Aachen Institute for Advanced Study in Computational Engineering Science

Preprint: AICES-2012/07-2

12/July/2012

Classification Algorithms using Adaptive Partitioning

P. Binev, A. Cohen, W. Dahmen, R. DeVore

RWTH AACHEN UNIVERSITY

AI CES

Financial support from the Deutsche Forschungsgemeinschaft (German Research Foundation) through grant GSC 111 is gratefully acknowledged.

List of AICES technical reports: http://www.aices.rwth-aachen.de/preprints

Classification Algorithms using Adaptive Partitioning

Peter Binev, Albert Cohen,
Wolfgang Dahmen, and Ronald DeVore *

July 12, 2012

Abstract

Algorithms for binary classification based on adaptive 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*. A general theory is developed to analyze the risk performance of set estimators with the goal of guaranteeing performance with high probability rather than in expectation. This analysis decouples the approximation and variance effects on the risk. Bounds are given for the variance in terms of VC dimension and margin conditions by introducing a new modulus and studying its relation to margin conditions. Bounds are given for the approximation term based on the smoothness of the regression function and margin conditions. When these approximation results are used with the variance bounds, an estimate of risk performance is obtained. A simple model selection is used to optimally balance the approximation and variance bounds. This general theory is then applied to the adaptive algorithms and results are formulated for the risk performance of these algorithms in terms of Besov smoothness of the regression function and margin conditions. The results of this paper are related to the work of Scott and Nowak [14] on tree based adaptive methods for classification, however with several important distinctions. In particular, our model selection utilizes a validation sample to avoid identifying suitable penalty terms. This allows us to employ wedge decorated trees that yield higher order performance.

Keywords: binary classification, adaptive methods, set estimators, tree based algorithms, analysis of risk
MSC numbers: 62M45, 65D05, 68Q32, 97N50.

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 families are based on some underlying form of approximation. In the case of set estimators, one

*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 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 Fondation "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).

directly approximates the *Bayes set*, using elements from a family $\mathcal{S}$ of sets. For plug-in estimators, one approximates the underlying regression function η, usually in a least squares sense, and builds the classifier as a level set of this approximation.

Generally speaking, it is not possible to compare the performance of different classifiers without some further knowledge or assumptions on the underlying probability measure from which the data is drawn. Each method has a particular class of probability measures (related to the underlying approximation process) on which it performs well. It could happen, however, that one method is based on a form of approximation that is always superior to the approximation method of the other. In this case, one would have a guarantee of better performance, provided there is a suitable control on the variance term. This is the case, for example, when nonlinear methods of approximation are used in place of linear methods. In classification based on set estimators, nonlinearity generally takes the form of some sort of adaptive partitioning. The purpose of this paper is to introduce classification algorithms using adaptive partitioning and to analyze the risk performance of these algorithms as well as their friendliness to numerical implementation.

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. We are given data $\mathbf{z} = (z_i)_{i=1}^n$, $z_i = (x_i, y_i)$, $i = 1, \ldots, n$, drawn independently according to ρ and we wish to construct a classifier based on this empirical data. Such 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.

The *risk* (probability of misclassification) of this classifier is defined as

$$R(\Omega) := \int_X \mathbb{P}\{T_\Omega(x) \neq y\} d\rho_X. \tag{1.2}$$

Note that for any measurable set $S \subset X$, we have

$$R(S) = \int_S (1 - p)\, d\rho_X + \int_{S^c} p\, d\rho_X = C - \int_S \eta\, d\rho_X, \tag{1.3}$$

with $C = \int_X p\, d\rho_X$.

A best classifier, i.e. one with minimal risk, is called a *Bayes classifier*. One choice is given by taking $\Omega = \Omega^* := \{x \,:\, \eta(x) \geq 0\}$. Its risk is

$$R(\Omega^*) = \int_X \min(p, 1 - p)\, d\rho_X. \tag{1.4}$$

Any other minimal risk set Ω differs from Ω^* only on sets of either measure zero or where η vanishes. In going further, we refer to T_{Ω^*} as *the Bayes classifier*, which is unknown to us. We seek to use

the data to build a classifier which is close to the Bayes classifier. The performance of any other classifier T_Ω is determined by the *excess risk*

$$R(\Omega) - R(\Omega^*) = \int\limits_{\Omega \Delta \Omega^*} |\eta| d\rho_X, \tag{1.5}$$

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

A natural way to empirically build a classifier is to consider a family $\mathcal{S}$ of sets and choose Ω as one of the sets from $\mathcal{S}$. Classification methods based on this strategy are called *set estimators*. Given such a family, we define $\Omega_\mathcal{S}$ to be the set from $\mathcal{S}$ that minimizes the risk over this family:

$$\Omega_\mathcal{S} := \operatorname*{argmin}_{S \in \mathcal{S}} R(S). \tag{1.6}$$

For any set S, we use the notation

$$\rho_S := \rho_X(S) = \int\limits_S d\rho_X \ \text{ and } \ \eta_S := \int\limits_S \eta \, d\rho_X. \tag{1.7}$$

Therefore, from (1.3), we also have

$$\Omega_\mathcal{S} = \operatorname*{argmax}_{S \in \mathcal{S}} \eta_S \ . \tag{1.8}$$

The set $\Omega_\mathcal{S}$ is also unknown to us, nevertheless it serves as a target for how well we can perform using the class $\mathcal{S}$. For any other $S \in \mathcal{S}$, we decompose the excess risk into

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

where both terms are nonnegative. The second term

$$a(\Omega^*, \mathcal{S}) := \inf_{\Omega \in \mathcal{S}} (R(\Omega) - R(\Omega^*)) = R(\Omega_\mathcal{S}) - R(\Omega^*) = \int\limits_{\Omega_\mathcal{S} \Delta \Omega^*} |\eta| d\rho_X, \tag{1.10}$$

is the error in approximating Ω^* by the sets in $\mathcal{S}$ and describes how well the family $\mathcal{S}$ can potentially approximate the Bayes classifier in excess risk. A classification algorithm uses the draw of the data $\mathbf{z}$ to find a set $\hat{\Omega} \in \mathcal{S}$ to be used as the empirical classifier. Since, the draw $\mathbf{z}$ gives us only limited information about η and ρ, we do not have $\hat{\Omega} = \Omega_\mathcal{S}$. The difference $R(\hat{\Omega}) - R(\Omega_\mathcal{S})$ appearing in (1.9) is now a random variable that depends on the draw, on the numerical method used to compute $\hat{\Omega}$, and also on the complexity or the size of $\mathcal{S}$. With some abuse of terminology, we call bounds for this random variable *variance estimates*. Such estimates are at the heart of classification theory and there are many papers written on this subject (see the survey [6] and the papers referenced therein).

A typical setting when building set classifiers is a nested sequence $(\mathcal{S}_m)_{m \geq 1}$ of families of sets, i.e. $\mathcal{S}_m \subset \mathcal{S}_{m+1}$ for each m. We use the family $\mathcal{S}_m$ for a certain value of m depending on the draw $\mathbf{z}$. The choice of m is made with the aim of balancing the two terms in (1.9) when $S = \hat{\Omega}$. The approximation term $a(\Omega^*, \mathcal{S}_m)$ decreases as m increases and the rate of decrease determines how effective this sequence is for building a classifier for ρ. On the other hand, bounds for $R(\hat{\Omega}) - R(\Omega_\mathcal{S})$ typically increase with m.

In view of (1.8), if $\hat{\eta}_S$ is any empirical estimator for η_S, a natural way to select a classifier within $\mathcal{S}$ is by

$$\hat{\Omega} := \hat{\Omega}_{\mathcal{S}} := \operatorname*{argmax}_{S \in \mathcal{S}} \hat{\eta}_S. \tag{1.11}$$

One of the most common strategies for building $\hat{\eta}_S$ is by introducing the empirical counterparts to (1.7),

$$\overline{\rho}_S := \frac{1}{n}\sum_{i=1}^{n} \chi_S(x_i) \ \text{ and } \ \overline{\eta}_S = \frac{1}{n}\sum_{i=1}^{n} y_i \chi_S(x_i). \tag{1.12}$$

The choice $\hat{\eta}_S = \overline{\eta}_S$ is equivalent to minimizing the empirical risk

$$\overline{R}(S) := \frac{1}{n}\#\{i : T_S(x_i) \neq y_i\}, \tag{1.13}$$

over the family $\mathcal{S}$ and therefore choose $\hat{\Omega}_{\mathcal{S}} = \overline{\Omega}_{\mathcal{S}}$ with

$$\overline{\Omega}_{\mathcal{S}} := \operatorname*{argmin}_{S \in \mathcal{S}} \overline{R}(S).$$

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.

Obtaining a concrete estimate of the decay of the excess risk as n grows requires assumptions on the underlying measure ρ. These 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 [13, 16]. We use a parameter α to delineate margin conditions and the parameter β to denote the smoothness assumption imposed on η. A common choice for (ii) is that η is in the Hölder class $\operatorname{Lip}\beta$ [1]. It is well-known in approximation theory that when using nonlinear methods, these assumptions can be weakened by considering smoothness in a certain scale of Besov spaces. This is important for us when we discuss classification methods built on adaptive partitioning since these are inherently nonlinear. The two assumptions (margin conditions and smoothness) have an intriguing interplay since they in some sense work against one another. One of the interests in using *Besov* in place of *Hölder* smoothness is to allow a more favorable trade-off between these assumptions, as we explain later.

The first part of this paper, from §2 to §6, gives an analysis of the risk performance of classifiers built according to (1.11). An important point is that we always seek results that hold with high probability rather than in expectation. Our typical bound in probability is of the form

$$\mathbb{P}\{R(\hat{\Omega}) - R(\Omega^*)) \geq C_0 n^{-\alpha}\} \leq C_1 n^{-r}, \tag{1.14}$$

from which one can obviously derive a bound in expectation of the form

$$\mathbb{E}(R(\hat{\Omega}) - R(\Omega^*)) \leq C_0 n^{-\alpha} + C_1 n^{-r}. \tag{1.15}$$

We begin in §2 by a derivation of uniform bounds for the approximation of η_S by the empirical estimator $\overline{\eta}_S$ that depend on the complexity of $\mathcal{S}$ either measured by its cardinality or its VC dimension. To provide variance estimates, we introduce in §3 a certain modulus, which is defined

on the available estimate between η_S and its estimator $\hat{\eta}_S$. We show in §4 how margin conditions can be used to estimate this modulus, and therefore the variance term.

While the analysis in §3 and §4 has many points in common with the existing literature, one of its specificities is that it can be applied to estimators $\hat{\eta}_S$ of η_S others than $\overline{\eta}_S$ and may therefore in principle be applied also to other types of classification algorithms than empirical risk minimization.

A general way to estimate the approximation term, based on the smoothness of η and the margin condition, is discussed in §5. The approximation estimate typically decays with the complexity parameter m, while the variance estimate grows. Given a model class $\mathcal{M} = \mathcal{M}(\alpha, \beta)$, the optimal balance between the approximation and variance terms requires a choice of m that typically depends on α and β. Such parameters being generally unknown, we propose in §6 a model selection procedure that allows us to simultaneously handle 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. However, in our view, the organization of the material in these sections help clarify various issues concerning the roles of approximation and variance estimates.

In §7, we propose numerical algorithms for classification based on adaptive partitioning, and analyze their performance using our previous results. Adaptive partitioning is a natural approach since it gives the flexibility of doing fine scale approximation near the decision boundary and coarse scale approximation away from this boundary. Our first algorithm builds set estimators using families $(\mathcal{S}_m)_{m\geq 1}$ of sets built on *tree based adaptive partitions*. In order to enhance approximation power, we develop a second adaptive partitioning algorithm based on *decorated trees*. Each cell corresponding to a leaf of the adaptive tree is now further subdivided using a hyperplane cut. Our results in this direction are motivated by [14]. However, the latter paper only considers non decorated trees (and therefore low order methods) and adaptive splitting rules based on specific penalty terms. Furthermore, the convergence analysis there assumes that the Bayes set Ω^* is a subgraph of a Hölder continuous function (horizon model). In contrast, our model selection procedure does not require the derivation of penalty terms and is applicable to more general decorated trees. Convergence rates can be obtained either under general approximability conditions or assumptions on the Besov smoothness of the regression function η. The numerical implementation and complexity of these algorithms are discussed in §8.

Adaptive partitioning classifiers can also be obtained through plug-in rules, using piecewise polynomials on adaptive partitions for the estimation of the regression function. The performance of this approach is studied in §9.

2 Empirical estimation of η_S

We begin by considering the particular estimator $\overline{\eta}_S$ of η_S given by (1.12). A critical issue for us is how well the empirical quantities $\overline{\rho}_S$ and $\overline{\eta}_S$ approximate the true values of ρ_S and η_S. This deviation can be controlled by Bernstein's inequality. Applying this inequality to the random variables $\chi_S(x)$ and $y\chi_S(x)$ respectively gives

$$\mathbb{P}\{|\rho_S - \overline{\rho}_S| > \delta\} \leq 2\exp\Big\{-\frac{n\delta^2}{2\rho_S + 2\delta/3}\Big\}, \tag{2.1}$$

and

$$\mathbb{P}\{|\eta_S - \overline{\eta}_S| > \delta\} \leq 2\exp\Big\{-\frac{n\delta^2}{2\rho_S + 2\delta/3}\Big\}. \tag{2.2}$$

Now suppose that $\mathcal{S}$ is any finite collection of sets of cardinality $\#\mathcal{S}$. Given a constant $r > 0$, we introduce the quantity

$$\varepsilon_n := \varepsilon_n(\mathcal{S}) := \frac{10(\log(\#\mathcal{S}) + r\log n)}{3n}. \tag{2.3}$$

Lemma 2.1 *Given any finite collection of sets $\mathcal{S}$ and ε_n as defined in (2.3), with probability at least $1 - 2n^{-r}$ on the draw $\mathbf{z}$, we have*

$$|\eta_S - \overline{\eta}_S| \le \sqrt{\rho_S \varepsilon_n} + \varepsilon_n, \quad \textit{for every} \quad S \in \mathcal{S}. \tag{2.4}$$

Proof: For any $S \in \mathcal{S}$, application of (2.2) gives

$$\mathbb{P}\{|\eta_S - \overline{\eta}_S| > \sqrt{\rho_S \varepsilon_n} + \varepsilon_n\} \le 2\exp\Big\{ - \frac{n(\sqrt{\rho_S \varepsilon_n} + \varepsilon_n)^2}{2\rho_S + 2(\sqrt{\rho_S \varepsilon_n} + \varepsilon_n)/3}\Big\}.$$

We next distinguish between two cases. If $\varepsilon_n \le \rho_S$ then the numerator in the exponential is at least $n\rho_S\varepsilon_n$ and the denominator is at most $10\rho_S/3$. If $\varepsilon_n > \rho_S$ then the numerator is at least $n\varepsilon_n^2$ and the denominator is at most $10\varepsilon_n/3$. Therefore in both case, we obtain

$$\mathbb{P}\{|\eta_S - \overline{\eta}_S| > \sqrt{\rho_S \varepsilon_n} + \varepsilon_n\} \le 2\exp\Big\{ - \frac{3n\varepsilon_n}{10}\Big\} \le 2(\#\mathcal{S})^{-1} n^{-r}.$$

Hence, this result also follows by a union bound. □

Since $R(S) - R(\Omega_{\mathcal{S}}) = \eta_{\Omega_{\mathcal{S}}} - \eta_S$, we also need estimates for how well we can empirically compute this quantity.

Lemma 2.2 *Given any finite collection of sets $\mathcal{S}$ and any $r > 0$, we define*

$$e_n(S) := \sqrt{\rho_{S\Delta\Omega_{\mathcal{S}}} \varepsilon_n} + \varepsilon_n, \quad \varepsilon_n := \frac{10(r\log n + \log(\#(\mathcal{S}))}{3n}. \tag{2.5}$$

Then, for all $S \in \mathcal{S}$, with probability at least $1 - 2n^{-r}$ on the draw $\mathbf{z}$, we have

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

Proof: The proof is similar to that of Lemma 2.1. We consider the random variable $y\chi_{\Omega_{\mathcal{S}}} - y\chi_S$, which has expectation $\eta_{\Omega_{\mathcal{S}}} - \eta_S$, sup norm less or equal to 1, and variance less or equal to $\rho_{S\Delta\Omega_{\mathcal{S}}}$. Thus, using Bernstein's inequality as in (2.2), we see that for any $\delta > 0$

$$\mathbb{P}\{|\eta_{\Omega_{\mathcal{S}}} - \eta_S - (\overline{\eta}_{\Omega_{\mathcal{S}}} - \overline{\eta}_S)| > \delta\} \le 2\exp\Big\{ - \frac{n\delta^2}{2\rho_{S\Delta\Omega_{\mathcal{S}}} + 2\delta/3}\Big\}. \tag{2.7}$$

Taking $\delta := \varepsilon_n(S)$, we conclude the proof as in Lemma 2.1. □

The assumption that $\mathcal{S}$ is finite in the above analysis is very strong and is not satisfied in many numerical methods of interest. However, as we now discuss, similar estimates hold in the case $\mathcal{S}$ is infinite but it has finite Vapnik-Chervonenkis dimension (VC dimension) $V_{\mathcal{S}}$. For example, it is known (see Theorem 3.4 of [6]) that

$$\mathbb{E}(\sup_{S\in\mathcal{S}} |\eta_S - \overline{\eta}_S|) \le 2\sqrt{\frac{2V_{\mathcal{S}}\log(n+1)}{n}}, \tag{2.8}$$

where the expectation is taken over all draws of size n. Here we instead search for estimates in probability which include a dependence on ρ_S, similar to Lemma 2.1. We begin with two lemmas about VC dimension.

Lemma 2.3 *If $\mathcal{S}$ is a collection of ρ_X measurable subsets of X with VC dimension V, then for any measurable set Ω, the collection of sets $\Lambda := \{S_\Delta\Omega : \ S \in \mathcal{S}\}$ has VC dimension at most $2V$.*

Proof: Suppose $\{x_1, \ldots, x_m\}$, $m > 2V$, is a set of points that is shattered by Λ. If $V+1$ of these points are not in Ω, then by relabeling, we can assume that $x_1, \ldots, x_{V+1}$ are not in Ω. For each $I = \{i_1, \ldots, i_j\}$, $1 \le i_1 < i_2 < \cdots < i_j \le V+1$, we know that there is a set $S_{I\Delta}\Omega$ from Λ which contains the x_j, $j \in I$, and does not contain the x_j, $j \notin I$. It follows that S_I contains all x_j, $j \in I$, and does not contain any x_j, $j \notin I \cap \{x_1, \ldots, x_{V+1}\}$. Hence, $\{x_1, \ldots, x_{V+1}\}$ is shattered by $\mathcal{S}$ which is a contradiction and proves that the assumption $m > 2V$ cannot hold in this case. On the other hand if $V+1$ of these points are in Ω, by again relabeling, we can assume $\{x_1, \ldots, x_{V+1}\}$ are all in Ω. For each $I = \{i_1, \ldots, i_j\}$, $1 \le i_1 < i_2 < \cdots < i_j \le V+1$, we know that there is a set $S_{I\Delta}\Omega$ from Λ which contains the x_j, $j \in I$, and does not contain the x_j, $j \notin I \cap \{x_1, \ldots, x_{V+1}\}$. Thus, the points x_j, $j \in \{1, \ldots, V+1\} \setminus I$ are all in S_I. Since any set $J \subset \{1, \ldots, V+1\}$ is the complement of an $I \subset \{1, \ldots, V+1\}$, we again get that $\{x_1, \ldots, x_{V+1}\}$ is shattered. This contradiction proves that there is no such set $\{x_1, \ldots, x_m\}$ with $m > 2V$ and confirms the assertion of the lemma. $\square$

The next lemma shows how well $\bar{\rho}_S$ approximates ρ_S for sets in a collection $\mathcal{S}$ with finite VC dimension.

Lemma 2.4 *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*

$$e_n(S) := e_n(S, r) := \sqrt{\rho_S \varepsilon_n} + \varepsilon_n, \quad \varepsilon_n := \varepsilon_{n,r} := A \max\{r+1, V\} \frac{\log n}{n}, \tag{2.9}$$

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

$$|\rho_S - \bar{\rho}_S| \le e_n(S), \quad S \in \mathcal{S}. \tag{2.10}$$

Proof: For any given $k = 1, \ldots, n$, let $\mathcal{S}_k$ be the collection of all sets $S \in \mathcal{S}$ for which $(k-1)\varepsilon_n < \rho_S \le k\varepsilon_n$. Note that since $\varepsilon_n \ge \frac{1}{n}$, we have $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_n$. We now fix $k \in \{1, \ldots, n\}$ and let $\mu := \sqrt{k}\varepsilon_n$. Observe that we have

$$e_n(S) = \sqrt{\rho_S \varepsilon_n} + \varepsilon_n \ge (\sqrt{k-1} + 1)\varepsilon_n \ge \mu, \tag{2.11}$$

and therefore

$$\mathbb{P}\left\{ \sup_{S \in \mathcal{S}_k} \left| \frac{1}{n} \sum_{i=1}^n \chi_S(x_i) - \rho_S \right| > e_n(S) \right\} \le \mathbb{P}\left\{ \sup_{S \in \mathcal{S}_k} \left| \frac{1}{n} \sum_{i=1}^n \chi_S(x_i) - \rho_S \right| > \mu \right\}. \tag{2.12}$$

We now apply Talagrand's concentration inequality in the form given in Theorem 1.3 of [2] applied to the set of functions $\mathcal{F} := \{\chi_S - \rho_S : \ S \in \mathcal{S}_k\}$. Each function $f \in \mathcal{F}$ has mean zero and

$\|f\|_{L_\infty} \le 1$. Considered as random variables over (X, ρ_X), they each have variance that does not exceed $\rho_S \le k\varepsilon_n$. If we define the random variables

$$Z(\mathbf{x}) := \sup_{S\in\mathcal{S}_k} \sum_{j=1}^n [\chi_S(x_j) - \rho_S], \quad \bar{Z}(\mathbf{x}) := \sup_{S\in\mathcal{S}_k} \Big|\sum_{j=1}^n [\chi_S(x_j) - \rho_S]\Big|, \quad \mathbf{x}\in X^n, \tag{2.13}$$

and their expectations $\mathbb{E}(Z), \mathbb{E}(\bar{Z})$, then according to the aforementioned Theorem 1.3, we have

$$\mathbb{P}\{|\bar{Z} - \mathbb{E}(\bar{Z})| > t\} \le C_0 \exp\Big\{ - c_0 t \log\Big(1 + \frac{t}{nk\varepsilon_n + \mathbb{E}(\bar{Z})}\Big)\Big\}, \tag{2.14}$$

where C_0, c_0 are absolute constants.

Next, we use an upper bound for $\mathbb{E}(\bar{Z})$ provided in Lemma 6.4 of [12]. To state this inequality, we use the related random variables

$$W^+(\mathbf{x}) := \frac{1}{n} Z(\mathbf{x}), \quad W^-(\mathbf{x}) := -\frac{1}{n} \inf_{S\in\mathcal{S}_k} \sum_{j=1}^n [\chi_S(x_j) - \rho_S].$$

Lemma 6.4 of [12] says that there is an absolute constant C_1 such that for

$$\sigma = C_1 \max\Big\{\sqrt{k\varepsilon_n}, \sqrt{\frac{V\log n}{n}}\ \Big\} = C_1\sqrt{k\varepsilon_n},$$

we have the bound

$$\mathbb{E}(W^\pm) \le C_1\sigma\sqrt{\frac{V\log n}{n}} = C_1^2\sqrt{\frac{k\varepsilon_n V\log n}{n}}. \tag{2.15}$$

Since $\bar{Z} \le n(W^+ + W^-)$, this gives the bound

$$\mathbb{E}(\bar{Z}) \le 2C_1^2 n\sqrt{\frac{k\varepsilon_n V\log n}{n}}. \tag{2.16}$$

Therefore, returning to (2.14), we have for any $t \ge 2\mathbb{E}(\bar{Z})$

$$\begin{aligned} \mathbb{P}\{\bar{Z} > t\}) &\le \mathbb{P}\{|\bar{Z} - E(\bar{Z})| > t/2\} \\ &\le C_0 \exp\left\{ -c_0 \frac{t}{2} \log\left(1 + \frac{t/2}{nk\varepsilon_n + 2C_1^2 n\sqrt{\frac{k\varepsilon_n V\log n}{n}}}\right)\right\}. \end{aligned} \tag{2.17}$$

We now take $t = n\mu = n\sqrt{k}\varepsilon_n$ and observe that $t \ge 2\mathbb{E}(\bar{Z})$ holds whenever

$$\sqrt{\varepsilon_n} \ge 4C_1^2\sqrt{\frac{V\log n}{n}}. \tag{2.18}$$

This is obviously true if the constant A in the definition of ϵ_n is larger than $16C_1^4$. With this stipulation on A, we can apply (2.17) and obtain

$$\begin{aligned} \mathbb{P}\{\sup_{S\in\mathcal{S}_k} [\hat{\rho}_S - \rho_S] > \mu\} &= \mathbb{P}\{\bar{Z} > n\mu\} \le \mathbb{P}\{\bar{Z} - \mathbb{E}(\bar{Z}\} > n\mu/2) \\ &\le C_0 \exp\left\{ -(c_0 n\mu/2) \log\left(1 + \frac{n\mu/2}{nk\varepsilon_n + 2C_1^2 n\sqrt{\frac{k\varepsilon_n V\log n}{n}}}\right)\right\}. \end{aligned} \tag{2.19}$$

The second term of the sum appearing in the denominator of the logarithm is smaller than the first because of (2.18). Therefore,

$$\begin{aligned}\mathbb{P}\Big\{ \sup_{S\in\mathcal{S}_k}[\hat{\rho}_S-\rho_S]>\mu\Big\} &\le C_0\exp\Big\{-(c_0n\mu/2)\log\Big(1+\frac{\mu}{4k\varepsilon_n}\Big)\Big\}\\ &\le C_0\exp\Big\{-c_0\frac{n\mu^2}{8k\varepsilon_n}\Big\}\\ &\le C_0\exp\Big\{-c_0\frac{n\varepsilon_n}{8}\Big\}\le C_0n^{-r-1}, \end{aligned} \tag{2.20}$$

provided A is chosen larger than $8/c_0$ which is another stipulation we impose on A.

As we have already noted, every $S\in\mathcal{S}$ is in one of the $\mathcal{S}_k$. Therefore, using (2.12) and a union bound over $1\le k\le n$, we arrive at (2.10). □

Theorem 2.5 *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*

$$e_n(S):=\sqrt{\rho_{S\Delta\Omega_{\mathcal{S}}}\varepsilon_n}+\varepsilon_n,\quad \varepsilon_n:=A\max\{r+1,V\}\frac{\log n}{n}, \tag{2.21}$$

where $r>0$ is arbitrary, then there is an absolute constant C_0 such that for any $n\ge 2$, with probability at least $1-C_0n^{-r}$ on the draw $\mathbf{z}\in Z^n$, we have

$$|\eta_S-\eta_{\Omega_{\mathcal{S}}}-(\bar{\eta}_S-\bar{\eta}_{\Omega_{\mathcal{S}}})|\le e_n(S),\quad S\in\mathcal{S}. \tag{2.22}$$

Proof: We know from Lemma 2.3 that the collection $\tilde{\mathcal{S}}:=\{S_\Delta\Omega_{\mathcal{S}}\,:\,S\in\mathcal{S}\}$ has VC dimension at most $2V$. We define $\tilde{\varepsilon}_n:=A\max\{r+1,2V\}\frac{\log n}{n}$ and the corresponding $\tilde{e}_n$, given by (2.5). We now apply Lemma 2.4 for $\tilde{\mathcal{S}}$ which gives that there is a set $E_0\subset X^n$ with $\rho^n(X)\le C_1n^{-\tilde{r}}$, such that for any draw $\mathbf{x}$ outside of E_0, we have

$$|\rho_{S\Delta\Omega_{\mathcal{S}}}-\bar{\rho}_{S\Delta\Omega_{\mathcal{S}}}|\le\tilde{e}_n(S)\le 2e_n(S),\ \ S\in\mathcal{S}. \tag{2.23}$$

Since $\mathcal{S}$ has VC dimension at most V, the set of functions $\mathcal{F}:=\{\chi_S:\ S\in\mathcal{S}\}$ satisfy $V_{\mathcal{F}^+}\le V$ where we follow the notation of [11]. In particular, $V_{\mathcal{F}^+}$ is the VC dimension of the set of epigraphs of $\mathcal{F}$. It follows from Lemma 9.2 and Theorem 9.4 of [11] that there is a cover $f_1,\dots,f_M$, $M\le C_2[n\log n]^V$ such that whenever $S\in\mathcal{S}$

$$\min_{1\le j\le M}\|\chi_S-f_j\|_{L_1(\rho_X)}\le 1/n\,. \tag{2.24}$$

It is easy to see that the f_j can each be chosen as $f_j=\chi_{S_j}$ with $S_j\in\mathcal{S}$ (perhaps at the expense of enlarging the constant C_2). Let $\Lambda:=\{S_1,\dots,S_M\}$. Hence for each $S\in\mathcal{S}$ there is an $S_j\in\Lambda$ such that

$$\rho_X(S_\Delta S_j)\le 1/n. \tag{2.25}$$

From Lemma 2.2 there is a set $E_1\subset Z^n$ with $\rho^n(E_1)\le 2n^{-\tilde{r}}$, such that for any draw $\mathbf{z}$ outside of E_1, we have

$$|\eta_{S_j}-\eta_{\Omega_{\mathcal{S}}}-(\bar{\eta}_{S_j}-\bar{\eta}_{\Omega_{\mathcal{S}}})|\le e_n(S_j),\quad j=1,\dots,M, \tag{2.26}$$

provided we choose the constant A larger than $\frac{20}{3}\Big(1+\frac{\log C_2+\log(\log n)}{\log n}\Big)$ for all n which is a stipulation we impose.

We now define the set $E \subset Z^n$ as the union of E_1 with the set of all points $\mathbf{z}$ whose $\mathbf{x}$ component is in E_0. Then $\rho^n(E) \leq (2 + C_1)n^{-\tilde{r}}$. In going further, we consider any draw $\mathbf{z}$ not in E and verify that (2.22) holds for such a draw. Given any $S \in \mathcal{S}$, we choose j so that (2.25) is valid. For this j, we have

$$\begin{aligned} |\eta_S - \eta_{\Omega_\mathcal{S}} - (\bar{\eta}_S - \bar{\eta}_{\Omega_\mathcal{S}})| &\leq |\eta_{S_j} - \eta_{\Omega_\mathcal{S}} - (\bar{\eta}_{S_j} - \bar{\eta}_{\Omega_\mathcal{S}})| + |\eta_S - \eta_{S_j}| + |\bar{\eta}_{S_j} - \bar{\eta}_S| \\ &\leq e_n(S_j) + \rho_X(S_j \triangle S) + |\bar{\eta}_{S_j} - \bar{\eta}_S| \\ &\leq 2e_n(S) + 1/n + |\bar{\eta}_{S_j} - \bar{\eta}_S| \\ &\leq 3e_n(S) + |\bar{\eta}_{S_j} - \bar{\eta}_S|. \end{aligned} \tag{2.27}$$

Here, we have used (2.26) in the second inequality, and in the third and last inequalities the fact that $\rho_{S_j \triangle \Omega_\mathcal{S}} \leq \rho_{S \triangle \Omega_\mathcal{S}} + 1/n$ and that $e_n(S) \geq \varepsilon_n \geq \frac{1}{n}$.

We are left with estimating the last term in (2.27). We have from (2.23) that

$$\begin{aligned} |\bar{\eta}_{S_j} - \bar{\eta}_S| &\leq \bar{\rho}_{S_j \triangle S} \\ &\leq [\bar{\rho}_{S_j \triangle S} - \rho_{S_j \triangle S}] + \rho_{S_j \triangle S} \\ &\leq 2e_n(S) + 1/n \leq 3e_n(S). \end{aligned} \tag{2.28}$$

When this estimate is inserted back into (2.27) and the constant A found so far is replaced by $36A$ we obtain the Theorem. □

3 A general estimate for variance in set estimators

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

$$|\eta_S - \eta_{\Omega_\mathcal{S}} - (\hat{\eta}_S - \hat{\eta}_{\Omega_\mathcal{S}})| \leq e_n(S), \tag{3.1}$$

for each set $S \in \mathcal{S}$. We have already proved such a bound for $\overline{\eta}_S$. We will also discuss similar bounds for plug-in estimators in §9. We develop our variance estimator assuming such a set valued function e_n.

To analyze the variance term in classifiers, we define the following modulus:

$$\omega(\rho, e_n) := \sup \left\{ \int\limits_{S \triangle \Omega_\mathcal{S}} |\eta| \ : \ S \in \mathcal{S} \ \text{ and } \int\limits_{S \triangle \Omega_\mathcal{S}} |\eta| \leq 3e_n(S) \right\}. \tag{3.2}$$

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 3.1 *Suppose that for each $S \in \mathcal{S}$, we have that* (3.1) *holds with probability $1 - \delta$. Then with this same probability, we have*

$$R(\hat{\Omega}_\mathcal{S}) - R(\Omega_\mathcal{S}) \leq \max\{\omega(\rho, e_n), a(\Omega^*, \mathcal{S})\}, \quad S \in \mathcal{S}, \tag{3.3}$$

with $a(\Omega^*, \mathcal{S})$ given by (1.10).

Proof: We consider any data $\mathbf{z}$ such that (3.1) holds and prove that (3.3) 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

$$R(\hat{\Omega}_{\mathcal{S}}) - R(\Omega_{\mathcal{S}}) = \eta_{\Omega_{\mathcal{S}}} - \eta_{\hat{\Omega}_{\mathcal{S}}} = \eta_{S_0} - \eta_{S_1}. \tag{3.4}$$

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 (3.1), we obtain

$$\eta_{S_0} - \eta_{S_1} = \eta_{\Omega_{\mathcal{S}}} - \eta_{\hat{\Omega}_{\mathcal{S}}} \leq e_n(\hat{\Omega}_{\mathcal{S}}). \tag{3.5}$$

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

$$\eta_{S_0} - \eta_{S_1} = A - B, \quad A := \eta_{S_0^+} - \eta_{S_1^-},\ B := \eta_{S_1^+} - \eta_{S_0^-}. \tag{3.6}$$

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

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

$$R(\widehat{\Omega}_{\mathcal{S}}) - R(\Omega_{\mathcal{S}}) = A - B \leq B \leq a(\Omega^*, \mathcal{S}), \tag{3.7}$$

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 (3.5) and (3.6),

$$\int\limits_{\hat{\Omega}_{\mathcal{S}} \Delta \Omega_{\mathcal{S}}} |\eta| = A + B \leq 3A/2 \leq 3(A - B) = 3(\eta_{S_0} - \eta_{S_1}) \leq 3e_n(\hat{\Omega}_{\mathcal{S}}). \tag{3.8}$$

This means that $\hat{\Omega}_{\mathcal{S}}$ is one of the sets appearing in the definition of $\omega(\rho, e_n)$ and (3.3) 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 3.1, we immediately obtain the following corollary which describes the performance of the set selection method.

Corollary 3.2 *Suppose that for each $S \in \mathcal{S}$,* (3.1) *holds with probability $1 - \delta$. Then with this same probability we have*

$$R(\hat{\Omega}_{\mathcal{S}}) - R(\Omega^*) \leq \omega(\rho, e_n) + 2a(\Omega^*, \mathcal{S}). \tag{3.9}$$

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 (3.3). □

Remark 3.3 *We close this section with some remarks on how our results compare with others in the literature.*

(i) *Theorem 3.1 can be applied to any classification method that is based on an estimation $\hat{\eta}_S$ of η_S, once the bounds for $|\eta_S - \eta_{\Omega_S} - (\hat{\eta}_S - \hat{\eta}_{\Omega_S})|$ 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 variance.*

(ii) *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.*

(iii) *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* [6]*, one first bounds variance under this assumption, and then later removes this assumption with additional arguments that employ margin conditions.*

4 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 following condition (sometimes referred to as the Tsybakov condition) requires that for any measurable set S, we have

$$\rho_S \le C_\rho \Big(\int_S |\eta| \Big)^\alpha \tag{4.1}$$

for some constant $C > 0$ and $\alpha \in [0, 1]$. This condition becomes more stringent as α tends to 1 and is known as the Massart condition when $\alpha = 1$. The Massart condition means that for some $A > 0$, we have $|\eta| > A$ almost everywhere. An equivalent form of (4.1) is that

$$\rho_X\{x \in X : \ |\eta(x)| \le t\} \le \overline{C}_\rho t^q, \quad q := \frac{\alpha}{1-\alpha}, \quad 0 < t \le 1. \tag{4.2}$$

In going further, we define $\mathcal{M}^\alpha$ as the set of all measures ρ such that ρ_X satisfies (4.1) or equivalently (4.2) and we define

$$|\rho|_{\mathcal{M}^\alpha} := \sup_{0<t\le 1} t^{-\frac{\alpha}{1-\alpha}} \rho_X\{x \in X : \ |\eta(x)| \le t\}. \tag{4.3}$$

We want to bring out the connection between the modulus ω and the condition (4.1). In the definition of ω and its application to bounds on the variance, we assume that, we have an empirical estimator for which (3.1) holds with probability $1 - \delta$. Notice that this is only assumed to hold for sets $S \in \mathcal{S}$ which is a distinction with (4.1). We shall make our comparison when e_n is of the form $e_n(S) = \sqrt{\varepsilon_n \rho_S} + \varepsilon_n$ as in the results of §2.

We introduce the function

$$\phi(\rho,t) := \sup_{\int_S |\eta| \le 3(t+\sqrt{t\rho_S})} \int_S |\eta|, \quad 0 < t \le 1, \tag{4.4}$$

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 $\omega(\rho,\varepsilon) \le \phi(\rho,\varepsilon_n)$ 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

$$\phi(\rho,t) \le C_0 t^s, \quad 0 < t \le 1. \tag{4.5}$$

for some constants C_0 and $s > 0$.

Lemma 4.1 *Suppose that ρ is a measure that satisfies* (4.1) *for a given value of $\alpha \in [0,1]$. Then ρ satisfies the ϕ-condition* (4.5) *for $s = \frac{1}{2-\alpha}$ with C_0 depending only on C_ρ and α. Conversely, if ρ satisfies the ϕ-condition with $s = \frac{1}{2-\alpha}$ and a constant $C_0 > 0$, then it satisfies* (4.1) *of order α with the constant C_ρ depending only on s and C_0.*

Proof: Suppose that ρ satisfies (4.1) for α and constant C_ρ. To check that the ϕ-condition is satisfied for $s = \frac{1}{2-\alpha}$, 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 (4.1),

$$\rho_S \le C_\rho \Big(\int_S |\eta| \Big)^\alpha \le C_\rho 3^\alpha (\sqrt{t\rho_S} + t)^\alpha. \tag{4.6}$$

From this, one easily derives

$$\rho_S \le M t^{\frac{\alpha}{2-\alpha}}, \tag{4.7}$$

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

$$\rho_S > M t^{\frac{\alpha}{2-\alpha}}. \tag{4.8}$$

Rewriting (4.6) as

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

and using (4.8) to estimate ρ_S on both sides from below, we obtain

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

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

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

When M is chosen large enough, we have $M^{-\frac{2-\alpha}{2\alpha}} C_\rho^{1/\alpha} 3(1 + M^{-1/2}) < 1$ which is a contradiction thereby proving (4.7).

It follows from (4.6) and (4.7) that

$$\int_S |\eta| \le 3(t + \sqrt{t\rho_S}) \le 3(t + M t^{\frac{1}{2-\alpha}}) \le C_0 t^{\frac{1}{2-\alpha}}, \tag{4.9}$$

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

$$\phi(\rho, t) \leq C_0 t^s, \quad s = \frac{1}{2-\alpha}, \tag{4.10}$$

which is the desired inequality.

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

$$\rho_X\{x : |\eta(x)| \leq y\} \leq \overline{C}_\rho y^{\frac{\alpha}{1-\alpha}}, \quad 0 \leq y \leq 1, \tag{4.11}$$

with $\overline{C}_\rho$ depending only on s and C_0. As we noted in (4.2), this is equivalent to condition (4.1) of order α. To prove (4.11), it is enough to prove

$$\rho_X\{x : y/2 \leq |\eta(x)| \leq y\} \leq \overline{C}'_\rho y^{\frac{\alpha}{1-\alpha}}, \quad 0 < y \leq 1, \tag{4.12}$$

since then (4.11) follows easily by a summation argument. We fix y and define $S := \{x : y/2 \leq |\eta(x)| \leq y\}$ and $t := y^2\rho_S \in (0, 1]$. Then, we have

$$\int_S |\eta| \leq y\rho_S = \sqrt{t\rho_S}. \tag{4.13}$$

This means that S is an admissible set in the definition of $\phi(\rho, t)$ in (4.4). Hence from the ϕ-condition (4.5), we know

$$y\rho_S/2 \leq \int_S |\eta| \ \leq \phi(\rho, t) \leq C_0 t^s = C_0(y^2\rho_S)^s. \tag{4.14}$$

In other words, we have

$$\rho_S \leq (2C_0)^{\frac{1}{1-s}} y^{\frac{2s-1}{1-s}} = (2C_0)^{\frac{1}{1-s}} y^{\frac{\alpha}{1-\alpha}}, \tag{4.15}$$

which completes the proof. □

5 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 $(\mathcal{S}_m)_{m=1}^\infty$ of families of sets: $\mathcal{S}_m \subset \mathcal{S}_{m+1}$, $m = 1, 2, \ldots$. 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

$$a_m(\rho) := a(\Omega^*, \mathcal{S}_m), \quad m = 1, 2, \ldots, \tag{5.1}$$

is monotonically decreasing and under very mild density assumptions tends to zero as $m \to \infty$. We define the approximation class $\mathcal{A}^s = \mathcal{A}^s((\mathcal{S}_m))$ as the set of all ρ for which

$$|\rho|_{\mathcal{A}^s} := \sup_{m \geq 1} m^s a_m(\rho) \tag{5.2}$$

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.

We suppose that we have a monotone sequence $(\mathcal{S}_m)_{m=1}^\infty$, where each $\mathcal{S}_m$ is a collections of sets. 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,

$$\{x : \ |\eta(x)| < t\} \subset \Omega(-t) \setminus \Omega(t) \subset \{x : \ |\eta(x)| \leq t\}. \tag{5.3}$$

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

$$t_m := t_m(\rho, \mathcal{S}_m) := \inf\{t > 0 : \text{ there exists } S \in \mathcal{S}_m \text{ such that } \Omega(t) \subset S \subset \Omega(-t)\}. \tag{5.4}$$

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 (5.6) given below). It follows that

$$\Omega^* \triangle S_m^* \subset \Omega(-t_m) \setminus \Omega(t_m). \tag{5.5}$$

If ρ satisfies the margin condition (4.2), then

$$a_m(\rho) \leq \int\limits_{\Omega^* \triangle S_m^*} |\eta| \, d\rho_X \leq \overline{C}_\rho \, t_m \cdot t_m^q = \overline{C}_\rho \, t_m^{q+1}. \tag{5.6}$$

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

We next give a simple example of how (5.6) can be utilized. Here $X = [0,1)^d$. 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 $\Lambda \subset \mathcal{D}_j$. Notice that $\#(\mathcal{D}_j) = 2^{jd}$ and $\#(\mathcal{S}_{2^{dj}}) = 2^{2^{jd}}$. We assume that ρ satisfies the two following properties:

- the regression function η is in the Lipschitz (or Hölder) space $\mathrm{Lip}\,\beta$ for some $0 < \beta \leq 1$, that is
$$|\eta|_{\mathrm{Lip}\,\beta} := \sup\{|\eta(x) - \eta(\tilde{x})| \, |x - \tilde{x}|^{-\beta} \ : \ x, \tilde{x} \in X\} < \infty\,;$$
- ρ satisfies the margin condition (4.2).

Then we claim that

$$a_{2^{dj}}(\rho) \leq (M 2^{-j\beta})^{q+1}, \quad j \geq 0, \tag{5.7}$$

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

$$|\eta(x) - \eta(\xi_Q)| \leq M 2^{-j\beta}. \tag{5.8}$$

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 $t := M 2^{-j\beta}$, then we claim that

$$\Omega(t) \subset S_j \subset \Omega(-t), \quad j \geq 0. \tag{5.9}$$

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 (5.9).

This example shows that the margin condition (4.2) combined with Hölder smoothness of order β for the regression function η, implies that ρ belongs to the approximation class $\mathcal{A}^s = \mathcal{A}^s((\mathcal{S}_{2^{dj}}))$ with $s := \frac{\beta(q+1)}{d}$.

It is well known that margin and smoothness conditions are coupled, in the sense that higher values of q 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 $\eta \in \operatorname{Lip} \beta$ with $0 \leq \beta \leq 1$, then $|\eta(x)|$ is smaller than t at any point x which is at distance less than $|\eta|^{1/\beta}_{\operatorname{Lip}\beta} t^{1/\beta}$ 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 $|\eta|_{\operatorname{Lip}\beta}$, showing that $\beta q \leq 1$. In such a case the approximation rate is therefore limited by $s \leq \frac{1+\beta}{d}$.

As observed in [1] 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 §7 that this constraint can also be broken when the Lipschitz spaces $\operatorname{Lip}\beta$ are replaced by certain Besov spaces $B^\beta_\infty(L_p)$ that govern the approximation rate when $\mathcal{S}_{2^{dj}}$ is replaced by a collection of adaptive partitions.

6 Risk performance and model selection

In this section, we combine our previous estimates for approximation error and variance bounds in order to obtain an estimate for risk performance of classification schemes.

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

$$VC(\mathcal{S}_m) \leq C_0 m, \ \ m \geq 1, \tag{6.1}$$

and we denote by $\overline{\Omega}_m$ the empirical risk minimization classifier picked in $\mathcal{S}_m$ according to (1.11) with $\hat{\eta}_S = \overline{\eta}_S$. We have shown in Theorem 2.5 that such an estimator provides a bound (2.22) 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. If $\rho \in \mathcal{A}^s((\mathcal{S}_m))$, for some $s > 0$, then according to Corollary 3.2, for any $m \geq 1$, we have with probability $1 - n^{-r}$,

$$R(\overline{\Omega}_m) - R(\Omega^*) \leq \omega(\rho, e_n) + 2|\rho|_{\mathcal{A}^s} m^{-s}. \tag{6.2}$$

If in addition ρ satisfies the margin condition of order $\alpha > 0$, then using Lemma 4.1 and the fact that $\omega(\rho, e_n) \leq C\phi(\rho, \varepsilon_n) \leq C\varepsilon_n^{\frac{1}{2-\alpha}}$, we obtain

$$R(\overline{\Omega}_m) - R(\Omega^*) \leq C\Big(\frac{m \log n}{n}\Big)^{\frac{1}{2-\alpha}} + 2|\rho|_{\mathcal{A}^s} m^{-s}, \tag{6.3}$$

where C depends on $|\rho|_{\mathcal{M}^\alpha}$. If we balance the two terms appearing on the right in (6.3) by taking $m = (\frac{n}{\log n})^{\frac{1}{(2-\alpha)s+1}}$, we obtain that with probability $1 - n^{-r}$

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

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 $\alpha = 1$ (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 (6.4), 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, as is usually done, we employ a form of model selection.

Let us assume that $\rho \in \mathcal{A}^s$ and that ρ also satisfies the margin condition (4.1) where both α and s are unknown to us. 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 $\overline{n}$. For each $1 \leq m \leq \overline{n}$, we let $\overline{\Omega}_m$ be defined by (1.11) with $\mathcal{S} = \mathcal{S}_m$ and $\mathbf{z}$ replaced by $\mathbf{z}'$. We know that for each m, $\overline{\Omega}_m$ satisfies (6.3) with n replaced by $\overline{n}$ with probability at least $1 - \overline{n}^{-r}$. Thus, with probability $1 - cn^{-r+1}$, we have

$$\int\limits_{\overline{\Omega}_m \Delta \Omega^*} |\eta| \, d\rho_X = R(\overline{\Omega}_m) - R(\Omega^*) \leq C\Big(m^{-s} + \Big(\frac{m \log n}{n}\Big)^{\frac{1}{2-\alpha}}\Big), \quad m = 1, \dots, \overline{n}. \tag{6.5}$$

We now let $\overline{\mathcal{S}} := \{\overline{\Omega}_1, \dots, \overline{\Omega}_{\overline{n}}\}$ and let $\overline{\Omega}_{m^*}$ be the set chosen from $\overline{\mathcal{S}}$ by (1.11) when using $\mathbf{z}''$. It follows from (6.5) that

$$a(\Omega^*, \overline{\mathcal{S}}) = \min_{1 \leq m \leq \overline{n}} \int\limits_{\overline{\Omega}_m \Delta \Omega^*} |\eta| \, d\rho_X \leq C \min_{1 \leq m \leq k} \left\{ m^{-s} + \Big(\frac{m \log n}{n}\Big)^{\frac{1}{2-\alpha}} \right\}. \tag{6.6}$$

Since $\#(\overline{\mathcal{S}}) = \overline{n} = n/2$, we have that $\varepsilon_n \leq C\frac{\log n}{n}$ in (2.3) when using $\overline{\mathcal{S}}$. Hence, from Corollary 3.2, we have

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

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

$$R(\overline{\Omega}_{m^*}) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{s}{(2-\alpha)s+1}} + C\Big(\frac{\log n}{n}\Big)^{\frac{1}{2-\alpha}} \leq C\Big(\frac{\log n}{n}\Big)^{\frac{s}{(2-\alpha)s+1}}. \tag{6.7}$$

Thus, we obtain the same estimate as if we knew α and s.

Remark 6.1 *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* (4.1).

7 Classification using tree based adaptive partitioning

We now turn to the main objective of this paper which is the construction and analysis of concrete algorithms for classification. 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 §5. 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 it 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

$$\#(\mathfrak{T}_m) \le \binom{2^d m}{m-1} \le \frac{(2^d m)^m}{(m-1)!} \le e^m 2^{dm}. \tag{7.1}$$

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

$$\#(\mathcal{S}_m) \le a^m \tag{7.2}$$

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 §8. This results in a set $\overline{\Omega}(\mathbf{z})$ and we have the following theorem for its performance.

Theorem 7.1 (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* (4.1)*, then with probability greater than $1 - cn^{-r+1}$, we*

have

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{s}{(2-\alpha)s+1}} \tag{7.3}$$

with C depending only on d, r, $|\rho|_{\mathcal{A}^s}$ and the constant in (4.1).
(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* (4.1)*, then with probability greater than $1 - cn^{-r+1}$, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{(2-\alpha)\beta+d(1-\alpha)}}, \tag{7.4}$$

with C depending only on d, r, $|\eta|_{B^\beta_\infty(L_p(X))}$ and the constant in (4.2).

Proof: Since $\log(\#(\mathcal{S}_m)) \le C_0 m$ where C_0 depends only on d, we have that $R(\Omega(\mathbf{z})) - R(\Omega^*)$ is bounded by the right side of (6.7) which proves (i). We can derive (ii) from (i) if we prove that the assumptions on ρ in (ii) imply that $\rho \in \mathcal{A}^s$, $s = \frac{\beta}{(1-\alpha)d} = \frac{(q+1)\beta}{d}$. 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 [9]) 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 $\|\eta - \eta_m\|_{L_\infty} \le \delta_m = C_1|\eta|_{B^\beta_\infty(L_p)} m^{-\beta/d}$ 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 $\Omega_m \Delta \Omega^* \subset \{x : \ |\eta(x)| \le \delta_m\}$ and so

$$a_m(\rho) \le \int_{\Omega_m\Delta\Omega^*} |\eta| \, d\rho_X \le \overline{C}_\rho \delta_m^{q+1} \le \overline{C}_\rho \Big(C_1|\eta|_{B^\beta_\infty(L_p)}\Big)^{q+1} m^{-s}, \tag{7.5}$$

as desired. □

Algorithm II: higher order methods via decorated trees

We want to remove the restriction $\beta \le 1$ that appears in Theorem 7.1 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 $m \ge 1$, we define

$$\mathcal{S}_m := \bigcup_{\mathcal{T}\in\mathfrak{T}_m} \mathcal{S}_\mathcal{T}. \tag{7.6}$$

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 7.2 *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 [11]. Let us consider any set of points $p_1, \ldots p_L$ from X. Then, from Theorem 9.2 in [11], 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 7.2 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 (7.1), 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

$$\mathrm{VC}(\mathcal{S}_\mathcal{T}) = (d+1)\#(\mathcal{L}(\mathcal{T})) \leq (d+1)2^d m. \tag{7.7}$$

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 (7.7). Lemma 7.2 now says that

$$\mathrm{VC}(\mathcal{S}_m) \leq \max\{8(d+2)m, 4(d+1)2^d m\} = C_d m, \tag{7.8}$$

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 §6 with $\mathcal{S}_m$ now given by (7.6). This results in a set $\overline{\Omega}(\mathbf{z})$ and we have the following theorem for the performance of this estimator.

Theorem 7.3 (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* (4.1)*, then with probability greater than $1 - cn^{-r+1}$, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{s}{(2-\alpha)s+1}} \tag{7.9}$$

with C depending only on d, r, $|\rho|_{\mathcal{A}^s}$ and the constant in (4.1)*.*
(ii) *If $\eta \in B^\beta_\infty(L_p(X))$ with $0 < \beta \leq 2$ and $p > d/\beta$ and if ρ satisfies the margin condition* (4.1)*, then with probability greater than $1 - cn^{-r+1}$, we have*

$$R(\overline{\Omega}(\mathbf{z})) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{(2-\alpha)\beta+d(1-\alpha)}}, \tag{7.10}$$

with C depending only on d, r, $|\eta|_{B^\beta_\infty(L_p(X))}$ and the constant in (4.1)*.*

Proof: In view of (7.8) we can invoke Theorem 2.5 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 (2.22) and hence is an admissible set function for the modulus (3.2). Now (i) follows now from (6.7).

To derive (ii) from (i), we prove that the assumptions on ρ in (ii) imply that $\rho \in \mathcal{A}^s$, $s = \frac{\beta}{(1-\alpha)d} = \frac{(q+1)\beta}{d}$, 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 [8]) 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 $\|\eta - \eta_m\|_{L_\infty} \le \delta_m = C_1|\eta|_{B^\beta(L_p(X))}m^{-\beta/d}$. 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

$$a_m(\rho) \le \int\limits_{\Omega_m\Delta\Omega^*} |\eta|\, d\rho_X \le \overline{C}_\rho\delta_m^{q+1} \le \overline{C}_\rho\Big(C_1|\eta|_{B^\beta_\infty(L_p)}\Big)^{q+1} m^{-s}, \tag{7.11}$$

as desired. □

Remark 7.4 *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* (7.10) *now holds for all $0 < \beta \le k+1$. 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 $s = \frac{\beta(q+1)}{d}$, but with β denoting the smoothness of η in the sense of the Lipschitz space $\mathrm{Lip}\,\beta$. 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 q in the margin condition are coupled, for example by the restriction $\beta q \le 1$ when ρ_X is bounded from below by the Lebesgue measure. Replacing the Lipschitz space $\mathrm{Lip}\,\beta$ 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

$$\eta(x) = \eta(x_1,\dots,x_d) = \mathrm{sign}(x_1 - 1/2)|x_1 - 1/2|^\gamma,$$

for some $0 < \gamma \le 1$, so that $\Omega^* = \{x \in X \ : \ x_1 > 1/2\}$ and the margin condition (4.2) holds with q up to $1/\gamma$. Then, one checks that $\eta \in B^\beta_\infty(L_p)$ for β and p such that $\beta \le \gamma + 1/p$. The constraint $1/p < \beta/d$ may then be rewritten as $\beta(1-1/d) < \gamma$ or equivalently

$$q\beta(1-1/d) < 1,$$

which is an improvement over $q\beta \le 1$.

8 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 $\overline{\Omega}_m \in \mathcal{S}_m$ given in (1.11) 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

$$\gamma_{\mathcal{T}} := \sum_{Q \in \mathcal{L}(\mathcal{T})} \max(\overline{\eta}_Q, 0), \tag{8.1}$$

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

$$\gamma_{Q,m} = \max_{\mathcal{T} \in \mathfrak{T}_m(Q)} \gamma_{\mathcal{T}}. \tag{8.2}$$

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

$$\mathcal{T}(Q, m) = \emptyset, \tag{8.3}$$

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

$$\overline{\mathcal{T}}_m := \mathcal{T}(X, m) \quad \text{and} \quad \overline{\Omega}_m := \bigcup_{Q \in \mathcal{L}(\overline{\mathcal{T}}_m)} \{Q : \overline{\eta}_Q > 0\}, \tag{8.4}$$

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

$$\mathcal{U}_l := \mathcal{T}'(\mathbf{z}) \cap \mathcal{D}_{\overline{n}-l}, \tag{8.5}$$

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 [7]), 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 $l = 1, \ldots, \overline{n}$: 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

$$(\ell_j^*(R))_{R\in\mathcal{C}'(Q)} := \operatorname{argmax}\Big\{ \sum_{R\in\mathcal{C}'(Q)} \gamma_{R,\ell'(R)} : \sum_{R\in\mathcal{C}'(Q)} \ell'(R) = j \Big\}, \tag{8.6}$$

  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 $1 \le j \le l$,

$$\gamma_{Q,j} = \sum_{R\in\mathcal{C}'(Q)} \gamma_{R,\ell_{j-1}^*(R)}, \quad \mathcal{T}(Q,j) = \{Q\} \cup \Big(\bigcup_{R\in\mathcal{C}'(Q)} \mathcal{T}(R, \ell_{j-1}^*(R)) \Big) \cup (\mathcal{C}(Q) \setminus \mathcal{C}'(Q)). \tag{8.7}$$

  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 $m = 0, \ldots, \overline{n}$. This provides the estimators $\overline{\Omega}_m$ for $m = 0, \ldots, \overline{n}$.

To estimate the complexity of the algorithm, we need to bound for each $l \in \{1, \ldots, \overline{n}\}$ the number of computations required by (8.6) and (8.7). With proper organization, the argmax in (8.6) can be found using at most $\mathcal{O}(\#(C'(Q))l^2)$ operations. We can execute (8.7) 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

$$\gamma_{\mathcal{T}} := \sum_{Q\in\mathcal{L}(\mathcal{T})} \max_{H\in\mathcal{H}, i=0,1} \max\{0, \overline{\eta}_{Q_i^H}\}. \tag{8.8}$$

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

$$\gamma_{Q,0} := \max_{H\in\mathcal{H}, i=0,1} \max\{0, \overline{\eta}_{Q_i^H}\} \tag{8.9}$$

and define as before $\gamma_{Q,m}$ and $\mathcal{T}(Q,m)$ by (8.2) and (8.4).

The procedure of determining the trees $\mathcal{T}(X,m)$ for $m = 0, \ldots, k$ 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 (8.9). 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 estimation of η on Q by a polynomial.

9 Plug-in classifiers

While, the main interest of this paper is in set estimators, it is useful to make some comments on plug-in estimators to help frame the results we have presented. Let $(V_m)_{m\geq 1}$ be a nested sequence of linear or nonlinear spaces with V_m of VC dimension at most m by which we mean the set of all epigraphs of the functions $g \in V_m$ has VC dimension at most m. A plug-in method uses the draw $\mathbf{z}$ to find an approximation $\tilde{\eta}_m \in V_m$ to η for each $1 \leq m \leq n$ and then uses model selection to choose m, thereby given an approximation $\tilde{\eta} := \tilde{\eta}_m$ to η. Typically, each $\tilde{\eta}_m$ is obtained by empirical least squares minimization over V_m, followed by a truncation. The classifier is then defined as

$$\tilde{\Omega} := \{x : \tilde{\eta}(x) \geq 0\}. \tag{9.1}$$

Let us first observe that the set $\tilde{\Omega}$, can also be viewed as obtained as an empirical set estimator. For each ρ_X-measurable function g on X, we define $S_g = \{x \in X : \ g(x) \geq 0\}$ and then define the collection of sets $\mathcal{S}_m := \{S_g : g \in V_m\}$ which have VC dimension at most m. If we define the set estimator

$$\tilde{\eta}_S := \int_S \tilde{\eta} d\rho_X \tag{9.2}$$

for each measurable S, then

$$\eta_{\tilde{\Omega}} = \max_{S\in\mathcal{S}} \tilde{\eta}_S. \tag{9.3}$$

Said, in other words, $\tilde{\Omega} = \Omega_{\mathcal{S}}$ for this family $\mathcal{S}$ of sets. Thus, plug-in estimators can always be viewed as set estimators and are therefore included in the analysis we give below.

The estimator $\tilde{\eta}_S$, defined in (9.2) is a way to approximate η_S and it satisfies

$$\begin{aligned} |\eta_S - \eta_{\tilde{\Omega}} - \tilde{\eta}_S + \tilde{\eta}_{\tilde{\Omega}}| &= \Big| \int_{S\setminus\tilde{\Omega}} [\eta - \tilde{\eta}] \, d\rho_X - \int_{\tilde{\Omega}\setminus S} [\eta - \tilde{\eta}] \, d\rho_X \Big| \\ &\leq \int_{S\Delta\tilde{\Omega}} |\eta - \tilde{\eta}| \, d\rho_X \leq \rho_{S\Delta\tilde{\Omega}}^{1/p'} \|\eta - \tilde{\eta}\|_{L_p(S\Delta\tilde{\Omega},\rho_X)}, \end{aligned} \tag{9.4}$$

for all measurable sets S. If we want this estimator to fall into the general theory we have developed, then we need to ensure that

$$\|\eta - \tilde{\eta}\|_{L_p(\rho_X)} \leq \varepsilon_n \quad \text{with high probability on the draw } \mathbf{z}, \tag{9.5}$$

for some sequence (ε_n) tending to zero. When this is the case, (3.1) holds for

$$e_n(S) := \rho_{S\Delta\hat{\Omega}}^{1/p'} \varepsilon_n, \tag{9.6}$$

and our general theory, via the modulus ω, can be applied to derive risk bounds for plug-in estimators.

There is, however, a more direct route to proving risk bounds for plug-in estimators which begins by observing that

$$R(\tilde{\Omega}) - R(\Omega^*) = \int_{\tilde{\Omega}\Delta\Omega^*} |\eta| \leq \int_{\tilde{\Omega}\Delta\Omega^*} |\eta - \tilde{\eta}| \, d\rho_X \leq \rho_X(\tilde{\Omega}\Delta\Omega^*)^{1/p'} \|\eta - \tilde{\eta}\|_{L_p(\rho_X)} \leq \|\eta - \tilde{\eta}\|_{L_p(\rho_X)}. \tag{9.7}$$

Hence, whenever (9.5) holds, we have with high probability on the draw $\mathbf{z}$ that

$$R(\tilde{\Omega}) - R(\Omega^*) \leq \varepsilon_n. \tag{9.8}$$

This can be improved if we assume in addition the margin condition (4.2). Indeed, again assuming (9.5), we have with high probability for all $t > 0$,

$$t^p \rho_X\{x : \ |\tilde{\eta}(x) - \eta(x)| > t\} \leq \int\limits_{\Omega} |\eta - \tilde{\eta}|^p \, d\rho_X \leq \varepsilon_n^p. \tag{9.9}$$

Next note that

$$\{x : \ \tilde{\eta}(x) \geq 0 \text{ and } \eta(x) < 0\} \subset \{x : |\eta(x)| \leq t\} \cup \{x : |\tilde{\eta}(x) - \eta(x)| > t\}. \tag{9.10}$$

Since a similar containment holds for the set $\{x : \ \tilde{\eta}(x) < 0 \text{ and } \eta(x) \geq 0\}$, we infer from (4.2) and (9.9) that

$$\rho_X(\tilde{\Omega}_\Delta \Omega^*) \leq \overline{C}_\rho t^q + t^{-p} \varepsilon_n^p. \tag{9.11}$$

If we take $t = \varepsilon_n^{\frac{p}{p+q}}$, we arrive at the estimate

$$\rho_X(\tilde{\Omega}_\Delta \Omega^*) \leq \overline{C} \varepsilon_n^{\frac{pq}{p+q}}. \tag{9.12}$$

When this is injected into (9.7), we arrive at the risk bound

$$R(\tilde{\Omega}) - R(\Omega^*) \leq \rho_X(\tilde{\Omega}_\Delta \Omega^*)^{1/p'} \|\eta - \tilde{\eta}\|_{L_p(\rho_X)} \leq \varepsilon_n^{\frac{1+q}{1+q/p}}. \tag{9.13}$$

Given our goal of obtaining risk estimates that hold with high probability on the draw $\mathbf{z}$, the critical question is when do we have plug-in estimators for which (9.5) is valid. We confine our discussion to the two most important cases $p = 2$ and $p = \infty$.

The case $p = 2$: Deriving $L_2(\rho_X)$ estimates for the empirical approximation of η is particularly well studied. The usual theory for regression (see e.g. [11]) proceeds as follows. For each $1 \leq m \leq n$, we find the best empirical least squares fit $\tilde{\eta}_m$ from V to the data $\mathbf{z}$. We then define $\tilde{\eta} := T_M \tilde{\eta}$, where T_M is the truncation operator $T_M z := \operatorname{sign} z \min\{|z|, M\}$ and M is an a priori bound for $\|\eta\|_{L_\infty}$ (in our case $M = 1$). One then uses model selection to find the appropriate choice of m and thereby proves that $\tilde{\eta} := T_m \tilde{\eta}_m$ satisfies

$$\mathbb{E}(\|\eta - \tilde{\eta}\|_{L_2(\rho_X)}) \leq \varepsilon_n, \tag{9.14}$$

where ε_n depends on the smoothness assumption on η and the particular approximation method. This does not satisfy our goals since we want results that hold with high probability rather than just in expectation. In fact, it is known that, *for general measures*, the above approach to defining $\tilde{\eta}$ *will not* give (9.5) (see [3]). However, some significant results are known in certain special cases. We will only discuss the case of approximation by piecewise polynomials as reported in [5] and its followups [3, 4].

One case, where one can obtain results like (9.14) that hold with high probability is when V is a space of *piecewise constants* on $X = [0, 1]^d$ (see [5]). In this case, the plug-in estimator gives sets to approximate Ω^* similar to our Algorithm 1. There are two types of piecewise constant

approximation. In linear approximation, one fixes a hierarchy of partitions $\mathcal{P}_m$ - typically with uniform spacing and then uses the linear spaces V_m of piecewise constant functions which are subordinate to the partition $\mathcal{P}_m$. In this case, one can use empirical least squares to generate the function $\tilde{\eta}$. The critical issue is what value should be chosen for m given the draw $\mathbf{z}$ of n points. If it is known that $\eta \in \operatorname{Lip} \beta$, then the best choice of m is $m \sim n^{\frac{d}{2\beta+d}}$. With this choice, one can prove that with high probability on the draw $\mathbf{z}$, we have

$$\|\eta - \tilde{\eta}\|_{L_2(\rho_X)} \leq C\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{2\beta+d}}, \tag{9.15}$$

provided $0 < \beta \leq 1$. This result holds with no additional assumptions on the measure ρ_X. We can obtain this same result without knowledge of β by using model selection. When this is used in (9.13), we obtain the risk estimate

$$R(\tilde{\Omega}) - R(\Omega^*) \leq C\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{(2-\alpha)\beta+d(1-\alpha/2)}}. \tag{9.16}$$

Note that (9.16) is always worse than the corresponding estimate given in (7.5) even though both use a similar family of sets to approximate Ω^*.

The second (*nonlinear*) form of piecewise constant approximation is to utilize adaptive partitioning results. Then, the space V_m consists of all piecewise constants which are subordinate to an allowable partition into at most m cells. The allowable partitions are the same as in our tree decompositions of §7. In this case, one can prove that with high probability the result (9.15) holds but now under the weaker assumption that η is in a Besov like space (depending on the measure ρ_X). One arrives at (9.16) for the risk estimate but now under the weaker Besov assumption. Again, (9.16) is worse than the corresponding bound given in (7.5).

If one considers piecewise polynomial approximation of order r (degree $r-1$) then bounds for the empirical least squares approximation are only known to hold with high probability when one imposes severe restrictions on the measure ρ_X (roughly speaking it should be equivalent to Lebesgue measure). Thus, it does not seem possible to obtain results comparable to our Algorithm 2 from this approach.

The case $p = \infty$: This case has been the subject of recent interest. For example, in Lemma 3.1 of [1], the authors consider regression functions η and approximation methods which take the data $\mathbf{z}$ and generate an $\tilde{\eta}$ from a linear space V chosen from the sequence (V_m) for which the following holds: for almost all $x \in X$,

$$\mathbb{P}\{|\eta(x) - \tilde{\eta}(x)| \geq \delta\} \leq C_0 e^{-a_n \delta^2}. \tag{9.17}$$

Here (a_n) is a sequence which is typically of the form $a_n = C_1 n^\gamma$ for some $\gamma > 0$ which depends on the smoothness of η. The authors of [1] go on to prove certain risk bounds when using $\tilde{\eta}$ as a plug-in classifier. On the surface, it seems that the condition (9.17) is a weaker assumption than requiring that (9.5) holds for $p = \infty$. However, we will now see this is not the case, at least when standard approximation methods are employed.

For simplicity, let us assume that the spaces V_m are linear and $\dim(V_m) = m$. We want to show that (9.17) actually implies (9.4) when using standard approximation spaces V_m. To see this, we recall that standard approximation spaces, and, in particular, the spaces used in [1] satisfiy what

are called *Bernstein inequalities* (which should be distinguished from the Bernstein's inequality about the concentration of measure). In the setting of current interest, this inequality would say that all functions $g \in V_m$ satisfy

$$\|g\|_{\mathrm{Lip}\,\beta} \leq C_B m^{\frac{\beta}{d}} \|g\|_{L_\infty(X)}, \quad g \in V_m, \tag{9.18}$$

with C_B an absolute constant. If (9.18) holds for a value β_0, then it is known to hold for smaller values $\beta < \beta_0$ as well.

Lemma 9.1 *Suppose that* (9.17) *holds for functions* $\eta \in \mathrm{Lip}\,\beta$ *with* $a_n = C_1 n^{\frac{2\beta}{2\beta+d}}$ *by choosing* $\tilde{\eta}$ *from a linear space* V_m *of dimension* $m = \lceil n^{\frac{d}{2\beta+d}} \rceil$ *which satisfies the Bernstein inequality for* β *with constant* C_B. *Then, for any* $r > 0$, $\tilde{\eta}$ *also satisfies*

$$\|\eta - \tilde{\eta}\|_{L_\infty(\rho_X)} \leq C_r (1 + |\eta|_{\mathrm{Lip}\,\beta}) n^{-\frac{\beta}{2\beta+d}} \sqrt{\log n} \tag{9.19}$$

with probability $\geq 1 - C_r n^{-r}$ *on the draw* $\mathbf{z}$.

Proof: We first take a set $X_0 \subset X$ with $\#(X_0) \leq C(\beta, d) m^2$ such that

$$\mathrm{dist}(x, X_0) \leq (2C_B)^{-1/\beta} m^{-2/d}, \quad x \in X.$$

For any $g \in V_m$ and any $x \in X$ and the point $x_0 \in X_0$ closest to x, we have

$$\begin{aligned} |g(x)| &\leq |g(x) - g(x_0)| + |g(x_0)| \\ &\leq \|g\|_{C\beta} |x - x_0|^\beta + \|g\|_{L_\infty(X_0)} \\ &\leq C_B m^{\beta/d} \|g\|_{L_\infty(X)} (2C_B)^{-1} m^{-\beta/d} + \|g\|_{L_\infty(X_0)} \\ &\leq (1/2)\|g\|_{L_\infty(X)} + \|g\|_{L_\infty(X_0)}, \end{aligned} \tag{9.20}$$

where we have used Bernstein's inequality in the third inequality. From (9.20), we find

$$\|g\|_{L_\infty(X)} \leq 2\|g\|_{L_\infty(X_0)}, \quad g \in V_m. \tag{9.21}$$

We now take

$$\delta := A m^{-\beta/d} \sqrt{\log n} \leq A n^{-\frac{\beta}{2\beta+d}} \sqrt{\log n},$$

with the constant $A \geq 1$ to be chosen in a moment. Then, the condition (9.17) gives that

$$|\eta(x_0) - \tilde{\eta}(x_0)| \leq \delta, \quad x_0 \in X_0 \tag{9.22}$$

holds with probability $\geq 1 - C\#(X_0) e^{-a_n \delta^2} \geq 1 - C n^{-r}$ provided the constant A is chosen large enough. We fix such an A, and in going further, we consider only draws $\mathbf{z}$ for which (9.22) is valid. Then, for any $x \in X$ and a point $x_0 \in X_0$ closest to x, we have from Bernstein's inequality

$$\begin{aligned} |\eta(x) - \tilde{\eta}(x)| &\leq |\eta(x) - \eta(x_0)| + |\eta(x_0) - \tilde{\eta}(x_0)| + |\tilde{\eta}(x_0) - \tilde{\eta}(x)| \\ &\leq \|\eta\|_{\mathrm{Lip}\,\beta} |x - x_0|^\beta + \delta + \|\tilde{\eta}\|_{\mathrm{Lip}\,\beta} |x - x_0|^\beta, \\ &\leq \|\eta\|_{\mathrm{Lip}\,\beta} m^{-2\beta/d} + \delta + C_B \|\tilde{\eta}\|_{L_\infty(X)} m^{\beta/d} (2C_B)^{-1} m^{-2\beta/d} \\ &\leq \{\|\eta\|_{\mathrm{Lip}\,\beta} + 1 + (1/2)\|\tilde{\eta}\|_{L_\infty(X)}\}\delta \\ &\leq \{\|\eta\|_{\mathrm{Lip}\,\beta} + 1 + \|\tilde{\eta}\|_{L_\infty(X_0)}\}\delta \\ &\leq \{\|\eta\|_{\mathrm{Lip}\,\beta} + 1 + \|\eta\|_{L_\infty(X)} + \delta\}\delta. \end{aligned} \tag{9.23}$$

Since $\|\eta\|_{L_\infty(X)} \le 1$, this completes the proof. □

Let us now compare the risk estimates obtained from the starting point of (9.17) with those given in this paper. A first point is that the risk bounds in [1] are in expectation while the results of this paper are with high probability. Also, the results in §7 are based on nonlinear methods and hence apply to the wider Besov classes in place of Lipschitz spaces considered in [1]. Perhaps the biggest distinction is that our results apply to arbitrary messures ρ_X, whereas those in [1] require that ρ_X is equivalent to Lebesgue measure on its support. Thus, the representative result in [1] is that when ρ_X is Lebesgue measure and η is a Lip β function, then whenever the margin condition (4.2) holds

$$\mathbb{E}(R(\tilde{\Omega}) - R(\Omega^*)) \le Cn^{-\frac{\beta(q+1)}{2\beta+d}} = Cn^{-\frac{\beta}{(2\beta+d)(1-\alpha)}}. \tag{9.24}$$

Note, that β and q cannot be chosen independently in this case.

On the other hand, in Theorem 7.3, we obtain with high probability the bound

$$R(\hat{\Omega}(\mathbf{z})) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{(2-\alpha)\beta+d(1-\alpha)}} = C\Big(\frac{\log n}{n}\Big)^{\frac{\beta}{(2\beta+d)(1-\alpha)+\beta\alpha}}, \tag{9.25}$$

This estimate is worse than (9.24) because it applies to arbitrary measures and does not take advantage of the fact that the underlying measure is Lebesgue. However, if one assumes at the outset that ρ_X is equivalent to Lebesgue measure then the bounds we obtain for set estimators can be improved. Let us consider a uniform partition $\mathcal{Q}_m$ of $[0,1]^d$ into cubes of side length $1/m$ and consider the collection $\mathcal{S}_m$ of all sets $S = \bigcup_{Q\in\Lambda} Q$ where Λ is an arbitrary subset of $\mathcal{Q}$. The sequence $(\mathcal{S}_m)$ is the linear analogue of the sets used in Algorithm I.

For simplicity, we assume ρ_X is Lebesgue measure and show that we can improve Lemma 2.2 by using $e_n(S) := \rho_S\sqrt{\overline{\varepsilon}_n}$ where $\overline{\varepsilon}_n = \frac{8(r+1)dm(1+\log n)}{3n}$. Indeed, from Bernstein's inequality, we have

$$\mathbb{P}\{|\eta_S - \hat{\eta}_S| > e_n(S)\} \le 2\exp\Big\{-\frac{3n\rho_S\overline{\epsilon}_m}{8}\Big\} = 2\exp\Big\{-(r+1)m\rho_S(1+\log n)\Big\}. \tag{9.26}$$

Now, for any $S \in \mathcal{S}_m$, we have $\rho_S = \frac{k}{m}$ for some integer $1 \le k \le m$ and the number of such sets $S \in \mathcal{S}_m$ is $\binom{m^d}{k} \le (\frac{em^d}{k})^k$. Therefore, using (9.26) and a union bound shows that $|\eta_S - \overline{\eta}_S| \le e_n(S)$ with probability greater than $1 - 2\sum_{k=1}^m n^{-(r+1)dk}\Big(\frac{em^d}{k}\Big)^k \ge 1 - cn^{-r}$.

If we now use $e_n(S) = \varepsilon_n\rho_S$ in the definition of $\omega(\rho, e_n)$ we obtain a new bound for the variance. To understand this bound and its relationship to margin conditions, leads us to consider

$$\phi'(\rho,t) := \sup_{\int_S |\eta|\le 3t\rho_S} \int_S |\eta|\, d\rho_X. \tag{9.27}$$

Indeed, $\omega(\rho, e_n) = \phi'(\rho, \overline{\varepsilon}_n)$. The followng lemma relates ϕ' to a margin relation.

Lemma 9.2 *For any $0 < t \le 1$,*

$$\phi'(\rho,t) \le 6t\rho_X\{x:\ 0 < |\eta(x)| \le 6t\}. \tag{9.28}$$

Proof: Let S be any set for which $\int_S |\eta|\, d\rho_X \le 3t\rho_S$ and define $S_0 := \{x \in S :\ 0 < |\eta(x)| \le 6t\}$ and $S_1 := \{x \in S : 6t < |\eta(x)|\}$. Then,

$$6t\rho_{S_1} \le \int_S |\eta|\, d\rho_X \le 3t\rho_S \le 3t[\rho_{S_0} + \rho_{S_1}].$$

Hence, $\rho_{S_1} \le \rho_{S_0}$ and $\rho_S \le 2\rho_{S_0}$. It follows that

$$\int_S |\eta|\, d\rho_X \le 3t\rho_S \le 6t\rho_{S_0} \le 6t\rho_X\{x : 0 < |\eta(x)| \le 6t\}.$$

If we take a supremum over all such S, we arrive at (9.28). □

Thus, using the set $\mathcal{S}_m$ in Theorem 3.1, gives the following estimate when $\eta \in \mathrm{Lip}\beta$ and ρ satisfies the margin condition (4.2)

$$R(\hat{\Omega}_{\mathcal{S}_m} - R(\Omega^*) \le \max\{\omega(\rho_m, (e_n)), a_m(\rho)\} \le C \max\left\{(m^{-\frac{\beta}{d}})^{q+1}, \Big(\frac{m(1+\log n)}{n}\Big)^{q+1}\right\}. \quad (9.29)$$

If we choose $m = n^{\frac{d}{2\beta+d}}$ to balance the two terms in (9.29), we obtain

$$R(\hat{\Omega}_{\mathcal{S}_m}) - R(\Omega^*) \le C\Big(\frac{\log n}{n}\Big)^{\frac{\beta(q+1)}{2\beta+d}}. \quad (9.30)$$

Since $q + 1 = 1/(1-\alpha)$ with α the parameter in (4.1), we have the same estimate as (9.24).

Acknowledment: The authors wish to thank Stephane Gaiffas for various valuable suggestions and references.

References

[1] J.-Y. Audibert and A.N. Tsybakov, *Fast learning rates for plug-in classifiers*, Ann. Statistics **35** (2007), 608-633.

[2] P. Bartlett and S. Mendelson, *Empirical minimization*, Prob. Theory and Related Fields **135** (2003), 311-334.

[3] P. Binev, A, Cohen, W. Dahmen, and R. DeVore, *Universal algorithms for learning theory, part II: piecewise polynomials*, Constructive Approximation, **26**(2007) 127–152.

[4] P. Binev, A, Cohen, W. Dahmen, and R. DeVore, *Universal Piecewise Polynomial Estimators for Machine Learning, in Curves and Surface Design*, Proceedings of the Avignon Conference (2006), (A. Cohen, J.L. Merrien, L. Shumaker, Eds.), Nashboro Press, 2007, 48–78.

[5] P. Binev, A. Cohen, W. Dahmen, and R. DeVore, V. Temlyakov, *Universal algorithms for learning theory, part I: piecewise constant functions*, J. Machine Learning, 6 (2005) 1297–1321.

[6] Olivier Bousquet, Stéphane Boucheron, Gabor Lugosi, *Theory of Classification: a Survey of Some Recent Advances*, ESAIM: PS 9 (2005) 323-375

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees.*, Wadsworth, 1984.

[8] A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, *Tree-structured approximation and optimal encoding*, Appl. Comp. Harm. Anal. **11** (2001), 192-226.

[9] R. DeVore, *Nonlinear approximation*, Acta Numerica **7** (1998), 51-150.

[10] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov, *On Mathematical Methods for Supervised Learning*, J. of FOCM, **6** (2006) 3–58.

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

[12] P. Massart, *Concentration Inequalities and Model Selection*, Springer, 2007.

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

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

[15] Vladimir Temlyakov, *Optimal Estimators in Learning Theory* , Approximation and Probability, Banach Center Publications, Volume 72, Inst. Math. Polish Academy of Sciences, Warsaw (2006), 341-366.

[16] Alexandre 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
binev@math.sc.edu

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

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

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