Risk Bounds for Embedded Variable Selection in Classification Trees

Servane Gey, Tristan Mary-Huard

Risk Bounds for Embedded Variable Selection in Classification Trees

Servane Gey[1] & Tristan Mary-Huard[2]

Servane.Gey@parisdescartes.fr, maryhuar@agroparistech.fr

[1]Laboratoire MAP5, UMR 8145, Université Paris Descartes, Paris, France

[2]UMR AgroParisTech INRA MIA 518, Paris, France

June 22, 2012

Abstract

The problems of model and variable selections for classification trees are jointly considered. A penalized criterion is proposed which explicitly takes into account the number of variables, and a risk bound inequality is provided for the tree classifier minimizing this criterion. This penalized criterion is compared to the one used during the pruning step of the CART algorithm. It is shown that the two criteria are similar under some specific margin assumptions. In practice, the tuning parameter of the CART penalty has to be calibrated by hold-out. Simulation studies are performed which confirm that the hold-out procedure mimics the form of the proposed penalized criterion.

Keywords: Classification Tree, Variable Selection, Statistical Learning Theory

1 Introduction

Since the pioneering work of Breiman *et al.* [6], classification trees have become a classical tool in machine learning. In particular, the Classification and Regression Tree (CART) algorithm is a well-established algorithm to build and prune tree predictors. This algorithm has been successfully applied in various fields, see for instance [1, 7, 10, 34].

1.1 Building/selecting a tree

The process of building (or choosing) a tree classifier from a training set can be summarized into an optimization problem, where the goal is to find the "best" tree classifier $\hat{f}$ satisfying

$$\hat{f} = \arg\min_{f_T} \left(\mathrm{P}_n f_T + pen(n, T) \right) \quad , \tag{1.1}$$

where n is the number of observations, $\mathrm{P}_n \hat{f}_T$ is the empirical risk of tree classifier $\hat{f}_T$ based on tree T, and $pen(n, T)$ is a penalty function based on the size of the training set and on the characteristics of T.

Obtaining the best tree classifier $\hat{f}$ necessitates to solve a non-convex function over a large set of trees, something unfeasible in practice. As an alternative, a 2-step heuristic approach to solve this problem has been proposed in [6], in the particular case where the penalized criterion is of the form

$$pen(n, T) = \alpha_n \times |T| \quad , \tag{1.2}$$

where α_n is a tuning parameter that depends on n, and $|T|$ is the size of the tree, i.e. the number of leaves (terminal nodes) of T. In the first step (called the *growing step*) a large tree T_{max} that achieves a perfect classification on the training set is built. Then, during the second step (called the *pruning step*), the optimal subtree is obtained from the large tree, where the optimal subtree satisfies

$$\hat{f}_{prun} = \operatorname*{arg\,min}_{f_T, \ T \subseteq T_{max}} \ \mathrm{P}_n f_T + \alpha_n \times |T| \quad .$$

While this heuristic approach is at the heart of the CART algorithm and is probably the most popular strategy to prune a tree, one should keep in mind that the actual goal is in fact to solve Problem (1.1), and to obtain the properties of $\hat{f}$, whatever the (approximate) strategy that is applied to find it.

From a theoretical point of view, many works have investigated the performance of the tree classifier resulting from the pruning step of CART rather than from the generic optimization problem. In the Gaussian or bounded regression context, penalty (1.2) was validated in [14] using model selection framework. Another validation was obtained in the classification framework in [28]. More recently, a refined analysis of the pruning step was proposed in [12], where margin adaptive risk bounds were obtained in the binary classification context. Importantly, these theoretical results are actually obtained conditionally to the construction of T_{max}. This means that only the performance of the pruning step is assessed, while the growing step is not taken into account.

1.2 Classification trees and variable selection

Because they are based on the 2-step heuristic of the CART algorithm, results obtained so far fail to take into account the complete process of obtaining a tree classifier. In particular, the embedded variable selection process that is inherent to tree classification algorithms has never been investigated. A variable selection process is called embedded when it is included in the training step of the classification algorithm. Therefore the learning and variable selection processes cannot be separated. This embedded property is actually one of the main arguments for the use of tree classifiers to deal with large dimension data (see

[5, 11, 13] for example). Note that in the CART algorithm, the inner selection process results from the recursive growing strategy of the tree: at each node, the "best" variable is selected among all for splitting. As a result, in many cases the maximal tree (and consequently all of its subtrees) only includes a small subset of the p initial variables. As a consequence, as long as tree classifiers are studied through the pruning step of the CART heuristic (hence conditionally to the growing step), it is impossible to investigate the complete variable selection process.

Although the embedded variable selection process is well-known ([8, 15, 20]), it may appear at first glance that it is not correctly handled in the optimization program

$$\hat{f} = \arg\min_{f_T} \left(\mathrm{P}_n f_T + \alpha_n \times |T| \right) \quad , \tag{1.3}$$

assuming the form of the penalty proposed in [6] is correct. Indeed, this penalized criterion does not obviously depend on the total number of covariates p. This can be astonishing: in both the regression and classification frameworks, theoretical studies have shown that in the variable selection context, an extra term should be added to the penalty that is used when only one model is considered per dimension ([2, 24]) to obtain oracle-type inequalities. Since the collection of possible trees increases with p, p should play a crucial role in the regularization term.

Since parameter p does not explicitly appear in criterion (1.3), one can argue that p is hidden in the constant term α_n. This argument is verified from at least two penalties that can be exhibited from previous works:

- In [28] (equation 4), the penalty term has the form

$$
\begin{aligned}
pen(|T|, n) &= C_1 \times \sqrt{\frac{|T| p \log n}{n}} \\
&= \sqrt{Cst \frac{p \log n}{n}} \times \sqrt{|T|} \\
&= \alpha(p, n) \sqrt{|T|} \quad ,
\end{aligned}
$$

- In [12] (Theorem 1), the penalty term is of order

$$
\begin{aligned}
pen(|T|, n) &\approx C_2 \times \frac{p \log (p)(1 + \log(n/\log (p)))}{n} \times |T| \\
&= \alpha(p, n)|T| \quad ,
\end{aligned}
$$

where C_1 and C_2 are known constants. While these two penalty functions depend on p, one can observe that their scaling order is much larger than the $\log(p)$ usually obtained in the variable selection context [2, 24].

3

1.3 Contribution

The goal of the present paper is to investigate the classification performance of the tree classifier obtained by solving Problem 1.1, and to decipher the exact impact of variable selection on tree classifier selection. While this impact is theoretically studied through an ideal exhaustive selection procedure (unfeasible in practice), it sheds light on the heuristic procedures currently used in practice to mimic the ideal one (see Section 3.2). From a theoretical point of view, we consider the model selection problem where the goal is to select a candidate from *all* possible tree classifiers. The strategy consists in choosing the candidate minimizing a penalized criterion that depends on parameters p and n. In this model selection context, we exhibit a penalization function where the variable selection process is explicitly taken into account, and provide performance guarantees for the candidate tree classifier through an upper bound of its risk. Then it is shown that the impact of variable selection, although investigated via the theoretical minimization problem (1.1), can also be exhibited in practice for practical heuristic approaches. More precisely, a simulation study is performed which shows that the proposed theoretical penalization function is actually the one that is implicitly used in the pruning step of the CART algorithm.

The paper is organized as follows. Section 2 presents the framework of binary classification and describes tree classifiers. The main theoretical contribution and the simulation study are presented in Section 3. Some discussion is developed in Section 4, and finally Section 5 gives the proofs of the results presented in Section 3.

2 Context

2.1 Classification framework

The considered classification framework is the following. Suppose one observes a sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of n independent copies of the random variable (X, Y), where the explanatory variable X takes values in a measurable space $\mathcal{X}$ of dimension $p \geqslant 2$, and is associated with a label Y taking values in $\{0, 1\}$. Suppose moreover that each coordinate of $\mathcal{X}$ is ordered (i.e. $\mathcal{X}$ is a product of p ordered subspaces). A classifier is then any function f mapping $\mathcal{X}$ into $\{0, 1\}$. The quality of a classifier is measured by its misclassification rate

$$\mathrm{P}f := \mathrm{P}(f(X) \neq Y) \ , \tag{2.1}$$

where P denotes the joint distribution of (X, Y). If the joint distribution of (X, Y) 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

$$f^*(x) = \mathbb{1}_{\eta(x) \geqslant 1/2} \ , \tag{2.2}$$

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

$$\eta(x) = \mathrm{P}\left[Y = 1 \mid X = x\right] \quad . \tag{2.3}$$

As P is unknown, the goal is to construct from 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

$$l(f^*, \tilde{f}) = \mathrm{P}(\tilde{f}(X) \neq Y) - \mathrm{P}(f^*(X) \neq Y) \tag{2.4}$$

is as small as possible.
Many strategies or classification algorithms have been proposed to build $\tilde{f}$ (see [16], [3] for an overview). The quality of a strategy is measured by its risk

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

where the expectation is taken with respect to the sample distribution. In the model selection framework, two strategies are usually considered:

- Empirical Risk Minimization: $\tilde{f}$ is chosen as the minimizer of

$$\mathrm{P}_n f := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{f(X_i) \neq Y_i\}} \quad , \tag{2.5}$$

  over all classifiers f belonging to a single class of classifiers,

- Structural Risk Minimization: $\tilde{f}$ is chosen as the minimizer of the penalized empirical risk over a collection of classes.

2.2 Margin assumptions

It is now well known that without any assumption on the joint distribution P, when considering a class of classifiers with finite Vapnik Chervonenkis (VC) dimension, the minimax convergence rate of the risk bound is of order $\mathcal{O}(1/\sqrt{n})$. It has also been shown that, under the overoptimistic zero-error assumption (that is $Y = \eta(X)$ almost surely, where η is defined by (2.3)), this minimax convergence rate is at best of order $\mathcal{O}(1/n)$ (see [33, 22] for example).

These two extreme cases can be modulated by so-called *margin assumptions* that make the link between the "global" pessimistic case (without any assumption on P) and the zero-error case ([18, 19, 23, 27, 26, 31, 32]).

In this paper, we consider the margin assumption proposed in [23]:

MA(1) There exist some constants $C_0 > 0$ and $\kappa > 1$ such that, for all $t > 0$,

$$\mathrm{P}\left(|2\eta(X) - 1| \leqslant t\right) \leqslant C_0 \, t^{\frac{1}{\kappa - 1}}, \tag{2.6}$$

5

Note that by taking $t = h \in]0,1[$ and the limit value $\kappa = 1$, we obtain the stronger assumption proposed in [27] (see also the slightly weaker condition proposed in [17]):

MA(2) There exists $h \in]0;1[$ such that

$$P\left(|2\eta(X) - 1| \leqslant h\right) = 0. \tag{2.7}$$

Assumption **MA(2)** has an intuitive interpretation. It means that (X, Y) is sufficiently well distributed to ensure that there is no region in 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. From a general point of view, the margin parameter quantifies the noise level of the classification problem, and may be understood as the equivalent of the variance parameter in the Gaussian model selection setting.

2.3 Tree classifiers, classes of tree classifiers

A tree T is a structure that can be represented as a hierarchy whose elements are called nodes. For binary trees, each node has either 0 or 2 children (called Left and Right). The initial node is called the root of the tree and a node with no child is called a leaf. The size of tree T is defined as the number of its leaves and noted $|T|$ in the following. In this paper, we define a tree $T_{c\ell}$ by two elements:

- its configuration c, i.e. the hierarchy between the nodes: for instance, in Figure 1, we know that node 6 is the Left child node of node 3, and so on,

- the ordered list ℓ of variables that appear at each node, i.e. the kth variable in the list appears in node k.
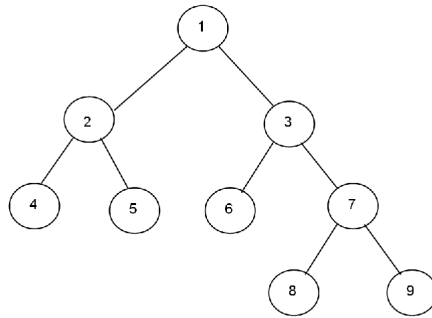

Figure 1: Tree configuration example: for each node, the parent and child nodes are known.

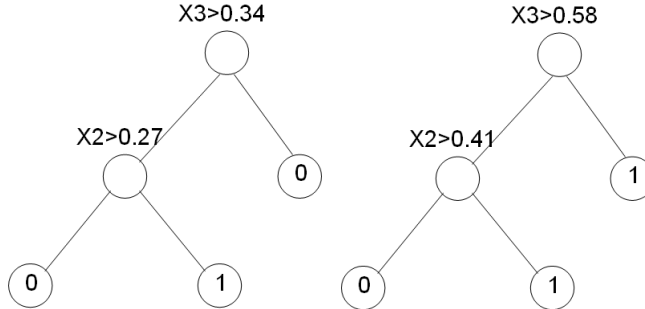A tree classifier f based on tree $T_{c\ell}$ associates

Figure 2: Two tree classifiers that belong to the same class.

- at each internal node a condition of the form "$X^{j_k} > s^{k}$", where j_k is the index of the variable associated with node k and s^k is a threshold,

- at each terminal node a label (here 0 or 1).

Therefore, an observation $x \in \mathcal{X}$ will be classified as follows: starting at the root, observation x will move from a node of f to another using the following rule: at node k, if "$x^{j_k} > s^{k}$" then x moves to Right, otherwise it moves to Left. At the end of the process, x will be classified according to the label of the leaf it reaches.

To summarize, a tree classifier associated with tree $T_{c\ell}$ splits $\mathcal{X}$ into $|T_{c\ell}|$ regions each associated with a label, and two classifiers associated with the same tree $T_{c\ell}$ differ in that the thresholds (for internal nodes) and labels (for leaves) are not the same. An example of two such tree classifiers is given in Figure 2. In the following, we will consider classes $\mathcal{C}_{c\ell} = \{f \ / f \text{ based on } T_{c\ell}\}$ of classifiers based on a same tree $T_{c\ell}$.

Finally, we define

$$\overline{f}_{c\ell} \in \arg\min_{f \in \mathcal{C}_{c\ell}} \mathrm{P}f, \tag{2.8}$$

where $\mathrm{P}f$ is defined by (2.1).

3 Results

3.1 Risk bounds

We first consider a single class $\mathcal{C}_{c\ell}$ of tree classifiers and its associated empirical risk minimizer

$$\widehat{f}_{c\ell} \in \arg\min_{f \in \mathcal{C}_{c\ell}} \mathrm{P}_n f,$$

where $\mathrm{P}_n f$ is defined by (2.5).

7

Proposition 1. *Assume that margin assumption* **MA(1)** *is verified. For all* $t_{c\ell} > 0$ *and* $\alpha > 0$, *there exist positive constants* K_1, K_2, K *depending on* α, C_0 *and* κ *such that, with probability at least* $1 - e^{-t_{c\ell}}$,

$$l(f^*, \widehat{f}_{c\ell}) \leqslant (1+\alpha)l(f^*, \overline{f}_{c\ell}) + K_1 \left(\frac{|T_{c\ell}| \log(2n)}{n} \right)^{\frac{\kappa}{2\kappa-1}} + K_2 \left(\frac{t_{c\ell}}{n} \right)^{\frac{\kappa}{2\kappa-1}} + K\frac{t_{c\ell}}{n} \quad .(3.1)$$

Moreover, we obtain the following upper bound

$$E\left[l(f^*, \widehat{f}_{c\ell}) \right] \leqslant (1+\alpha)l(f^*, \overline{f}_{c\ell}) + K_1 \left(\frac{|T_{c\ell}| \log(2n)}{n} \right)^{\frac{\kappa}{2\kappa-1}} + Cn^{-\frac{\kappa}{2\kappa-1}} \quad . (3.2)$$

The obtained bound is in keeping with classical results already given in [23]. In particular, if the Bayes classifier belongs to class $\mathcal{C}_{c\ell}$, the rate of convergence for the risk associated with estimator $\widehat{f}_{c\ell}$ is of order $(\log(2n)/n)^{\frac{\kappa}{2\kappa-1}}$.

In practice, since no information is available about how to choose class $\mathcal{C}_{c\ell}$, one needs to consider the collection $\mathcal{M}$ of all possible configurations and variable lists. In each class $\mathcal{C}_{c\ell}$, a candidate $\widehat{f}_{c\ell}$ is chosen by empirical risk minimization, then the final classifier $\tilde{f}$ is selected among all class candidates by minimization of a penalized criterion:

$$\widehat{c\ell} = \arg\min_{c,\ell} \left(\mathrm{P}_n \widehat{f}_{c\ell} + pen(c,\ell) \right) \quad ,$$

$$\tilde{f} = \widehat{f}_{\widehat{c\ell}} \quad .$$

The following result provides insight about how the penalty should be chosen to ensure good performance for $\tilde{f}$.

Proposition 2. *Assume that margin assumption* **MA(1)** *is verified. If*

$$\tilde{f} = \underset{\{\widehat{f}_{c\ell} , (c,\ell)\in\mathcal{M}\}}{\operatorname{argmin}} \left(\mathrm{P}_n \widehat{f}_{c\ell} + pen(c,\ell) \right), \tag{3.3}$$

where

$$pen(c,\ell) = C'_\kappa \left(\frac{|T_{c\ell}| \log(2n)}{n} \right)^{\frac{\kappa}{2\kappa-1}} + C''_\kappa \left(\frac{|T_{c\ell}| \log(p)}{n} \right)^{\frac{\kappa}{2\kappa-1}} \tag{3.4}$$

with constants C'_κ *and* C''_κ *depending on* C_0 *and* κ *appearing in the margin condition, then there exist positive constants* C'_1, C'_2 *and* Σ *such that with probability at least* $1 - 3\Sigma e^{-x}$

$$l(f^*, \tilde{f}) \leqslant C'_1 \inf_{c,\ell} \left\{ \inf_{f\in\mathcal{C}_{c\ell}} l(f^*, f) + pen(c,\ell) \right\} + C'_2 \left(\left(\frac{x}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \frac{x}{n} \right) \quad .$$

Moreover, we obtain the following upper bound:

$$\mathbb{E}[l(f^*, \tilde{f})] \leqslant C'_1 \inf_{(c,\ell)\in\mathcal{M}} \left\{ \inf_{f\in\mathcal{C}_{c\ell}} l(f^*, f) + pen(c,\ell) \right\} + \frac{C''_2 \times \Sigma}{n^{\frac{\kappa}{2\kappa-1}}}. \tag{3.5}$$

The proofs of Propositions 1 and 2 are given in Section 5.

Several comments can be made about the result of Proposition 2:

Quality of the upper bound Compared with previous results [28, 12], the upper bound for the risk is improved in two different ways. First, since all possible binary trees are considered, in the present result the complete construction path of the tree classifier is taken into account: the infimum in equation (3.5) is taken on all possible classes of tree classifiers. Conversely, in previous results only the performance of the pruning step was assessed, i.e. the corresponding infimum was restricted to the list of classes associated with subtrees of the maximal tree. Second, thanks to the margin hypothesis, the convergence rate of the upper bound is faster than $\mathcal{O}(1/\sqrt{n})$ as soon as $\kappa < +\infty$.

Margin parameter The proposed penalty (3.4) depends on the margin parameter κ, that is usually unknown in practice. From a theoretical point of view, because this parameter quantifies the noise level of the classification problem, it necessarily appears in the ideal penalty function (as does the unknown variance in Gaussian model selection). From a practical point of view, it has to be estimated from the data. Obtaining this estimate in the general case is an open question.

Strong margin assumption In the particular case of margin assumption **MA(2)** given by equation (2.7), penalty (3.4) becomes (taking $\kappa = 1$):

$$
\begin{aligned}
pen(c, \ell) &= \frac{C_h^1 \log(2n) + C_h^2 \log(p)}{n} |T_{c\ell}| \\
&= \alpha_n |T_{c\ell}|.
\end{aligned}
$$

This corresponds exactly to the penalty proposed in [6] for the CART algorithm (see equation (1.2)). This penalty function has already been validated for the pruning step of the CART algorithm, (see [14] for the regression framework and [12] for the binary classification framework). A similar result is established by Proposition 2 when considering the exact optimization problem (1.1). Also note that in this context the margin parameter only appears in constant a_n. Because this constant will be tuned accordingly to the data (using cross-validation for instance), the problem of estimating the margin parameter is discarded.

Variable selection In comparison with the upper bound obtained in Proposition 1, one can observe in (3.5) the impact of parameter p that appears through the penalty. This quantity arises during the union bound step of the proof (see Section 5.3), where one has to count the number of classes sharing the same complexity. This conveys the fact that to build an optimal tree of size k, one has to choose k variables among p (with replacement). This is obviously a much easier task when $p = 100$ than when $p = 10,000$. This is where the variable selection task is taken into account. Moreover, the penalty term can be upper bounded by

$$
pen(c, \ell) \leqslant C_\kappa' \left(\frac{|T_{c\ell}| \log(2n)}{n} \right)^{\frac{\kappa}{2\kappa - 1}} + \log(p) \left(C_\kappa'' \left(\frac{|T_{c\ell}|}{n} \right)^{\frac{\kappa}{2\kappa - 1}} + C_\kappa''' \frac{|T_{c\ell}|}{n} \right),
$$

advocating for a penalty that should be linear with respect to $\log(p)$. This linear relationship is investigated in Section 3.2.

Oracle-type inequality Vapnik-Chervonenkis bounds for binary classification without any margin assumption give the following penalty form (see [9] for instance)

$$pen_V(c, \ell) = C_V^1 \sqrt{\frac{|T_{c\ell}| \log(n)}{n}} + C_V^2 \frac{|T_{c\ell}|}{n}.$$

This implies that, for classes associated with trees of large size, $pen(c, \ell)$ given in (3.4) becomes larger than $pen_V(c, \ell)$. Therefore, to obtain an oracle-type inequality, $pen(c, \ell)$ can be replaced by $\min \{pen_V(c, \ell), pen(c, \ell)\}$.

3.2 Illustration on simulated data

3.2.1 Practical determination of $\tilde{f}$

The application of the strategy described in Proposition 2 necessitates finding the empirical risk minimizer in each class $\mathcal{C}_{c\ell}$, and then comparing all the candidates $\widehat{f}_{c\ell}$ using the penalized criterion given by (3.3). From a computational point of view, the exhaustive comparison among all classes is an NP-hard problem. Therefore we need heuristic algorithms to obtain a sequence of near-optimal penalized risk minimizers $\left(\widehat{f}_k \right)_{k \geqslant 1}$ such that

$$\widehat{f}_k \approx \underset{\{\widehat{f}_{c\ell}, \ |T_{c\ell}|=k\}}{\operatorname{argmin}} P_n \widehat{f}_{c\ell} \ .$$

The CART algorithm, when applied with the empirical risk as an impurity measure at each node (see [16]), may be understood as a forward heuristic algorithm to build the sequence of optimal tree classifiers. In particular, the subtree classifier $\widehat{f}_k$ of size k extracted from the maximal tree can be interpreted as the (approximate) optimizer of the empirical risk over all the possible trees of size k.

This new understanding of the CART algorithm as a heuristic approach to obtain the sequence of subtree minimizers is important, because it points out that these subtree classifiers $\widehat{f}_k$ should be penalized as if the exhaustive search were performed, *i.e.* using penalty given by (3.4).

In most applications, when dealing with the construction of a tree classifier, experimenters use criterion (1.2) in a growing-pruning strategy, and the unknown parameter α_n is chosen by hold-out or Q-fold cross-validation. This estimated value can be compared with its theoretical counterpart given in (3.4). To this end, we perform a simulation study and compare the α_n obtained by cross-validation to its theoretical form

$$\frac{C_h^1 \log(2n) + C_h^2 \log(p)}{n}$$

obtained under the strong margin assumption **MA(2)**.

3.2.2 Simulations

We consider four simulation designs:

Design 1 Variables $X^1, ..., X^p$ are independently generated with distribution $\mathcal{N}(0,1)$. The label is generated as follows: If $X^1 > 0$ and $X^2 > 0$ then $Y = 1$ with probability q, otherwise $Y = 1$ with probability $1 - q$. Therefore only variables X^1 and X^2 are informative. In this design, the Bayes classifier can be represented as a tree with 3 leaves, hence it belongs to the considered collection of classes. Moreover, variables are independent, and margin assumption **MA(2)** is satisfied.

Design 2 First the labels are generated according to a Bernoulli distribution with parameter $1/2$. Then variable X^1 is generated such that $X^1|Y = 0$ and $X^1|Y = 1$ are normally distributed with means 0 and 1, respectively, and variance σ^2. Variables $X^2, ..., X^p$ are independent with distribution $\mathcal{N}(0,1)$ and are non-informative. As for design 1, the Bayes classifier can be represented as a tree and variables are independent, but it is easy to show that margin assumption **MA(2)** is not satisfied.

Design 3 Labels are simulated as in design 2. Then variables X^1 and X^2 are generated such that, for $j = 1, 2$, $X^j|Y = 0$ and $X^j|Y = 1$ are normally distributed with means 0 and 1, respectively, and variance σ^2. The last $p - 2$ variables are independent and non-informative. Here the Bayes classifier no longer belongs to the collection of tree classes, and margin assumption **MA(2)** is not satisfied.

Design 4 Three independent variables X^1, X^2, X^3 are generated with distribution $\mathcal{N}(0,1)$. Each additional variable X^j is then simulated as a noisy copy of $(X^1+X^2+X^3)/\sqrt{3}$. The label is generated as follows: If $(X^1)^2+(X^2)^2+(X^3)^2 > 2.5$ then $Y = 1$, else $Y = 0$. Here, all the variables are correlated (with a strong correlation between the extra variables), the Bayes classifier cannot be represented as a tree, and margin assumption **MA(2)** is not satisfied.

For designs 1 to 3, 400 samples are generated, and 1000 for design 4. On each of them, a tree classifier is selected using the growing/pruning strategy, where parameter α_n is selected by 10-fold cross-validation. Different values of parameters n ($n = 50, 100, 200$) and noise ($q = 0.1, 0.2, 0.3$ in design 1, $\sigma^2 = 0.5, 1, 2$ in designs 2 and 3, and $\sigma^2 = 0.2$ in design 4) are used. The number of variables considered to build the classifiers grows from $p = 30$ to $p = 10^3$.

Figure 3 displays the average value (on 400 simulations) of α_n versus the log-number of variables for the different designs. Parameter α_n decreases with respect to n, and the relationship between the selected α_n and $\log p$ is linear. These behaviors are observed whatever the level of noise (not shown) and whatever the design. This confirms that variable selection is taken into account by the pruning procedure of the CART algorithm through the choice of α_n. This also suggests that the penalty function proposed in (3.4) is relevant regarding its dependency on $\log p$.
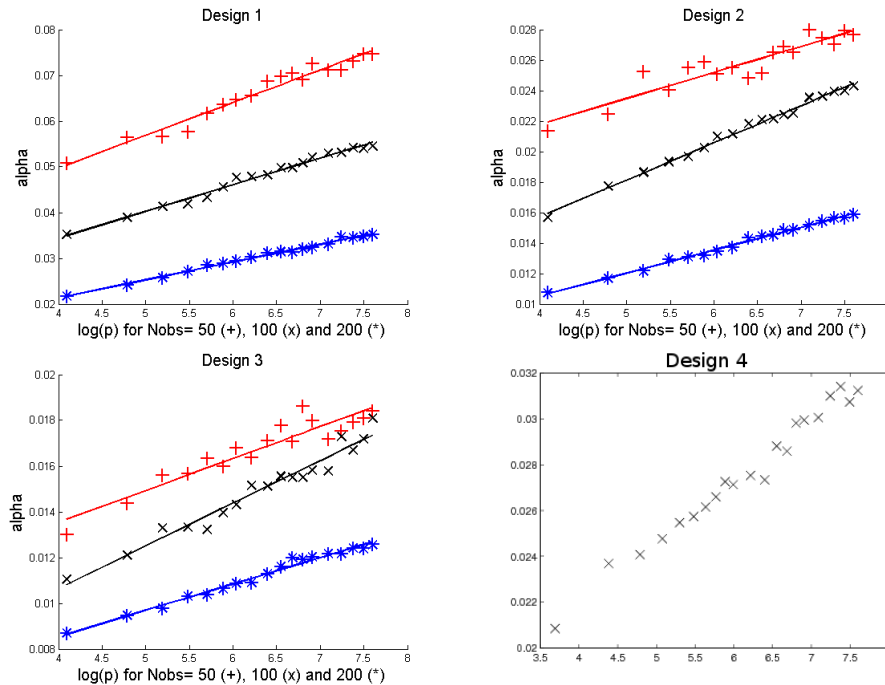

Figure 3: Average value of α_n with respect to $\log p$, for $n = 50$ (+), $n = 100$ (x) and $n = 200$ (*). Data are simulated from design 1 with $q = 0.3$ (Top Left), design 2 (Top Right), design 3 (Bottom Left) with $\sigma^2 = 2$. For design 4 (Bottom Right) the average α_n is obtained over 1000 samples, for $n = 100$.

4 Discussion

As stated in the Introduction, most previous results are related to the pruning step of the CART algorithm rather than considering the general optimization problem (1.1). For instance, in [12] and [28], risk bounds are obtained for the collection of CART pruned subtrees, which itself depends on the data at hand: the collection of models includes classes $\mathcal{C}_0, ..., \mathcal{C}_{K-1}, \mathcal{C}_K$ of tree classifiers

built on the maximal tree T_K, obtained from the training set, and its subtrees $T_0 \preccurlyeq ... \preccurlyeq T_{K-1}$. Thus the conditional risk bounds provided in previous articles only guarantee that the risk of the candidate is at most of the order of the risk of class $\mathcal{C}_{k^*}$ corresponding to the best subtree T_{k^*}. While this exactly describes the process of the CART algorithm, the guarantee may be poor if the best subtree of the collection is far from the best tree among all possible trees. Conversely, the approach presented here guarantees that the risk bound for the selected tree classifier is comparable to the risk of the class corresponding to the optimal tree (among all possible trees).

Proposition 2 generalizes the results obtained in [30] in two ways. First, Scott and Nowak considered the particular case where the tree classifiers are constructed on a fixed dyadic grid. In dyadic trees, the choice of the threshold at each internal node is deterministic, instead of being optimally tuned on the training set. This optimization is taken into account in the results presented here. Second, as recalled in Section 2, without any margin assumption, the penalty functions obtained in [30] are naturally proportional to the square root of the tree size over n. A $\sqrt{\log p}$ factor also appears in the resulting penalties. In comparison, the results presented here exhibit a range of penalty function from square root to linear depending on the margin assumption. If **MA(2)** is satisfied, this validates the form of the penalty implemented in the CART algorithm. If **MA(1)** is satisfied, it leads to better convergence rates for the risk bound.

Whenever margin assumption **MA(1)** is satisfied, the penalty suggested in Proposition 2 is sublinear. In this case the heuristic approach of the CART algorithm can still be employed to obtain an approximate version of $\hat{f}$. Indeed, as proved in [29], pruning with subadditive penalties produces sequences of pruned subtrees included in the sequence obtained through pruning with a linear penalty. This means that one can obtain an approximate optimizer of criterion (3.3), to the condition that the margin parameter is known.

The theoretical form of the penalty term (3.4) derived in Proposition 2 is of practical interest. First, it shows that sequential selection algorithms, such as stepwise or backward variable selection methods, can be easily studied in the model selection framework where the selection is supposed to be exhaustive. In the particular case of tree classification, the simulation study confirms that the penalty derived under the hypothesis of exhaustive variable selection is the one that is used in practice by the CART algorithm, that proceeds as a forward variable selection process. Second, it provides an interesting insight into the CART variable selection process. Indeed, the definition of the classes comes from the fact that a single variable may appear at different nodes, a specificity that changes the classical way of taking into account variable selection in the penalty term: in trees the variable list is ordered (the first variable of the list is associated with the first node) and a variable may be associated with several nodes. Therefore the classical $\binom{p}{k-1}$ term that appears in penalties in [2] or [24]

(i.e. the number of samplings without replacements and unordered sample) is replaced with p^{k-1} (i.e. the number of sampling with replacements and ordered sample).

In [18], Koltchinskii provides a synthesis of oracle inequalities in classification. In particular, the author considers margin assumptions more general than the margin assumption **MA(1)** given in [23]. The in-probability upper bounds for the loss $l(f^*, \tilde{f})$ given in Propositions 1 and 2 can be straightforwardly generalized using Koltchinskii's margin definition. This would lead to improved in-probability upper bounds for the loss $l(f^*, \tilde{f})$, similar to the one given in Theorem 6 of [18]. However, unlike hypothesis **MA(1)** considered here, it would not permit one to obtain explicit rates of convergence for the risk. Importantly, using a more general margin assumption would provide no improvement concerning the embedded selection aspect that we investigated here. From this aspect the results obtained are tight, as illustrated by the simulation study.

Acknowledgements

5 Proofs

5.1 Preliminary results

We provide two lemmas regarding the Vapnik entropy and the cardinality of tree class collections.
Note $H_{c\ell}$ the Vapnik-Chervonenkis log-entropy of class $\mathcal{C}_{c\ell}$:

$$H_{c\ell} = \log |\{A(f) \cap \{X_1, \ldots, X_n\}, \ f \in \mathcal{C}_{c\ell}\}|,$$

where $A(f) = \{x \in \mathcal{X} : f(x) = 1\}$.

Lemma 1. *For a tree class $\mathcal{C}_{c\ell}$, one has*

$$E(H_{c\ell}) \leqslant |T_{c\ell}| \log(2n)$$

This is obtained from lemma (2) in [14]. For a tree with $|T_{c\ell}|$ leaves, there are $|T_{c\ell}| - 1$ nodes for which the thresholds have to be estimated, leading to at most n ways to split the training sample. The possible number of splittings is bounded by $n^{|T_{c\ell}|-1}$. A given splitting shatters the sample into $|T_{c\ell}|$ subsamples, and each of these subsamples receive label 0 or 1. There are $2^{|T_{c\ell}|}$ ways to label the subsamples, hence

$$
\begin{aligned}
H_{c\ell} &< \log\left(n^{|T_{c\ell}|-1} \times 2^{|T_{c\ell}|}\right) \\
&< |T_{c\ell}| \log(2n) \ .
\end{aligned}
$$

Taking the expectation leads to the result.

Lemma 2. *The number of classes of trees of size k is*

$$p^{k-1} N_{(k)}, \quad with \quad N_{(k)} = \frac{1}{k} \binom{2k-2}{k-1} \quad .$$

First note that counting the number of classes amounts to counting the number of trees. A tree $T_{c\ell}$ is defined by a configuration c combined with a variable list ℓ. The total number of tree configurations of size k is given by the Catalan number $N_{(k)}$. The total number of lists of $k-1$ variables is p^{k-1}, because at each node we have to choose between the p available variables. Combined with the total number of tree configurations, this leads to the proposed lemma.

Remark In contrast with the classical variable selection framework, in trees the variable list is ordered (the first variable of the list is associated with the first node) and a variable may be associated with several nodes. Therefore the classical $\binom{p}{k-1}$ term that appears in penalties in [2] or [24] (i.e. the number of samplings without replacements and unordered sample) is replaced with p^{k-1} (i.e. the number of sampling with replacements and ordered sample).

5.2 Proof of Proposition 1

A classical way to bound $l(f^*, \widehat{f}_{c\ell})$ is to use the following decomposition:

$$l(f^*, \widehat{f}_{c\ell}) = l(f^*, \overline{f}_{c\ell}) + \mathrm{P}\widehat{f}_{c\ell} - \mathrm{P}\overline{f}_{c\ell},$$

and then to upper bound the variance term $\mathrm{P}\widehat{f}_{c\ell} - \mathrm{P}\overline{f}_{c\ell}$. In the case where class $\mathcal{C}_{c\ell}$ is finite, an upper bound can be obtained by using Bernstein inequality, as developped in [21] for instance. In our setting, because there may be (at least) one continuous coordinate (i.e. one continuous variable), classes $\mathcal{C}_{c\ell}$ are not finite. In this case, the upper bounding can be done using Theorem 2 from [18], which can be restated for our purpose as follows:

Theorem 5.2.1 (Koltchinskii, 2006). *If there exists a nondecreasing strictly concave function $\psi_{c\ell} : \mathbb{R}_+ \to \mathbb{R}_+$ such that with probability at least $1 - e^{-t_{c\ell}}$*

$$\sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} |(\mathrm{P}_n - \mathrm{P})(f - g)| \leqslant \psi_{c\ell}(\delta) \quad ,$$

and if $\psi_{c\ell}^\sharp$ is defined as

$$\psi_{c\ell}^\sharp(\varepsilon) = \inf\{\delta > 0 \ s.t. \ \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma} \leqslant \varepsilon\} \quad ,$$

then for all $\delta \geq \psi_{c\ell}^\sharp(1/q)$

$$\mathrm{P}\left[\mathrm{P}\widehat{f}_{c\ell} - \mathrm{P}\overline{f}_{c\ell} > \delta \right] \leqslant e^{-t_{c\ell}} \quad .$$

In order to use Theorem 5.2.1, we need to provide an explicit expression for $\psi_{c\ell}$. To proceed, we start from the following probabilistic upper bound given in [18] and derived from Talagrand's inequality for bounded processes (see [4] for more details):

$$\sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} |(\mathrm{P}_n - \mathrm{P})(f-g)| \leqslant 2 \left(\mathbb{E}\left[\sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} |(\mathrm{P}_n - \mathrm{P})(f-g)| \right] + D(\mathcal{C}_{c\ell}(\delta))\sqrt{\frac{t_{c\ell}}{n}} + \frac{t_{c\ell}}{n} \right) (5.1)$$

with probability larger than $1 - e^{-t_{c\ell}}$, where

$$\mathcal{C}_{c\ell}(\delta) = \{ f \in \mathcal{C}_{c\ell} \text{ s.t. } Pf - P\overline{f}_{c\ell} \leqslant \delta \}$$

and

$$
\begin{aligned}
D(\mathcal{C}_{c\ell}(\delta)) &= \sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} \sqrt{\mathbb{E}((f-g)^2)} \\
&= \sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} d(f,g)
\end{aligned}
$$

This last term can be upper-bounded in expression (5.1) using the margin assumption **MA(1)** described by (2.6):

$$d^2(f,f^*) \leqslant C_\kappa l(f^*,f)^{\frac{1}{\kappa}}$$

where $C_\kappa = (\kappa-1)^{\frac{1}{\kappa}} C^{\frac{\kappa-1}{\kappa}} \dfrac{\kappa}{\kappa-1}$. Hence

$$
\begin{aligned}
d(f,g) &\leqslant 2\sqrt{C_\kappa} \left(l(f^*,\overline{f}_{c\ell})^{\frac{1}{2\kappa}} + \delta^{\frac{1}{2\kappa}} \right) \\
\Rightarrow D(\mathcal{C}_{c\ell}(\delta)) &\leqslant 2\sqrt{C_\kappa} \left(l(f^*,\overline{f}_{c\ell})^{\frac{1}{2\kappa}} + \delta^{\frac{1}{2\kappa}} \right) = D \quad .
\end{aligned}
\tag{5.2}
$$

Now because

$$\mathbb{E}\left[\sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} |(\mathrm{P}_n - \mathrm{P})(f-g)| \right] \leqslant \mathbb{E}\left[\sup_{d(f,g) \leqslant D} |(\mathrm{P}_n - \mathrm{P})(f-g)| \right]$$

we can use the result of [25] (p295) to obtain

$$\mathbb{E}\left[\sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} |(\mathrm{P}_n - \mathrm{P})(f-g)| \right] \leqslant 24 D \sqrt{\frac{E[H_{c\ell}]}{n}} \quad , \tag{5.3}$$

where $H_{c\ell}$ is the Vapnik-Chervonenkis log-entropy of $\mathcal{C}_{c\ell}$. Combining (5.2) and (5.3), then using lemma 5 of [32], we obtain for all $\alpha \in]0,1[$

$$
\begin{aligned}
\sup_{f,g \in \mathcal{C}_{c\ell}(\delta)} |(\mathrm{P}_n - \mathrm{P})(f-g)| &\leqslant 2\left[2\sqrt{C_\kappa}\left(24\sqrt{\frac{E[H_{c\ell}]}{n}} + \sqrt{\frac{t_{c\ell}}{n}} \right)\left(l(f^*,\overline{f}_{c\ell})^{\frac{1}{2\kappa}} + \delta^{\frac{1}{2\kappa}} \right) + \frac{t_{c\ell}}{n} \right] \\
&\leqslant 4\sqrt{C_\kappa}\left(24\sqrt{\frac{E[H_{c\ell}]}{n}} + \sqrt{\frac{t_{c\ell}}{n}} \right) \delta^{\frac{1}{2\kappa}} \\
&\quad + 2\frac{t_{c\ell}}{n} + \alpha l(f^*,\overline{f}_{c\ell}) + \beta_{\kappa,\alpha}\left(\frac{E[H_{c\ell}]}{n} \right)^{\frac{\kappa}{2\kappa-1}} + \frac{\beta_{\kappa,\alpha}}{24}\left(\frac{t_{c\ell}}{n} \right)^{\frac{\kappa}{2\kappa-1}} \quad .
\end{aligned}
$$

In the present framework, we then have

$$
\begin{aligned}
\psi_{c\ell}(\delta) &= 4\sqrt{C_\kappa}\left(24\sqrt{\frac{E[H_{c\ell}]}{n}} + \sqrt{\frac{t_{c\ell}}{n}}\right)\delta^{\frac{1}{2\kappa}} + 2\frac{t_{c\ell}}{n} + \alpha l(f^*, \overline{f}_{c\ell}) + \beta_{\kappa,\alpha}\left(\frac{E[H_{c\ell}]}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \frac{\beta_{\kappa,\alpha}}{24}\left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}} \\
&= \psi_1(\delta) + \psi_2(\delta) + K
\end{aligned}
$$

where

$$
\begin{aligned}
\psi_1(\delta) &= 96\sqrt{C_\kappa}\sqrt{\frac{E[H_{c\ell}]}{n}}\delta^{\frac{1}{2\kappa}} \\
\psi_2(\delta) &= 4\sqrt{C_\kappa}\sqrt{\frac{t_{c\ell}}{n}}\delta^{\frac{1}{2\kappa}} \\
\text{and } K &= 2\frac{t_{c\ell}}{n} + \alpha l(f^*, \overline{f}_{c\ell}) + \beta_{\kappa,\alpha}\left(\frac{E[H_{c\ell}]}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \frac{\beta_{\kappa,\alpha}}{24}\left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}}
\end{aligned}
$$

Moreover, $\psi_{c\ell}^\sharp(\varepsilon) \leqslant \psi_1^\sharp(\varepsilon/3) + \psi_1^\sharp(\varepsilon/3) + \frac{3K}{\varepsilon}$, and $\psi_1^\sharp$ and $\psi_2^\sharp$ can be determined using the following characterization (available for all strictly concave functions ψ):

$$
\forall \varepsilon > 0, \ \psi\left(\psi^\sharp(\varepsilon)\right) = \psi^\sharp(\varepsilon)\varepsilon \ .
$$

Solving this last equation for the particular form of functions ψ_1 and ψ_2, we obtain

$$
\begin{aligned}
\psi_{c\ell}^\sharp(\varepsilon) &\leqslant \left(\frac{288\sqrt{C_\kappa}\sqrt{E[H_{c\ell}]}}{\varepsilon\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa-1}} + \left(\frac{12\sqrt{C_\kappa}\sqrt{t_{c\ell}}}{\varepsilon\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa-1}} \\
&\quad + \left(2\frac{t_{c\ell}}{n} + \alpha l(f^*, \overline{f}_{c\ell}) + \beta_{\kappa,\alpha}\left(\frac{E[H_{c\ell}]}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \frac{\beta_{\kappa,\alpha}}{24}\left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right)\frac{3}{\varepsilon}
\end{aligned}
$$

Taking $\varepsilon = 1/q$ one has with probability larger than $1 - e^{-t_{c\ell}}$

$$
\begin{aligned}
\mathrm{P}\widehat{f}_{c\ell} - \mathrm{P}\overline{f}_{c\ell} &\leqslant \left(\left(q288\sqrt{C_\kappa}\right)^{\frac{2\kappa}{2\kappa-1}} + 3q\beta_{\kappa,\alpha}\right)\left(\frac{E[H_{c\ell}]}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \left(\left(q12\sqrt{C_\kappa}\right)^{\frac{2\kappa}{2\kappa-1}} + \frac{3q\beta_{\kappa,\alpha}}{24}\right)\left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}} \\
&\quad + 6q\frac{t_{c\ell}}{n} + 3q\alpha l(f^*, \overline{f}_{c\ell}) \ .
\end{aligned}
$$

Using Lemma 1 and rescaling α properly, this leads to

$$
l(f^*, \widehat{f}_{c\ell}) \leqslant (1 + \alpha)l(f^*, \overline{f}_{c\ell}) + K^1_{\alpha,\kappa,q}\left(\frac{|T_{c\ell}|\log(2n)}{n}\right)^{\frac{\kappa}{2\kappa-1}} + K^2_{\alpha,\kappa,q}\left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}} + K_q\frac{t_{c\ell}}{n} \quad .(5.4)
$$

Renaming $K^1_{\alpha,\kappa,q} = K_1$, $K^2_{\alpha,\kappa,q} = K_2$ and $K_q = K$ leads to the first expression in Proposition 1. The risk bound follows by integration.

5.3 Proof of Proposition 2

We first choose the weights $t_{c\ell} = x_{c\ell} + x$ associated with classes $\mathcal{C}_{c\ell}$ such that $x_{c\ell}$ and x are positive and

$$\sum_{c,\ell} e^{-x_{c\ell}} = \Sigma < +\infty \quad .$$

The exact form of the weights will be chosen later. Furthermore, we will use lemma 4 of [18], reformulated here for our purpose:

Lemma 5.3.1 (Koltchinskii, 2006). *Consider a class $\mathcal{C}_{c\ell}$ and assume that* **MA(1)** *is satisfied. For all $t_{c\ell} > 0$ and $\alpha \in]0, 2/5[$, with probability at least $1 - 2e^{-t_{c\ell}}$, one has*

$$P_n \overline{f}_{c\ell} - P_n f^* \leqslant (1+\alpha)(P\overline{f}_{c\ell} - Pf^*) + K_\alpha \left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \frac{t_{c\ell}}{n} \tag{5.5}$$

and

$$P\overline{f}_{c\ell} - Pf^* \leqslant \left(1 - \frac{5}{2}\alpha\right)^{-1} \left(P_n \widehat{f}_{c\ell} - P_n f^* + \frac{3}{2}K_1 \left(\frac{|T_{c\ell}| \log(2n)}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 3K_2 \left(\frac{t_{c\ell}}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 3K\frac{t_{c\ell}}{n}\right) \tag{5.6}$$

with the same notations as above.

We start the proof from the result obtained in Proposition 1. Combining equation (3.1) of Proposition 1 and a classical union bound argument, one has with probability larger than $1 - \Sigma e^{-x}$

$$l(f^*, \tilde{f}) \leqslant (1+\alpha)l(f^*, \overline{f}_{\widehat{c\ell}}) + K_1 \left(\frac{|T_{\widehat{c\ell}}| \log(2n)}{n}\right)^{\frac{\kappa}{2\kappa-1}} + K_2 \left(\frac{x_{\widehat{c\ell}} + x}{n}\right)^{\frac{\kappa}{2\kappa-1}} + K\frac{x_{\widehat{c\ell}} + x}{n} \quad ,$$

where $\alpha \in]0, 2/5[$. We now use equation (5.6) from Lemma 5.3.1 to obtain with probability larger than $1 - 3\Sigma e^{-x}$

$$
\begin{aligned}
l(f^*, \tilde{f}) &\leqslant \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(P_n \widehat{f}_{\widehat{c\ell}} - P_n f^* + \frac{5K_1}{2} \left(\frac{|T_{\widehat{c\ell}}| \log(2n)}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 4K_2 \left(\frac{x_{\widehat{c\ell}} + x}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 4K\frac{x_{\widehat{c\ell}} + x}{n}\right) \\
&\leqslant \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(P_n \widehat{f}_{\widehat{c\ell}} - P_n f^* + \frac{5K_1}{2} \left(\frac{|T_{\widehat{c\ell}}| \log(2n)}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 4K_2 \left(\frac{x_{\widehat{c\ell}}}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 4K\frac{x_{\widehat{c\ell}}}{n}\right) \\
&\quad + \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(4K_2 \left(\frac{x}{n}\right)^{\frac{\kappa}{2\kappa-1}} + 4K\frac{x}{n}\right)
\end{aligned}
$$

In the context of variable selection, one has to choose the weights such that

$$\sum_{c,\ell} e^{-x_{c\ell}} < +\infty \Rightarrow \sum_k \sum_{\mathcal{C}_{c\ell} \text{ s.t. } |T_{c\ell}|=k} e^{-x_{c\ell}} < +\infty \quad .$$

Giving equal weights x_k to classes of same complexity k (i.e. classes $\mathcal{C}_{c\ell}$ and $\mathcal{C}_{c'\ell'}$ such that $|T_{c\ell}| = |T_{c'\ell'}| = k$), one obtains from Lemma 2:

$$\sum_k \sum_{\mathcal{C}_{c\ell} \text{ s.t. } |T_{c\ell}|=k} e^{-x_{c\ell}} = \sum_k p^{k-1} \frac{1}{k} \binom{2k-2}{k-1} e^{-x_k}$$

$$\leqslant \sum_k \frac{(4p)^k}{k} e^{-x_k} \quad .$$

The choice $x_{c\ell} = x_{|T_{c\ell}|} = \lambda |T_{c\ell}| \log(p)$ with $\lambda > 3$ ensures that the sum is finite. Hence,

$$l(f^*, \tilde{f}) \leqslant \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(P_n \widehat{f}_{\widehat{c\ell}} - P_n f^* + \frac{5K_1}{2} \left(\frac{|T_{\widehat{c\ell}}| \log(2n)}{n} \right)^{\frac{\kappa}{2\kappa-1}} + 4K_2 \left(\frac{\lambda |T_{\widehat{c\ell}}| \log(p)}{n} \right)^{\frac{\kappa}{2\kappa-1}} \right.$$

$$\left. + 4K \frac{\lambda |T_{\widehat{c\ell}}| \log(p)}{n} \right) + \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(4K_2 \left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + 4K \frac{x}{n} \right)$$

$$\leqslant \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(P_n \widehat{f}_{\widehat{c\ell}} - P_n f^* + C'_\kappa \left(\frac{|T_{\widehat{c\ell}}| \log(2n)}{n} \right)^{\frac{\kappa}{2\kappa-1}} + C''_\kappa \left(\frac{|T_{\widehat{c\ell}}| \log(p)}{n} \right)^{\frac{\kappa}{2\kappa-1}} + C'''_\kappa \left(\frac{|T_{\widehat{c\ell}}| \log(p)}{n} \right) \right)$$

$$+ \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(4K_2 \left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + 4K \frac{x}{n} \right) \quad ,$$

for a proper choice of constants C'_κ, C''_κ, and C'''_κ. This leads to

$$l(f^*, \tilde{f}) \leqslant \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \inf_{c,\ell} \left(P_n \widehat{f}_{c\ell} - P_n f^* + pen(c,\ell) \right) + \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(4K_2 \left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + 4K \frac{x}{n} \right) \quad .$$

Since $P_n \widehat{f}_{c\ell} - P_n f^* \leqslant P_n \overline{f}_{c\ell} - P_n f^*$ (by definition of $\widehat{f}_{c\ell}$), this last expression can be upper bounded (with probability larger than $1 - 3\Sigma e^{-x}$) thanks to equation (5.5) of Lemma 5.3.1:

$$l(f^*, \tilde{f}) \leqslant \frac{(1+\alpha)^2}{1 - \frac{5\alpha}{2}} \inf_{c,\ell} \left(P \overline{f}_{c\ell} - P f^* + K_\alpha \left(\frac{x_{c\ell}}{n} \right)^{\frac{\kappa}{2\kappa-1}} + \left(\frac{x_{c\ell}}{n} \right) + K_\alpha \left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + \left(\frac{x}{n} \right) + pen(c,\ell) \right)$$

$$+ \frac{(1+\alpha)}{1 - \frac{5\alpha}{2}} \left(4K_2 \left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + 4K \frac{x}{n} \right)$$

$$\leqslant \frac{2(1+\alpha)^2}{1 - \frac{5\alpha}{2}} \inf_{c,\ell} \left(P \overline{f}_{c\ell} - P f^* + pen(c,\ell) \right) + \frac{2(1+\alpha)^2}{1 - \frac{5\alpha}{2}} \left(4K_2 \left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + 4K \frac{x}{n} \right)$$

$$\leqslant C'_1 \inf_{c,\ell} \left(P \overline{f}_{c\ell} - P f^* + pen(c,\ell) \right) + C'_2 \left(\left(\frac{x}{n} \right)^{\frac{\kappa}{2\kappa-1}} + \frac{x}{n} \right) \quad .$$

The last inequality corresponds to the first equation of Proposition 2. The risk bound follows by integration.

References

[1] L. Bel, D. Allard, J.M. Laurent, R. Cheddadi, and A. Bar-Hen. CART algorithm for spatial data: application to environmental and ecological data. *Computational Statistics and Data Analysis*, 53(8):3082–3093, 2009.

[2] L. Birgé and P. Massart. A generalized Cp criterion for Gaussian model selection. Technical Report 647, Universités de Paris 6 et 7, 2001.

[3] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic), 2005.

[4] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[5] L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[6] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Chapman & Hall, 1984.

[7] Philip A. Chou, Tom Lookabaugh, and Robert M. Gray. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory*, 35(2):299–315, 1989.

[8] T. Czekaj, W. Wu, and B. Walczak. Classification of genomic data: Some aspects of feature selection. *Talanta*, 76:564–574, 2008.

[9] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[10] S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, 97:77–87, 2002.

[11] B.D. Dyer, M.J. Kahn, and M.D. Leblanc. Classification and regression tree (CART) analyses of genomic signatures reveal sets of tetramers that discriminate temperature optima of archaea and bacteria. *Archaea*, 2:159–167, 2007.

[12] Servane Gey. Risk bounds for cart classifiers under a margin condition. *Pattern Recognition*, 45:3523–3534, 2012.

[13] Servane Gey and Emilie Lebarbier. Using CART to detect multiple change-points in the mean for large samples. Technical Report 12, SSB, 2008.

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

[15] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

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

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

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

[20] T.L. Lal, O. Chapelle, J. Weston, and A. Elisseeff. *Feature Extraction: Foundations and Applications, Studies in Fuzziness and Soft Computing*, chapter Embedded methods., pages 137–165. Springer, 2006.

[21] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

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

[23] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[24] T. Mary-Huard, S. Robin, and J.J. Daudin. A penalized criterion for variable selection in classification. *Journal of Multivariate Analysis*, 98(4):695–705, 2007.

[25] Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, 2000.

[26] Pascal Massart. *Concentration inequalities and model selection*, volume 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.

[27] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

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

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

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

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

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

[33] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley Inter-Sciences, 1998.

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