Risk Bounds for CART Classifiers under a Margin Condition

Servane Gey

Risk Bounds for CART Classifiers under a Margin Condition

Servane Gey*

March 1, 2012

Abstract

Non asymptotic risk bounds for Classification And Regression Trees (CART) classifiers are obtained in the binary supervised classification framework under a margin assumption on the joint distribution of the covariates and the labels. These risk bounds are derived conditionally on the construction of the maximal binary tree and allow to prove that the linear penalty used in the CART pruning algorithm is valid under the 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.

Keywords: Classification, CART, Pruning, Margin, Risk Bounds.
MSC 2010 classification: P2010 62G99 62H99

1 Introduction

The Classification And Regression Trees (CART) method proposed by Breiman, Friedman, Olshen and Stone [7] in 1984 consists in constructing an efficient procedure that gives a piecewise constant estimator of a classifier or a regression function from a training sample of observations. This procedure is based on binary tree-structured partitions and on a penalized criterion that

*Servane.Gey@parisdescartes.fr, Laboratoire MAP5 - UMR 8145, Université Paris Descartes, 75270 Paris Cedex 06, France

selects "good" tree-structured estimators among a huge collection of trees. It currently yields some easy-to-interpret and easy-to-compute estimators which are widely used in many applications in Medicine, Meteorology, Biology, Pollution or Image Coding (see [8], [38] for example). This type of procedure is often performed when the space of explanatory variables is high-dimensional. Due to its recursive computation, CART needs few computations to provide classifiers, which accelerates the computation time drastically when the number of variables is large. It is now widely used in the genetics framework (see [12] for example), or more generally to reduce variable dimension (see [30] [22] for example).

To construct a decision tree from a training sample of observations, the CART algorithm consists in constructing a deep dyadic recursive tree T_{max} from the observations by minimizing some local impurity function at each step. Then, T_{max} is pruned to obtain an uniquely defined finite sequence of nested trees thanks to a penalized criterion, whose penalty term is of the form

$$\mathrm{pen}_n(T) = \alpha \, \frac{|\widetilde{T}|}{n}, \tag{1}$$

where α is a tuning parameter, n is the number of observations, and $|\widetilde{T}|$ is the size of the tree T, i.e. the number of leaves (terminal nodes) of T. Thus the CART algorithm can be viewed as a model selection procedure, where the collection of models is a collection of random decision trees constructed on the training sample of observations. In its pruning procedure, CART selects a small collection of trees within the whole collection of random trees. Then, a final tree belonging to the small collection thus constructed is selected either by cross-validation or by test sample. The present paper focuses on the test sample method.

CART differs from the procedure proposed by Blanchard *et al.* [4] in that the first large tree is constructed locally, and not in a global way by minimizing some loss function on the whole sample. For further results on the construction of the deep tree T_{max}, we refer to Nobel [26, 27], and Nobel and Olshen [28] about Recursive Partitioning.
In this paper, our concern is the pruning step which entails the choice of the penalty function (1): the linearity of the penalty term is fundamental to ensure that the whole information is kept in the obtained sequence. Gey *et*

al. [14] addressed this question in the regression framework. Following this previous work, the present paper aims at validating the choice of the penalty in the two class classification framework. Former results on binary classification (see Nobel [27], or Scott *et al.* [33] in the image context) provide optimal trees in terms of risk conditionally on the construction of the first large dyadic tree T_{max}. These trees are obtained by penalizing the empirical misclassification rate with a penalty term of the form

$$\mathrm{pen}_n(T) \quad = \quad \alpha \sqrt{\frac{|\widetilde{T}| \log n}{n}}. \tag{2}$$

Unfortunately, as discussed by Scott in [32], the pruning algorithm computed with non-linear penalties is computationally slower than the one using linear penalties, and provides subtrees that are not necessarily unique nor nested.

The latter results are obtained without making any assumption on the joint distribution P of the variables. By adding an assumption on P, we exhibit non-asymptotic conditional risk bounds for the tree chosen thanks to the usual CART algorithm as described above. These risk bounds improve those obtained in previous papers (see [27], [32], [33] for instance); they validate the form of the penalty (1) used in the pruning step, and show that the impact of the selection via test sample is conveniently controlled.

In this paper, we leave aside the problem of consistency of CART. CART is known to be non-consistent in many cases. Some results and conditions to obtain consistency can be found in Devroye *et al.* [9]. Furthermore, Section 4 briefly presents consistent results for CART based on the risk bounds obtained.

The outline is the following. Section 2 gives the general framework of binary classification, an overview of the CART procedure, and introduces the methods and notations used in the following sections. Section 4 presents the main theoretical results for classification trees: Theorem 1 bears on the whole procedure, while Propositions 1, 2 concern the pruning procedure and Proposition 3 concerns the final step. Section 5 offers propects about the margin effect on classification trees. Proofs are gathered in Section 6.

2 Classification with CART

2.1 Binary classification

The CART method is used in the following general classification framework. Suppose one observes a sample of N independent 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 measurable space $\mathcal{X}$ and is associated with a label Y taking values in $\{0, 1\}$. A classifier is then any function f mapping $\mathcal{X}$ into $\{0, 1\}$. Its quality is measured by its misclassification rate

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

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

$$f^*(x) = \mathbb{1}_{\eta(x) \geqslant 1/2}, \tag{3}$$

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

$$\eta(x) = \mathrm{P}\left[Y = 1 \mid X = x\right], \tag{4}$$

and $\mathbb{1}$ denotes the indicator function. As P is unknown, the goal is to construct from the sample $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ 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 loss

$$l(f^*, \tilde{f}) = \mathrm{P}(\tilde{f}(X) \neq Y) - \mathrm{P}(f^*(X) \neq Y) \tag{5}$$

is as small as possible. Then, the quality of $\tilde{f}$ will be measured by its risk, i.e, the expectation with respect to the sample distribution

$$\mathbb{E}[l(f^*, \tilde{f})]. \tag{6}$$

Numerous papers have dealt with the issue of predicting a label from an input $x \in \mathcal{X}$ via the construction of a classifier (see for example [1], [37], [9], [31], [15]). There is a large collection of methods coming both from computational and statistical areas and based on learning a classifier from a learning sample, where the inputs and labels are known. For a non exhaustive yet

extensive bibliography on this subject, we refer to Boucheron *et al.* [5].
The classifiers considered in the present paper are classical empirical risk minimizers (also referred to as ERM classifiers), where the empirical misclassification rate on a sample $\mathcal{E}$ of size m is defined, for any classifier f, by

$$\mathrm{P}_m(f) = \frac{1}{m} \sum_{(X_i,Y_i)\in\mathcal{E}} \mathbb{1}_{Y_i \neq f(X_i)}. \tag{7}$$

The ERM classifier $\tilde{f}$ studied here is computed by classical hold out: the sample $\{(X_1, Y_1); \ldots; (X_N, Y_N)\}$ of the random variable $(X, Y) \in \mathcal{X} \times \{0, 1\}$ is split in two independent subsamples: a learning sample $\mathcal{L}$ of size n_l and a test sample $\mathcal{T}$ of size n_t, with $n_l + n_t = N$. A collection of ERM classifiers is computed by minimizing P_{n_l} (equation (7) with $\mathcal{E} = \mathcal{L}$) on a collection of models, and the final classifier $\tilde{f}$ is computed by minimizing P_{n_t} (equation (7) with $\mathcal{E} = \mathcal{T}$) over the collection obtained in that way.

2.2 CART classifiers

The CART algorithm provides piecewise constant classifiers represented by binary decision trees. An example of the latter is given in Figure 1 for a couple of covariates (X^1, X^2) belonging to $\mathcal{X} = [0; 1]^2$.
The tree on the left hand side of Figure 1 defines the partition of $\mathcal{X}$ represented on the right hand side of Figure 1: each question asked on an internal node relates to a split in $\mathcal{X}$. If the answer to the question is positive, go to the left child node, if not, go to the right child node. Hence the first question corresponds to a two-part partition of the covariate space. Then, each part is split into two subparts, and so on. Thus $\mathcal{X}$ is associated to the so called *root* of the tree, and the final partition is associated to the terminal nodes, also called *leaves*, of the tree. Hence each node of the tree represents a subset of the covariates space defined by the successive questions. The final partition is given by the leaves of the tree. Finally, a predictive value for the dependent variable is associated to each leaf. Thus, if $\widetilde{T}$ denotes the set of leaves of a decision tree T, the classifier $f_T : \mathcal{X} \mapsto \{0; 1\}$ defined on $\widetilde{T}$ can be written as

$$f_T = \sum_{t\in\widetilde{T}} a_t \mathbb{1}_t, \tag{8}$$

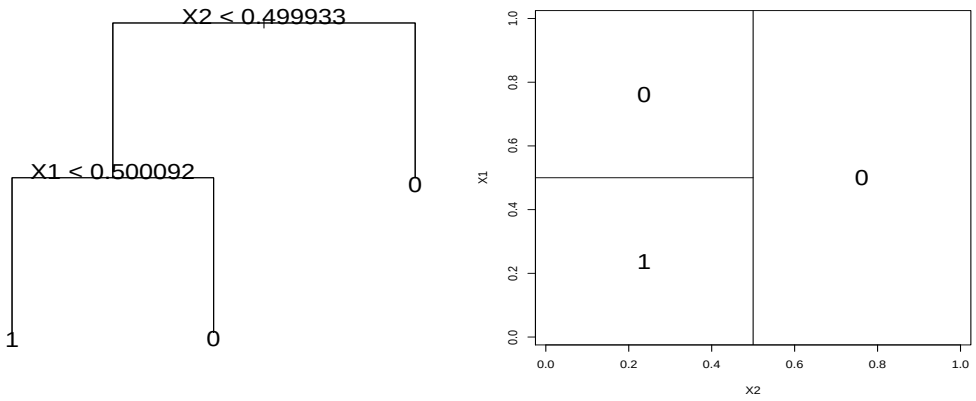

Figure 1: Decision tree example (left) and its associated partition (right).

where $a_t \in \{0; 1\}$ and $\mathbb{1}_t(x) = 1$ if x falls in the leaf t, $\mathbb{1}_t(x) = 0$ otherwise.

2.3 The CART algorithm

CART is based on a recursive partitioning using a class $\mathcal{S}$ of subsets of $\mathcal{X}$ which determines the question to be asked at each internal node of the tree. Below, we consider general classes $\mathcal{S}$ with finite Vapnik-Chervonenkis dimension, henceforth referred to as VC-dimension (for a complete overview of the VC-dimension see [36]). Let us notice that, theoretically, CART can be performed with any kind of split class $\mathcal{S}$, but, in practice, the more frequently used class is that of half spaces of $\mathcal{X}$ with axis-parallel frontiers (which corresponds to axis-parallel cuts) for computational reasons.

To begin with, a collection of CART classifiers is constructed by using learning sample $\mathcal{L}$. This collection is computed in two steps, called the growing algorithm and the pruning algorithm. The growing algorithm allows to construct a maximal binary tree T_{max} from the data by recursive partitioning, and then the pruning algorithm allows to select a finite collection of subtrees of T_{max}.
Since our main interest in this paper is the pruning algorithm, we skip the growing algorithm (for more details about the growing algorithm, see [7]).

Just notice that the maximal tree T_{max} is constructed from the learning sample in such a way that, at the end of the algorithm, its leaves are pure, i.e, contain only observations having the same label.

Then, to avoid overfitting, a decision tree having good predictive performance has to be selected among all possible subtrees pruned from T_{max}. Let us recall that a pruned subtree of T_{max} is defined as any binary subtree of T_{max} having the same root (denoted t_1) as T_{max}. As mentioned in [7], looking at the whole family of subtrees pruned from T_{max} is an NP-hard problem. Then, a good alternative to the exhaustive search is the pruning algorithm, which is computed as follows.

First, let us introduce some notations:

(i) For a tree T, t is the general notation for a node of T and, if t is an internal node, T_t denotes the *branch* of T issued from t, that is the subtree of T whose root is t.

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

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

In the meantime, let us denote by n the size of the sample used to prune T_{max}; in the methods detailed below, we will see that, in any case, $n \leqslant n_l < N$, where n_l is the size of the learning sample $\mathcal{L}$.
Second, let us notice that, given a tree T and $\mathcal{F}_T$ the set of classifiers defined on $\widetilde{T}$ as defined by (8), the ERM classifier on $\mathcal{F}_T$ is

$$\begin{aligned} \hat{f}_T &= \operatorname{argmin}_{f \in \mathcal{F}_T} \mathrm{P}_n(f) \\ &= \sum_{t \in \widetilde{T}} \hat{y}_t \mathbb{1}\!\mathrm{l}_t, \end{aligned}$$

where P_n is the empirical misclassification rate defined by (7), and $\hat{y}_t \in \{0; 1\}$ is the majority vote inside the leaf t. Thus, if t is an internal node of T, $\hat{f}_{|T_t}$ denotes the restriction of $\hat{f}_T$ to the sub-partition associated with the leaves of the branch T_t, and $\mathrm{P}_n(t) = n^{-1} \sum_{\{X_i \in t\}} \mathbb{1}\!\mathrm{l}_{\hat{y}_t \neq Y_i}$ denotes the weighted

misclassification rate inside the node t.
Third, given any subtree $T \preceq T_{max}$ and $\alpha > 0$, one defines

$$\mathrm{crit}_\alpha(T) = \mathrm{P}_n(\hat{f}_T) + \alpha \frac{|\widetilde{T}|}{n}. \tag{9}$$

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

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

(ii) if $\mathrm{crit}_\alpha(T) = \mathrm{crit}_\alpha(T_\alpha)$, 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 proved in [7, pp 284-290].
The pruning algorithm's principle is to raise temperature α, and to record the corresponding T_α. The algorithm is summarized in Table 1 (see [7, pp 59-92] for a complete overview).

Remark 1.

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

2) T_{max} are T_1 are constructed in such a way that, for all $T \preceq T_1$ and all internal node t of T, $\mathrm{P}_n(t) > \mathrm{P}_n(\hat{f}_{|T_t})$; hence, $\alpha_k > 0$ for all $k > 1$.

3) The pruning algorithm is designed to catch, at each iteration k, the minimal temperature α_{k+1} for which the overall energy is kept, that is for which $\mathrm{crit}_{\alpha_{k+1}}(T_{k+1}) = \mathrm{crit}_{\alpha_{k+1}}(T_k)$. This property results directly from the linearity of the penalty used in criterion (9).

Finally, the selection of a tree among the sequence $(T_k)_{1 \leqslant k \leqslant K}$ is made by using test sample $\mathcal{T}$: choose $\hat{k}$ as

$$\hat{k} = \underset{\{1 \leqslant k \leqslant K\}}{\mathrm{argmin}} \left[\mathrm{P}_{n_t}(\hat{f}_{T_k}) \right], \tag{10}$$

where P_{n_t} is the empirical misclassification rate on $\mathcal{T}$ as defined by (7). Then, the final CART classifier is

$$\tilde{f} = \hat{f}_{T_{\hat{k}}}.$$

	Pruning algorithm
Input	Binary decision tree T_{max}.
Initialization	$\alpha_1 = 0$, $T_1 = T_{\alpha_1} = \text{argmin}_{T \preceq T_{max}} \mathrm{P}_n(\hat{f}_T)$. `Set` $T = T_1$ and $k = 1$.
Iteration	`While` $\lvert\widetilde{T}\rvert > 1$, `Compute` $\alpha_{k+1} = \min_{\{t \text{ internal node of } T\}} \dfrac{\mathrm{P}_n(t) - \mathrm{P}_n(\hat{f}_{\lvert T_t})}{\lvert\widetilde{T}_t\rvert - 1}$. `Prune` all branches T_t of T verifying $\mathrm{P}_n(\hat{f}_{\lvert T_t}) + \alpha_{k+1}\lvert\widetilde{T}_t\rvert = \mathrm{P}_n(t) + \alpha_{k+1}$ `Set` T_{k+1} the pruned subtree obtained in that way. `Set` $T = T_{k+1}$ and $k = k + 1$.
Output	Trees $T_1 \succ \ldots \succ T_K = \{t_1\}$, Temperatures $0 = \alpha_1 < \ldots < \alpha_K$.

Table 1: CART pruning algorithm.

2.4 Properties of the pruned subtrees sequence

It may be easily seen that the computational complexity of the pruning algorithm is linear with respect to the number of nodes of T_{max}. Hence, the pruning algorithm is interesting in two ways:

1) It reduces drastically the computational complexity of the exhaustive search from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$ (see [32] for instance),

2) It provides a small collection of trees that can be easily evaluated on $\mathcal{T}$.

Thus, to ensure that the CART algorithm provides good classifiers, it is important to verify that

- pruning is like looking at the entire family of pruned subtrees according to penalized criterion (9),

- pruning provides trees having good performance in term of risk conditionally on the growing algorithm,
- using a test sample does not alter too much the performance of the tree thus selected.

The first point has already been established by Breiman *et al.* [7]:

Theorem 2.4.1 (Breiman, Friedman, Olshen, Stone [7])**.**
For all $k \in \{1, \ldots, K\}$, $T_k = T_{\alpha_k}$ *and, for all* $\alpha > 0$, *there exists* $k \in \{1, \ldots, K\}$ *satisfying* $T_\alpha = T_k$.

Theorem 2.4.1 ensures that

1) the trees of the sequence are unique and minimize penalized criterion (9) for known temperatures,
2) whatever the choice of the temperature α used in the penalized criterion (9), T_α belongs to the sequence.

Thus, the definition of T_α leads to an infinite collection of trees over all real α, but only finitely many trees are possible according to criterion (9).

To the best of our knowledge, the fact that the classifiers provided by CART perform well in terms of conditional risk remains to be seen. To proceed, two methods are applied to construct the sequence $(T_k)_{1 \leqslant k \leqslant K}$. These methods, as well as the general notations and assumptions refered to in this paper, are presented in the next section.

3 Methods, Notations and Assumptions

3.1 Methods and notations

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

$$\mathcal{F}_T = \left\{ \sum_{t \in \widetilde{T}} a_t \mathbb{1}_t \ ; \ (a_t) \in \{0,1\}^{|\widetilde{T}|} \right\}. \tag{11}$$

Thus $\hat{f}_T = \sum_{t \in \widetilde{T}} \hat{y}_t \mathbb{1}_t$ is the ERM classifier on $\mathcal{F}_T$.

The two different methods applied in the CART pruning algorithm are:

M1: $\mathcal{L}$ is split in two independent parts $\mathcal{L}_1$ and $\mathcal{L}_2$ containing respectively n_1 and n_2 observations, with $n_1 + n_2 = n_l = N - n_t$. Hence T_{max} is constructed using $\mathcal{L}_1$, then pruned using $\mathcal{L}_2$. This method is applied in Gelfand *et al.* [11] for instance.

M2: T_{max} is constructed and pruned using sample $\mathcal{L}$ entirely. This is the most commonly used method in the CART literature and its applications.

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 (see [7]). So making a selection by using a test sample without penalizing requires visiting all the models. As mentioned above, looking for the best model in the collection of all subtrees pruned from the maximal one becomes explosive. Hence pruning allows to reduce significantly the number of trees taken into account. With both M1 and M2 methods, $\mathcal{T}$ is used to select a tree among the pruned sequence. Let us mention that $\mathcal{T}$ usually represents 10% of the data and is randomly taken in the original sample, except if the design is fixed. In that case one takes, for example, one observation out of ten to obtain the test sample. In a similar way, for the M1 method $\mathcal{L}_1$ and $\mathcal{L}_2$ are taken randomly in $\mathcal{L}$, except if the design is fixed, in which case one takes one observation out of two for instance.

Methods M1 and M2 involve different treatments for the risks of the CART classifiers thus obtained. Indeed, by conditioning with respect to the sample used to perform the growing algorithm, T_{max} becomes deterministic with M1, while it implies random models depending on the sample used to prune T_{max} with M2. In the latter case, union bounds on the family of all possible trees that can be constructed on the grid $\{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\}$ are used to obtain risk bounds. This allows to obtain risk bounds only conditionally on this grid instead of conditionally on the grid and the labels. To simplify the notations, we define the loss and the $\mathbb{L}^2$ distance corresponding with either method M1 or M2.

Definition 1. *The loss of a classifier f is defined by $\lambda(f^*, f)$, and is computed as follows:*
(i) if $\tilde{f}$ is constructed via M1, $\lambda(f^*, f) := l(f^*, f)$, with l defined by (5).
(ii) if $\tilde{f}$ is constructed via M2,

$$\begin{aligned}\lambda(f^*, f) &:= \mathbb{E}\left[\mathrm{P}_{n_l}(f) - \mathrm{P}_{n_l}(f^*) \mid X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\right] \\ &= \frac{1}{n_l} \sum_{\{X_i \ ; \ (X_i,Y_i)\in\mathcal{L}\}} |2\eta(X_i) - 1| \, \mathbb{1}_{f(X_i)\neq f^*(X_i)},\end{aligned}$$

where P_{n_l} is the empirical misclassification rate on $\mathcal{L}$ defined by (7), and η is defined by (4).

Since $l(f^*, f) = \mathbb{E}\left[|2\eta(X) - 1| \, \mathbb{1}_{f(X)\neq f^*(X)}\right]$ for all classifier f (see [9] for instance), λ is just the empirical version of l on the grid $\{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\}$ in the M2 case.

Definition 2. *The $\mathbb{L}^2$ distance between two classifiers f and g is defined by $d(f, g)$, and is computed as follows:*
(i) if $\tilde{f}$ is constructed via M1, $d^2(f, g) := \mathbb{E}\left[(f(X) - g(X))^2\right]$.
(ii) if $\tilde{f}$ is constructed via M2,

$$d^2(f, g) := d^2_{n_l}(f, g) = \frac{1}{n_l} \sum_{\{X_i \ ; \ (X_i,Y_i)\in\mathcal{L}\}} \left(f(X_i) - g(X_i)\right)^2,$$

As for λ, d is the empirical version of the $\mathbb{L}^2$ distance on the grid $\{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\}$ with M2.

Remark 2. If the design is fixed, λ and d are different according to the method only through the grid on which they are computed (the grid of method M1 being obtained from the one of method M2 by taking one point out of two). In this case, λ and d are no more random.

We based our computation of risk bounds for the ERM classifiers provided by CART on recent results (see for instance [21], [34], [35], [25], [17, 18], [24], [19], [16]). They stem from Vapnik's results (see [36], [20] for example), showing that, without any assumption on the joint distribution P, the penalty term used in the penalized criterion for the model selection procedure should be taken proportional to $\sqrt{|\widetilde{T}|/n_l}$ to obtain classifiers optimal in term of conditional risk (see [27, 33] for instance). Nevertheless, it has also been

shown that, under the overoptimistic zero-error assumption (that is $Y = \eta(X)$ almost surely, where η is defined by (4)), this penalty term should be taken proportional to $|\widetilde{T}|/n_l$, as done in criterion (9). Since we aim at validating the choice of the penalized criterion (9) in contexts less restrictive than the zero-error one, we consider weaker assumptions on P.

3.2 Margin assumptions

Margin assumptions are now widely known to improve risk bounds of ERM classifiers in the binary classification context. One of the best-known margin assumptions is that of Mammen and Tsybakov [21] that may be written as follows:

MA(MT) There exist some constants $C > 0$ and $\kappa > 1$ such that, for all $t > 0$,

$$\mathrm{P}\left(|2\eta(X) - 1| \leqslant t\right) \leqslant C\ t^{\frac{1}{\kappa - 1}}, \tag{12}$$

where η is defined by (4). **MA(MT)** implies the more intuitive assumption considered by Massart and Nedelec in [25] (see also the slightly weaker condition proposed in [16]): taking $t = h \in]0;1[$ and the limit value $\kappa = 1$, **MA(MT)** leads to

MA(MN) $\exists h \in]0;1[\quad \mathrm{P}\left(|2\eta(X) - 1| \leqslant h\right) = 0.$

Assumption **MA(MN)** means that (X, Y) is sufficiently well distributed to ensure that there is no region in $\mathcal{X}$ for which the toss-up strategy could be favored over others: h can be viewed as a measurement of the gap between labels 0 and 1 in the 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.
In assumption **MA(MT)**, η can be continuous, but has to cross the line $\eta(x) = 1/2$ in a non smooth way.

From this simple example, the so called *margin* h can be viewed as a noise level for the classification problem. From this point of view, margin assumptions have been generalized by Koltchinskii in [17]; they compare directly the loss l defined by (5) with some kind of "noise variance" related to the $\mathbb{L}^2$ distance to the Bayes classifier f^*:

MA(K) There exists some strictly convex positive function φ satisfying $\varphi(0) = 0$ such that,

$$\forall f : \mathcal{X} \ \rightarrow \ \{0;1\} \quad l(f^*, f) \geqslant \varphi\left(\sqrt{\mathbb{E}\left[(f(X) - f^*(X))^2\right]}\right)$$

It is easy to check that **MA(MT)** and **MA(MN)** imply **MA(K)** with $\varphi(x) = C_\kappa x^{\frac{2\kappa}{2\kappa-1}}$ and $\varphi(x) = hx^2$ respectively.

Remark 3. Taking $h > 1$ in **MA(MN)** (or more generally $\varphi(x) > x^2$ in **MA(K)**) has no sense since, for any classifier f, (see [9] for instance)

$$l(f^*, f) = \mathbb{E}\left[|2\eta(X) - 1| \left(f(X) - f^*(X)\right)^2\right] \leqslant \mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right].$$

MA(MT) (with $\kappa > 1$) and **MA(MN)** (with $\kappa = 1$) lead to risk bounds suggesting that the empirical misclassification rate of $\hat{f}_T$ have to be penalized by a term proportional to $\left(|\widetilde{T}|/n_l\right)^{\kappa/(2\kappa-1)}$ to obtain ERM classifiers optimal in terms of risk (see also [35] for instance), while **MA(K)** leads to more general penalty terms given by strictly concave functions of $|\widetilde{T}|/n_l$. Hence these margin assumptions make the link between the "global" pessimistic case (without any assumption on P) and the zero-error case by considering some noise level of the classification problem. More recent results (see [17, 18], [2] for instance) deal with data-driven penalties based on local Rademacher complexities also derived from margin assumptions.

As it can be seen in [7], the CART pruning algorithm looks at the entire family of pruned subtrees according to criterion (9) only if the penalty taken in the criterion is linear. Thus, it follows from the above mentioned results that the following margin assumption has to be fulfilled:

MA(1) $\exists h \in]0;1[\quad \forall f : \mathcal{X} \mapsto \{0;1\} \quad \lambda(f^*, f) \geqslant h d^2(f^*, f),$

where λ and d are defined in Definitions 1 and 2 respectively.

Examples:

1) Take $X = (X^1, \ldots, X^d)$ uniformly distributed on $[0;1]^d$. The associated label is designed as follows: if $X^j \leqslant 1/2$ or $X^j > 1/2$ for all $j = 1, \ldots, d$, then $Y = 1$ with probability q; otherwise $Y = 1$ with probability $1 - q$.

2) Take $X = (X^1, X^2)$ such that X^1 and X^2 are independently generated with gaussian distribution $\mathcal{N}(0,1)$. The associated label is designed as follows: If $X^1 > 0$ and $X^2 > 0$ then $Y = 1$ with probability q, otherwise $Y = 1$ with probability $1 - q$.

In these two simple examples, if $q \neq 1/2$, **MA(MN)**, and consequently **MA(1)**, is satisfied with any value of h satisfying $0 < h < |2q - 1|$ in both M1 and M2 cases; indeed $\eta(X) = q$ or $\eta(X) = 1 - q$, depending on where X falls. Examples in which **MA(1)** fails can be found in [2].

Below, we prove that, under **MA(1)**, the penalty used by CART in criterion (9) for the pruning step leads to classifiers having good performance.
In the remaining part of this paper, the constant h will denote the so called *margin*.

4 Risk Bounds

This section is devoted to the results obtained on the performance of the CART classifiers for both M1 and M2 methods. These performance are regarded from the risk viewpoint presented in paragraph 2.1, where classifiers are considered as estimators of the Bayes classifier f^*. The risk of the classifier $\tilde{f}$ provided by the CART algorithm is compared to those of the collection $\left(\hat{f}_T\right)_{T \preceq T_{max}}$ conditionally on the construction of T_{max}.
We shall first present a general theorem, then give more precise results about the last two parts of the algorithm, which are the pruning algorithm and the final selection by test sample.

Theorem 1. *Given N independent 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 $\tilde{f}$ (10) of the Bayes classifier f^* (3) obtained via the CART algorithm as defined in section 2. Then we have the following results.*

(i) if $\tilde{f}$ is constructed via M1:
Suppose that margin assumption **MA(1)** *is satisfied. Then, there exist some*

absolute constants C, C_1 and C_2 such that

$$\mathbb{E}\left[\lambda(f^*, \tilde{f}) \mid \mathcal{L}_1\right] \leqslant C \inf_{T \preceq T_{max}} \left\{ \inf_{f \in \mathcal{F}_T} \mathbb{E}\left[\lambda(f^*, f) \mid \mathcal{L}_1\right] + \frac{|\widetilde{T}|}{hn_2} \right\} + \frac{C_1}{hn_2} \quad (13)$$

$$+ C_2 \frac{\log(n_l)}{hn_t}. \quad (14)$$

(ii) if $\tilde{f}$ is constructed via M2:
Let $P_{\mathcal{L}}$ be the $\mathcal{L}$ sample distribution. Let V be the Vapnik-Chervonenkis dimension of the set of splits used to construct T_{max} and suppose that $n_l \geqslant V$. Let K be the number of pruned subtrees of the sequence provided by the pruning algorithm, and suppose that margin assumption **MA(1)** *is satisfied. Then, there exist some absolute constants C', C'_1, C''_1 and C_2 such that, for every $\delta \in]0;1[$, on a set Ω_δ verifying $\mathrm{P}_{\mathcal{L}}(\Omega_\delta) \geq 1 - \delta$,*

$$\mathbb{E}\left[\lambda(f^*, \tilde{f}) \mid \mathcal{L}\right] \leqslant C' \inf_{T \preceq T_{max}} \left\{ \inf_{f \in \mathcal{F}_T} \lambda(f^*, f) + \log\left(\frac{n_l}{V}\right) \frac{|\widetilde{T}|}{hn_l} \right\} + \frac{C_\delta}{hn_l} \quad (15)$$

$$+ C_2 \frac{\log K}{hn_t}, \quad (16)$$

with $C_\delta = C'_1 + C''_1 \log(1/\delta)$.

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.

Several comments can be made on the basis of the results from Theorem 1:

Methods 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.
- On the other hand, the M2 method permits to keep the whole information given by $\mathcal{L}$. Indeed, 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 former. We focus on this method to ensure that it provides classifiers that have good performance in terms of risk.

Interpretation of the bounds For both M1 and M2 methods, the inequality of Theorem 1 may be divided into two parts:

- (13) and (15) correspond to the pruning algorithm. 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 as 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 (9) used in the pruning algorithm. This shows that, for a sufficiently large temperature α, this criterion allows to select convenient subtrees in term of conditional risk.
  Let us emphasize that the remainder term driving the choice of the penalty is directly proportional to the number of leaves in the M1 method, whereas a multiplicative logarithmic term appears in the M2 method. This term is due to the randomness of the models considered, since the samples used to construct and prune T_{max} are no longer independent.

- (14) and (16) correspond to the final selection of $\tilde{f}$ among the collection of pruned subtrees using $\mathcal{T}$. As $K \leqslant n_l$, this selection adds a term proportional to $\log n_l/n_t$ for both methods, showing that not much is lost when a test sample is used provided that n_t is sufficiently large with respect to $\log n_l$. Nevertheless, since we have no idea of the size of the constant C_2, it is difficult to deduce a general way of choosing $\mathcal{T}$ from this upper bound.

Consistency results Since growing and pruning are independent when applying M1, the VC-dimension V of the set of splits $\mathcal{S}$ only appears with M2. Thus, in this case, the term $\log(n_l/V)$ in the infimum has to be taken into account if V is negligeable in front of n_l. 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 $\tilde{f}$ provided that n_t is conveniently chosen with respect to $\log n_l$.

Role of the margin It has been shown in [25] and in [21] that, under margin assumptions **MA(MN)** and **MA(MT)** respectively, the ERM estimator of f^* on one model is minimax if f^* belongs to some Hölder classes. This means that, under margin assumption **MA(1)**, the upper bound obtained in Theorem 1 for the CART classifier can not be improved. On the other hand, if margin assumption **MA(MT)** is fulfilled, similar bounds are obtained with a remainder term in the infimum proportional to $\left(|\widetilde{T}|/n_l\right)^{\kappa/(2\kappa-1)}$. Since $\kappa > 1$, this term is subbaditive with respect to $|\widetilde{T}|$ (see [32] for full description of subbaditive penalties), so results of [32] can be applied: the subtrees pruned by minimizing a penalized criterion with a penalty proportional to $\left(|\widetilde{T}|/n_l\right)^{\kappa/(2\kappa-1)}$ are subtrees of the CART sequence $(T_k)_{1\leqslant k\leqslant K}$. So, if κ is known, the best solution is to prune T_{max} with the usual pruning algorithm, and then to extract from the sequence obtained in that way the subsequence minimizing the criterion penalized by the subadditive penalty.

Margin dependent penalties It is important to point out that the penalty term suggested by the risk bounds depends on margin parameters, which are usually unknown in practice. To withdraw the margin parameter h under margin assumption **MA(1)**, one prunes T_{max} with the pruning algorithm given in Table 1, and then one uses a test sample or cross-validation to select a subtree. If no margin assumption is fulfilled, the procedure of Scott [32] can be applied, with a penalty term proportional to $\sqrt{|\widetilde{T}|/n_l}$. Otherwise, the margin parameters have to be estimated.

Optimality of the bounds Theorem 1 also shows that the higher the margin, the smaller the risk, which is intuitive since the inverse of the margin plays the role of the classification noise. Actually, to reach optimality in terms of conditional risk, the penalty should be taken as $cst \times \left(h^{-1}|\widetilde{T}|/n_l \wedge \sqrt{|\widetilde{T}|/n_l}\right)$ since, in any case, the remainder term inside the in-

fimum is, at worst, proportional to $\sqrt{|\widetilde{T}|/n_l}$. Hence CART will underpenalize trees for which $h \leqslant \sqrt{|\widetilde{T}|/n_l}$, leading to classifiers having an excessive number of leaves. Nevertheless, the condition $h > \sqrt{|\widetilde{T}_{max}|/n_l}$ can be controlled during the growing algorithm by forcing the maximal tree's construction to stop earlier, for example. This is obviously difficult to do in practice since it heavily depends on the data and on the size of the learning sample, and is worth being investigated more thoroughly.

The two following subsections give 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 4.2 validates the discrete selection by test-sample.

4.1 Validation of the Pruning algorithm

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 2.3. We show that, for a convenient constant α, $\hat{f}_{T_\alpha}$ is not far from f^* in terms of its conditional risk. Let us emphasize that the subsample $\mathcal{T}$ plays no role in the two following results.

4.1.1 $\tilde{f}$ constructed via M1

Here we assume that $\mathcal{L} = \mathcal{L}_1 \cup \mathcal{L}_2$. Thus 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 $\mathcal{L}_2$, the selection is made 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 (11). Let P_{n_2} be the empirical misclassification rate on $\mathcal{L}_2$ as defined by (7). Then let us consider the following:

- For $T \preceq T_{max}$, $\hat{f}_T = \operatorname{argmin}_{f \in \mathcal{F}_T} [\mathrm{P}_{n_2}(f)]$,
- For $\alpha > 0$, T_α is the smallest minimizing subtree for the temperature α as defined in subsection 2.3 and $\hat{f}_{T_\alpha} = \operatorname{argmin}_{f \in \mathcal{F}_{T_\alpha}} [\mathrm{P}_{n_2}(f)]$.

Proposition 1. *Let $P_{\mathcal{L}_2}$ be the product distribution on $\mathcal{L}_2$ and let h be the margin given by* **MA(1)**. *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

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

on a set Ω_ξ such that $P_{\mathcal{L}_2}(\Omega_\xi) \geqslant 1 - \Sigma_\alpha e^{-\xi}$, where l is defined by (5), $C_1(\alpha) > \alpha_0$ and Σ_α are increasing with α.

We obtain a trajectorial non-asymptotic risk bound on a large probability set, leading to the conclusions given for Theorem 1. Nevertheless, taking an excessive temperature α will overpenalize and 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 suggests that both sides of the inequality grow with α. The choice of the convenient temperature is then critical 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.

4.1.2 $\tilde{f}$ constructed via M2

Here we define the different empirical risks, expected loss and estimators exactly in the same way as in subsection 4.1.1, although l is replaced by the empirical expected loss λ on $X_1^{n_l} = \{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\}$ defined in Definition 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_l:

Proposition 2. *Let $P_{\mathcal{L}}$ be the product distribution on $\mathcal{L}$, λ be the empirical expected loss computed on $\{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\}$, and let h be the margin given by* **MA(1)**. *Let $\xi > 0$ and*

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

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

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

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

We obtain a similar trajectorial non-asymptotic risk bound on a large probability set. The same conclusions as those derived from M1 hold in this case. Let us just mention that the remainder term $h^{-1}\alpha_{n_l,V}|\widetilde{T}|/n_l$ in the risk bound takes into account the complexity of the collection of trees having $|\widetilde{T}|$ leaves which can be constructed on $\{X_i \ ; \ (X_i, Y_i) \in \mathcal{L}\}$. 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 a uniform bound.

Example: Let us consider the case where $\mathcal{S}$ is the set of all half-spaces of $\mathcal{X} = \mathbb{R}^d$ with axis-parallel frontiers. In this case, if $d \geqslant 3$,

$$\frac{\log(d)}{\log 2} - 1.18 \leqslant V \leqslant d,$$

consequently, if $n_l \geqslant d$, we obtain a penalty proportional to

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

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

4.2 Final Selection

We focus here on the selection of the classifier $\tilde{f}$ among the collection $(\hat{f}_{T_k})_{1\leqslant k\leqslant K}$ provided by the pruning algorithm as defined in subsection 2.3. Let us recall that $\tilde{f}$ is defined by

$$\tilde{f} = \underset{\{\hat{f}_{T_k};1\leqslant k\leqslant K\}}{\operatorname{argmin}} \left[\mathrm{P}_{n_t}(\hat{f}_{T_k})\right],$$

where P_{n_t} is the empirical misclassification rate on $\mathcal{T}$ defined by (7).
The performance of this classifier can be compared to the performance of the collection $(\hat{f}_{T_k})_{1\leqslant k\leqslant K}$ by the following:

Proposition 3.
Let λ be the loss defined in Definition 1. For both methods M1 and M2, there

exist three absolute constants $C'' > 1$, $C_1' > 3/2$ *and* $C_2' > 3/2$ *such that*

$$\mathbb{E}\left[\lambda(f^*, \tilde{f}) \mid \mathcal{L}\right] \leqslant C'' \inf_{1 \leqslant k \leqslant K} \lambda(f^*, \hat{f}_{T_k}) + C_1' \, h^{-1} \frac{\log K}{n_t} + h^{-1} \frac{C_2'}{n_t},$$

where K *is the number of pruned subtrees extracted during the pruning algorithm.*

5 Concluding Remarks

We have proven that CART provides convenient classifiers in terms of conditional risk under the margin assumption **MA(1)**. As for the regression case, the properties of the growing algorithm need to be analyzed to obtain full unconditional upper bounds. Results on the performance of theoretical procedures in which CART is viewed as a forward algorithm to approximate an ideal, but intractable, binary tree are given in [13]. Although they do not validate any concrete algorithm as done here, these results confirm that the penalty term used in penalized criterion (9) is well chosen under **MA(1)**.

The remarks made after Theorem 1 on the size of the margin h enlarge our perspectives for the application of CART in practice. Among such perspective, we may

- use the slope heuristic (see for example [3]) to select a classifier among a collection,
- search for a robust manner to determine if the margin assumption is fulfilled, allowing to use the blind selection by test sample.

Some track to estimate the margin h if assumption **MA(1)** is fulfilled could be to use mixing procedures as boosting (see [6] [10] for example). Hence, this estimate could be used in the penalized criterion to help find the convenient temperature. It could also give an idea of the difficulty to classify the considered data and henceforth to help choose the most adapted classification method.

Acknowledgements

I would like to thank an anonymous referee for numerous remarks and suggestions which helped to improve the presentation of this paper.

6 Proofs

Let us start with a preliminary result.

6.1 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 independent copies of (X, Y). Then given two classifiers f and g, let us define

$$d_n^2(f, g) = \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - g(X_i)\right)^2 .$$

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 $\mathcal{M}_n^*$ only depends 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 $f \in \mathcal{F}_T$ and any $\sigma > 0$, define

$$B_T(f, \sigma) \ = \ \{g \in \mathcal{F}_T \ ; \ d_n(f, g) \leqslant \sigma\}$$

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

$$\overline{\mathrm{P}}_n(f) = \mathrm{P}_n(f) - \mathbb{E}[\mathrm{P}_n(f) \mid X_1^n], \tag{17}$$

where P_n is defined for any given classifier f by

$$\mathrm{P}_n(f) \ = \ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) \neq Y_i}.$$

Remark 4. If P_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 population distance

$$d^2(f, g) = \mathbb{E}\left[(f(X) - g(X))^2\right]$$

instead of its empirical version d_n.

We have the following result:

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

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

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

$$\mathbb{1}_{g(x)\neq y} - \mathbb{1}_{f(x)\neq y} = (g(x) - f(x))(1 - 2\mathbb{1}_{y=1}),$$

yielding

$$\begin{aligned}\overline{\mathrm{P}}_n(g) - \overline{\mathrm{P}}_n(f) &= \frac{1}{n}\sum_{i=1}^n (g(X_i) - f(X_i))\,(1 - 2\mathbb{1}_{Y_i=1}) \\ &\quad -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (g(X_i) - f(X_i))\,(1 - 2\mathbb{1}_{Y_i=1}) \mid X_1^n\right].\end{aligned}$$

Let us now consider a Rademacher sequence of random signs $(\varepsilon_i)_{1\leqslant i\leqslant n}$ independent of $(X_i, Y_i)_{1\leqslant i\leqslant n}$. Then, one has by a symmetrization argument

$$\mathbb{E}\left[\sup_{g \in B_T(f,\sigma)} |\overline{\mathrm{P}}_n(g) - \overline{\mathrm{P}}_n(f)| \mid X_1^n\right] \leqslant \mathbb{E}\left[\sup_{g \in B_T(f,\sigma)} \frac{2}{n}\left|\sum_{i=1}^n \varepsilon_i (g(X_i) - f(X_i))(1 - 2\mathbb{1}_{Y_i=1})\right| \mid X_1^n\right].$$

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

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

where each (a_t, b_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 $g \in B_T(f,\sigma)$, $d_n^2(f,g) = \sum_{t\in\widetilde{T}}(a_t - b_t)^2 \leqslant \sigma^2$, we obtain that

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

Finally, since $(\varepsilon_i)_{1\leqslant i\leqslant n}$ and $(1-2\mathbb{1}_{Y_i=1})_{1\leqslant i\leqslant n}$ take their values in $\{-1;1\}$, $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are centered and independent of $(X_i,Y_i)_{1\leqslant i\leqslant n}$, and since, by definition, for each $t\in\widetilde{T}$ $n^{-1}\sum_{i=1}^n\varphi_t^2(X_i)=1$, Jensen's inequality implies

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

□

6.2 Proof of Proposition 1

To prove Proposition 1, we adapt results from Massart [23, Theorem 4.2], and Massart and Nédélec [25] (see also Massart *et.al.* [24]).

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 $f^*\in\mathcal{F}\subset\{f:\mathcal{X}\mapsto[0,1]\ ;\ f\in\mathbb{L}^2(\mathcal{X})\}$ 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 $\mathrm{pen}_n:\mathcal{M}_n\longrightarrow\mathbb{R}_+$, and $\gamma:\mathcal{F}\times(\mathcal{X}\times[0,1])\longrightarrow\mathbb{R}_+$ a contrast function, i.e. γ such that $f\mapsto\mathbb{E}\left[\gamma(f,(X,Y))\right]$ is convex and minimum at point f^*. Hence define for all $f\in\mathcal{F}$ the expected loss $l(f^*,f)=\mathbb{E}\left[\gamma(f,(X,Y))-\gamma(f^*,(X,Y))\right]$.
Finally let

$$\gamma_n=\frac{1}{n}\sum_{i=1}^n\gamma(.,(X_i,Y_i)) \tag{18}$$

be the empirical contrast associated with γ. For example, in the classification context, $\gamma(f,(x,y))=\mathbb{1}_{f(x)\neq y}$, leading to the classical loss as defined by (5), and the classical empirical misclassification rate P_n as defined by (7). Hence, if the collection of models $\mathcal{M}_n$ has finite-dimensional models with dimension $|m|$, the penalty function can be taken as $\mathrm{pen}_n(m)=cst\ \times\ |m|$ for instance. Then let $\hat{m}$ be defined as

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

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

$$\tilde{f}=\hat{f}_{\hat{m}}. \tag{19}$$

One makes the following assumptions:
$\mathbf{H_1}$: γ is bounded by 1, which is not a restriction since all the functions we consider take values in $[0,1]$.
$\mathbf{H_2}$: Assume there exist $c \geqslant (2\sqrt{2})^{-1/2}$ and some (pseudo-)distance d such that, for every pair $(f,g) \in \mathcal{F}^2$, one has

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

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

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

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

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

where d is given by assumption $\mathbf{H_2}$. 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

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

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 $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ be a sample of independent 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 $\mathcal{F} \subset \{f : \mathcal{X} \mapsto [0,1] \ ; \ f \in \mathbb{L}^2(\mathcal{X})\}$. Consider some penalty function $\mathrm{pen}_n : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ and the corresponding penalized estimator $\tilde{f}$ (19) of the target function f^*. Take a family of weights $(x_m)_{m \in \mathcal{M}_n}$ such that*

$$\Sigma = \sum_{m \in \mathcal{M}_n} e^{-x_m} < +\infty. \tag{20}$$

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 $1 - \Sigma e^{-\xi}$,

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

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. The proof is inspired from Massart [23] and Massart *et.al.* [24]. We give only sketches of proofs since those are now routine results in the model selection area (see [24] for a fuller overview).

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

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

lead to the following inequality:

$$l(f^*, \tilde{f}) \leqslant l(f^*, f_m) + \bar{\gamma}_n(f_m) - \bar{\gamma}_n(\tilde{f}) + \text{pen}_n(m) - \text{pen}_n(\hat{m}) \qquad (21)$$

where $\bar{\gamma}_n$ is defined by (17). 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 $\text{pen}_n(\hat{m})$. 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 $f \in \mathcal{F}_{m'}$, let us define

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

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

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

Then (21) becomes

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

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

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

where ν_n is the recentered empirical measure, we bound $V_{m'}$ uniformly in $m' \in \mathcal{M}_n$ by using Rio's version of Talagrand's inequality, whose first version can

be found in [29], and 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 $f \in \mathcal{F}_{m'}$,

$$\left|\frac{\gamma(f,.) - \gamma(f_m,.)}{w_{m'}(f)}\right| \leqslant \frac{1}{y_{m'}^2}. \tag{22}$$

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

$$\mathrm{Var}\left[\frac{\gamma(f,(X,Y)) - \gamma(f_m,(X,Y))}{w_{m'}(f)}\right] \leqslant \frac{c^2}{4y_{m'}^2}. \tag{23}$$

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 (20). 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 now need 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 techniques similar to Massart *et al.*'s [25], 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 $m' \in \mathcal{M}_n$, let $u_{m'} \in \mathcal{F}_{m'}$ be defined by

$$l(f^*, u_{m'}) \leqslant 2 \inf_{z\in\mathcal{F}_{m'}} l(f^*, z).$$

Then we have

$$\mathbb{E}(V_{m'}) \leqslant \mathbb{E}\left[\sup_{z\in\mathcal{F}_{m'}} \frac{|\bar{\gamma}_n(z)-\bar{\gamma}_n(u_{m'})|}{w_{m'}(z)}\right] + \mathbb{E}\left[\frac{|\bar{\gamma}_n(u_{m'})-\bar{\gamma}_n(f_m)|}{\inf_{z\in\mathcal{F}_{m'}}[w_{m'}(z)]}\right].$$

For every $z \in \mathcal{F}_{m'}$, let

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

Then, since

$$\begin{aligned} l(f^*,z) &= \mathbb{E}\left[\gamma(z,(X,Y)) - \gamma(f^*,(X,Y))\right] \\ l(f^*,z) &= l(f^*,u_{m'}) + \mathbb{E}\left[\gamma(z,(X,Y)) - \gamma(u_{m'},(X,Y))\right], \end{aligned}$$

Then we have

$$l(f^*,z) \leqslant \omega^2_{m'}(z) \leqslant 5\ l(f^*,z). \tag{24}$$

On the one hand we have $w_{m'}(z) \geqslant l(f^*,z) + y^2_{m'} \geqslant (1/5)\omega^2_{m'}(z) + y^2_{m'}$ for every $z \in \mathcal{F}_{m'}$. Hence

$$\mathbb{E}\left[\sup_{z\in\mathcal{F}_{m'}} \frac{|\bar{\gamma}_n(z)-\bar{\gamma}_n(u_{m'})|}{w_{m'}(z)}\right] \leqslant 5\ \mathbb{E}\left[\sup_{z\in\mathcal{F}_{m'}} \frac{|\bar{\gamma}_n(z)-\bar{\gamma}_n(u_{m'})|}{\omega^2_{m'}(z)+5y^2_{m'}}\right].$$

Furthermore we have

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

and, if $l(f^*,z) \leqslant \varepsilon^2$, then $l(f^*,u_{m'}) \leqslant 2\varepsilon^2$ and $d(z,u_{m'}) \leqslant d(f^*,z) + d(s*,u_{m'}) \leqslant c\varepsilon + c\varepsilon\sqrt{2}$. Hence we get that $d(z,u_{m'}) \leqslant (1+\sqrt{2})c\varepsilon \leqslant 2c\varepsilon\sqrt{2}$. Let us now suppose that $\varepsilon \geqslant \varepsilon_{m'}$. Then we have by monoticity of $x \mapsto \phi(x)/x$ and by definition of $\varepsilon_{m'}$ that

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

since $c \geqslant (2\sqrt{2})^{-1/2}$.
So, by assumption $\mathbf{H_3}$, we finally obtain that, for all $\varepsilon \geqslant \varepsilon_{m'}$,

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

So we can apply Lemma 5.5 in [25] and use the monoticity of $x \mapsto \phi_{m'}(x)/x$ to obtain that

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

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

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

On the other hand, let us notice that

$$\begin{aligned}\inf_{z\in\mathcal{F}_{m'}} w_{m'}(z) &\geqslant 2y_{m'} \inf_{z\in\mathcal{F}S_{m'}} [\sqrt{l(f^*,z)}+\sqrt{l(f^*,f_m)}] \\ &\geqslant \frac{y_{m'}\sqrt{2}}{c} d(u_{m'},f_m),\end{aligned}$$

hence

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

leading by Jensen's inequality to

$$\mathbb{E}\left[\frac{|\bar{\gamma}_n(u_{m'})-\bar{\gamma}_n(f_m)|}{\inf_{z\in\mathcal{F}_{m'}}[w_{m'}(z)]}\right] \leqslant c(y_{m'}\sqrt{2})^{-1}\frac{\sqrt{\mathrm{Var}\left[\bar{\gamma}_n(u_{m'})-\bar{\gamma}_n(f_m)\right]}}{d(u_{m'},f_m)} \leqslant \frac{c}{y_{m'}\sqrt{2n}}.$$

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

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

So we finally obtain that, on the set Ω_ξ,

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

with

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

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, on the one hand,

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

and, on the other hand,

$$w_{\hat{m}}(\tilde{f}) \leqslant 2l(f^*, \tilde{f}) + 2l(f^*, f_m) + y_{\hat{m}}^2.$$

Hence the following inequality holds on Ω_ξ for any $m \in \mathcal{M}_n$ and any $f_m \in \mathcal{F}_m$:

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

with

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

□

Application to classification trees:

Let us now suppose that (X, Y) takes values in $\mathcal{X} \times \{0, 1\}$. The contrast is taken as $\gamma(f, (X, Y)) = \mathbb{1}_{f(X) \neq Y}$, the expected loss is defined by (5), 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 2 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 f and g

$$\begin{aligned}(\gamma(f, (X, Y)) - \gamma(g, (X, Y)))^2 &= \left(\mathbb{1}_{Y \neq f(X)} - \mathbb{1}_{Y \neq g(X)}\right)^2 &(26)\\ &= (f(X) - g(X))^2. &(27)\end{aligned}$$

Then, if we take $d^2(f,g) = \mathbb{E}\left[(f(X) - g(X))^2\right]$, we have that, for all classifiers f and g, $\mathrm{Var}\left[\gamma(g,(X,Y)) - \gamma(f,(X,Y))\right] \leqslant d^2(f,g)$. Moreover, with the margin condition **MA(1)**, we have that

$$l(f^*, f) \;\geqslant\; h d^2(f^*, f), \tag{28}$$

hence assumption $\mathbf{H_2}$ is checked with d and $c^2 = 1/h$, where h is the margin. By definition of h, we have $h \leqslant 1 \leqslant 2\sqrt{2}$, and then $c \geqslant (2\sqrt{2})^{-1/2}$.
Then, assumption $\mathbf{H_3}$ is checked by Lemma 1 with $\phi_T(x) = 2x\sqrt{|\widetilde{T}|/n}$. Hence, Theorem 2 is verified with $\varepsilon_T = \sqrt{1/h}\sqrt{|\widetilde{T}|/n}$.
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 [14], we immediately obtain $\Sigma_\alpha = \Sigma_\theta < +\infty$. Then, we get proposition 1 by Theorem 2.

6.3 Proof of Proposition 2

Let $n = n_l$ and let X_1^n denote the sample $\{X_i \,;\, (X_i, Y_i) \in \mathcal{L}\}$.

First we generalize Theorem 2 to random models, and then we apply it to CART. Let (X,Y), $\mathcal{F}$, $f^* \in \mathcal{F}$, $\mathcal{L} = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, γ and γ_n be defined as in subsection 6.2. Finally, let us rewrite the expected loss of $f \in \mathcal{F}$ conditionally on X_1^n as in Definition 1, that is

$$\lambda(f^*, f) = \mathbb{E}\left[\mathrm{P}_{n_l}(f) - \mathrm{P}_{n_l}(f^*) \mid X_1^n\right].$$

Let us consider a collection of at most countable models $(\mathcal{F}_m)_{m \in \mathcal{M}_n^*}$ and a sub-collection $(\mathcal{F}_m)_{m \in \mathcal{M}_n}$, 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 $\hat{f}_m = \mathrm{argmin}_{f \in \mathcal{F}_m} \gamma_n(f)$ 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 $(g,f) \in \mathcal{F}^2$, one has

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

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

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

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

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

where d_n is given by assumption $\mathbf{H_2}$. Let $\bar{\gamma}_n$ be defined as (17). 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

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

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 3. *Let $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample of independent 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 $\mathcal{F} \subset \{f : \mathcal{X} \mapsto [0, 1] \ ; \ f \in \mathbb{L}^2(\mathcal{X})\}$ (which may depend 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 $\mathcal{L}$, and some penalty function $\mathrm{pen}_n : \mathcal{M}_n \longrightarrow \mathbb{R}_+$. Let $\tilde{f}$ (19) be the corresponding penalized estimator of the target function f^*. Take a family of weights $(x_m)_{m \in \mathcal{M}_n^*}$ such that*

$$\sum_{m \in \mathcal{M}_n^*} e^{-x_m} \leqslant \Sigma < +\infty, \tag{29}$$

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 $1 - 2\Sigma e^{-\xi}$,

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

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

Proof. The proof is highly similar to that of Theorem 2. 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 finally obtain 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 $f_m \in \mathcal{F}_m$. Starting from (21), we have

$$\lambda(f^*, \tilde{f}) \quad \leqslant \quad \lambda(f^*, f_m) + w_{\hat{m},m}(\tilde{f}) V_{\hat{m},m} + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}), \qquad (30)$$

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

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

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

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 2 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 $\mathbf{H_2}$ allows 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 (22) and (23). By using the same techniques as in the proof of Theorem 2 and the same considerations as in [14], we obtain that

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

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

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

So we have

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

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

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

on a set Ω_ξ such that $P\left(\Omega_\xi \mid X_1^n\right) \geqslant 1-2\Sigma e^{-\xi}$. 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 2. □

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 algorithm in CART (see subsection 4.1) as recalled in subsection 6.2. Since the growing and the pruning algorithms are made on the same sample $\mathcal{L}$, the conditions of Theorem 3 hold. Since $n = n_l$ 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 $\mathcal{S}$ 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 $(Y_1, \ldots, Y_n)$. Then $\{T \preceq T_{max}\} \subset \mathcal{M}_n^*$ and we are able to apply Theorem 3. Considering (26), we take

$$d_n^2(f,g) = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i)-g(X_i)\right)^2,$$

corresponding with the distance d given in Definition 2. Using the margin condition **MA(1)**, (28) is also verified for λ and d_n, and we have assumption $\mathbf{H_2}$ with $c^2 = 1/h$. 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 $\varepsilon_T = \sqrt{1/h}\sqrt{|\widetilde{T}|/n}$.
Finally, to choose a convenient family of weights $(x_{_T})_{T\in\mathcal{M}_n^*}$, taking (see [14])

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

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.

6.4 Proof of Proposition 3

Proposition 3 is a direct application of the theorem obtained by Blanchard and Massart in [18], reformulated for our purpose here: assume that we observe $N + n$ independent 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 remaining 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 (18) (with (X, Y) replaced by Z). Hence, we have the following result.

Theorem 6.4.1 (Blanchard and Massart [18])**.**
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 non-increasing, and which satisfies for all $\varepsilon > 0$

$$\sup_{\{f\in\mathcal{F}\ ;\ l(f^*,f)\leqslant\varepsilon^2\}} \mathrm{Var}\left[\gamma(f, Z) - \gamma(f^*, Z)\right] \leqslant w(\varepsilon). \tag{31}$$

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

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

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

6.5 Proof of Theorem 1

We are now able to prove Theorem 1 via propositions 1, 2 and 3. The beginning of the proof remains the same if $\tilde{f}$ is constructed either via M1 or M2. So we just give the first step of 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 $(T_k)_{1 \leqslant k \leqslant K}$,

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

Starting from this inequality, if $\tilde{f}$ is constructed via M1, 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.

Yet, if $\tilde{f}$ is constructed via M2, we apply Proposition 2 with $\alpha = 2\alpha_0\alpha_{n_1,V}$ and, for each $\delta \in]0;1[$, $\xi = \log(2\Sigma_\alpha/\delta)$. Then, we obtain Theorem 1 with the appropriate constants.

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. *Journal of Machine Learning Research 10* (2009), 245–279.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] Gey, S., and Mary Huard, T. Risk bounds for embedded variable selection in classification trees. Tech. rep., arxiv, 1108.0757v1, 2011.

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

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

[16] 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.

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

[18] 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.

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

[20] 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.

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

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

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

[24] 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.

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

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

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

[28] 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.

[29] 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.

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

[31] 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.

[32] Scott, C. Tree pruning with subadditive penalties. *IEEE Transactions on Signal Processing 53*, 14 (2005), 4518–4525.

[33] Scott, C., and Nowak, R. Minimax-optimal classification with dyadic decision trees. *IEEE Trans. on Information Theory 52*, 4 (2006), 1335–1353.

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

[35] 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.

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

[37] Vapnik, V. N., and Chervonenkis, A. Y. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow, 1974.

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