



HAL
open science

Margin Adaptive Risk Bounds for Classification Trees

Servane Gey

► **To cite this version:**

| Servane Gey. Margin Adaptive Risk Bounds for Classification Trees. 2009. hal-00362281v2

HAL Id: hal-00362281

<https://hal.science/hal-00362281v2>

Preprint submitted on 18 Feb 2009 (v2), last revised 1 Mar 2012 (v5)

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

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

Margin Adaptive Risk Bounds for Classification Trees

Servane Gey*

February 18, 2009

Abstract

Margin adaptive risk bounds for Classification and Regression Trees (CART, Breiman *et. al.* 1984) classifiers are obtained 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 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 [8] 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 [9], [29] for example). From a more general point of view on classification methods, this kind of algorithm is often performed when the space of explanatory variables is high-dimensional. Indeed CART needs fewer operations than other usual methods to provide classifiers. It is now widely used in the genetics framework (see [13] for example), or more generally to reduce variable dimension (see [26] [18] 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* [4] 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* [14] 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 [23].

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

$$\mathbb{P}(z(X) \neq Y),$$

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

$$s^*(x) = \mathbb{1}_{\eta(x) \geq 1/2}, \quad (1)$$

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

$$\eta(x) = \mathbb{P}[Y = 1 \mid X = x]. \quad (2)$$

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

$$l(s^*, \tilde{s}) = \mathbb{P}(\tilde{s}(X) \neq Y) - \mathbb{P}(s^*(X) \neq Y) \quad (3)$$

is as small as possible. Then, given an estimator \tilde{s} of s^* , the quality of \tilde{s} will be measured by its risk, i.e. the expectation $\mathbb{E}[l(s^*, \tilde{s})]$.

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. 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 first place. For a non exhaustive, but very complete bibliography on this subject, the interested reader can refer to Boucheron *et. al* [5]. 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.

The results we obtain on the pruning procedure rely on a new oversight of the two class classification problem initiated by Mammen and Tsybakov [17, 27] in 1999, who introduced the margin condition. A slighter condition has been introduced by Massart *et. al* [21] [20] in 2006 in the following way : there exists some absolute constant $h \in [0; 1/2]$ such that, for every x in \mathcal{X} ,

$$|\eta(x) - 1/2| \geq h, \quad (4)$$

where η is the conditional expectation defined by (2). Assuming that this condition is 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. A complete overview of this margin condition can be found in [20]. General considerations and results about margin adaptive model selection can be found in the upcoming paper by Arlot *et. al* [1].

The purpose of this paper being the pruning procedure in CART, we leave aside the construction of the first deep tree. Some results and discussions on this topic can be found in the papers by Nobel and Olshen [24] and Nobel [22] about Recursive Partitioning. Hence all our upper bounds for the risk of the classifier obtained by CART are considered conditionally to the construction of the first tree, called maximal tree.

Furthermore, Breiman *et. al.* [8] 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.* [12]), or a large tree is constructed and pruned using the same subsample $\mathcal{L}_1 \cup \mathcal{L}_2$ (as done in [8]).

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 z the empirical misclassification rate of z on \mathcal{L}_k

$$\gamma_{n_k}(z) = \frac{1}{n_k} \sum_{(X_i, Y_i) \in \mathcal{L}_k} \mathbb{1}_{Y_i \neq z(X_i)}, \quad (5)$$

take the final estimator of s^* as follows :

$$\tilde{s} = \underset{\{\hat{s}_{T_i}; 1 \leq i \leq K\}}{\operatorname{argmin}} [\gamma_{n_3}(\hat{s}_{T_i})], \quad (6)$$

where \hat{s}_{T_i} is the piecewise binary estimator of s^* 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 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 [8].

CART is based on a recursive partitioning using a training sample $\tilde{\mathcal{L}}$ of the random variable $(X, Y) \in \mathcal{X} \times \{0, 1\}$ (we shall take as $\tilde{\mathcal{L}} = \mathcal{L}_1$ or $\tilde{\mathcal{L}} = \mathcal{L}_1 \cup \mathcal{L}_2$), and a class \mathcal{G} of subsets of \mathcal{X} which tells us how to split at each step. Usually \mathcal{G} 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 [8], [10]). In our framework, we consider a class \mathcal{G} with finite Vapnik-Chervonenkis dimension, henceforth referred to as VC-dimension (for a complete overview of the VC-dimension see [28]).

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 [8].

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{G} . 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 [8], [16]) . 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 $\tilde{\mathcal{L}}$ (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{G} . 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 , \tilde{T} denotes the set of its leaves and $|\tilde{T}|$ the cardinality of \tilde{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 binary piecewise functions in $\mathbb{L}^2(\mathcal{X})$

defined on the partition given by the leaves of T , one has

$$\begin{aligned}\hat{s}_T &= \operatorname{argmin}_{z \in S_T} \gamma_n(z) \\ &= \sum_{t \in \tilde{T}} \operatorname{argmax}_{\{Y_i ; X_i \in t\}} |\{Y_i ; X_i \in t\}| \mathbb{1}_t,\end{aligned}$$

where γ_n is the empirical misclassification rate defined by (5). Then, given $T \preceq T_{max}$ and $\alpha > 0$, one defines

$$\operatorname{crit}_\alpha(T) = \gamma_n(\hat{s}_T) + \alpha \frac{|\tilde{T}|}{n} \quad (7)$$

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 [8, 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}

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

corresponding to a finite increasing sequence of temperatures

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

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 [8] justifies this algorithm :

Theorem 2.1.1 (Breiman, Friedman, Olshen, Stone).

The sequence $(\alpha_k)_{1 \leq k \leq K}$ is nondecreasing, the sequence $(T_k)_{1 \leq k \leq K}$ 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 $(T_k)_{1 \leq k \leq K}$.

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 (7). 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{s} , as mentioned in the Introduction, via equality (6). 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{s} .

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 $\mathcal{L} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$ such that all the $(X_i, Y_i)_{1 \leq i \leq N}$ belong to $\mathcal{X} \times \{0; 1\}$ and have common unknown joint distribution \mathbb{P} . Let s^* be the Bayes classifier (1) and $l(s^*, \tilde{s})$ be the expected loss (3) of the final estimator \tilde{s} associated to \mathbb{P} . Then the risk of \tilde{s} becomes

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

Next, for a given tree T , S_T will denote the set of classifiers defined on the partition given by the leaves of T , that is

$$S_T = \left\{ \sum_{t \in \tilde{T}} a_t \mathbb{1}_t ; (a_t) \in \{0, 1\}^{|\tilde{T}|} \right\}, \quad (8)$$

where \tilde{T} refers the set of the leaves of T . Thus \hat{s}_T will be the empirical risk minimizer classifier 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

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

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

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 \hat{T} is selected among the sequence of pruned subtrees thanks to \mathcal{L}_3 , and we define $\tilde{s} = \hat{s}_{\hat{T}}$.

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 \hat{T} is selected among the sequence of pruned subtrees thanks to \mathcal{L}_3 , and we define $\tilde{s} = \hat{s}_{\hat{T}}$.

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. As we will see in Section 3.2, 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 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.* [8] in the pruning algorithm is well-chosen under some conditions on the distribution of Y conditionnally on $X = x$, 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{s} 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 margin condition (4) mentioned by Massart and Nédélec [21] is fulfilled: there exists some absolute constant $h \in [0; 1/2]$ such that, for every x in \mathcal{X} ,

$$|\eta(x) - 1/2| \geq h,$$

where η is the conditional expectation (2) defined in Section 1.

As mentioned in [21], this assumption permits to obtain more refined upper bounds for the risk of empirical risk minimizer classifiers than the universal bound given by the Vapnik-Chervonenkis (VC) theory (see Lugosi [15] for instance). The so-called margin h measures 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 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 the model selection theorems of Massart *et al.* [21] [20] 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 [26]. 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 $((X_i, Y_i))_{1 \leq i \leq N}$ of common distribution \mathbb{P} , with $(X_i, Y_i) \in \mathcal{X} \times \{0, 1\}$, let us consider the estimator \tilde{s} (6) of the Bayes classifier s^* (1) obtained via the CART procedure as defined in section 2. Let h (4) be the margin associated with \mathbb{P} . Then we have the following results.*

(i) *if \tilde{s} is constructed via M1 :*

Let $l(s^, \tilde{s})$ be the expected loss (3) of \tilde{s} . Assume that $h > 2^{-1} \sqrt{|\tilde{T}_1|/n_2}$. Then there exist some absolute constants C, C_1 and C_2 such that*

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

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

(ii) *if \tilde{s} is constructed via M2 :*

Let $l_{n_1}(s^, \tilde{s}) = \mathbb{P}[\tilde{s}(X) \neq Y \mid X_1^{n_1}] - \mathbb{P}[s^*(X) \neq Y \mid X_1^{n_1}]$ be the empirical expected loss conditionally on the grid $X_1^{n_1} = \{X_i ; (X_i, Y_i) \in \mathcal{L}_1\}$. Let V be the Vapnik-Chervonenkis dimension of the set of splits used to construct T_{max} and suppose that $n_1 \geq V$. Assume that $h > 2^{-1} \sqrt{|\tilde{T}_{max}|/n_1}$. Then there exist some absolute constants C', C'_1 and C_2 such that*

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

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

Let us remark that the assumption $h > 2^{-1} \sqrt{|\tilde{T}_{max}|/n}$ (where $n = n_2$ for M1 or $n = n_1$ for M2) is not limiting, 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 to stop earlier.

Let us comment a little the results given in Theorem 1. For both the M1 and M2 methods, the inequality can be separated in 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 \mathcal{T}_{max} . The term inside the infimum is of the same form as the penalized criterion (7) 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 \mathcal{T}_{max} .

- (10) and (12) correspond to the final selection of \tilde{s} among the collection of pruned tree structured classifiers 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.

Finally, let us emphasize 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 separable the labels are, and the easier the classification shall be. This confirms the fact that CART does a convenient job if the margin condition is fulfilled.

Furthermore, let us give a short account on the lower bound assumed on the margin. Massart *et. al* [21] show that, if $h \leq 2^{-1} \sqrt{|\tilde{T}|/n}$ for one model S_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 $\sqrt{|\tilde{T}|/n}$. 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 $|\tilde{T}|/n$, 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 $\sqrt{|\tilde{T}|/n}$.

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 $\sqrt{|\tilde{T}|/n} > |\tilde{T}|/n$. 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 (7). 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 $|\tilde{T}|/n$ to ensure that the sequence of pruned subtrees contains the whole statistical information. Hence, 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 $\sqrt{|\tilde{T}|/n}$. This issue will be investigated in an upcoming paper cowritten with T. Mary-Huard. Some heuristic alternative to the use of a test sample will be proposed, that seems to automatically compensate the lack of margin by selecting a tree having smaller number of leaves, without changing the tree selected by test sample if the margin is large enough.

Hence, taking $h > 2^{-1} \sqrt{|\tilde{T}_{max}|/n}$ in the theorem permits to ensure that $h > 2^{-1} \sqrt{|\tilde{T}|/n}$ for all the models S_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{s}_{T_α} is not far from s^* 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{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.

For any subtree T of T_{max} , let S_T be the model defined on the leaves of T given by (8). s^* will then be estimated on S_T , which dimension is $|\tilde{T}|$.

Then we choose the estimators as follows : let γ_{n_2} be the empirical contrast as defined by (5).

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

Let us now consider the behaviour of such \hat{s}_{T_α} .

Proposition 1. *Let $P_{\mathcal{L}_2}$ be the product distribution on \mathcal{L}_2 and let h (4) be the margin associated with the distribution of $(X, Y) \in \mathcal{X} \times \{0, 1\}$. Assume that $h > 2^{-1} \sqrt{|\tilde{T}_{max}|/n_2}$. 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

$$l(s^*, \hat{s}_{T_\alpha}) \leq C_1(\alpha) \inf_{T \preceq T_{max}} \left\{ \inf_{z \in S_T} l(s^*, z) + h^{-1} \frac{|\tilde{T}|}{n_2} \right\} + C h^{-1} \frac{1 + \xi}{n_2}$$

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

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

We obtain here a trajectorial non asymptotic risk bound on a large probability 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 $\mathbb{E}[l(s^*, \hat{s}_{T_\alpha}) \mid \mathcal{L}_1]$. 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 $\mathbb{E}[l(s^*, \hat{s}_{T_\alpha}) \mid \mathcal{L}_1]$ 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{s} 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 $X_1^{n_1} = \{X_i ; (X_i, Y_i) \in \mathcal{L}_1\}$,

$$l_{n_1}(s^*, z) = \mathbb{E}[\gamma(z, (X, Y)) - \gamma(s^*, (X, Y)) \mid X_1^{n_1}], \quad (13)$$

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{s}_{T_α} 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 (4) be the margin associated with the distribution of $(X, Y) \in \mathcal{X} \times \{0, 1\}$. Assume that $h > 2^{-1} \sqrt{|\tilde{T}_{max}|/n_1}$. Let l_{n_1} (13) be the empirical expected loss computed on $\{X_i \mid (X_i, Y_i) \in \mathcal{L}_1\}$. 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

$$l_{n_1}(s^*, \hat{s}_{T_\alpha}) \leq C'_1(\alpha) \inf_{T \preceq T_{max}} \left\{ \inf_{z \in S_T} l_{n_1}(s^*, z) + h^{-1} \alpha_{n_1, V} \frac{|\tilde{T}|}{n_1} \right\} + C' h^{-1} \frac{1 + \xi}{n_1}$$

on a set Ω_ξ such that $P_{\mathcal{L}_1}(\Omega_\xi) \geq 1 - 2\Sigma_\alpha e^{-\xi}$, 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 probability 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 $\{X_i \mid (X_i, Y_i) \in \mathcal{L}_1\}$. 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.

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{s} among the collection of pruned subtrees given by the pruning procedure by using a test sample \mathcal{L}_3 . Given the sequence $(T_k)_{1 \leq k \leq K}$ pruned from T_{max} as defined in subsection 3.1, let us recall that \tilde{s} is defined by

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

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

Proposition 3.

- (i) *if \tilde{s} is constructed via M1, let $\lambda = l$ and $R_{n_3}(s^*, \tilde{s}) = \mathbb{E}[\lambda(s^*, \tilde{s}) \mid \mathcal{L}_1, \mathcal{L}_2]$.*
- (ii) *if \tilde{s} is constructed via M2, let $\lambda = l_{n_1}$ and $R_{n_3}(s^*, \tilde{s}) = \mathbb{E}[\lambda(s^*, \tilde{s}) \mid \mathcal{L}_1]$, 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

$$R_{n_3}(s^*, \tilde{s}) \leq C'' \inf_{1 \leq k \leq K} \lambda(s^*, \hat{s}_{T_k}) + C'_1 h^{-1} \frac{\log K}{n_3} + h^{-1} \frac{C'_2}{n_3}.$$

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{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 $K \leq n_1$. 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 $(T_k)_{1 \leq k \leq K}$,

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

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 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 results obtained on the size of the margin give some prospects for the application of CART in practice. These prospects may be for example

- using the slope heuristic (see for example [3] [2]) 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,
- estimating the margin if the condition is satisfied.

Some track to estimate the margin could be to use mixing procedures as boosting (see [7] [11] 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 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be n independant copies of (X, Y) . Let μ denote the marginal distribution of X and $\|\cdot\|_1$ denote the empirical norm on $X_1^n = (X_i)_{1 \leq i \leq n}$. Then given two classifiers z and u , let us define

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

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{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, \dots, Y_n) . Hence, for a tree $T \in \mathcal{M}_n^*$, define

$$S_T = \left\{ \sum_{t \in \tilde{T}} a_t \mathbb{1}_t ; (a_t) \in \{0, 1\}^{|\tilde{T}|} \right\},$$

where \tilde{T} refers the set of the leaves of T . Then, for any $u \in S_T$ and any $\sigma > 0$, define

$$B_T(u, \sigma) = \{z \in S_T ; d(u, z) \leq \sigma\}$$

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

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

where γ_n is defined for any given classifier z by

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

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 $\|\cdot\|_1$.

We have the following result :

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

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

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 $y \in \{0, 1\}$

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

yielding

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

where ν_n is the recentered empirical measure on X_1^n .

Let us now consider a Rademacker sequence of random signs $(\varepsilon_i)_{1 \leq i \leq n}$ independent of X_1^n . Then one has by a symmetrization argument

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

Since z and u belong to S_T , we have that

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

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

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

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

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

And the proof is achieved. \square

5.2 Proofs

All the proofs are based on results obtained by Massart *et. al.* [19] [21]. The main difference between our viewpoint and the one of Sauv e *et. al.* [26] lies in the concentration inequality used. Arlot *et. al.* [1] give margin adaptive upper bounds for classifiers obtained from nested models under some more general margin condition. Since the models we consider are not nested in a first place, we choose to keep the viewpoint taken by Massart *et. al.* in this paper, what is sufficient to obtain the validation of the pruning algorithm in CART.

5.2.1 Proof of Proposition 1

To prove Proposition 1, we first adapt a theorem by Massart [19, Theorem 4.2] in the same way as Massart and N ed elec [21] by using the Rio's version of the Talagrand concentration inequality [25] instead of the version used by Massart in [19], mainly to obtain sharper upper bounds. Let us give this result.

Let $n = n_2$. Let us give a sample $\mathcal{L}_2 = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of the random variable $(X, Y) \in \mathcal{X} \times [0, 1]$, where \mathcal{X} is a measurable space and let $s^* \in \mathcal{S} \subset \{z : \mathcal{X} \mapsto [0, 1] ; z \in \mathbb{L}^2(\mathcal{X})\}$ be the unknown function to be recovered. Assume $(\mathcal{S}_m)_{m \in \mathcal{M}_n}$ is a countable collection of at most countable models included in \mathcal{S} . Let us give a penalty function $\text{pen}_n : \mathcal{M}_n \rightarrow \mathbb{R}_+$, and $\gamma : \mathcal{S} \times (\mathcal{X} \times [0, 1]) \rightarrow \mathbb{R}_+$ a contrast function, i.e. γ such that $z \mapsto \mathbb{E}[\gamma(z, (X, Y))]$ is convex and minimum at point s^* . Hence define for all $z \in \mathcal{S}$ the expected loss $l(s^*, z) = \mathbb{E}[\gamma(z, (X, Y)) - \gamma(s^*, (X, Y))]$.

Finally let

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

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

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

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

$$\tilde{s} = \hat{s}_{\hat{m}}. \quad (16)$$

One makes the following assumptions:

H₁: γ is bounded by 1 (what is not restricting since all the functions we consider take values in $[0, 1]$).

H₂: Assume there exist $c \geq (2\sqrt{2})^{-1/2}$ and some pseudo-distance d such that, for every pair $(z, u) \in \mathcal{S}^2$, one has

$$\operatorname{Var} [\gamma(z, (X, Y)) - \gamma(u, (X, Y))] \leq d^2(z, u),$$

and particularly for all $z \in \mathcal{S}$

$$d^2(s^*, z) \leq c^2 l(s^*, z).$$

H₃ : For any positive σ and for any $u \in \mathcal{S}_m$, let us define

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

where d is given by assumption **H₂**. Let $\bar{\gamma}_n = \gamma_n(\cdot) - \mathbb{E}[\gamma_n(\cdot)]$. 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

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

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

One gets the following result :

Theorem 5.2.1 (adapted from Massart [19] and Massart, Nédélec [21]).

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

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

Assume that assumptions **H₁**, **H₂** and **H₃** 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 $m \in \mathcal{M}_n$

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

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

$$l(s^*, \tilde{s}) \leq C \inf_{m \in \mathcal{M}_n} [l(s^*, S_m) + \operatorname{pen}_n(m)] + C' c^2 \frac{1 + \xi}{n},$$

where $l(s^*, S_m) = \inf_{s \in S_m} l(s^*, s)$ and the constant C' only depends on C .

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

$$\gamma_n(\tilde{s}) + \text{pen}_n(\hat{m}) \leq \gamma_n(s_m) + \text{pen}_n(m)$$

lead to the following inequality:

$$l(s^*, \tilde{s}) \leq l(s^*, s_m) + \bar{\gamma}_n(s_m) - \bar{\gamma}_n(\tilde{s}) + \text{pen}_n(m) - \text{pen}_n(\hat{m}) \quad (18)$$

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

$$w_{m'}(z) = \left[\sqrt{l(s^*, s_m)} + \sqrt{l(s^*, z)} \right]^2 + y_{m'}^2,$$

with $y_{m'} \geq \varepsilon_{m'}$, where $\varepsilon_{m'}$ is defined by assumption **H₃**. Hence let us define

$$V_{m'} = \sup_{z \in S_{m'}} \frac{\bar{\gamma}_n(s_m) - \bar{\gamma}_n(z)}{w_{m'}(z)}.$$

Then (18) becomes

$$l(s^*, \tilde{s}) \leq l(s^*, s_m) + V_{\hat{m}} w_{\hat{m}}(\tilde{s}) + \text{pen}_n(m) - \text{pen}_n(\hat{m})$$

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

$$V_{m'} = \sup_{z \in S_{m'}} \nu_n \left(\frac{\gamma(s_m, \cdot) - \gamma(z, \cdot)}{w_{m'}(z)} \right),$$

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 $f \in \mathcal{F}$ $P(f^2) \leq v$ and $\|f\|_\infty \leq b$, then for every positive y , the following inequality holds for $Z = \sup_{f \in \mathcal{F}} (P_n - P)(f)$

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

To proceed, we need to check the two bounding assumptions. First, since by assumption **H₁** the contrast γ is bounded by 1, we have that, for each $z \in S_{m'}$,

$$\left| \frac{\gamma(z, \cdot) - \gamma(s_m, \cdot)}{w_{m'}(z)} \right| \leq \frac{1}{y_{m'}^2}. \quad (19)$$

Second, by using assumption **H₂**, we have that, for each $z \in S_{m'}$,

$$\text{Var} \left[\frac{\gamma(z, (X, Y)) - \gamma(s_m, (X, Y))}{w_{m'}(z)} \right] \leq \frac{d^2(z, s_m)}{w_{m'}^2(z)} \quad (20)$$

$$\leq \frac{c^2}{4y_{m'}^2}. \quad (21)$$

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

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

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 $m' \in \mathcal{M}_n$

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

on a set Ω_ξ such that $P(\Omega_\xi) \geq 1 - \Sigma e^{-\xi}$. 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 . Let $u_{m'} \in S_{m'}$ be defined by

$$l(s^*, u_{m'}) \leq 2 \inf_{z \in S_{m'}} l(s^*, z).$$

Then we have

$$\mathbb{E}(V_{m'}) \leq \mathbb{E} \left[\sup_{z \in S_{m'}} \frac{|\bar{\gamma}_n(z) - \bar{\gamma}_n(u_{m'})|}{w_{m'}(z)} \right] + \mathbb{E} \left[\frac{|\bar{\gamma}_n(u_{m'}) - \bar{\gamma}_n(s_m)|}{\inf_{z \in S_{m'}} [w_{m'}(z)]} \right].$$

Let

$$\omega_{m'}^2(z) = l(s^*, u_{m'}) + \mathbb{E} [\gamma(z, (X, Y)) - \gamma(u_{m'}, (X, Y))]_+.$$

Then, since

$$l(s^*, z) = l(s^*, u_{m'}) + \mathbb{E} [\gamma(z, (X, Y)) - \gamma(u_{m'}, (X, Y))],$$

we have

$$l(s^*, z) \leq \omega_{m'}^2(z) \leq 5 l(s^*, z). \quad (22)$$

On the one hand we have $w_{m'}(z) \geq l(s^*, z) + y_{m'}^2 \geq (1/5)\omega_{m'}^2(z) + y_{m'}^2$ for every $z \in S_{m'}$. Hence

$$\mathbb{E} \left[\sup_{z \in S_{m'}} \frac{|\bar{\gamma}_n(z) - \bar{\gamma}_n(u_{m'})|}{w_{m'}(z)} \right] \leq 5 \mathbb{E} \left[\sup_{z \in S_{m'}} \frac{|\bar{\gamma}_n(z) - \bar{\gamma}_n(u_{m'})|}{\omega_{m'}^2(z) + 5y_{m'}^2} \right].$$

Furthermore we have

$$\mathbb{E} \left[\sup_{\{z; \omega_{m'}(z) \leq \varepsilon\}} |\gamma_n(z) - \gamma_n(u_{m'})| \right] \leq \mathbb{E} \left[\sup_{\{z; l(s^*, z) \leq \varepsilon^2\}} |\gamma_n(z) - \gamma_n(u_{m'})| \right],$$

and, if $l(s^*, z) \leq \varepsilon^2$, then $l(s^*, u_{m'}) \leq 2\varepsilon^2$ and $d(z, u_{m'}) \leq d(s^*, z) + d(s^*, u_{m'}) \leq c\varepsilon + c\varepsilon\sqrt{2}$. Hence we get that $d(z, u_{m'}) \leq (1 + \sqrt{2})c\varepsilon \leq 2c\varepsilon\sqrt{2}$.

Let us now suppose that $\varepsilon \geq \varepsilon_{m'}$. Then we have by monotonicity of $x \mapsto \phi(x)/x$ and by definition of $\varepsilon_{m'}$ that

$$\frac{\phi_{m'}(2c\varepsilon\sqrt{2})}{(2c\varepsilon\sqrt{2})^2} \leq \frac{\phi_{m'}(c\varepsilon)}{c^2\varepsilon^2 2\sqrt{2}} \leq \frac{\phi_{m'}(c\varepsilon_{m'})}{c^2\varepsilon_{m'}^2 2\sqrt{2}} \leq 1$$

since $c \geq (2\sqrt{2})^{-1/2}$.

So, by assumption **H₃**, we finally obtain that, for all $\varepsilon \geq \varepsilon_{m'}$,

$$\mathbb{E} \left[\sup_{\{z; \omega_{m'}(z) \leq \varepsilon\}} |\gamma_n(z) - \gamma_n(u_{m'})| \right] \leq \mathbb{E} \left[\sup_{\{z; d(z, u_{m'}) \leq 2c\varepsilon\sqrt{2}\}} |\gamma_n(z) - \gamma_n(u_{m'})| \right] \leq \phi_{m'}(2c\varepsilon\sqrt{2}).$$

So we can apply Lemma 5.5 in [21] and use the monotonicity of $x \mapsto \phi_{m'}(x)/x$ to obtain that

$$\mathbb{E} \left[\sup_{z \in S_{m'}} \frac{|\bar{\gamma}_n(z) - \bar{\gamma}_n(u_{m'})|}{w_{m'}(z)} \right] \leq 4 \frac{\phi_{m'}(2c\sqrt{10}y_{m'})}{y_{m'}^2} \leq 8\sqrt{10} \frac{\phi_{m'}(cy_{m'})}{y_{m'}^2}.$$

Hence, since $y_{m'} \geq \varepsilon_{m'}$ and $x \mapsto \phi_{m'}(cx)/x$ is nonincreasing, we get by definition of $\varepsilon_{m'}$

$$\mathbb{E} \left[\sup_{z \in S_{m'}} \frac{|\bar{\gamma}_n(z) - \bar{\gamma}_n(u_{m'})|}{w_{m'}(z)} \right] \leq 8\sqrt{10} \frac{\phi_{m'}(c\varepsilon_{m'})}{y_{m'}\varepsilon_{m'}} \leq 8\sqrt{10} \frac{\varepsilon_{m'}}{y_{m'}}.$$

On the other hand, let us notice that

$$\begin{aligned} \inf_{z \in S_{m'}} w_{m'}(z) &\geq 2y_{m'} \inf_{z \in S_{m'}} [\sqrt{l(s^*, z)} + \sqrt{l(s^*, s_m)}] \\ &\geq \frac{y_{m'}\sqrt{2}}{c} d(u_{m'}, s_m), \end{aligned}$$

hence

$$\mathbb{E} \left[\frac{|\bar{\gamma}_n(u_{m'}) - \bar{\gamma}_n(s_m)|}{\inf_{z \in S_{m'}} [w_{m'}(z)]} \right] \leq c(y_{m'}\sqrt{2})^{-1} \mathbb{E} \left[\frac{|\bar{\gamma}_n(u_{m'}) - \bar{\gamma}_n(s_m)|}{d(u_{m'}, s_m)} \right],$$

leading by Jensen's inequality to

$$\mathbb{E} \left[\frac{|\bar{\gamma}_n(u_{m'}) - \bar{\gamma}_n(s_m)|}{\inf_{z \in S_{m'}} [w_{m'}(z)]} \right] \leq c(y_{m'}\sqrt{2})^{-1} \frac{\sqrt{\text{Var} [\bar{\gamma}_n(u_{m'}) - \bar{\gamma}_n(s_m)]}}{d(u_{m'}, s_m)} \leq \frac{c}{y_{m'}\sqrt{2n}}.$$

Then we get for all $m' \in \mathcal{M}_n$

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

Hence we have that, on the set Ω_ξ , for all $m' \in \mathcal{M}_n$,

$$V_{m'} \leq \frac{8\sqrt{10}\varepsilon_{m'} + c(2n)^{-1/2}}{y_{m'}} + \sqrt{\frac{c^2 + 16(8\sqrt{10}\varepsilon_{m'} + c(2n)^{-1/2})y_{m'}^{-1}}{2ny_{m'}^2}} (x_{m'} + \xi) + \frac{x_{m'} + \xi}{ny_{m'}^2}.$$

Let us define

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

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

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

So we finally obtain that, on the set Ω_ξ ,

$$l(s^*, \tilde{s}) \leq l(s^*, s_m) + K' w_{\hat{m}}(\tilde{s}) + \text{pen}_n(m) - \text{pen}_n(\hat{m}), \quad (23)$$

with

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

We are now able to achieve the proof. By using repeatedly the elementary inequality $(\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2$ we derive that, on the one hand,

$$y_{\hat{m}}^2 \leq 4K^2 \left[640\varepsilon_{\hat{m}}^2 + \frac{c^2}{2n} + c^2 \frac{x_{\hat{m}} + \xi}{2n} \right],$$

and, on the other hand,

$$w_{\hat{m}}(\tilde{s}) \leq 2l(s^*, \tilde{s}) + 2l(s^*, s_m) + y_{\hat{m}}^2.$$

Hence the following inequality holds on Ω_ξ :

$$\begin{aligned} (1 - 2K') l(s^*, \tilde{s}) &\leq (1 + 2K') l(s^*, s_m) + \text{pen}_n(m) + 2K'K^2 \frac{\xi}{n} + \frac{2c^2K'K^2}{n} \\ &\quad + 5 \times 2^9 K'K^2 \varepsilon_{\hat{m}}^2 + 2c^2K'K^2 \frac{x_{\hat{m}}}{n} - \text{pen}_n(\hat{m}). \end{aligned}$$

We can now choose the constants K' (and then K), K_1 and K_2 in order to control the multiplier of $l(s^*, s_m)$ and the last term of (23). If we take

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

then we obtain that, on the set Ω_ξ ,

$$l(s^*, \tilde{s}) \leq C l(s^*, s_m) + \text{pen}_n(m) + C' c^2 \frac{1 + \xi}{n},$$

with $C' = 4\sqrt{2}K'K^2$. Since this inequality is true for any $m \in \mathcal{M}_n$ and any $s_m \in S_m$, the proof is achieved. \square

Application to classification trees:

Let us now suppose that (X, Y) takes values in $\mathcal{X} \times \{0, 1\}$. The contrast is taken as $\gamma(z, (X, Y)) = \mathbb{1}_{z(X) \neq Y}$, the expected loss is defined by (3), and the collection of models is $(S_T)_{T \leq T_{max}}$. The models and the collection are countable since there is a finite number of functions in each S_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 **H₂**, 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 z

$$|\gamma(u, (X, Y)) - \gamma(z, (X, Y))| = |\mathbb{1}_{Y \neq u(X)} - \mathbb{1}_{Y \neq z(X)}| \quad (24)$$

$$= |u(X) - z(X)|. \quad (25)$$

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

$$l(s^*, z) = \mathbb{E}[|2\eta(X) - 1|(z(X) - s^*(X))^2] \geq 2h\|z - s^*\|^2, \quad (26)$$

hence assumption **H₂** is checked with $d^2(u, z) = \|u - z\|^2$ and $c^2 = 1/2h$, where h is the margin. By definition of h , we have $h \leq 1 \leq \sqrt{2}$, and then $c \geq (2\sqrt{2})^{-1/2}$.

Then assumption **H₃** is checked by Lemma 1 with $\phi_T(x) = 2x\sqrt{|\tilde{T}|/n}$. 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 $\varepsilon_T = \sqrt{2/h} \sqrt{|\tilde{T}|/n \wedge (|\tilde{T}|/n)^{1/4}}$. If it is assumed that $h \geq 2^{-1} \sqrt{|\tilde{T}|/n}$ for all $T \preceq T_{max}$, then the first term in the penalty function is proportional to $|\tilde{T}|/(hn)$. The assumption on the margin is automatically satisfied if we suppose that $h \geq 2^{-1} \sqrt{|\tilde{T}_{max}|/n}$.

Finally, to choose a convenient family of weights $(x_T)_{T \preceq T_{max}}$, taking x_T function of the number of leaves of T , one has:

$$\begin{aligned} \Sigma &= \sum_{T \preceq T_{max}} e^{-x_T} \\ &\leq \sum_{D \geq 1} \left| \{T \preceq T_{max} ; |\tilde{T}| = D\} \right| e^{-x_D}. \end{aligned}$$

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

$$\begin{aligned} \Sigma &\leq \sum_{D \geq 1} \frac{1}{D} \binom{2(D-1)}{D-1} e^{-x_D} \\ &\leq \sum_{D \geq 1} \frac{1}{D} \exp(2 \log 2D - x_D). \end{aligned}$$

Taking $x_D = \theta D$, with $\theta > 2 \log 2$ independent of D , 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 $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of size n of the random variable (X, Y) , and by X_1^n the sample $\{X_1, \dots, X_n\}$.

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

$$l_n(s^*, z) = \mathbb{E}[\gamma(z, (X, Y)) - \gamma(s^*, (X, Y)) \mid X_1^n].$$

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

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

where $\hat{s}_m = \operatorname{argmin}_{z \in S_m} \gamma_n(z)$ is the minimum contrast estimator of s^* on S_m . Then $\tilde{s} = \hat{s}_{\hat{m}}$.

Let us make the following assumptions.

H₁ : γ is bounded by 1.

H₂ : Assume there exist $c \geq (2\sqrt{2})^{-1/2}$ and some pseudo-distance d_n (that may depend on X_1^n) such that, for every pair $(z, u) \in \mathcal{S}^2$, one has

$$\operatorname{Var} [\gamma(z, (X, Y)) - \gamma(u, (X, Y)) \mid X_1^n] \leq d_n^2(z, u),$$

and particularly for all $z \in \mathcal{S}$

$$d_n^2(s^*, z) \leq c^2 l_n(s^*, z).$$

H₃ : For any positive σ and for any $u \in \mathcal{S}_m$, let us define

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

where d_n is given by assumption **H₂**. 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

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

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

One gets the following result.

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

$$\sum_{m \in \mathcal{M}_n^*} e^{-x_m} \leq \Sigma < +\infty, \quad (27)$$

with Σ deterministic. Assume that assumptions **H₁**, **H₂** and **H₃** 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 $m \in \mathcal{M}_n$

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

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

$$l_n(s^*, \tilde{s}) \leq C \inf_{m \in \mathcal{M}_n} [l_n(s^*, S_m) + \text{pen}_n(m)] + C' c^2 \frac{1 + \xi}{n},$$

where $l_n(s^*, S_m) = \inf_{s_m \in S_m} l_n(s^*, s_m)$ and the constant C' only depends on C .

Proof. Since there are just a few lines that change from the proof of Theorem 5.2.1, we just give a sketch of proof. The main differences are in the conditioning and the fact that the collection of models $(S_m)_{m \in \mathcal{M}_n}$ 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 [14].

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

$$l_n(s^*, \tilde{s}) \leq l_n(s^*, s_m) + w_{\hat{m}, m}(\tilde{s}) V_{\hat{m}, m} + \text{pen}_n(m) - \text{pen}_n(\hat{m}), \quad (28)$$

where for all m' and M in \mathcal{M}_n^* , for all $z \in S_{m'}$ and $s_M \in S_M$,

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

$$V_{m',M} = \sup_{z \in S_{m'}} \left[\frac{\bar{\gamma}_n(s_M) - \bar{\gamma}_n(z)}{w_{m',M}(z)} \right],$$

with $y_{m'} \geq \varepsilon_{m'}$ and $y_M \geq \varepsilon_M$. 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-probability inequality that does not depend on X_1^n .

Assumption **H₂** 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). Then we bound $\mathbb{E}[V_{m',M} | X_1^n]$ to apply the Rio's version of the Talagrand inequality conditionally on X_1^n . By using the same technique as in the proof of Theorem 5.2.1, we obtain that

$$\mathbb{E} [V_{m',M} | X_1^n] \leq 8\sqrt{10} \frac{\phi_{m'}(cy_{m'} + cy_M)}{(y_{m'} + y_M)^2} + \frac{c}{(y_{m'} + y_M)\sqrt{2n}}.$$

Then, since $y_{m'} + y_M \geq y_{m'} \geq \varepsilon_{m'}$ and $\varepsilon_M > 0$, we get by definition of $\varepsilon_{m'}$

$$8\sqrt{10} \frac{\phi_{m'}(cy_{m'} + cy_M)}{(y_{m'} + y_M)^2} \leq 8\sqrt{10} \frac{\phi(c\varepsilon_{m'})}{(y_{m'} + y_M)\varepsilon_{m'}} \leq 8\sqrt{10} \frac{\varepsilon_{m'} + \varepsilon_M}{y_{m'} + y_M}.$$

So we have

$$\mathbb{E} [V_{m',M} | X_1^n] \leq \frac{8\sqrt{10}(\varepsilon_{m'} + \varepsilon_M) + c(2n)^{-1/2}}{y_{m'} + y_M}.$$

Summing up over $m' \in \mathcal{M}_n^*$ and $M \in \mathcal{M}_n^*$, that leads by Rio's inequality, to

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

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

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^* ,

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

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

Application to classification trees :

Let us consider the classification framework and the collection of models $(S_T)_{T \leq T_{max}}$ 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, \dots, Y_n) . Then $\{T \leq T_{max}\} \subset \mathcal{M}_n^*$ and we are able to apply Theorem 2. Considering (24), we take $d_n(z, u) = \mathbb{E} [(z(X) - u(X))^2 | X_1^{n_1}] = \|u - z\|_1^2$, where $\|\cdot\|_1$ is the empirical norm on $X_1^{n_1}$. Using the margin condition (4), (26) is also verified for l_n , and we have assumption **H₂** with $c = 1/2h$. Then, by Lemma 1, assumption **H₃** is checked with $\phi_T(x) = 2x\sqrt{|\tilde{T}|/n}$ and, in the same way as in the proof of Proposition 1, ε_T is taken as $\varepsilon_T = \sqrt{2/h}\sqrt{|\tilde{T}|/n}$ under assumption $h \geq 2^{-1}\sqrt{|\tilde{T}|/n}$ for all $T \leq T_{max}$.

Finally, to choose a convenient family of weights $(x_T)_{T \in \mathcal{M}_n^*}$, taking x_T function of the number of leaves of T , one has:

$$\sum_{T \in \mathcal{M}_n^*} e^{-x_T} \leq \sum_{D \geq 1} \left| \{T \in \mathcal{M}_n^* ; |\tilde{T}| = D\} \right| e^{-x_D}.$$

We now use a result on the number of trees that can be constructed on n_1 points in \mathcal{X} (see [14, Lemma 2] for more details). Let V be the Vapnik-Chervonenkis dimension of the set of splits \mathcal{G} used to construct T_{max} and let us recall that $n \geq V$. Then we have

$$\left| \{T \in \mathcal{M}_n^* ; |\tilde{T}| = D\} \right| \leq \left(\frac{en_1}{V} \right)^{DV}.$$

So it suffices to take for all $T \in \mathcal{M}_n^*$

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

with $\theta > 1$, to obtain

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

And we have Proposition 2.

5.2.3 Proof of Proposition 3

We apply a result obtained by Boucheron, Bousquet and Massart [6] in the framework of data-driven penalties. For a sake of completeness, we recall this result in the hold-out case.

Assume that we observe $N + n$ independant random variables with common distribution P depending on a parameter s^* to be estimated. Suppose the first N observations $Z' = Z'_1, \dots, Z'_N$ are used to build some preliminary collection of estimators $(\hat{s}_m)_{m \in \mathcal{M}_n}$ and the remaining observations Z_1, \dots, Z_n are used to select an estimator \tilde{s} 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 [6]).

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 $\varepsilon > 0$

$$\sup_{\{z \in \mathcal{S} ; l(s^*, z) \leq \varepsilon^2\}} \text{Var} [\gamma(z, Z) - \gamma(s^*, Z)] \leq w(\varepsilon). \quad (29)$$

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

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

where δ_* satisfies $\sqrt{n}\delta_*^2 = w(\delta_*)$.

We are exactly in the conditions of Theorem 5.2.2 since the collection of estimators $(\hat{s}_{T_k})_{1 \leq k \leq K}$ is built on $\mathcal{L}_1 \cup \mathcal{L}_2$ if \tilde{s} is constructed via M1, and on \mathcal{L}_1 if \tilde{s} is constructed via M2. Furthermore, we derive from (26) that w can be taken as $w(\varepsilon) = (1/\sqrt{2h})\varepsilon$ for both methods M1 and M2, where h is the margin. Then we have $\delta_* = 1/\sqrt{2hn}$ and the proof of proposition 3 is achieved by Theorem 5.2.2 with

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

References

- [1] ARLOT, S., AND BARTLETT, P. Margin adaptive model selection in statistical learning. Tech. rep., arXiv : 0804.2937, 2008.
- [2] ARLOT, S., AND MASSART, P. Data-driven calibration of penalties for least squares regression. Tech. rep., arXiv : 0802.0837, 2008.
- [3] BIRGÉ, L., AND MASSART, P. A generalized cp criterion for gaussian model selection. Tech. Rep. 647, Universités de Paris 6 et 7, 2001.
- [4] BLANCHARD, G., SCHAFER, C., ROZENHOLC, Y., AND MULLER, K.-R. Optimal dyadic decision trees. *Machine Learning* 66, 2-3 (2007), 209–242.
- [5] BOUCHERON, S., BOUSQUET, O., AND LUGOSI, G. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.* 9 (2005), 323–375 (electronic).
- [6] BOUCHERON, S., BOUSQUET, O., AND MASSART, P. Data-driven penalties. Unpublished manuscript, 2004.
- [7] BREIMAN, L. Arcing classifiers. *Ann. Statist.* 26, 3 (1998), 801–849. With discussion and a rejoinder by the author.
- [8] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. *Classification And Regression Trees*. Chapman & Hall, 1984.
- [9] CHOU, P. A., LOOKABAUGH, T., AND GRAY, R. M. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory* 35, 2 (1989), 299–315.

- [10] DONOHO, D. L. CART and best-ortho-basis : A connection. *The Annals of Statistics* 25, 5 (1997), 1870–1911.
- [11] 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.
- [12] 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.
- [13] GEY, S., AND LEBARBIER, E. Using cart to detect multiple change-points in the mean for large samples. Tech. Rep. 12, SSB, 2008.
- [14] GEY, S., AND NEDELEC, E. Model selection for CART regression trees. *IEEE Trans. Inform. Theory* 51, 2 (2005), 658–670.
- [15] 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.
- [16] LUGOSI, G., AND VAYATIS, N. On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.* 32, 1 (2004), 30–55.
- [17] MAMMEN, E., AND TSYBAKOV, A. B. Smooth discrimination analysis. *Ann. Statist.* 27, 6 (1999), 1808–1829.
- [18] MARY-HUARD, T. *Reduction de la Dimension et Selection de Modeles en Classification Supervisee*. PhD thesis, Universite de Paris-Sud, nb 8303, July 2006.
- [19] MASSART, P. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse* (2000).
- [20] 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.
- [21] MASSART, P., AND NÉDÉLEC, É. Risk bounds for statistical learning. *Ann. Statist.* 34, 5 (2006), 2326–2366.
- [22] NOBEL, A. B. Recursive partitioning to reduce distortion. *IEEE Trans. on Inform. Theory* 43, 4 (1997), 1122–1133.
- [23] NOBEL, A. B. Analysis of a complexity-based pruning scheme for classification trees. *IEEE Trans. Inform. Theory* 48, 8 (2002), 2362–2368.
- [24] 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.
- [25] 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.

- [26] SAUVÉ, M., AND TULEAU, C. Variable selection through cart. Tech. Rep. 5912, Institut National de Recherche en Informatique et en Automatique, 2006.
- [27] TSYBAKOV, A. B. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* 32, 1 (2004), 135–166.
- [28] VAPNIK, V. N. *Statistical Learning Theory*. Wiley Inter-Sciences, 1998.
- [29] WERNECKE, POSSINGER, KALB, AND STEIN. Validating classification trees. *Biometrical Journal* 40, 8 (1998), 993–1005.