Classification algorithms using adaptive partitioning

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald Devore

▶ **To cite this version:**

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald Devore. Classification algorithms using adaptive partitioning. Annals of Statistics, 2014, 42 (6), pp.2141-2163. 10.1214/14-AOS1234SUPP . hal-01217382

HAL Id: hal-01217382
https://hal.science/hal-01217382

Submitted on 20 Oct 2015

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

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

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

CLASSIFICATION ALGORITHMS USING ADAPTIVE PARTITIONING[??]

By Peter Binev[*,†], Albert Cohen[*,‡] Wolfgang Dahmen[*,§] and Ronald DeVore[*,¶]

University of South Carolina[†], Université Pierre et Marie Curie[‡], RWTH Aachen[§], Texas A & M University[¶]

Algorithms for binary classification based on adaptive tree partitioning are formulated and analyzed for both their risk performance and their friendliness to numerical implementation. The algorithms can be viewed as generating a set approximation to the Bayes set and thus fall into the general category of *set estimators*. In contrast with the most studied tree based algorithms, which utilize piecewise constant approximation on the generated partition [15, 6], we consider decorated trees, which allow us to derive higher order methods. Convergence rates for these methods are derived in terms the parameter α of margin conditions and a rate s of best approximation of the Bayes set by decorated adaptive partitions. They can also be expressed in terms of the Besov smoothness β of the regression function that governs its approximability by piecewise polynomials on adaptive partition. The execution of the algorithms does not require knowledge of the smoothness or margin conditions. Besov smoothness conditions are weaker than the commonly used Hölder conditions, which govern approximation by non-adaptive partitions, and therefore for a given regression function can result in a higher rate of convergence. This in turn mitigates the compatibility conflict between smoothness and margin parameters.

1. Introduction. A large variety of methods have been developed for classification of randomly drawn data. Most of these fall into one of two basic categories: *set estimators* or *plug-in estimators*. Both of these fami-

[*]This research was supported by the Office of Naval Research Contracts ONR-N00014-08-1-1113, ONR N00014-09-1-0107; the AFOSR Contract FA95500910500; the ARO/DoD Contract W911NF-07-1-0185; the NSF Grants DMS 0915231, DMS 1222390, and DMS 0915104; the Special Priority Program SPP 1324, funded by DFG; the French-German PROCOPE contract 11418YB; the Agence Nationale de la Recherche (ANR) project ECHANGE (ANR-08-EMER-006); the excellence chair of the Foundation "Sciences Mathématiques de Paris" held by Ronald DeVore. This publication is based on work supported by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

MSC 2010 subject classifications: 62M45, 65D05, 68Q32, 97N50

Keywords and phrases: binary classification, adaptive methods, set estimators, tree based algorithms

lies are based on some underlying form of approximation. In the case of set estimators, one directly approximates the *Bayes set*, using elements from a given family of sets. For plug-in estimators, one approximates the underlying *regression function* and builds the classifier as a level set of this approximation.

The purpose of this paper is to introduce a family of classification algorithms using tree-based adaptive partitioning and to analyze their risk performance as well as their friendliness to numerical implementation. These algorithms fall in the category of set estimators. Tree-based classification algorithms have been well studied since their introduction in [8], and their convergence properties have been discussed both in terms of oracle inequalities and minimax convergence estimates, see e.g. [15] and [6]. Among the specific features of the approach followed in our paper are (i) the use of decorated trees which allow us to derive faster rates for certain classes of distributions than obtained when using standard trees and (ii) a convergence analysis based on nonlinear approximation theory which allows us to significantly weaken the usual assumptions that are made to establish a given convergence rate. More detailed comparisons with existing methods and results are decribed later.

We place ourselves in the following setting of binary classification. Let $X \subset \mathbb{R}^d$, $Y = \{-1, 1\}$ and $Z = X \times Y$. We assume that $\rho = \rho_X(x) \cdot \rho(y|x)$ is a probability measure defined on Z. We denote by $p(x)$ the probability that $y = 1$ given x and by $\eta(x)$ the regression function

$$\eta(x) := \mathbb{E}(y|x) = p(x) - (1 - p(x)) = 2p(x) - 1, \tag{1.1}$$

where $\mathbb{E}$ denotes expectation. For any set S, we use the notation

$$\rho_S := \rho_X(S) = \int_S d\rho_X \quad \text{and} \quad \eta_S := \int_S \eta \, d\rho_X. \tag{1.2}$$

A classifier returns the value $y = 1$ if x is in some set $\Omega \subset X$ and $y = -1$ otherwise. Therefore, the classifier is given by a function $T_\Omega = \chi_\Omega - \chi_{\Omega^c}$ where Ω is some ρ_X measurable set and Ω^c is its complement. With a slight abuse of notation, we sometimes refer to the set Ω itself as the classifier. We denote by $R(\Omega) := \mathbb{P}\{T_\Omega(x) \neq y\}$ the risk (probability of misclassification) of this classifier, and by $\Omega^* := \{x \ : \ \eta(x) \geq 0\}$ the Bayes classifier which minimizes this risk $R(\Omega)$, or equivalently, maximizes the quantity η_Ω among all possible sets Ω.

We measure the performance of a classifier Ω by the *excess risk*

$$R(\Omega) - R(\Omega^*) = \int_{\Omega \triangle \Omega^*} |\eta| \, d\rho_X, \tag{1.3}$$

with $A \triangle B := (A - B) \cup (B - A)$ the symmetric difference between A and B.

Given the data $\mathbf{z} = (z_i)_{i=1}^n$, $z_i = (x_i, y_i)$, $i = 1, \dots, n$, drawn independently according to ρ, a classification algorithm uses the draw to find a set $\hat{\Omega} = \hat{\Omega}(\mathbf{z})$ to be used as a classifier. Obtaining a concrete estimate of the decay of the excess risk for a given classification algorithm as n grows requires assumptions on the underlying measure ρ. These conditions are usually spelled out by assuming that ρ is in a *model class* $\mathcal{M}$. Model classes are traditionally formed by two ingredients: (i) assumptions on the behavior of ρ near the boundary of the Bayes set Ω^* and (ii) assumptions on the smoothness of the regression function η.

Conditions that clarify (i) are called margin conditions and are an item of many recent papers, see e.g. [14, 16]. One formulation (sometimes referred to as the Tsybakov condition) requires that

$$\rho_X\{x \in X : \ |\eta(x)| \leq t\} \leq C_\alpha t^\alpha, \quad 0 < t \leq 1, \tag{1.4}$$

for some constant $C_\alpha > 0$ and $\alpha \geq 0$. This condition becomes more stringent as α tends to $+\infty$. The limiting case $\alpha = \infty$, known as Massart condition, means that for some $A > 0$, we have $|\eta| > A$ almost everywhere. A common choice for (ii) in the classification literature, see e.g. [2], is that η belongs to the Hölder space C^β. This space can be defined for any $\beta > 0$ as the set of functions f such that

$$\|\Delta_h^m f\|_{L_\infty} \leq C|h|^\beta, \ \ h \in \mathbb{R}^d, \tag{1.5}$$

where Δ_h^m is the m-th power of the finite difference operator defined by $\Delta_h f = f(\cdot + h) - f$, with m being the smallest integer such that $m \geq \beta$. The Hölder class may be viewed intuitively as the set of functions whose derivatives of fractional order β belong to L_∞.

An important observation is that there is a conflict between margin and smoothness assumptions, in the sense that raising the smoothness β limits the range of α in the margin condition. For example, when ρ_X is the Lebesgue measure on X, it is easily checked that the constraint $\alpha\beta \leq 1$ must hold as soon as the Bayes boundary $\partial\Omega^*$ has non-zero $(d-1)$-dimensional Hausdorff measure.

An instance of a convergence result exploiting (i) and (ii), Theorem 4.3 in [2], is that under the assumption that the density of ρ_X with respect to the Lebesgue measure is uniformly bounded, certain classifiers based on plug-in rules achieve in expectation the rate

$$\mathbb{E}(R(\hat{\Omega}) - R(\Omega^*)) \leq C n^{-\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}} \tag{1.6}$$

if the margin assumption holds with parameter α, and if η belongs to the Hölder class C^β.

The classification algorithms that we study in this paper have natural links with the process of approximating the regression function using piecewise constant or piecewise polynomials on adaptive partitions, which is an instance of *nonlinear approximation*. It is well-known in approximation theory that, when using nonlinear methods, the smoothness condition needed to attain a specified rate can be dramatically weakened. This state of affairs is reflected in the convergence results for our algorithms that are given in Theorems 6.1 and 6.3. These results say that with high probability (larger than $1 - Cn^{-r}$ where $r > 0$ can be chosen arbitrarily large), our classifiers achieve the rate

$$R(\hat{\Omega}) - R(\Omega^*) \leq C\Big(\frac{n}{\log n}\Big)^{-\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}} \tag{1.7}$$

if the margin assumption holds with parameter α and if η belongs to the Besov space $B^\beta_\infty(L_p)$ with $p > 0$ such that $\beta p > d$. This Besov space is defined by the condition

$$\|\Delta_h^m f\|_{L_p} \leq C|h|^\beta, \quad h \in \mathbb{R}^d, \tag{1.8}$$

with m being an integer such that $m > \beta$ and may be viewed as the set of function whose derivatives of fractional order β belong to L_p. Notice that the constraint $\beta p > d$ ensures that $B^\beta_\infty(L_p)$ is compactly embedded in L_∞. Therefore our algorithm achieves the same rate as (1.6), save for the logarithm, however with a significant weakening on the smoothness condition imposed on the regression function because of the use of adaptive partitioning. In particular, an individual regression function may have significantly higher smoothness β in this scale of Besov spaces than in the scale of Hölder spaces, resulting in a better rate when using our classifier.

In addition, the weaker smoothness requirement for a given convergence rate allows us to alleviate the conflict between smoothness and margin conditions in the sense that the constraint $\alpha\beta \leq 1$ can be relaxed when using the space $B^\beta_\infty(L_p)$, see (6.12)). Let us also observe that our risk bound in (1.7) holds in the stronger sense of high probability, rather than expectation, and that no particular assumption (such as equivalence with Lebesgue measure) is made on the density of ρ_X. Finally, let us stress that our algorithms are numerically implementable and do not require the a-priori knowledge of the parameters α and β.

The distinction between Theorems 6.1 and 6.3 is the range of β for which they apply. Theorem 6.1 only applies to the range $\beta \leq 1$ and can be seen

as the analog of using piecewise constant approximation on adaptive partition for plug-in estimators. On the other hand, Theorem 6.3 applies for any $\beta \leq 2$. The gain in the range of β results from the fact that the algorithm uses decorated trees. This correspond to piecewise affine approximation for plug-in methods. In principle, one can extend the values of β arbitrarily by using higher polynomial order decorated trees. However, the numerical implementation of such techniques becomes more problematic and is therefore not considered in this paper. In the regression context, piecewise polynomial estimators on adaptive partitions have been considered in [1, 4].

Set estimators aim at approximating the Bayes set Ω^* by elements S from a family of sets $\mathcal{S}$ in the sense of the distance defined by the excess risk. Our approach to deriving the risk bounds in Theorems 6.1 and 6.3 is by splitting this risk into

$$R(\hat{\Omega}) - R(\Omega^*) = \Big(R(\hat{\Omega}) - R(\Omega_{\mathcal{S}})\Big) + \Big(R(\Omega_{\mathcal{S}}) - R(\Omega^*)\Big), \tag{1.9}$$

where

$$\Omega_{\mathcal{S}} := \operatorname*{argmin}_{S\in\mathcal{S}} R(S) = \operatorname*{argmax}_{S\in\mathcal{S}} \eta_S. \tag{1.10}$$

The two terms are positive, and are respectively referred to as the estimation error and approximation error. We bound in §2 the estimation error by introducing of a certain modulus which is defined utilizing any available uniform estimate between the quantity $\eta_S - \eta_{\Omega_S}$ and its empirical counterpart computed from the draw. For set estimators based on empirical risk minimization, we show in §3 how margin conditions can be used to bound this modulus, and therefore the estimation error term.

In §4, we turn to estimates for the approximation term. This analysis is based on the smoothness of η and the margin condition. A typical setting when building set classifiers is a nested sequence $(\mathcal{S}_m)_{m\geq 1}$ of families of sets, i.e. $\mathcal{S}_m \subset \mathcal{S}_{m+1}$, where m describes the complexity of $\mathcal{S}_m$ in the sense of VC dimension. The value of m achieving between the optimal balance between the estimation and approximation terms depends on the parameters α and β which are unknown. A standard model selection procedure is discussed in §5 that reaches this balance for a variety of model classes $\mathcal{M} = \mathcal{M}(\alpha, \beta)$ over a range of α and β.

Many ingredients of our analysis of general classification methods appear in earlier works, see e.g. [7, 11]. However, in our view, the organization of the material in these sections helps clarify various issues concerning the roles of approximation and estimation error bounds.

In §6, we turn to our proposed classification methods based on adaptive partitioning. We analyze their performance using the results from the previous sections and arrive at the aforementioned Theorems 6.1 and 6.3. The numerical implementation and complexity of these algorithms are discussed in §7.

2. A general bound for the estimation error in set estimators. In view of $\Omega^* = \operatorname{argmax}_{\Omega \subset X} \eta_\Omega$, if $\hat{\eta}_S$ is any empirical estimator for η_S, a natural way to select a classifier within $\mathcal{S}$ is by

$$\hat{\Omega} := \hat{\Omega}_{\mathcal{S}} := \operatorname*{argmax}_{S \in \mathcal{S}} \hat{\eta}_S. \tag{2.1}$$

One of the most common strategies for building $\hat{\eta}_S$ is by introducing the empirical counterparts to (1.2),

$$\overline{\rho}_S := \frac{1}{n}\sum_{i=1}^n \chi_S(x_i) \ \text{ and } \ \overline{\eta}_S = \frac{1}{n}\sum_{i=1}^n y_i \chi_S(x_i). \tag{2.2}$$

The choice $\hat{\eta}_S = \overline{\eta}_S$ is equivalent to minimizing the empirical risk over the family $\mathcal{S}$, namely choosing

$$\hat{\Omega}_{\mathcal{S}} = \overline{\Omega}_{\mathcal{S}} := \operatorname*{argmin}_{S \in \mathcal{S}} \overline{R}(S), \qquad \overline{R}(S) := \frac{1}{n}\#\{i : T_S(x_i) \neq y_i\}, \tag{2.3}$$

However, other ways of defining $\hat{\eta}_S$ are conceivable leading to different types of classifiers. Of course, an important point is whether such classifiers have a reasonable numerical implementation.

We give in this section a general method for bounding the estimation error, whenever we have an empirical estimator $\hat{\eta}_S$ for η_S, with a bound of the form

$$|\eta_S - \eta_{\Omega_S} - (\hat{\eta}_S - \hat{\eta}_{\Omega_S})| \leq e_n(S), \tag{2.4}$$

for each set $S \in \mathcal{S}$. In the case where we use for $\hat{\eta}_S$ the set estimators $\bar{\eta}_S$ defined in (2.2), we have the following bound.

THEOREM 2.1. *For any sufficiently large constant $A > 0$ the following holds. If $\mathcal{S}$ is a collection of ρ_X measurable sets $S \subset X$ with finite VC dimension $V := V_{\mathcal{S}}$, and if*

$$e_n(S) := \sqrt{\rho_{S\Delta\Omega_S}\varepsilon_n} + \varepsilon_n, \quad \varepsilon_n := A\max\{r+1, V\}\frac{\log n}{n}, \tag{2.5}$$

where $r > 0$ is arbitrary, then there is an absolute constant C_0 such that for any $n \geq 2$, with probability at least $1 - C_0 n^{-r}$ on the draw $\mathbf{z} \in Z^n$, we have

$$|\eta_S - \eta_{\Omega_\mathcal{S}} - (\bar{\eta}_S - \bar{\eta}_{\Omega_\mathcal{S}})| \leq e_n(S), \quad S \in \mathcal{S}. \tag{2.6}$$

The techniques for proving this result are well known in classification but we could not find any reference which gives the bounds in this theorem in the above form. For this reason, we give a proof of this theorem in the appendix which also shows the conditions on the constant A.

REMARK 2.2. *The above theorem covers, in particular, the case where $\mathcal{S}$ is* finite *collection of sets, since then trivially $V_\mathcal{S} \leq \#\mathcal{S}$. Alternatively, in this case, a straightforward argument using Bernstein's inequality yields the same result with the explicit expression $\varepsilon_n := \frac{10(\log(\#\mathcal{S}) + r \log n)}{3n}$ and probability at least $1 - 2n^{-r}$.*

To analyze the estimation error in classifiers, we define the following modulus:

$$\omega(\rho, e_n) := \sup \left\{ \int\limits_{S \Delta \Omega_\mathcal{S}} |\eta| \ : \ S \in \mathcal{S} \ \text{ and } \int\limits_{S \Delta \Omega_\mathcal{S}} |\eta| \leq 3 e_n(S) \right\}. \tag{2.7}$$

Notice that the second argument e_n is not a number but rather a set function. In the next section, we discuss this modulus in some detail and bring out its relation to other ideas used in classification, such as margin conditions. For now, we use it to prove the following theorem.

THEOREM 2.3. *Suppose that for each $S \in \mathcal{S}$, we have that (2.4) holds with probability $1 - \delta$. Then with this same probability, we have*

$$R(\hat{\Omega}_\mathcal{S}) - R(\Omega_\mathcal{S}) \leq \max\{\omega(\rho, e_n), a(\Omega^*, \mathcal{S})\} \tag{2.8}$$

with $a(\Omega^*, \mathcal{S}) := R(\Omega_\mathcal{S}) - R(\Omega^*)$ being the approximation error from (1.9).

Proof: We consider any data $\mathbf{z}$ such that (2.4) holds and prove that (2.8) holds for such $\mathbf{z}$. Let $S_0 := \Omega_\mathcal{S} \setminus \hat{\Omega}_\mathcal{S}$ and $S_1 := \hat{\Omega}_\mathcal{S} \setminus \Omega_\mathcal{S}$ so that $S_0 \cup S_1 = \hat{\Omega}_\mathcal{S} \Delta \Omega_\mathcal{S}$. Notice that, in contrast to $\Omega_\mathcal{S}$ and $\hat{\Omega}_\mathcal{S}$, the sets S_0, S_1 are generally not in $\mathcal{S}$. We start from the equality

$$R(\hat{\Omega}_\mathcal{S}) - R(\Omega_\mathcal{S}) = \eta_{\Omega_\mathcal{S}} - \eta_{\hat{\Omega}_\mathcal{S}} = \eta_{S_0} - \eta_{S_1}. \tag{2.9}$$

We can assume that $\eta_{S_0} - \eta_{S_1} > 0$ since otherwise we have nothing to prove. From the definition of $\hat{\Omega}_{\mathcal{S}}$, we know that

$$\hat{\eta}_{\Omega_{\mathcal{S}}} - \hat{\eta}_{\hat{\Omega}_{\mathcal{S}}} \leq 0.$$

Using this in conjunction with (2.4), we obtain

$$\eta_{S_0} - \eta_{S_1} = \eta_{\Omega_{\mathcal{S}}} - \eta_{\hat{\Omega}_{\mathcal{S}}} \leq e_n(\hat{\Omega}_{\mathcal{S}}). \tag{2.10}$$

In going further, we introduce the following notation. Given a set $S \subset X$, we denote by $S^+ := S \cap \Omega^*$ and $S^- := S \cap (\Omega^*)^c$. Thus, $\eta \geq 0$ on S^+ and $\eta < 0$ on S^-. Also $S = S^+ \cup S^-$ and $S^+ \cap S^- = \emptyset$. Hence we can write

$$\eta_{S_0} - \eta_{S_1} = A - B, \quad A := \eta_{S_0^+} - \eta_{S_1^-},\ B := \eta_{S_1^+} - \eta_{S_0^-}. \tag{2.11}$$

Note that $A, B \geq 0$. We consider two cases.

Case 1: If $A \leq 2B$, then

$$R(\widehat{\Omega}_{\mathcal{S}}) - R(\Omega_{\mathcal{S}}) = A - B \leq B \leq a(\Omega^*, \mathcal{S}), \tag{2.12}$$

where we have used the fact that $S_1^+ \subset \Omega^* \setminus \Omega_{\mathcal{S}}$ and $S_0^- \subset \Omega_{\mathcal{S}} \setminus \Omega^*$.

Case 2: If $A > 2B$, then, by (2.10) and (2.11),

$$\int\limits_{\hat{\Omega}_{\mathcal{S}} \Delta \Omega_{\mathcal{S}}} |\eta| = A + B \leq 3A/2 \leq 3(A - B) = 3(\eta_{S_0} - \eta_{S_1}) \leq 3e_n(\hat{\Omega}_{\mathcal{S}}). \tag{2.13}$$

This means that $\hat{\Omega}_{\mathcal{S}}$ is one of the sets appearing in the definition of $\omega(\rho, e_n)$ and (2.8) follows in this case from the fact that

$$\eta_{S_0} - \eta_{S_1} = A - B \leq \int\limits_{\hat{\Omega}_{\mathcal{S}} \Delta \Omega_{\mathcal{S}}} |\eta| \leq \omega(\rho, e_n).$$

□

From Theorem 2.3, we immediately obtain the following corollary.

COROLLARY 2.4. *Suppose that for each $S \in \mathcal{S}$,* (2.4) holds *with probability $1 - \delta$. Then with this same probability we have*

$$R(\hat{\Omega}_{\mathcal{S}}) - R(\Omega^*) \leq \omega(\rho, e_n) + 2a(\Omega^*, \mathcal{S}). \tag{2.14}$$

Proof: We have $R(\hat{\Omega}_{\mathcal{S}}) - R(\Omega^*) = R(\hat{\Omega}_{\mathcal{S}}) - R(\Omega_{\mathcal{S}}) + R(\Omega_{\mathcal{S}}) - R(\Omega^*)$. The second term equals $a(\Omega^*, \mathcal{S})$ and the first term is bounded by (2.8). □

REMARK 2.5. *The corollary does not impose any particular assumption on ρ and $\mathcal{S}$, apart from finite VC dimension. For later comparisons with existing results, we briefly illustrate how $e_n(S)$ can be sharpened if one imposes additional assumptions on ρ_X. Assume that ρ_X is (or is equivalent to) the Lebesgue measure and for any arbitrary integer $l \geq 1$ consider a uniform partition $\mathcal{Q}$ of $X = [0,1]^d$ into $m = l^d$ cubes of side length l^{-1}, providing the collection $\mathcal{S}$ of all sets S that are unions of cubes from $\mathcal{Q}$. Then, defining*

$$e_n(S) := \rho_{S\Delta\Omega_{\mathcal{S}}}\sqrt{\varepsilon_n}, \quad \text{where} \quad \varepsilon_n := \frac{8(r+1)m(1+\log n)}{3n}, \tag{2.15}$$

we claim that

$$\mathbb{P}\{|\eta_S - \overline{\eta}_S - (\eta_{\Omega_{\mathcal{S}}} - \overline{\eta}_{\Omega_{\mathcal{S}}})| \leq e_n(S) \,:\, S \in \mathcal{S}\} \geq 1 - Cn^{-r}, \tag{2.16}$$

where C is an absolute constant depending on r. In order to prove this, we may assume that $\varepsilon_n \leq 1$, and in particular that $m \leq n$, since otherwise the result is trivial. For any $S \in \mathcal{S}$, application of Bernstein's inequality to the random variable $y\chi_{\Omega_{\mathcal{S}}}(x) - y\chi_S(x)$ gives

$$\begin{aligned}
\mathbb{P}\{|\eta_S - \overline{\eta}_S - (\eta_{\Omega_{\mathcal{S}}} - \overline{\eta}_{\Omega_{\mathcal{S}}})| > e_n(S)\} &\leq 2\exp\Big\{-\frac{ne_n(S)^2}{2\rho_{S\Delta\Omega_{\mathcal{S}}} + 2e_n(S)/3}\Big\} \\
&\leq 2\exp\Big\{-\frac{3n\rho_{S\Delta\Omega_{\mathcal{S}}}\varepsilon_n}{8}\Big\} \\
&\leq 2\exp\Big\{-(r+1)m\rho_{S\Delta\Omega_{\mathcal{S}}}(1+\log n)\Big\}.
\end{aligned}$$

Now, for any $S \in \mathcal{S}$, we have $\rho_{S\Delta\Omega_{\mathcal{S}}} = \frac{k}{m}$ for some integer $1 \leq k \leq m$ and the number of such sets $S \in \mathcal{S}$ is at most $\binom{m}{k}$, which itself is bounded by $(\frac{em}{k})^k$. Therefore, a union bound shows that $|\eta_S - \overline{\eta}_S - (\eta_{\Omega_{\mathcal{S}}} - \overline{\eta}_{\Omega_{\mathcal{S}}})| \leq e_n(S)$ with probability at least $1 - 2\sum_{k=1}^m (en)^{-(r+1)k}\Big(\frac{em}{k}\Big)^k \geq 1 - Cn^{-r}$.

REMARK 2.6. *Theorem 2.3 can be applied to any classification method that is based on an estimation $\hat{\eta}_S$ of η_S, once the bounds for $|\eta_S - \eta_{\Omega_{\mathcal{S}}} - (\hat{\eta}_S - \hat{\eta}_{\Omega_{\mathcal{S}}})|$ in terms of $e_n(S)$ have been established for all $S \in \mathcal{S}$. This determines $\omega(\rho, e_n)$ and thereby gives a bound for the estimation error. In particular we show in [3] that risk bounds for plug-in estimators can be obtained from this general approach and compare them with existing results as well as with the set estimator results derived in this paper.*

REMARK 2.7. *The usual approach to obtaining bounds on the performance of classifiers is to assume at the outset that the underlying measure ρ satisfies a margin condition. Our approach is motivated by the desire to obtain bounds with no assumptions on ρ. This is accomplished by introducing the modulus ω. As we discuss in the following section, a margin assumption allows one to obtain an improved bound on ω and thereby recover existing results in the literature. Another point about our result is that we do not assume that the Bayes classifier Ω^* lies in $\mathcal{S}$. In some approaches, as discussed in the survey [7], one first bounds the estimation error under this assumption, and then later removes this assumption with additional arguments that employ margin conditions.*

3. Margin conditions. The modulus ω introduced in the previous section is not transparent and, of course, depends on the set function $e_n(S)$. However, as we now show, for the types of e_n that naturally occur, the modulus is intimately connected with margin conditions. Margin assumptions are one of the primary ingredients in obtaining estimates on the performance of empirical classifiers. The margin condition (1.4) recalled in the introduction has the following equivalent formulation: for any measurable set S, we have

$$\rho_S \le C_\gamma \Big(\int_S |\eta| \Big)^\gamma, \quad \gamma := \frac{\alpha}{1+\alpha} \in [0,1]. \tag{3.1}$$

for some constant $C_\gamma > 0$ and $\gamma \in [0,1]$. This condition is void when $\gamma = 0$ and becomes more stringent as γ tends to 1. The case $\gamma = 1$ gives Massart condition.

In going further, we define $\mathcal{M}^\alpha$ as the set of all measures ρ such that ρ_X satisfies (1.4) or equivalently (3.1) and we define

$$|\rho|_{\mathcal{M}^\alpha} := \sup_{0<t\le 1} t^{-\alpha} \rho_X\{x \in X : \ |\eta(x)| \le t\}. \tag{3.2}$$

We want to bring out the connection between the modulus ω and the condition (3.1). In the definition of ω and its application to bounds on the estimation error, we assume that, we have an empirical estimator for which (2.4) holds with probability $1-\delta$. Notice that this is only assumed to hold for sets $S \in \mathcal{S}$ which is a distinction with (3.1). We shall make our comparison first when e_n is of the form $e_n(S) = \sqrt{\varepsilon_n \rho_S} + \varepsilon_n$ as appears for set estimators in Theorem 2.1.

We introduce the function

$$\phi(\rho, t) := \sup_{\int_S |\eta| \le 3(t+\sqrt{t\rho_S})} \int_S |\eta|, \quad 0 < t \le 1, \tag{3.3}$$

where now in this definition we allow arbitrary measurable sets S (not necessarily from $\mathcal{S}$). Under our assumption on the form of e_n, we have $\omega(\rho, e_n) \le \phi(\rho, \varepsilon_n)$ and so the decay of ϕ gives us a bound on the decay of ω. We say that ρ satsifies the ϕ-condition of order $s > 0$ if

$$\phi(\rho, t) \le C_0 t^s, \quad 0 < t \le 1. \tag{3.4}$$

for some constants C_0 and $s > 0$.

LEMMA 3.1. *Suppose that ρ is a measure that satisfies (1.4) for a given value of $0 \le \alpha \le \infty$. Then ρ satisfies the ϕ-condition (3.4) for $s = \frac{1+\alpha}{2+\alpha}$ with C_0 depending only on C_α and α. Conversely, if ρ satisfies the ϕ-condition with $s = \frac{1+\alpha}{2+\alpha}$ and a constant $C_0 > 0$, then it satisfies (1.4) for the value α with the constant C_α depending only on s and C_0.*

Proof: Suppose that ρ satisfies (1.4) for α and constant C_α, which equivalently means that it satisfies (3.1) for $\gamma := \frac{\alpha}{1+\alpha}$ and a constant C_γ. To check that the ϕ-condition is satisfied for $s = \frac{1+\alpha}{2+\alpha} = \frac{1}{2-\gamma}$, we let $t \in (0, 1]$ be fixed and let S be such that $\int_S |\eta| \le 3(\sqrt{t\rho_S} + t)$. From (3.1),

$$\rho_S \le C_\gamma \Big(\int_S |\eta| \Big)^\gamma \le C_\gamma 3^\gamma (\sqrt{t\rho_S} + t)^\gamma. \tag{3.5}$$

From this, one easily derives

$$\rho_S \le M t^{\frac{\gamma}{2-\gamma}}, \tag{3.6}$$

with a constant M depending only on C_γ and γ. To see this, suppose to the contrary that for some (arbitrarily large) constant M

$$\rho_S > M t^{\frac{\gamma}{2-\gamma}}. \tag{3.7}$$

Rewriting (3.5) as

$$\rho_S^{\frac{2-\gamma}{2\gamma}} \le C_\gamma^{1/\gamma} 3 \big(t^{1/2} + t \rho_S^{-1/2} \big),$$

and using (3.7) to estimate ρ_S on both sides from below, we obtain

$$M^{\frac{2-\gamma}{2\gamma}} t^{1/2} \le C_\gamma^{1/\gamma} 3 \big(t^{1/2} + M^{-1/2} t^{\frac{4-3\gamma}{4-2\gamma}} \big).$$

Since $0 < \gamma \le 1$, we have $\frac{4-3\gamma}{4-2\gamma} \ge \frac{1}{2}$, which yields

$$t^{1/2} \le M^{-\frac{2-\gamma}{2\gamma}} C_\gamma^{1/\gamma} 3 \big(1 + M^{-1/2} \big) t^{1/2}.$$

When M is chosen large enough, we have $M^{-\frac{2-\gamma}{2\gamma}} C_\gamma^{1/\gamma} 3\big(1 + M^{-1/2}\big) < 1$ which is a contradiction thereby proving (3.6).

It follows from (3.5) and (3.6) that

$$\int_S |\eta| \le 3(t + \sqrt{t\rho_S}) \le 3(t + Mt^{\frac{1}{2-\gamma}}) \le C_0 t^{\frac{1}{2-\gamma}}, \tag{3.8}$$

where C_0 depends on C_γ and γ. Taking now a supremum over all such sets S gives

$$\phi(\rho, t) \le C_0 t^s, \quad s = \frac{1}{2-\gamma}, \tag{3.9}$$

which is the desired inequality.

We now prove the converse. Suppose that ρ satisfies the ϕ-condition of order $s = \frac{1+\alpha}{2+\alpha}$ with constant C_0. We want to show that

$$\rho_X\{x : |\eta(x)| \le h\} \le C_\alpha h^\alpha, \quad 0 \le h \le 1, \tag{3.10}$$

with C_α depending only on s and C_0. As we noted before, this is equivalent to condition (3.1) of order $\gamma = \frac{\alpha}{\alpha+1}$. To prove (3.10), it is enough to prove

$$\rho_X\{x : h/2 \le |\eta(x)| \le h\} \le C'_\alpha h^\alpha, \quad 0 < h \le 1, \tag{3.11}$$

since then (3.10) follows easily by a summation argument. We fix h and define $S := \{x : h/2 \le |\eta(x)| \le h\}$ and $t := h^2\rho_S \in (0, 1]$. Then, we have

$$\int_S |\eta| \le h\rho_S = \sqrt{t\rho_S}. \tag{3.12}$$

This means that S is an admissible set in the definition of $\phi(\rho, t)$ in (3.3). Hence from the ϕ-condition (3.4), we know

$$h\rho_S/2 \le \int_S |\eta| \ \le \phi(\rho, t) \le C_0 t^s = C_0(h^2\rho_S)^s. \tag{3.13}$$

In other words, we have

$$\rho_S \le (2C_0)^{\frac{1}{1-s}} h^{\frac{2s-1}{1-s}} = (2C_0)^{\frac{1}{1-s}} h^\alpha, \tag{3.14}$$

which completes the proof. □

REMARK 3.2. *The purpose of this section is to show the connection of the modulus $\omega(\rho, e_n)$ with the existing and well studied margin conditions. However, the estimates for performance given in (2.14) can be applied without any specific assumption such as a margin condition, which corresponds to the case $\gamma = 0$. One could also examine other types of bounds for $\phi(\rho, t)$ than the power bound (3.4) and obtain similar results.*

4. Bounds for the approximation error $a(\Omega^*, \mathcal{S})$. The approximation error $a(\Omega^*, \mathcal{S})$ depends on ρ and the richness of the collection $\mathcal{S}$. A typical setting starts with a nested sequence $(\mathcal{S}_m)_{m\geq 1}$ of families of sets, that is such that $\mathcal{S}_m \subset \mathcal{S}_{m+1}$ for all $m \geq 1$. The particular value of m and the collection $\mathcal{S}_m$ that is used for a given draw of the data depends on n and properties of ρ (such as the smoothness of η and margin conditions) and is usually chosen through some form of model selection as discussed further. In order to analyze the performance of such classification algorithms, we would like to know conditions on ρ that govern the behavior of the approximation error as $m \to \infty$. We study results of this type in this section.

The error

$$a_m(\rho) := a(\Omega^*, \mathcal{S}_m), \quad m \geq 1, \tag{4.1}$$

is monotonically decreasing. We define the approximation class $\mathcal{A}^s = \mathcal{A}^s((\mathcal{S}_m)_{m\geq 1})$ as the set of all ρ for which

$$|\rho|_{\mathcal{A}^s} := \sup_{m\geq 1} m^s a_m(\rho) \tag{4.2}$$

is finite. Our goal is to understand what properties of ρ guarantee membership in $\mathcal{A}^s$. In this section, we give sufficient conditions for ρ to be in an approximation classes $\mathcal{A}^s$ for both set estimators and plug in estimators. These conditions involve the smoothness (or approximability) of η and margin conditions.

Given a measure ρ, it determines the regression function η and the Bayes set $\Omega^* := \{x : \eta(x) > 0\}$. We fix such a ρ and for each $t \in \mathbb{R}$, we define the level set $\Omega(t) := \{x : \ \eta(x) \geq t\}$. Notice that $\Omega(t) \subset \Omega(t')$ if $t \geq t'$. Also,

$$\{x : \ |\eta(x)| < t\} \subset \Omega(-t) \setminus \Omega(t) \subset \{x : \ |\eta(x)| \leq t\}. \tag{4.3}$$

For each $m = 1, 2, \ldots,$ we define

$$t_m := t_m(\rho, \mathcal{S}_m) := \inf\{t > 0 : \ \exists S \in \mathcal{S}_m \text{ s.t. } \Omega(t) \subset S \subset \Omega(-t)\}. \tag{4.4}$$

For convenience, we assume that there is always an $S_m^* \in \mathcal{S}_m$ such that $\Omega(t_m) \subset S_m^* \subset \Omega(-t_m)$. (If no such set exists then one replaces t_m by $t_m + \varepsilon$ with $\varepsilon > 0$ arbitrarily small and arrives at the same conclusion (4.6) given below). It follows that

$$\Omega^* \triangle S_m^* \subset \Omega(-t_m) \setminus \Omega(t_m). \tag{4.5}$$

If ρ satisfies the margin condition (1.4), then

$$a_m(\rho) \leq \int\limits_{\Omega^* \triangle S_m^*} |\eta| \, d\rho_X \leq C_\alpha \, t_m \cdot t_m^\alpha = C_\alpha \, t_m^{\alpha+1}. \tag{4.6}$$

Thus, a sufficient condition for ρ to be in $\mathcal{A}^s$ is that $t_m^{\alpha+1} \leq Cm^{-s}$.

The following example illustrates how the margin condition (1.4) combined with Hölder smoothness of the regression function implies that ρ belongs to the approximation class $\mathcal{A}^s$ where s depends on the margin and smoothness parameters. To be specific, let $X = [0,1]^d$. Let $\mathcal{D}$ be the collection of dyadic cubes Q contained in X, i.e., cubes $Q \subset X$ of the form $Q = 2^{-j}(k + [0,1]^d)$ with $k \in \mathbb{Z}^d$ and $j \in \mathbb{Z}$. Let $\mathcal{D}_j$, $j = 0, 1, \ldots$, be the collection of dyadic cubes of sidelength 2^{-j}. Let $\mathcal{S}_{2^{dj}}$ be the collection of all sets of the form $S_\Lambda = \cup_{Q\in\Lambda} Q$, where $\Lambda \subset \mathcal{D}_j$. This corresponds to the family $\mathcal{S}$ considered in Remark 2.5 for $m = 2^{jd}$. In fact, $\#(\mathcal{D}_j) = 2^{jd}$ and $\#(\mathcal{S}_{2^{dj}}) = 2^{2^{jd}}$. We complete the family $(\mathcal{S}_m)_{m\geq 1}$ by setting $\mathcal{S}_m = \mathcal{S}_{2^{dj}}$ when $2^{dj} \leq m < 2^{d(j+1)}$.

PROPOSITION 4.1. *We assume that ρ has the two following properties:* (i) *the regression function η is in the Lipschitz (or Hölder) space C^β for some $0 < \beta \leq 1$, that is*

$$|\eta|_{C^\beta} := \sup\{|\eta(x) - \eta(\tilde{x})|\,|x - \tilde{x}|^{-\beta} \ : \ x, \tilde{x} \in X\} < \infty\,;$$

(ii) *ρ satisfies the margin condition (1.4).*

Then one has

$$\rho \in \mathcal{A}^s = \mathcal{A}^s((\mathcal{S}_m)_{m\geq 1}) \quad \textit{with} \ \ s := \frac{\beta(\alpha+1)}{d}. \tag{4.7}$$

Proof: We claim that

$$a_{2^{dj}}(\rho) \leq (M2^{-j\beta})^{\alpha+1}, \quad j \geq 0, \tag{4.8}$$

with $M := 2^{-\beta} d^{\beta/2} |\eta|_{C^\beta}$. To this end, we first note that when $Q \in \mathcal{D}_j$, and ξ_Q is the center of Q, then

$$|\eta(x) - \eta(\xi_Q)| \leq M2^{-j\beta}. \tag{4.9}$$

We define $S_j \in \mathcal{S}_{2^{dj}}$ as the union of all $Q \in \mathcal{D}_j$ for which $\eta(\xi_Q) \geq 0$. If $t := M2^{-j\beta}$, then we claim that

$$\Omega(t) \subset S_j \subset \Omega(-t), \quad j \geq 0. \tag{4.10}$$

For example, if $x \in \Omega(t)$ then $\eta(x) \geq t$. So, if $x \in Q$, then $\eta(\xi_Q) \geq 0$ and hence $Q \subset S_j$. Similarly, if $x \in Q \subset S_j$ then $\eta(\xi_Q) \geq 0$ and hence $\eta(x) \geq -t$ for all $x \in Q$ and this implies the right containment in (4.10). □

It is well known that margin and smoothness conditions are coupled, in the sense that higher values of α force the regression function to have a sharper transition near the Bayes boundary, therefore putting restrictions on its smoothness. As an example, assume that ρ_X is bounded from below by the Lebesgue measure, i.e., there exists a constant $c > 0$ such that for any $S \in \mathcal{S}$

$$\rho_X(S) \geq c|S| = c \int_S dx.$$

In the most typical setting, the Bayes boundary $\partial\Omega^*$ is a $d-1$ dimensional surface of non-zero $\mathcal{H}^{d-1}$ Hausdorff measure. If $\eta \in C^\beta$ with $0 \leq \beta \leq 1$, then $|\eta(x)|$ is smaller than t at any point x which is at distance less than $|\eta|_{C^\beta}^{1/\beta} t^{1/\beta}$ from this boundary. It follows that

$$\rho_X\{x \in X \ : |\eta(x)| \leq t\} \geq c_0 t^{1/\beta},$$

where c_0 depends on $\mathcal{H}^{d-1}(\partial\Omega^*)$ and $|\eta|_{C^\beta}$, showing that $\alpha\beta \leq 1$. In such a case the approximation rate is therefore limited by $s \leq \frac{1+\beta}{d}$.

As observed in [2] one can break this constraint either by considering pathological examples, such as regression functions that satisfy $\mathcal{H}^{d-1}(\partial\Omega^*) = 0$, or by considering marginal measures ρ_X that vanish in the vicinity of the Bayes boundary. We show in §6 that this constraint can also be broken when the Hölder spaces C^β are replaced by the Besov spaces $B^\beta_\infty(L_p)$, defined by (1.8), that govern the approximation rate when $\mathcal{S}_{2^{dj}}$ is replaced by a collection of adaptive partitions.

5. Risk performance and model selection. In this section, we combine our previous bounds for approximation and estimation errors in order to obtain an estimate for risk performance of classification schemes.

Let us assume that we have a sequence $(\mathcal{S}_m)_{m\geq 1}$ of families $\mathcal{S}_m$ of sets that are used to develop a binary classification algorithm. We suppose that for some constant C_0,

$$VC(\mathcal{S}_m) \leq C_0 m, \quad m \geq 1, \tag{5.1}$$

and we denote by $\overline{\Omega}_m$ the empirical risk minimization classifier picked in $\mathcal{S}_m$ according to (2.1) with $\hat{\eta}_S = \overline{\eta}_S$. Theorem 2.1 gives that such an estimator provides a bound (2.6) with

$$e_n(S) = \sqrt{\rho_{S\Delta\Omega_S}\varepsilon_n} + \varepsilon_n, \quad \varepsilon_n = C\,\frac{m \log n}{n}$$

and C depending only on r and C_0. If $\rho \in \mathcal{A}^s((\mathcal{S}_m)_{m\geq 1})$, for some $s > 0$, then according to Corollary 2.4, for any $m \geq 1$, we have with probability $1 - n^{-r}$,

$$R(\overline{\Omega}_m) - R(\Omega^*) \leq \omega(\rho, e_n) + 2|\rho|_{\mathcal{A}^s} m^{-s}. \tag{5.2}$$

If in addition ρ satisfies the margin condition (1.4) of order $\alpha > 0$, then using Lemma 3.1 and the fact that $\omega(\rho, e_n) \leq C\phi(\rho, \varepsilon_n) \leq C\varepsilon_n^{\frac{1+\alpha}{2+\alpha}}$, we obtain

$$R(\overline{\Omega}_m) - R(\Omega^*) \leq C\Big(\frac{m \log n}{n}\Big)^{\frac{1+\alpha}{2+\alpha}} + 2|\rho|_{\mathcal{A}^s} m^{-s}, \tag{5.3}$$

where C depends on $|\rho|_{\mathcal{M}^\alpha}$. If we balance the two terms appearing on the right in (5.3) by taking $m = (\frac{n}{\log n})^{\frac{1+\alpha}{(2+\alpha)s+1+\alpha}}$, we obtain that with probability $1 - n^{-r}$

$$R(\overline{\Omega}_m) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)s}{(2+\alpha)s+1+\alpha}}, \tag{5.4}$$

where C depends on $|\rho|_{\mathcal{M}^\alpha}$ and $|\rho|_{A^s}$. The best rates that one can obtain from the above estimate correspond to $\alpha = \infty$ (Massart's condition) and $s \to \infty$ (the regression function η has arbitrarily high smoothness), and are limited by the so-called fast rate $\mathcal{O}(\frac{\log n}{n})$.

To obtain the bound (5.4), we need to know both s and α in order to make the optimal choice of m and $\mathcal{S}_m$. Of course, these values are not known to us and to circumvent this we employ a standard model selection technique based on an independant validation sample.

Model Selection: Let $(\mathcal{S}_m)_{m\geq 1}$ be any collection of set estimators. For notational convenience, we assume that n is even, i.e. $n = 2\overline{n}$. Given the draw $\mathbf{z}$, we divide $\mathbf{z}$ into two independent sets $\mathbf{z}'$ and $\mathbf{z}''$ of equal size $\overline{n}$.

Step 1: For each $1 \leq m \leq \overline{n}$, we let $\overline{\Omega}_m$ be defined by (2.1) with $\mathcal{S} = \mathcal{S}_m$ and $\mathbf{z}$ replaced by $\mathbf{z}'$.

Step 2: We now let $\overline{\mathcal{S}} := \{\overline{\Omega}_1, \ldots, \overline{\Omega}_{\overline{n}}\}$ and let $\hat{\Omega} := \overline{\Omega}_{m^*}$ be the set chosen from $\overline{\mathcal{S}}$ by (2.1) when using $\mathbf{z}''$.

The set $\hat{\Omega}$ is our empirical approximation of Ω^* obtained by this model selection procedure. To see how well it performs, let us now assume that

$\rho \in \mathcal{A}^s$ and that ρ also satisfies the margin condition (1.4) for α. In **Step 1**, we know that for each m, $\overline{\Omega}_m$ satisfies (5.3) with n replaced by $\overline{n}$ with probability at least $1 - \overline{n}^{-r}$. Thus, with probability $1 - cn^{-r+1}$, we have

$$R(\overline{\Omega}_m) - R(\Omega^*) \leq C\Big(m^{-s} + \Big(\frac{m \log n}{n}\Big)^{\frac{1+\alpha}{2+\alpha}}\Big), \quad m = 1, \dots, \overline{n}. \tag{5.5}$$

It follows that for $\overline{\mathcal{S}}$ of **Step 2**, we have

$$a(\Omega^*, \overline{\mathcal{S}}) = \min_{1 \leq m \leq \overline{n}} \int\limits_{\overline{\Omega}_m \Delta \Omega^*} |\eta| \, d\rho_X \leq C \min_{1 \leq m \leq \overline{n}} \left\{ m^{-s} + \Big(\frac{m \log n}{n}\Big)^{\frac{1+\alpha}{2+\alpha}} \right\}.$$

Since $\#(\overline{\mathcal{S}}) = \overline{n} = n/2$, we can take $\varepsilon_n \leq C\frac{\log n}{n}$ in Remark 2.2 and a suitable constant C when bounding performance on $\overline{\mathcal{S}}$. Hence, from Corollary 2.4, we have for the set $\hat{\Omega}$ given by **Step 2**,

$$\begin{aligned} R(\hat{\Omega}) - R(\Omega^*) &\leq 2a(\Omega^*, \overline{\mathcal{S}}) + C\Big(\frac{\log n}{n}\Big)^{\frac{1+\alpha}{2+\alpha}} \\ &\leq C \min_{1 \leq m \leq \overline{n}} \left\{ m^{-s} + \Big(\frac{m \log n}{n}\Big)^{\frac{1+\alpha}{2+\alpha}} \right\} + C\Big(\frac{\log n}{n}\Big)^{\frac{1+\alpha}{2+\alpha}}. \end{aligned}$$

In estimating the minimum, we choose m that balances the two terms and obtain

$$R(\hat{\Omega}) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)s}{(2+\alpha)s+1+\alpha}}. \tag{5.6}$$

Thus, the set $\hat{\Omega}$, while not knowing α and s gives the same estimate we obtained earlier when assuming we knew α and s.

REMARK 5.1. *Note that we have done our model selection without using a penalty term. The use of a penalty term would have forced us to know the value of α in (3.1). A discussion of why penalty approaches may still be of interest in practice can be found in* [5].

A simple application of (5.6) and Proposition 4.1 gives an estimate in the general case. In the case of Remark 2.5 one has $\omega(\rho, e_n) \leq C\varepsilon_n^{\frac{\alpha+1}{2}}$ and can balance the terms in the estimate corresponding to (5.6) by taking $m := \Big(\frac{n}{(\log n)^{1/(2+d)}}\Big)^{\frac{d}{2\beta+d}}$. These give the following result.

COROLLARY 5.2. *Let $(\mathcal{S}_m)_{m \geq 1}$ be the sequence of family of sets built from uniform partitions as are used in Proposition 4.1.*

(i) *Assume that ρ satisfies the margin condition (1.4) of order α and that η is Hölder continuous of order β. Then, the classifier resulting from the above model selection satisfies*

$$\text{Prob}\left\{R(\overline{\Omega}_{m^*}) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}}\right\} \geq 1 - Cn^{-r}, \tag{5.7}$$

where C depends on $r, |\rho|_{\mathcal{M}^\alpha}, |\eta|_{C^\beta}$.

(ii) *If one assumes in addition that ρ_X is equivalent to the Lebesgue measure, one obtains*

$$\text{Prob}\left\{R(\overline{\Omega}_{m^*}) - R(\Omega^*) \leq C\Big(\frac{(\log n)^{\frac{1}{2+d}}}{n}\Big)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta}}\right\} \geq 1 - Cn^{-r}. \tag{5.8}$$

Case (ii) illustrates the improvement of the rates that results from constraining the marginal ρ_X. In view of our earlier comments on the conflict between margin and Hölder smoothness conditions of high order, the main deficiency of both results is the underlying strong assumption of Hölder smoothness. The sequence $(\mathcal{S}_m)_{m\geq 1}$ is based on uniform partitions and does not allow us to exploit weaker Besov-smoothness conditions. In what follows, we remedy this defect by turning to classification algorithms based on *adaptive* partitioning. In doing so, we avoid any a priori constraints on ρ_X and hence use the set function e_n given by (2.5).

6. Classification using tree based adaptive partitioning. One of the most natural ways to try to capture Ω^* is through adaptive partitioning. Indeed, such partitioning methods have the flexibility to give fine scale approximation near the boundary of Ω^* but remain coarse away from the boundary. We now give two examples. The first is based on simple dyadic tree partitioning, while the second adds wedge ornation on the leaves of the tree to enhance risk performance. For simplicity of presentation, we only consider dyadic partitioning on the specific domain $X = [0,1]^d$, even though our analysis covers far greater generality.

Algorithm I: dyadic tree partitioning

We recall the dyadic cubes $\mathcal{D}$ introduced in §4. These cubes organize themselves into a tree with root X. Each $Q \in \mathcal{D}_j$ has 2^d children which are its dyadic subcubes from $\mathcal{D}_{j+1}$. A finite subtree $\mathcal{T}$ of $\mathcal{D}$ is a finite collection of cubes with the property that the root X is in $\mathcal{T}$ and whenever $Q \in \mathcal{T}$ its parent is also in $\mathcal{T}$. We say a tree is *complete* if, whenever Q is in $\mathcal{T}$, then all

of its siblings are also in $\mathcal{T}$. The set $\mathcal{L}(\mathcal{T})$ of leaves of such a tree $\mathcal{T}$ consists of all the cubes $Q \in \mathcal{T}$ such that no child of Q is in $\mathcal{T}$. The set of all such leaves of a complete tree forms a partition of X.

Any finite complete tree is the result of a finite number of successive cube refinements. We denote by $\mathfrak{T}_m$ the collection of all complete trees $\mathcal{T}$ that can be obtained using m refinements. Any such tree $\mathcal{T} \in \mathfrak{T}_m$ has $(2^d - 1)m + 1$ leaves. We can bound the number of trees in $\mathcal{T} \in \mathfrak{T}_m$ by assigning a bitstream that encodes, i.e. precisely determines, $\mathcal{T}$ as follows. Let $\mathcal{T} \in \mathfrak{T}_m$. We order the children of X lexicographically and assign a one to every child which is refined in $\mathcal{T}$ and a zero otherwise. We now consider the next generation of cubes (i.e. the grandchildren of X) in $\mathcal{T}$. We know these grandchildren from the bits already assigned. We arrange the grandchildren lexicographically and again assign them a one if they are refined in $\mathcal{T}$ and a zero otherwise. We continue in this way and receive a bitstream which exactly determines $\mathcal{T}$. Since $\mathcal{T}$, has exactly $2^d m + 1$ cubes, every such bitstream has length $2^d m$ and has a one in exactly $m - 1$ positions. Hence, we have

$$\#(\mathfrak{T}_m) \le \binom{2^d m}{m-1} \le \frac{(2^d m)^m}{(m-1)!} \le e^m 2^{dm}. \tag{6.1}$$

For each $\mathcal{T} \in \mathfrak{T}_m$ and any $\Lambda \subset \mathcal{L}(\mathcal{T})$, we define $S = S_\Lambda := \bigcup_{Q \in \Lambda} Q$. We denote by $\mathcal{S}_m$ the collection of all such sets S that can be obtained from a $\mathcal{T} \in \mathfrak{T}_m$ and some choice of Λ. Once $\mathcal{T}$ is chosen there are $2^{\#(\mathcal{L}(\mathcal{T}))} \le 2^{2^d m}$ choices for Λ. Hence

$$\#(\mathcal{S}_m) \le a^m \tag{6.2}$$

with $a := e2^{d+2^d}$.

Given our draw $\mathbf{z}$, we use the set estimator and model selection over $(\mathcal{S}_m)_{m \ge 1}$ as described in the previous section. We discuss the numerical implementation of this algorithm in §7. This results in a set $\overline{\Omega}(\mathbf{z})$ and we have the following theorem for its performance.

THEOREM 6.1. (i) *For any $r > 0$, there is a constant $c > 0$ such that the following holds. If $\rho \in \mathcal{A}^s$, $s > 0$, and ρ satisfies the margin condition (1.4), then with probability greater than $1 - cn^{-r+1}$, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)s}{(2+\alpha)s+1+\alpha}} \tag{6.3}$$

with C depending only on d, r, $|\rho|_{\mathcal{A}^s}$ and the constant in (1.4).
(ii) *If $\eta \in B^\beta_\infty(L_p(X))$ with $0 < \beta \le 1$ and $p > d/\beta$ and if ρ satisfies the*

margin condition (1.4)*, then with probability greater than* $1 - cn^{-r+1}$*, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}}, \tag{6.4}$$

with C *depending only on* d, r*,* $|\eta|_{B^\beta_\infty(L_p(X))}$ *and the constant in* (1.4)*.*

Proof: Since $\log(\#(\mathcal{S}_m)) \leq C_0 m$ where C_0 depends only on d, we have that $R(\Omega(\mathbf{z})) - R(\Omega^*)$ is bounded by the right side of (5.6) which proves (i). We can derive (ii) from (i) if we prove that the assumptions on ρ in (ii) imply that $\rho \in \mathcal{A}^s$, $s = \frac{(\alpha+1)\beta}{d}$. To see that this is the case, we consider the approximation of η by piecewise constants subordinate to partitions $\mathcal{L}(\mathcal{T})$, $\mathcal{T} \in \mathfrak{T}_m$. It is known (see [10]) that the Besov space assumption on η implies that there is a tree $\mathcal{T}_m$ and piecewise constant η_m on $\mathcal{L}(\mathcal{T}_m)$ that satisfies $\|\eta - \eta_m\|_{L_\infty} \leq \delta_m = C_1|\eta|_{B^\beta_\infty(L_p)} m^{-\beta/d}$ with C_1 depending on p, β, and d. Let $\Lambda := \{Q \in \mathcal{L}(\mathcal{T}_m) : \ \eta_m(x) > 0, \ x \in Q\}$ and $\Omega_m := \bigcup_{Q\in\Lambda_m} Q$. Then $\Omega_m \in \mathcal{S}_m$ and $\Omega_m \triangle \Omega^* \subset \{x : \ |\eta(x)| \leq \delta_m\}$ and so

$$a_m(\rho) \leq \int_{\Omega_m \triangle \Omega^*} |\eta| \, d\rho_X \leq C_\alpha \delta_m^{\alpha+1} \leq C_\alpha \Big(C_1|\eta|_{B^\beta_\infty(L_p)}\Big)^{\alpha+1} m^{-s}, \tag{6.5}$$

as desired. □

Algorithm II: higher order methods via decorated trees

We want to remove the restriction $\beta \leq 1$ that appears in Theorem 6.1 by enhancing the family of sets $\mathcal{S}_m$ of the previous section. This enhancement can be accomplished by choosing, for each $Q \in \mathcal{L}(\mathcal{T})$, a subcell of Q obtained by a hyperplane cut (henceforth called an *H-cell*) and then taking a union of such subcells. To describe this, we note that, given a dyadic cube Q, any $d-1$ dimensional hyperplane H partitions Q into at most two disjoint sets Q_0^H and Q_1^H which are the intersections of Q with the two open half spaces generated by the hyperplane cut. By convention we include $Q \cap H$ in Q_0^H. Given a tree $\mathcal{T} \in \mathfrak{T}_m$, we denote by $\zeta_\mathcal{T}$ any mapping that assigns to each $Q \in \mathcal{L}(\mathcal{T})$ an H-cell $\zeta_\mathcal{T}(Q)$. Given such a collection $\{\zeta_\mathcal{T}(Q)\}_{Q\in\mathcal{L}(\mathcal{T})}$, we define

$$S := S(\mathcal{T}, \zeta) := \bigcup_{Q\in\mathcal{L}(\mathcal{T})} \zeta_\mathcal{T}(Q).$$

For any given tree $\mathcal{T}$, we let $\mathcal{S}_\mathcal{T}$ be the collection of all such sets that result from arbitrary choices of ζ. For any $m \geq 1$, we define

$$\mathcal{S}_m := \bigcup_{\mathcal{T} \in \mathfrak{T}_m} \mathcal{S}_\mathcal{T}. \tag{6.6}$$

Thus, any such $S \in \mathcal{S}_m$ is the union of H-cells of the $Q \in \mathcal{L}(T)$, with one H-cell chosen for each $Q \in \mathcal{L}(\mathcal{T})$. Clearly $\mathcal{S}_m$ is infinite, however, the following lemma shows that $\mathcal{S}_m$ has finite VC dimension.

LEMMA 6.2. *If $\Gamma_1, \ldots, \Gamma_N$ are each collections of sets from X with VC dimension $\leq k$, then the collection $\Gamma := \bigcup_{i=1}^N \Gamma_i$ has VC dimension not greater than* $\max\{8 \log N, 4k\}$.

Proof: We follow the notation of Section 9.4 in [12]. Let us consider any set of points $p_1, \ldots p_L$ from X. Then, from Theorem 9.2 in [12], the shattering number of Γ for this set of point satisfies

$$s(\Gamma_j, \{p_1, \ldots, p_L\}) \leq \sum_{i=0}^{k} \binom{L}{i} =: \Phi(k, L)$$

and therefore

$$s(\Gamma, \{p_1, \ldots, p_L\}) \leq N\Phi(k, L).$$

By Hoeffding's inequality, if $k \leq L/2$ we have $2^{-L}\Phi(k, L) \leq \exp(-2L\delta^2)$ with $\delta := \frac{1}{2} - \frac{k}{L}$. It follows that if $L > \max\{8 \log N, 4k\}$, we have

$$s(\Gamma, \{p_1, \ldots, p_L\}) < 2^L N \exp(-L/8) < 2^L,$$

which shows that $\mathrm{VC}(\Gamma) \leq \max\{8 \log N, 4k\}$. □

We apply Lemma 6.2 with the role of the Γ_j being played by the collection $\mathcal{S}_\mathcal{T}$, $\mathcal{T} \in \mathfrak{T}_m$. As shown in (6.1), we have $N = \#(\mathfrak{T}_m) \leq e^m 2^{dm}$. We note next that the VC dimension of each $\mathcal{S}_\mathcal{T}$ is given by

$$\mathrm{VC}(\mathcal{S}_\mathcal{T}) = (d+1)\#(\mathcal{L}(\mathcal{T})) \leq (d+1)2^d m. \tag{6.7}$$

In fact, given $\mathcal{T}$ placing $d+1$ points in every $Q \in \mathcal{L}(\mathcal{T})$ shows that $(d+1)\#(\mathcal{L}(\mathcal{T}))$ points can be shattered since $d+1$ points can be shattered by hyperplanes in $\mathbb{R}^d$. No matter how one distributes more than $(d+1)\#(\mathcal{L}(\mathcal{T}))$ points in X, at least one $Q \in \mathcal{L}(\mathcal{T})$ contains more than $d+1$ points. These points can no longer be shattered by a hyperplane which confirms (6.7). Lemma 6.2 now says that

$$\mathrm{VC}(\mathcal{S}_m) \leq \max\{8(d+2)m, 4(d+1)2^d m\} = C_d m, \tag{6.8}$$

where $C_d := \max\{8(d+2), 4(d+1)2^d\}$.

Given our draw $\mathbf{z}$, we use the set estimator and model selection as described in §5 with $\mathcal{S}_m$ now given by (6.6). This results in a set $\overline{\Omega}(\mathbf{z})$ and we have the following theorem for the performance of this estimator.

THEOREM 6.3. (i) *For any $r > 0$, there is a constant $c > 0$ such that the following holds. If $\rho \in \mathcal{A}^s$, $s > 0$, and ρ satisfies the margin condition (1.4), then with probability greater than $1 - cn^{-r+1}$, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)s}{(2+\alpha)s+1+\alpha}} \tag{6.9}$$

with C depending only on d, r, $|\rho|_{\mathcal{A}^s}$ and the constant in (1.4).
(ii) *If $\eta \in B^\beta_\infty(L_p(X))$ with $0 < \beta \le 2$ and $p > d/\beta$ and if ρ satisfies the margin condition (3.1), then with probability greater than $1 - cn^{-r+1}$, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}}, \tag{6.10}$$

with C depending only on d, r, $|\eta|_{B^\beta_\infty(L_p(X))}$ and the constant in (1.4).

Proof: In view of (6.8) we can invoke Theorem 2.1 with $\varepsilon_n = Cm\log n/n$, where C depends on d and r, to conclude that $e_n(S) = \sqrt{\rho_{S\Delta\Omega_{\mathcal{S}_m}}\varepsilon_n} + \varepsilon_n$ satisfies (2.6) and hence is an admissible set function for the modulus (2.7). Now (i) follows from (5.6).

To derive (ii) from (i), we prove that the assumptions on ρ in (ii) imply that $\rho \in \mathcal{A}^s$, $s = \frac{(\alpha+1)\beta}{d}$, for $\beta \in (0, 2]$. To see that this is the case, we consider the approximation of η by piecewise *linear* functions subordinate to partitions $\mathcal{L}(\mathcal{T})$, $\mathcal{T} \in \mathfrak{T}_m$. It is known (see [9]) that the Besov space assumption on η implies that there is a tree $\mathcal{T}_m$ and a piecewise linear function η_m on $\mathcal{L}(\mathcal{T}_m)$ that satisfies $\|\eta - \eta_m\|_{L_\infty} \le \delta_m = C_1|\eta|_{B^\beta_\infty(L_p(X))}m^{-\beta/d}$. Now for any cube Q consider the H-cell mapping $\zeta_{\mathcal{T}}(Q) := \{x \in Q : \eta_m(x) \ge 0\}$. Then

$$\Omega_m := \bigcup_{Q\in\mathcal{L}(\mathcal{T})} \zeta_{\mathcal{T}}(Q)$$

is in $\mathcal{S}_m$ and $\Omega_m \Delta \Omega^* \subset \{x : \ |\eta(x)| \le \delta_m\}$ so that

$$a_m(\rho) \le \int_{\Omega_m\Delta\Omega^*} |\eta|\, d\rho_X \le C_\alpha\delta_m^{\alpha+1} \le C_\alpha\Big(C_1|\eta|_{B^\beta_\infty(L_p)}\Big)^{\alpha+1} m^{-s}, \tag{6.11}$$

as desired. □

REMARK 6.4. *It is in theory possible to further extend the range of β by considering more general decorated trees, where for each considered cube Q, we use an algebraic surface A of degree $k > 1$ instead of a hyperplane H that corresponds to the case $k = 1$. The resulting families $\mathcal{S}_m$ consist of level sets of piecewise polynomials of degree k on adaptive partitions obtained by m splits. From this one easily shows that the corresponding VC dimension is again controlled by m (with multiplicative constants now depending both on d and k) and that (6.10) now holds for all $0 < \beta \leq k + 1$. However, the practical implementation of such higher order classifiers appears to be difficult.*

We have seen in §5 that the approximation rate for non-adaptive partitioning is also given by $s = \frac{\beta(\alpha+1)}{d}$, but with β denoting the smoothness of η in the sense of the Hölder space C^β. The results established in this section show that the same approximation rate is obtained under the weaker constraint that $\eta \in B^\beta_\infty(L_p)$ with $p > d/\beta$ if we use adaptive partitioning.

We also observed in §5 that the Hölder smoothness β and the parameter α in the margin condition are coupled, for example by the restriction $\alpha\beta \leq 1$ when ρ_X is bounded from below by the Lebesgue measure. Replacing the Hölder space C^β by a Besov space $B^\beta_\infty(L_p)$ with $p > d/\beta$ allows us to relax the above constraint. As a simple example consider the case where ρ_X is the Lebesgue measure and

$$\eta(x) = \eta(x_1, \ldots, x_d) = \text{sign}(x_1 - 1/2)|x_1 - 1/2|^\delta,$$

for some $0 < \delta \leq 1$, so that $\Omega^* = \{x \in X \ : \ x_1 > 1/2\}$ and the margin condition (1.4) holds with α up to $1/\delta$. Then, one checks that $\eta \in B^\beta_\infty(L_p)$ for β and p such that $\beta \leq \delta + 1/p$. The constraint $1/p < \beta/d$ may then be rewritten as $\beta(1 - 1/d) < \delta$ or equivalently

$$\alpha\beta(1 - 1/d) < 1, \tag{6.12}$$

which is an improvement over $\alpha\beta \leq 1$.

7. Numerical Implementation. The results we have presented thus far on adaptive partitioning do not constitute a numerical algorithm since we have not discussed how one would find the sets $\overline{\Omega}_m \in \mathcal{S}_m$ required by (2.1) and used in the model selection. We discuss this issue next.

Given the draw $\mathbf{z}$, we consider the collection of all dyadic cubes in $\mathcal{D}_0 \cup \cdots \cup \mathcal{D}_{\overline{n}}$ with $\overline{n} = n/2$ which contain an x_i, $i = 1, \ldots, \overline{n}$. These cubes form a tree $\mathcal{T}'(\mathbf{z})$ which we call the *occupancy tree*. Adding to all such cubes their siblings, we obtain a complete tree $\mathcal{T}(\mathbf{z})$ whose leaves form a partition of X.

Let us first discuss the implementation of Algorithm I. For each complete subtree $\mathcal{T} \subset \mathcal{T}(\mathbf{z})$ we define

$$\gamma_{\mathcal{T}} := \sum_{Q \in \mathcal{L}(\mathcal{T})} \max(\overline{\eta}_Q, 0), \tag{7.1}$$

which we call the *energy* in $\mathcal{T}$. The set estimator $\overline{\Omega}_m$ corresponds to a complete tree $\overline{\mathcal{T}}_m \in \mathfrak{T}_m$ which maximizes the above energy. Note that several different trees may attain the maximum. Since only the values $m = 1, \ldots, \overline{n}$ are considered in the model selection procedure, and since there is no gain in subdividing a non-occupied cube, a maximizing tree is always a subtree of $\mathcal{T}(\mathbf{z})$.

Further, for each cube $Q \in \mathcal{T}(\mathbf{z})$, we denote by $\mathfrak{T}_m(Q)$ the collection of all complete trees $\mathcal{T}$ with root Q obtained using at most m subdivisions and being contained in $\mathcal{T}(\mathbf{z})$. We then define

$$\gamma_{Q,m} = \max_{\mathcal{T} \in \mathfrak{T}_m(Q)} \gamma_{\mathcal{T}}. \tag{7.2}$$

Again, this maximum may be attained by several trees in $\mathfrak{T}_m(Q)$. In fact, if for instance for a maximizer $\mathcal{T} \in \mathfrak{T}_m(Q)$, $\overline{\eta}_R > 0$ holds for all $R \in \mathcal{C}(R') \subset \mathcal{L}(\mathcal{T})$, the children of some parent node $R' \in \mathcal{T}$, then the subtree $\tilde{\mathcal{T}}$ of $\mathcal{T}$ obtained by removing $\mathcal{C}(R')$ from $\mathcal{T}$, has the same energy. We denote by $\mathcal{T}(Q, m)$ any tree in $\mathfrak{T}_m(Q)$ that attains the maximum $\gamma_{Q,m}$. By convention, we set

$$\mathcal{T}(Q, m) = \emptyset, \tag{7.3}$$

when Q is not occupied. With this notation, we define

$$\overline{\mathcal{T}}_m := \mathcal{T}(X, m) \quad \text{and} \quad \overline{\Omega}_m := \bigcup_{Q \in \mathcal{L}(\overline{\mathcal{T}}_m)} \{Q : \overline{\eta}_Q > 0\}, \tag{7.4}$$

to be used in the model selection discussed earlier.

We now describe how to implement the maximization that gives $\overline{\mathcal{T}}_m$ and therefore $\overline{\Omega}_m$. Notice that $\overline{\eta}_Q = \gamma_{Q,m} = 0$ and $\mathcal{T}(Q, m)$ is empty when Q is not occupied and therefore these values are available to us for free. Thus, the computational work in this implementation is solely determined by the occupied cubes that form $\mathcal{T}'(\mathbf{z})$. For $l = 0, \ldots, \overline{n}$, we define

$$\mathcal{U}_l := \mathcal{T}'(\mathbf{z}) \cap \mathcal{D}_{\overline{n}-l}, \tag{7.5}$$

the set of occupied cubes of resolution level $\overline{n}-l$. Notice that $\mathcal{U}_0 = \mathcal{L}(\mathcal{T}'(\mathbf{z}))$. We work from the leaves of $\mathcal{T}'(\mathbf{z})$ towards the root, in a manner similar to CART optimal pruning (see [8]), according to the following steps:

- $l = 0$: We compute for each $Q \in \mathcal{U}_0$ the quantities $\overline{\eta}_Q$ and define $\gamma_{Q,0} := \max\{0, \overline{\eta}_Q\}$, $\mathcal{T}(Q,0) := \{Q\}$. This requires at most $\overline{n}$ arithmetic operations.
- for $l = 1, \dots, \overline{n}$: Suppose we have already determined the quantities $\gamma_{Q,j}$ and $\overline{\eta}_Q$, as well as the trees $\mathcal{T}(Q,j)$, for all $Q \in \mathcal{U}_{l-1}$ and $0 \le j \le l-1$. Recall that $\mathcal{T}(Q,j)$ is a complete subtree. Now for all $0 \le j \le l$ and all cubes $Q \in \mathcal{U}_l$, we compute

(7.6)
$$(\ell_j^*(R))_{R\in\mathcal{C}'(Q)} := \operatorname{argmax}\Big\{ \sum_{R\in\mathcal{C}'(Q)} \gamma_{R,\ell'(R)} : \sum_{R\in\mathcal{C}'(Q)} \ell'(R) = j \Big\},$$

where $\mathcal{C}'(Q) := \mathcal{C}(Q) \cap \mathcal{T}'(\mathbf{z})$ denotes the set of occupied children of Q. Notice that the above argmax may not be unique, in which case we can pick any maximizer. We obviously have for each $Q \in \mathcal{U}_l$ and any $1 \le j \le l$,

(7.7)
$$\gamma_{Q,j} = \sum_{R\in\mathcal{C}'(Q)} \gamma_{R,\ell_{j-1}^*(R)},$$

with

$$\mathcal{T}(Q,j) = \{Q\} \cup \Big(\bigcup_{R\in\mathcal{C}'(Q)} \mathcal{T}(R, \ell_{j-1}^*(R)) \Big) \cup (\mathcal{C}(Q) \setminus \mathcal{C}'(Q)).$$

For $j = 0$, we compute the $\overline{\eta}_Q$ for all $Q \in \mathcal{U}_l$ by summing the $\overline{\eta}_R$ for $R \in \mathcal{C}'(Q)$ and define $\gamma_{Q,0} = \max\{0, \overline{\eta}_Q\}$ and $\mathcal{T}(Q,0) = \{Q\}$.

- At the final step $l = \overline{n}$, the set $\mathcal{U}_{\overline{n}}$ consists only of the root X and we have computed $\mathcal{T}(X,m)$ for $m = 0, \dots, \overline{n}$. This provides the estimators $\overline{\Omega}_m$ for $m = 0, \dots, \overline{n}$.

To estimate the complexity of the algorithm, we need to bound for each $l \in \{1, \dots, \overline{n}\}$ the number of computations required by (7.6) and (7.7). With proper organization, the argmax in (7.6) can be found using at most $\mathcal{O}(\#(C'(Q))l^2)$ operations. We can execute (7.7) with the same order of computation. The total complexity over all levels is therefore at most $\mathcal{O}(n^4)$ (a finer analysis can reduce it to $\mathcal{O}(n^3)$). Also each optimal tree $\mathcal{T}(Q,m)$ can be recorded with at most dm bits. It should be noted that the complexity with respect to the data size n is independent of the spatial dimension d which only enters when encoding the optimal trees $\mathcal{T}(X,m)$.

We turn now to the implementation of Algorithm II. We denote by $\mathcal{H}$ the set of all $d-1$ dimensional hyperplanes. Using the notations therein, for any subtree $\mathcal{T}$ of $\mathcal{T}(\mathbf{z})$ and any $Q \in \mathcal{L}(\mathcal{T})$, the energy is now defined as

(7.8)
$$\gamma_{\mathcal{T}} := \sum_{Q\in\mathcal{L}(\mathcal{T})} \max_{H\in\mathcal{H}, i=0,1} \max\{0, \overline{\eta}_{Q_i^H}\}.$$

The set estimator $\overline{\Omega}_m$ corresponds to a tree $\overline{\mathcal{T}}_m \in \mathfrak{T}_m$ which maximizes the above energy. Similar to the previous discussion, we define

$$\gamma_{Q,0} := \max_{H\in\mathcal{H},i=0,1} \max\{0, \overline{\eta}_{Q_i^H}\} \tag{7.9}$$

and define as before $\gamma_{Q,m}$ and $\mathcal{T}(Q,m)$ by (7.2) and (7.4).

The procedure of determining the trees $\mathcal{T}(X,m)$ for $m = 0,\dots,k$ is then, in principle, the same as above, however with a significant distinction due to the search for a "best" hyperplane $H = H_Q$ that attains the maximum in (7.9). Since a cube Q contains a finite number n_Q of data, the search can be reduced to $\binom{n_Q}{d}$ hyperplanes and the cost of computing $\gamma_{Q,0}$ is therefore bounded by n_Q^d. In addition the search of H_Q needs to be performed on *every* cube $Q \in \mathcal{T}(\mathbf{z})$, so that a crude global bound for this cost is given by n^{d+2}. This additional cost is affordable for small d but becomes prohibitive in high dimension. An alternate strategy is to to rely on more affordable classifiers to produce an affine (or even higher order algebraic) decision boundary on each Q. Examples are plug-in classifiers that are based on least-square estimation of η on Q by a polynomial.

APPENDIX A: PROOF OF THEOREM 2.1

For a given collection of sets $\mathcal{S}$ with VC-dimension $V = VC(\mathcal{S})$ let $f_S(x,y) := y(\chi_S(x) - \chi_{\Omega_\mathcal{S}}(x))$, $S \in \mathcal{S}$, where again $\Omega_\mathcal{S}$ is a best approximation to the Bayes set from $\mathcal{S}$. Since in these terms $\mathbb{E}(f_S) = \eta_S - \eta_{\Omega_\mathcal{S}}$, and $\frac{1}{n}\sum_{j=1}^n f_S(x_j,y_j) = \bar{\eta}_S - \bar{\eta}_{\Omega_\mathcal{S}}$, we need to estimate

$$\mathbb{P}\{\exists\, S \in \mathcal{S} : |\eta_S - \eta_{\Omega_\mathcal{S}} - (\bar{\eta}_S - \bar{\eta}_{\Omega_\mathcal{S}})| > e_n(S)\},$$

where

$$e_n(S) := e_n(S,r) := \sqrt{\rho_{S\Delta\Omega_\mathcal{S}}\varepsilon_n} + \varepsilon_n, \quad \varepsilon_n := \varepsilon_{n,r} := \frac{K\log n}{n}, \tag{A.1}$$

where $K := A\max\{r+1, V\}$. We want to show that the above probability is small provided A is chosen large enough.

We use the notation $\sigma_S^2 := \mathbb{E}_\rho(f_S^2)$, so that

$$\sigma_S^2 \le \rho_{S\Delta\Omega_\mathcal{S}}. \tag{A.2}$$

Rather than estimating the excess probability directly over all of $\mathcal{S}$ we first decompose the collection $\{f_S : S \in \mathcal{S}\}$ into the following slices. For any given $k = 1,\dots,n$, we define

$$\mathcal{S}_k := \{S \in \mathcal{S} : \varepsilon_n(k-1) \le \sigma_S^2 \le \varepsilon_n k\}, \tag{A.3}$$

Since $\varepsilon_n \geq \frac{1}{n}$, we have $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_n$. For later bounds (see (A.11) below) to remain well defined we remark that $\mathcal{S}_k = \emptyset$ for $k > n/(K \log n)$.

We now fix k and let

$$\mu := \varepsilon_n \sqrt{k}, \tag{A.4}$$

we observe that, by (A.2),

$$e_n(S) = \sqrt{\rho_{S\Delta\Omega_S}\varepsilon_n} + \varepsilon_n \geq \sigma_S\sqrt{\varepsilon_n} + \varepsilon_n \geq (\sqrt{(k-1)} + 1)\varepsilon_n \geq \mu, \quad S \in \mathcal{S}_k, \tag{A.5}$$

which yields

$$\mathbb{P}\left\{\exists\, S \in \mathcal{S}_k : |\eta_S - \eta_{\Omega_S} - (\bar\eta_S - \bar\eta_{\Omega_S})| > e_n(S)\right\} \leq \mathbb{P}\left\{\sup_{S\in\mathcal{S}_k} |\eta_S - \eta_{\Omega_S} - (\bar\eta_S - \bar\eta_{\Omega_S})| > \mu\right\}. \tag{A.6}$$

We define the random variable

$$Z(\mathbf{x}) := \sup_{S\in\mathcal{S}_k} \Big|\sum_{j=1}^{n} (f_S(x_j) - \mathbb{E}(f_S))\Big| = n \sup_{S\in\mathcal{S}_k} \big|\eta_S - \eta_{\Omega_S} - (\bar\eta_S - \bar\eta_{\Omega_S})\big|, \quad \mathbf{x} \in X^n, \tag{A.7}$$

and note that

$$\sup_{S\in\mathcal{S}_k} \sigma_S \leq \sqrt{\varepsilon_n k} = \sqrt{\frac{kK\log n}{n}} =: \sigma_k. \tag{A.8}$$

Since $\|f_S - \mathbb{E}(f_S)\|_{L_\infty} \leq 2$, Talagrand's inequality as stated in Theorem 1.3 of [5], adapted to the present situation, asserts that

$$\mathbb{P}\{|Z - \mathbb{E}(Z)| > t\} \leq C_0 \exp\Big\{ - c_0 t \log\Big(1 + \frac{2t}{nk\varepsilon_n + \mathbb{E}(Z)}\Big)\Big\}, \tag{A.9}$$

where c_0, C_0 are absolute constants. We next bound $\mathbb{E}(Z)$ by resorting to known bounds on expected sup-norms of empirical processes. Specifically, noting that

$$\|f_S\|_{L_\infty} \leq 1, \quad |f_S(x,y)| \leq \chi_{\Omega_S}(x) \leq 1, \quad \forall\ (x,y) \in X \times Y,\ S \in \mathcal{S}, \tag{A.10}$$

the bound from [13, (3.17),p. 46] yields

$$\mathbb{E}(Z) \leq n\, C_1 \max\left\{\sigma_k\sqrt{\frac{V\log(C_2\sigma_k^{-1})}{n}}, \frac{V\log(C_2\sigma_k^{-1})}{n}\right\}, \tag{A.11}$$

where C_1, C_2 are absolute constants. Observe that by (A.8) and (A.1), the first term on the right hand side of (A.11) exceeds the second one for each provided that $kK\log n/2V \geq \log\left(\frac{C_2^2 n}{kK\log n}\right)$, $k = 1, \ldots, \lceil n/(K\log n)\rceil$. We now set $K = C_3 V$ and observe that by choosing A large we can attain any value of c_3. So the first term of the max in (A.11) is attained by the first term for all relevant k whenever

$$C_3 \geq C_2^2/V. \tag{A.12}$$

Under this condition we infer from (A.11) that

$$\mathbb{E}(Z) \leq n\, C_1 \sqrt{k\varepsilon_n}\sqrt{\frac{V}{2n}\log\left(\frac{C_2^2 n}{kC_3 V\log n}\right)} =: nB_k \tag{A.13}$$

Therefore, returning to (A.9), we have for any $t \geq 2\mathbb{E}(Z)$

$$\begin{aligned} \mathbb{P}\{Z > t\} &\leq \mathbb{P}\{|Z - \mathbb{E}(Z)| > t/2\} \\ &\leq C_0 \exp\left\{-c_0 \frac{t}{2}\log\left(1 + \frac{t}{nk\varepsilon_n + nB_k.}\right)\right\}. \end{aligned} \tag{A.14}$$

Recalling (A.4) and taking $t = n\mu = n\varepsilon_n\sqrt{k}$, we observe that $t \geq 2\mathbb{E}(Z)$ holds, by (A.13), whenever $\varepsilon_n\sqrt{k} \geq 2B_k$. In view of (A.13) and the definition of ε_n, this is indeed the case for all $k \leq \lceil n/(K\log n)\rceil$ whenever

$$\varepsilon_n = \frac{C_3 V\log n}{n} \geq \frac{2C_1^2 V}{n}\log\left(\frac{C_2^2 n}{C_3 V\log n}\right) \tag{A.15}$$

holds. This is certainly true when in addition to (A.12)

$$C_3 \geq 2C_1^2 \tag{A.16}$$

holds. Thus, (A.14) takes the form

$$\begin{aligned} \mathbb{P}\{Z > n\mu\} &\leq \mathbb{P}\{|Z - \mathbb{E}(Z)| > n\mu/2\} \\ &\leq C_0 \exp\left\{-(c_0 n\mu/2)\log\left(1 + \frac{\mu}{k\varepsilon_n + B_k}\right)\right\}. \end{aligned} \tag{A.17}$$

Since, as noted earlier, $\varepsilon_n\sqrt{k} \geq 2B_k$, the second term of the sum appearing in the denominator of the logarithm is smaller than the first one. Therefore, recalling (A.7),

$$\begin{aligned} \mathbb{P}\Big\{\sup_{S\in\mathcal{S}_k}\left|\eta_S - \eta_{\Omega_S} - (\bar\eta_S - \bar\eta_{\Omega_S})\right| > \mu\Big\} &\leq C_0\exp\Big\{-(c_0 n\mu/2)\log\Big(1 + \frac{\mu}{2k\varepsilon_n}\Big)\Big\} \\ &\leq C_0\exp\Big\{-c_0\frac{n\mu^2}{4k\varepsilon_n}\Big\} \\ &\leq C_0\exp\Big\{-c_0\frac{n\varepsilon_n}{4}\Big\} \leq C_0 n^{-r-1}, \end{aligned} \tag{A.18}$$

provided that

$$C_3 \geq \frac{4(r+1)}{c_0}\frac{4(r+1)}{c_0 V}. \tag{A.19}$$

The second inequality in (A.19) is obviously true since $V \geq 1$ and the first is true if C_3 (respectively A) is large enough. As we have already noted, every $S \in \mathcal{S}$ is in one of the $\mathcal{S}_k$. Therefore, using (A.6) and a union bound over $1 \leq k \leq \lceil n/(K \log n)\rceil \leq n$, collection the stipulations from (A.12), (A.16), (A.19), we arrive at the statement of the theorem with $\varepsilon_n = K \log n/n$ provided that for $K = C_3 V$

$$K \geq \max\left\{\frac{4(r+1)}{c_0}, 2C_1^2 V, C_2^2\right\}, \tag{A.20}$$

where c_0, C_1, C_2 are the constants from (A.11) and (A.9). □

ACKNOWLEDGEMENTS

The authors wish to thank Stephane Gaiffas and László Györfi for various valuable suggestions and references, as well as the anonymous referees for their constructive comments.

REFERENCES

[1] N. Akakpo, *Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection*, Mathematical methods of statistics **28** (2012), 1-28.
[2] J.-Y. Audibert and A.N. Tsybakov, *Fast learning rates for plug-in classifiers*, Ann. Statistics **35** (2007), 608-633.
[3] P. Binev, A. Cohen, W. Dahmen and R. DeVore, *Analysis of risk bounds for set and plug-in estimators - application to adaptive partitioning classifiers*, submitted to proceedings of SAMPTA, 2014.
[4] P. Binev, A, Cohen, W. Dahmen, and R. DeVore, *Universal algorithms for learning theory, part II: piecewise polynomials*, Constructive Approximation, **26** (2007) 127-152.
[5] G. Blanchard and P. Massart. Discussion of V.Koltchinskii's 2004 IMS Medallion Lecture paper *Local Rademacher complexities and oracle inequalities in risk minimization*, Annals of Statistics **34** (2006) 2664-2671.
[6] G.Blanchard, C.Schfer, Y.Rozenholc, and K-R.Mller, *Optimal Dyadic Decision Trees*, Machine Learning, **66** (2007), 209-242.
[7] O. Bousquet, S. Boucheron and G. Lugosi, *Theory of Classification: a Survey of Some Recent Advances*, ESAIM: PS **9** (2005) 323-375.
[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees.*, Wadsworth, 1984.
[9] A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, *Tree-structured approximation and optimal encoding*, Appl. Comp. Harm. Anal. **11** (2001), 192-226.
[10] R. DeVore, *Nonlinear approximation*, Acta Numerica **7** (1998), 51-150.

[11] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics, Stochastic Modelling and Applied Probability, **31**, Springer-Verlag, 1996.

[12] L. Györfy, M. Kohler, A. Krzyzak, A. and H. Walk *A distribution-free theory of nonparametric regression*, Springer, Berlin, 2002.

[13] V. Kolchinskii, Oracle inequalities in empirical risk minimization and sparse recovery problems, Lecture Notes in Mathematics, **2033**, École d'Été de Probabilités de Saint Flour, Springer-Verlag, 2011.

[14] P. Massart and E. Nedelec, *Risk bounds for statistical learning*, Ann. Statistics, **34** (2006), 2326-2366.

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

[16] A. B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Annals of Statistics **32** (2004), 135-166.

Peter Binev
Department of Mathematics,
University of South Carolina,
Columbia, SC 29208, USA
E-mail: binev@math.sc.edu

Albert Cohen
UPMC Univ Paris 06, UMR 7598,
Laboratoire Jacques-Louis Lions,
F-75005, Paris, France
E-mail: cohen@ann.jussieu.fr

Wolfgang Dahmen
Institut für Geometrie und Praktische Mathematik,
RWTH Aachen,
Templergraben 55, D-52056, Aachen, Germany
E-mail: dahmen@igpm.rwth-aachen.de

Ronald DeVore
Department of Mathematics,
Texas A&M University,
College Station, TX 77840, USA
E-mail: rdevore@math.tamu.edu