Oracle bounds and exact algorithm for dyadic classification trees

Gilles Blanchard[1]*, Christin Schäfer[1]**, and Yves Rozenholc[2]

[1] Fraunhofer–Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany,
{blanchar,christin}@first.fhg.de,
[2] Laboratoire de Probabilités et Modèles aléatoires,
Université Pierre et Marie Curie, BC 188, 75252 Paris Cedex 05, France,
rozen@math.jussieu.fr

Abstract. This paper introduces a new method using dyadic decision trees for estimating a classification or a regression function in a multiclass classification problem. The estimator is based on model selection by penalized empirical loss minimization. Our work consists in two complementary parts: first, a theoretical analysis of the method leads to deriving oracle-type inequalities for three different possible loss functions. Secondly, we present an algorithm able to compute the estimator in an exact way.

1 General setup

1.1 Introduction

In this paper we introduce a new method using dyadic decision trees for estimating a classification or a regression function in a multiclass classification problem. The two main focuses of our work are a theoretical study of the statistical properties of the estimator, and an exact algorithm used to compute it.

The theoretical part (section 2) is centered around the convergence properties of piecewise constant estimators on abstract partition models (generalized histograms) for estimating either a classification function or the conditional probability distribution (cpd) $P(Y|X)$ for a classification problem. A suitable partition is selected by a penalized minimum empirical loss method and we derive oracle inequalities for different possible loss functions: for classification, we use the 0-1 loss; for cpd estimation, we consider the minus-log loss, and the square error loss. These general results are then applied to dyadic decision trees. In section 3, we present an algorithm able to compute in an exact way the solution of the minimization problem that defines the estimator in this case.

* Supported by a grant of the Humboldt Foundation.

** This research was partly supported through grants of the *Bundesministerium für Bildung und Forschung* FKZ 01–BB02A and FKZ 01–SC40A.

1.2 Related work and novelty of our approach

The oracle-style bounds presented here for generalized histograms for multiclass problems are novel up to our knowledge. Our analysis relies heavily on [1] which contains the fundamental tools used to prove Theorems 1-3. For classification, Theorem 1 presents a bound for a penalty which is *not* inverse square-root in the sample size (as is the case for example in classical VC theory for *consistent* bounds, i.e. bounds that show convergence to the Bayes classifier of a SRM procedure when sample size grows to infinity) but inverse linear, thus of strictly lower order. This holds under an identifiability assumption of the maximum class, akin to Tsybakov's condition (see [2] and [3]). For cpd estimation, result of Theorem 3 seems entirely novel in that it states an oracle inequality with the Kullback-Leibler (K-L) divergence on both sides. In contrast, related results in [4, 5] for density estimation had the Hellinger distance on the left-hand side. Dyadic trees for density estimation have also been recently studied in [6] with a result for convergence in L^2.

Traditional CART-type algorithms [7] adopt a similar penalized loss approach, but do not solve exactly the minimization problem. Instead, they grow a large tree in a greedy way, and prune it afterwards. Some statistical properties of this pruning procedure have been studied in [8]. More recently, an exact algorithm for dyadic trees and related theoretical analysis for classification loss has been proposed in [9, 10]. It differs fundamentally from the algorithm presented here in that the directions of the splits are fixed in advance in the latter work, so that the procedure essentially reduces to a pruning. It is also different in that the authors do not make any identifiability assumption and therefore use a square-root type penalty (see discussion in section 2.3).

On the algorithmic side, the novelty of our work resides on the fact that we are able to treat the case of arbitrary direction choice for the splits in the tree. This allows for a much increased adaptivity of the estimators to the problem as compared to a fixed-directions architecture, particularly if the target function is very anisotropic, e.g. if there are irrelevant input features.

1.3 Goals

We consider a multiclass classification problem modeled by a couple of variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = [0, 1]^d$ and a finite class set $\mathcal{Y} = \{1, \ldots, t\}$. We assume that we observe a training sample $(X_i, Y_i)_{i=1,\ldots,n}$ of size n, drawn i.i.d. from some unknown probability $P(X, Y)$. We are interested in estimating either a classification function or the cpd $P(Y|X)$. Estimation of the cpd can be of practical interest of its own or can be used to form a derived classifier by "plug-in". It is generally argued that such plug-in estimates can be suboptimal and that one should directly try to estimate the classifier if it is the final aim (see [11]). However, even if classification is the goal, there is also some important added value in estimating $P(Y|X)$:

- it gives more information to the user than the classification function, allowing for a finer appreciation of ambiguous cases;

– it allows to deal with cases where the classification loss is not the same for all classes. In particular, it is more adapted when performance is measured by a ROC curve.

To qualitatively measure the fit of a function f to a data point (X, Y), a loss function $\ell(f, X, Y) \in \mathbb{R}$ is used. The goal is to be as close as possible to the function f^* minimizing the average loss:

$$f^* = \operatorname*{Arg\,Min}_{f \in \mathcal{F}} E\left[\ell(f, X, Y)\right],$$

where the minimum is taken over some suitable subset $\mathcal{F}$ of all measurable functions. We consider several possible loss functions, this will be detailed in section 1.6.

If a function $\widehat{f}$ is selected by some method using the training sample, it is coherent to measure its closeness to f^* by the means of its excess (average) loss (also called *risk*):

$$L(\ell, \widehat{f}, f^*) = E\left[\ell(\widehat{f}, X, Y)\right] - E\left[\ell(f^*, X, Y)\right];$$

our theoretical study is focused on this quantity.

1.4 Bin estimation and model selection

We focus on bin estimation, i.e. the estimation of the target function using a piecewise constant function with a finite number of pieces, which can be seen as a generalized histogram. Such a piecewise constant function f is therefore characterized by a finite measurable partition $\mathcal{B}$ of the input space $\mathcal{X}$ – each piece of the partition will hereafter be called a bin – and by the values $f_{b,y}$ taken on the bins for $b \in \mathcal{B}, y \in \mathcal{Y}$:

$$f(x, y) = \sum_{b \in \mathcal{B}} \mathbb{I}_{\{x \in b\}} f_{b,y}. \tag{1}$$

Once a partition is fixed, it is natural to estimate the parameters $f_{b,y}$ using the training sample points which are present in the bin: we therefore define the following counters for all $b \in \mathcal{B}, y \in \mathcal{Y}$:

$$N_{b,y} = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in b; Y_i = y\}} \quad \text{and} \quad N_b = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in b\}} = \sum_{y \in \mathcal{Y}} N_{b,y}.$$

Of course, the crucial problem here is the choice of a suitable partition, which is a problem of *model selection.* Hereafter, we identify a model with a partition: an abstract model will be denoted by m, and the associated partition by $\mathcal{B}_m$; $|m|$ denotes the number of pieces in $\mathcal{B}_m$. The set of piecewise constant real functions on bins of $\mathcal{B}_m$ (i.e. of the form (1)) will be denoted $\mathcal{G}_m$. Similarly, the set of classification functions which are piecewise constant on $\mathcal{B}_m$ will be denoted $\mathcal{C}_m$. Finally, the set of piecewise constant densities on $\mathcal{B}_m$ will be denoted $\mathcal{F}_m$:

$$\mathcal{F}_m = \left\{ f \in \mathcal{G}_m \,\middle|\, \forall x \in \mathcal{X}, \ \sum_y f(x, y) = 1 \right\}.$$

1.5 Dyadic decisions trees

Our goal is to consider specific partition models generated by dyadic decision trees. A dyadic decision tree is a binary tree structure T such that each internal node of T is "colored" with an element of $\{1, \ldots, d\}$ (recall d is the dimension of $\mathcal{X} = [0,1]^d$). To each node (internal or terminal) of T is then associated a certain bin obtained by recursively splitting $[0,1]^d$ in half along the axes, according the colors at the internal nodes of T. This is defined formally in the following way:

1. To the root of T is associated $[0,1]^d$.

2. Suppose s is an interal node of T, and that a bin of the form $b(s) = \prod_{j=1}^d I_j$ is associated to s, where the (I_j) are dyadic intervals on the different axes of $\mathcal{X}$. Let k_s be the color of s, then the bins associated to the right and left children nodes r_s, ℓ_s of s are obtained by cutting $b(s)$ at its midpoint perpendicular to axis k_s; in other words, $b(r_s)$ is obtained by replacing in the product defining $b(s)$ interval I_{k_s} by its right half-interval, and correspondingly for $b(\ell_s)$.

Finally, the partition model generated by T is the set of bins attached to the terminal nodes (leaves) of T.

1.6 Loss functions

We investigate three possible loss functions. For classification problems, we consider the set of classifier functions $\mathcal{F}_{class.} = \{f : \mathcal{X} \to \mathcal{Y}\}$ and the 0-1 loss:

$$\ell_{class.}(f, X, Y) = \mathbb{I}_{\{f(X) \neq Y\}}. \tag{2}$$

The corresponding minimizer $f^*_{class.}$ of the average loss among all functions from $\mathcal{X}$ to Y is given by the *Bayes classifier* $f^*_{class.}(x) = \operatorname*{Arg\,Max}_{y \in \mathcal{Y}} P(Y = y|X = x)$ (see e.g. [11]).

For cpd estimation, we consider the set $\mathcal{F}_{cpd}$ of functions which are conditional probabilities of Y given X, i.e. functions $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ which are measurable and satisfy $\sum_{y \in \mathcal{Y}} f(x, y) = 1$ for all $x \in \mathcal{X}$. In this case we use one of two possible loss functions: the minus-log loss

$$\ell_{log}(f, X, Y) = -\log(f(X, Y)), \tag{3}$$

(which can possibly take the value $+\infty$) and the square loss

$$\ell_{sq}(f, X, Y) = (1 - f(X, Y))^2 + \sum_{j \neq Y} f(X, j)^2 = \left\| f(X, \cdot) - \overline{Y} \right\|_t^2, \tag{4}$$

where $\|\cdot\|_t$ is the standard Euclidian norm in $\mathbb{R}^t$ and $\overline{Y}$ is the Y-th canonical base vector of $\mathbb{R}^t$. It is easy to check that the function f^*_{cpd} minimizing the average losses $E\ell_{log}(f, X, Y)$ and $E\ell_{sq}(f, X, Y)$ over $\mathcal{F}_{cpd}$ is indeed $f^*_{cpd}(x, y) = P(Y = y|X = x)$. The corresponding excess losses from f to f^*_{cpd} are then given, respectively, by the average K-L divergence given X:

$$L(\ell_{log}, f, f^*_{cpd}) = E_P \left[\log \left(\frac{P(Y|X)}{f(X, Y)} \right) \right] \doteq KL(P, f|X), \tag{5}$$

and the averaged squared euclidian distance in $\mathbb{R}^t$:

$$L(\ell_{sq}, f, f^*_{cpd}) = E_{P(X)}\left[\|f(X,\cdot) - P(Y = \cdot|X)\|_t^2\right] \doteq \left\|f - f^*_{cpd}\right\|^2_{t,2}. \qquad (6)$$

Finally, we will make use of the following additional notation: $\ell(f)$ is a shortcut for $\ell(f,\cdot,\cdot)$ as a function of X and Y; we denote the expectation of a function f with respect to P either by $E_P[f]$ or Pf; P_n denotes the empirical distribution associated to the sample.

2 Theoretical results for the bin estimators

2.1 Fixed model m

First let us assume that some fixed model m is chosen. We now define an estimator associated to this model and depending on the loss function used. The classical *empirical risk minimization* method consists in considering the empirical (or training) loss

$$P_n\ell(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, X_i, Y_i), \qquad (7)$$

and selecting the function attaining the minimum of this empirical loss over the set of functions $\mathcal{F}_m$ in the model. When using the classification loss, this gives rise to the classifier minimizing the training error:

$$\widetilde{f}_m(x) = \operatorname*{Arg\,Max}_{y\in Y} \sum_{b\in\mathcal{B}_m} N_{b,y}\mathbb{I}_{\{x\in b\}}; \qquad (8)$$

when using the square loss or the minus-log loss (3), this gives rise to

$$\widehat{f}_m(x,y) = \sum_{b\in\mathcal{B}_m} \frac{N_{b,y}}{N_b}\mathbb{I}_{\{x\in b\}}. \qquad (9)$$

In case of an undefinite ratio $0/0$ in the formula above, one can choose arbitrary values for this bin, say $1/t$ for all classes.

In the case of the minus-log loss, notice that the loss has infinite average whenever there is a bin b such that $N_{b,y} = 0$ but $P(Y = y|X \in b) \neq 0$. This motivates to consider the following slightly modified estimator which bypasses this problem:

$$\widehat{f}^{\rho}_m = (1 - t\rho)\widehat{f}_m + \rho, \qquad (10)$$

where ρ is some small positive constant. Typically, we can choose ρ of order $\mathcal{O}(n^{-k})$ (see discussion after Theorem 3) for some arbitrary but fixed k (to fix ideas, say $k = 3$), so that the two functions will be very close in all cases.

2.2 Model selection via penalization

Now we address the problem of choosing a model m. A common approach is to use a penalized empirical loss criterion, namely selecting the model $\widehat{m}$ such that

$$\widehat{m} = \operatorname*{Arg\,Min}_{m \in \mathcal{M}} \left\{ P_n \ell(\widehat{f}^{\rho}_m) + \mathrm{pen}(m) \right\}, \tag{11}$$

where pen is a suitable penalization function. For the standard CART algorithm, the penalization is of order $\alpha|m|$. The goal of the theoretical study to come is to justify that penalties of this order with estimators defined by (11) lead to oracle-type bounds for the respective excess losses. Note that we must assume that the *exact* minimization of (11) is found, or at least with a known error margin, which typically is not the case for the greedy CART algorithm. We will show in section 3 how the minimization can be solved effectively for dyadic trees.

2.3 Oracle inequalities for the penalized estimators

Classification loss. In the case of classification loss, it has been known for some time [2, 3] that the best convergence rates in classification strongly depend on the behavior of $P(Y|X)$ and in particular of the identifiability of the majority class. Without any assumption to this regard, the minimax rate of convergence for classification error is of order $\mathcal{O}(\sqrt{D/n})$ for a model of VC-dimension D (see e.g. [11]), and thus the penalty should be at least of this order. Such an analysis has been used in [9] for dyadic classification trees. Presently, we will assume instead that we are in a favorable case in which the majority class is always identifiable[3] with a fixed known "margin" η_0, which allows to use a smaller order penalty ($\mathcal{O}(|m|/n)$). Moreover, this additive (wrt. the size of the model) penalty makes the minimization problem (11) easier to solve practically. Note that the identifiability assumption is only necessary for classifier estimation in Theorem 1, not for cpd estimation in Theorems 2-3.

Theorem 1. *Assume the following identifiability condition: there exists some* $\eta_0 > 0$ *such that*

$$\forall x \in \mathcal{X}, \qquad P(Y = f^*_{class}(x)|X = x) - \max_{y \neq f^*(x)} P(Y = y|X = x) \geq \eta_0. \tag{12}$$

Let $(x_m)_{m \in \mathcal{M}}$ *be real numbers with* $\sum_{m \in \mathcal{M}} \exp(-x_m) \leq 1$. *Then for any* $K > 1$, *there exist absolute constants* C_1, C_2, C_3 *such that, if*

$$\forall m \in \mathcal{M} \qquad \mathrm{pen}(m) \geq C_1 \frac{|m| \log t}{\eta_0 n} + C_2 \frac{x_m}{\eta_0 n} \tag{13}$$

[3] Note that this identifiability assumption (12) below is much weaker than the assumption that the Bayes error is zero, which appears in classical VC theory to justify non-square-root penalties for *consistent* bounds and SRM procedures.

then the penalized estimator $\widetilde{f}_{\widehat{m}}$ *satisfies*

$$E\left[\mathbf{err}(\widetilde{f}_{\widehat{m}}) - \mathbf{err}(f^*_{class})\right] \leq K \inf_{\substack{m\in\mathcal{M}\\ f\in\mathcal{C}_m}} \left(\mathbf{err}(f) - \mathbf{err}(f^*_{class}) + 2\mathrm{pen}(m) + \frac{C_3}{n}\right),$$

where **err** *denotes the generalization error and the expectation on the left-hand side is over training sets* $(X_i, Y_i)_{i=1\ldots n}$.

Square loss.

Theorem 2. *Let* $(x_m)_{m\in\mathcal{M}}$ *be real numbers with* $\sum_{m\in\mathcal{M}} \exp(-x_m) \leq 1$. *Then for any* $K > 1$, *there exist absolute constants* C_1, C_2, C_3 *such that, if*

$$\forall m \in \mathcal{M} \qquad \mathrm{pen}(m) \geq C_1 \frac{t|m|}{n} + C_2 \frac{x_m}{n} \tag{14}$$

then the penalized estimator $\widehat{f}_{\widehat{m}}$ *satisfies*

$$E\left[\left\|\widehat{f}_{\widehat{m}} - f^*_{cpd}\right\|^2_{t,2}\right] \leq K \inf_{m\in\mathcal{M}} \left(\inf_{f\in\mathcal{F}_m} \left\|f - f^*_{cpd}\right\|^2_{t,2} + 2\mathrm{pen}(m) + \frac{C_3}{n}\right).$$

Minus-log loss.

Theorem 3. *Let* $(x_m)_{m\in\mathcal{M}}$ *be real numbers with* $\sum_{m\in\mathcal{M}} \exp(-x_m) \leq 1$. *Then for any* $K > 1$, *there exist absolute constants* C_1, C_2, C_3 *such that, if*

$$\forall m \in \mathcal{M} \qquad \mathrm{pen}(m) \geq C_1 \frac{t|m| \log \rho}{n} + C_2 \frac{x_m \log \rho}{n} \tag{15}$$

then the penalized estimator $\widehat{f}^{\rho}_{\widehat{m}}$ *satisfies*

$$E\left[KL(P, \widehat{f}^{\rho}_{\widehat{m}}|X)\right] \leq K \inf_{\substack{m\in\mathcal{M}\\ f\in\mathcal{F}_m}} \left(KL(P, f|X) + 2\mathrm{pen}(m) + \frac{C_3}{n} - 3\log(1 - t\rho)\right).$$

Note that the typical values of ρ should be of order n^{-k} for some arbitrary $k > 0$. Assuming the number of models per dimension is at most exponential, the penalty function is then of order $t|m| \log n/n$, and the trailing term $\log(1-t\rho)$ is of order t/n^k.

Application to dyadic decision trees.

Corollary 1. *For dyadic decision trees in dimension* d, *Theorems 1-3 apply with the choice*

$$x_m = C|m| \log(d), \tag{16}$$

where C *is a universal constant.*

Proof. The point here is only to count the number of models of size $|m| = D$. An upper bound can be obtained the following way: the number of binary trees with $D + 1$ leaves is given by the Catalan number $Cat(D) = (D + 1)^{-1}\binom{2D}{D}$; such a tree has D internal nodes and we can therefore label these nodes in d^D different ways. It can be shown that $Cat(n) \leq C' 4^n / n^{3/2}$ for some constant C'; hence for C big enough in (16), $\sum_m \exp(-x_m) \leq 1$ is satisfied. □

3 Implementation of the estimator

Principle and naive approach. We hereafter assume that the penalization function is on the form $\text{pen}(m) = \alpha|m|$ for some α (possibly depending on the sample size n).

In traditional CART, no exact minimization is performed. The split at each node is determined in a greedy way in order to yield the best local reduction of some empirical criterion (the entropy criterion corresponds to ℓ_{log} while the Gini criterion corresponds to ℓ_{sq}). In contrast, we introduce a method to find the global solution of (11) for dyadic decision trees by dynamic programming. This method is strongly inspired from an algorithm proposed by Donoho [12] for image compression.

We assume that there is a fixed bound k_{max} on the maximal numbers of cuts along a same dimension. Therefore, the smallest possible bins are those obtained with k_{max} cuts in every dimension, i.e. small hypercubes of edge length $2^{-k_{max}}$. We represent any achievable bin by a d-tuple $b = (L_1(b), \ldots, L_d(b))$, where for each i, $L_i(b)$ is a finite list of length $0 \leq |L_i| \leq k_{max}$, with elements in $\{r, \ell\}$. Each of these (possibly empty) lists contains the successions of cuts in the corresponding dimension needed to obtain the bin; each element of the list indicates if the left or the right child is selected after a cut, see section 1.5. Note that, while the order of the sequence of cuts along a same dimension is important, the order in which the cuts along different dimensions are performed is not relevant for the definition of the bin. Finally, we will denote $|b| = \sum_i |L_i(b)|$ and call it the depth of cell b, and $\mathcal{B}_{k_{max}}$ the set of achievable bins, i.e. such that $|L_i(b)| \leq k_{max}$ for all $1 \leq i \leq d$.

The principle of the method is simple, and is based on the additive property of the function to be optimized. If b is a bin, denote T_b a "local" dyadic tree rooted in b, i.e. a dyadic tree starting at bin b and splitting it recursively, while still satisfying the assumption that the bins attached to its leaves belong to $\mathcal{B}_{k_{max}}$. Furthermore we assume that to each terminal bin a value is associated estimated from the data, such as (10), so that T_b can be considered as a piecewise constant function on b. Denote $|T_b|$ the number of leaves of T_b and define

$$\mathcal{E}(T_b) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in b\}} \ell(T_b, X_i, Y_i) + n\alpha|T_b|.$$

Note that when $b = [0,1]^d$, finding the minimum of $\mathcal{E}(T)$ is equivalent to the minimization problem (11). Moreover, whenever T_b is not reduced to its root (hereafter we will call such a tree *nondegenerate*), if we denote u and v the bins attached to the left and right children of the root and T_u, T_v the corresponding subtrees, then we have

$$\mathcal{E}(T_b) = \mathcal{E}(T_u) + \mathcal{E}(T_v).$$

For a bin b, let T_b^* denote the local dyadic tree minimizing $\mathcal{E}(T_b)$. Finally, let us denote by b_ℓ^i, b_r^i the left and right sub-bins obtained by splitting b in half along direction i. Then from the above observations it is straightforward that

$$\mathcal{E}(T_b^*) = \min\left(\mathcal{E}(R_b), \min\left\{ \mathcal{E}(T^*_{b_\ell^i}) + \mathcal{E}(T^*_{b_r^i}) \middle| i : |L_i(B)| < k_{max} \right\}\right), \tag{17}$$

where R_b denotes the degenerate local tree $\{b\}$.

From this it is quite simple to develop the following naive bottom-up approach to solving the optimization (11): suppose we know the optimal local tree T^*_b for every bin of depth $|b| = k$, then using (17) we can compute the optimal local trees for all bins at depth $k-1$. Starting with the deepest bins (the hypercubes of side length 2^{-kmax}) for which the optimal local trees are degenerate, it is possible to compute recursively optimal trees for lower depth bins, finally finding the optimal tree T^* for $[0,1]^d$.

Dictionary-based approach. The naive approach proposed above however has a significant drawback, namely its complexity; there are already $2^{dk_{max}}$ smallest bins at depth dk_{max}, and even more bins for intermediate depth values, due to the combinatorics in the choice of cuts. We therefore put forward an improved approach, based on the following observation: if $2^{dk_{max}} > n$, then some (possibly a lot) of the smallest bins are actually empty, and so are bins at intermediate depths as well. Furthermore, for an empty bin b at any depth the optimal local tree is obviously the degenerate tree $T^*_b = R_b$. Therefore, it is sufficient to keep track of the *non-empty* bins along the process. This can be done using a dictionary $\mathcal{D}_k$ of non-empty bins of depth k; the algorithm is then as follows:

```
**Initialization**: construct dictionary $\mathcal{D}_{dk_{max}}$ by finding the minimal bins (hypercubes of edge length $2^{-k_{max}}$) containing at least one datapoint, and inserting them in $\mathcal{D}_{dk_{max}}$. For each of these bins $b$, also store that $T^*_b = R_b$.
**Loop on depth,** $D = dk_{max}, \ldots, 1$**:**
  Initialize $\mathcal{D}_{D-1} = \emptyset$.
  **Loop on elements** $b \in \mathcal{D}_D$**:**
    **Loop on dimension** $k \in \{1, \ldots, d\}$ **and** $|L_k(b)| > 0$**:**
    Let $b'$ denote the sibling of $b$ along dimension $k$, i.e. the bin obtained from $b$ by flipping the last element of $L_k(b)$. Let $u$ denote the direct common ancestor-bin of $b$ and $b'$.
    If $u$ is already stored in $\mathcal{D}_{D-1}$ with a (provisional) $T^*_u$, then replace
        $T^*_u \longleftarrow \operatorname{Arg\,Min}\left(\mathcal{E}(T^*_u), \mathcal{E}(T^*_b) + \mathcal{E}(T^*_{b'})\right).$
    If $u$ is not yet stored in $\mathcal{D}_{D-1}$, store it along with the provisional
        $T^*_u \longleftarrow \operatorname{Arg\,Min}\left(\mathcal{E}(R_u), \mathcal{E}(T^*_b) + \mathcal{E}(T^*_{b'})\right).$
    **Endloop on** $k$
  **Endloop on** $b$
**Endloop on** $D$
```

It is straightforward to prove that at the end of each loop over b, $\mathcal{D}_{D-1}$ contains all nonempty bins of depth $D-1$ with the corresponding optimal local trees. Therefore at the end of the procedure $\mathcal{D}_0$ contains the tree minimizing the optimization problem (11).

We now give a result about the complexity of our procedure:

Proposition 1. *For fixed training sample size $n \geq 1$, input dimension $d \geq 1$, maximum number of splits along each dimension $k_{max} \geq 1$, the complexity $\mathcal{C}(n, d, k_{max})$ of the dictionary-based algorithm satisfies*

$$\mathcal{O}\left(dk_{max}^d\right) \leq \mathcal{C}(n, d, k_{max}) \leq \mathcal{O}\left(ndk_{max}^d \log(nk_{max}^d)\right). \tag{18}$$

Proof. For a given training point (X_i, Y_i), the exact number of bins (at any depth) that contain this point is $(k_{max}+1)^d$. Namely, there is a unique bin b_0 of maximal depth dk_{max} containing (X_i, Y_i); then, any other bin b containing this point must be an "ancestor" of b_0 in the sense that for all $1 \leq k \leq d$, $L_k(b)$ must be a prefix list of $L_k(b_0)$. Bin b is uniquely determined by the length of the prefix lists $|L_k(b)|$, $1 \leq k \leq d$; for each length there are $(k_{max}+1)$ possible choices, hence the result.

Since the algorithm must loop at least through all of these bins, and makes an additional loop on dimension for each bin, this gives the lower bound. For the upper bound, we bound the total number of bins for all training points by $\mathcal{O}(nk^d)$. Note that we can implement a dictionary $\mathcal{D}$ such that search and insert operations are of complexity $\mathcal{O}(\log(|\mathcal{D}|))$ (for example an AVL tree, [13]). Coarsely upper-bounding the size of the dictionaries used by the total number of bins, we get the announced upper bound. □

Retaining nk_{max}^d as the leading factor of the upper bound, we see that the complexity of the dictionary-based algorithm is still exponential in the dimension d. To fix ideas, assume that we choose k_{max} so that the projection of the training set on any coordinate axis is totally separated by the regular grid of size $2^{-k_{max}}$. If the distribution of X has a bounded density wrt. Lebesgue measure, k_{max} should be of order $\log(n)$ and the complexity of the algorithm of order $n \log^d(n)$ (in the sense of logarithmic equivalence). Although it is much better than looping through every possible bin (which gives rise to a complexity of order $2^{d(k_{max}+1)} \overset{\log}{\approx} n^d$), it means that the algorithm will only be viable for low dimensional problems, or by imposing restrictions on k_{max} for moderate dimensional problems. Note however that other existing algorithms for dyadic decision trees [9, 10, 6] are all of complexity $2^{dk_{max}}$, but that the authors choose k_{max} of the order of $d^{-1} \log n$. This makes sense in [10], because the cuts are fixed in advance and the algorithm is not adaptive to anisotropy. However, in [6] the author notices that k_{max} should be chosen as large as the computational complexity permits to take full advantage of the anisotropy adaptivity.

4 Discussion and future directions

The two main points of our work are a theoretical study of the estimator and a practical algorithm. On the theoretical side, Theorems 1-2 are "true" oracle inequalities in the sense that the convergence rates for each of the models considered is of the order of the minimax rate (for a study of minimax rates for classification on finite VC-dimension models under the identifiability condition (12), see [3]). Theorem 3 misses the minimax rate, which is known to be of order

$\mathcal{O}(|m|/n)$, by a logarithmic factor. We do not know at this point if this factor can be alleviated. Another interesting future direction is to derive from these inequalities convergence rates for anisotropic regularity function classes, similarly to what was done in [6, 12].

From the algorithmic side, our algorithm is arguably only viable for low- or moderate-dimensional problems (we tested it on 10-dimensional datasets). For application to high-dimensional problems, some partly-greedy heuristic appears as an interesting strategy, for example by splitting the algorithm into several lower-dimensional problems on which we can can run the exact algorithm. We are currently investigating this direction.

Acknowledgments The authors want to thank Lucien Birgé and Klaus-Robert Müller for valuable discussions.

A Proofs of Theorems 1-3

The proofs for our results are based on a general model selection theorem appearing in [14], which is a generalization of an original theorem of Massart [1]. We quote it here in a slightly modified and shortened form tailored for our needs (see also [15] for a similar form of the theorem).

Theorem 4. *Let $\ell(\cdot,\cdot)$ be a loss function defined on $\mathcal{S} \times \mathcal{X}$; denote $f^* = \operatorname*{Arg\,Min}_{f \in \mathcal{S}} P\ell(f)$ and $L(f, f^*) = P\ell(f) - P\ell(f^*)$. Let $(S_m)_{m\in\mathcal{M}}$, $S_m \subset \mathcal{S}$ be a countable collection of classes of functions and assume that there exists*

- *a pseudo-distance d on $\mathcal{S}$;*
- *a sequence of sub-root* [4] *functions $(\phi_m), m \in \mathcal{M}$;*
- *two positive constants b and R ;*

such that

(H1) $\quad \forall f \in \mathcal{S},\ \forall x \in \mathcal{X}, \quad |\ell(f, x)| \leq b$;

(H2) $\quad \forall f, f' \in \mathcal{S}, \quad \mathrm{Var}_P[\ell(f) - \ell(f')] \leq d^2(f, f')$;

(H3) $\quad \forall f \in \mathcal{S}, \quad d^2(f, f^*) \leq RL(f, f^*)$;

and, if r_m^* denotes the solution of $\phi_m(r) = r/R$,

(H4) $\forall m \in \mathcal{M},\ \forall f_0 \in \mathcal{F}_m,\ \forall r \geq r_m^*$

$$E\left[\sup_{\substack{f\in\mathcal{F}_m \\ d^2(f,f_0)\leq r}} (P - P_n)(\ell(f) - \ell(f_0))\right] \leq \phi_m(r).$$

Let $(x_m)_{m\in\mathcal{M}}$ be real numbers with $\sum_{m\in\mathcal{M}} e^{-x_m} \leq 1$. Let $\varepsilon \geq 0$ and $\widetilde{f}$ denote an ε-approximate penalized minimum loss estimator over the family $(\mathcal{F}_m)$ with the penalty function $\mathrm{pen}(m)$*, that is, such that there exists $\widetilde{m}$ with $\widetilde{f} \in \mathcal{F}_{\widetilde{m}}$ and*

$$P_n\ell(\widetilde{f}) + \mathrm{pen}(\widetilde{m}) \leq \inf_{m\in\mathcal{M}} \inf_{f\in\mathcal{F}_m} \left(P_n\ell(f) + \mathrm{pen}(m) + \varepsilon\right).$$

[4] A function ϕ on $\mathbb{R}_+$ is subroot if it is positive, nondecreasing and $\phi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Given $K > 1$, there exist constants C_1, C_2, C_3 (depending on K only) such that, if the penalty function $\operatorname{pen}(m)$ *satisfies for each $m \in \mathcal{M}$:*

$$\operatorname{pen}(m) \geq C_1 \frac{r_m^*}{R} + C_2 \frac{(R+b)x_m}{n},$$

then the following inequality holds:

$$EL(\widetilde{f}, f^*) \leq K \inf_{m \in \mathcal{M}} \left(\inf_{f \in \mathcal{F}_m} L(f, f^*) + 2\operatorname{pen}(m) + \frac{C_3}{n} + \varepsilon \right).$$

Proof outline for Theorem 1. We will apply Theorem 4 to the set of models $(\mathcal{C}_m)$. Checking for hypothesis (H1) is obvious. To check (H2)-(H3), we choose the distance $d(f,g) = E\left[(\ell_{class}(f,X,Y) - \ell_{class}(g,X,Y))^2\right]$, so that (H2) is trivially satisfied. To check (H3), denote $\eta(x,i) = P(Y = i|X = x)$ and $\eta^*(x) = \max_{i \in \mathcal{Y}} \eta(i,x)$; we then have

$$\begin{aligned} E\left[\mathbb{I}_{\{f(X)\neq Y\}} - \mathbb{I}_{\{f^*(X)\neq Y\}}\right] &= E\left[(\eta^*(X) - \eta(X, f(X)))\,\mathbb{I}_{\{f(X)\neq f^*(X)\}}\right] \\ &\geq \eta_0 E\left[\mathbb{I}_{\{f(X)\neq f^*(X)\}}\right], \end{aligned}$$

where we have used hypothesis (12). On the other hand,

$$\begin{aligned} E\left[(\mathbb{I}_{\{f(X)\neq Y\}} - \mathbb{I}_{\{f^*(X)\neq Y\}})^2\right] &= E\left[(\eta^*(X) + \eta(X, f(X)))\,\mathbb{I}_{\{f(X)\neq f^*(X)\}}\right] \\ &\leq 2E\left[\mathbb{I}_{\{f(X)\neq f^*(X)\}}\right], \end{aligned}$$

which proves that (H3) is satisfied with $R = 2/\eta_0$. Finally, for hypothesis (H4), we can follow the same reasoning as in [1], p. 294-295; in this reference the empirical shattering coefficient is taken into account, but the present case is even simpler since model $\mathcal{C}_m$ is finite with cardinality $t^{|m|}$, leading to

$$E\left[\sup_{f \in \mathcal{C}_m, d^2(f,f_0)\leq r} (P - P_n)(\ell_{class}(f) - \ell_{class}(f_0))\right] \leq C\sqrt{\frac{r|m|\log t}{n}},$$

for some universal constant C. This leads to the conclusion. □

Proof outline for Theorem 2. We apply Theorem 4 to the set of models $(\mathcal{F}_m)$. For (H1), it is easy to check that

$$\forall f \in \mathcal{F}_{cpd}, \qquad \ell_{sq}(f,X,Y) = \left\| f(X,\cdot) - \overline{Y} \right\|_t^2 = \|f(X,\cdot)\|_t^2 + 1 - 2f(X,Y) \leq 2.$$

For (H2), we note that $\ell_{sq}(f,X,Y) - \ell_{sq}(g,X,Y) = \|f(X,\cdot)\|_t^2 - \|g(X,\cdot)\|_t^2 - 2(f(X,Y) - g(X,Y))$. Using the equality $\operatorname{Var}[F] = E[\operatorname{Var}[F|X]] + \operatorname{Var}[E[F|X]]$, we deduce that

$$\begin{aligned} &\operatorname{Var}\left[\ell_{sq}(f,X,Y) - \ell_{sq}(g,X,Y)\right] \\ &\qquad = E\left[\operatorname{Var}\left[2(f(X,Y) - g(X,Y))|X\right]\right] + \operatorname{Var}\left[\|f(X,\cdot)\|_t^2 - \|g(X,\cdot)\|_t^2\right] \\ &\qquad \leq 4E\left[(f-g)^2\right] + E\left[\|f(X,\cdot) - g(X,\cdot)\|_t^2 \, \|f(X,\cdot) + g(X,\cdot)\|_t^2\right] \\ &\qquad \leq 8E\left[\|f(X,\cdot) - g(X,\cdot)\|_t^2\right] \doteq d^2(f,g); \end{aligned}$$

this proves that (H2) is satisfied for the above choice of d; recalling (6), (H3) is then satisfied with $R = 1/8$. Finally, for hypothesis (H4) is is possible to show that

$$E\left[\sup_{f\in\mathcal{G}_m, d^2(f,f_0)\le r}(P-P_n)(\ell_{sq}(f)-\ell_{sq}(f_0))\right]\le C\sqrt{\frac{rt|m|}{n}},$$

using local Rademacher and Gaussian complexities, using a method similar to [14]. □

Proof of Theorem 3. To apply Theorem 4, we define the ambient space

$$\mathcal{S}^\rho = \{f\in\mathcal{F}_{cpd}|\forall(x,y)\in\mathcal{X}\times\mathcal{Y},\ f(x,y)\ge\rho\}$$

and the models as $S^\rho_m = \mathcal{S}^\rho\cap\mathcal{F}_m$, which will insure boundedness of the loss. As a counterpart of using these restricted ambient space and models, the application of Theorem 4 will result in an inequality involving not f^*_{cpd}, but the minimizer of the average loss on $\mathcal{S}^\rho$, denoted f^*_ρ, and the model-wise minimizers of the loss on $\mathcal{S}^\rho_m$ instead of $\mathcal{F}_m$. However, it is easy to show the following inequalities:

$$\forall f\in\mathcal{F}_{cpd},\qquad L(f,f^*_{cpd})\le L(f,f^*_\rho)-\log(1-t\rho);$$

$$\forall m\in\mathcal{M},\qquad \inf_{f\in\mathcal{S}^\rho_m}L(f,f^*_\rho)\le\inf_{f\in\mathcal{F}_m}L(f,f^*_{cpd})-\log(1-t\rho);$$

finally, it can be shown that $\widehat{f}^\rho_{\widehat{m}}$ is a $-\log(1-t\rho)$-approximate penalized estimator. Therefore, if Theorem 4 applies, these inequalities lead to the conclusion of Theorem 3.

We now turn to verifying the main assumptions of the abstract model selection theorem.

- Check for (H1): boundedness of the loss on the models. Obviously, we have

$$\forall f\in\mathcal{S}^\rho,\ \forall(x,y)\in\mathcal{X}\times\mathcal{Y}\qquad 0\le\ell_{log}(f,x,y)\le-\log\rho$$

- Check for (H2)-(H3): distance linking the risk and its variance. We choose the distance d as the $L^2(P)$ distance between logarithms of the functions:

$$d(f,g)=E_P\left[(\ell_{log}(f,x,y)-\ell_{log}(g,x,y))^2\right]=E_P\left[\log^2\frac{f}{g}\right].$$

Obviously we have $Var[\ell_{log}(f,x,y)-\ell_{log}(g,x,y)]\le d(f,g)$ with this choice; the problem is then to compare $E_P\left[\log^2\frac{P(Y|X)}{f}\right]$ to $E_P\left[\log\frac{P(Y|X)}{f}\right]$. Denoting $Z(x,i)=f(x,k)/P(Y=k|X=x)$, we therefore have to compare $E[\log^2 Z]$ to $E[-\log Z]$ with the expectation taken wrt. P, so that $E[Z]=1$. Note that $Z\ge\rho$. Using Lemma 1 below, we deduce that

$$d(P(Y|X),f)\le\frac{\log^2\rho}{\rho-1-\log\rho}KL(P,f|X),$$

Note that typically when ρ is small the factor R in (H3) is therefore of order $-\log\rho$.

- Check for (H4): d-local risk control on models. For any $f, g \in \mathcal{S}_m^\rho$, $F = \log \dfrac{f}{g} \in \mathcal{G}_m$. For $A \in \mathcal{B}_m, i \in \mathcal{Y}$, denote $P_{A,i} = P[X \in A, Y = i]$ and

$$\varphi_{A,i}(x,y) = \frac{\mathbb{I}\{x \in A\}\,\mathbb{I}\{Y = i\}}{\sqrt{P_{A,i}}};$$

note that the family $(\varphi(A,i))_{A,i}$ is an orthonormal basis (for the $L^2(P)$ structure) of $\mathcal{G}_m$, hence any function $f \in \mathcal{G}_m$ can be written under the form

$$f = \sum_{A,i} \alpha_{A,i} \varphi_{A,i}\ ,$$

with $Pf^2 = \sum \alpha_{A,i}^2$. Putting $\nu_n = (P - P_n)$, we then have for any $f \in \mathcal{S}_m^\rho$

$$E_P\left[\sup_{\substack{g \in S_m^\rho \\ d^2(f,g) \le r}} |\nu_n(\ell(f,x) - \ell(g,x))|\right] \le E_P\left[\sup_{\substack{F \in \mathcal{G}_m \\ E[F^2] \le r}} |\nu_n F|\right] = \Xi;$$

$$\begin{aligned}
\Xi = E_P\left[\sup_{\substack{(\alpha_{A,i}): \\ \sum_{A,i} \alpha_{A,i}^2 \le r}} \left|\sum_{A,i} \alpha_{A,i}\, \nu_n \phi_{A,i}\right|\right] &\le \sqrt{r} E_P\left[\left(\sum_{A,i} (\nu_n \varphi_{A,i})^2\right)^{\frac{1}{2}}\right] \\
&\le \sqrt{r} E_P\left[\left(\sum_{A,i} (\nu_n \varphi_{A,i})^2\right)\right]^{\frac{1}{2}} \\
&= \sqrt{r \sum_{A,i} \frac{1}{n} \frac{P_{A,i}(1 - P_{A,i})}{P_{A,i}}} \le \sqrt{\frac{rt|m|}{n}}.
\end{aligned}$$

□

The following Lemma is inspired by similar techniques appearing in [4, 16].

Lemma 1. *Let Z be a real, positive random variable such that $E[Z] = 1$ and $Z \ge \eta$ a.s. Then the following inequality holds:*

$$\frac{E\left[\log^2 Z\right]}{E\left[-\log Z\right]} \le \frac{\log^2 \eta}{\eta - 1 - \log \eta}.$$

Proof. Let $u = -\log Z \le -\log \eta$; we have

$$\begin{aligned}
E[-\log Z] = E[u] = E[e^{-u} - 1 + u] &= E\left[u^2 \frac{e^{-u} - 1 + u}{u^2}\right] \\
&\ge E\left[u^2\right] \frac{\eta - 1 - \log \eta}{\log^2 \eta},
\end{aligned}$$

where the first line comes from the fact that $E\left[e^{-u}\right] = E\left[Z\right] = 1$, and the last inequality from the fact that the function $g(x) = x^{-2}(e^{-x} - 1 + x)$ is positive and decreasing on $\mathbb{R}$. □

References

1. Massart, P.: Some applications of concentration inequalities in statistics. Ann. Fac. Sci. Toulouse Math. **9** (2000) 245–303
2. Tsybakov, A.: Optimal aggregation of classifiers in statistical learning. Annals of Statistics **32** (2004)
3. Massart, P., Nédélec, E.: Risk bounds for statistical learning. Technical report, Laboratoire de mathématiques, Université Paris-Sud (2004)
4. Castellan, G.: Histograms selection with an Akaike type criterion. C. R. Acad. Sci., Paris, Sér. I, Math. **330** (2000) 729–732
5. Barron, A., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. Probability theory and related fields **113** (1999) 301–413
6. Klemelä, J.: Multivariate histograms with data-dependent partitions. Technical report, Institut für angewandte mathematik, Universität Heidelberg (2003)
7. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression Trees. Wadsworth, Belmont, California (1984)
8. Gey, S., Nédélec, E.: Risk bounds for CART regression trees. In: Nonlinear Estimation and Classification. Volume 171 of Lecture Notes in Statistics. Springer (2003) 369–380
9. Scott, C., Nowak, R.: Dyadic classification trees via structural risk minimization. In: Proc. Neural Information Processing Systems (NIPS). (2002)
10. Scott, C., Nowak, R.: Near-minimax optimal classification with dyadic classification trees. In: Proc. Neural Information Processing Systems (NIPS). (2003)
11. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition. Volume 31 of Applications of Mathematics. Springer (1996)
12. Donoho, D.L.: CART and best ortho-basis: a connection. Annals of Statistics **25** (1997) 1870–1911
13. Adelson-Velskii, G.M., Landis, E.: An algorithm for the organization of information. Soviet Math. Doclady **3** (1962) 1259–1263
14. Blanchard, G., Bousquet, O., Massart, P.: Statistical performance of Support Vector Machines. Technical report, Laboratoire de mathématiques, Université Paris-Sud (2004)
15. Blanchard, G., Lugosi, G., Vayatis, N.: On the rate of convergence of regularized Boosting classifiers. Journal of Machine Learning Research **4** (2003) 861–894
16. Barron, A., Sheu, C.: Approximation of density functions by sequences of exponential families. Annals of Statistics **19** (1991) 1347–1369