{\mathcal{P}(\mathcal{C}_n(\mathbf{d}_n)) : \mathbf{d}_n \in \mathcal{S}^n\}$ ;

$$\forall n \in \mathbb{N}^*, \forall \mathbf{d}_n \in \mathcal{S}^n, \{\mathbf{r} \times \mathbb{R}, \mathbf{r} \in \mathcal{C}_n(\mathbf{d}_n)\} \subseteq \mathcal{B} \tag{H9}$$

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

Assume moreover that the estimator  $(\sigma_n^2)$  involved in Definition 2.1 is such that

$$|\sigma_n^2 - \sigma^2| = O_{\mathbb{P}}(n^{\alpha-1/2}).$$

Then

$$g_n = \sum_{\substack{A \in \mathcal{P}(\mathcal{C}_n) \\ \mathbb{Q}_n(A) > 0}} \mathbb{E}_n[Y | \mathbf{X} \in A] \mathbf{1}_A$$

is weakly consistent:

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

*Remark 6.* Many algorithms as CART generate rules as in (1.2) where each simple condition ( $\mathbf{X} \in c$ ) consists in some distinct coordinate of  $\mathbf{X}$  belonging

to an interval of  $\mathbb{R}$ . In Section 4 we use such rule generators for which (H9) is automatically satisfied. Indeed let  $\mathcal{H}$  be the set of all hyperrectangles of  $\mathbb{R}^d$ :

$$\mathcal{H} := \left\{ I_1 \times \dots \times I_d \subset \mathbb{R}^d : I_i \text{ is an interval of } \mathbb{R} \right\} \cup \{\emptyset\}.$$

Any set of rules  $\mathcal{C}_n$  such that  $\mathcal{C}_n \subseteq \mathcal{H}$ , which is always the case of the sets of rules considered in Section 4, fulfills (H9) since it is a VC class (see Wenocur and Dudley, 1981) and VC classes are  $\mathbb{Q}$ -Donsker (see for example Van der Vaart, 2000, Lemma 19.15 and comments).

### 3 Proof of Theorem 2.1

In order to prove the main theorem, we need some preliminary results based on the 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 usual notion of boundedness in probability for sequences of random variables is generalized because sequences of maps are 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. See (Van der Vaart, 2000, Chapter 18) for details.

**Definition 3.1.** (Van der Vaart, 2000, Chapter 18) A sequence  $(M_n)_{n \in \mathbb{N}}$  of maps defined on  $\Omega$  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  $K^\delta = \{y \in \mathbb{D} : d(y, K) < \delta\}$ .

*Remark 7.* If  $\mathbb{D} = \mathbb{R}$ ,  $(M_n)$  is asymptotically tight if and only if

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

The notation  $O_{\mathbb{P}^*}(1)$  stands for *asymptotically tight* instead of the usual  $O_{\mathbb{P}}(1)$  (*bounded in probability*).

For  $f : \mathcal{S} \rightarrow \mathbb{R}$  in  $\mathcal{L}^1(\mathbb{Q})$  we define  $v_n f := \sqrt{n}(\int f d\mathbb{Q}_n - \int f d\mathbb{Q})$  and consider the empirical process indexed by a set  $\mathcal{F}$  of such functions:  $\{v_n f : f \in \mathcal{F}\}$ .

**Definition 3.2.** (Van der Vaart, 2000, Section 19.2) A class of functions  $\mathcal{F}$  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 3.3.** 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.

If  $\mathcal{F}$  is a  $\mathbb{Q}$ -Donsker class of functions, then the empirical process  $((v_n f)_{f \in \mathcal{F}})_{n \in \mathbb{N}}$  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 Van der Vaart, 2000, Theorem 18.12). 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 8.* If  $\mathcal{B} \subseteq \mathcal{B}_S$  is a  $\mathbb{Q}$ -Donsker class of sets, where  $\mathcal{B}_S$  is the Borel set on  $S$ , 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 9.* 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 (non necessarily measurable) such that  $M_n = O_{\mathbb{P}^*}(1)$  and  $Z_n \leq a_n M_n$  for any  $n$ , then  $Z_n \xrightarrow[n \rightarrow +\infty]{\mathbb{P}} 0$ .

### 3.1 Empirical estimation of conditional expectations

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

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

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

**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*

$$\sup_{A \in \mathcal{B}_n} |\mathbb{E}_n[Y^i | (\mathbf{X}, Y) \in A] - \mathbb{E}[Y^i | (\mathbf{X}, Y) \in A]| = O_{\mathbb{P}^*}(n^{\alpha-1/2}), \quad (3.1)$$

and

$$\sup_{A \in \mathcal{B}_n} |\mathbb{V}_n[Y | (\mathbf{X}, Y) \in A] - \mathbb{V}[Y | (\mathbf{X}, Y) \in A]| = O_{\mathbb{P}^*}(n^{\alpha-1/2}). \quad (3.2)$$

Proofs of these results are postponed 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 sequence of partitions with shrinking cell diameters is a suitable covering. We define the diameter of a cell  $\mathbf{r}$  as  $\text{Diam}(\mathbf{r}) = \sup_{\mathbf{x} \in \mathbf{r}, \mathbf{x}' \in \mathbf{r}} \|\mathbf{x} - \mathbf{x}'\|$ , 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 (H3) with  $\alpha \in [0, \frac{1}{2})$  and such that

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

is a.s. a  $\mathbb{Q}$ -Donsker class. Suppose that  $\sigma^2$  is known,  $g^*$  is uniformly continuous and

$$\max_{\mathbf{r} \in \mathcal{P}_n} \text{Diam}(\mathbf{r}) = o_{\mathbb{P}}(1). \quad (3.3)$$

Then the sequence  $(\mathcal{P}_n)$  is  $\alpha$ -suitable.

*Proof.* Let us show that each cell is either significant or insignificant. Thanks to Condition (H3), Corollary 3.1 (Eq. (3.2)) and Remark 9,

$$\max_{\mathbf{r} \in \mathcal{P}_n} |\mathbb{V}_n(Y | \mathbf{X} \in \mathbf{r}) - \mathbb{V}(Y | \mathbf{X} \in \mathbf{r})| = O_{\mathbb{P}}(n^{\alpha-1/2}). \quad (3.4)$$

Moreover  $\mathbb{V}(Y | \mathbf{X} \in \mathbf{r}) = \mathbb{V}(g^*(\mathbf{X}) | \mathbf{X} \in \mathbf{r}) + \sigma^2$ . Thus, as the redundancy condition (H6) is automatically satisfied for cells of a partition, the desired result will follow if we check that  $\varepsilon_n = o_{\mathbb{P}}(1)$  with

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

From (3.4) we remark that

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

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

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

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

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

By uniform continuity, the condition (3.3) 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 (2.2), each cell which is not significant is insignificant and the corresponding covering sequence is  $\alpha$ -suitable.  $\square$

### 3.2 Estimation-approximation decomposition

Let  $\mathcal{G}_n$  denote the set of piecewise constant functions on the partition  $\mathcal{P}(\mathcal{C}_n) \cup \{\mathbf{c}_n^c\}$  such that  $\forall g \in \mathcal{G}_n, \forall \mathbf{x} \in \mathbb{R}^d, |g(\mathbf{x})| \leq L$  and  $\forall g \in \mathcal{G}_n, \forall \mathbf{x} \in \mathbf{c}_n^c, g(\mathbf{x}) = 0$ .

The excess risk  $\ell(g^*, g_n)$  can be decomposed into two terms mimicking Lemma 10.1 of Györfi et al. (2006). First notice that under the conditions of Theorem 2.1 on  $g_n$  we have

$$\frac{1}{n} \sum_{i=1}^n (g_n(\mathbf{X}_i) - Y_i)^2 \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}_i) \leq \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}_i), \quad \forall g \in \mathcal{G}_n. \quad (3.5)$$

Now, since moreover  $g_n \in \mathcal{G}_n$ ,

$$\begin{aligned} \ell(g^*, g_n) &= \mathbb{E}[(g_n(\mathbf{X}) - g^*(\mathbf{X}))^2] \\ &= \mathbb{E}[(g_n(\mathbf{X}) - Y)^2] - \mathbb{E}[(g^*(\mathbf{X}) - Y)^2] \\ &= \mathbb{E}[(g_n(\mathbf{X}) - Y)^2] - \inf_{g \in \mathcal{G}_n} \mathbb{E}[(g(\mathbf{X}) - Y)^2] \\ &\quad + \inf_{g \in \mathcal{G}_n} \mathbb{E}[(g(\mathbf{X}) - Y)^2] - \mathbb{E}[(g^*(\mathbf{X}) - Y)^2] \\ &= \sup_{g \in \mathcal{G}_n} \left\{ \mathbb{E}[(g_n(\mathbf{X}) - Y)^2] - \frac{1}{n} \sum_{i=1}^n (g_n(\mathbf{X}_i) - Y_i)^2 \right. \\ &\quad + \frac{1}{n} \sum_{i=1}^n (g_n(\mathbf{X}_i) - Y_i)^2 - \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 \\ &\quad \left. + \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 - \mathbb{E}[(g(\mathbf{X}) - Y)^2] \right\} \\ &\quad + \inf_{g \in \mathcal{G}_n} \mathbb{E}[(g(\mathbf{X}) - g^*(\mathbf{X}))^2] \\ &\leq 2 \sup_{g \in \mathcal{G}_n} \left| \mathbb{E}[(g(\mathbf{X}) - Y)^2] - \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 \right| \\ &\quad + \underbrace{\frac{1}{n} \sum_{i=1}^n (g_n(\mathbf{X}_i) - Y_i)^2 \mathbf{1}_{\mathbf{c}_n^c}(\mathbf{X}_i)}_{\leq 4L^2 \mathbb{Q}_n(\mathbf{c}_n^c)} \\ &\quad + \inf_{g \in \mathcal{G}_n} \mathbb{E}[(g(\mathbf{X}) - g^*(\mathbf{X}))^2]. \end{aligned}$$

Up to the term  $4L^2 \mathbb{Q}_n(\mathbf{c}_n^c)$  which converges to zero by assumption, this is the standard decomposition of the risk into the estimation error and the approximation error. To prove the theorem it is sufficient to prove that

$$\inf_{g \in \mathcal{G}_n} \mathbb{E}[(g(\mathbf{X}) - g^*(\mathbf{X}))^2] = o_{\mathbb{P}}(1) \quad (3.6)$$

and

$$\sup_{g \in \mathcal{G}_n} \left| \mathbb{E}[(g(\mathbf{X}) - Y)^2] - \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 \right| = o_{\mathbb{P}}(1). \quad (3.7)$$

### 3.3 Approximation error

Let us define  $\mathcal{C}_n^{i \setminus s} = \{\mathbf{r} \setminus \mathbf{c}_n^s : \mathbf{r} \in \mathcal{C}_n^i\}$  where  $\mathbf{c}_n^s = \cup_{\mathbf{r} \in \mathcal{C}_n^s} \mathbf{r}$ . Notice that  $\mathcal{P}(\mathcal{C}_n^{i \setminus s}) = \{A \setminus \mathbf{c}_n^s : A \in \mathcal{P}(\mathcal{C}_n^i)\}$ . Indeed, by Proposition 1.1:

$$\begin{aligned}
\mathcal{P}(\mathcal{C}_n^{i \setminus s}) &= \left\{ \bigcap_{\mathbf{r} \in \tilde{\mathcal{C}}} \mathbf{r} \setminus \bigcup_{\mathbf{r} \in \mathcal{C}_n^{i \setminus s} \setminus \tilde{\mathcal{C}}} \mathbf{r} : \tilde{\mathcal{C}} \subseteq \mathcal{C}_n^{i \setminus s} \right\} \\
&= \left\{ \bigcap_{\mathbf{r} \in \tilde{\mathcal{C}}^i} (\mathbf{r} \setminus \mathbf{c}_n^s) \setminus \bigcup_{\mathbf{r} \in \mathcal{C}_n^i \setminus \tilde{\mathcal{C}}^i} (\mathbf{r} \setminus \mathbf{c}_n^s) : \tilde{\mathcal{C}}^i \subseteq \mathcal{C}_n^i \right\} \\
&= \left\{ \left( \bigcap_{\mathbf{r} \in \tilde{\mathcal{C}}^i} \mathbf{r} \setminus \bigcup_{\mathbf{r} \in \mathcal{C}_n^i \setminus \tilde{\mathcal{C}}^i} \mathbf{r} \right) \setminus \mathbf{c}_n^s : \tilde{\mathcal{C}}^i \subseteq \mathcal{C}_n^i \right\} \\
&= \left\{ A \setminus \mathbf{c}_n^s : A \in \mathcal{P}(\mathcal{C}_n^i) \right\}. \tag{3.8}
\end{aligned}$$

So that, by Proposition 1.1 again,  $\mathcal{P}(\mathcal{C}_n)$  is a partition of  $\mathbf{c}_n$  finer than  $\mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s})$ . Hence, with

$$\tilde{g}_n = \sum_{\substack{A \in \mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s}) \\ \mathbb{Q}(A) > 0}} \mathbb{E}[Y | \mathbf{X} \in A] \mathbf{1}_A,$$

we have  $\tilde{g}_n \in \mathcal{G}_n$  and

$$\begin{aligned}
\inf_{g \in \mathcal{G}_n} \sqrt{\mathbb{E} \left[ (g(\mathbf{X}) - g^*(\mathbf{X}))^2 \right]} &\leq \sqrt{\mathbb{E} \left[ (\tilde{g}_n(\mathbf{X}) - g^*(\mathbf{X}))^2 \right]} \\
&\leq \sqrt{\mathbb{E} \left[ (\tilde{g}_n(\mathbf{X}) - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \right]} \\
&\quad + \underbrace{\sqrt{\mathbb{E} \left[ g^*(\mathbf{X})^2 \mathbf{1}_{\mathbf{c}_n^c}(\mathbf{X}) \right]}}_{\leq L \sqrt{\mathbb{Q}(\mathbf{c}_n^c)}}.
\end{aligned}$$

Thus, to prove (3.6), it suffices to show that  $\mathbb{W}_n = o_{\mathbb{P}}(1)$  where

$$\mathbb{W}_n := \mathbb{E} \left[ (\tilde{g}_n(\mathbf{X}) - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \right].$$

We will repeatedly make use of the following equalities:

$$\begin{cases} \mathbb{E}[g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) | \mathbf{X} \in \mathbf{r}] = \mathbb{E}[g^*(\mathbf{X}) | \mathbf{X} \in \mathbf{r}] = \mathbb{E}[Y | \mathbf{X} \in \mathbf{r}], \\ \mathbb{V}(g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) | \mathbf{X} \in \mathbf{r}) = \mathbb{V}(g^*(\mathbf{X}) | \mathbf{X} \in \mathbf{r}), \quad \forall \mathbf{r} \subseteq \mathbf{c}_n \end{cases} \tag{3.9}$$

and

$$\mathbb{V}(Y | \mathbf{X} \in \mathbf{r}) = \mathbb{V}(g^*(\mathbf{X}) | \mathbf{X} \in \mathbf{r}) + \sigma^2, \quad \forall \mathbf{r} \subseteq \mathbb{R}^d. \tag{3.10}$$

Let us first remark that

$$\begin{aligned}
\mathbb{W}_n &= \mathbb{E} \left[ \left( \sum_{A \in \mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s})} \mathbb{E}[Y \mid \mathbf{X} \in A] \mathbf{1}_A(\mathbf{X}) - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \right] \\
&= \sum_{A' \in \mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s}) \cup \{\mathbf{c}_n^c\}} \mathbb{E} \left[ \left( \sum_{A \in \mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s})} \mathbb{E}[g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \mid \mathbf{X} \in A] \mathbf{1}_A(\mathbf{X}) \right. \right. \\
&\quad \left. \left. - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \mathbf{1}_{A'}(\mathbf{X}) \right] \\
&= \sum_{A' \in \mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s}) \cup \{\mathbf{c}_n^c\}} \mathbb{E} \left[ \left( \mathbb{E}[g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \mid \mathbf{X} \in A'] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \mathbf{1}_{A'}(\mathbf{X}) \right] \\
&= \sum_{A' \in \mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s}) \cup \{\mathbf{c}_n^c\}} \mathbb{E} \left[ \left( \mathbb{E}[g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \mid \mathbf{X} \in A'] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \mid \mathbf{X} \in A' \right] \\
&\quad \times \mathbb{P}(\mathbf{X} \in A') \tag{3.11}
\end{aligned}$$

which shows that  $\mathbb{W}_n$  is a within-group variance for the variable  $g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X})$  and the groups  $\mathcal{P}(\mathcal{C}_n^s) \cup \mathcal{P}(\mathcal{C}_n^{i \setminus s}) \cup \{\mathbf{c}_n^c\}$ .

According to the decomposition

$$\begin{aligned}
\mathbb{W}_n &= \underbrace{\sum_{A \in \mathcal{P}(\mathcal{C}_n^s)} \mathbb{E} \left[ \left( \mathbb{E}[Y \mid \mathbf{X} \in A] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \mid \mathbf{X} \in A \right] \mathbb{P}(\mathbf{X} \in A)}_{\mathbb{W}_n^s} \\
&\quad + \underbrace{\sum_{A \in \mathcal{P}(\mathcal{C}_n^{i \setminus s})} \mathbb{E} \left[ \left( \mathbb{E}[Y \mid \mathbf{X} \in A] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \mid \mathbf{X} \in A \right] \mathbb{P}(\mathbf{X} \in A)}_{\mathbb{W}_n^{i \setminus s}} \\
&\quad + \underbrace{\mathbb{E} \left[ \left( \mathbb{E}[g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \mid \mathbf{X} \in \mathbf{c}_n^c] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) \right)^2 \mid \mathbf{X} \in \mathbf{c}_n^c \right] \mathbb{P}(\mathbf{X} \in \mathbf{c}_n^c)}_0,
\end{aligned}$$

it is sufficient to prove that  $\mathbb{W}_n^s \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0$  and  $\mathbb{W}_n^{i \setminus s} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0$ .

To deal with  $\mathbb{W}_n^s$ , we start from the decomposition of the total variance into the within-group and the between-group variances:

$$\begin{aligned}
\mathbb{W}_n^s &= \mathbb{E}[(g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}) - \mathbb{E}[g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X})])^2 \mathbf{1}_{\mathbf{c}_n^s}(\mathbf{X})] - \mathbb{B}_n^s \\
&\leq \mathbb{V}(g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X})) - \mathbb{B}_n^s \tag{3.12}
\end{aligned}$$

where

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

The between group variance  $\mathbb{B}_n^s$  will be lower estimated by a function of  $\mathbb{W}_n^s$  from which a lower bound of  $\mathbb{W}_n^s$  will follow. The key point of this lower bound is to use the definition of  $\mathcal{C}_n^s$  in (H5). To make the terms controlled by this significant condition appear, we lower bound  $\mathbb{B}_n^s$  by a sum over the elements of  $\mathcal{C}_n^s$  instead of elements of  $\mathcal{P}(\mathcal{C}_n^s)$  and the expectations is replaced by empirical expectations. After applying the inequality provided by the significant condition we'll have to do the reverse operation to make  $\mathbb{W}_n^s$  appear again.

Let us first lower bound  $\mathbb{B}_n^s$  by a sum over the elements of  $\mathcal{C}_n^s$  instead of elements of  $\mathcal{P}(\mathcal{C}_n^s)$  and replace the expectations by empirical expectations:

$$\begin{aligned}
\mathbb{B}_n^s &= \sum_{\mathbf{r} \in \mathcal{C}_n^s} \sum_{A \in \mathcal{P}_{\mathcal{C}_n^s}(\mathbf{r})} \frac{1}{\#\varphi_{\mathcal{C}_n^s}(A)} (\mathbb{E}[Y | \mathbf{X} \in A] - \mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{x})}])^2 \mathbb{P}(\mathbf{X} \in A) \\
&\geq \sum_{\mathbf{r} \in \mathcal{C}_n^s} \frac{1}{M(\mathcal{C}_n^s, \mathbf{r})} \sum_{A \in \mathcal{P}_{\mathcal{C}_n^s}(\mathbf{r})} (\mathbb{E}[Y | \mathbf{X} \in A] - \mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{x})}])^2 \mathbb{P}(\mathbf{X} \in A) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} \sum_{A \in \mathcal{P}_{\mathcal{C}_n^s}(\mathbf{r})} (\mathbb{E}[Y | \mathbf{X} \in A] - \mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{x})}])^2 \mathbb{P}(\mathbf{X} \in A | \mathbf{X} \in \mathbf{r}) \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} \left( \sum_{A \in \mathcal{P}_{\mathcal{C}_n^s}(\mathbf{r})} \mathbb{E}[Y | \mathbf{X} \in A] \mathbb{P}(\mathbf{X} \in A | \mathbf{X} \in \mathbf{r}) - \mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{x})}] \right)^2 \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&= \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} \left( \mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - \mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{x})}] \right)^2 \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&\geq \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} (V_{n, \mathbf{r} \times \mathbb{R}}^2 - \Delta_n) \mathbb{P}(\mathbf{X} \in \mathbf{r}) \tag{3.13}
\end{aligned}$$

where we applied Jensen's inequality for the third to last inequality and where

$$\Delta_n := \sup_{A \in \mathcal{B}_n} \{V_{n,A}^2 - U_{n,A}^2\}$$

with for any  $A \in \mathcal{B}$ ,

$$\begin{aligned}
U_{n,A} &:= \mathbb{E}[Y | (\mathbf{X}, Y) \in A] - \mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{X})}] \\
V_{n,A} &:= \mathbb{E}_n[Y | (\mathbf{X}, Y) \in A] - \mathbb{E}_n[Y]
\end{aligned}$$

and  $\mathcal{B}_n = \{A \in \mathcal{B} \text{ s.t. } \mathbb{Q}_n(A) > n^{-\alpha}\}$ .

Continuing (3.13) with the definition of  $\mathcal{C}_n^s$  in (H5) in mind,

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

Since  $\mathbb{W}_n^s = \sum_{A \in \mathcal{P}_{\mathcal{C}_n}(\mathcal{C}_n^s)} \mathbb{V}(g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X}) | \mathbf{X} \in A) \mathbb{P}(\mathbf{X} \in A)$ , this last term can be lower bounded by  $\mathbb{W}_n^s$  if the empirical variances are replaced by variances and the sum over  $\mathcal{C}_n^s$  by a sum over  $\mathcal{P}(\mathcal{C}_n^s)$ . Let us then define

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

and write

$$\begin{aligned} \mathbb{B}_n^s &\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) - \beta_n^{-2}|\sigma_n^2 - \sigma^2| - \Delta_n) \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\ &= \frac{1}{M(\mathcal{C}_n^s)} \sum_{\mathbf{r} \in \mathcal{C}_n^s} (\beta_n^{-2}(\mathbb{V}(g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X}) | \mathbf{X} \in \mathbf{r}) - \Delta'_n) - \beta_n^{-2}|\sigma_n^2 - \sigma^2| - \Delta_n) \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\ &\hspace{15em} \text{by (3.9) and (3.10)} \\ &= \frac{\beta_n^{-2}}{M(\mathcal{C}_n^s)} \times \sum_{\mathbf{r} \in \mathcal{C}_n^s} \left( \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X}))^2 | \mathbf{X} \in \mathbf{r} \right] \right. \\ &\quad \left. - (\Delta'_n + |\sigma_n^2 - \sigma^2| + \beta_n^2 \Delta_n) \right] \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\ &\geq \frac{\beta_n^{-2}}{M(\mathcal{C}_n^s)} \times \sum_{\mathbf{r} \in \mathcal{C}_n^s} \sum_{A \in \mathcal{P}_{\mathcal{C}_n^s}(\mathbf{r})} \left( \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X}))^2 | \mathbf{X} \in A \right] \right) \mathbb{P}(\mathbf{X} \in A) \\ &\quad - (\beta_n^{-2} \Delta'_n + \beta_n^{-2} |\sigma_n^2 - \sigma^2| + \Delta_n) \quad \text{since } \sum_{\mathbf{r} \in \mathcal{C}_n^s} \mathbb{P}(\mathbf{X} \in \mathbf{r}) \leq M(\mathcal{C}_n^s) \\ &\geq \frac{\beta_n^{-2}}{M(\mathcal{C}_n^s)} \times \sum_{\mathbf{r} \in \mathcal{C}_n^s} \sum_{A \in \mathcal{P}_{\mathcal{C}_n^s}(\mathbf{r})} \left( \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A] - g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X}))^2 | \mathbf{X} \in A \right] \right) \mathbb{P}(\mathbf{X} \in A) \\ &\quad - (\beta_n^{-2} \Delta'_n + \beta_n^{-2} |\sigma_n^2 - \sigma^2| + \Delta_n) \\ &\geq \beta_n^{-2} \frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)} \times \sum_{A \in \mathcal{P}(\mathcal{C}_n^s)} \left( \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A] - g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X}))^2 | \mathbf{X} \in A \right] \right) \mathbb{P}(\mathbf{X} \in A) \\ &\quad - (\beta_n^{-2} \Delta'_n + \beta_n^{-2} |\sigma_n^2 - \sigma^2| + \Delta_n) \\ &= \beta_n^{-2} \frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)} \mathbb{W}_n^s - (\beta_n^{-2} \Delta'_n + \beta_n^{-2} |\sigma_n^2 - \sigma^2| + \Delta_n). \end{aligned}$$

Together with (3.12), this yields

$$\mathbb{W}_n^s \leq \frac{\mathbb{V}(g^*(\mathbf{X})\mathbf{1}_{\mathcal{C}_n}(\mathbf{X})) + \beta_n^{-2} \Delta'_n + \beta_n^{-2} |\sigma_n^2 - \sigma^2| + \Delta_n}{1 + \beta_n^{-2} \frac{m(\mathcal{C}_n^s)}{M(\mathcal{C}_n^s)}}.$$

Under (H3) and (H9), Corollary 3.1 applies and we obtain

$$\begin{aligned}
\Delta_n &= \sup_{A \in \mathcal{B}_n} \{(V_{n,A} - U_{n,A})(V_{n,A} + U_{n,A})\} \\
&\leq \sup_{A \in \mathcal{B}_n} \left\{ |\mathbb{E}_n[Y | (\mathbf{X}, Y) \in A] - \mathbb{E}[Y | (\mathbf{X}, Y) \in A]| \right. \\
&\quad \left. + |\mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{X})}] - \mathbb{E}_n[Y]| \right\} \times 4L \\
&\leq 4L \times \left( \sup_{A \in \mathcal{B}_n} |\mathbb{E}_n[Y | (\mathbf{X}, Y) \in A] - \mathbb{E}[Y | (\mathbf{X}, Y) \in A]| \right. \\
&\quad \left. + |\mathbb{E}[g^*(\mathbf{X})\mathbf{1}_{\mathbf{c}_n(\mathbf{X})}] - \mathbb{E}[g^*(\mathbf{X})]| + \underbrace{|\mathbb{E}[g^*(\mathbf{X})] - \mathbb{E}_n[Y]|}_{\mathbb{E}[Y]} \right) \\
&\leq 4L \times (O_{\mathbb{P}^*}(n^{\alpha-1/2}) + L\mathbb{Q}(\mathbf{c}_n^c) + O_{\mathbb{P}}(n^{-1/2})) \\
&= O_{\mathbb{P}^*}(n^{\alpha-1/2})
\end{aligned}$$

since  $\mathbb{Q}(\mathbf{c}_n^c) = O_{\mathbb{P}^*}(n^{\alpha-1/2})$ . Corollary 3.1 also yields

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

Since it is moreover assumed that  $|\sigma_n^2 - \sigma^2| = O_{\mathbb{P}}(n^{\alpha-\frac{1}{2}})$ , (2.3) of Condition (H6) and Remark 9 lead to

$$\mathbb{W}_n^s \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0.$$

To deal with  $\mathbb{W}_n^{i \setminus s}$ , remember (3.8) and write

$$\begin{aligned}
\mathbb{W}_n^{i \setminus s} &= \sum_{A \in \mathcal{P}(\mathcal{C}_n^{i \setminus s})} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_A(\mathbf{X}) \right] \\
&= \sum_{A \in \mathcal{P}(\mathcal{C}_n^i)} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A \setminus \mathbf{c}_n^s] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_{A \setminus \mathbf{c}_n^s}(\mathbf{X}) \right] \\
&\leq \sum_{A \in \mathcal{P}(\mathcal{C}_n^i)} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_{A \setminus \mathbf{c}_n^s}(\mathbf{X}) \right] \\
&\leq \sum_{A \in \mathcal{P}(\mathcal{C}_n^i)} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_A(\mathbf{X}) \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^i}(\mathbf{r})} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in A] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_A(\mathbf{X}) \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^i}(\mathbf{r})} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_A(\mathbf{X}) \right] \\
&\leq \frac{1}{m(\mathcal{C}_n^i)} \sum_{\mathbf{r} \in \mathcal{C}_n^i} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \mathbf{1}_{\mathbf{r}}(\mathbf{X}) \right] \\
&= \frac{1}{m(\mathcal{C}_n^i)} \sum_{\mathbf{r} \in \mathcal{C}_n^i} \mathbb{E} \left[ (\mathbb{E}[Y | \mathbf{X} \in \mathbf{r}] - g^*(\mathbf{X}) \mathbf{1}_{\mathbf{c}_n}(\mathbf{X}))^2 \middle| \mathbf{X} \in \mathbf{r} \right] \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&= \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}) \quad \text{by (3.9) and (3.10)} \\
&\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_n^2 + |\sigma_n^2 - \sigma^2| + \Delta'_n) \mathbb{P}(\mathbf{X} \in \mathbf{r}) \\
&\leq \frac{1}{m(\mathcal{C}_n^i)} \sum_{\mathbf{r} \in \mathcal{C}_n^i} (\varepsilon_n^2 + |\sigma_n^2 - \sigma^2| + \Delta'_n) \mathbb{P}(\mathbf{X} \in \mathbf{r}) \quad \text{under (H5)} \\
&\leq \frac{M(\mathcal{C}_n^i)}{m(\mathcal{C}_n^i)} (\varepsilon_n^2 + |\sigma_n^2 - \sigma^2| + \Delta'_n) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0
\end{aligned}$$

under (2.4) of Condition (H6) and with Remark 9.

### 3.4 Estimation error

The proof of (3.7) is inspired by (Györfi et al., 2006, Theorem 13.1) and its proof. It is not needed here to truncate the functions in  $\mathcal{G}_n$  nor  $Y$  since they are all bounded by assumption. Thus the assumptions (H8) do not need to involve the truncation constant.

Recall that  $\mathcal{G}_n$  is the set of piecewise constant functions with values in  $[-L, L]$  on the elements of the partition  $\mathcal{P}(\mathcal{C}_n) \cup \{\mathbf{c}_n^c\}$  such that  $\forall g \in \mathcal{G}_n, \forall \mathbf{x} \in \mathbf{c}_n^c, g(\mathbf{x}) =$



0. Then  $\mathcal{G}_n$  is a subset of

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

where  $\Pi_n$  is a refinement of  $\{\mathcal{P}(\mathcal{C}_n(\mathbf{d}_n)) : \mathbf{d}_n \in \mathcal{S}^n\}$  (see Definition 2.3) and  $\mathcal{G}_c$  is the set of constant functions  $\mathbb{R}^d \rightarrow [-L, L]$ . Then

$$\begin{aligned} \sup_{g \in \mathcal{G}_n} \left| \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 - \mathbb{E}[(g(\mathbf{X}) - Y)^2] \right| \\ \leq \sup_{g \in \mathcal{G}_c \circ \Pi_n} \left| \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 - \mathbb{E}[(g(\mathbf{X}) - Y)^2] \right|, \end{aligned}$$

According to (Györfi et al., 2006, Theorem 9.1 and Problem 10.4) we have,

$$\begin{aligned} \mathbb{P} \left\{ \sup_{g \in \mathcal{G}_c \circ \Pi_n} \left| \frac{1}{n} \sum_{i=1}^n (g(\mathbf{X}_i) - Y_i)^2 - \mathbb{E}[(g(\mathbf{X}) - Y)^2] \right| > \varepsilon \right\} \\ \leq 8\mathbb{E} \left[ \mathcal{N}_1 \left( \frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, \mathbf{X}_1^n \right) \right] \exp \left\{ \frac{-n\varepsilon^2}{128.(4L^2)^2} \right\}, \quad (3.14) \end{aligned}$$

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  $\varepsilon$ -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) (it can be checked to hold in the current situation, in particular for a family of finite collections of disjoint sets instead of a family of partitions of  $\mathbb{R}^d$ ) we have

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

According to (Györfi et al., 2006, Lemma 9.2) for any set of functions  $\mathcal{G}$  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  $L_1$   $\varepsilon$ -packing of  $\mathcal{G}$  on  $z_1^m$ . See (Györfi et al., 2006, Definition 9.4 (c)).

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

$$\sup_{z_1, \dots, z_m \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}, m \leq n} \mathcal{M}_1(\varepsilon, \mathcal{G}_c, z_1^m) \leq \left\lceil \frac{2L}{\varepsilon} \right\rceil + 1.$$

Finally,

$$\sup_{z_1, \dots, z_m \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}, m \leq n} \mathcal{N}_1\left(\frac{\varepsilon}{32L}, \mathcal{G}_c \circ \Pi_n, z_1^m\right) \leq \tilde{\Delta}(\Pi_n) \left( \left\lceil \frac{64L^2}{\varepsilon} \right\rceil + 1 \right)^{\tilde{\mathcal{M}}(\Pi_n)}. \quad (3.15)$$

According to (3.14) and (3.15) we have:

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

and since

$$\begin{aligned} & 8 \tilde{\Delta}_n(\Pi_n) \left( \left\lceil \frac{64L^2}{\varepsilon} \right\rceil + 1 \right)^{\tilde{\mathcal{M}}(\Pi_n)} \exp \left( -\frac{n\varepsilon^2}{2048L^4} \right) \\ & = 8 \exp \left( -\frac{n}{L^4} \left( \frac{\varepsilon^2}{2048} - \frac{\log \tilde{\Delta}_n(\Pi_n)L^4}{n} - \frac{\tilde{\mathcal{M}}(\Pi_n)L^4 \log \left( \left\lceil \frac{64L^2}{\varepsilon} \right\rceil + 1 \right)}{n} \right) \right), \end{aligned}$$

this concludes the proof of (3.7) and of Theorem 2.1.

## 4 Illustrations

In this section we propose a simple algorithm to generate data-dependent coverings using either Random Forests (RF, Breiman (2001)) or Gradient Boosted tree (GB, Friedman (2001)) or Stochastic Gradient Boosting tree (SGB, Friedman (2002)) as rule generator. The interest is double; First, it exhibits examples of data-dependent quasi-coverings that are likely to be suitable as in Definition 2.1. Second, we observe that the interpretability indices are small when applying our algorithm predicting real data with significant rules. Note that any rules considered here are based on intervals of  $\mathbb{R}$  as in Remark 6 and then their length is automatically smaller than or equal to  $d$ .

## 4.1 Covering Algorithm

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

1. The algorithm generates an ensemble of trees with a given method among RF, GB or SGB with a maximal size by tree, *tree\_size*, and a maximal number of generated rules, *max\_rules*.
2. For a chosen  $\alpha \in (0, 1/2)$ , the algorithm sets  $\beta_n = n^{\alpha/2-1/4}$  and  $\varepsilon_n = \beta_n s_n$ , where  $s_n$  is the empirical standard deviation of  $y_1, \dots, y_n$ . It keeps all nodes and leaves of all trees generated by the rule generator that have a length lower than or equal to  $l_{max} \in \{1, \dots, d\}$  and that fulfill (H3).
3. This set of rules is divided into two sets: The set of significant rules  $S_n$  (*i.e.*, rules  $\mathbf{r}$  that fulfill  $\beta_n |\mathbb{E}_n[Y|\mathbf{X} \in \mathbf{r}] - \mathbb{E}_n[Y]| \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma_n^2)_+}$ ) and  $I_n$  the set of insignificant rules (*i.e.*, rules  $\mathbf{r}$  that are not in  $S_n$  and fulfill  $\varepsilon_n \geq \sqrt{(\mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r}) - \sigma_n^2)_+}$ ).
4. The algorithm selects a minimal set of rules  $\mathcal{C}_n$  using Algorithm 1 in Appendix. If, at any step, the set of selected rules forms a covering the selection process is stopped. A rule is added to the current selected set of rules if and only if it has at least a rate  $1 - \gamma \in (0, 1)$  of points not covered by the current set of rules<sup>2</sup>. The set  $S_n$  is browsed first, sorted by empirical coverage rate and  $I_n$  is browsed next, if necessary, sorted by increasing empirical variance.

As explained in Definition 2.1 the set  $\mathcal{C}_n$  is the union of two sets of rules: The significant set  $\mathcal{C}_n^s$  and the insignificant set  $\mathcal{C}_n^i$ . In practice the set  $\mathcal{C}_n^s$  is the most interesting one because it identifies the rules where the conditional mean is prominent.

The datasets and the code of the illustrations are available on GitHub. The code mixed Python and R languages.

## 4.2 Datasets

We consider seven public datasets presented in Table 1. These datasets have different dimensions and number of observations. For the dataset *Student* we have removed variables  $G1$  and  $G2$  which are the first and the second grade respectively because the target attribute  $G3$  has a strong correlation with attributes  $G2$  and  $G1$ . In Cortez and Silva (2008) authors argue that, despite the complexity of the regression model, the prediction of  $G3$  without  $G2$  and  $G1$  is much more useful in practice.

<sup>2</sup>This step was already described in Remark 4 in order to fulfill the redundancy condition (H6).

Name	$(n \times d)$	Short description
Diabetes	$443 \times 10$	Prediction of quantitative measure of disease progression one year after baseline (Efron et al., 2004).
Prostate	$97 \times 8$	Prediction of the level of prostate-specific antigen based on clinical measures in men who were about to receive a radical prostatectomy (Hastie et al., 2001).
Ozone	$330 \times 9$	Prediction of atmospheric ozone concentration from daily meteorological measurements (Hastie et al., 2001).
Machine	$209 \times 8$	Prediction of published relative performance (Dua and Graff, 2017).
MPG	$398 \times 8$	Prediction of city-cycle fuel consumption in miles per gallon (Dua and Graff, 2017).
Boston	$506 \times 13$	Prediction of the median price of neighborhoods, (Harrison Jr and Rubinfeld, 1978).
Student	$649 \times 32$	Prediction of the final grade of student based on attributes collected by reports and questionnaires (Cortez and Silva, 2008).

Table 1: Brief description of the public regression datasets used.

### 4.3 Execution

For each dataset we realize 20 executions. For each execution the data are randomly split into a training set and a test set, with ratios of 70% and 30%, respectively. As a benchmark of accuracy we consider RF (Breiman, 2001) and as a benchmark of interpretability we consider the CART algorithm (Breiman et al., 1984). We compare Covering Algorithm with RF as rule generator (CA RF), GB trees as rule generator (CA GB), SGB trees as rule generator (CA SGB) with RuleFit (Friedman and Popescu, 2008) (with rules only), Node harvest (Meinshausen, 2010) (Nh) and SIRUS (Bénard et al., 2020). These algorithms are the main existing interpretable rule-based algorithms for regression settings that are based on tree ensembles.

RuleFit is a very accurate rule-based algorithm. First it generates a list of rules by considering all nodes and leaves of a boosted tree ensemble ISLE (Friedman et al., 2003). Then rules are used as features in a sparse linear regression model obtained by Lasso (Tibshirani, 1996).

Node harvest uses also a tree ensemble as rule generator. The algorithm

considers all nodes and leaves of RF as rules and solves a linear quadratic problem to fit a weight to each rule. Hence, the prediction is a convex combination of rules.

SIRUS (Stable and Interpretable RULe Set) has been designed to be a stable predictive algorithm. It uses a modified RF to generate a large number of rules as nodes and leaves. Then it selects the rules with a rate of occurrence greater than a tuning parameter  $p_0$ . To ensure a large number of occurrences on rules the features are discretized.

The maximal number of leaves of CART is set at  $tree\_size = 10$  and the number of tree for RF is set at  $nb\_tree = 1000$ . For CA and RuleFit we set similarly the maximal number of rules generated by the rule generator  $max\_rules = 4000$  and the size of a tree  $tree\_size = 8$ . For CA, Node harvest and SIRUS we set identically the maximum rule length  $l\_max = 3$ . For CA we also set  $\alpha = 1/2 - 1/100$  and  $\gamma = 0.90$ . In these applications the real value of  $\sigma^2$  is unknown. We estimate it by  $\sigma_n^2$  the minimal variance of the generated rules fulfilling the covering condition (H3). We have no guarantee that this estimator is good enough (see Remark 3). For SIRUS, the hyperparameter  $p_0$  is estimated by 10-fold cross-validation and the maximal number of rules is set at 25. The parameters setting is summarized in Table 2.

Algorithm	Parameters
CART	$tree\_size = 10$ .
RF	$nb\_tree = 1000$ .
RuleFit	$tree\_size = 8$ , $max\_rules = 4000$ .
Node harvest (Nh)	$l\_max = 3$ .
SIRUS	$l\_max = 3$ , $max\_selected\_rules = 25$ , $ncv = 10$ .
CA	$max\_rules = 4000$ , $tree\_size = 8$ , $alpha = 1/2 - 1/100$ , $gamma = 0.90$ , $l\_max = 3$ .

Table 2: Algorithm parameter settings.

#### 4.4 Results

Experimental results are gathered in Table 3. For each dataset and algorithm we present an average over the 20 executions of the number of selected rules (Nb rules), the empirical coverage rate on the training set (Cov), the interpretability

index (1.4) (Int) and the mean squared error on the test set divided by the empirical variance of the target (MSE).

These results emphasize that suitable data-dependent quasi-coverings (see Definition 2.1) are very efficient to generate an interpretable rule-based model, see also Section 4.5. They generate very simple models with an interpretability index (1.4) much lower than the other algorithms. Moreover the sets of selected rules  $\mathcal{C}_n$  form in most cases a covering of the training set. CA achieves a good interpretability-accuracy tradeoff with emphasis on interpretability except for the dataset *Student* that might be too complex to be interpretable. SIRUS appears as a good challenger achieving a more balanced tradeoff. Finally, these results show that the choice of the rule generator for the CA has an impact on the interpretability index and on the accuracy.

## 4.5 Examples of interpretation

Any model generated by a CA can be summarized in a table of rules. We present one selected set of rules by the CA RF in Table 4 for the dataset *Ozone*. The measure  $\Delta_n$  is the empirical mean deviation ratio of the prediction from the mean of  $Y$ .

Table 4 is a description of the model. Together with Table 5 we are able to translate in natural language the significant rules in Table 4, accordingly with the definition of interpretability of Biran and Cotton (2017).

- Rule *R1* suggests that if the *temp* variable is reasonably low then the daily maximum of the hourly average ozone concentrations in Upland tends to be low.
- Rule *R2* suggests that if the *ibt* variable is not too high then the target tends to be low.
- Rules *R3* and *R4* suggests that CA has detected an interaction between high *ibt* and the *humidity* variable. Indeed, high value of *ibt* and high *humidity* indicate a high hourly average ozone concentrations and high value of *ibt* and low *humidity* indicate a low hourly average ozone concentrations.

## 4.6 Comments

In Figure 5 and Figure 6 we present boxplots of the MSE over the 20 executions for each algorithm for the datasets *Machine* and *Ozone*. These two datasets has been chosen to show that CA may have very variable accuracy (Figure 5) as well as stable and good accuracy (Figure 6) for two datasets with the same dimension.

One reason might be that rule generators of CA are random generators of rules that might not capture the good features at each run. In particular we identified the problem of the predictions of the empty cells of the partition generated by the quasi-coverings. Then by convention  $g_n(\mathbf{x}) = 0$  for  $\mathbf{x} \in A$

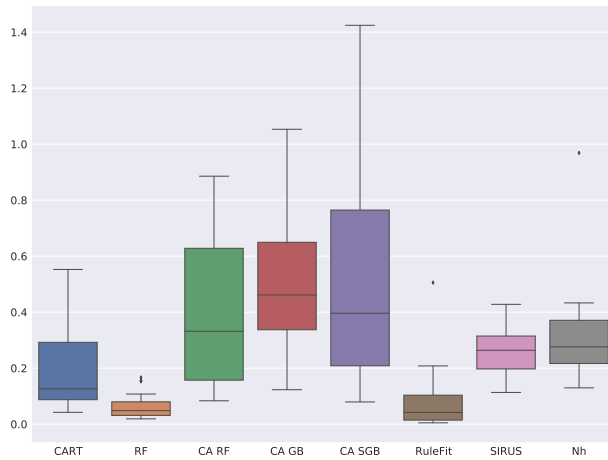


Figure 5: Boxplots of the ratio mean square error/empirical variance of the target (MSE) over 20 executions of algorithms for the *Machine* dataset.

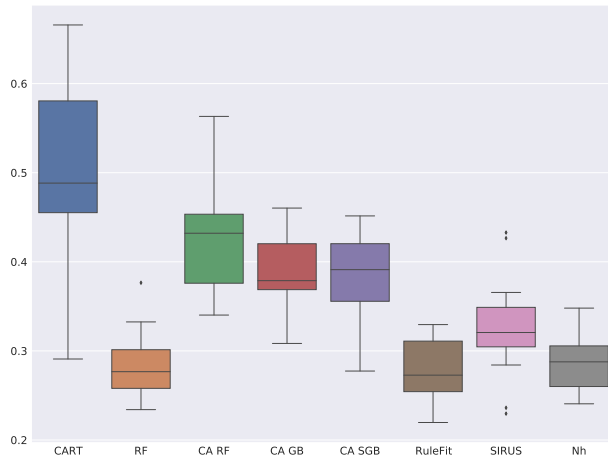


Figure 6: Boxplots of the ratio mean square error/empirical variance of the target (MSE) over 20 executions of algorithms for the *Ozone* dataset.

with  $\mathbb{Q}_n(A) = 0$ . Thus CA accuracy highly depends on the coverage  $\mathbb{Q}_n(A)$  of the sets  $A \in \mathcal{P}(\mathcal{C}_n)$  which can be too small to be stable. In Definition 2.1 of suitable data-dependent quasi-coverings only the coverage of the elements of  $\mathcal{C}_n$  is controlled. The issue might be solved by considering a deterministic rule generator.

## 5 Conclusion and perspectives

In this paper we provide a general setting for studying the consistency of interpretable rule-based estimators. We introduce the notion of suitable quasi-covering consisting of two kinds of sets, the significant and the insignificant ones. The significant sets are thought of as interpretable sets by construction. The insignificant ones are sets which variance tends to zero. We provide Covering Algorithm that extracts from any rule generator a suitable data-dependent covering. We apply it to Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2001) and Stochastic Gradient Boosting (Friedman, 2002) and we compare its result with the ones of CART (Breiman et al., 1984), Random Forest (Breiman, 2001), RuleFit (Friedman and Popescu, 2008), Node harvest (Meinshausen, 2010) and SIRUS (Bénard et al., 2020). The loss of accuracy in the prediction is the cost to pay to have an interpretable model according to our definition of interpretability. We broaden the accuracy-interpretability trade-off of classical algorithms by providing a much more interpretable method that remains consistent.

Our methodology based on quasi-coverings is very effective for generating interpretable models. The use of tree ensemble such as RF, GB or SGB as the rule-generator of Covering Algorithm is questionable; In Section 4.6, we pointed out the possible negative effect of the randomization procedure, when combined with the ERM principle rather than averaging, for identifying informative variables. We observed that the choice of the rule generator has an important effect on the accuracy, the interpretability and the stability. We look for an algorithm that generates deterministically significant, insignificant rules and a suitable sequence of data-dependent coverings satisfying (H8) on its own.

In practice the variance  $\sigma^2$  is unknown and has to be estimated efficiently enough. We let for future work the difficult question of defining a rule-based estimator of the variance with rate of convergence  $O_{\mathbb{P}}(n^{\alpha-1/2})$  as required in Theorem 2.1. Our setting could also be broadened; 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 regularity conditions on  $g^*$ . Finally, the scope could be easily adapted 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 the Covering Algorithm's selection process.



## Proof of Proposition 1.1

*Proof.* By the definition of  $\varphi_C$ , for any  $\mathbf{x} \in \mathbb{R}^d$ ,  $\varphi_C^{-1}(\varphi_C(\mathbf{x})) = \bigcap_{\mathbf{r} \in \mathcal{C}: \mathbf{x} \in \mathbf{r}} \mathbf{r} \setminus \bigcup_{\mathbf{r} \in \mathcal{C}: \mathbf{x} \notin \mathbf{r}} \mathbf{r}$ . Thus

$$\begin{aligned} A \in \mathcal{P}(\mathcal{C}) = \varphi_C^{-1}(\text{Im}(\varphi_C)) &\iff \exists \mathbf{x} \in \mathbb{R}^d / A = \varphi_C^{-1}(\varphi_C(\mathbf{x})) = \bigcap_{\substack{\mathbf{r} \in \mathcal{C} \\ \mathbf{x} \in \mathbf{r}}} \mathbf{r} \setminus \bigcup_{\substack{\mathbf{r} \in \mathcal{C} \\ \mathbf{x} \notin \mathbf{r}}} \mathbf{r} \\ &\iff \exists \tilde{\mathcal{C}} \subseteq \mathcal{C} / A = \bigcap_{\mathbf{r} \in \tilde{\mathcal{C}}} \mathbf{r} \setminus \bigcup_{\mathbf{r} \in \mathcal{C} \setminus \tilde{\mathcal{C}}} \mathbf{r} \text{ and } A \neq \emptyset. \end{aligned}$$

□

## Proof of Proposition 2.1

*Proof.* 1. • Lower bound: Consider the set  $\mathcal{E}_1$  of the points  $x \in \mathbb{R}^d$  whose coordinates are all zero except  $x[i] = 2$ , for  $1 \leq i \leq d$  and the set  $\mathcal{E}_2$  of the points  $x \in \mathbb{R}^d$  whose coordinates are all zero except  $x[i] = 1$  and  $x[i+1] = -1$ , for  $1 \leq i \leq d$  (with the convention  $x[d+1] = x[1]$ ). If  $\mathcal{P}$  is a partition by hyperrectangles and  $C \in \mathcal{P}$ , then no pair of the  $2d$  points of  $\mathcal{E}_1 \cup \mathcal{E}_2$  can be in the same element of  $\mathcal{P} \setminus \{C\}$ . Indeed, suppose that  $(x, y) \in \mathcal{E}_1^2$  (or analogously  $\in \mathcal{E}_1^2$ ) lies in some  $D \in \mathcal{P} \setminus \{C\}$ , with  $x[i] = 2$  and  $y[j] = 2$ . Since  $R$  is an hyperrectangle, the point  $z$  whose coordinates are all zero except  $z[i] = \frac{x[i]+y[i]}{2} = 1$  and  $z[j] = \frac{x[j]+y[j]}{2} = 1$  is in  $R$  too. Thus  $C$  and  $R$  intersect. In the same way, suppose that  $(x, y) \in \mathcal{E}_1 \times \mathcal{E}_2$  lies in some  $D \in \mathcal{P} \setminus \{C\}$ , with  $x[i] = 2$ ,  $y[j] = 1$  and  $y[j+1] = -1$ . Then, the point  $z$  whose coordinates are all zero except  $z[i] = 1$  (while  $z[j] = x[j] = 0$  and  $z[j+1] = x[j+1] = 0$ ) is in  $R$  too. Thus  $C$  and  $R$  intersect. Thus, the  $2d$  points of  $\mathcal{E}_1 \cup \mathcal{E}_2$  induce  $2d$  distinct elements of  $\mathcal{P}$  that are added to  $C$ .

• Upper bound: The partition

$$\begin{aligned} \mathcal{P} = \{ &(-\infty, 0) \times \mathbb{R}^{d-1}; (1, +\infty) \times \mathbb{R}^{d-1}; \\ &[0, 1] \times (-\infty, 0) \times \mathbb{R}^{d-2}; [0, 1] \times (1, +\infty) \times \mathbb{R}^{d-2}; \\ &\dots; \\ &[0, 1]^{d-1} \times (-\infty, 0); [0, 1]^{d-1} \times (1, +\infty); \\ &[0, 1]^d \} \end{aligned}$$

fulfills the conditions of the proposition.

2. The covering  $\mathcal{C} = \{[0, 1]^d; \mathbb{R}^d\}$  fulfills the conditions of the proposition.

□

### Proof of Proposition 3.2

For any  $f : \mathcal{S} \rightarrow \mathbb{R}$  in  $\mathcal{L}^1(\mathbb{Q})$  we note classically  $\mathbb{Q}f := \int f d\mathbb{Q}$ ,  $\mathbb{Q}_n f := \int f d\mathbb{Q}_n$ .

*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  $\mathbb{Q}(A) > 0$ ,

$$\begin{aligned}
& |\mathbb{E}_n[f | A] - \mathbb{E}[f | A]| \\
&= \left| \frac{\int_A f d\mathbb{Q}_n}{\mathbb{Q}_n(A)} - \frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)} \right| \\
&= \left| \frac{\mathbb{Q}(A) (\int_A f d\mathbb{Q}_n - \int_A f d\mathbb{Q}) + (\mathbb{Q}(A) - \mathbb{Q}_n(A)) \int_A f d\mathbb{Q}}{\mathbb{Q}(A)\mathbb{Q}_n(A)} \right| \\
&\leq \left| \frac{\int_A f d\mathbb{Q}_n - \int_A f d\mathbb{Q}}{\mathbb{Q}_n(A)} \right| + \left| (\mathbb{Q}(A) - \mathbb{Q}_n(A)) \frac{\int_A f d\mathbb{Q}}{\mathbb{Q}(A)\mathbb{Q}_n(A)} \right|. \tag{5.1}
\end{aligned}$$

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 7, 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

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

Then (5.1) yields, with  $c := \sup_{f \in \mathcal{F}, \mathbf{x} \in \mathcal{S}} |f(\mathbf{x})| < \infty$  and since  $\mathbb{Q}_n(A) \geq n^{-\alpha}$ , for  $n$  large enough, in the event  $\Omega_n$ ,

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

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

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

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

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

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

which, together with Remark 7 again, proves the proposition.  $\square$

### Proof of Corollary 3.1

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

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

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

*Proof of (3.2).* This part follows from Proposition 3.2 since  $Y$  is bounded and

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

$\square$

Dataset	Regression Tree (CART)			Random Forest (RF)		
	Nb rules	Int	MSE	Nb rules	Int	MSE
Diabetes	10	30.2	1.14	191190.6	1070454.75	<b>0.58</b>
Prostate	10	25.3	1.11	42047.95	169328.6	<b>0.56</b>
Ozone	10	32	0.5	108082.95	572067.75	<b>0.28</b>
Machine	10	24.2	0.18	73754.25	319376.45	<b>0.06</b>
MPG	10	29.0	0.27	148954.4	623266.0	<b>0.14</b>
Boston	10	26.3	0.29	212961.05	1270252.25	<b>0.14</b>
Student	10	37.75	1.55	199278.65	1838107.25	<b>0.74</b>

Dataset	CA RF				CA GB			
	Nb rules	Cov	Int	MSE	Nb rules	Cov	Int	MSE
Diabetes	<b>7.6</b>	0.99	<b>17.95</b>	0.70	9.1	0.99	19.7	0.74
Prostate	<b>5.8</b>	1.0	<b>13.95</b>	0.85	<b>6.05</b>	1.0	<b>13.35</b>	0.86
Ozone	<b>3.3</b>	0.99	<b>5.85</b>	0.42	3.95	0.99	6.55	0.39
Machine	<b>2.85</b>	0.99	<b>5.05</b>	0.40	3.25	0.99	6.05	0.50
MPG	<b>3.1</b>	0.99	<b>4.3</b>	0.28	<b>3.65</b>	0.99	6.85	0.30
Boston	<b>4.9</b>	0.99	<b>10.9</b>	0.45	6.8	0.99	15.15	0.45
Student	18.45	0.99	51.4	0.86	32	0.98	86.75	1.01

Dataset	CA SGB				RuleFit			
	Nb rules	Cov	Int	MSE	Nb rules	Cov	Int	MSE
Diabetes	8.75	0.99	19.8	0.67	249.35	1.0	1168.7	0.71
Prostate	6.5	1.0	15.05	0.76	78.4	1.0	268.8	0.66
Ozone	4.05	0.99	6.5	0.38	187.45	1.0	851	<b>0.28</b>
Machine	3.2	0.99	6.1	0.52	99.85	1.0	278.15	0.08
MPG	3.75	0.99	7.25	0.29	174.7	1.0	745.15	<b>0.13</b>
Boston	6.35	0.99	13.85	0.46	264.95	1.0	1207.65	<b>0.12</b>
Student	37.3	0.98	100.05	1.01	277.95	1.0	1280.4	0.90

Dataset	NodeHarvest				SIRUS			
	Nb rules	Cov	Int	MSE	Nb rules	Cov	Int	MSE
Diabetes	227.15	1.0	613.05	<b>0.59</b>	21.25	0.99	32.35	<b>0.57</b>
Prostate	66	1.0	160.35	<b>0.57</b>	23.25	0.99	33.7	<b>0.59</b>
Ozone	181.45	1.0	472.9	<b>0.29</b>	24.9	1.0	39.4	0.32
Machine	67.15	1.0	159.0	0.32	13.9	1.0	21.85	0.26
MPG	100.15	1.0	257.85	0.15	25.0	1.0	45.2	0.21
Boston	139.85	1.0	345.4	0.20	25.0	10.99	38.5	0.25
Student	78.5	1.0	228.45	<b>0.76</b>	<b>15.35</b>	1.0	<b>26.2</b>	<b>0.77</b>

Table 3: Average of the number of rules (NB rules), the empirical coverage rate on the training set (Cov), the interpretability index (Int) and the ratio mean square error/empirical variance of the target (MSE) over 20 executions of usual interpretable algorithms for various regression public datasets. Best values are in bold, as well as values within 10% of the best for each dataset.

Rule	Conditions	Coverage	Prediction	Std	$\Delta_n$
<i>R1</i>	$temp \in [30, 67.5]$	0.65	7.99	4.57	-0.2
<i>R2</i>	$ibt \in [-15, 191]$	0.60	8.02	4.5	-0.2
<i>R3</i>	$ibt \in [172, 326]$ $humidity \in [35.5, 92]$	0.41	18.61	7	0.8
<i>R4</i>	$ibt \in [172, 326]$ $humidity \in [19, 35.5]$	0.08	6.11	2.64	-0.4

Table 4: Summary of the rules selected by CA RF on the dataset *Ozone*. All rules are significant. The *temp* variable is the Sandburg Air Force base temperature in degrees Fahrenheit, the *ibt* variable is the inversion base temperature at LAX in degrees Fahrenheit and the *humidity* variable is the humidity in percent at LAX.

	Y	<i>temp</i>	<i>ibt</i>	<i>humidity</i>
mean	11.78	61.75	161.16	58.13
std	8.01	14.46	76.68	19.87
min	1.00	25.00	-25.00	19.00
25%	5.00	51.00	107.00	47.00
50%	10.00	62.00	167.50	64.00
75%	17.00	72.00	214.00	73.00
max	38.00	93.00	332.00	93.00

Table 5: Description of the variables selected by CA RF for the dataset *Ozone*. The *Y* variable is the daily maximum of the hourly average ozone concentrations in Upland.

## Covering Algorithm selection process

---

**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  $\mathcal{C}_n \leftarrow \arg \max_{\mathbf{r} \in S} \mathbb{Q}_n(\mathbf{r})$ ;
2  $S \leftarrow S \setminus \mathcal{C}_n$ ;
3 while  $\sum_{\mathbf{r} \in \mathcal{C}_n} \mathbb{Q}_n(\mathbf{r}) < 1$  do
4    $\mathbf{r}^* \leftarrow \arg \max_{\mathbf{r} \in S} \mathbb{Q}_n(\mathbf{r})$ ;
5   if  $\mathbb{Q}_n(\mathbf{r}^* \cap \{\cup_{\mathbf{r} \in \mathcal{C}_n} \mathbf{r}\}) \leq \gamma \mathbb{Q}_n(\mathbf{r}^*)$  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  $\sum_{\mathbf{r} \in \mathcal{C}_n} \mathbb{Q}_n(\mathbf{r}) < 1$  then
12   while  $\sum_{\mathbf{r} \in \mathcal{C}_n} \mathbb{Q}_n(\mathbf{r}) < 1$  do
13      $\mathbf{r}^* \leftarrow \arg \min_{\mathbf{r} \in I} \mathbb{V}_n(Y|\mathbf{X} \in \mathbf{r})$ ;
14     if  $\mathbb{Q}_n(\mathbf{r}^* \cap \{\cup_{\mathbf{r} \in \mathcal{C}_n} \mathbf{r}\}) \leq \gamma \mathbb{Q}_n(\mathbf{r}^*)$  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  $\mathcal{C}_n$ ;

```

---

## References

- Bénard, C., Biau, G., Da Veiga, S., and Scornet, E. (2020). Interpretable random forests via rule extraction. *arXiv preprint arXiv:2004.14841*.
- Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, page 1.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. CRC press.
- Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*.
- Denil, M., Matheson, D., and Freitas, N. (2013). Consistency of online random forests. In *International Conference on Machine Learning*, pages 1256–1264.
- Dua, D. and Graff, C. (2017). Uci machine learning repository.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.
- Friedman, J. and Popescu, B. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954.
- Friedman, J., Popescu, B., et al. (2003). Importance sampled learning ensembles. *Journal of Machine Learning Research*, 94305.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.
- Fürnkranz, J. and Kliegr, T. (2015). A brief overview of rule learning. In *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, pages 54–69. Springer.
- Grunewalder, S. (2018). Plug-in estimators for conditional expectations and probabilities. In *International Conference on Artificial Intelligence and Statistics*, pages 1513–1521.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.
- Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.
- Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.
- Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer series in statistics Springer, Berlin.
- Holmes, G., Hall, M., and Prank, E. (1999). Generating rule sets from model trees. In *Australasian Joint Conference on Artificial Intelligence*, pages 1–12. Springer.

- Karalič, A. and Bratko, I. (1997). First order regression. *Machine Learning*, 26(2-3):147–176.
- Liitiäinen, E., Verleysen, M., Corona, F., and Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72(16-18):3692–3703.
- Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.
- Margot, V., Baudry, J.-P., Guilloux, F., and Wintenberger, O. (2018). Rule induction partitioning estimator. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 288–301. Springer.
- Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, 4(4):2049–2072.
- Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.
- Quinlan, J. R. (1993). C4.5: Programs for machine learning.
- Ramosaj, B. and Pauly, M. (2019). Consistent estimation of residual variance with random forest out-of-bag errors. *Statistics & Probability Letters*, 151:49–57.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). 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.
- Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.
- Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.



- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
- Wenocur, R. S. and Dudley, R. M. (1981). Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33(3):313–318.
- Zhao, Q. and Bhowmick, S. S. (2003). Association rule mining: A survey. *Nanyang Technological University, Singapore*.