ON THRESHOLD-BASED CLASSIFICATION RULES

LEILA MOHAMMADI AND SARA VAN DE GEER
Leiden University

Suppose we have n i.i.d. copies $\{(X_i, Y_i); i = 1, \ldots, n\}$ of an example (X, Y), where $X \in \mathcal{X}$ is an instance and $Y \in \{-1, 1\}$ is a label. A decision function (or classifier) f is a function $f\colon \mathcal{X} \to [-1, 1]$. Based on f, the example (X, Y) is misclassified if $Yf(X) \leq 0$. In this paper, we first study the case $\mathcal{X} = \mathbb{R}$, and the simple decision functions $h_a(x) = 2\mathrm{l}\{x \geq a\} - 1$ based on a threshold $a \in \mathbb{R}$. We choose the threshold $\hat{a}_n$ that minimizes the classification error in the sample, and derive its asymptotic distribution. We also show that, under monotonicity assumptions, $\hat{a}_n$ is a nonparametric maximum likelihood estimator. Next, we consider more complicated classification rules based on averaging over a class of base classifiers. We allow that certain examples are not classified due to lack of evidence, and provide a uniform bound for the margin. Moreover, we illustrate that when using averaged classification rules, maximizing the number of examples with margin above a given value, can overcome the problem of overfitting. In our illustration, the classification problem then boils down to optimizing over certain threshold-based classifiers.

1. Introduction

Suppose we have n i.i.d. realizations $\{(X_i, Y_i); i = 1, \ldots, n\}$ of an example (X, Y), where $X \in \mathcal{X}$ is an instance and $Y \in \{-1, 1\}$ is a label. A decision function f is a function $f\colon \mathcal{X} \to [-1, 1]$. We will also refer to f as a classifier. The decision rule based on f is to attach to an instance $x \in \mathcal{X}$ the label $y = 1$ if $f(x) > 0$, and otherwise to attach the label $y = -1$. Using this rule, the example (X, Y) is misclassified if $Yf(X) \leq 0$. A base classifier h is a function $h\colon \mathcal{X} \to \{-1, 1\}$, attaching the label $h(x)$ to the instance x.

Given a class of classifiers $\mathcal{F}$, the problem is to choose the "best" one. Let $L(f) = P(Yf(X) \leq 0)$ be the theoretical loss, or prediction error, of the classifier $f \in \mathcal{F}$. Thus, if X_{n+1} is a new instance for which we want to predict the label, $L(f)$ is the mean error of the prediction

$$\hat{Y}_{n+1} = \begin{cases} 1, & \text{if } f(X_{n+1}) > 0 \\ -1, & \text{if } f(X_{n+1}) \leq 0. \end{cases}$$

The smallest possible prediction error over $\mathcal{F}$ is $\min_{f \in \mathcal{F}} L(f)$. Consider now the empirical loss $L_n(f)$ of a particular classifier f, which is defined as the fraction of examples in the sample that have been misclassified by f. We will study, for some choices of $\mathcal{F}$, the classifier $\hat{f}_n$ that minimizes $L_n(f)$ over

AMS subject classifications: 62G05, 62G20.

Keywords and phrases: bounded variation; classification; classification error; convex hull; cube root asymptotics; entropy; threshold; VC class.

$f \in \mathcal{F}$. Following Vapnik (1995), we call the classifier $\hat{f}_n$ the *empirical risk minimizer.*

The overall minimizer

$$f_* = \arg \min_{\textit{all} \text{ classifiers}} L(f),$$

is called *Bayes* rule. The empirical counterpart of this rule, i.e., the minimizer of $L_n(f)$ over *all* f, may overfit the data, and thus have no predictive power at all. More generally, to avoid overfitting, the class $\mathcal{F}$ of classifiers cannot be chosen too large (or too complex).

Our aim in this paper is to provide results in a relatively simple situation, proving limit theorems in the "finite dimensional case" (see Section 2), and doing complexity regularization in "infinite-dimensional cases" (see Section 4). We believe that these results lead to a better understanding of more complicated classification problems as well.

In most of the paper we consider the case where the instance space $\mathcal{X}$ is the real line or a subset thereof. In Section 2, we will examine the class of base classifiers of the form

$$h_a(x) = \begin{cases} 1, & \text{if } x \geq a \\ -1, & \text{if } x < a, \end{cases}$$

with $a \in \mathbb{R}$. We first recall some parametric approaches, and then derive the asymptotic behavior of the (nonparametric) empirical risk minimizer $\hat{a}_n$.

In Section 3 we study the problem of *averaging* over base classifiers. The idea there is as follows. If one has several base classifiers, say $h_1, \ldots, h_T$, one may let them vote. A simple majority vote is to label x with $y = 1$ if 50% or more of the h_t vote for $y = 1$, i.e., if $(1/T) \sum_{t=1}^T h_t(x) > 0$. However, one may find that certain base classifiers are more important than others. This can be expressed by using a weighted average $f(x) = \sum_{t=1}^T \alpha_t h_t(x)$, where α_t expresses the relative importance of base classifier h_t, $0 \leq \alpha_t \leq 1$, $\sum_{t=1}^T \alpha_t = 1$. In most situations, little is known a priori about the relative importance of the base classifiers. The weights are then chosen according to some data-dependent criterion (for instance using AdaBoost (see, e.g., Freund and Schapire, 1997). This leads to considering all possible convex combinations

$$\mathcal{C} = \text{conv}(\mathcal{H}),$$

where

$$\text{conv}(\mathcal{H}) = \left\{ f = \sum_{t=1}^T \alpha_t h_t, 0 \leq \alpha_t \leq 1, \sum_{t=1}^T \alpha_t = 1, T \in \mathbb{N} \right\}, \tag{1.1}$$

is the convex hull of $\mathcal{H}$.

Section 3 sketches a general result on averaging and the complexity of the resulting $\mathcal{C}$, together with some examples. These examples illustrate that averaging can lead to overfitting. To overcome this problem, one may consider allowing that certain examples are not classified due to lack of evidence (i.e., no sufficient majority). We will establish a uniform bound for the fraction of well-classified examples, i.e., the margin, for the case where $\mathcal{C}$ is a class of functions of bounded variation. The results in Section 3 are inspired by the paper of Schapire et al (1998). In Section 4, we put the idea of *maximizing the margin* in the context of *regularization*. We examine functions of bounded variation, and show that maximizing the number of examples with margin above a given value θ is the same as empirical risk minimization using more than one threshold. Section 5 concludes. The proofs are in the appendix.

In our study, the empirical behavior of certain quantities is compared to the theoretical counterpart. The notation used in the theory of empirical processes is helpful to express the notions in a consistent way. We let

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$$

be the empirical distribution based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Thus P_n puts mass $1/n$ on each example (X_i, Y_i), $i = 1, \ldots, n$. The distribution of (X, Y) is denoted by P.

Given a classifier f, we write its empirical error as

$$L_n(f) = P_n(\mathbf{y}f \leq 0) = \#\{Y_i f(X_i) \leq 0, 1 \leq i \leq n\}/n, \tag{1.2}$$

and its theoretical error as

$$L(f) = P(\mathbf{y}f \leq 0) = P(Yf(X) \leq 0). \tag{1.3}$$

Moreover, we let

$$F_0(x) = P(Y = 1 | X = x) \tag{1.4}$$

be the conditional probability of the label $Y = 1$ if the instance X has value x. Thus, Bayes rule f_* is

$$f_*(x) = 2\mathrm{l}\{F_0(x) > \tfrac{1}{2}\} - 1, \quad x \in \mathcal{X}$$

(where we just fixed some choice for values x with $F_0(x) = \frac{1}{2}$). The distribution of X is denoted by G.

We consider asymptotics as $n \to \infty$, regarding the sample $(X_1, Y_1), \ldots,$ (X_n, Y_n) as first n of an infinite sequence of i.i.d. copies of (X, Y). The distribution of the infinite sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ is denoted by $\mathbf{P}$. Convergence in distribution (law) of a sequence Z_n to Z is denoted by $Z_n \xrightarrow{\mathcal{L}} Z$, and $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

2. A simple base classifier

In this section, we assume that $\mathcal{X} = \mathbb{R}$, and we consider the set of base classifiers

$$h_a(x) = 2\,\mathrm{l}\{x \geq a\} - 1,$$

i.e.,

$$\mathcal{H} = \{h_a : \ a \in \mathbb{R}\}. \tag{2.1}$$

In this case $F_0 = P(Y = 1|X = \cdot)$ is some function on $\mathbb{R}$. If $F_0(x)$ is an increasing function of x, then large values of x make the value $Y = 1$ more likely. It then indeed makes sense to predict the value $Y = 1$ when x exceeds a certain threshold a.

Let us use the short-hand (abuse of) notation

$$L_n(a) = L_n(h_a) = P_n(\mathbf{y}h_a \leq 0), \tag{2.2}$$

for the empirical error using the threshold a, and

$$L(a) = L(h_a) = P(\mathbf{y}h_a \leq 0), \tag{2.3}$$

for its theoretical counterpart.

The problem is now how one should choose a. The best threshold is the theoretical minimizer

$$a_0 = \arg\min_{a \in \mathbb{R}} L(a).$$

But L is not known. However, we can use the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ to estimate L.

2.1. Parametric models

Note that $(X_1, Y_1), \ldots, (X_n, Y_n)$ consists of two samples, one sample with $Y_i = 1$, and one with $Y_i = -1$. In the classical setup (see Duda and Hart, 1973), a parametric form for the distribution of both samples is assumed. For instance, let us suppose that the sample with $Y_i = 1$ is normally distributed with mean μ and variance σ^2, and the one with $Y_i = -1$ is normally distributed with mean ν and the same variance σ^2. Consider the situation where $\mu > \nu$. Write $P(Y = 1) = 1 - P(Y = -1) = p$. Let ϕ be the standard normal density. Then

$$F_0(x) = \frac{p\phi([x-\mu]/\sigma)}{p\phi([x-\mu]/\sigma) + (1-p)\phi([x-\nu]/\sigma)}. \tag{2.4}$$

It is easy to see that F_0 is a distribution function and that its density f_0, with respect to Lebesgue measure, exists.

The median

$$a_0 = F_0^{-1}(\tfrac{1}{2})$$

of F_0, minimizes the probability of misclassification $L(a)$, over $a \in \mathbb{R}$. When $p = \frac{1}{2}$, one obtains $a_0 = (\mu + \nu)/2$.

We may estimate a_0 using the maximum likelihood estimator. This estimator converges with rate $\sqrt{n}$ and is asymptotically normal.

Another classical approach to the classification problem is to assume that F_0 is the logistic distribution function with shift parameter a_0:

$$F_0(x) = 1 - \frac{1}{1 + \mathrm{e}^{\lambda(x-a_0)}}, \quad x \in \mathbb{R}. \tag{2.5}$$

One may (again) use the maximum likelihood method to estimate a_0. Note that (2.4) coincides with (2.5) when we take $\lambda = (\mu - \nu)/\sigma^2$, and $a_0 = (1/\lambda)\log\bigl(p/(1-p)\bigr) + (\mu+\nu)/2$. However, in the two normal samples model, the distribution G of X clearly depends in the parameters μ, ν, p (and σ^2), and hence on the parameter a_0, whereas in the logistic regression model, G is treated as completely unknown.

2.2. Cube root asymptotics in classification

In this subsection, and throughout the rest of the paper, we do not assume a parametric form for the distribution and thus for the function F_0. We consider the *direct* approach as advocated by Vapnik (1995), where the classification rule is solely based on minimizing the error in the sample. For our simple classifiers this means taking the empirical risk minimizer

$$\hat{a}_n = \arg\min_a L_n(a).$$

We will derive the asymptotic distribution of $\hat{a}_n$. Moreover, we show that $\hat{a}_n$ is the nonparametric maximum likelihood estimator of a_0, when F_0 is known to be a distribution function.

We first need to show that when a_0 is uniquely defined, the estimator $\hat{a}_n$ converges to a_0.

Lemma 2.1. *Suppose that $L(a)$ has a unique minimum at a_0, and that $L(\cdot)$ is continuous. Then $\hat{a}_n \to a_0$ almost surely.*

We now establish that under regularity conditions, $\hat{a}_n$ converges with rate $n^{1/3}$ to a_0, and derive the asymptotic distribution. This result follows from an application of Kim and Pollard (1990). Such an application can also be found in Bühlmann and Yu (2002), where another classification problem is studied.

Let $W(t)$: $t \in \mathbb{R}$ be a two-sided Brownian motion, and

$$Z = \arg\max_t [W(t) - t^2].$$

Theorem 2.1. *Suppose*

$$F_0(x)\begin{cases} < \frac{1}{2}, & \textit{if } x < a_0 \\ = \frac{1}{2}, & \textit{if } x = a_0 \\ > \frac{1}{2}, & \textit{if } x > a_0. \end{cases}$$

Assume moreover that F_0 *has derivative* f_0 *at* a_0 *and* G *has derivative* g *at* a_0 *satisfying* $g(a_0) > 0$. *Then*

$$n^{1/3}(\hat{a}_n - a_0) \to^{\mathcal{L}} [f_0(a_0)\sqrt{g(a_0)}]^{-2/3} Z. \tag{2.6}$$

It may be verified that the conditions of Theorem 2.1 imply that $L(a_0)$ is Bayes risk.

The asymptotic distribution of the empirical prediction error $L_n(\hat{a}_n)$ now follows easily, and also the asymptotic distribution of $L(\hat{a}_n)$, the prediction error of the empirically best classifier. Let us summarize these asymptotics in a corollary.

Corollary 2.1. *As a by-product of the proof of Theorem* 2.1, *one finds that under its conditions,*

$$n^{2/3}(L_n(\hat{a}_n) - L_n(a_0)) \xrightarrow{\mathcal{L}} -\left(\frac{g(a_0)}{f_0(a_0)}\right)^{1/3} \max_t [W(t) - t^2]. \tag{2.7}$$

Moreover,

$$\begin{aligned} \sqrt{n}(L_n(\hat{a}_n) - L(\hat{a}_n)) &= \sqrt{n}(L_n(\hat{a}_n) - L(a_0)) + O_{\mathbf{P}}(n^{-1/6}) \\ &= \sqrt{n}(L_n(a_0) - L(a_0)) + O_{\mathbf{P}}(n^{-1/6}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, L(a_0)(1 - L(a_0))), \end{aligned} \tag{2.8}$$

and

$$n^{2/3}(L(\hat{a}_n) - L(a_0)) \to^{\mathcal{L}} \left(\frac{g(a_0)}{f_0(a_0)}\right)^{1/3} Z^2. \tag{2.9}$$

For practical purposes, result (2.8) is of interest, because it gives you for example a 95% upper bound for the prediction risk of the estimator. From a theoretical point of view, (2.9) is also of importance, as it shows that the prediction risk of $\hat{a}_n$ is very close to the Bayes risk, the difference being in order much smaller than $n^{-1/2}$.

We end this section by showing that the empirical risk minimizer is a nonparametric maximum likelihood estimator. Note first that F_0 is in whole or in part an unknown function. If it is known that $F_0 \in \Lambda$, with Λ being some given class of functions, one may estimate it using the maximum likelihood estimator

$$\hat{F}_n = \arg\max_{F \in \Lambda} \left\{ \sum_{Y_i = 1} \log F(X_i) + \sum_{Y_i = -1} \log(1 - F(X_i)) \right\}. \tag{2.10}$$

In the case where Λ is the class of all distribution functions, the (nonparametric) maximum likelihood estimator $\hat{F}_n$ is studied in Groeneboom and Wellner (1992).

One may call the approach where the classification rule is based on an estimator of F_0 the *function estimation* approach. It is known that function estimation is generally harder than classification (see Devroye, Györfi and Lugosi, 1996; Mammen and Tsybakov, 1999). We now show that in our case, the two problems actually give the same answer. Theorem 2.2 namely shows that the maximum likelihood estimator $\hat{a}_n^* = \hat{F}_n^{-1}(1/2)$ (defined as $\hat{a}_n^* = \inf\{x : \hat{F}_n(x) \geq \frac{1}{2}\}$) can be chosen to be equal to $\hat{a}_n$. In other words, minimizing the classification error using classifiers based on thresholds is the same as classifying $Y = 1$ when $2\hat{F}_n - 1 > 0$. The latter is the classifier which fits the data best (in, e.g., least squares sense).

Theorem 2.2. *Let $\hat{F}_n$ be the maximum likelihood estimator defined in* (2.10), *where Λ is the class of all distribution functions, and* $\hat{a}_n^* = \inf\{x : \hat{F}_n(x) \geq \frac{1}{2}\}$. *Then $\hat{a}_n^*$ minimizes $L_n(a)$, i.e., we may take $\hat{a}_n = \hat{a}_n^*$.*

3. Averaging classification rules and complexity

Let $\mathcal{H}$ be a set of base classifiers; each one is a map $h\colon \mathcal{X} \to \{-1, 1\}$. From $\mathcal{H}$, we form the set of all convex combinations $\mathcal{C} = \operatorname{conv}(\mathcal{H})$ (the convex hull of $\mathcal{H}$). We studied the following example in Section 2:

Example 3.1. Let $\mathcal{X} = \mathbb{R}$ and let

$$\mathcal{H} = \{h(x) = 2 1\{x \geq a\} - 1 : a \in \mathbb{R}\}.$$

Then it is easily seen that $\mathcal{C} = \operatorname{conv}(\mathcal{H})$ is a class of functions f which increase from -1 to 1. In fact, it is the class $\{2F - 1\colon F \in \Lambda\}$ where Λ is the set of all distribution functions, as in Theorem 2.2.

Now, in Example 3.1, the base classifiers are given by half-intervals, and half-intervals form a Vapnik–Chervonenkis, or VC, class.

Definition 3.1. A collection $\mathcal{A}$ of sets is called a *VC class* if for some finite V, and for all $n > V$ and all n distinct points $x_1, \ldots, x_n$, *not* all of the 2^n subsets of those points can be written in the form $A \cap \{x_1, \ldots, x_n\}$ for some $A \in \mathcal{A}$. Then V is called the *VC dimension* of $\mathcal{A}$.

It is not within the scope of this paper to give a full account of the concept of VC classes. More details can be found in the original paper by Vapnik and Chervonenkis (1971) and later work, e.g., van der Vaart and Wellner (1996) or van de Geer (2000) and the references therein. The concept plays an important role in classification, because when using base classifiers $\mathcal{H} = \{2 1_A - 1 : A \in \mathcal{A}\}$ given by a VC class $\mathcal{A}$, the empirical risk minimizer over $\mathcal{H}$ will not overfit the data. The VC dimension V of $\mathcal{H}$ is the size of the largest set of points that can be classified in all possible ways, by choosing a suitable $h \in \mathcal{H}$. So as soon as the size n of the data set is larger than V, a perfect fit is no longer possible and one starts learning from the data.

Let us now put the situation of Example 3.1 in a more general setting. The following terminology is used in, e.g., Dudley (1984, 1999).

Definition 3.2. The convex hull of a Vapnik-Chervonenkis class (VC class) is called a *VC hull* class. A class of functions $\mathcal{F}$ is called a *VC major* class if the collection of major sets $\big\{\{f > t\}, f \in \mathcal{F}, t \in \mathbb{R}\big\}$ is a VC class.

One may say that a VC class is relatively *simple* or *small.* In many classification problems, one starts out with base classifiers $\mathcal{H}$ given by a VC class, and then forms the convex hull $\mathcal{C}$. The classification rules based on $\mathcal{C}$ are now in general no longer given by a VC class, because Dudley (1984, 1999) shows that VC major implies VC hull, but not the other way around. In other words, by averaging classification rules, one may lose the VC property. Let us illustrate this in some examples.

Example 3.2 (continued). The class of majors of the sets of increasing functions $\mathcal{C}$ is again $\mathcal{H}$. So in this example, averaging leaves the complexity untouched.

Example 3.3. Let $\mathcal{X} = \mathbb{R}$ and let

$$\mathcal{H} = \{h(x) = b(2\mathrm{l}\{x \geq a\} - 1), b \in \{-1, 1\}, a \in \mathbb{R}\}. \tag{3.1}$$

Then $\mathcal{C} = \mathrm{conv}(\mathcal{H})$ is a class of functions f with $-1 \leq f \leq 1$ and with total variation $\mathrm{TV}(f) = \int |df| \leq 2$ (see also Example 3.5). However, it is easy to see that given n distinct points $x_1, \dots, x_n$, any of the 2^n subsets of those points can be written in the form $\{f > 0\} \cap \{x_1, \dots, x_n\}$ for some $f \in \mathcal{C}$. So $\{\{f > 0\} : f \in \mathcal{C}\}$ is far from being VC. Thus, when G is continuous, the empirical risk minimizer based on $\mathcal{C}$ will overfit the data.

Example 3.4. Let $\mathcal{X} = \mathbb{R}^2$ and let

$$\mathcal{H} = \{h(x) = 2\mathrm{l}\{x \geq a\} - 1 : a \in \mathbb{R}^2\},$$

where $\{x \geq a\} = \{x = (\xi_1, \xi_2) : \xi_1 \geq a_1, \xi_2 \geq a_2\}$, $a = (a_1, a_2)$. In this case $\mathcal{C} = \mathrm{conv}(\mathcal{H})$ is a rescaled set of two-dimensional distribution functions. One may check that $\{\{f > 0\} : f \in \mathcal{C}\}$ is a class of monotone sets (a set $A \subset \mathbb{R}^2$ is monotone if $(\xi_1, \xi_2) \in A$ implies $(\tilde{\xi}_1, \tilde{\xi}_2) \in A$ for all $\tilde{\xi}_1 \leq \xi_1$, $\tilde{\xi}_2 \leq \xi_2$). The class $\big\{\{f > 0\} : f \in \mathcal{C}\big\}$ is not VC: if $x_1, \dots, x_n$ are on a decreasing line $\{x = (\xi_1, \xi_2) : \xi_2 = g(\xi_1)\}$, g decreasing, then any subset of $x_1, \dots, x_n$ can be written in the form $\{f > 0\} \cap \{x_1, \dots, x_n\}$ for some $f \in \mathcal{C}$.

Examples 3.3 and 3.4 show that averaging may increase the complexity substantially. But so far, we have not quantified what we mean by complexity. It can be described in terms of covering numbers and entropy.

Definition 3.3. Let (T, τ) be a (subset of a) metric space (T is a set and τ is a metric on T). The *ϵ-covering number* $N(\epsilon, T, \tau)$ is the minimum number of balls with radius ϵ, necessary to cover T. The *ϵ-entropy* is $H(\epsilon, T, \tau) = \log N(\epsilon, T, \tau)$.

Let Q be some finite measure on $\mathcal{X}$. A class of bounded functions $\mathcal{F}$ on $\mathcal{X}$ can be considered as a subset of the metric space $L_p(Q)$, $1 \leq p \leq \infty$. We then write $N_p(\epsilon, \mathcal{F}, Q)$ ($H_p(\epsilon, \mathcal{F}, Q)$) for its covering number (entropy).

The following theorem from Ball and Pajor (1990) shows that a VC hull class is typically infinite dimensional. It derives the entropy of $\mathcal{C} = \text{conv}(\mathcal{H})$ from the covering number of $\mathcal{H}$, in the case $\mathcal{H}$ is of finite metric dimension d. Recall that a VC class with VC dimension V is of metric dimension $2(V-1)$ when considered as subset of $L_2(Q)$ (see, e.g., van der Vaart and Wellner, 1996).

Theorem 3.1. *Let Q be a finite measure. Suppose that for some positive constants c and d,*

$$N_2(\epsilon, \mathcal{H}, Q) \leq c\epsilon^{-d}, \quad \epsilon > 0. \tag{3.2}$$

Then for some constant A,

$$H_2(\epsilon, \mathcal{C}, Q) \leq A\epsilon^{-2d/(2+d)}, \quad \epsilon > 0. \tag{3.3}$$

Indeed, a VC class of (indicators of) sets satisfies (3.2), for some c and d (see, e.g., van der Vaart and Wellner, 1996). By (3.3) we see that the entropy of the convex hull, which is the logarithm of the covering number, has still a negative power of ϵ in the bound. It means that $\mathcal{C}$ can be infinite-dimensional.

To overcome the problem of overfitting, it is proposed to use the so-called *margin* (see, e.g., Schapire et al., 1998). For a decision rule f, the *margin* of an example (x, y) is $yf(x)$. Now, fix a value $0 \leq \theta \leq 1$. If for some decision function f, $Y_i f(X_i) > \theta$, we say that the example (X_i, Y_i) is θ-well classified, whereas if $0 \leq Y_i f(X_i) \leq \theta$ we consider it as a case where there is actually not enough evidence (when using a $(1+\theta)50\%$ majority rule). We call the latter case θ-undecided.

Instead of the empirical classification error, consider now the fraction number of θ-undecided examples

$$P_n(\mathbf{y}f \leq \theta). \tag{3.4}$$

We will study a uniform (in $f \in \mathcal{C}$) bound for

$$P(\mathbf{y}f \leq 0) - P_n(\mathbf{y}f \leq \theta).$$

(This bound will be non-asymptotic, in the sense that it holds for each n.) Such a bound is of interest, because if we take the classifier f which maximizes the $P_n(\mathbf{y}f > \theta)$, a uniform bound may be used to evaluate its prediction error (see Freund, Mansour and Schapire, 2001, and see also Section 4).

The following definition from Schapire et al. (1998) is now of relevance.

Definition 3.4. We say that $\hat{\mathcal{C}}$ forms an *ϵ-sloppy θ-cover* of $\mathcal{C}$ for the measure Q on $\mathcal{X}$, if for all $f \in \mathcal{C}$ there is a g in $\hat{\mathcal{C}}$ such that

$$Q(|f - g| > \theta) \leq \epsilon.$$

We let $\mathcal{N}(\epsilon, \theta, \mathcal{C}, Q)$ be the size of the smallest ϵ-sloppy θ-cover.

Here is a slight extension of Theorem 4 given in Schapire et al. (1998). Related results are in Kearns and Schapire (1990) and Bartlett (1998).

Theorem 3.2. *Let $\epsilon > 0$, $\theta > 0$. We have*

$$\text{(3.5)} \quad \mathbf{P}\Big(\sup_{f \in \mathcal{C}}\big(P(\mathbf{y}f \leq 0) - P_n(\mathbf{y}f \leq \theta)\big) > \epsilon\Big) \\ \leq 2\mathbf{E}\big(\mathcal{N}(\epsilon/8, \theta/2, \mathcal{C}, P_{2n})\big) \exp[-n\epsilon^2/32].$$

The theorem can be used to derive uniform bounds by calculating ϵ-sloppy θ-covers of $\mathcal{C} = \text{conv}(\mathcal{H})$ from the VC dimension of $\mathcal{H}$. However, a direct study of the properties of $\mathcal{C}$, without referring to the VC dimension of $\mathcal{H}$, may lead to better bounds.

We will apply a uniform bound using the entropy of $\mathcal{C}$, for the sup-norm on a finite subset of $\mathcal{X}$.

Definition 3.5. Let $\mathcal{S} \subset \mathcal{X}$. Let $\|\cdot\|_{\mathcal{S}}$ be the sup-norm on $\mathcal{S}$, i.e., $\|f\|_{\mathcal{S}} = \sup_{x \in \mathcal{S}} |f(x)|$. The entropy for the sup-norm of $\mathcal{C}$ restricted to $\mathcal{S}$ is

$$\text{(3.6)} \qquad H_\infty(\theta, \mathcal{C}, \mathcal{S}) = H(\theta, \mathcal{C}, \|\cdot\|_{\mathcal{S}}), \theta > 0.$$

Moreover, let

$$H_n(\theta, \mathcal{C}) = \sup\{H_\infty(\theta, \mathcal{C}, \{x_1, \ldots, x_n\}) : x_1, \ldots, x_n \in \mathcal{X}\}.$$

One easily sees that

$$\log \mathcal{N}(\epsilon, \theta, \mathcal{C}, P_n) \leq H_\infty(\theta, \mathcal{C}, \{X_1, \ldots, X_n\}) \\ \leq H_n(\theta, \mathcal{C}), \quad \theta > 0.$$

It is now not difficult to establish the following lemma.

Lemma 3.1. *We have for all $\delta > 0$,*

$$\mathbf{P}\Big(\sup_{f\in\mathcal{C}}(P(\mathbf{y}f \le 0) - P_n(\mathbf{y}f \le \theta)) > \epsilon_n(\theta)\Big) \le \delta, \tag{3.7}$$

where

$$\epsilon_n(\theta) = 8\sqrt{\frac{1}{n}\left[H_{2n}(\theta/2, \mathcal{C}) + \log\frac{2}{\delta}\right]}. \tag{3.8}$$

Example 3.5. Let $\mathcal{X} = \mathbb{R}$. Let $\mathrm{TV}(f) = \int |df|$ be the total variation of a function f on $\mathbb{R}$. Consider the class

$$\mathcal{C}_{\mathrm{BV}} = \{f : \mathcal{X} \to [-1,1] : TV(f) \le 2\}. \tag{3.9}$$

Recall that if $\mathrm{TV}(f) \le 2$, we may write f as $f = f_1 - f_2 + c$, where f_1 and f_2 are increasing functions with $-1 \le f_k \le 1$, $k = 1, 2$, and where c is a constant. The entropy for the sup-norm $H_\infty(\cdot, \mathcal{C}_{\mathrm{BV}}, \mathcal{X})$ is infinite (see Clements, 1963). The entropy $H_n(\cdot, \mathcal{C}_{\mathrm{BV}})$ can be bounded using, e.g., a result from van de Geer (2000).

Lemma 3.2. *We have*

$$H_n(\theta, \mathcal{C}_{\mathrm{BV}}) \le \frac{8}{\theta}\log\Big(n + \frac{4}{\theta}\Big), \quad \theta > 0. \tag{3.10}$$

Corollary 3.1. *For the class of functions of bounded variation $\mathcal{C}_{\mathrm{BV}}$, we have for all $\delta > 0$,*

$$\mathbf{P}\Big(\sup_{f\in\mathcal{C}_{\mathrm{BV}}}(P(\mathbf{y}f \le 0) - P_n(\mathbf{y}f \le \theta)) > \epsilon_n(\theta)\Big) \le \delta, \tag{3.11}$$

where

$$\epsilon_n(\theta) = 8\sqrt{\frac{1}{n}\left[\frac{16}{\theta}\log\Big(2n + \frac{8}{\theta}\Big) + \log\frac{2}{\delta}\right]}. \tag{3.12}$$

4. Maximizing the number of examples with large enough margin

Recall that the margin of the decision rule f at the example (x, y) is defined as $yf(x)$. Let $\mathcal{C}$ be a given class of classifiers and $\theta \ge 0$ be a given number. In this section, we propose to maximize, over $f \in \mathcal{C}$, the quantity

$$P_n(\mathbf{y}f \ge \theta),$$

or equivalently, to minimize the number of examples that are either misclassified or θ-undecided. As decision functions we take the class of functions

$\mathcal{C}_{\mathrm{BV}}$ with total variation bounded by M. It turns out that this reduces the optimization problem to finding optimal thresholds. We let

$$\mathcal{C}_{\mathrm{BV}} = \{f : [0,1] \to [-1,1] : \mathrm{TV}(f) \leq 2\},$$

that is, we take $M = 2$ to simplify the exposition, and to be sure that the base classifiers $h_a(x) = 2\mathrm{l}\{x \geq a\} - 1$ are in $\mathcal{C}_{\mathrm{BV}}$. Moreover, we let $\mathcal{X} = [0,1]$.

Example 3.3 shows that the majority votes based on $\mathcal{C}_{BV}$ is not VC. Indeed, it can be easily seen that when G is continuous, the empirical risk minimizer over $\mathcal{C}_{\mathrm{BV}}$ will yield a perfect fit, i.e., then overfitting occurs.

Lemma 4.1 illustrates that maximizing the number of examples with large enough margin, over decision rules $f \in \mathcal{C}_{\mathrm{BV}}$, instead of empirical risk minimization over the same $f \in \mathcal{C}_{\mathrm{BV}}$, is actually a kind of regularization method. In the empirical risk minimization problem, the original class $\mathcal{C}_{\mathrm{BV}}$ is replaced by a smoother version where the functions are not allowed to have too many sign changes.

Lemma 4.1. *Suppose* $0 < \theta \leq 1$ *and the class of base classifiers is*

$$\text{(4.1)}\quad \mathcal{H} = \left\{ h(x) = \sum_{k=1}^{K+1} b_k \mathrm{l}\{a_{k-1} \leq x < a_k\}, \right.$$
$$\left. 0 = a_0 < \cdots < a_K < a_{K+1} = 1, (b_1, \ldots, b_{K+1}) \in \{-1,1\}^{K+1} \right\},$$

with

$$\text{(4.2)}\qquad K = \left\lfloor \frac{1}{\theta} \right\rfloor.$$

Then

$$\min_{h \in \mathcal{H}} P_n(\mathbf{y}h \leq 0) = \min_{f \in \mathcal{C}_{\mathrm{BV}}} P_n(\mathbf{y}f < \theta).$$

We note that when $\theta = 1$, then in (4.2) $K = 1$, i.e., the class $\mathcal{H}$ in (4.1) is the class of base classifiers h_a based on only one threshold a. This classifier h_a is either of the form: classify $y = 1$ if $x \geq a$, or of the form: classify $y = 1$ if $x < a$. So we are back in the situation of Section 2, albeit that we now do not impose a priori the restriction that the large values of x (rather than the small values of x) make $y = 1$ more likely.

Recall now the bound we presented in Lemma 3.2. Lemma 4.1 shows that looking directly at the maximizer of the margin asks for a uniform bound, not so much over $f \in \mathcal{C}_{\mathrm{BV}}$, but rather over all base classifiers with at most $K = \lfloor 1/\theta \rfloor$ thresholds. Lemma 4.1 also indicates that the choice of the parameter θ in the margin is related to the amount of regularization that one is imposing. One may consider a data dependent choice for θ.

Here, an extension of Lemma 3.2 can be valuable. A uniform in θ version of (3.7) (which can be derived by considering a finite grid of values $\theta \in \{1/n, 2/n, \ldots, 1\}$) suggests to maximize the margin over all $f \in \mathcal{C}$ and $\theta \geq 1/n$, but applying the penalty $\bar{\epsilon}_n(\theta)$ (apart from logarithmic factors) of the form $\bar{\epsilon}_n(\theta) \approx \sqrt{1/(n\theta)}$. In Lemma 4.1, this corresponds to empirical risk minimization with K thresholds and a penalty of the form $\approx \sqrt{K/n}$ (again apart from logarithmic factors).

5. Conclusion

We studied the problem of estimating a classification rule from a set of training examples. In the simple case of classifiers with just a threshold, i.e.,

$$\mathcal{H} = \{h_a(x) = 2\mathrm{l}\{x \geq a\} - 1 : \; a \in \mathbb{R}\}, \tag{5.1}$$

we briefly considered some parametric models, where the maximum likelihood estimator of a_0 (the value of a that gives minimal prediction error) generally converges with rate $\sqrt{n}$. We also considered the nonparametric model, and the estimator $\hat{a}_n$, defined as the minimizer of the misclassification error (error in the sample). In the case that $F_0 = P(Y = 1 \mid X = \cdot)$ is known to be a distribution function on $\mathbb{R}$, we proved that the empirical risk minimizer $\hat{a}_n$ can be chosen equal to the nonparametric maximum likelihood estimator. In the nonparametric set up, the convergence is slower, namely $n^{1/3}$. But this is the price to pay for not assuming a parametric model but just using the data.

The ideas of Section 2 can be extended to a higher-dimensional instance space, say $\mathcal{X} = \mathbb{R}^d$, where instead of thresholds, one considers separating hyperplanes $\{x : a^T x \leq 1\}$, $a \in \mathbb{R}^d$. This is related to assuming a single index model for the regression of Y on X, i.e.,

$$P(Y = 1|X = x) = F_0(a_0^T x),$$

where a_0 is an unknown parameter, and F_0 is an unknown (monotone) function. Minimizing the empirical classification error over the base classifiers $h_a(x) = 2\mathrm{l}\{a^T x \leq 1\} - 1$ will again lead to cube root asymptotics.

We compared the complexity of the set $\mathcal{H}$ with its convex hull $\mathcal{C}$ in some particular cases. Using averaged classification rules, the difference between theoretical and empirical errors can be very large. We considered the case of a class of functions of bounded variation as an example. Here, the empirical risk minimizer will generally overfit the data.

We also considered uniform bounds in the spirit of Schapire et al. (1998), which are, e.g., of interest when maximizing the margin.

In the particular case of bounded variation functions, we showed that maximizing the number of examples with large enough margin is equivalent to empirical risk minimization using the base classifiers which have K

thresholds where $K = \lfloor 1/\theta \rfloor$. This puts the idea in the light of complexity regularization.

As explained in Schapire et al., the AdaBoost algorithm tends to increase margins. Directly maximizing the number of examples with large margin can sometimes be of too large computational complexity, in which case the AdaBoost algorithm is a good alternative. Another alternative is density estimation or curve estimation. For example, using averaged classifiers, one may also first choose the one which fits the data best in least squares sense, and then use the majority vote rule. When considering averaged classification rules in the class $\mathcal{C}_{\mathrm{BV}}$, the rate of convergence for the least squares estimator (when also $H_0 = 2F_0 - 1$ is in $\mathcal{C}_{\mathrm{BV}}$) is again of order $n^{1/3}$ (see, e.g., van de Geer, 2000). This indicates that the rate of convergence of the prediction error of corresponding classification rule to Bayes prediction error, is also of order $n^{1/3}$. The latter is much slower than the rate of order $n^{2/3}$ we found in (2.9) of Corollary 2.1. This would confirm indeed that although function estimation can be computationally easier than classification, it is theoretically much harder.

APPENDIX

Proofs

Proof of Lemma 2.1. Since $\{\{x : x \geq a\} : a \in \mathbb{R}\}$ is a VC class, we know that

$$\sup_{a\in\mathbb{R}} |L_n(a) - L(a)| \to 0, \quad \text{a.s.}$$

(see, e.g., Pollard, 1984). So,

$$0 \leq L_n(a_0) - L_n(\hat{a}_n) = L(a_0) - L(\hat{a}_n) + o(1) \leq o(1), \quad \text{a.s.}$$

so that $L(\hat{a}_n) \to L(a_0)$, a.s.. By the uniqueness of a_0 and the continuity of L, this gives $\hat{a}_n \to a_0$, a.s.. □

Proof of Theorem 2.1. We have

$$\begin{aligned} L(a) &= P(\mathbf{y}h_a < 0) \\ &= \int_{x<a} P(Y = 1 \mid X = x)\, dG(x) + \int_{x\geq a} P(Y = -1 \mid X = x)\, dG(x) \\ &= \int_{x\leq a} (2F_0(x) - 1)\, dG(x) + \int \bigl(1 - F_0(x)\bigr)\, dG(x). \end{aligned}$$

Differentiating this with respect to a and putting the result equal to zero yields

$$(2F_0(a_0) - 1)g(a_0) = 0.$$

By straightforward calculations, we find

$$P(\mathbf{y}h_a \le 0) - P(\mathbf{y}h_{a_0} \le 0) = (a - a_0)^2 f_0(a_0)g(a_0) + o(|a - a_0|^2).$$

Consider now the process

$$\mathcal{W}_n(\tau) = \sqrt{n^{1/3}}\{\nu_n(\mathbf{y}h_{a_0+n^{-1/3}\tau} \le 0) - \nu_n(\mathbf{y}h_{a_0} \le 0)\}$$

where

$$\nu_n(\mathbf{y}h_a \le 0) = \sqrt{n}[P_n(\mathbf{y}h_a \le 0) - P(\mathbf{y}h_a \le 0)].$$

The covariance structure of ν_n can be calculated easily. We have for $a_0 < a_1 \le a_2$,

$$\begin{aligned}
&\mathrm{cov}([\mathrm{l}\{Yh_{a_1}(X)\le 0\} - \mathrm{l}\{Yh_{a_0}(X)! \le 0\}], [\mathrm{l}\{Yh_{a_2}(X)\le 0\} - \mathrm{l}\{Yh_{a_0}(X)\le 0\}]) \\
&\quad = P(a_0 < X \le a_1,\ Y = 1) + P(a_0 < X \le a_1, Y = -1) \\
&\qquad - [P(\mathbf{y}h_{a_1} \le 0) - P(\mathbf{y}h_{a_0} \le 0)][P(\mathbf{y}h_{a_2} \le 0) - P(\mathbf{y}h_{a_0} \le 0)] \\
&\quad = \int_{a_0}^{a_1} F_0(x)\,dG(x) + \int_{a_0}^{a_1} (1 - F_0(x))\,dG(x) + o(|a_1 - a_0|) \\
&\quad = \int_{a_0}^{a_1} dG(x) + o(|a_1 - a_0|) = (a_1 - a_0)g(a_0) + o(|a_1 - a_0|).
\end{aligned}$$

Similar expressions can be found for $a_2 \le a_1 < a_0$. For $a_2 < a_0 < a_1$ (or $a_1 < a_0 < a_2$) the covariance is $o(|a_1 - a_2|)$. Thus, for $0 < \tau_1 \le \tau_2$ or $\tau_2 \le \tau_1 < 0$, one has $\mathrm{cov}(\mathcal{W}_n(\tau_1), \mathcal{W}_n(\tau_2)) = \tau_1 g(a_0) + o(1)$, and otherwise the covariance is $o(1)$.

Invoking the theory of Kim and Pollard (1990), we find

$$\mathcal{W}_n(\tau) \xrightarrow{\mathcal{L}} \mathcal{W}(\tau), \quad \tau \in \mathbb{R},$$

where the convergence in distribution is to be understood as convergence in distribution of $\mathcal{W}_n$ as a stochastic process, and where

$$\mathcal{W}(\tau) = \sqrt{g(a_0)}W(\tau),$$

with W a two-sided Brownian motion.

Let $\hat{\tau}_n = n^{1/3}(\hat{a}_n - a_0)$. Applying the arguments in Kim and Pollard (1990), we obtain

$$\begin{aligned}
\hat{\tau}_n &= \arg\min_{\tau} P_n(\mathbf{y}h_{a_0+n^{-1/3}\tau} \le 0) \\
&= \arg\max_{\tau}[-\mathcal{W}_n(\tau) - f_0(a_0)g(a_0)\tau^2 + o(|\tau|^2)] \\
&\qquad \xrightarrow{\mathcal{L}} \arg\max_{\tau}[\sqrt{g(a_0)}W(\tau) - f_0(a_0)g(a_0)\tau^2] \\
&\qquad\qquad = \arg\max_{\tau}[W(\tau) - f_0(a_0)\sqrt{g(a_0)}\tau^2].
\end{aligned}$$

Now, use the time change $t = \tau b^{2/3}$, where $b = f_0(a_0)\sqrt{g(a_0)}$. Then $W(\tau) = W(t/b^{2/3}) =^{\mathcal{L}} W(t)/b^{1/3}$. So

$$W(\tau) - b\tau^2 \overset{\mathcal{L}}{=} \frac{W(t)}{b^{1/3}} - \frac{bt^2}{b^{4/3}} = \frac{W(t) - t^2}{b^{1/3}}.$$

Thus,

$$\arg\max_{\tau}[W(\tau) - b\tau^2] \overset{\mathcal{L}}{=} b^{-2/3} \arg\max_{t}[W(t) - t^2] = b^{-2/3} Z,$$

where

$$Z = \arg\max_{t}[W(t) - t^2]. \qquad \square$$

Proof of Theorem 2.2. Recall that $\hat{F}_n$ maximizes

$$\sum_{Y_i=1} \log F(X_i) + \sum_{Y_i=-1} \log\bigl(1 - F(X_i)\bigr)$$

over all $F \in \Lambda$, where Λ is the class of all distribution functions. This is equivalent to minimizing

$$P_n(\mathbf{y} - H)^2 = \frac{1}{n}\sum_{i=1}^{n}\bigl(Y_i - H(X_i)\bigr)^2$$

over $H = 2F - 1$, $F \in \Lambda$ (see Robertson et al., 1988).

Let $X_{(1)} \le \cdots \le X_{(n)}$ be the order statistics and let $Y_{(1)}, \ldots, Y_{(n)}$ be the corresponding labels. For ease of notation, let us—only in this proof—simply write $X_1 \le \cdots \le X_n$ for the order statistics, with corresponding labels $Y_1, \ldots, Y_n$.

It is known that $\hat{F}_n$ only jumps at the instances $X_1, \ldots, X_n$ (see Groeneboom and Wellner, 1992). So

$$\hat{F}_n(x) = \sum_{j=1}^{n} \hat{\alpha}_j \, \mathrm{l}\{x \ge X_j\},$$

where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n) \in S$, and where S is the simplex

$$S = \left\{\alpha = (\alpha_1, \ldots, \alpha_n) : \alpha_j \ge 0, j = 1, \ldots, n, \sum_{j=1}^{n} \alpha_j = 1\right\}.$$

Define now $h_j(x) = 2\,\mathrm{l}\{x \ge X_j\} - 1$. This is the base classifier using threshold X_j. Then $\hat{H}_n = 2\hat{F}_n - 1 = \sum_{j=1}^{n} \hat{\alpha}_j h_j$. Thus $\hat{\alpha}$ minimizes

$$P_n(\mathbf{y} - \sum_{j=1}^{n} \alpha_j h_j)^2$$

over all $\alpha \in S$.

It is moreover not difficult to see that $\hat{a}_n^* = X_{\hat{k}^*}$, where $\hat{k}^*$ minimizes

$$P_n(h_j - \hat{H}_n)^2$$

over $j \in \{1, \ldots, n\}$, and that $\hat{a}_n$ can be chosen in $\{X_1, \ldots, X_n\}$, in which case $\hat{a}_n = X_{\hat{k}}$, where $\hat{k}$ minimizes

$$P_n(\mathbf{y} - h_j)^2$$

over $j \in \{1, \ldots, n\}$.

Now, since $\sum_{j=1}^n \alpha_j = 1$ for $\alpha \in S$, we have

$$P_n\Big(\mathbf{y} - \sum_{j=1}^n \alpha_j h_j\Big)^2 = P_n\Big(\sum_{j=1}^n \alpha_j(\mathbf{y} - h_j)\Big)^2.$$

The derivative of the right hand side with respect to α_k is

$$2P_n\Big(\sum_{j=1}^n \alpha_j(\mathbf{y} - h_j)(\mathbf{y} - h_k)\Big),$$

so that $\hat{H}_n$ is given by

$$2P_n\big((\mathbf{y} - \hat{H}_n)(\mathbf{y} - h_k)\big)\begin{cases} \geq \lambda & \\ = \lambda & \text{if } \hat{\alpha}_k > 0, \end{cases} \quad k = 1, \ldots, n$$

(Rockafellar, 1970). Here λ is the Lagrange multiplier for the restriction that $\sum_{j=1}^n \alpha_j = 1$. We know moreover that $\hat{F}_n$ jumps at $X_{\hat{k}^*}$ (by the definition of $\hat{F}_n^{-1}(1/2)$), i.e., $\hat{\alpha}_{\hat{k}^*} > 0$.

We may conclude for each j,

$$\begin{aligned} P_n(\mathbf{y} - h_j)^2 &= P_n(h_j - \hat{H}_n)^2 - P_n(\mathbf{y} - \hat{H}_n)^2 + 2P_n\big((\mathbf{y} - \hat{H}_n)(\mathbf{y} - h_j)\big) \\ &\geq P_n(h_j - \hat{H}_n)^2 - P_n(\mathbf{y} - \hat{H}_n)^2 + \lambda \\ &\geq \min_{k=1,\ldots,n} P_n(h_k - \hat{H}_n)^2 - P_n(\mathbf{y} - \hat{H}_n)^2 + \lambda \\ &= P_n(h_{\hat{k}^*} - \hat{H}_n)^2 - P_n(\mathbf{y} - \hat{H}_n)^2 + \lambda. \end{aligned}$$

Write this as

$$P_n(\mathbf{y} - h_j)^2 \geq C, \quad j = 1, \ldots, n,$$

where

$$C = P_n(h_{\hat{k}^*} - \hat{H}_n)^2 - P_n(\mathbf{y} - \hat{H}_n)^2 + \lambda.$$

It follows that

$$P_n(\mathbf{y} - h_{\hat{k}})^2 \geq C.$$

For the case $j = \hat{k}^*$, we find

$$P_n(\mathbf{y} - h_{\hat{k}^*})^2 = C.$$

In other words, $\hat{k}^*$ is a minimizer of $P_n(\mathbf{y} - h_j)^2$ over $j \in \{1, \dots, n\}$. □

Proof of Lemma 3.1. From Theorem 3.2, we know that,

$$\begin{aligned}\mathbf{P}\Big(\sup_{f \in \mathcal{C}}\big(P(\mathbf{y}f \le 0) - P_n(\mathbf{y}f \le \theta)\big) > \epsilon_n(\theta)\Big) &\le 2\exp[H_{2n}(\theta/2, \mathcal{C}) - n\epsilon_n^2(\theta)/32] \\ &\le 2\exp[-n\epsilon_n(\theta)^2/64] \le \delta,\end{aligned}$$

since $n\epsilon_n^2(\theta)/32 \ge 2H_{2n}(\theta/2, \mathcal{C})$ and $n\epsilon_n^2(\theta)/64 \ge \log(2/\delta)$. □

Proof of Lemma 3.2. Let $\mathcal{F}$ be the class of all monotone functions f on $\mathbb{R}$ with $0 \le f \le 1$. It is shown in van de Geer (2000, Lemma 2.2) that

$$H_n(\theta, \mathcal{F}) \le \frac{1}{\theta}\log\left(n + \frac{1}{\theta}\right), \quad \theta > 0.$$

The result easily follows from this. □

Proof of Lemma 4.1. We first show that for any $f \in \mathcal{C}_{\mathrm{BV}}$ there exists an $h \in \mathcal{H}$ such that

$$\text{(A.1)} \qquad P_n(\mathbf{y}h \le 0) \le P_n(\mathbf{y}f < \theta).$$

It is clear that we may restrict attention to all classifiers $f \in \mathcal{C}_{\mathrm{BV}}$ with $|f(x)| \le \theta$, and that maximizing $P_n(\mathbf{y}f \ge \theta)$ amounts to maximizing the number of examples (X_i, Y_i) with $f(X_i) = \theta Y_i$. Now, suppose $0 \le a_0 < a_1 < \cdots < a_K \le 1$, and that $|f(a_k) - f(a_{k-1})| = 2\theta$, $k = 1, \dots, K$. Then obviously, if $f \in \mathcal{C}_{\mathrm{BV}}$,

$$2 \ge TV(f) \ge \sum_{k=1}^{K} |f(a_k) - f(a_{k-1})| = 2\theta K,$$

so we must have $K \le 1/\theta$.

Suppose $|f(x)| < \theta$ for all x in a certain interval, say $(a, b) \subset [0, 1]$. Then according to f, instances X_i that lie in (a, b) do not contribute to the number of θ-well classified examples. Thus, choosing f monotone on $[a, b]$ will give the smallest total variation and has no impact on the number of θ-well classified examples.

Since according to f, X_i is θ-undecided if $|f(X_i)| < \theta$, choosing $|f(X_i)| = \theta$, for all $i = 1, \dots, n$ is optimal. This is indeed possible by the following argument. Suppose f is monotone on $[a, b]$. The total variation of f is then not changed if we change the values of $f(x)$ for $x \in (a,b)$, as long as we do not disturb the monotonicity. Therefore, we may choose $f(x) = f(a)$ for all $x \in [a,c)$ and $f(x) = f(b)$ for all $x \in [c,b]$. It follows from (A.1) that

$$\min_{h\in\mathcal{H}} P_n(\mathbf{y}h \le 0) \le \min_{f\in\mathcal{C}_{\mathrm{BV}}} P_n(\mathbf{y}f < \theta).$$

On the other hand, $\mathcal{H}_\theta = \{\theta h : h \in \mathcal{H}\} \subset \mathcal{C}_{\mathrm{BV}}$ so that

$$\begin{aligned}\min_{f\in\mathcal{C}_{\mathrm{BV}}} P_n(\mathbf{y}f < \theta) &\le \min_{f\in\mathcal{H}_\theta} P_n(\mathbf{y}f < \theta)\\ &= \min_{h\in\mathcal{H}} P_n(\mathbf{y}h < 1) = \min_{h\in\mathcal{H}} P_n(\mathbf{y}h \le 0). \qquad \square\end{aligned}$$

REFERENCES

Ball, K. and Pajor, A. (1990). The entropy of convex bodies with "few" extreme points. *Geometry of Banach Spaces* (P.F.X. Müller and W. Schachermayer, eds.). London Math. Soc. Lecture Note Ser., vol. 158. Cambridge Univ. Press, Cambridge, pp. 25–32.

Bartlett, P.L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Inform. Theory* **44**, 525–536.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.* **30**, 927–961.

Clements, G.F. (1963). Entropies of sets of functions of bounded variation. *Canad. J. Math.* **15**, 422–432.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Appl. Math., vol. 31. Springer-Verlag, New York.

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis.* Wiley, New York.

Dudley, R.M. (1984). A course on empirical processes. *École d'Été de Probabilités de Saint-Flour XII*—1982 (P.L. Hennequin, ed.). Lecture Notes in Math., vol. 1097, Springer-Verlag, New York, pp. 1—142.

Dudley, R.M. (1999). *Uniform Central Limit Theorems.* Cambridge Stud. Adv. Math., vol. 63, Cambridge Univ. Press, Cambridge

Freund, Y., Mansour, Y. and Schapire, R. (2001). Why averaging classifiers can protect against overfitting. *Artificial Intelligence and Statistics 2001: Proceedings of the Eighth Internatinoal Workshop* (T. Jaakkola and T. Richardson, eds.) Morgan Kaufmann, San Francisco, CA.

Freund, Y. and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Sytem Sci.* **55**, 119–139.

Groeneboom, P. and Wellner, J. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* DMV Seminar, vol. 19, Birkhäuser Verlag, Basel.

Kearns, M.J. and Schapire, R.E. (1990). Efficient distribution-free learning of probabilistic concepts. In: *Proc. 31st Symp. on Foundations of Computer Science*, IEEE Computer Soc. Press, Los Alamitos, CA, pp. 382–391.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191–219.

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

Rockafellar, R.T. (1970). *Convex Analysis.* Princeton Math. Ser., vol. 28, Princeton Univ. Press, Princeton, NJ.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference.* Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., Wiley, New York.

Schapire, R., Freund, Y., P. Bartlett P. and Lee, W.S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.* **5**, 1651–1686.

van de Geer, S.A. (2000). *Empirical Processes in M-Estimation.* Empirical Processes, Cambridge Univ. Press, Cambridge.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics.* Springer Ser. Statist., Springer, New York.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

Vapnik, V.N. and Chervonenkis, A.Ja. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16**, 264–280.

Leila Mohammadi
Mathematical Institute
Leiden University
P.O. Box 9512
2300 RA Leiden
The Netherlands
leila@math.leidenuniv.nl

Sara van de Geer
Mathematical Institute
Leiden University
P.O. Box 9512
2300 RA Leiden
The Netherlands
geer@math.leidenuniv.nl