Model Selection for CART Regression Trees

Servane Gey and Elodie Nedelec

Abstract—**The performance of the classification and regression trees (CART) pruning algorithm and the final discrete selection by test sample as a functional estimation procedure are considered. The validation of the pruning procedure applied to Gaussian and bounded regression is of primary interest. On the one hand, the paper shows that the complexity penalty used in the pruning algorithm is valid in both cases and, on the other hand, that, conditionally to the construction of the maximal tree, the final selection does not alter dramatically the estimation accuracy of the regression function. In both cases, the risk bounds that are proved, obtained by using the penalized model selection, validate the CART algorithm which is used in many applications such as meteorology, biology, medicine, pollution monitoring, or image coding.**

Index Terms—**Bounded regression, CART, Gaussian regression, model selection, pruning.**

I. Introduction

THE aim of classification and regression trees (CART) proposed by Breiman, Friedman, Olshen, and Stone [1] in 1984 is to construct an efficient algorithm which gives a piecewise-constant estimator of a classifier or a regression function from a training sample of observations. This algorithm is based on binary tree-structured partitions and on a penalized criterion that permits to select some "good" tree-structured estimators among a huge collection of trees. In practice, it yields some easy-to-interpret and easy-to-compute estimators which are widely used in many applications such as medicine, meteorology, biology, pollution, or image coding (see [2], [3] for example). From a more general point of view on regression methods, this kind of algorithm is often performed when the space of explanatory variables is high dimensional. Indeed, due to its local splitting, CART needs fewer operations than other usual methods to provide estimators.

More precisely, given a training sample of observations, the CART algorithm consists in constructing a large tree from the observations by minimizing at each step some impurity function, and then, in pruning the thus constructed tree to obtain a finite sequence of nested trees thanks to a penalized criterion, whose penalty term is proportional to the number of leaves.

This raises the question of "why" this penalty is well chosen. This paper aims at validating the choice of the penalty in the Gaussian and bounded regression frameworks. In the classification case, it is not that clear for a good penalty to be proportional to the number of leaves. The interested reader will find some discussions and results about this topic in the paper by Nobel [4].

Let $\mathcal{L} = \{(X_1, Y_1); \ldots; (X_N, Y_N)\}$ be a set of independent random variables, where each $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ follows a regression model with a common regression function s. Let $\tilde{s}$ be the piecewise-constant estimator of s provided by CART. We measure the performance of $\tilde{s}$ by the risk defined as follows:

$$R(\tilde{s}, s) = \mathbb{E}\left[(\tilde{s}(X) - s(X))^2\right] \tag{1}$$

where $\mathbb{E}$ denotes the expectation with respect to the current distribution of (X, Y).

In this paper, we leave aside the analysis of the growing procedure to focus on the pruning procedure. We show that this method, used to reduce the complexity of the problem, is well chosen in the sense that it guarantees a good performance of the selected estimator $\tilde{s}$ in terms of its risk $R(\tilde{s}, s)$. All our upper bounds for the risk are considered conditionally to the growing procedure. For results about the growing procedure see the papers by Nobel and Olshen [5] and Nobel [6] about recursive partitioning.

Furthermore, Breiman *et al.* [1] propose two algorithms in their book, one using a test sample and another using cross-validation. We focus on two methods that use a test sample and give about the same results: let us split $\mathcal{L}$ in three independent subsamples $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$, containing, respectively, n_1, n_2, and n_3 observations, with $n_1 + n_2 + n_3 = N$. $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ are randomly taken in $\mathcal{L}$, except if the design is fixed. In that case, one takes, for example, one observation out of three to obtain each subsample. Given these three subsamples, suppose that either a large tree is constructed using $\mathcal{L}_1$ and then pruned using $\mathcal{L}_2$ (as done in Gelfand *et al.* [7]), or a large tree is constructed and pruned using the same subsample $\mathcal{L}_1$ (as done in Breiman *et al.* [1]). Then the final step used in both cases is to choose a subtree among the sequence obtained after the pruning procedure. The method we will study in the rest of the paper is to make $\mathcal{L}_3$ go down each tree of the sequence and to select the tree which has the minimum empirical quadratic contrast, i.e., given for any $k = 1, 2, 3$ and any $u \in \mathbb{L}^2(\mathcal{X})$ the empirical quadratic contrast

$$\gamma_{n_k}(u) = \frac{1}{n_k} \sum_{(X_i, Y_i) \in \mathcal{L}_k} (Y_i - u(X_i))^2 \tag{2}$$

to take the final estimator of s as follows:

$$\tilde{s} = \underset{\{\hat{s}_{T_i}; 1 \leqslant i \leqslant K\}}{\operatorname{argmin}} \left[\gamma_{n_3}(\hat{s}_{T_i})\right] \tag{3}$$

where $\hat{s}_{T_i}$ is the piecewise-constant estimator of s defined on the leaves of the tree T_i and K is the number of trees appearing in the sequence.

Manuscript received January 30, 2002; revised October 4, 2004.

S. Gey is with the Laboratoire MAP5–Université Paris V, 75270 Paris Cedex 06, France (e-mail: Servane.Gey@math-info.univ-paris5.fr).

E. Nedelec is with the Laboratoire de Mathématiques–U.M.R. 8628 Université Paris XI, 91405 Orsay Cedex, France (e-mail: Elodie.Nedelec@math.u-psud.fr).

Communicated by G. Lugosi, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2004.840903

0018-9448/$20.00 © 2005 IEEE

In this paper, we analyze the risk of $\tilde{s}$ and we prove that the penalty used in the pruning algorithm for the two above mentioned cases is well-chosen, using some of Birgé, Massart's [8] and Massart's [9] results on model selection via dimensional penalization. Engel [10] and Donoho [11] obtain some results of consistency in the regression case for estimators by histograms constructed via binary partitioning on a dyadic deterministic grid of points $(x_i)_{1\leqslant i\leqslant N}$ (in dimension one for Engel and dimension two for Donoho). This framework differs from ours in the sense that the grid $(X_i)_{1\leqslant i\leqslant N}$ we consider can be random. Moreover, the results we obtain are nonasymptotic upper bounds for the risk of the resulting histogram estimator. For more details about asymptotic results, see also Nobel [12].

The paper is organized as follows. In Section II, we recall some facts about the CART algorithm and give some notation used in the rest of the paper. In Section III, we study the Gaussian regression framework, in which we validate the pruning algorithm taking either $\mathcal{L}_1$ independent of $\mathcal{L}_2$ or $\mathcal{L}_1 = \mathcal{L}_2$ and give an upper bound concerning the final selection using $\mathcal{L}_3$ as test sample. In Section IV, we perform the same program for the bounded regression framework. Section V is devoted to some open questions and the proofs of the results obtained in the previous sections are given in the last sections.

II. Preliminaries and Notation

A. The CART Algorithm

Let us give a short account of the CART algorithm in the regression case and recall the results associated with it, which are fully explained in [1].

CART is based on recursive partitioning using a training sample $\widetilde{\mathcal{L}}$ of the random variable $(X,Y) \in \mathcal{X} \times \mathbb{R}$ (we shall take as $\widetilde{\mathcal{L}} = \mathcal{L}_1$ or $\widetilde{\mathcal{L}} = \mathcal{L}_1 \cup \mathcal{L}_2$), and a class $\mathcal{S}$ of subsets of $\mathcal{X}$ which tells us how to split at each step. Usuall,y $\mathcal{S}$ is taken as some class of half-spaces of $\mathcal{X}$, for example the half-spaces of $\mathcal{X}$ with frontiers parallel to the axes (see, for example, [1], [11]). In our framework, we consider a class $\mathcal{S}$ with finite Vapnik–Chervonenkis (VC) dimension, henceforth refered to as VC-dimension (for a complete overview of the VC-dimension see [13]).

The algorithm is computed in two steps, that we call *growing procedure* and *pruning procedure*. The growing procedure permits to construct, from the data, a maximal binary tree $T_{\max}$ by recursive partitioning, and then the pruning procedure permits to select, among all the subtrees of $T_{\max}$, a sequence which contains the entire statistical information.

1) Growing Procedure: The aim of the growing procedure is to construct by recursive partitioning a maximal binary tree $T_{\max}$ based on the data composing $\mathcal{L}_1$ and on the class $\mathcal{S}$ of subsets of $\mathcal{X}$. This algorithm yields a sharp partition of $\mathcal{X}$, providing a large collection of estimators.

The first step is computed as follows: the whole space $\mathcal{X}$ is assimilated to the *root* of the tree, denoted by t_1, so that every observation $X_i (1 \leqslant i \leqslant n_1)$ belongs to t_1. The next step starts by computing the first *split* as

$$\hat{sp} = \operatorname*{argmin}_{sp\in\mathcal{S}} \{\gamma_{n_1}(\hat{s}_{|sp}) + \gamma_{n_1}(\hat{s}_{|sp^c})\}.$$

Here, for any subset sp of $\mathcal{X}$, $\hat{s}_{|sp}$ is the minimum least-squares estimator of s on the set of constant functions on sp, that is,

$$\begin{aligned}\hat{s}_{|sp} &= \operatorname{argmin}_{\{a\mathbb{1}_{sp};a\in\mathbb{R}\}}\gamma_{n_1}(a\mathbb{1}_{sp})\\ &= \overline{Y}_{sp}\mathbb{1}_{sp}\end{aligned}$$

where, for all $x \in \mathcal{X}$, $\mathbb{1}_{sp}(x) = 1$ if $x \in sp$ and $\mathbb{1}_{sp}(x) = 0$ otherwise, and

$$\overline{Y}_{sp} = \frac{1}{|X_i \in sp|}\sum_{X_i\in sp} Y_i.$$

Hence, noticing that $\gamma_{n_1}(\hat{s}_{|sp}+\hat{s}_{|sp^c}) = \gamma_{n_1}(\hat{s}_{|sp})+\gamma_{n_1}(\hat{s}_{|sp^c})$, the data of $\mathcal{L}_1$ are split in such a way that the interclass variance between $\{Y_i; X_i \in \hat{sp}\}$ and $\{Y_i; X_i \in \hat{sp}^c\}$ is maximal. In the tree terminology, one adds to the root t_1 a left node t_L (assimilated to $\hat{sp}$) and a right node t_R (assimilated to $\hat{sp}^c$). In what follows, we always assimilate a tree node with its corresponding subset in $\mathcal{S}$.

Then the same elementary step is applied recursively to the two generated subsamples $\{(X_i,Y_i); X_i \in \hat{sp}\}$ and $\{(X_i,Y_i); X_i \in \hat{sp}^c\}$ until some convenient stopping condition is satisfied. This provides the maximal tree $T_{\max}$ and one calls *terminal nodes* or *leaves* the final nodes of $T_{\max}$.

2) Pruning Procedure: First let us recall that a pruned subtree of $T_{\max}$ is defined as any binary subtree of $T_{\max}$ having the same root t_1 as $T_{\max}$.

Then, let us introduce the following notation.

i) Take two trees T_1 and T_2. Then, if T_1 is a pruned subtree of T_2, write $T_1 \preceq T_2$.

ii) For a tree T, $\widetilde{T}$ denotes the set of its leaves and $|T|$ the cardinality of $\widetilde{T}$.

To prune $T_{\max}$, one proceeds as follows. First, simply denote by n the number of data used. Notice that, given a tree T and S_T a set of piecewise functions in $\mathbb{L}^2(\mathcal{X})$ defined on the partition given by the leaves of T, one has

$$\begin{aligned}\hat{s}_T &= \operatorname{argmin}_{z\in S_T}\gamma_n(z)\\ &= \sum_{t\in\widetilde{T}} \overline{Y}_t\mathbb{1}_t.\end{aligned}$$

Then, given $T \preceq T_{\max}$ and $\alpha > 0$, one defines

$$\operatorname{crit}_\alpha(T) = \gamma_n(\hat{s}_T) + \alpha\frac{|T|}{n} \tag{4}$$

the penalized criterion for the so-called temperature α, and T_α the subtree of $T_{\max}$ satisfying:

i) $T_\alpha = \operatorname{argmin}_{T\preceq T_{\max}}\operatorname{crit}_\alpha(T)$;

ii) if $\operatorname{crit}_\alpha(T) = \operatorname{crit}_\alpha(T_\alpha)$, then $T_\alpha \preceq T$.

Thus, T_α is the smallest minimizing subtree for the temperature α. The existence and the unicity of T_α are given in [1, pp. 284–290].

The aim of the pruning procedure is to make the temperature α increase and to take at each time the corresponding T_α. The algorithm is an iterative one consisting in minimizing at each step a function of the nodes, which leads to a finite decreasing sequence of subtrees pruned from $T_{\max}$

$$T_{\max} \succeq T_1 \succ \cdots \succ T_{K-1} \succ T_K = \{t_1\}$$

corresponding to a finite increasing sequence of temperatures

$$0 = \alpha_1 < \alpha_2 < \cdots < \alpha_{K-1} < \alpha_K.$$

Remark 1: T_1 is the smallest subtree for the temperature 0, so it is not necessarily equal to $T_{\max}$.

Breiman, Friedman, Olshen, and Stone's theorem [1] justifies this algorithm.

Theorem II-A.1: (Breiman, Friedman, Olshen, Stone) The sequence $(\alpha_k)_{1 \leqslant k \leqslant K}$ is nondecreasing, the sequence $(T_k)_{1 \leqslant k \leqslant K}$ is nonincreasing, and, given $k \in \{1, \ldots, K\}$, if $\beta \in [\alpha_k, \alpha_{k+1}[$, then $T_\beta = T_{\alpha_k} = T_k$.

By this theorem, it is easy to check that, for any $\alpha > 0$, T_α belongs to the sequence $(T_k)_{1 \leqslant k \leqslant K}$.

It is easily seen that this algorithm reduces the complexity of the choice of a subtree pruned from $T_{\max}$ efficiently, since by Theorem II-A.1 the sequence of pruned subtrees contains the whole statistical information according to the choice of the penalty function used in (4). Consequently, it is useless to look at all the subtrees. Hence, to validate this algorithm completely, it remains to show that this choice of penalty is convenient.

The final step is to choose a suitable temperature α. Instead of minimizing over α, this issue is dealt with by using a test sample to provide the final estimator $\tilde{s}$, as mentioned in the Introduction, via equality (3). The results given in Sections III and IV deal, on the one hand, with the performance of the piecewise-constant estimators given by T_α for α fixed and, on the other hand, with the performance of $\tilde{s}$.

3) Notation: Assume we observe a set of independent random variables $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ such that

$$Y_i = s(X_i) + \varepsilon_i$$

where (X_i, Y_i) lies in $\mathcal{X} \times \mathbb{R}$, ε_i is a noise centered conditionally to X_i, and s is the regression function to be estimated. Let us define by μ the common distribution of the $(X_i)_{1 \leqslant i \leqslant N}$ and by $\|\cdot\|$ the $\mathbb{L}^2(\mathcal{X}, \mu)$-norm. Then the risk (1) of the final estimator $\tilde{s}$ becomes

$$R(\tilde{s}, s) = \mathbb{E}[\|\tilde{s} - s\|^2].$$

Next, for a given tree T, S_T will denote the set of some piecewise-constant functions defined on the partition given by the leaves of T. Thus, $\hat{s}_T$ will be the minimum quadratic contrast estimator of s on S_T. Then a tree-structured estimator $\hat{s}$ of s is said to satisfy an oracle inequality if there exists some nonnegative constant C, such that

$$\mathbb{E}\left[\|s - \hat{s}\|^2 \mid \mathcal{L}_1\right] \leqslant C \inf_{T \preceq T_{\max}} R_{\mathcal{L}_1}(\hat{s}_T, s)$$

where, for each subtree T pruned from $T_{\max}$,

$$R_{\mathcal{L}_1}(\hat{s}_T, s) = \mathbb{E}\left[\|s - \hat{s}_T\|^2 \mid \mathcal{L}_1\right].$$

To estimate s using the CART algorithm and to compare the performance of $\tilde{s}$ with those of each $\hat{s}_T$, two different methods can be applied.

M1: $\mathcal{L}$ is split in three independent parts $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ containing, respectively, n_1, n_2, and n_3 observations. Hence, $T_{\max}$ is constructed using $\mathcal{L}_1$, then pruned using $\mathcal{L}_2$, and finally a best subtree $\widehat{T}$ is selected among the sequence of pruned subtrees thanks to $\mathcal{L}_3$, and we define $\tilde{s} = \hat{s}_{\widehat{T}}$.

M2: $\mathcal{L}$ is split in two independent parts $\mathcal{L}_1$ and $\mathcal{L}_3$ containing, respectively, n_1 and n_3 observations. Hence, $T_{\max}$ is constructed and pruned using $\mathcal{L}_1$ and finally, a best subtree $\widehat{T}$ is selected among the sequence of pruned subtrees thanks to $\mathcal{L}_3$, and we define $\tilde{s} = \hat{s}_{\widehat{T}}$.

Note that a penalty is needed in both methods in order to reduce the number of candidate tree-structured models contained in $T_{\max}$. Indeed, if one does not penalize, the number of models to be considered grows exponentially with N, so results such as the ones of Wegkamp [14] cannot be applied. Then, making a selection by using a test sample without penalizing requires to visit all the models. As we will see in Sections III-B and IV-B, since in that case the number of models considered occurs via its logarithm in the upper bound of the risk, the resulting estimator will have a significantly large upper bound for its risk. Hence, penalizing permits to reduce significantly the number of trees taken into account and then to get a convenient risk for $\tilde{s}$. Both methods M1 and M2 are considered for the following reasons.

- Since all the risks are considered conditionnally to the growing procedure, the M1 method permits to make a deterministic penalized model selection and then to obtain sharper upper bounds than the M2 method.
- To the contrary, the M2 method permits to keep the whole information given by $\mathcal{L}_1$, since, in that case, the sequence of pruned subtrees is not obtained via some plug-in method using a first split of the sample to provide the collection of tree-structured models. This method is the one proposed by Breiman *et al.* and it is more commonly applied in practice than the M1 one. We focus on this method to ensure that it provides estimators that have good performance in terms of risk.

Let us recall that the aim of this paper is to prove on the one hand that the complexity penalty used by Breiman *et al.* [1] in the pruning algorithm is well-chosen, and, on the other hand, that the final selection among the pruned subtrees is, in terms of risk, not far from being optimal. We focus more particularly on Gaussian and bounded regression. The Gaussian case is classical and Birgé, Massart [8] obtain optimal constants for the risk of the penalized estimator in this case. The bounded case can be viewed as a first step to obtain similar results for the two-class classification problem, for which the penalty term is not obviously proportional to the number of leaves. From this viewpoint, the quadratic risk is equal to the misclassification cost. This is why we do not address here the issue of other estimation methods, as for example the maximum-likelihood estimation which is used in logistic regression and can sometimes do better than least squares estimators in this case.

Sections III and IV are, respectively, devoted to the two above-mentioned cases and consider separately the pruning procedure and the final selection by test sample. We will see that, conditionally to the construction of $T_{\max}$, the final estimator $\tilde{s}$ satisfies some oracle-type inequalities for the Gaussian case when using either method M1 or M2. Moreover, the penalty term is the same with the two methods, although a factor $\log n_1$ can occur in the temperature when $\mathcal{L}_1 = \mathcal{L}_2$. In addition, the penalized model selection is made via pruning on random models defined on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$. Then, by using Birgé and Massart's results for Gaussian regression on fixed design and working conditionally to $\mathcal{L}_1$, as we will see in Section III, the norm occurring in the risk for pruning is the empirical norm $\|\cdot\|_2$ on $\{X_i; (X_i, Y_i) \in \mathcal{L}_2\}$ for M1 and the empirical norm $\|\cdot\|_1$ on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$ for M2. On the other hand, the norm occurring in the risk for the discrete selection is the empirical norm $\|\cdot\|_3$ on $\{X_i; (X_i, Y_i) \in \mathcal{L}_3\}$. Nevertheless, under some truncation arguments, the results of Baraud [15] or Wegkamp [14] can be applied and the results will be true under the underlying norm $\|\cdot\|$. The results are slightly different for the bounded case since the norms that will occur are $\|\cdot\|$ for M1 and $\|\cdot\|_1$ for M2, the selection being made on a deterministic grid conditionally to $\mathcal{L}_1$ in the M1 case. In that case, a connection can be made between pruning and final selection by test sample.

Note that the constants appearing in the upper bounds for the risks are not sharp. We do not investigate the sharpness of the constants here.

III. Gaussian Regression

Let us consider the Gaussian regression framework, where, for a given $i \in \{1; \ldots; N\}$, ε_i is $\mathcal{N}(0, \sigma^2)$-distributed conditionnally to X_i, with σ^2 known. The two following subsections give some more precise results on the pruning algorithm for both the M1 and M2 methods, and particularly on the constants appearing in the penalty function. The last subsection validates the discrete selection by test sample. Note that the two results obtained for the validation of the pruning algorithm also hold in the case of deterministic X_i's.

A. Validation of the Pruning Algorithm

In this subsection, we focus on the pruning algorithm and show that, for a convenient constant α, $\hat{s}_{T_\alpha}$ (where T_α is the smallest minimizing subtree for the temperature α as defined in Section II-A) is not far from s in terms of its risk conditionally to $\mathcal{L}_1$. Let us emphasize that the subsample $\mathcal{L}_3$ plays no role in the two following results.

1) $\tilde{s}$ Constructed Via M1: Here we consider the second subsample $\mathcal{L}_2$ of n_2 observations. We assume that $T_{\max}$ is constructed on the first set of observations $\mathcal{L}_1$ and then pruned with the second set $\mathcal{L}_2$ independent of $\mathcal{L}_1$. Since the set of pruned subtrees is deterministic according to $\mathcal{L}_2$, we make a selection among a deterministic collection of models. By this way, since $T_{\max}$ is fixed, we do not have to look at the manner that $T_{\max}$ is constructed. Hence, in contrast to Proposition 2 in the following, where the growing and pruning procedures are made on the same sample, the parameters occurring in the growing procedure, as the VC-dimension of the set of split used, play no role in the bounds or constants we obtain here.

In the rest of the paper, given a subtree T of $T_{\max}$, we write S_T the linear subspace of $\mathbb{L}^2(\mathcal{X}, \mu)$ composed by all the piecewise-constant functions defined on the partition associated with the leaves of T. S_T is then a model on which s will be estimated, and its dimension is $|T|$. Then we choose the estimators as follows:

- for $T \preceq T_{\max}$, $\hat{s}_T = \operatorname{argmin}_{t \in S_T} [\gamma_{n_2}(t)]$;
- for $\alpha > 0$, T_α is the smallest minimizing subtree for the temperature α as defined in Section II-A and $\hat{s}_{T_\alpha} = \operatorname{argmin}_{t \in S_{T_\alpha}} [\gamma_{n_2}(t)]$.

Let us now consider the behavior of such $\hat{s}_{T_\alpha}$.

Taking (1) into account, the following upper bound is actually an upper bound for the risk of $\hat{s}_{T_\alpha}$ conditionally to $\mathcal{L}_1$:

Proposition 1: Let $\|\cdot\|_2$ be the empirical norm on $\{X_i; (X_i, Y_i) \in \mathcal{L}_2\}$ and, for each $T \preceq T_{\max}$ and each $u \in S_T$, let

$$R^{(2)}_{\mathcal{L}_1}(s, u) = \mathbb{E}\left[\|s - u\|_2^2 \mid \mathcal{L}_1\right].$$

If $\alpha > \sigma^2(1 + 4\log 2 + 2\sqrt{2\log 2})$, then there exist some nonnegative constants Σ_α and C'_2 such that

$$R^{(2)}_{\mathcal{L}_1}(s, \hat{s}_{T_\alpha}) \leqslant C'_1(\alpha) \inf_{T \preceq T_{\max}} \left\{ \inf_{u \in S_T} R^{(2)}_{\mathcal{L}_1}(s, u) + \sigma^2 \frac{|T|}{n_2} \right\} + C'_2 \sigma^2 \frac{\Sigma_\alpha}{n_2}$$

where $C'_1(\alpha) > (1 + 4\log 2 + 2\sqrt{2\log 2})$ and Σ_α are increasing with α.

A proof of this proposition is given in Appendix II-A.

To conclude:

— the penalty term is the same as the one proposed by Breiman *et al.* [1] in their pruning algorithm;

— the loss of $\hat{s}_T$ with respect to s is

$$R^{(2)}_{\mathcal{L}_1}(s, \hat{s}_T) = \inf_{u \in S_T} R^{(2)}_{\mathcal{L}_1}(s, u) + \sigma^2 \frac{|T|}{n_2}. \tag{5}$$

Thus, for a large enough α, $\hat{s}_{T_\alpha}$ satisfies in this case an oracle inequality up to some additive constants;

— the inequality holds only for large enough temperatures α. Nevertheless, when α becomes too large, the models are overpenalized, and the left-hand side $\mathbb{E}\left[\|s - \hat{s}_{T_\alpha}\|_2^2 \mid \mathcal{L}_1\right]$ will grow with α. The main issue at this stage is to choose a temperature α making a good compromise between the size of $\mathbb{E}\left[\|s - \hat{s}_{T_\alpha}\|_2^2 \mid \mathcal{L}_1\right]$ and a large enough penalty term. This issue is partially addressed in Section V. $C'_1(\alpha)$ and Σ_α are increasing with α, so both sides of the inequality grow with α.

Note that under the following condition on the distribution μ of the $(X_i)_{1 \leqslant i \leqslant N}$

$$\inf_{t \in \widetilde{T}_{\max}} \mu(t) > \frac{(\log n_1)^3}{n_1} \tag{6}$$

and using truncation, the results of Baraud [15] can be applied and the same inequality holds under the $\mathbb{L}^2(\mathcal{X}, \mu)$-norm $\|\cdot\|$ on a large probability set.

2) $\tilde{s}$ Constructed Via M2: In this subsection, we define the different estimators and projections exactly in the same way as in Section III-A1, where $\|\cdot\|_2$ is replaced by the empirical norm $\|\cdot\|_1$ on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$ since the models and the evaluations of the empirical errors $\gamma_{n_1}(\hat{s}_T)$ are computed on the same grid $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$. In this case, we obtain nearly the same performance for $\hat{s}_{T_\alpha}$ despite the fact that the constants are not so accurate and can depend on n_1.

Proposition 2: Let $P_{\mathcal{L}_1}$ denote the product distribution on $\mathcal{L}_1$ and let $\|\cdot\|_1$ be the empirical norm on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$.

For $T \preceq T_{\max}$, let $\bar{s}_T$ verify

$$\|s - \bar{s}_T\|_1^2 = \inf_{u \in S_T} \|s - u\|_1^2.$$

Let $\mathcal{S}$ denote the set of all splits used in the growing procedure and let V denote the VC-dimension of $\mathcal{S}$. Suppose that $V < +\infty$ and that $n_1 \geqslant V$. Let $\xi > 0$, $L_{n_1,V} = V\left(2\log 2 + \log\left(n_1/V\right)\right)$, and

$$\alpha_{n_1,V} = 1 + 2L_{n_1,V} + 2\sqrt{L_{n_1,V}}.$$

If $\alpha > \sigma^2 \alpha_{n_1,V}$, then there exist nonnegative constant Σ_α and C_2' such that

$$\|s - \hat{s}_{T_\alpha}\|_1^2 \leqslant C_1'(\alpha) \inf_{T \preceq T_{\max}} \left\{ \|s - \bar{s}_T\|_1^2 + \sigma^2 \alpha_{n_1,V} \frac{|T|}{n_1} \right\} + C_2' \frac{\sigma^2}{n_1} \xi$$

on a set Ω_ξ such that $P_{\mathcal{L}_1}(\Omega_\xi) \geqslant 1 - 2\Sigma_\alpha e^{-\xi}$, where $C_1'(\alpha) > 1$ and Σ_α are increasing with α.

A proof of this proposition is given in Appendix II-B.

The same conclusions as the ones of the M1 case hold in this case. Note the following.

- The penalty term takes into account the complexity of the collection of trees having fixed number of leaves which can be constructed on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$. Since this complexity is controlled via the VC-dimension V, V necessarily appears in the penalty term. It differs from Proposition 1 in the sense that the models we consider are random, so this complexity has to be taken into account to obtain an uniform bound.
- Baraud [15] can no longer be applied in this case since the size $|T_{\max}|$ of the maximal tree is not easily controlled without any assumption on the distribution μ nor on the construction of $T_{\max}$.

Example: Let us consider the case where $\mathcal{S}$ is the set of all half-spaces of $\mathcal{X} = \mathbb{R}^d$ (which is more often used in the CART algorithm). In this case, $V = d + 1$, consequently, if $n_1 > V$, we obtain a penalty of the form

$$\mathrm{pen}_n(T) = \beta \frac{\sigma^2}{n_1} |T| \left(1 + 2(d+1)\left(2\log 2 + \log \frac{n_1}{d+1}\right)\right) + 2\beta \frac{\sigma^2}{n_1} |T| \left(\sqrt{(d+1)\left(2\log 2 + \log \frac{n_1}{d+1}\right)}\right)$$

with $\beta > 1$. So, if CART provides some minimax estimator on a class of functions, the $\log n_1$ term always appears for s in this class when working in a linear space of low dimension (on a signal, for example). On the other hand, Birgé *et al.* [8] show that the risk of $\hat{s}_{T_\alpha}$ explodes if $\beta \leqslant 1$.

Thus, in both cases, the penalty of Breiman *et al.* [1] is well chosen and the pruning algorithm is valid. Theorem II-A.1 gives another important piece of information: the sequence of pruned subtrees contains all the information, so it is useless to look at all the subtrees. To select a subtree, or equivalently, a suitable temperature α, one just has to consider those that appear in the sequence.

In practice, as the suitable temperature α is unreachable, a test sample must be used to select a subtree. This particular method is examined in Section III-B.

B. Final Selection

Given the sequence $(T_k)_{1 \leqslant k \leqslant K}$ pruned from $T_{\max}$ as defined in Section II-A2, let us recall that the final estimator $\tilde{s}$ provided by CART is defined by

$$\tilde{s} = \operatorname*{argmin}_{\{\hat{s}_{T_k}; 1 \leqslant k \leqslant K\}} \left[\gamma_{n_3}(\hat{s}_{T_k})\right].$$

The performance of this estimator can be compared to the performance of the subtrees $(T_k)_{1 \leqslant k \leqslant K}$ by the following.

Proposition 3: Let $\|\cdot\|_3$ denote the empirical norm on $\{X_i; (X_i, Y_i) \in \mathcal{L}_3\}$.

i) if $\tilde{s}$ is constructed via M1

$$\mathbb{E}\left[\|s - \tilde{s}\|_3^2 \mid \mathcal{L}_1, \mathcal{L}_2\right] \leqslant C \inf_{1 \leqslant k \leqslant K} \mathbb{E}\left[\|s - \hat{s}_{T_k}\|_3^2 \mid \mathcal{L}_1, \mathcal{L}_2\right] + C' \frac{\log K}{n_3}.$$

ii) if $\tilde{s}$ is constructed via M2

$$\mathbb{E}\left[\|s - \tilde{s}\|_3^2 \mid \mathcal{L}_1\right] \leqslant C \inf_{1 \leqslant k \leqslant K} \mathbb{E}\left[\|s - \hat{s}_{T_k}\|_3^2 \mid \mathcal{L}_1\right] + C' \frac{\log K}{n_3}.$$

A proof of this proposition is given in Appendix II-C.

Note that under condition (6) and using truncation, if $n_1 \leqslant n_3$, the results of Baraud [15] can also be applied in both cases, and the same inequality holds under the norm $\|\cdot\|$ on a large probability set. Let us also remark that the results of Wegkamp [14] can be applied here since the number of models is small. Nevertheless, since the different norms cannot be compared easily, these results cannot be connected to the results on the pruning procedure.

We can now conclude the following.

- Except for the first trees of the sequence $(T_k)_{1 \leqslant k \leqslant K}$ for which $\alpha_k \leqslant \sigma^2 \alpha_{n_1,V}$, all the other trees have conditional risks controlled by the infimum of the errors that can be made on all the subtrees pruned from $T_{\max}$.
- The conditional risk of the final estimator $\tilde{s}$ with respect to $\|\cdot\|_3$ is controlled by the infimum of the errors that can be made on the subtrees of the sequence $(T_k)_{1 \leqslant k \leqslant K}$.
- The discrete selection adds a term of order $\log(n_1)/n_3$, which is at worst of the same order as the penalty. Thus, it does not alter dramatically the accuracy of the estimation.

In addition, if CART provides a collection of models S_T such that

- the maximal dimension of the models is $D_N = \mathrm{o}(N/\log N)$;
- the approximation properties of the models are convenient enough to ensure that the bias tends to zero with increasing sample size N,

then the upper bound of the risk tends to zero with N, providing a result of consistency for $\tilde{s}$.

Consequently, if we take the pruning and selection procedures separately, each of them has a convenient behavior. Nevertheless, having $\alpha_{n_1,V}$ and $\alpha > \sigma^2\alpha_{n_1,V}$ could permit, via Theorem II-A.1, to choose a model without $\mathcal{L}_3$. In that case, a general bound could be established for the final estimator.

IV. Bounded Regression

In this section, we consider the bounded regression framework, where, for a given $i \in \{1;\ldots;N\}$, $|Y_i| \leqslant 1$ and ε_i is an unknown bounded noise, centered conditionally to X_i. The three following subsections yield about the same results as those of Section III.

A. Validation of the Pruning Algorithm

We will follow exactly the same lines and use the same notation as in Section III-A. All the remarks made in the Gaussian case on the way each model selection is made are still valid in this case.

1) $\tilde{s}$ Constructed Via M1: We have the following upper bound for the risk of $\hat{s}_{T_\alpha}$ conditionally to $\mathcal{L}_1$ and $\mathcal{L}_2$.

Proposition 4: Let $P_{\mathcal{L}_2}$ be the product distribution on $\mathcal{L}_2$. Let $\xi > 0$.

There exists a nonexplicit positive constant α_0 such that, if $\alpha > \alpha_0$, then there exist some nonnegative constants Σ_α and C_2' such that

$$\|s - \hat{s}_{T_\alpha}\|^2 \leqslant C_1'(\alpha) \inf_{T \preceq T_{\max}} \left\{ \inf_{u \in S_T} \|s-u\|^2 + \frac{|T|}{n_2} \right\} + C_2' \frac{1+\xi}{n_2}$$

on a set Ω_ξ such that $P_{\mathcal{L}_2}(\Omega_\xi) \geqslant 1 - 2\Sigma_\alpha e^{-\xi}$, where $C_1'(\alpha) > \alpha_0$ and Σ_α are increasing with α.

A proof of this proposition is given in Appendix III-A.

Remark 2: The fact that we do not know anything about the noise (except that it is bounded) leads to a minimal temperature α_0 that we cannot reach.

The conclusions concerning the bounded case are the same as those of Section III-A, except that $\hat{s}_{T_\alpha}$ does not obviously satisfy an oracle inequality since the true risk is unknown, but the inequality obtained is sufficient to validate the pruning procedure.

2) $\tilde{s}$ Constructed Via M2: One gets the following upper bound for the risk of $\hat{s}_{T_\alpha}$ conditionally to $\mathcal{L}_1$.

Proposition 5: Let $P_{\mathcal{L}_1}$ denote the product distribution on $\mathcal{L}_1$ and $\|\cdot\|_1$ be the empirical norm on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$.

For $T \preceq T_{\max}$, let $\bar{s}_T$ verify $\|s - \bar{s}_T\|_1^2 = \inf_{u \in S_T} \|s-u\|_1^2$.

Let $\xi > 0$ and

$$\alpha_{n_1,V} = 1 + V\left(1 + \log\frac{n_1}{V}\right).$$

There exists a nonexplicit positive constant α_0 such that, if $\alpha > \alpha_0\alpha_{n_1,V}$, then there exist some nonnegative constants Σ_α and C_2' such that

$$\|s - \hat{s}_{T_\alpha}\|_1^2 \leqslant C_1'(\alpha) \inf_{T \preceq T_{\max}} \left\{ \|s - \bar{s}_T\|_1^2 + \alpha_{n_1,V}\frac{|T|}{n_1} \right\} + C_2' \frac{1+\xi}{n_1}$$

on a set Ω_ξ such that $P_{\mathcal{L}_1}(\Omega_\xi) \geqslant 1 - 2\Sigma_\alpha e^{-\xi}$, where $C_1'(\alpha) > \alpha_0$ and Σ_α are increasing with α.

A proof of this proposition is given in Appendix III-B.

We can conclude exactly in the same way as for the pruning validation of the Gaussian regression framework (Section III-A), except that we do not know anything about the minimal temperature to be chosen in the sequence given by the pruning algorithm. It is therefore necessary to choose a method to select the suitable subtree among the sequence. One method consists in proceeding by test sample.

B. Final Selection

In this framework, our goals are exactly the same as in the Gaussian regression one. We define the final estimator given by the CART algorithm as (3) and we analyze the behavior of $\tilde{s}$ during the final step as in the Gaussian regression case.

Proposition 6: Let $\|\cdot\|_1$ denote the empirical norm on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$.

i) If $\tilde{s}$ is constructed via M1

$$\mathbb{E}\left[\|s - \tilde{s}\|^2 \mid \mathcal{L}_1, \mathcal{L}_2\right] \leqslant C \inf_{1 \leqslant i \leqslant K} \|s - \hat{s}_{T_i}\|^2 + C' \frac{\log K}{n_3}.$$

ii) If $\tilde{s}$ is constructed via M2

$$\mathbb{E}\left[\|s - \tilde{s}\|_1^2 \mid \mathcal{L}_1\right] \leqslant C \inf_{1 \leqslant k \leqslant K} \mathbb{E}\left[\|s - \hat{s}_{T_k}\|_1^2 \mid \mathcal{L}_1\right] + C' \frac{\log K}{n_3}.$$

A proof of this proposition is given in Appendix III-C.

We obtain similar bounds for the Gaussian and bounded cases, then the conclusions concerning the performance of $\tilde{s}$ are the same for both cases.

In addition, the following result holds for bounded regression. It is a consequence of Propositions 4–6.

Theorem 1: Given for $i = 1, \ldots, N$

$$Y_i = s(X_i) + \varepsilon_i$$

with $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ and ε_i centered conditionally to X_i, we assume that the $(X_i)_{1 \leqslant i \leqslant N}$ are identically distributed with common unknown distribution μ and that $|Y_i| \leqslant 1$. We consider both methods M1 and M2. Let $\|\cdot\|$ be the $\mathbb{L}^2(\mathcal{X}, \mu)$-norm and $\|\cdot\|_1$ be the empirical norm on $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$. Let l^2 be the square distance associated with $\|\cdot\|$ if $\tilde{s}$ is constructred via M1 and with $\|\cdot\|_1$ if $\tilde{s}$ is constructed via M2.

Then there exist some nonnegative constants C_1, C_2, and C_3 such that

$$\mathbb{E}\left[l^2(\tilde{s},s)\mid\mathcal{L}_1\right]\leqslant C_1\inf_{T\preceq T_{\max}}\left[\inf_{u\in S_T} l^2(u,s)+\frac{|T|}{n}\right] +\frac{C_2}{n}+C_3\frac{\log n_1}{n_3}$$

where $n=n_2$ if $\tilde{s}$ is constructed via M1 and $n=n_1$ if $\tilde{s}$ is constructed via M2.

Proof: The proof remains the same if $\tilde{s}$ is constructed either via M1 or M2. So we just give the proof for the M1 method.

Actually, since we have at most one model per dimension in the pruned subtree sequence, it suffices to note that $K\leqslant n_1$. Then let α_0 be the minimal constant given by Proposition 4. Hence, since for a given $\alpha>0, T_\alpha$ belongs to the sequence $(T_k)_{1\leqslant k\leqslant K}$

$$\mathbb{E}\left[l^2(\tilde{s},s)\mid\mathcal{L}_1,\ \mathcal{L}_2\right]\leqslant C\inf_{\alpha>\alpha_0} l^2(\hat{s}_{T_\alpha},s)+C'\frac{\log K}{n_3}.$$

Then, by using Proposition 4 with $\alpha=2\alpha_0$ and by taking the expectation according to $\mathcal{L}_2$, we obtain Theorem 1. □

V. Open Questions

We can conclude that pruning a maximal tree is a convenient algorithm in terms of model selection for the two regression contexts mentioned above. But two questions remain: first, "how to choose a convenient tree in the pruned sequence ?" The method we studied in this paper gives positive results, but could it be possible to remove the third (or second) subsample in order to obtain a better upper bound for the risk of $\tilde{s}$? Actually, considering the different results we obtained, if we had the true constant α occurring in the penalty, we would only have to take, in the sequence, the subtree T_k such that $\alpha_k\leqslant\alpha<\alpha_{k+1}$. Then the last term in the upper bound for the risk could be removed. But in theory this α is unreachable since it depends on too many unknown parameters, such as noise variance σ^2. We only have a minimal constant, which can be interpreted as follows: when the temperature increases, the number of leaves decreases. But it follows from Propositions 1, 2, 4, and 5 that a "good" subtree is associated with a large enough temperature. Consequently, a jump in the number of leaves could occur when the temperature becomes higher than the minimal constant. At this stage, we hope that the "good" subtree is above this temperature. An answer could be to extract from the data the right temperature for the penalized criterion. So far, there exists no general method to do this, but there are some heuristic ones based on the theoretical results of Birgé and Massart [8] and simulations (see Gey and Lebarbier [16], for example).

Second, "how to analyze the approximation quality of CART to obtain an upper bound for the complete risk ?" Nobel, Olshen [5] and Nobel [6] give some asymptotic results on recursive partitioning. Engel [10] and Donoho [11] obtain some upper bounds for the risk of the penalized estimator in the particular construction obtained via a recursive partitioning on a fixed dyadic grid. But we lack approximation results concerning CART as introduced by Breiman *et al.* [1]. This aspect of the problem remains to be analyzed.

Appendix I

A. Local Bound for Some Empirical Processes

Let $(X,Y)\in\mathcal{X}\times[-1,1]$ be defined as $Y=s(X)+\varepsilon$, where s takes values in $[-1,1]$ and ε is a noise centered conditionally to X and bounded by 1. Let $\{(X_1,Y_1);\ldots;(X_n,Y_n)\}$ be an n-sample of (X,Y). Let μ_n denote the empirical distribution on $X_1^n=(X_i)_{1\leqslant i\leqslant n}$ and $\|\cdot\|_1$ denote the empirical norm on X_1^n. Then, given z and u in $\mathbb{L}^2(\mu_n)$, define

$$d^2(z,u)=16\|z-u\|_1^2.$$

For any tree T constructed on X_1^n, define S_T as the set of all piecewise-constant functions bounded by 1 defined on the partition associated with the leaves of T. Then, for any $u\in S_T$ and any $\sigma>0$, define

$$\begin{aligned}B_T(u,\sigma)&=\{z\in S_T; d(u,z)\leqslant\sigma\}\\&=\{z\in S_T;\|u-z\|_1\leqslant\sigma/4\}.\end{aligned}$$

Finally, for $z\in\mathbb{L}^2(\mathcal{X})$, define the centered empirical quadratic contrast of z by

$$\bar{\gamma}_n(z)=\gamma_n(z)-\mathbb{E}\left[\gamma_n(z)\mid X_1^n\right] \quad (7)$$

where γ_n is defined for any given $z\in\mathbb{L}^2(\mathcal{X},\mu)$ by

$$\gamma_n(z)=\frac{1}{n}\sum_{i=1}^n\left(Y_i-z(X_i)\right)^2.$$

Remark 3: If γ_n is evaluated on a sample (X_i') independent of X_1^n, it is easy to check that the bounds we obtain in what follows are still valid by taking the marginal distribution μ of X instead of μ_n, and the distance d associated with the $\mathbb{L}^2(\mathcal{X},\mu)$-norm instead of the empirical norm $\|\cdot\|_1$.

Then we have the following result.

Lemma 1: For any $u\in S_T$ and any $\sigma>0$

$$\mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}|\bar{\gamma}_n(z)-\bar{\gamma}_n(u)|\mid X_1^n\right]\leqslant(7/2)\,\sigma\sqrt{\frac{|T|}{n}}.$$

Proof: We have

$$\bar{\gamma}_n(z)-\bar{\gamma}_n(u)=\frac{2}{n}\sum_{i=1}^n\epsilon_i(z(X_i)-u(X_i)) +2\nu_n((s-u)(z-u))-\nu_n((z-u)^2)$$

where ν_n is the recentered empirical measure. So we have three terms to study, that we simply denote by

- $A_1=\mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}\left|\frac{2}{n}\sum_{i=1}^n\epsilon_i(z(X_i)-u(X_i))\right|\mid X_1^n\right]$;
- $A_2=\mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}\left|2\nu_n\left((s-u)(z-u)\right)\right|\mid X_1^n\right]$;
- $A_3=\mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}\left|\nu_n\left((z-u)^2\right)\right|\mid X_1^n\right]$.

Then we fix an orthonormal basis of S_T denoted by $(\varphi_l)_{l\in\widetilde{T}}$ adapted to $\widetilde{T}$ (i.e., some normalized characteristic functions), and we have

$$B_T(u,\sigma)=\left\{z\in S_T; z-u=\sum_{l\in\widetilde{T}}a_l\varphi_l\,,\ \sum_{l\in\widetilde{T}}a_l^2\leqslant\frac{\sigma^2}{16}\right\}. \quad (8)$$

We will now bound each A_i, $i=1,2,3$.

Upper bound for A_1:

Using the Cauchy–Schwarz inequality, for any z in $B_T(u,\sigma)$ such that $z-u=\sum_{l\in\widetilde{T}} a_l\varphi_l$, we get

$$\left|\sum_{i=1}^n \epsilon_i(z(X_i)-u(X_i))\right| \leqslant \sqrt{\sum_{l\in\widetilde{T}} a_l^2}\sqrt{\sum_{l\in\widetilde{T}}\left(\sum_{i=1}^n \epsilon_i\varphi_l(X_i)\right)^2}$$

$$\leqslant \frac{\sigma}{4}\sqrt{\sum_{l\in\widetilde{T}}\left(\sum_{i=1}^n \epsilon_i\varphi_l(X_i)\right)^2}.$$

Since the ϵ_i are centered random variables bounded by 1, Jensen's inequality implies

$$A_1 \leqslant \frac{\sigma}{2n}\sqrt{\sum_{l\in\widetilde{T}}\sum_{i=1}^n \mathbb{E}(\epsilon_i^2 \mid X_1^n)\varphi_l(X_i)^2} \leqslant \frac{\sigma}{2}\sqrt{\frac{|T|}{n}}.$$

Upper bound for A_2.

Given independent random signs $(\zeta_1,\zeta_2,\ldots)$, independent from $(X_1,..,X_n)$, for any z in $B_T(u,\sigma)$ let

$$\psi(z)=\sum_{i=1}^n \zeta_i\left(s(X_i)-u(X_i)\right)\left(z(X_i)-u(X_i)\right).$$

By a symmetrization argument (see [17] for more details about symmetrization arguments) one has

$$A_2 \leqslant \frac{4}{n}\mathbb{E}\left[\sup_{z\in B_T(u,\sigma)} |\psi(z)| \mid X_1^n\right].$$

For all i we have $|s(X_i)-u(X_i)|\leqslant 1$ and, if $z\in B_T(u,\sigma)$

$$\|u-z\|_1=\sqrt{\sum_{l\in\widetilde{T}} a_l^2}\leqslant \sigma/4$$

(8). So, using the Cauchy–Schwarz inequality, we have

$$A_2 \leqslant \frac{\sigma}{n}\sqrt{\sum_{l\in\widetilde{T}}\sum_{i=1}^n \mathbb{E}(\zeta_i^2)\varphi_l(X_i)^2}.$$

We can remark that the upper bound of A_2 is, up to a factor 2, the same as the upper bound of A_1 and we can conclude

$$A_2 \leqslant \sigma\sqrt{\frac{|T|}{n}}$$

upper bound for A_3.

Given independent random signs $(\zeta_1,\zeta_2,\ldots)$, independent from $(X_1,\ldots,X_n)$, one has by a symmetrization argument

$$A_3 \leqslant \mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}\left|\frac{2}{n}\sum_{i=1}^n \zeta_i(z(X_i)-u(X_i))^2\right| \mid X_1^n\right].$$

We now consider the contraction θ defined by $\theta(x)=(x^2\wedge 1)/2$. Then, since $|(z(X_i)-u(X_i))|\leqslant 1$, we have

$$A_3 \leqslant \mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}\left|\frac{2}{n}\sum_{i=1}^n \zeta_i\theta(z(X_i)-u(X_i))\right| \mid X_1^n\right].$$

We can now use a contraction inequality established by Ledoux and Talagrand ([17, Lemma 6.3 and Theorem 4.12]) and conclude

$$A_3 \leqslant \mathbb{E}\left[\sup_{z\in B_T(u,\sigma)}\left|\frac{8}{n}\sum_{i=1}^n \zeta_i(z(X_i)-u(X_i))\right| \mid X_1^n\right].$$

Finally, the upper bound is the same as the upper bound of the first term A_1 up to a constant. Then we get

$$A_3 \leqslant 2\sigma\sqrt{\frac{|T|}{n}}.$$

So, combining the three inequalities, we have

$$A_1+A_2+A_3 \leqslant (1/2+1+2)\sigma\sqrt{\frac{|T|}{n}}$$

and the lemma is proven. □

B. A Complexity Bound

Let $\mathcal{S}$ be a class of subsets of $\mathcal{X}$ and $(S_m)_{m\in\mathcal{M}_n^*}$ a collection of tree-structured models constructed on n points of $\mathcal{X}$ using $\mathcal{S}$. Then we have the following.

Lemma 2: Let V denote the VC-dimension of $\mathcal{S}$ and suppose $n\geqslant V$. Let $D\in\mathbb{N}^*$ and, for $m\in\mathcal{M}_n^*$, $D_m=\mathrm{Dim}(S_m)$. Then

$$|\{m\in\mathcal{M}_n^*; D_m=D\}|\leqslant\left(\frac{ne}{V}\right)^{DV}.$$

Proof: Let $\{x_1;\ldots;x_n\}\in\mathcal{X}^n$. We want to bound uniformly in $\{x_1;\ldots;x_n\}$ the number of ways to construct a tree having D leaves on these n points. Then we will have Lemma 2.

Let D be some positive integer. For a tree-structured model S_m, D_m is the number of leaves of S_m. Thus, a D-dimensional model is a tree having D leaves. For such a tree, there are $D-1$ nonterminal nodes, which implies that there are $D-1$ splits.

To prove the lemma, we use Sauer's lemma that gives a relationship between the different ways to split r points of $\mathcal{X}$ in two parts using $\mathcal{S}$, and the VC-dimension V of $\mathcal{S}$.

For A, a subset of $\mathcal{X}$, we define $A\cap\mathcal{S}=\{A\cap S; S\in\mathcal{S}\}$ and $\Delta(A)=|A\cap\mathcal{S}|$. Then, for any integer r, we define $m(r)=\max\{\Delta(A); A\subset\mathcal{X}, |A|=r\}$. Consequently, the number of ways to cut $\{x_1;\ldots;x_n\}$ in two parts using $\mathcal{S}$ is at most $m(n)$. But one has by Sauer's lemma that, for any integer r

$$m(r)\leqslant\sum_{j=0}^V\binom{r}{j}.$$

Thus, we obtain

$$|\{m\in\mathcal{M}_n^*\,;\,D_m=D\}|\leqslant\left(\sum_{j=0}^V\binom{n}{j}\right)^D.$$

Moreover, since $n\geqslant V$

$$\sum_{j=0}^V\binom{n}{j}\leqslant\sum_{j=0}^V\frac{n^j}{j!}\leqslant\left(\frac{en}{V}\right)^V$$

and the proof is achieved. □

Appendix II

In the following sections, we denote by $\mathcal{T}$ the set of all subtrees pruned from $T_{\max}$ and consider μ, P, and $\|\cdot\|$ as defined in Section II-B.

A. Proof of Proposition 1

We use the result established by Birgé and Massart in [8, Theorem 2] on Gaussian model selection. For the sake of completeness, let us recall this result.

Let $n = n_2$ and $\|\cdot\|_n$ be the empirical norm associated with the empirical distribution μ_n on the grid $(X_1, \dots, X_n)$. Let us give a collection of linear deterministic models $(\mathcal{S}_m)_{m\in\mathcal{M}_n}$, a penalty function $\mathrm{pen}_n : \mathcal{M}_n \longrightarrow \mathbb{R}_+$, and a sample $\mathcal{L}_2$ from the random variable (X, Y) defined as in Section III. Let $\bar{s}_m$ denote the $\mathbb{L}^2(\mu_n)$-projection of s on the model $\mathcal{S}_m$, and D_m denote the dimension of $\mathcal{S}_m$. Let $\hat{m}$ be defined by

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n} \left[\gamma(\hat{s}_m) + \mathrm{pen}_n(m)\right]$$

where $\hat{s}_m$ is the minimum contrast estimator of s on $\mathcal{S}_m$. Then one gets the following.

Theorem VII-A.1 (Birgé, Massart): Let $\xi > 0$, $\eta \in]0, 1[$, $K > 2 - \eta$, and $(L_m)_{m\in\mathcal{M}_n}$ be a family of weights such that

$$\Sigma = \sum_{m\in\mathcal{M}_n} e^{-L_m D_m} < +\infty.$$

If

$$\mathrm{pen}_n(m) \geq \sigma^2 \frac{D_m}{n} \left(K + 2(2-\eta)L_m + \frac{2}{\eta}\sqrt{L_m} \right)$$

then

$$\begin{aligned} \mathbb{E}\left[\|s - \hat{s}_{\hat{m}}\|_n^2\right] &\leqslant C(K, \eta) \\ &\times \inf_{m\in\mathcal{M}_n} \left\{ \mathbb{E}\left[\|s - \bar{s}_m\|_n^2\right] + \mathrm{pen}_n(m) \right\} + \sigma^2 C'(K, \eta)\frac{\Sigma}{n}. \end{aligned}$$

The collection of models we consider is $\mathcal{T}$ (depending on the first subsample) and is deterministic conditionnally to $\mathcal{L}_1$. For $T \in \mathcal{T}$, the model considered is

$$S_T = \left\{ \sum_{t\in\widetilde{T}} a_t \mathbb{1}_t ; \forall\, t \in \widetilde{T}\ a_t \in \mathbb{R} \right\}$$

and its dimension is $|T|$. Given this collection, to apply the result of Birgé and Massart, we need to choose a convenient family of weights $(L_T)_{T\preceq T_{\max}}$.

Taking L as a function of the dimension, one has

$$\begin{aligned} \Sigma &= \sum_{T\preceq T_{\max}} e^{-L_T |T|} \\ &\leqslant \sum_{D\geq 1} |\{T \preceq T_{\max}\ ;\ |T| = D\}| e^{-L(D)D}. \end{aligned}$$

Furthermore, for any given dimension D, the number of balanced binary trees having D final nodes is the Catalan number $(1/D)\binom{2(D-1)}{D-1}$. Thus, we have

$$\begin{aligned} \Sigma &\leqslant \sum_{D\geq 1} \frac{1}{D} \binom{2(D-1)}{D-1} e^{-L(D)D} \\ &\leqslant \sum_{D\geq 1} \frac{1}{D} \exp\left[(2\log 2 - L(D))D\right]. \end{aligned}$$

Taking $L(D) = \theta$, with $\theta > 2\log 2$ independent of D, we immediately obtain $\Sigma_\alpha = \Sigma_\theta < +\infty$. Then we get Proposition 1 by [8, Theorem 2].

B. Proof of Proposition 2

Let us denote by X_1^n the sample $(X_1, \dots, X_n)$ of size n of the random variable (X, Y) and by μ_n the empirical distribution on X_1^n.

First we generalize Theorem VII-A.1 to random models, and then we apply it to CART.

Theorem 2: Let $n \in \mathbb{N}$. Consider the Gaussian regression model defined in Section III. Then take an n-sample of the random variable (X, Y) and $(S_m)_{m\in\mathcal{M}_n^*}$ a countable family of finite-dimensional linear subspaces with respective dimensions D_m constructed on the grid X_1^n. Fix $(L_m)_{m\in\mathcal{M}_n^*}$ a family of weights satisfying the condition

$$\Sigma = \sum_{m\in\mathcal{M}_n^*,\, D_m>0} e^{-L_m D_m} < +\infty$$

where Σ is deterministic.

Given a subspace $\mathcal{M}_n \subset \mathcal{M}_n^*$ that can also depend on $(Y_1, \dots, Y_n)$, we select the estimators as follows:

- $\hat{s}_m = \operatorname*{argmin}_{t\in S_m} [\gamma_n(t)]$;
- $\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n} [\gamma_n(\hat{s}_m) + \mathrm{pen}_n(m)]$ and then $\tilde{s} = \hat{s}_{\hat{m}}$.

Let $\eta \in]0, 1[$ and $K > 2 - \eta$. Let us consider a penalty function on $\mathcal{M}_n^*$ such that

$$\mathrm{pen}_n(m) \geqslant \frac{\sigma^2}{n} D_m \left(K + 2(2-\eta)L_m + \frac{2}{\eta}\sqrt{L_m} \right)$$

for all $m \in \mathcal{M}_n^*$. Let $\xi > 0$, $\|\cdot\|_n$ the empirical norm on X_1^n, and $s_m = \operatorname*{argmin}_{u\in S_m} \|s - u\|_n$.

Then the penalized estimator satisfies, for all $m \in \mathcal{M}_n$

$$\begin{aligned} \|s - \tilde{s}\|_n^2 &\leqslant C_1(K, \eta)\ \left\{ \|s - s_m\|_n^2 + \mathrm{pen}_n(m) \right\} \\ &\quad + C_2(K, \eta)\ \frac{\sigma^2}{n}\xi \end{aligned}$$

on a set Ω_ξ such that $P(\Omega_\xi) \geqslant 1 - 2\Sigma e^{-\xi}$ and for suitable constants $C_1(K, \eta)$ and $C_2(K, \eta)$.

Proof: We follow exactly the same lines as in [8], the only difference being that all our upper bounds are obtained by conditioning with respect to X_1^n, so we skip the proof. Note that the result holds on a set Ω_ξ having probability measure P unconditional to X_1^n. This is due to the fact that Σ is deterministic and does not depend on X_1^n. □

Application to tree partitions:

In that case, we have $n = n_1$. We consider $\mathcal{M}_n = \mathcal{T}$ and we take $\mathcal{M}_n^*$ as all the tree-structured partitions constructed on the grid X_1^n using $\mathcal{S}$. Taking Theorem 2 into account with $n = n_1$, it suffices to choose the weights $(L_m)_{m\in\mathcal{M}_n^*}$ to obtain Proposition 2.

Taking the weights as a function of the dimension, we have by Lemma 2

$$\Sigma \leqslant \sum_{D\geqslant 1} \exp\left(-L_D D + DV + DV \log \frac{n_1}{V}\right).$$

Then, we take $L_D = V(\theta + \log(n_1/V))$, with $\theta > 1$ and we obtain Proposition VII-B.

C. Proof of Proposition 3

Let us call $n = n_3$. Then let us note that, for $u \in \mathbb{L}^2(\mu_n)$

$$\|s-u\|_n^2 = \mathbb{E}\left[\gamma_n(u) - \gamma_n(s) \mid X_1^n\right]. \tag{9}$$

Since this equality depends only on $\mathcal{L}_3$, the same proof can be achieved for M1 as for M2, the only difference being in the conditioning which depends on $\mathcal{L}_1$ and $\mathcal{L}_2$ for M1 and only on $\mathcal{L}_1$ for M2. Consequently, we just give the proof for the M1 method.

Let $k \in \{1, \ldots, K\}$ and take $\bar{\gamma}_n$ as (7).

Then we have by (9)

$$\begin{aligned}\|s-\tilde{s}\|_n^2 &\leqslant \|s-\hat{s}_{T_k}\|_n^2 + \bar{\gamma}_n(\hat{s}_{T_k}) - \bar{\gamma}_n(\tilde{s}) \\ &\leqslant \|s-\hat{s}_{T_k}\|_n^2 + 2\frac{\sigma}{\sqrt{n}}(Z(\tilde{s}) - Z(\hat{s}_{T_k}))\end{aligned}$$

where

$$Z_{k,j} = \left(Z(\hat{s}_{T_j}) - Z(\hat{s}_{T_k})\right)/\|\hat{s}_{T_j} - \hat{s}_{T_k}\|_n$$

is $\mathcal{N}(0,1)$-distributed, knowing subsamples $\mathcal{L}_1$ and $\mathcal{L}_2$.

The general principle is to use the fact that $Z_{k,j}$ is a Gaussian variable to bound it uniformly in k, j. The result will be an in-probability uniform upper bound for $Z_{k,j}$ that will be integrated to obtain Proposition 3.

Since $Z_{k,j}$ is a Gaussian variable conditionally to $\mathcal{L}_1$ and $\mathcal{L}_2$, for all $x \in \mathbb{R}$ we have

$$P\left[Z_{k,j} \geqslant x \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant e^{-x^2/2}.$$

Taking $\xi > 0$ and setting $x = \sqrt{2(\log K + \xi)}$ we get

$$P\left[Z_{k,j} \geqslant \sqrt{2(\log K + \xi)} \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant \exp(-\log K - \xi).$$

Thus,

$$P\left[\sup_{1\leqslant j\leqslant K} Z_{k,j} \geqslant \sqrt{2(\log K + \xi)} \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant e^{-\xi}$$

and

$$\sup_{1\leqslant j\leqslant K}\left\{\left(Z(\hat{s}_{T_j}) - Z(\hat{s}_{T_k})\right)/\|\hat{s}_{T_j} - \hat{s}_{T_k}\|_n\right\} \geqslant \sqrt{2(\log K + \xi)}$$

on a set Ω_ξ such that $P(\Omega_\xi \mid \mathcal{L}_1,\ \mathcal{L}_2) \leqslant e^{-\xi}$. So, given $0 < \eta < 1$, using the two inequalities

$$2ab \leqslant (1-\eta)a^2 + (1/(1-\eta))b^2$$

and

$$(a+b)^2 \leqslant (1+\eta)a^2 + (1+1/\eta)b^2$$

we obtain, on Ω_ξ^c

$$\begin{aligned}2\frac{\sigma}{\sqrt{n}}\left(Z(\tilde{s}) - Z(\hat{s}_{T_k})\right) \leqslant &(1-\eta^2)\|s-\tilde{s}\|_n^2 \\ &+ \left(\frac{1}{\eta} - \eta\right)\|s - \hat{s}_{T_k}\|_n^2 \\ &+ \left(\frac{1}{1-\eta}\right)\frac{\sigma^2}{n}2(\log K + \xi).\end{aligned}$$

We can now integrate the first inequality with respect to the third subsample and we obtain

$$\begin{aligned}\mathbb{E}\left[\|s-\tilde{s}\|_n^2 \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant &C_1(\eta)\|s-\hat{s}_{T_k}\|_n^2 \\ &+ C_2(\eta)\frac{\log K}{n} + C_3(\eta)\frac{2\sigma^2}{n}.\end{aligned}$$

This yields Proposition 3.

Appendix III

A. Proof of Proposition 4

We apply the result established by Massart [9, Theorem 4.2] on bounded regression model selection. For the sake of completeness, let us recall this result.

Let $n = n_2$. Assume $(\mathcal{S}_m)_{m\in\mathcal{M}_n}$ is a collection of models and that one has a sample $\mathcal{L}_2$ of the random variable (X, Y) defined as in Section IV-A. The contrast function is here bounded by 1, and pen_n, $\bar{s}_m$, and $\hat{m}$ are defined as in the proof of Proposition 1.

Suppose that, for $m \in \mathcal{M}_n$

$$\mathcal{S}_m = \left\{\sum_{t\in C} a_t \mathbb{1}_t\ ;\ C \in \mathcal{C}_m,\ (a_t) \in [0,1]^{|C|}\right\}$$

where each $\mathcal{C}_m$ is a countable class of subsets of $\mathbb{R}$. Then we define $D_m = \mathrm{Dim}(\mathcal{S}_m)$ as the dimension of the linear space associated with $\mathcal{S}_m$. One makes the following assumptions:

$\mathbf{H_1}$: There exist some pseudodistance d and some absolute constant c such that for every $(t,u) \in (\mathbb{L}^2(\mu))^2$, one has $\mathrm{Var}_s[\gamma(t,X) - \gamma(u,X)] \leqslant d^2(u,t)$, and particularly

$$\mathrm{Var}_s[\gamma(t,X) - \gamma(s,X)] \leqslant d^2(s,t) \leqslant c\|s-t\|^2$$

$\mathbf{H_2}$: For any positive σ and for any $u \in \mathcal{S}_m$, let us define

$$B_m(u,\sigma) = \{t \in \mathcal{S}_m; d(u,t) \leqslant \sigma\}$$

where d is given by assumption $\mathbf{H_1}$.

We now assume that for any $m \in \mathcal{M}_n$, there exists some continuous function ϕ_m mapping $\mathbb{R}_+$ onto $\mathbb{R}_+$ such that $\phi_m(0) = 0$, $\phi_m(x)/x$ is nonincreasing and

$$\mathbb{E}\left[\sup_{t\in B_m(u,\sigma)}|\bar{\gamma}_n(t) - \bar{\gamma}_n(u)|\right] \leqslant \phi_m(\sigma)$$

for all $\sigma \geqslant \sigma_m$, where σ_m is the solution of the equation $\phi_m(x) = x^2$, $x > 0$. One gets the following result.

Theorem VIII-A.1 (Massart): Let $\xi > 0$. Let K_1 and K_2 be two constants, with $K_i > 0$, $i = 1, 2$. Take a family of weights $(x_m)_{m \in \mathcal{M}_n}$ such that $\Sigma = \sum_{m \in \mathcal{M}_n} e^{-x_m} < +\infty$. Then, for some nonnegative constant K_3, for every $m \in \mathcal{M}_n$, and every positive ξ, if

$$\mathrm{pen}_n(m) \geqslant K_1 \sigma_m^2 + K_2(x_m/n) - K_3(\xi/n)$$

with probability larger than $1 - \exp(-x_m - \xi)$

$$\|s - \hat{s}_{\hat{m}}\|^2 \leqslant C(K_1, K_2) \inf_{m \in \mathcal{M}_n} \left[\|s - \bar{s}_m\|^2 + \mathbb{E}[\mathrm{pen}_n(m)]\right] + C'(K_1, K_2)\frac{\xi + 1}{n}$$

on a set Ω_ξ such that $P_{\mathcal{L}_2}(\Omega_\xi) \geqslant 1 - 2\Sigma e^{-\xi}$.

Here again the collection of models we consider is $\mathcal{T}$ and, for $T \in \mathcal{T}$

$$S_T = \left\{ \sum_{t \in \widetilde{T}} a_t \mathbb{1}_t;\ \forall\, t \in \widetilde{T}\ a_t \leqslant 1 \right\}$$

and its dimension is $|T|$. Given this collection, to apply [9, Theorem 4.2] we first choose the pseudodistance d defined on $\mathbb{L}^2(\mathcal{X}, \mu)$ in the following way: since Y is bounded by 1, for all u and t in $\mathbb{L}^2(\mathcal{X}, \mu)$ we have

$$\mathrm{Var}\left[(Y - u(X))^2 - (Y - t(X))^2\right] \leqslant 16\,\|u - t\|^2. \tag{10}$$

Then, given $T \in \mathcal{T}$, by Lemma 1 we have

$$\mathbb{E}\left[\sup_{t \in B_T(u,\sigma)} |\bar{\gamma}_n(t) - \bar{\gamma}_n(u)| \mid \mathcal{L}_1\right] \leqslant \frac{7}{2}\sigma\sqrt{|T|/n}$$

where, for $u \in S_T$

$$B_T(u, \sigma) = \{t \in S_T; \|t - u\| \leqslant \sigma/2\}.$$

Hence, the solution of the equation $\sigma^2 = (7/2)\sigma\sqrt{|T|/n}$ is $\sigma_T = (7/2)\sqrt{|T|/n}$.

The last step consists in choosing the sequence of weights $(x_T)_{T \preceq T_{\max}}$ such that the family $(e^{-x_T})_{T \preceq T_{\max}}$ is summable. Considering the same argument as in the proof of Proposition 1 and taking x as a function of the dimension, we choose $x(D) = \theta D$, with $\theta > 2\log 2$ independent of D.

Thus, we get Proposition 4 by [9, Theorem 4.2].

B. Proof of Proposition 5

In what follows, we denote by X_1^n the sample $(X_1, \ldots, X_n)$ of size n of the random variable X and by μ_n the empirical distribution on X_1^n.

First, we generalize Theorem VIII-A.1 to random models, and then we apply it to CART.

Theorem 3: Consider the bounded regression model defined in Section IV and $(S_m)_{m \in \mathcal{M}_n^*}$ a countable random family of finite-dimensional subspaces constructed on the X_1^n with respective dimensions D_m. Fix $(x_m)_{m \in \mathcal{M}_n^*}$ a family of weights satisfying the condition

$$\sum_{m \in \mathcal{M}_n^*,\ D_m > 0} e^{-x_m} \leqslant \Sigma$$

with Σ deterministic.

Given a subspace $\mathcal{M}_n \subset \mathcal{M}_n^*$ that can also depend on $(Y_1, \ldots, Y_n)$, we select the estimators as follows:

- $\hat{s}_m = \mathrm{argmin}_{t \in S_m}[\gamma_n(t)]$;
- $\hat{m} = \mathrm{argmin}_{m \in \mathcal{M}_n}[\gamma_n(\hat{s}_m) + \mathrm{pen}_n(m)]$ and then $\tilde{s} = \hat{s}_{\hat{m}}$.

Moreover, we make the following assumptions.

H1 : The contrast is bounded by some constant b.

H2 : Let $Z_i = (X_i, Y_i)$. There exists a nonnegative constant c_1 such that, for all t and u in $\mathbb{L}^2(\mathcal{X}, \mu_n)$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\left[\gamma(t, Z_i) - \gamma(u, Z_i) \mid X_1^n\right] \leqslant c_1 \|t - u\|_n^2$$

almost surely, where $\|\cdot\|_n$ is the empirical norm on X_1^n.

H3 : Take $\bar{\gamma}_n$ as (7) and for $m \in \mathcal{M}_n^*$ define

$$B_m(u, \sigma) = \{t \in S_m\ ;\ \sqrt{c_1}\|t - u\|_n \leqslant \sigma\}.$$

Then for all $m \in \mathcal{M}_n^*$, there exists some continuous function ϕ_m mapping $\mathbb{R}_+$ onto $\mathbb{R}_+$ such that $\phi_m(0) = 0$, $\phi_m(x)/x$ is nonincreasing, and

$$\mathbb{E}\left[\sup_{t \in B_m(u,\sigma)} |\bar{\gamma}_n(t) - \bar{\gamma}_n(u)| \mid X_1^n\right] \leqslant \phi_m(\sigma)$$

for all $\sigma \geqslant \sigma_m$ where σ_m is such that $\phi_m(\sigma_m) = \sigma_m^2$.

Given **H1**, **H2**, and **H3**, given $\xi > 0$, if for all $m \in \mathcal{M}_n^*$

$$\mathrm{pen}_n(m) \geqslant K_1 \sigma_m^2 + K_2 \frac{x_m}{n}$$

for some constants K_1 and K_2, then

$$\|s - \tilde{s}\|_n^2 \leqslant C_1 \inf_{m \in \mathcal{M}_n} \left\{\|s - s_m\|_n^2 + \mathrm{pen}_n(m)\right\} + C_2 \frac{1 + \xi}{n}$$

on a set Ω_ξ such that $P(\Omega_\xi) \geqslant 1 - 2\Sigma e^{-\xi}$.

Proof: Since there are just a few lines that change from the proof of [9, Theorem 4.2], we just give a sketch of proof. Note that assumption **H2** permits to give exactly the same upper bounds (except that they depend on X_1^n) for the variance as in [9]. We denote $\sqrt{c_1}\|\cdot\|_n$ by d_n.

Taking (9) into account, we have the following upper-bounding:

$$\|s - \tilde{s}\|_n^2 \leqslant \|s - s_m\|_n^2 + w_{\hat{m},m}(\tilde{s})V_{\hat{m},m} \tag{11}$$

$$+ \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}) \tag{12}$$

where for m' and M in $\mathcal{M}_n^*$

$$w_{m',M}(t) = (d_n(s, s_M) + d_n(s, s_{m'}))^2 + (y_{m'} + y_M)^2$$

$$V_{m',M} = \sup_{t \in S_{m'}} \left[\frac{|\bar{\gamma}_n(t) - \bar{\gamma}_n(s_M)|}{w_{m',M}(t)}\right]$$

with $y_{m'} \geqslant \sigma_{m'}$ and $y_M \geqslant \sigma_M$ to be chosen later.

Since the noise is unknown, we take $V_{\hat{m},m}$ to ensure that we have a bounded term that can be locally controlled. Then the principle will be to bound $V_{m',M}$ uniformly in m', M in order to offset the penalty term $\mathrm{pen}_n(\hat{m})$. This will be done by concentrating $V_{m',M}$ around its expectation uniformly in m', M. A uniform in-probability upper bound will be obtained and the weights $y_{m'}$ and y_M will be chosen to offset $\mathrm{pen}_n(\hat{m})$ in such a way that only $\|s - s_m\|_n^2 + K\, w_{\hat{m},m}(\tilde{s}) + \mathrm{pen}_n(m)$ remains in the upper bound of (11) on a large probability set. Let us notice that this set will be unconditional to X_1^n because Σ is deterministic by assumption.

We control $V_{m',M}$ for all possible values of m' and M in $\mathcal{M}_n^*$ by using Talagrand's inequality for empirical processes. Since $\mathbb{E}\left[V_{m',M} \mid X_1^n\right]$ is involved in this inequality, we control it by using assumption **H3**. Indeed, considering the same arguments as in [9], we have

$$\mathbb{E}\left[V_{m',M} \mid X_1^n\right] \leqslant 4\frac{\phi_{m'}(3y_{m'}+3y_M)}{(y_{m'}+y_M)^2} + (y_{m'}+y_M)^{-1} n^{-1/2}.$$

Hence, using the monoticity assumption on $\phi_{m'}(x)/x$, since $y_{m'} + y_M \geqslant y_{m'} \geqslant \sigma_{m'}$ and $\sigma_M > 0$, we get by definition of $\sigma_{m'}$

$$4\frac{\phi_{m'}(3y_{m'} + 3y_M)}{(y_{m'} + y_M)^2} \leqslant 12\frac{\phi_{m'}(\sigma_{m'})}{(y_{m'} + y_M)\sigma_{m'}} \leqslant 12\frac{\sigma_{m'} + \sigma_M}{y_{m'} + y_M}.$$

Then, we finally have

$$\mathbb{E}\left[V_{m',M} \mid X_1^n\right] \leqslant (y_{m'} + y_M)^{-1}\left[12(\sigma_{m'} + \sigma_M) + n^{-1/2}\right].$$

Hence Talagrand's inequality leads, for $\xi > 0$ and appropriate constants κ_1 and κ_2, for all m' and M in $\mathcal{M}_n^*$, to

$$\begin{aligned} V_{m',M} \leqslant & \frac{\kappa_1}{y_{m'} + y_M}\left[12\sigma_{m'} + \frac{n^{-1/2}}{2} + 12\sigma_M + \frac{n^{-1/2}}{2}\right] \\ & + \frac{\kappa_2}{y_{m'} + y_M}\left[\left(\sqrt{\frac{x_{m'} + \xi/2}{4n}} + \sqrt{\frac{x_M + \xi/2}{4n}}\right)\right] \\ & + \frac{\kappa_2 b}{y_{m'}^2 + y_M^2}\left[\frac{x_{m'} + x_M + \xi}{n}\right] \end{aligned}$$

on an event $\widetilde{\Omega}_\xi$ such that $P(\widetilde{\Omega}_\xi \mid X_1^n) \geqslant 1 - 2\Sigma e^{-\xi}$. Then, since Σ is deterministic, we have $P(\widetilde{\Omega}_\xi) \geqslant 1 - 2\Sigma e^{-\xi}$.

Hence, if we define for all $m' \in \mathcal{M}_n^*$

$$\begin{aligned} y_{m'} = 2K & \left[\kappa_1\left(12\sigma_{m'} + \frac{n^{-1/2}}{2}\right) + \kappa_2\sqrt{\frac{x_{m'} + \xi/2}{4n}}\right] \\ & + 2K\left[\sqrt{\kappa_2 b\frac{x_{m'} + \xi/2}{n}}\right] \end{aligned}$$

so that on $\widetilde{\Omega}_\xi$, one has $V_{m',M} \leqslant 1/K$ for all m' and M in $\mathcal{M}_n^*$, we derive from (11) that

$$\|s - \tilde{s}\|_n^2 \leqslant \|s - s_m\|_n^2 + w_{\hat{m},m}(\tilde{s})K^{-1} + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}).$$

Thus, using the same technique as in [9] and assumption **H2**, and taking $\Omega_\xi = \widetilde{\Omega}_\xi \cap \Omega_n$, the proof is achieved. □

Application to the risk bound of the CART estimator

In that case, we have $n = n_1$, $\mathcal{M}_n = \mathcal{T}$, and $\mathcal{M}_n^*$ as the collection of all trees that can be constructed on the grid $\{X_1; \ldots; X_{n_1}\}$ using $\mathcal{S}$. Taking Theorem 3 into account, we have to check assumptions **H1**, **H2**, and **H3** and then to choose the family of weights $(x_m)_{m \in \mathcal{M}_n^*}$.

Since Y is supposed to be bounded by 1 and since we consider all the functions in $\mathbb{L}^2(\mathcal{X}, \mu_n)$ also bounded by 1, the contrast is bounded by 1 and we have assumption **H1**. Then **H2** is checked with $c_1 = 16$. Furthermore, in the same manner as in the proof of Proposition 4, since Lemma 1 is still valid when working with $\|\cdot\|_n$, we have **H3** with $\phi_T(\sigma) = (7/2)\sigma\sqrt{|T|/n}$ and $\sigma_T = (7/2)\sqrt{|T|/n}$.

Finally, since Lemma 2 is true uniformly on $(x_1, \ldots, x_n)$, we choose the weights $(x_T)_{T \preceq T_{\max}}$ in the same manner as in the proof of Proposition 2 and we obtain $x_T = V\left(\theta + \log(n/V)\right)|T|$ with $\theta > 1$.

And the proof is achieved by Theorem 3.

C. *Proof of Proposition 6*

As for the proof of Proposition 3, the only difference being in the norms used, it suffices to give the proof of Proposition 6 for the M1 method. We use the same definitions and notation.

We have

$$\|s - \tilde{s}\|^2 \leqslant \|s - \hat{s}_{T_k}\|^2 + V_k w_k,$$

where

$$\begin{aligned} w_k &= (d(s, \hat{s}_{T_k}) + d(s, \tilde{s}))^2 + C^2 \\ V_k &= \frac{\bar{\gamma}_n(\hat{s}_{T_k}) - \bar{\gamma}_n(\tilde{s})}{(d(s, \hat{s}_{T_k}) + d(s, \tilde{s}))^2 + C^2} \end{aligned}$$

with $d^2(t, u) = E[(\gamma(t, \cdot) - \gamma(u, \cdot))^2]$ satisfying $d^2(t, s) \leqslant 2\,\|t - s\|^2$ (see [9]), and C a nonnegative constant we will choose later in the proof. The road map of this proof is exactly the same as the one of the proof of Proposition 5. Note that, since the collection of models considered is finished, we will use Berstein's instead of Talagrand's inequality to bound V_k uniformly in k.

Let

$$V_{k,j} = \frac{\bar{\gamma}_n(\hat{s}_{T_k}) - \bar{\gamma}_n(\hat{s}_{T_j})}{(d(s, \hat{s}_{T_k}) + d(s, \hat{s}_{T_j}))^2 + C^2}.$$

We use Bernstein's concentration inequality for centered and bounded random variables in order to bound the random variable $V_{k,j}$ uniformly on k and j to obtain an uniform upper bound for V_k. To proceed, note that

$$n\, V_{k,j} = \sum_{i=1}^{n} \frac{G_i - E(G_i)}{w_{k,j}}$$

with $-1 \leqslant G_i \leqslant 1$. Then, since for all $l \geqslant 2$

$$\left(\frac{|G_i|}{w_{k,j}}\right)^l \leqslant \left(\frac{|G_i|}{w_{k,j}}\right)^2 \left(\frac{1}{C^2}\right)^{l-2}$$

by Bernstein's inequality we obtain, for $x > 0$

$$P\left[n\, V_{k,j} \geq \sqrt{2vx} + \frac{x}{C^2} \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant e^{-x} \qquad (13)$$

where $v = \sum_{i=1}^{n} E\left[|G_i|^2/w_{k,j}^2\right]$ is bounded by

$$v = \frac{nd^2(\hat{s}_{T_k}, \hat{s}_{T_j})}{w_{k,j}^2} \leqslant \frac{n}{4C^2}.$$

Then, taking $\xi > 0$ and setting $x = \log K + \xi$ in (13), we get

$$P\left[V_{k,j} \geq \sqrt{\frac{(\log K + \xi)}{2nC^2}} + \frac{(\log n + \xi)}{nC^2} \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant \frac{e^{-\xi}}{K}.$$

Thus, except on a set Ω_ξ with probability lower than $e^{-\xi}$, we have

$$V_k \leqslant \frac{1}{C}\sqrt{\frac{(\log K + \xi)}{n}}\left(1 + \frac{1}{C}\sqrt{\frac{(\log K + \xi)}{n}}\right).$$

Then, taking $C = B\sqrt{(\log K + \xi)/n}$, where B will be chosen later, we get

$$\|s - \tilde{s}\|^2 \leqslant \|s - \hat{s}_{T_k}\|^2 + \frac{1}{B}\left(1 + \frac{1}{B}\right) w_k$$

except on Ω_ξ.

Then, using the condition satisfied by d and the inequality $(a+b)^2 \leqslant 2(a^2 + b^2)$, we obtain

$$\begin{aligned}\left(1 - \frac{4}{B}\left(1 + \frac{1}{B}\right)\right)&\|s - \tilde{s}\|^2 \\ &\leqslant \left(1 + \frac{4}{B}\left(1 + \frac{1}{B}\right)\right)\|s - \hat{s}_{T_k}\|^2 \\ &\quad + B\left(1 + \frac{1}{B}\right)\frac{(\log K + \xi)}{n}\end{aligned}$$

except on Ω_ξ.

Given $B \geqslant 5$ to ensure that $1 - (4/B)(1 + 1/B) > 0$, we finally obtain on Ω_ξ^c

$$\|s - \tilde{s}\|^2 \leqslant C_1(B)\,\|s - \hat{s}_{T_k}\|^2 + C_2(B)\,\frac{\log K}{n} + C_3(B)\,\frac{\xi}{n}.$$

Taking the expectation with respect to the third subsample, we get Proposition 6.

REFERENCES

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. London, U.K.: Chapman & Hall, 1984.

[2] P. A. Chou, T. Lookabaugh, and R. M. Gray, “Optimal pruning with applications to tree-stuctured source coding and modeling,” *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 299–315, Mar. 1989.

[3] Wernecke, Possinger, Kalb, and Stein, “Validating classification trees,” *Biometrical J.*, vol. 40, no. 8, pp. 993–1005, 1998.

[4] A. B. Nobel, “Analysis of a complexity-based pruning scheme for classification trees,” *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2362–2368, Aug. 2002.

[5] A. B. Nobel and R. A. Olshen, “Termination and continuity of greedy growing for tree-structured vector quantizers,” *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 191–205, Jan. 1996.

[6] A. B. Nobel, “Recursive partitioning to reduce distortion,” *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1122–1133, Jul. 1997.

[7] S. B. Gelfand, C. S. Ravishankar, and E. J. Delp, “An iterative growing and pruning algorithm for classification tree design,” *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 13, no. 2, pp. 163–174, Feb. 1991.

[8] L. Birgé and P. Massart, “A Generalized C_p Criterion for Gaussian Model Selection,” Université Paris, Paris, France, Tech. Rep. 647, 2001.

[9] P. Massart, “Some applications of concentration inequalities to statistics,” *Ann. Faculté des Sciences de Toulouse*, 2000.

[10] J. Engel, “A simple wavelet approach to nonparametric regression from recursive partitioning schemes,” *J. Multivariate Anal.*, vol. 49, pp. 242–254, 1994.

[11] D. L. Donoho, “CART and best-ortho-basis: A connection,” *Ann. Statist.*, vol. 25, no. 5, pp. 1870–1911, 1997.

[12] A. B. Nobel, “Histogram regression estimation using data-dependent partitions,” *Ann. Statist.*, vol. 24, no. 3, pp. 1084–1105, 1996.

[13] V. N. Vladimir, *Statistical Learning Theory*. New York: Wiley Interscience, 1998.

[14] M. Wegkamp, “Model selection in nonparametric regression,” *Ann. Statist.*, vol. 31, pp. 252–273, 2003.

[15] Y. Baraud, “Model selection for regression on a random design,” *ESAIM Probability & Statistics*, vol. 6, pp. 127–146, 2002.

[16] S. Gey and E. Lebarbier, “A CART based algorithm for detection of mutiple change points in the mean,” unpublished manuscript.

[17] Ledoux and Talagrand, *Probability in Banach Spaces (Isoperimetry and Processes)*. Berlin, Germany: Springer-Verlag, 1991.