Risk Bounds for Classification Trees under a Margin Condition

Servane Gey

Risk Bounds for Classification Trees under a Margin Condition

Servane Gey*

Abstract

Risk bounds for Classification and Regression Trees (CART, Breiman *et. al.* 1984) classifiers are obtained under a margin condition in the binary supervised classification framework. These risk bounds are obtained conditionally on the construction of the maximal deep binary tree and permit to prove that the linear penalty used in the CART pruning algorithm is valid under a margin condition. It is also shown that, conditionally on the construction of the maximal tree, the final selection by test sample does not alter dramatically the estimation accuracy of the Bayes classifier.
In the two-class classification framework, the risk bounds that are proved, obtained by using penalized model selection, validate the CART algorithm which is used in many data mining applications such as Biology, Medicine or Image Coding.

Keywords: Classification, CART, Pruning, Margin, Risk Bounds.

AMS 2000 subject classifications : primary 62G99, 62H30 *and secondary* 62-07.

1 Introduction

The main purpose of this paper is the Classification And Regression Trees (CART) method proposed by Breiman, Friedman, Olshen and Stone [9] in 1984. This method consists in constructing 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 [10], [40] for example). This kind of algorithm is often performed when the space of explanatory variables is high-dimensional. Due to its recursive computation, CART needs few computations to provide convenient classifiers, accelerating drastically the computation time when the number of variables is large. It is now widely used in the genetics framework (see [16] for example), or more generally to reduce variable dimension (see [34] [26] for example).

Let us give a short account of the CART algorithm, that will be described in a precise manner in Section 2. Given a training sample of observations, the CART algorithm consists in constructing a large dyadic recursive 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

*Laboratoire MAP5 - Université Paris Descartes, 75270 Paris Cedex 06, France. Servane.Gey@parisdescartes.fr

term is proportional to the number of leaves. This differs from the algorithm proposed by Blanchard *et. al* [5] by the fact that the first large tree is locally constructed, and not in a global way by minimizing some loss function on the whole sample.
Hence the pruning step raises the question of "why" this linear penalty is well-chosen. Gey *et. al* [17] gave an answer to this question in the regression framework, but were unable to have similar results in the classification framework. Following this previous work, this paper aims at validating the choice of the penalty in the two class classification framework. The interested reader can also find some previous discussions and results about this topic in the paper by Nobel [31].

The CART method takes place in the following general classification framework. Suppose one observes a sample $\mathcal{L}$ of N independant copies $(X_1, Y_1), \ldots, (X_N, Y_N)$ of the random variable (X, Y), where the explanatory variable X takes values in a mesurable space $\mathcal{X}$ and is associated with a label Y taking values in $\{0, 1\}$. A classifier is then any function g mapping $\mathcal{X}$ into $\{0, 1\}$ and its quality is measured by its misclassification rate

$$\mathrm{P}(g(X) \neq Y),$$

where P denotes the joint distribution of (X, Y). If P were known, the problem of finding an optimal classifier minimizing the misclassification rate would be easily solved by considering the Bayes classifier f^* defined for every $x \in \mathcal{X}$ by

$$f^*(x) = \mathbb{1}_{\eta(x) \geqslant 1/2}, \tag{1}$$

where $\eta(x)$ is the conditional expectation of Y given $X = x$, that is

$$\eta(x) = \mathrm{P}\left[Y = 1 \mid X = x\right]. \tag{2}$$

P is unknown, so the goal is to construct from the sample $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ a classifier $\tilde{f}$ that is as close as possible to f^* in the following sense: since f^* minimizes the misclassification rate, $\tilde{f}$ will be chosen in such a way that its misclassification rate is as close as possible to the misclassification rate of f^*, i.e. in such a way that the expected loss

$$l(f^*, \tilde{f}) = \mathrm{P}(\tilde{f}(X) \neq Y) - \mathrm{P}(f^*(X) \neq Y) \tag{3}$$

is as small as possible. Then, given an estimator $\tilde{f}$ of f^*, the quality of $\tilde{f}$ will be measured by its risk, i.e. the expectation with respect to the product distribution $\mathbb{E}[l(f^*, \tilde{f})]$.
Many works deal with the issue of predicting a label from an input $x \in \mathcal{X}$ via the construction of a classifier having good quality (see for example [1], [39], [11], [35], [18]). Coming both from computational and statistical areas, there exists a large collection of methods based on learning a classifier with respect to a learning sample, where the inputs and labels are known in the fisrt place. For a non exhaustive, but very complete bibliography on this subject, the interessed reader can refer to Boucheron *et. al* [6]. The aforesaid article focuses more on algorithms such as Boosting or Support Vector Machines, but gives the main ideas that are behind the viewpoint we choose to take in this paper.

In order to obtain a linear penalty function in the CART pruning algorithm, an additional hypothesis on the distribution of (X, Y) is needed. Indeed, it has been shown that, without any assumption on P, the penalty function shall be proportional to the squareroot of the number of leaves (see [38], [23] for example). Introducing some margin condition on P permits to obtain various results on estimators of the Bayes classifier improving

the results of Vapnik, unimprovable without any condition. Mammen and Tsybakov [25] introduced a margin condition in 1999 to obtain sharper rates of convergence for their estimators (see also [36]). Further works deal with same kinds of margin conditions to obtain sharper risk bounds for various estimators of the Bayes classifier (see for example [37], [29], [28], [22], [19]), and we will focus here on this kind of margin.
Several margin conditions have been recently generalized by Koltchinskii in [20, 21]. This paper exhibits several manners to obtain risk upper bounds via local rademacher complexities computed on the chosen model or collection of models. In particular, it makes correspondance between margin conditions and local rademacher complexities (see also [2] for data dependent penalties based on local rademacher complexities). To obtain a linear penalty in CART, it appears that the margin condition shall be of the following form (see [20]): there exists some absolute constant $h \in [0; 1/2]$ such that, for all tree structured classifier g,

$$l(f^*, g) \geqslant 2h\mathbb{E}\left[(g(X) - f^*(X))^2\right],$$

where l is the expected loss defined by (3). Assuming that similar conditions are fulfilled will permit to obtain upper bounds for the risk of the CART classifier, leading to the validation of the penalty chosen in the pruning algorithm. In the rest of the paper, the constant h will denote the so-called margin. Of course this margin condition is chosen due to its relevantness in the particular framework of CART and shall be adapted or simply ignored, depending on the problem studied.

The purpose of this paper being the pruning procedure in CART, we leave aside the construction of the first large tree. Some results and discussions on this topic can be found in the papers by Nobel and Olshen [32] and Nobel [30] about Recursive Partitioning. Hence all our upper bounds for the risk of the classifier obtained by CART are considered conditionally on the construction of the first large tree, called maximal tree. We neither focus on concistency for CART to focus on non asymptotic risk bounds. CART is known to be nonconsistent in many cases. Some results and conditions to obtain the consistency can be found in the paper by Devroye *et. al.* [11]. We only give an idea to obtain consistent results for CART based on the risk bounds obtained in Section 3.
Furthermore, Breiman *et. al.* [9] 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.* [14]), or a large tree is constructed and pruned using the same subsample $\mathcal{L}_1 \cup \mathcal{L}_2$ (as done in [9]).
Then the final step used in both cases is to choose a subtree among the sequence by making $\mathcal{L}_3$ go down each tree of the sequence and selecting the tree having the minimum empirical misclassification rate : given for any $k = 1, 2, 3$ and any classifier g the empirical misclassification rate of g on $\mathcal{L}_k$

$$\gamma_{n_k}(g) = \frac{1}{n_k} \sum_{(X_i, Y_i) \in \mathcal{L}_k} \mathbb{1}_{Y_i \neq g(X_i)}, \tag{4}$$

take the final estimator of f^* as follows :

$$\tilde{f} = \underset{\{\hat{f}_{T_i}; 1 \leqslant i \leqslant K\}}{\operatorname{argmin}} \left[\gamma_{n_3}(\hat{f}_{T_i}) \right], \tag{5}$$

where $\hat{f}_{T_i}$ is the piecewise binary estimator of f^* defined on the leaves of the tree T_i and K is the number of trees appearing in the sequence.

The paper is organized as follows. Section 2 gives a slight overview of the CART algorithm, and introduces the methods and notations used in the following sections. Section 3 gives the main theoretical results for classification trees using the two methods mentioned above. This gives the main theorem on the whole algorithm, and more precisely upper bounds for each part of the algorithm we consider, that is the pruning procedure and the final choice by test sample. Finally, Section 4 gives some propects about the margin effect on classification trees, and Section 5 is the appendix where all the proofs of the results of Section 3 are given.

2 The CART Procedure

Let us give a short account of the CART procedure in the classification case and recall the results associated with it, which are fully explained in [9].
CART is based on a recursive partitioning using a training sample $\widetilde{\mathcal{L}}$ of the random variable $(X, Y) \in \mathcal{X} \times \{0, 1\}$ (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. Usually $\mathcal{S}$ 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 [9], [12]). In our framework, we consider a class $\mathcal{S}$ with finite Vapnik-Chervonenkis dimension, henceforth refered to as VC-dimension (for a complete overview of the VC-dimension see [38]).
The procedure 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.

2.1 Growing and pruning procedures

2.1.1 Growing Procedure

Since what mainly interests us in this paper is the pruning procedure, we just give the general idea of the growing procedure to focus on the pruning procedure. For more details about the growing procedure, see [9].
The growing procedure is based on a recursive binary partitioning of $\mathcal{X}$. To give the idea of the construction, let us start with the first step : $\mathcal{X}$ is split into two parts by minimizing some empirical convex function on $\mathcal{S}$. The general idea is to use a strictly convex function in order to avoid ties, what is systematically the case by using the simplest empirical misclassification rate (see [9], [24]). Thus this function is chosen in such a way that the data are split into two groups where the labels of the data in each group are as similar as possible. It implies that the empirical misclassification rate in each subgroup is largely reduced. By the way, the sum of the empirical misclassification rates of each subgroup (called *node*) is always smaller than the global empirical misclassification rate on the data of $\widetilde{\mathcal{L}}$ (called the *root* t_1 of the tree). In the tree terminology, one adds to the root t_1 a left

node t_L and a right node t_R. In what follows, we always assimilate a tree node with its corresponding subset in $\mathcal{S}$. Finally, a label is given to each node by majority vote (what corresponds to minimize the empirical misclassification rate in each node).

Then the same elementary step is applied recursively to the two generated subsamples $\{(X_i, Y_i)\ ;\ X_i \in t_L\}$ and $\{(X_i, Y_i)\ ;\ X_i \in t_R\}$ 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}.

2.1.2 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 $|\widetilde{T}|$ 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 $\mathcal{F}_T$ a set of binary piecewise functions in $\mathbb{L}^2(\mathcal{X})$ defined on the partition given by the leaves of T, one has

$$\begin{aligned} \hat{f}_T &= \operatorname{argmin}_{g \in \mathcal{F}_T} \gamma_n(g) \\ &= \sum_{t \in \widetilde{T}} \operatorname{argmax}_{\{Y_i\ ;\ X_i \in t\}} |\{Y_i\ ;\ X_i \in t\}|\, \mathbb{1}_t, \end{aligned}$$

where γ_n is the empirical misclassification rate defined by (4) and $\mathbb{1}_t(x) = 1$ if x falls in the leaf t, $\mathbb{1}_t(x) = 0$ otherwise.
Then, given $T \preceq T_{max}$ and $\alpha > 0$, one defines

$$\operatorname{crit}_\alpha(T) = \gamma_n(\hat{f}_T) + \alpha \frac{|\widetilde{T}|}{n} \tag{6}$$

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 [9, 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}

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

corresponding to a finite increasing sequence of temperatures

$$0 = \alpha_1 < \alpha_2 < \ldots < \alpha_{K-1} < \alpha_K,$$

where t_1 corresponds to the root of T_{max} as defined in the growing procedure.

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 [9] justifies this algorithm :

Theorem 2.1.1 (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 $k \in \{1,\dots,K\}$, 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 2.1.1 the sequence of pruned subtrees contains the whole statistical information according to the choice of the penalty function used in (6). Consequently it is useless to look at all the subtrees. Let us also recall that the form of the penalized criterion is essential to obtain Theorem 2.1.1. 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 $\tilde{f}$, as mentioned in the Introduction, via equality (5). The results given in Sections 3.1 and 3.2 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 $\tilde{f}$.

Before focusing on risk bounds, let us give the methods and notations used to obtain these bounds.

2.2 Methods and Notations

Assume we observe a set of independent random variables $\mathcal{L} = \{(X_1, Y_1), \cdots, (X_N, Y_N)\}$ such that all the $(X_i, Y_i)_{1\leqslant i\leqslant N}$ belong to $\mathcal{X} \times \{0; 1\}$ and have common unknown joint distribution P. Let f^* be the Bayes classifier (1) and $l(f^*, \tilde{f})$ be the expected loss (3) of the final estimator $\tilde{f}$ associated to P. Then the risk of $\tilde{f}$ becomes

$$R(\tilde{f}, f^*) = \mathbb{E}\,[l(f^*, \tilde{f})].$$

Next, for a given tree T, $\mathcal{F}_T$ will denote the set of classifiers defined on the partition given by the leaves of T, that is

$$\mathcal{F}_T = \left\{ \sum_{t\in\widetilde{T}} a_t \mathbb{1}_t \ ; \ (a_t) \in \{0,1\}^{|\widetilde{T}|} \right\}, \tag{7}$$

where $\widetilde{T}$ refers the set of the leaves of T. Thus $\hat{f}_T$ will be the empirical risk minimizer classifier on $\mathcal{F}_T$. Then a tree-structured estimator $\hat{f}$ of f^* is said to satisfy an oracle inequality if there exists some nonnegative constant C, such that

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

where, for each subtree T pruned from T_{max}, $R_{\mathcal{L}_1}(\hat{f}_T, f^*) = \mathbb{E}\left[l(f^*, \hat{f}_T) \mid \mathcal{L}_1\right]$.

To estimate f^* using the CART algorithm and to compare the performance of $\tilde{f}$ with those of each $\hat{f}_T$, 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 $\tilde{f} = \hat{f}_{\widehat{T}}$.

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 $\tilde{f} = \hat{f}_{\widehat{T}}$.

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 making a selection by using a test sample without penalizing requires to visit all the models. In that case, looking for the best model in the collection of all subtrees pruned from the maximal one becomes explosive. Hence penalizing permits to reduce significantly the number of trees taken into account and then to get also a convenient risk for $\tilde{f}$. Both methods M1 and M2 are considered for the following reasons :

- Since all the risks are considered conditionally on 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.
- A contrario, 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 classifiers 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.* [9] in the pruning algorithm is well-chosen under some conditions on the distribution of (X, Y), and, on the other hand, that the final selection among the pruned subtrees is, in terms of risk, not far from being optimal. Section 3 is 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 on the construction of T_{max}, the final classifier $\tilde{f}$ satisfies some oracle-type inequalities 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 $\{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}_1\}$, so all the risks are taken conditionally on this random grid. Hence a connection can be made between pruning and final selection by test-sample since $\mathcal{L}_3$ is independent of $\mathcal{L}_1$ and $\mathcal{L}_2$.

3 Risk Bounds

This section is devoted to the results obtained on the performance of the CART classifiers for both methods M1 and M2. We give a general theorem in a first place, then we give

some more precise results on the two last parts of the algorithm, that are the pruning procedure and the final selection by test sample.

Assume that the following margin condition is fulfilled: there exists some absolute constant $h \in [0; 1/2]$ such that, for all tree structured classifier g,

$$\begin{array}{rcll} l(f^*, g) & \geqslant & 2h\mathbb{E}\left[(g(X) - f^*(X))^2\right] & \text{if } \tilde{f} \text{ is constructed via M1} \\ l_{n_1}(f^*, g) & \geqslant & 2h\dfrac{1}{n_1}\displaystyle\sum_{X_1^{n_1}}(g(X_i) - f^*(X_i))^2 & \text{if } \tilde{f} \text{ is constructed via M2} \end{array} \tag{8}$$

where l is the expected loss (3) defined in Section 1, $X_1^{n_1} = \{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}_1\}$ and

$$l_{n_1}(f^*, g) = n_1^{-1} \sum_{X_1^{n_1}} \left(\mathrm{P}\left[g(X_i) \neq Y_i \mid X_1^{n_1}\right] - \mathrm{P}\left[f^*(X_i) \neq Y_i \mid X_1^{n_1}\right]\right)$$

is the empirical expected loss conditionally on the grid $X_1^{n_1}$. Let us emphasize that this condition cannot be verified since it relies on the unknown distribution P of (X, Y) (or at least on the unknown conditional distribution). Nevertheless let us remark that this is not a too restrictive condition. To understand a little better what h represents, let us introduce a stronger margin condition in the M1 case, leading to (8) (see also the slightly weaker condition proposed in [19]):

$$\mathrm{P}\left(|\eta(X) - 1/2| > h\right) = 0,$$

where η is the regression function defined by (2). Hence, the so-called margin h can be viewed as a measurement of the gap between labels 0 and 1 in that sense that, if $\eta(x)$ is too close to $1/2$, then choosing 0 or 1 will not make a real difference for that x.
So the margin condition (8) permits to ensure that the model is sparse enough to well-separate the labels with respect to the marginal distribution of X. We will see in the following of the paper that the connection between the zero error case and the non-zero error case is made via the margin by extracting some thresholds for h, essentially depending on n, the dimension of the model used, and the VC dimension of the class of classifiers used.

All the results given in this section are obtained with adapted model selection theorems of Massart *et al.* [29] [28] by conditioning with respect to $\mathcal{L}_1$ for the M2 method, or $\mathcal{L}_1$ and $\mathcal{L}_2$ for the M1 one. Similar methods are used in [34]. All the proofs will be found in Section 5.

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.

Theorem 1. *Given N independant pairs of variables $((X_i, Y_i))_{1 \leqslant i \leqslant N}$ of common distribution* P*, with $(X_i, Y_i) \in \mathcal{X} \times \{0, 1\}$, let us consider the estimator $\tilde{f}$ (5) of the Bayes classifier f^* (1) obtained via the CART procedure as defined in section 2. Then we have the following results.*

(i) if $\tilde{f}$ is constructed via M1 :
Let $l(f^*, \tilde{f})$ be the expected loss (3) of $\tilde{f}$ and h (8) be the margin associated with l. Assume that $h > 2^{-1}\sqrt{|\widetilde{T}_{max}|/n_2}$. Then there exist some absolute constants C, C_1 and C_2 such

that

$$\mathbb{E}\left[l(f^*,\tilde{f}) \mid \mathcal{L}_1\right] \leqslant C \inf_{T \preceq T_{max}} \left\{ \mathbb{E}\left[l(f^*,\hat{f}_T) \mid \mathcal{L}_1\right] + h^{-1}\frac{|\widetilde{T}|}{n_2} \right\} + h^{-1}\frac{C_1}{n_2} \tag{9}$$

$$+h^{-1}C_2\frac{\log(n_1)}{n_3}. \tag{10}$$

(ii) if $\tilde{f}$ is constructed via M2 :
Let $l_{n_1}(f^*,\tilde{f}) = n_1^{-1}\sum_{X_1^{n_1}}\left(\mathrm{P}\left[\tilde{f}(X_i) \neq Y \mid X_1^{n_1}\right] - \mathrm{P}\left[f^*(X_i) \neq Y \mid X_1^{n_1}\right]\right)$ be the empirical expected loss conditionally on the grid $X_1^{n_1} = \{X_i \ ; \ (X_i,Y_i) \in \mathcal{L}_1\}$ and h (8) be the margin associated with l_{n_1}. Let V be the Vapnik-Chervonenkis dimension of the set of splits used to construct T_{max} and suppose that $n_1 \geqslant V$. Assume that $h > 2^{-1}\sqrt{|\widetilde{T}_{max}|/n_1}$. Then there exist some absolute constants C', C_1' and C_2 such that

$$\mathbb{E}\left[l_{n_1}(f^*,\tilde{f}) \mid \mathcal{L}_1\right] \leqslant C' \inf_{T \preceq T_{max}} \left\{ l_{n_1}(f^*,\hat{f}_T) + h^{-1}\log\left(\frac{n_1}{V}\right)\frac{|\widetilde{T}|}{n_1} \right\} + h^{-1}\frac{C_1'}{n_1} \tag{11}$$

$$+h^{-1}C_2\frac{\log(n_1)}{n_3}. \tag{12}$$

Let us remark that the assumption $h > 2^{-1}\sqrt{|\widetilde{T}_{max}|/n}$ (where $n = n_2$ for M1 or $n = n_1$ for M2) is not limitating, since the growing procedure cannot give a maximal tree T_{max} having more than n leaves. Furthermore, this assumption can be controlled during the procedure by forcing the maximal tree's construction to stop earlier.

Let us comment the results given in Theorem 1:

1) For both the M1 and M2 methods, the inequality can be separated into two parts :

   - (9) and (11) correspond to the pruning procedure. They show that, up to some absolute constant and the final selection, the conditional risk of the final classifier is approximately of the same order than the infimum of the penalized risks of the collection of subtrees of T_{max}. The term inside the infimum is of the same form as the penalized criterion (6) used in the pruning procedure. This shows that, for a sufficiently large temperature α, this criterion permits to select convenient subtrees in term of conditional risk. Let us emphasize that the penalty term is directly proportional to the number of leaves in the M1 method, whereas a multiplicative logarithmical term appears in the M2 method. This term is due to the randomness of the models considered, since there is no more independency between the samples used to construct and prune T_{max}.
   - (10) and (12) correspond to the final selection of $\tilde{f}$ among the collection of pruned tree structured classifers using $\mathcal{L}_3$. This selection adds a term proportional to $\log n_1/n_3$, what shows that one does not lose too much by using a test sample if n_3 is sufficiently large with respect to $\log n_1$. Nevertheless, since we have no idea of the size of the constant C_2, it is difficult to deduce a general way of choosing $\mathcal{L}_3$ from this upper bound.

2) Let us comment the role of the Vapnik-Chervonenkis dimension of the set of splits $\mathcal{S}$ used to construct T_{max}. Let us take the more often used case in CART, where $\mathcal{S}$ is

the set of all half-spaces of $\mathcal{X} = \mathbb{R}^d$. In this particular case, we have $V = d + 1$. So, if $\mathcal{X}$ is low dimensional, the $\log n_1$ term has to be taken into account in the risk bound. Nevertheless, if CART provides models 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 we have a result of consistency for $\tilde{f}$ if n_3 is conveniently chosen with respect to $\log n_1$.

3) Let us emphazise the role of the margin in the quality of the selected classifier. Theorem 1 shows that the higher the margin, the smaller the risk, what is intuitive since the more seperable the labels are, and the easier the classification shall be. This confirms the fact that CART does a convenient job if margin condition (8) is fulfilled.
Furthermore, let us give a short account on the lower bound assumed on the margin. Massart *et. al* [29] show that, if $h \leqslant 2^{-1}\sqrt{|\widetilde{T}|/n}$ for one model $\mathcal{F}_T$ (where $n = n_2$ for M1 and $n = n_1$ for M2), then the upper bound for the risk on this model (and then the penalty term in our framework) is of order $\sqrt{|\widetilde{T}|/n}$. They obtain this result via minimax bounds for the risk that make a connection between the zero error case (corresponding to $h = 1/2$), with a minimax risk of order $|\widetilde{T}|/n$, and the "global" pessimistic case (corresponding to $h = 0$, or h too small to have an effect on the minimax risk), with a minimax risk of order $\sqrt{|\widetilde{T}|/n}$.
These results tell us that CART will underpenalize and select classifiers having larger number of leaves if the margin is too small, since in that case the penalty term should be of order $\sqrt{|\widetilde{T}|/n} > |\widetilde{T}|/n$. Let us recall that the pruning procedure and consequently the results of Theorem 2.1.1 heavily depends on the linearity of the penalized criterion (6). It is not clear that these results remain valid by using a non linear penalty function, so we have to keep a penalty term of order $|\widetilde{T}|/n$ to ensure that the sequence of pruned subtrees contains the whole statistical information. So, even if it could be proven that the margin is too small, we cannot replace the penalty term in the algorithm by a penalty of order $\sqrt{|\widetilde{T}|/n}$.
Hence, taking $h > 2^{-1}\sqrt{|\widetilde{T}_{max}|/n}$ in the theorem permits to ensure that $h > 2^{-1}\sqrt{|\widetilde{T}|/n}$ for all the models $\mathcal{F}_T$ given by the pruning procedure.

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. Subsection 3.2 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.

3.1 Validation of the Pruning Procedure

In this section, we focus more particularly on the pruning algorithm and give trajectorial risk bounds for the classifier associated with T_α, the smallest minimizing subtree for the temperature α defined in subsection 3.1. We show that, for a convenient constant α, $\hat{f}_{T_\alpha}$ is not far from f^* in terms of its risk conditionally on $\mathcal{L}_1$. Let us emphasize that the subsample $\mathcal{L}_3$ plays no role in the two following results.

3.1.1 $\tilde{f}$ 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.
For any subtree T of T_{max}, let $\mathcal{F}_T$ be the model defined on the leaves of T given by (7). f^* will then be estimated on $\mathcal{F}_T$, which dimension is $|\widetilde{T}|$.
Then we choose the estimators as follows : let γ_{n_2} be the empirical contrast as defined by (4).

- For $T \preceq T_{max}$, $\hat{f}_T = \text{argmin}_{g \in \mathcal{F}_T} [\gamma_{n_2}(g)]$,
- For $\alpha > 0$, T_α is the smallest minimizing subtree for the temperature α as defined in subsection 2.1.2 and $\hat{f}_{T_\alpha} = \text{argmin}_{g \in S_{T_\alpha}} [\gamma_{n_2}(g)]$.

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

Proposition 1. *Let $P_{\mathcal{L}_2}$ be the product distribution on $\mathcal{L}_2$ and let h (8) be the margin associated with the distribution of $(X,Y) \in \mathcal{X} \times \{0,1\}$. Assume that $h > 2^{-1}\sqrt{|\widetilde{T}_{max}|/n_2}$. Let $\xi > 0$.*
There exists a large enough positive constant $\alpha_0 > 2 + \log 2$ such that, if $\alpha > \alpha_0$, then there exist some nonnegative constants Σ_α and C such that

$$l(f^*, \hat{f}_{T_\alpha}) \quad \leqslant \quad C_1(\alpha) \inf_{T \preceq T_{max}} \left\{ \inf_{g \in \mathcal{F}_T} l(f^*, g) + h^{-1} \frac{|\widetilde{T}|}{n_2} \right\} + C\ h^{-1} \frac{1+\xi}{n_2}$$

on a set Ω_ξ such that $P_{\mathcal{L}_2}(\Omega_\xi) \geqslant 1 - \Sigma_\alpha e^{-\xi}$, where l is defined by (3), $C_1(\alpha) > \alpha_0$ and Σ_α are increasing with α.

A proof of this proposition can be found in paragraph 5.2.1.

We obtain here a trajectorial non asymptotic risk bound on a large probabilty set, leading to the conclusions given for Theorem 1. Nevertheless, taking a too large temperature α will lead to overpenalize and to select a classifier having high risk $\mathbb{E}[l(f^*, \hat{f}_{T_\alpha}) \mid \mathcal{L}_1]$. Furthermore, the fact that $C_1(\alpha)$ and Σ_α are increasing with α tells us that both sides of the inequality grow with α. This is at this stage that the choice of the convenient temperature takes its whole sense in order to make a good compromise between the size of $\mathbb{E}[l(f^*, \hat{f}_{T_\alpha}) \mid \mathcal{L}_1]$ and a large enough penalty term.
In practice, since this temperature depends on the unknown margin h and some unknown constants, the use of a test sample as described in Section 1 is a convenient choice, as shown by Proposition 3.

3.1.2 $\tilde{f}$ constructed via M2

In this subsection we define the different contrasts, expected loss and estimators exactly in the same way as in subsection 3.1.1, despite the fact that l is replaced by the empirical expected loss on $X_1^{n_1} = \{X_i\ ;\ (X_i, Y_i) \in \mathcal{L}_1\}$,

$$l_{n_1}(f^*, g) = \mathbb{E}\left[\gamma(g, (X,Y)) - \gamma(f^*, (X,Y)) \mid X_1^{n_1}\right], \tag{13}$$

since the models and the evaluations of the empirical errors $\gamma_{n_1}(\hat{s}_T)$ are computed on the same grid $X_1^{n_1}$. In this case, we obtain nearly the same performance for $\hat{f}_{T_\alpha}$ despite the fact that the constant appearing in the penalty term can now depend on n_1:

Proposition 2. *Let $P_{\mathcal{L}_1}$ be the product distribution on $\mathcal{L}_1$ and let h (8) be the margin associated with l_{n_1}. Assume that $h > 2^{-1}\sqrt{|\widetilde{T}_{max}|/n_1}$. Let l_{n_1} (13) be the empirical expected loss computed on $\{X_i \ \ (X_i, Y_i) \in \mathcal{L}_1\}$. Let $\xi > 0$ and*

$$\alpha_{n_1,V} = 2 + V/2\left(1 + \log\frac{n_1}{V}\right).$$

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

$$l_{n_1}(f^*, \hat{f}_{T_\alpha}) \ \leqslant \ C_1'(\alpha) \inf_{T \preceq T_{max}} \left\{ \inf_{g \in \mathcal{F}_T} l_{n_1}(f^*, g) + h^{-1}\alpha_{n_1,V}\frac{|\widetilde{T}|}{n_1} \right\} + C' \ h^{-1}\frac{1+\xi}{n_1}$$

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 can be found in paragraph 5.2.2.

We obtain again a trajectorial non asymptotic risk bound on a large probabilty set. The same conclusions as the one of the M1 case hold in this case. Let us just mention that 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.

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 $n_1 > d + 1$, we obtain a penalty proportional to

$$\left(\frac{4 + (d+1)\left(1 + \log\left[n_1/(d+1)\right]\right)}{2h}\right)\frac{|\widetilde{T}|}{n_1}.$$

So, if CART provides some minimax estimator on a class of functions, the $\log n_1$ term always appears for f^* in this class when working in a linear space of low dimension.

As for the M1 case, since the temperature α depends on the unknown margin h and some unknown constants, the use of a test sample to select the final classifier among the sequence of pruned subtrees is a convenient choice, as shown by Proposition 3.

3.2 Final Selection

We focus here on the final step of the CART procedure: the selection of the classifier $\tilde{f}$ among the collection of pruned subtrees given by the pruning procedure by using a test sample $\mathcal{L}_3$. Given the sequence $(T_k)_{1 \leqslant k \leqslant K}$ pruned from T_{max} as defined in subsection 3.1, let us recall that $\tilde{f}$ is defined by

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

The performance of this classifier can be compared to the performance of the collection of classifiers $(\hat{s}_{T_k})_{1 \leqslant k \leqslant K}$ by the following :

Proposition 3.
(i) if $\tilde{f}$ is constructed via M1, let $\lambda = l$ and $R_{n_3}(f^*, \tilde{f}) = \mathbb{E}\left[\lambda(f^*, \tilde{f}) \mid \mathcal{L}_1,\ \mathcal{L}_2\right]$.
(ii) if $\tilde{f}$ is constructed via M2, let $\lambda = l_{n_1}$ and $R_{n_3}(f^*, \tilde{f}) = \mathbb{E}\left[\lambda(f^*, \tilde{f}) \mid \mathcal{L}_1\right]$, where l_{n_1} in defined by (13).
For both cases, there exist three absolute constants $C'' > 1$, $C_1' > 3/2$ and $C_2' > 3/2$ such that

$$R_{n_3}(f^*, \tilde{f}) \quad \leqslant \quad C'' \inf_{1 \leqslant k \leqslant K} \lambda(f^*, \hat{f}_{T_k}) + C_1'\ h^{-1} \frac{\log K}{n_3} + h^{-1} \frac{C_2'}{n_3}.$$

A proof of this proposition can be found in paragraph 5.2.3.

We are now able to prove Theorem 1 via propositions 1, 2 and 3. The proof remains the same if $\tilde{f}$ 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 1. Hence, since for a given $\alpha > 0$ T_α belongs to the sequence $(T_k)_{1 \leqslant k \leqslant K}$,

$$\mathbb{E}\left[l(f^*, \tilde{f}) \mid \mathcal{L}_1,\ \mathcal{L}_2\right] \leqslant C'' \inf_{\alpha > \alpha_0} l(f^*, \hat{f}_{T_\alpha}) + C_1'\ h^{-1} \frac{\log K}{n_3} + h^{-1} \frac{C_2'}{n_3}.$$

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

4 Prospects

We have proven that CART provides convenient classifiers in term of conditional risk under a margin condition. Nevertheless, as for the regression case, it remains to analyze the properties of the growing procedure to obtain full unconditional upper bounds. The assumptions on the size of the margin give some prospects for the application of CART in practice. These propects may be for example

- using the slope heuristic (see for example [4] [3]) to select a classifier among a collection,
- searching for a robust manner to determine if the margin hypothesis is fulfilled, permitting to use the blind selection by test sample,

Some track to estimate the margin could be to use mixing procedures as boosting (see [8] [13] for example). Hence this estimate could be used in the penalized criterion to help finding the convenient temperature. It also could permit to give an idea of the difficulty to classify the data considered and henceforth to help choosing the most adapted classification method.

5 Appendix

5.1 Local Bound for Tree Structured Classifiers

Let $(X, Y) \in \mathcal{X} \times \{0, 1\}$ be a pair of random variables and $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be n independant copies of (X, Y). Let μ denote the marginal distribution of X and $\|.\|_1$

denote the empirical norm on $X_1^n = (X_i)_{1 \leqslant i \leqslant n}$. Then given two classifiers g and u, let us define

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

Let $\mathcal{M}_n*$ be the set of all possible tree structured partitions that can be constructed on the grid X_1^n, corresponding to trees having all possible splits in $\mathcal{S}$ and all possible forms without taking account of the response variable Y. So $\mathcal{M}_n*$ depends only on the grid X_1^n and is independent of the variables $(Y_1, \ldots, Y_n)$. Hence, for a tree $T \in \mathcal{M}_n*$, define

$$\mathcal{F}_T = \left\{ \sum_{t \in \widetilde{T}} a_t \mathbb{1}_t \ ; \ (a_t) \in \{0,1\}^{|\widetilde{T}|} \right\},$$

where $\widetilde{T}$ refers the set of the leaves of T. Then, for any $u \in \mathcal{F}_T$ and any $\sigma > 0$, define

$$B_T(u,\sigma) \quad = \quad \{g \in \mathcal{F}_T \ ; \ d(u,g) \leqslant \sigma\}$$

For each classifier $g : \mathcal{X} \to \{0,1\}$, let us define the empirical contrast of g recentered conditionally on X_1^n

$$\bar{\gamma}_n(g) = \gamma_n(g) - \mathbb{E}[\gamma_n(g) \mid X_1^n], \tag{14}$$

where γ_n is defined for any given classifier g by

$$\gamma_n(g) \quad = \quad \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{g(X_i) \neq Y_i}.$$

Remark 2. 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 the empirical distribution, and the distance d associated with the $\mathbb{L}^2(\mathcal{X}, \mu)$-norm instead of the empirical norm $\|.\|_1$.

We have the following result :

Lemma 1. *For any $u \in \mathcal{F}_T$ and any $\sigma > 0$*

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

Proof. First of all, let us mention that, since the different variables we consider take values in $\{0.1\}$, we have for all $x \in \mathcal{X}$ and all $y \in \{0,1\}$

$$\mathbb{1}_{g(x) \neq y} - \mathbb{1}_{u(x) \neq y} = -(g(x) - u(x)),$$

yielding

$$|\bar{\gamma}_n(g) - \bar{\gamma}_n(u)| = |\nu_n(g-u)|,$$

where ν_n is the recentered empirical measure on X_1^n.
Let us now consider a Rademacker sequence of random signs $(\varepsilon_i)_{1 \leqslant i \leqslant n}$ independent of X_1^n. Then one has by a symmetrization argument

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

Since g and u belong to $\mathcal{F}_T$, we have that

$$g - u = \sum_{t \in \widetilde{T}} (a_t - u_t)\varphi_t,$$

where each (a_t, u_t) takes values in $[0,1]^2$ and $(\varphi_t)_{t\in\widetilde{T}}$ is an orthonormal basis of $\mathcal{F}_T$ adapted to $\widetilde{T}$ (i.e. some normalized characteristic functions). Then by applying the Cauchy-Schwarz inequality, since $g \in B_T(u,\sigma)$, $\|g-u\|_1 = d^2(g,u) = \sum_{t\in\widetilde{T}}(a_t-u_t)^2 \leqslant \sigma^2$, we obtain that

$$\begin{aligned}\left|\sum_{i=1}^n \varepsilon_i (g(X_i) - u(X_i))\right| &\leqslant \sqrt{\sum_{t\in\widetilde{T}}(a_t-u_t)^2}\sqrt{\sum_{t\in\widetilde{T}}\left(\sum_{i=1}^n \varepsilon_i\varphi_t(X_i)\right)^2} \\ &\leqslant \sigma\sqrt{\sum_{t\in\widetilde{T}}\left(\sum_{i=1}^n \varepsilon_i\varphi_t(X_i)\right)^2}.\end{aligned}$$

Finally, since $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are centered random variables with variance equal to 1, independent with X_1^n, and since for each $t \in \widetilde{T}$ $\|\varphi_t\|_1 = 1$, Jensen's inequality implies

$$\mathbb{E}\left[\sup_{g\in B_T(u,\sigma)} |\bar{\gamma}_n(g) - \bar{\gamma}_n(u)| \mid X_1^n\right] \leqslant 2\frac{\sigma}{n}\sqrt{\sum_{t\in\widetilde{T}}\sum_{i=1}^n \varphi_t^2(X_i)} \leqslant 2\sigma\sqrt{\frac{|\widetilde{T}|}{n}}.$$

And the proof is achieved. □

5.2 Proofs

All the proofs are based on results obtained by Massart *et. al.* [28]. The main difference between our viewpoint and the one of Sauvé *at. al.* [34] lies in the concentration inequality used. Since the margin condition we consider can easily be derived from the one in [28], we choose to keep the viewpoint taken by Massart *et. al.*, and generalized by Koltchinskii [20], what is sufficient to obtain the validation of the pruning algorithm in CART.

We give here only sketches of proofs since this kind of results are now routines in the model selection area (see [28] for a more complete overview). The interested reader can find the detailed demonstrations in the first version of the paper [15].

5.2.1 Proof of Proposition 1

To prove Proposition 1, we use a theorem firstly obtained by Massart [27, Theorem 4.2] and adapted by Massart and Nédélec [29] in the classification framework. This theorem uses the Rio's version of the Talagrand concentration inequality [33] instead of the version used by Massart in [27], mainly to obtain sharper upper bounds. This result is discussed in [28], and we adapt it to our case in order to apply it to CART. Let us recall the assumptions and the theorem that we shall use.

Let $n = n_2$. Let us give a sample $\mathcal{L}_2 = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ of the random variable $(X,Y) \in \mathcal{X}\times[0,1]$, where $\mathcal{X}$ is a measurable space and let $f^* \in \mathcal{F} \subset \{g : \mathcal{X} \mapsto [0,1]\ ;\ g \in \mathbb{L}^2(\mathcal{X})\}$ be the unknown function to be recovered. Assume $(\mathcal{F}_m)_{m\in\mathcal{M}_n}$ is a countable collection of countable models included in $\mathcal{F}$. Let us give a penalty function pen_n :

$\mathcal{M}_n \longrightarrow \mathbb{R}_+$, and $\gamma : \mathcal{F} \times (\mathcal{X} \times [0,1]) \longrightarrow \mathbb{R}_+$ a contrast function, i.e. γ such that $g \mapsto \mathbb{E}\left[\gamma(g,(X,Y))\right]$ is convex and minimum at point f^*. Hence define for all $g \in \mathcal{F}$ the expected loss $l(f^*,g) = \mathbb{E}\left[\gamma(g,(X,Y)) - \gamma(f^*,(X,Y))\right]$.
Finally let

$$\gamma_n = \frac{1}{n}\sum_{i=1}^{n} \gamma(.,(X_i,Y_i)) \tag{15}$$

be the empirical contrast associated with γ. Let $\hat{m}$ be defined as

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left[\gamma_n(\hat{f}_m) + \operatorname{pen}_n(m)\right]$$

where $\hat{f}_m = \operatorname{argmin}_{g \in \mathcal{F}_m} \gamma_n(g)$ is the minimum empirical contrast estimator of f^* on $\mathcal{F}_m$. Then the final estimator of f^* is

$$\tilde{f} = \hat{f}_{\hat{m}}. \tag{16}$$

One makes the following assumptions:
H$_1$: γ is bounded by 1 (what is not restricting since all the functions we consider take values in $[0,1]$).
H$_2$: Assume there exist $c \geqslant (2\sqrt{2})^{-1/2}$ and some (pseudo-)distance d such that, for every pair $(g,u) \in \mathcal{F}^2$, one has

$$\operatorname{Var}\left[\gamma(g,(X,Y)) - \gamma(u,(X,Y))\right] \leqslant d^2(g,u),$$

and particularly for all $g \in \mathcal{F}$

$$d^2(f^*,g) \leqslant c^2 l(f^*,g).$$

H$_3$: For any positive σ and for any $u \in \mathcal{F}_m$, let us define

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

where d is given by assumption **H$_2$**. Let $\bar{\gamma}_n = \gamma_n(.) - \mathbb{E}[\gamma_n(.)]$. 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 non-increasing and

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

for every positive σ such that $\phi_m(\sigma) \leqslant \sigma^2$. Let ε_m be the unique solution of the equation $\phi_m(cx) = x^2$, $x > 0$.

One gets the following result :

Theorem 5.2.1 (Massart *et. al.* [28])**.**
Let $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ be a sample of independant realizations of the random pair $(X,Y) \in \mathcal{X} \times [0,1]$. Let $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$ be a countable collection of models included in some countable family $\mathcal{F} \subset \{g : \mathcal{X} \mapsto [0,1] \ ; \ g \in \mathbb{L}^2(\mathcal{X})\}$. Consider some penalty function $\operatorname{pen}_n : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ and the corresponding penalized estimator $\tilde{f}$ (16) of the target function f^*. Take a family of weights $(x_m)_{m \in \mathcal{M}_n}$ such that

$$\Sigma = \sum_{m \in \mathcal{M}_n} e^{-x_m} < +\infty. \tag{17}$$

Assume that assumptions $\mathbf{H_1}$, $\mathbf{H_2}$ *and* $\mathbf{H_3}$ *hold.*
Let $\xi > 0$. *Hence, given some absolute constant* $C > 1$, *there exist some positive constants* K_1 *and* K_2 *such that, if for all* $m \in \mathcal{M}_n$

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

then, with probability larger than $1 - \Sigma e^{-\xi}$,

$$l(f^*, \tilde{f}) \leqslant C \inf_{m \in \mathcal{M}_n} [l(f^*, \mathcal{F}_m) + \mathrm{pen}_n(m)] + C' \, c^2 \frac{1+\xi}{n},$$

where $l(f^*, \mathcal{F}_m) = \inf_{f_m \in \mathcal{F}_m} l(f^*, f_m)$ *and the constant* C' *only depends on* C.

Proof. Let $m \in \mathcal{M}_n$ and $f_m \in \mathcal{F}_m$. The definition of the expected loss and the fact that

$$\gamma_n(\tilde{f}) + \mathrm{pen}_n(\hat{m}) \leqslant \gamma_n(f_m) + \mathrm{pen}_n(m)$$

lead to the following inequality:

$$l(f^*, \tilde{f}) \leqslant l(f^*, f_m) + \bar{\gamma}_n(f_m) - \bar{\gamma}_n(\tilde{f}) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}) \tag{18}$$

where $\bar{\gamma}_n$ is defined by (14). The general principle is now to concentrate $\bar{\gamma}_n(f_m) - \bar{\gamma}_n(\tilde{f})$ around its expectation in order to offset the term $\mathrm{pen}_n(\hat{m})$. Since $\hat{m} \in \mathcal{M}_n$, we proceed by bounding $\bar{\gamma}_n(f_m) - \bar{\gamma}_n(\hat{f}_{m'})$ uniformly in $m' \in \mathcal{M}_n$. For $m' \in \mathcal{M}_n$ and $g \in \mathcal{F}_{m'}$, let us define

$$w_{m'}(g) = \left[\sqrt{l(f^*, f_m)} + \sqrt{l(f^*, g)}\right]^2 + y_{m'}^2,$$

with $y_{m'} \geqslant \varepsilon_{m'}$, where $\varepsilon_{m'}$ is defined by assumption $\mathbf{H_3}$. Hence let us define

$$V_{m'} = \sup_{g \in \mathcal{F}_{m'}} \frac{\bar{\gamma}_n(f_m) - \bar{\gamma}_n(z)}{w_{m'}(z)}.$$

Then (18) becomes

$$l(f^*, \tilde{f}) \quad \leqslant \quad l(f^*, f_m) + V_{\hat{m}} w_{\hat{m}}(\tilde{f}) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m})$$

Since $V_{m'}$ can be written as

$$V_{m'} = \sup_{g \in \mathcal{F}_{m'}} \nu_n \left(\frac{\gamma(f_m, .) - \gamma(g, .)}{w_{m'}(z)} \right),$$

where ν_n is the recentered empirical measure, we bound $V_{m'}$ uniformly in $m' \in \mathcal{M}_n$ by using the Rio's version of the Talagrand's inequality recalled here : if $\mathcal{F}$ is a countable family of measurable functions such that, for some positive constants v and b, one has for all $f \in \mathcal{F}$ $P(f^2) \leqslant v$ and $\|f\|_\infty \leqslant b$, then for every positive y, the following inequality holds for $Z = \sup_{f \in \mathcal{F}} (P_n - P)(f)$

$$\mathbb{P}\left[Z - \mathbb{E}(Z) \geqslant \sqrt{2 \frac{(v + 4b\mathbb{E}(Z))y}{n}} + \frac{by}{n} \right] \leqslant e^{-y}.$$

To proceed, we need to check the two bounding assumptions. First, since by assumption $\mathbf{H_1}$ the contrast γ is bounded by 1, we have that, for each $g \in \mathcal{F}_{\Updownarrow'}$,

$$\left| \frac{\gamma(g, .) - \gamma(f_m, .)}{w_{m'}(z)} \right| \quad \leqslant \quad \frac{1}{y_{m'}^2}. \tag{19}$$

Second, by using assumption $\mathbf{H_2}$, we have that, for each $g \in \mathcal{F}_{\Downarrow'}$,

$$\mathrm{Var}\left[\frac{\gamma(g,(X,Y)) - \gamma(f_m,(X,Y))}{w_{m'}(z)}\right] \leqslant \frac{c^2}{4y_{m'}^2}. \tag{20}$$

Then, by Rio's inequality, we have for every $x > 0$

$$P\left[V_{m'} \geqslant \mathbb{E}(V_{m'}) + \sqrt{\frac{c^2 + 16\mathbb{E}(V_{m'})}{2ny_{m'}^2}x} + \frac{x}{ny_{m'}^2}\right] \leqslant e^{-x}.$$

Let us take $x = x_{m'} + \xi$, $\xi > 0$, where $x_{m'}$ is given by (17). Then by summing up over $m' \in \mathcal{M}_n$, we obtain that for all $m' \in \mathcal{M}_n$

$$V_{m'} \leqslant \mathbb{E}(V_{m'}) + \sqrt{\frac{c^2 + 16\mathbb{E}(V_{m'})}{2ny_{m'}^2}(x_{m'} + \xi)} + \frac{x_{m'} + \xi}{ny_{m'}^2}$$

on a set Ω_ξ such that $P(\Omega_\xi) \geqslant 1 - \Sigma e^{-\xi}$. We have now to bound $\mathbb{E}(V_{m'})$ in order to obtain an upper bound for $V_{m'}$ on the set of large probability Ω_ξ. By using technics similar to Massart *et. al.* [29], we obtain the following inequality via the monoticity of $x \mapsto \phi(x)/x$ and the assumption $c \geqslant (2\sqrt{2})^{-1/2}$: for all $m' \in \mathcal{M}_n$,

$$\mathbb{E}[V_{m'}] \leqslant \frac{8\sqrt{10}\varepsilon_{m'} + c(2n)^{-1/2}}{y_{m'}}.$$

Hence, taking

$$y_{m'} = K\left[8\sqrt{10}\varepsilon_{m'} + c(2n)^{-1/2} + c\sqrt{\frac{x_{m'} + \xi}{n}}\right]$$

with $K > 0$, we obtain that, on Ω_ξ, for all $m' \in \mathcal{M}_n$,

$$V_{m'} \leqslant \frac{1}{K}\left[1 + \sqrt{\frac{1}{2}\left(1 + \frac{8}{K\sqrt{2}}\right)} + \frac{1}{2K\sqrt{2}}\right].$$

Finally, by using repeatedly the elementary inequality $(\alpha + \beta)^2 \leqslant 2\alpha^2 + 2\beta^2$ to bound $y_{\hat{m}}^2$ and $w_{\hat{m}}(\tilde{f})$, we derive that the following inequality holds on Ω_ξ for any $m \in \mathcal{M}_n$ and any $f_m \in \mathcal{F}_m$:

$$\begin{aligned}\left(1 - 2K'\right) l(f^*, \tilde{f}) \leqslant\ & \left(1 + 2K'\right) l(f^*, f_m) + \mathrm{pen}_n(m) + 2K'K^2\frac{\xi}{n} + \frac{2c^2K'K^2}{n} \\ & + 5 \times 2^9 K'K^2\varepsilon_{\hat{m}}^2 + 2c^2K'K^2\frac{x_{\hat{m}}}{n} - \mathrm{pen}_n(\hat{m}),\end{aligned}$$

with

$$K' = \frac{C-1}{2(C+1)}, \quad K_1 = 5 \times 2^9 K'K^2, \quad K_2 = 2K'K^2,$$

achieving the proof. □

Application to classification trees:

Let us now suppose that (X, Y) takes values in $\mathcal{X} \times \{0, 1\}$. The contrast is taken as $\gamma(g,(X,Y)) = \mathbb{1}_{g(X)\neq Y}$, the expected loss is defined by (3), and the collection of models is $(\mathcal{F}_T)_{T \preceq T_{max}}$. The models and the collection are countable since there is a finite number of functions in each $\mathcal{F}_T$, and a finite number of nodes in T_{max}. Since we are working

conditionally on $\mathcal{L}_1$, we can apply Theorem 5.2.1 directly with $\mathcal{L}_2$. To check assumption $\mathbf{H_2}$, let us first note that, since all the variables we consider take values in $\{0,1\}$, we have the following for all classifiers u and g

$$\begin{aligned} |\gamma(u,(X,Y)) - \gamma(g,(X,Y))| &= \left|\mathbb{1}_{Y\neq u(X)} - \mathbb{1}_{Y\neq g(X)}\right| \quad (21)\\ &= |u(X) - g(X)|. \quad (22)\end{aligned}$$

Then if we take $d^2(u,g) = \mathbb{E}\left((u(X)-g(X))^2\right) = \|u-g\|^2$, where $\|.\|$ is the $\mathbb{L}^2$-norm with respect to the marginal distribution of X, we have that, for all classifiers u and g, $\mathrm{Var}\left[\gamma(g,(X,Y)) - \gamma(u,(X,Y))\right] \leqslant d^2(u,g)$. Moreover, with the margin condition, we have that

$$l(f^*,g) \geqslant 2h\|g-f^*\|^2, \quad (23)$$

hence assumption $\mathbf{H_2}$ is checked with $d^2(u,g) = \|u-g\|^2$ and $c^2 = 1/2h$, where h is the margin. By definition of h, we have $h \leqslant 1 \leqslant \sqrt{2}$, and then $c \geqslant (2\sqrt{2})^{-1/2}$.

Then assumption $\mathbf{H_3}$ is checked by Lemma 1 with $\phi_T(x) = 2x\sqrt{|\widetilde{T}|/n}$. Then, to ensure that the upper bound in Theorem 5.2.1 covers the global bound given by the Vapnik theory, Theorem 5.2.1 is verified with $\varepsilon_T = \sqrt{2/h}\sqrt{|\widetilde{T}|/n} \wedge (|\widetilde{T}|/n)^{1/4}$. If it is assumed that $h \geqslant 2^{-1}\sqrt{|\widetilde{T}|/n}$ for all $T \preceq T_{max}$, then the first term in the penalty function is proportional to $|\widetilde{T}|/(hn)$. The assumption on the margin is automatically satisfied if we suppose that $h \geqslant 2^{-1}\sqrt{|\widetilde{T}_{max}|/n}$.

Finally, to choose a convenient family of weights $(x_T)_{T\preceq T_{max}}$, taking $x_T = \theta|\widetilde{T}|$, with $\theta > 2\log 2$ independent of $|\widetilde{T}|$ as done in [17], we immediately obtain $\Sigma_\alpha = \Sigma_\theta < +\infty$. Then we get proposition 1 by Theorem 5.2.1.

5.2.2 Proof of Proposition 2

In what follows, we denote by $\mathcal{L}_1$ the sample $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ of size n of the random variable (X,Y), and by X_1^n the sample $\{X_1,\ldots,X_n\}$.

First we generalize Theorem 5.2.1 to random models, and then we apply it to CART. Let (X,Y), $\mathcal{F}$ $f^* \in \mathcal{F}$, $\mathcal{L}_1 = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, γ and γ_n be defined as in sebsection 5.2.1. Finally let us rewrite the expected loss of $g \in \mathcal{F}$ conditionally on X_1^n as

$$l_n(f^*,g) = \mathbb{E}\left[\gamma(g,(X,Y)) - \gamma(f^*,(X,Y)) \mid X_1^n\right].$$

Let us consider an at most countable collection of at most countable models $(\mathcal{F}_m)_{m\in\mathcal{M}_n^*}$ and a subcollection $(\mathcal{F}_m)_{m\in\mathcal{M}_n}$, where $\mathcal{M}_n \subset \mathcal{M}_n^*$ may depend on $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. Finally let us consider a penalty function $\mathrm{pen}_n : \mathcal{M}_n \mapsto \mathbb{R}_+$ and let us define the estimator $\tilde{f}$ of f^* as follows : let

$$\hat{m} = \mathrm{argmin}_{m\in\mathcal{M}_n}[\gamma_n(\hat{f}_m) + \mathrm{pen}_n(m)],$$

where $\hat{f}_m = \mathrm{argmin}_{g\in\mathcal{F}_m}\gamma_n(g)$ is the minimum contrast estimator of f^* on $\mathcal{F}_m$. Then $\tilde{f} = \hat{f}_{\hat{m}}$.

Let us make the following assumptions.
$\mathbf{H_1}$: γ is bounded by 1.

$\mathbf{H_2}$: Assume there exist $c \geqslant (2\sqrt{2})^{-1/2}$ and some (pseudo-)distance d_n (that may depend on X_1^n) such that, for every pair $(g,u) \in \mathcal{F}^2$, one has

$$\mathrm{Var}\left[\gamma(g,(X,Y)) - \gamma(u,(X,Y)) \mid X_1^n\right] \leqslant d_n^2(g,u),$$

and particularly for all $g \in \mathcal{F}$

$$d_n^2(f^*,g) \leqslant c^2 l_n(f^*,g).$$

$\mathbf{H_3}$: For any positive σ and for any $u \in \mathcal{F}_m$, let us define

$$B_m(u,\sigma) = \{g \in \mathcal{F}_m \ ; \ d_n(u,g) \leqslant \sigma\}$$

where d_n is given by assumption $\mathbf{H_2}$. Let $\bar{\gamma}_n$ be defined as (14). 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 non-increasing and

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

for every positive σ such that $\phi_m(\sigma) \leqslant \sigma^2$. Let ε_m be the unique solution of the equation $\phi_m(cx) = x^2$, $x > 0$.

One gets the following result.

Theorem 2. *Let $\mathcal{L}_1 = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ be a sample of independant realizations of the random pair $(X,Y) \in \mathcal{X} \times [0,1]$. Let $(\mathcal{F}_m)_{m \in \mathcal{M}_n^*}$ be a countable collection of models included in some countable family $\mathcal{F} \subset \{g : \mathcal{X} \mapsto [0,1] \ ; \ g \in \mathbb{L}^2(\mathcal{X})\}$ (depending eventually on X_1^n). Consider some subcollection of models $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$, where $\mathcal{M}_n \subset \mathcal{M}_n^*$ may depend on $\mathcal{L}_1$, and some penalty function $\mathrm{pen}_n : \mathcal{M}_n \longrightarrow \mathbb{R}_+$. Let $\tilde{f}$ (16) be the corresponding penalized estimator of the target function f^*. Take a family of weights $(x_m)_{m \in \mathcal{M}_n^*}$ such that*

$$\sum_{m \in \mathcal{M}_n^*} e^{-x_m} \leqslant \Sigma < +\infty, \tag{24}$$

with Σ deterministic. Assume that assumptions $\mathbf{H_1}$, $\mathbf{H_2}$ and $\mathbf{H_3}$ hold.
Let $\xi > 0$. Hence, given some absolute constant $C > 1$, there exist some positive constants K_1 and K_2 such that, if for all $m \in \mathcal{M}_n$

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

then, with probability larger than $1 - 2\Sigma e^{-\xi}$,

$$l_n(f^*,\tilde{f}) \leqslant C \inf_{m \in \mathcal{M}_n} \left[l_n(f^*,\mathcal{F}_m) + \mathrm{pen}_n(m)\right] + C' \ c^2 \frac{1+\xi}{n},$$

where $l_n(f^*,\mathcal{F}_m) = \inf_{f_m \in \mathcal{F}_m} l_n(f^*,f_m)$ and the constant C' only depends on C.

Proof. There are just a few lines that change from the proof of Theorem 5.2.1. The main differences are in the conditioning and the fact that the collection of models $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$ is random. To remove these issues, all the bounds are computed uniformly on $\mathcal{M}_n^*$ so that the probability of the set we obtain at the end is unconditional to X_1^n since Σ is deterministic. The inequalities are obtained by the same techniques as the ones used for

the proof of the results on model selection on random models done by Gey and Nédélec in [17].

Let $m \in \mathcal{M}_n$ and $f_m \in \mathcal{F}_m$. Starting from (18), we have

$$l_n(f^*, \tilde{f}) \leqslant l_n(f^*, f_m) + w_{\hat{m},m}(\tilde{f}) V_{\hat{m},m} + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}), \tag{25}$$

where for all m' and M in $\mathcal{M}_n^*$, for all $g \in \mathcal{F}_{\Updownarrow'}$ and $f_M \in \mathcal{F}_{\mathcal{M}}$,

$$w_{m',M}(z) = \left[\sqrt{l(f^*, g)} + \sqrt{l_n(f^*, f_M)}\right]^2 + (y_{m'} + y_M)^2,$$

$$V_{m',M} = \sup_{g \in \mathcal{F}_{m'}} \left[\frac{\bar{\gamma}_n(f_M) - \bar{\gamma}_n(g)}{w_{m',M}(z)}\right],$$

with $y_{m'} \geqslant \varepsilon_{m'}$ and $y_M \geqslant \varepsilon_M$. The general principle is now exactly the same as in the proof of Theorem 5.2.1 despite the fact that we have to bound $V_{m',M}$ not only uniformly in $m' \in \mathcal{M}_n^*$, but also in $M \in \mathcal{M}_n^*$ in order to have an in-probabilty inequality that does not depend on X_1^n.

Assumption $\mathbf{H_2}$ permits to give exactly the same upper bounds (except that they depend on X_1^n and that $y_{m'}$ is replaced by $y_{m'} + y_M$) as (19) and (20). By using the same technics as in the proof of Theorem 5.2.1 and the same condiderations as in [17], we obtain that

$$\begin{aligned} V_{m',M} \leqslant\ & \frac{1}{y_{m'} + y_M}\left(8\sqrt{10}\varepsilon_{m'} + \frac{c(2n)^{-1/2}}{2} + 8\sqrt{10}\varepsilon_M + \frac{c(2n)^{-1/2}}{2}\right) \\ & + \sqrt{\frac{c^2 + 16(8\sqrt{10}(\varepsilon_{m'} + \varepsilon_M) + c(2n)^{-1/2})(y_{m'} + y_M)^{-1}}{2n(y_{m'}^2 + y_M^2)}(x_{m'} + x_M + \xi)} \\ & + \frac{1}{y_{m'}^2 + y_M^2}\left(\frac{x_{m'} + \xi/2}{n} + \frac{x_M + \xi/2}{n}\right) \end{aligned}$$

on a set Ω_ξ such that $P(\Omega_\xi \mid X_1^n) \geqslant 1 - 2\Sigma e^{-\xi}$. Then, since Σ is deterministic, we get that $P(\Omega_\xi) \geqslant 1 - 2\Sigma e^{-\xi}$.

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

$$y_{m'} = 2K\left[8\sqrt{10}\varepsilon_{m'} + \frac{c(2n)^{-1/2}}{2} + c\sqrt{\frac{x_{m'} + \xi/2}{n}}\right],$$

we obtain that, on Ω_ξ, for all m' and M in $\mathcal{M}_n^*$,

$$V_{m',M} \leqslant \frac{1}{K}\left[1 + \sqrt{\frac{1}{2}\left(1 + \frac{8}{K\sqrt{2}}\right)} + \frac{1}{2K\sqrt{2}}\right].$$

Finally the proof is achieved in the same way as the proof of Theorem 5.2.1. □

Application to classification trees:

Let us consider the classification framework and the collection of models $(\mathcal{F}_T)_{T \preceq T_{max}}$ obtained via the growing procedure in CART (see subsection 3.1) as recalled in subsection 5.2.1. Since the growing and the pruning procedures are made on the same sample $\mathcal{L}_1$, we are exactly in the conditions of Theorem 2. Since n_1 is fixed, let us consider $\mathcal{M}_n*$ as the set of all possible tree structured partitions that can be constructed on the grid

X_1^n, corresponding to trees having all possible splits in $\mathcal{G}$ and all possible forms without taking account of the response variable Y. So $\mathcal{M}_n*$ depends only on the grid X_1^n and is independent of the variables $(Y_1, \ldots, Y_n)$. Then $\{T \preceq T_{max}\} \subset \mathcal{M}_n^*$ and we are able to apply Theorem 2. Considering (21), we take $d_n(g,u) = \mathbb{E}\left[(g(X) - u(X))^2 \mid X_1^{n_1}\right] = \|u-g\|_1^2$, where $\|.\|_1$ is the empirical norm on $X_1^{n_1}$. Using the margin condition (8), (23) is also verified for l_n and d_n, and we have assumption $\mathbf{H_2}$ with $c = 1/2h$. Then, by Lemma 1, assumption $\mathbf{H_3}$ is checked with $\phi_T(x) = 2x\sqrt{|\widetilde{T}|/n}$ and, in the same way as in the proof of Proposition 1, ε_T is taken as $\varepsilon_T = \sqrt{2/h}\sqrt{|\widetilde{T}|/n}$ under assumption $h \geqslant 2^{-1}\sqrt{|\widetilde{T}|/n}$ for all $T \preceq T_{max}$.
Finally, to choose a convenient family of weights $(x_T)_{T\in\mathcal{M}_n^*}$, taking (see [17])

$$x_T = V\left(\theta + \log\frac{n_1}{V}\right)|\widetilde{T}|,$$

where V is the VC-dimension of the set of splits $\mathcal{S}$ used to construct T_{max} and $\theta > 1$, we obtain

$$\Sigma_\alpha = \Sigma_\theta = \sum_{D\geqslant 1} \exp\left(-(\theta-1)DV\right) < +\infty.$$

And we have Proposition 2.

5.2.3 Proof of Proposition 3

Proposition 3 is a direct application of the theorem obtained by Boucheron, Bousquet and Massart [7] recalled here: assume that we observe $N + n$ independant random variables with common distribution P depending on a parameter f^* to be estimated. Suppose the first N observations $Z' = Z'_1, \ldots, Z'_N$ are used to build some preliminary collection of estimators $(\hat{f}_m)_{m\in\mathcal{M}_n}$ and the remainding observations $Z_1, \ldots, Z_n$ are used to select an estimator $\tilde{f}$ among this collection by minimizing the empirical contrast as defined by (15) (with (X,Y) replaced by Z). Hence we have the following result.

Theorem 5.2.2 (Boucheron, Bousquet, Massart [7])**.**
Suppose that $\mathcal{M}_n$ is finite with cardinal K. Assume that there exists some continuous function w mapping $\mathbb{R}_+$ onto $\mathbb{R}_+$ such that $x \mapsto w(x)/x$ is nonincreasing, and which satisfies for all $\varepsilon > 0$

$$\sup_{\{g\in\mathcal{F}\ ;\ l(f^*,g)\leqslant\varepsilon^2\}} \mathrm{Var}\left[\gamma(g,Z) - \gamma(f^*,Z)\right] \leqslant w(\varepsilon). \tag{26}$$

Then one has for every $\theta \in (0,1)$

$$(1-\theta)\mathbb{E}\left[l(f^*,\tilde{f}) \mid Z'\right] \leqslant (1+\theta)\inf_{m\in\mathcal{M}_n} l(f^*,\hat{f}_m) + \delta_*^2\left(2\theta + (1+\log(K))(\frac{1}{3}+\frac{1}{\theta})\right),$$

where l is defined by (3) and δ_* satisfies $\sqrt{n}\delta_*^2 = w(\delta_*)$.

Taking $w(\varepsilon) = (1/\sqrt{2h})\varepsilon$ for both methods M1 and M2, where h is the margin, leads to proposition 3 with

$$C = \frac{1+\theta}{1-\theta}, \qquad C_1 = \frac{\theta+3}{2\theta(1-\theta)}, \qquad C_2 = C_1 + \frac{\theta}{1-\theta}.$$

References

[1] Aïzerman, M. A., Braverman, E. M., and Rozonoèr, L. I. *Method of Potential Functions in the Theory of Learning Machines*. Nauka, Moscow (in Russian). 1970.

[2] Arlot, S., and Bartlett, P. Margin adaptive model selection in statistical learning. Tech. Rep. 0804.2937, arXiv, 2008.

[3] Arlot, S., and Massart, P. Data-driven calibration of penalties for least squares regression. Tech. Rep. 0802.0837, arXiv, 2008.

[4] Birgé, L., and Massart, P. A generalized cp criterion for gaussian model selection. Tech. Rep. 647, Universités de Paris 6 et 7, 2001.

[5] Blanchard, G., Schafer, C., Rozenholc, Y., and Muller, K.-R. Optimal dyadic decision trees. *Machine Learning 66*, 2-3 (2007), 209–242.

[6] Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat. 9* (2005), 323–375 (electronic).

[7] Boucheron, S., Bousquet, O., and Massart, P. Data-driven penalties. Unpublished manuscript, 2004.

[8] Breiman, L. Arcing classifiers. *Ann. Statist. 26*, 3 (1998), 801–849. With discussion and a rejoinder by the author.

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

[10] Chou, P. A., Lookabaugh, T., and Gray, R. M. Optimal pruning with applications to tree-stuctured source coding and modeling. *IEEE Transactions on Information Theory 35*, 2 (1989), 299–315.

[11] Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, vol. 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[12] Donoho, D. L. CART and best-ortho-basis : A connection. *The Annals of Statistics 25*, 5 (1997), 1870–1911.

[13] Freund, Y., and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*, 1 (1997), 119–139.

[14] Gelfand, S. B., Ravishankar, C., and Delp, E. J. An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on PAMI 13*, 2 (1991), 163–174.

[15] Gey, S. Margin adaptive risk bounds for classification trees. Tech. Rep. 0902.3130, arXiv, 2009.

[16] Gey, S., and Lebarbier, E. Using cart to detect multiple change-points in the mean for large samples. Tech. Rep. 12, SSB, 2008.

[17] Gey, S., and Nedelec, E. Model selection for CART regression trees. *IEEE Trans. Inform. Theory 51*, 2 (2005), 658–670.

[18] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning*. Springer, 2001.

[19] Kohler, M., and Krzyżak, A. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Trans. Inform. Theory 53*, 5 (2007), 1735–1742.

[20] Koltchinskii, V. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist. 34*, 6 (2006), 2593–2656.

[21] Koltchinskii, V. Rejoinder: "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656]. *Ann. Statist. 34*, 6 (2006), 2697–2706.

[22] Lecué, G. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist. 35*, 4 (2007), 1698–1721.

[23] Lugosi, G. Pattern classification and learning theory. In *Principles of nonparametric learning (Udine, 2001)*, vol. 434 of *CISM Courses and Lectures*. Springer, Vienna, 2002, pp. 1–56.

[24] Lugosi, G., and Vayatis, N. On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist. 32*, 1 (2004), 30–55.

[25] Mammen, E., and Tsybakov, A. B. Smooth discrimination analysis. *Ann. Statist. 27*, 6 (1999), 1808–1829.

[26] Mary-Huard, T. *Reduction de la Dimension et Selection de Modeles en Classification Supervisee*. PhD thesis, Universite de Paris-Sud, nb 8303, July 2006.

[27] Massart, P. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse* (2000).

[28] Massart, P. *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[29] Massart, P., and Nédélec, É. Risk bounds for statistical learning. *Ann. Statist. 34*, 5 (2006), 2326–2366.

[30] Nobel, A. B. Recursive partitioning to reduce distortion. *IEEE Trans. on Inform. Theory 43*, 4 (1997), 1122–1133.

[31] Nobel, A. B. Analysis of a complexity-based pruning scheme for classification trees. *IEEE Trans. Inform. Theory 48*, 8 (2002), 2362–2368.

[32] Nobel, A. B., and Olshen, R. A. Termination and continuity of greedy growing for tree-structured vector quantizers. *IEEE Trans. on Inform. Theory 42*, 1 (1996), 191–205.

[33] Rio, E. Une inégalité de Bennett pour les maxima de processus empiriques. *Ann. Inst. H. Poincaré Probab. Statist. 38*, 6 (2002), 1053–1057. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

[34] Sauvé, M., and Tuleau, C. Variable selection through cart. Tech. Rep. 5912, Institut National de Recherche en Informatique et en Automatique, 2006.

[35] Schapire, R. E., Freund, Y., Bartlett, P., and Sun Lee, W. Boosting the margin : a new explanation for the effectiveness of voting methods. *The Annals of Statistics 26*, 5 (1998), 1651–1686.

[36] Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *Ann. Statist. 32*, 1 (2004), 135–166.

[37] Tsybakov, A. B., and van de Geer, S. A. Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist. 33*, 3 (2005), 1203–1224.

[38] Vapnik, V. N. *Statistical Learning Theory*. Wiley Inter-Sciences, 1998.

[39] Vapnik, V. N., and Chervonenkis, A. Y. *Statistical problems of learning*. Nauka, Moscow (in Russian). Moscow, 1974.

[40] Wernecke, Possinger, Kalb, and Stein. Validating classification trees. *Biometrical Journal 40*, 8 (1998), 993–1005.