Journal of Machine Learning Research 4 (2003) ??–?? Submitted ??; Published ??

On the rate of convergence of regularized boosting classifiers

Gilles Blanchard BLANCHAR@MATH.U-PSUD.FR
CNRS Laboratoire de Mathématiques
Université Paris-Sud
Bâtiment 425
91405 Orsay Cedex, France

Gábor Lugosi LUGOSI@UPF.ES
Department of Economics,
Pompeu Fabra University,
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain

Nicolas Vayatis VAYATIS@CCR.JUSSIEU.FR
Université Paris 6–Pierre et Marie Curie
Laboratoire de Probabilités et Modèles Aléatoires
4, place Jussieu - Boite courrier 188
75252 Paris cedex 05, France

Editor: Thore Graepel and Ralf Herbrich

Abstract

A regularized boosting method is introduced, for which regularization is obtained through a penalization function. It is shown through oracle inequalities that this method is model adaptive. The rate of convergence of the probability of misclassification is investigated. It is shown that for a quite large class of distributions, the probability of error converges to the Bayes risk at a rate faster than $n^{-(V+2)/(4(V+1))}$ where V is the VC dimension of the "base" class whose elements are combined by boosting methods to obtain an aggregated classifier. The dimension-independent nature of the rates may partially explain the good behavior of these methods in practical problems. Under Tsybakov's noise condition the rate of convergence is even faster. We investigate the conditions necessary to obtain such rates for different base classes. The special case of boosting using decision stumps is studied in detail. We characterize the class of classifiers realizable by aggregating decision stumps. It is shown that some versions of boosting work especially well in high-dimensional logistic additive models. It appears that adding a limited labelling noise to the training data may in certain cases improve the convergence, as has been also suggested by other authors.

Keywords: classification, boosting, consistency, rates of convergence, decision stumps

1. Introduction

The statistical and learning-theoretical literature has witnessed a recent explosion of theoretical work attempting to explain the often surprisingly good behavior of classification methods related to boosting and other algorithms based on weighted voting schemes. Boosting algorithms, originally introduced by Freund and Schapire (see Freund (1995), Freund and Schapire (1997), and Schapire (1990)), are based on an adaptive aggregation of simple classifiers contained in a small "base class". Originally, theoretical analysis was based on the observation that AdaBoost and related methods tend to produce large-margin classifiers in a certain sense (see Schapire, Freund, Bartlett, and Lee (1998), Koltchinskii and Panchenko (2002)). This view was complemented by Breiman's observation (Breiman, 1998) that boosting performs gradient descent optimization of an empirical cost function different from the number of misclassified samples, see also Mason, Baxter, Bartlett, and Frean (1999), Collins, Schapire, and Singer (2000), Friedman, Hastie, and Tibshirani (2000). Based on this new view, various versions of boosting algorithms have been shown to be consistent in different settings, see Breiman (2000), Bühlmann and Yu (2003), Jiang (2003), Lugosi and Vayatis (2003), Mannor and Meir (2001), Mannor, Meir, and Zhang (2002), Zhang (2003).

The purpose of this paper is a deeper investigation of the convergence of the probability of error of regularized boosting classifiers by deriving bounds for the rate of convergence. The main point is the introduction of a boosting procedure with regularization by a penalty function depending on the ℓ_1 norm of the boosting coefficients. The main result of the paper is an oracle inequality showing that this procedure is model adaptive, and stating in particular that the rate of convergence for the probability of error of the associated classification rule converges to that of the Bayes classifier at a dimension-independent rate faster than $n^{-(V+2)/(4(V+1))}$ – where V is the VC dimension of the base classifiers – for a large class of distributions. The class of distributions for which this rate holds is defined in terms of properties of the function f^* minimizing the expected cost function. If the base classifier set if sufficiently rich, the class turns out to be quite large. The analysis also points out a curious behavior of boosting methods: in some cases the rate of convergence can be speeded up by adding (limited) random noise to the data!

We also note that under some additional natural assumption on the distribution, considered by Tsybakov (2003), Nedelec and Massart (2003), and Bartlett, Jordan, and McAuliffe (2003), the rate of convergence may be even faster.

One of the main objectives of this paper is to better understand the behavior of boosting methods using decision stumps. This special case is studied in detail first in a simple one-dimensional setting and then in general. We characterize the class of classifiers realizable by aggregating decision stumps. It is shown that some versions of boosting work especially well in high-dimensional logistic additive models in that they do not suffer from the "curse of dimensionality".

The paper is organized as follows. In Section 2.1 our mathematical model of boosting classification is described. The main results are stated in Section 3. In particular, rates of convergence of a regularized boosting classifier are established under certain assumptions on the distribution. The main result, Corollary 7 is then discussed in subsequent sections in which various concrete examples are considered. Our introductory example is a one-dimensional problem in which "decision stumps" are used as a base class. This example, detailed in Section 4, sheds some light on the nature of the assumption guaranteeing a fast rate of convergence. Also, this example reveals some interesting and surprising phenomena inherent in boosting classifiers. In particular, it is pointed out that adding random noise to the labels in the data may improve the performance of regularized boosting. In Section 5 we investigate, in detail, the example of boosting using decision stumps in higher-dimensional problems. We point out that a sufficient condition for fast rates of convergence is that the conditional probability function belongs to a logistic additive model, verifying the observation of Friedman, Hastie, and Tibshirani (2000) that boosting using decision stumps works especially well in logistic additive models. We point out (see Corollary 12) that regularized boosting using the logistic cost function and decision stumps has a remarkably good behavior under the additive logistic model in high dimensional problems. We also characterize the class of classifiers that can be realized by a convex combination of decision stumps. In Section 6 several important special cases of base classes are studied briefly. These classes are rich enough so that they allow universally consistent classification and have a fast rate of convergence for a large classes of distributions. We also emphasize the scale and rotation invariance of boosting methods based on several of these base classes. The proof of Theorem 1 is given in Section 7.

2. Setup

2.1 Problem formulation and notation

The binary classification problem we consider is described as follows. Let (X,Y) be a pair of random variables taking values in $\mathcal{X} \times \{-1,1\}$ where $\mathcal{X}$ is a measurable feature space. Given a training data of n independent, identically distributed observation/label pairs $D_n = (X_1,Y_1),\ldots,(X_n,Y_n)$, having the same distribution as (X,Y), the problem is to design a classifier $g_n : \mathcal{X} \to \{-1,1\}$ which assigns a label to each possible value of the observation. The loss of g_n is measured by

$$L(g_n) = \mathbb{P}\left[g_n(X) \neq Y | D_n\right] \ .$$

The minimal possible probability of error is the Bayes risk, denoted by

$$L^* = \inf_g L(g) = \mathbb{E}\min(\eta(X), 1-\eta(X))$$

where the infimum is taken over all measurable classifiers $g : \mathcal{X} \to \{-1,1\}$ and $\eta(x) = \mathbb{P}[Y=1|X=x]$ denotes the posterior probability function. The infimum is achieved by the Bayes classifier $g^*(x) = \mathbb{I}_{[\eta(x)>1/2]} - \mathbb{I}_{[\eta(x)\leq 1/2]}$ (where $\mathbb{I}$ denotes the indicator function).

The voting classifiers studied in this paper combine their decisions based on a weighted majority vote of classifiers from a base class of classifiers $\mathcal{C}$, whose elements $g : \mathcal{X} \to \{-1,1\}$ of $\mathcal{C}$ are called the *base classifiers*. We denote the VC dimension of $\mathcal{C}$ by V and assume it is finite. For simplicity we assume that $\mathcal{C}$ is symmetric in the sense that for any $g \in \mathcal{C}$ we also have $-g \in \mathcal{C}$. (This is equivalent to allowing negative weights in the voting schemes.)

We define by $\mathcal{F}_\lambda$ the class of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ obtained as nonnegative linear combinations of the classifiers in $\mathcal{C}$ with the sum of the coefficients equal to $\lambda > 0$:

$$\mathcal{F}_\lambda = \left\{ f(x) = \sum_{j=1}^N w_j g_j(x) \ : \ N \in \mathbb{N};\ \forall 1 \leq j \leq N\,,\ g_j \in \mathcal{C}\,,\ w_j \geq 0\,;\ \sum_{j=1}^N w_j = \lambda \right\} .$$

Note that the symmetry of $\mathcal{C}$ implies that $\mathcal{F}_{\lambda_1} \subset \mathcal{F}_{\lambda_2}$ whenever $\lambda_1 < \lambda_2$. Each $f \in \mathcal{F}_\lambda$ defines a classifier g_f by

$$g_f(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ -1 & \text{otherwise.} \end{cases}$$

To simplify notation, we write $L(f) = L(g_f) = \mathbb{P}[g_f(X) \neq Y]$ and

$$\widehat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[g_f(X_i)\neq Y_i]} \ .$$

As mentioned in the introduction, boosting methods may be viewed as iterative methods for optimizing a convex empirical cost function. The approach taken in this paper is similar to that of Lugosi and Vayatis (2003) in that we ignore the dynamics of the optimization procedure and simply consider minimizers of an empirical cost function.

To this end, let $\phi : \mathbb{R} \to \mathbb{R}^+$ be a twice differentiable, strictly increasing and strictly convex function such that $\phi(0) = 1$ and $\lim_{x\to-\infty}\phi(x) = 0$ which we call the *cost function*. The corresponding risk functional and empirical risk functional are defined by

$$A(f) = \mathbb{E}\phi(-Y f(X)) \quad \text{and} \quad A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(-Y_i f(X_i)) \ .$$

We recall from Lugosi and Vayatis (2003) the simple fact that there exists an extended-real-valued function f^* minimizing $A(f)$ over all measurable function, given by

$$f^*(x) = \arg\min_{\alpha\in\mathbb{R}} \{\eta(x)\phi(-\alpha) + (1-\eta(x))\phi(\alpha)\} \ .$$

We write $A^* = A(f^*) = \inf_f A(f)$.

The estimates we consider take the form

$$\widehat{f}_n^{\lambda} = \arg\min_{f\in\mathcal{F}_\lambda} A_n(f) \ .$$

(Note that the minimum may not be achieved in $\mathcal{F}_\lambda$. However, to simplify the arguments we implicitly assume that the minimum in fact exists. All proofs may be adjusted, in a straightforward way, to handle appropriate approximate minimizers of the empirical cost functional.) As argued in Lugosi and Vayatis (2003), the parameter λ may be regarded as a smoothing parameter. Large values of λ improve the approximation properties of the class $\mathcal{F}_\lambda$ at the price of making the estimation problem more difficult.

The estimators considered in this paper use a value of λ chosen empirically, by minimizing a penalized value of the empirical cost $A_n(\widehat{f}_n^{\lambda})$. To this end, consider a sequence of real numbers $(\lambda_k)_{k\in\mathbb{N}}$ increasing to $+\infty$ and let $\zeta : \mathbb{R}^+ \to \mathbb{R}^+$ be a so-called penalty (or regularization) function. Define the penalized estimator by

$$\widehat{f}_n = \arg\min_{k\geq 1}\{A_n(\widehat{f}_n^{\lambda_k}) + \zeta(\lambda_k)\} \ . \tag{1}$$

The role of the penalty is to compensate for overfitting which helps find an adequate value of λ_k. For larger values of λ_k the class $\mathcal{F}_{\lambda_k}$ is larger, and therefore $\zeta(\lambda_k)$ should be larger as well. By a careful choice of the penalty, specified in Theorem 1 below, one may find a close-to-optimal balance between estimation and approximation errors.

The main purpose of this paper is to investigate the probability of error $L(\widehat{f}_n)$ of the classifier $g_{\widehat{f}_n}$ induced by the penalized estimator. The decision function $g_{\widehat{f}_n}$ may be regarded as a regularized boosting classifier where the regularization parameter λ controls the sum of the weights of the aggregated classifiers, and it is chosen by minimizing a penalized value of the empirical cost function.

Remark 1. Choosing λ in a countable set is done here to simplify the proof of the oracle inequality in Theorem 1; the minimum over $\lambda \in \mathbb{R}^+$ could also be considered with similar results up to minor additional terms in the penalty.

Remark 2. For simplicity we assume that the base class $\mathcal{C}$ contains binary-valued functions and that the class has a finite VC dimension. However, the results may be generalized in a straightforward way to the case when $\mathcal{C}$ contains real-valued functions taking values in $[-1, 1]$. The assumption of finite VC dimension may be replaced by the more general assumption that the covering numbers $\mathcal{N}(\varepsilon, \mathcal{C}, L_2(Q))$ are bounded by $c\varepsilon^{-V}$ for some positive constants c and V for any probability distribution Q.

Remark 3. (COMPUTATIONAL ISSUES.) To compute the penalized estimator $\widehat{f}_n$ in practice, one may proceed by computing, for each λ_k, the minimizer $\widehat{f}_n^{\lambda_k}$ of the empirical cost function. This may be done using iterative boosting algorithms which limit the sum of the weights of the base classifiers, such as MARGINBOOST.L1 proposed by Mason, Baxter, Bartlett, and Frean (1999). Furthermore, many other algorithms have also been proposed to solve directly the regularized boosting problems of the type (1) when the minimization is performed over all $\lambda > 0$. We refer the reader to the recent comprehensive review of Meir and Rätsch (2003). For additional discussion on the algorithmic issues we refer to Bennett, Demiriz, and Rätsch (2002), Lugosi and Vayatis (2003).

2.2 Relation to earlier work

Margin bounds. The first theoretical bounds about boosting-type methods are so-called "margin bounds". Although the motivation for deriving these bounds was initially to study the AdaBoost algorithm, these bounds are "agnostic" in the sense that they do not depend on the precise algorithm used, and can be applied for any algorithm which returns an estimator belonging to $\bigcup_{\lambda>0} \mathcal{F}_\lambda$. These bounds rely on the complexity of the base class $\mathcal{C}$ and on an empirical quantity, called margin. Schapire, Freund, Bartlett, and Lee (1998) proved the first bound of this type for boosting algorithms, and improved rates were obtained by Koltchinskii and Panchenko (2002). Duffy and Helmbold (2000) used the former result to study boosting-type algorithms with more general potential functions (such as the function ϕ considered in this paper). Margin bounds provide an explicit confidence interval for the generalization error, although it is recognized that the bounds obtained are generally too loose to be of practical interest.

Oracle inequalities. As opposed to margin bounds, oracle-type inequalities refer to a precise algorithm, usually some adaptive empirical loss minimization procedure over a family of models. Oracle inequalities ensure that the adaptive estimator does "almost" as well (up to additional terms that should be as small as possible) as the best possible function inside each model. Oracle inequalities do not provide an explicit confidence interval, but a guarantee about the performance and good behavior of the estimator with respect to a given collection of models. They allow, in particular, to derive bounds about convergence rates of the procedures considered. This type of bound will be our main focus in this paper.

Convergence rates and model adaptivity. An oracle inequality for the estimator defined by (1) was derived by Lugosi and Vayatis (2003) (see also Zhang (2003) for oracle inequalities in a related but different framework), when the penalty function ζ is of order $n^{-1/2}$. However, it was proved by Bartlett, Jordan, and McAuliffe (2003) that, when λ is fixed, the rate of convergence of $\mathcal{A}(\widehat{f}_n^\lambda)$ towards $\inf_{f\in\mathcal{F}_\lambda} A(f)$ is of order $n^{-(V+2)/(2(V+1))}$ — hence strictly smaller than $O(n^{-1/2})$. This result can be compared to the the improved rates — which were of the same order — obtained by Koltchinskii and Panchenko (2002) for margins bounds. One goal of the present paper is to provide an improved oracle inequality that shows the adaptivity (and consistency) of the penalized estimator with respect to these faster rates for a corresponding lighter penalty function (of order strictly smaller than $O(n^{-1/2})$) . Note that — up to our knowledge — it is *not* straightforward to build an adaptive estimator over the different models $\mathcal{F}_\lambda$ directly from the single-model analysis of Bartlett, Jordan, and McAuliffe (2003). In the present paper, although we use similar techniques, we require to use additional machinery and slightly different hypotheses for the model adaptive estimator. Additional discussion can be found in Section 7.

3. Main results

To study the probability of error of the classifier $g_{\widehat{f}_n}$, we first investigate the magnitude of $A(\widehat{f}_n) - A^*$ which is well-known to be related to the difference $L(\widehat{f}_n) - L^*$. All subsequent results are based on the following theorem.

Theorem 1 *Assume that the cost function ϕ is twice differentiable, strictly increasing and strictly convex with $\phi(0) = 1$ and $\lim_{x\to-\infty}\phi(x) = 0$ such that the constant*

$$L_\phi = 0 \vee \max_{x\in\mathbb{R}} \left(\frac{2(\phi'(x) + \phi'(-x))}{\frac{\phi''}{\phi'}(x) + \frac{\phi''}{\phi'}(-x)} - (\phi(x) + \phi(-x)) \right) \tag{2}$$

is finite. (Here $a \vee b$ denotes the maximum of a and b.) Define

$$R(\lambda, n) = (V+2)^{\frac{V+2}{V+1}} ((L_\phi + 2)\phi(\lambda))^{\frac{1}{V+1}} (\lambda\phi'(\lambda))^{\frac{V}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}} ,$$

$$b(\lambda) = (L_\phi + 2)\phi(\lambda) ,$$

and let $(\lambda_k)_{k\in\mathbb{N}}$ be an increasing sequence in $(1, +\infty)$ such that $\sum_{k\in\mathbb{N}} \lambda_k^{-\alpha} \le 1$ for some $\alpha > 0$. Then there exist positive constants c_1, c_2 such that if $\zeta : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies

$$\forall\lambda > 0, \quad \zeta(\lambda) \ge c_1 R(\lambda, n) + \frac{c_2 b(\lambda)(\alpha\log(\lambda) + \xi + \log(2))}{n}$$

for some positive number ξ, then, with probability at least $1 - \exp(-\xi)$, the penalized estimator $\widehat{f}$ defined by (1) satisfies

$$A(\widehat{f}_n) - A(f^*) \le 2 \inf_{k\ge1} \left\{ \inf_{f\in\mathcal{F}_{\lambda_k}} (A(f) - A(f^*)) + 2\zeta(\lambda_k) \right\} .$$

The proof of this theorem is given in Section 7. A few remarks are in order.

Remark 1. (constants.) Concrete values of the constants c_1 and c_2 may be obtained from the proof. However, these values are not optimal and for clarity, and because our main focus here is on the rate of convergence as a function of the sample size, we preferred not to specify their values.

Remark 2. (confidence.) The definition of the penalty given in the theorem depends on the confidence parameter ξ. However, note that its role is minor since it only appears in the smaller order second term. Indeed, for concreteness, one may take, for example, $\xi = 2\log n$ without altering the obtained rate of convergence. This choice also allows one to deduce "almost sure" convergence results by an application of the Borel-Cantelli lemma. The theorem presented here is derived as a consequence of Theorem 7 in Blanchard, Bousquet, and Massart (2003). (The statement of the cited result is given in Appendix A below.) It is also possible, with a penalty function of the same order up to logarithmic terms, to derive similar nonasymptotic upper bounds for the expected difference $\mathbb{E}A(\widehat{f}_n) - A(f^*)$ using Theorem 8 of Blanchard, Bousquet, and Massart (2003). The corresponding result is omitted for brevity.

Remark 3. (cost functions.) The properties required by Theorem 1 of the cost function are not claimed to be necessary to derive the result. Especially the condition involving the constant L_ϕ may seem to be unnatural, although it is not overly restrictive. In particular the most widely used strictly convex cost functions, the exponential and the "logit" functions satisfy the property. Indeed, it is straightforward to check that for $\phi(x) = e^x$, $L_\phi = 0$ while for the logit cost $\phi = \log_2(1+e^x)$, $L_\phi = 2 - 2\log 2$. We give the corresponding explicit corollary for these two cost functions (using some straightforward upper bounds and the fact that $\lambda > 1$):

Corollary 2 *For the exponential cost function $\phi(x) = \exp(x)$, the penalty function*

$$\zeta(\lambda) = c_1(V+2)\exp(\lambda)\lambda^{\frac{V}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}} + c_2 \frac{\exp(\lambda)(\alpha\log\lambda + \xi)}{n},$$

and for the logit cost $\phi(x) = \log(1+e^x)$ the penalty function

$$\zeta(\lambda) = c_3(V+2)\lambda n^{-\frac{1}{2}\frac{V+2}{V+1}} + c_4 \frac{\lambda(\alpha\log\lambda + \xi)}{n},$$

(where c_1, c_2, c_3, c_4 are appropriate constants) satisfy the requirements of Theorem 1.

In particular, for the logit cost, it is interesting to note that a penalization which behaves linearly (up to a logarithmic factor) in λ is sufficient. This corresponds to a regularization function proportional to $||w||_1$, where w is the collection of coefficients defining a positive linear combination of base class functions. This type of regularization has been proposed by various authors (see, e.g., Meir and Rätsch (2003) for an overview).

How restrictive is condition (2) in the case of more general cost functions? Since we assumed that ϕ is twice differentiable, strictly increasing and convex, L_ϕ is finite if and only if the limsup of the expression inside the maximum in (2), when $x \to \pm\infty$, is not $+\infty$. A simple sufficient condition for this to hold is that there exists some $L > 0$ such that

$\liminf_{x\to-\infty}(\phi''/\phi')(x) > L$ and $\limsup_{x\to+\infty}(\phi'/\phi)(x) < L/2$. Furthermore, if we assume that $\eta(X)$ takes values in $[\epsilon, 1-\epsilon]$ almost surely, then by a straightforward modification of the proof of Theorem 1 (or, to be more precise, of Lemma 19 in Section 7) one sees that in the definition of L_ϕ, the maximum can be restricted to $x \in [-f^*_\varepsilon, f^*_\varepsilon]$, where f^*_ε is the value of f^* at a point x such that $\eta(x) = 1-\varepsilon$. In this case L_ϕ is necessarily finite. Note that this assumption on η can be enforced by adding a small flipping noise on the data labels (see the related discussion below).

We note that Bartlett, Jordan, and McAuliffe (2003) study the role of the cost function in depth and derive convergence results on a fixed model $\mathcal{F}_\lambda$ for much more general cost functions. The more restrictive conditions needed here come from the fact that we consider an *adaptive* estimator over the set of models.

In the case when the distribution of the (X,Y) happens to be such that the "approximation error" $\inf_{f\in\mathcal{F}_{\lambda_k}} A(f) - A^*$ vanishes for some value of λ, the above theorem implies the following immediate corollary for the rate of convergence of $A(\widehat{f}_n)$ to A^*.

Corollary 3 *Assume that the distribution of (X,Y) is such that there exists a $\lambda_0 > 0$ such that $\inf_{f\in\mathcal{F}_{\lambda_0}} A(f) = A(f^*)$. Under the conditions of Theorem 1, if the penalty is chosen to be*

$$\zeta(\lambda) = c_1 R(\lambda, n) + \frac{c_2 b(\lambda)(\alpha\log(\lambda) + 2\log n + \log 2)}{n}$$

then for every n, with probability at least $1 - 1/n^2$,

$$A(\widehat{f}_n) - A(f^*) \le C n^{-\frac{1}{2}\left(\frac{V+2}{V+1}\right)}$$

where the constant C depends on the distribution, on the class $\mathcal{F}$, and on the cost function ϕ.

Note that the penalty function does *not* depend on λ_0 above, so that the procedure is truly adaptive.

Of course, our main concern is not the behavior of the expected cost $A(\widehat{f}_n)$ but the probability of error $L(\widehat{f}_n)$ of the corresponding classifier. However for most cost functions the difference $L(\widehat{f}_n) - L^*$ may directly be related to $A(\widehat{f}_n) - A^*$. Next we recall a simple but very useful inequality due to Zhang (2003). This result has been generalized to a great extent by Bartlett, Jordan, and McAuliffe (2003) for very general cost functions but we do not use the full power of their result.

Lemma 4 (ZHANG) *Let ϕ be a nonnegative convex nondecreasing cost function such that there exist constants c and $s \ge 1$ satisfying, for any $\eta \in [0,1]$,*

$$\left|\frac{1}{2} - \eta\right|^s \le c^s(1 - H(\eta))$$

where $H(\eta) = \inf_{\alpha\in\mathbb{R}}(\eta\phi(-\alpha) + (1-\eta)\phi(\alpha))$. Then for any real-valued measurable function f,

$$\begin{aligned} L(f) - L(f^*) &\le 2c\left(\mathbb{E}\left[(1 - H(\eta(X)))\mathbb{I}_{[g_f(X)\neq g^*(X)]}\right]\right)^{1/s} \\ &\le 2c(A(f) - A(f^*))^{1/s} . \end{aligned}$$

We note here that for both the exponential and the logit cost functions the condition of the lemma is satisfied with $c = \sqrt{2}$ and $s = 2$.

Lemma 4 implies that the rate of convergence of $L(f) - L(f^*)$ to zero is at least as fast as the sth root of the rate of $A(f) - A(f^*)$ to zero. The next lemma shows that, in fact, the excess probability of error $L(f) - L(f^*)$ always goes to zero strictly faster than $(A(f) - A(f^*))^{1/s}$ whenever s is strictly greater than one. (Recall that this is the case for the exponential and logit cost functions that are our main concern in this paper.)

Lemma 5 *Let ϕ be a nonnegative convex nondecreasing cost function such that there exist constants c and $s > 1$ satisfying, for any $\eta \in [0,1]$,*

$$\left|\frac{1}{2} - \eta\right|^s \leq c^s(1 - H(\eta)) \ .$$

Let $\{f_n\}$ be a sequence of real-valued measurable functions with $\lim_{n\to\infty} A(f_n) = A(f^*)$. Then, as $n \to \infty$,

$$\frac{L(f_n) - L(f^*)}{(A(f_n) - A(f^*))^{1/s}} \to 0 \ .$$

PROOF. The proof is based on Lemma 4 and ideas from Devroye, Györfi, and Lugosi (1996, Theorem 6.5). Let $\epsilon \in (0, 1/2)$ be an arbitrary number. Then

$$\begin{aligned}
& L(f_n) - L(f^*) \\
&= \mathbb{E}\left[|2\eta(X) - 1|\mathbb{I}_{[g_{f_n}(X)\neq g^*(X)]}\right] \\
&\quad \text{(see, e.g., Devroye, Györfi, and Lugosi (1996, Theorem 2.2))} \\
&= \mathbb{E}\left[|2\eta(X) - 1|\mathbb{I}_{[g_{f_n}(X)\neq g^*(X)]}\mathbb{I}_{[|\eta(X)-1/2|\leq\epsilon]}\right] \\
&\quad +\mathbb{E}\left[|2\eta(X) - 1|\mathbb{I}_{[g_{f_n}(X)\neq g^*(X)]}\mathbb{I}_{[|\eta(X)-1/2|>\epsilon]}\right] \\
&\leq \mathbb{E}\left[|2\eta(X) - 1|^s\mathbb{I}_{[g_{f_n}(X)\neq g^*(X)]}\right]^{1/s} \\
&\quad \cdot\left(\mathbb{P}\left[g_{f_n}(X) \neq g^*(X), |\eta(X) - 1/2| \leq \epsilon, \eta(X) \neq \frac{1}{2}\right]^{(s-1)/s}\right. \\
&\quad \left. +\mathbb{P}\left[g_{f_n}(X) \neq g^*(X), |\eta(X) - 1/2| > \epsilon\right]^{(s-1)/s}\right) \\
&\quad \text{(by Hölder's inequality applied for both terms)}
\end{aligned}$$

Using the assumption on ϕ,

$$\begin{aligned}
\mathbb{E}\left[|2\eta(X) - 1|^s\mathbb{I}_{[g_{f_n}(X)\neq g^*(X)]}\right] &\leq (2c)^s\mathbb{E}\left[(1 - H(\eta(X)))\mathbb{I}_{[g_{f_n}(X)\neq g^*(X)]}\right] \\
&\leq (2c)^s(A(f_n) - A(f))
\end{aligned}$$

by Lemma 4. Thus, it suffices to prove that the sum of the two probabilities above may be made arbitrarily small for large n, by an appropriate choice of ϵ. To this end, first note that for any fixed ϵ,

$$\lim_{n\to\infty} \mathbb{P}\left[g_{f_n}(X) \neq g^*(X), |\eta(X) - 1/2| > \epsilon\right] = 0$$

because otherwise $L(f_n) - L(f^*)$ would not converge to zero, contradicting the assumption that $A(f_n) - A(f^*)$ converges to zero (by Lemma 4). On the other hand,

$$\mathbb{P}\left[g_{f_n}(X) \neq g^*(X), |\eta(X) - 1/2| \leq \epsilon, \eta(X) \neq \frac{1}{2}\right] \leq \mathbb{P}\left[|\eta(X) - 1/2| \leq \epsilon, \eta(X) \neq \frac{1}{2}\right]$$

which converges to zero as $\epsilon \to 0$, and the proof is complete. ∎

Thus, when $s > 1$, $L(f_n) - L(f^*)$ converges to zero faster than $(A(f_n) - A(f^*))^{1/s}$ for all distributions. However, to obtain nontrivial bounds for the ratio of these two quantities, one has to impose some assumptions on the underlying distribution. This may be done by following Tsybakov (2003) who pointed out that under certain low-noise assumptions on the distribution much faster rates of convergence may be achieved. Tsybakov's condition requires that there exist constants $\alpha \in [0, 1]$ and $\beta > 0$ such that for any real-valued measurable function f,

$$\mathbb{P}[g_f(X) \neq g^*(X)] \leq \beta \left(L(f) - L^*\right)^{\alpha} . \tag{3}$$

Notice that all distributions satisfy this condition with $\alpha = 0$ and $\beta = 1$, while larger values of α place more restriction on the distribution. Intuitively, a large value of α means that the probability that $\eta(X)$ is close to $1/2$ is small. In the extreme case of $\alpha = 1$ it is easy to see that $\eta(X)$ stays bounded away from $1/2$ with probability one. For more discussion on the meaning of this condition we refer to Tsybakov (2003) and Bartlett, Jordan, and McAuliffe (2003). In Bartlett, Jordan, and McAuliffe (2003) it is shown that under Tsybakov's noise condition, the rate in Lemma 4 may be improved as follows.

Lemma 6 (BARTLETT, JORDAN, AND MCAULIFFE) *Let ϕ be a cost function satisfying the conditions of Lemma 4 and assume that condition (3) holds for some $\alpha \in [0, 1]$ and $\beta > 0$. Then*

$$L(f) - L(f^*) \leq \left(\frac{2^s c}{\beta^{1-s}}(A(f) - A(f^*))\right)^{1/(s - s\alpha + \alpha)} .$$

For the cost functions that are most important for the present paper, $s = 2$ and in that case, as α moves from zero to one, the exponent $1/(s - s\alpha + \alpha)$ changes from $1/2$ to 1. Thus, large values of α significantly improve the rates of convergence of $L(f)$ to L^*.

Combining Corollary 3 with Lemmas 4, 5, and 6 we obtain the following result. Even though it may be generalized trivially for other cost functions, for concreteness and simplicity we only state it for the two cost functions that have been most important in various versions of boosting classifiers. Recall that for both of these cost functions the condition of Lemma 4 is satisfied with $s = 2$.

Corollary 7 *Let ϕ be either the exponential or the logit cost function and consider the penalized estimate $\widehat{f}_n$ of Corollary 3. Assume that the distribution of (X, Y) is such that there exists a $\lambda > 0$ such that $\inf_{f \in \mathcal{F}_\lambda} A(f) = A(f^*)$. Then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq C n^{-\frac{1}{4}\left(\frac{V+2}{V+1}\right)}$$

where the constant C depends on the distribution, on the class $\mathcal{F}$, and on the cost function ϕ. Also, with probability one,

$$\lim_{n\to\infty}\left(L(\widehat{f}_n)-L^*\right)n^{\frac{1}{4}\left(\frac{V+2}{V+1}\right)}=0\ .$$

If, in addition, condition (3) holds for some $\alpha\in[0,1]$ and $\beta>0$, then with probability at least $1-1/n^2$,

$$L(\widehat{f}_n)-L^*\le Cn^{-\frac{1}{2(2-\alpha)}\left(\frac{V+2}{V+1}\right)}\ .$$

Corollary 7 is the main result of this paper on which the rest of the discussion is based. The remarkable fact about this corollary is that the obtained rate of convergence is independent of the dimension of the space in which the observations take their values. The rates depend on the VC dimension of the base class which may be related to the dimension of the input space. However, this dependence is mild and even if V is very large, the rates are always faster than $n^{-1/(2(2-\alpha))}$. In the rest of the paper we consider concrete examples of base classes and argue that the class of distributions for which such surprisingly fast rates can be achieved can be quite large. The dependence on the dimension is mostly reflected in the value of the constant C. Recall from Theorem 1 that the value of C is determined by the smallest value of λ for which $\inf_{f\in\mathcal{F}_\lambda}A(f)=A(f^*)$ and its dependence on λ is determined by the cost function ϕ. For complex distributions, high-dimensional input spaces, and simple base classes, this constant will be very large. The main message of Corollary 7 is that, as a function of the sample size n, the probability of error converges at a fast rate, independently of the dimension. To understand the meaning of this result, we need to study the main condition on the distribution, that is, that the minimizer f^* of the expected cost falls in the closure of $\mathcal{F}_\lambda$ (in the sense that $\inf_{f\in\mathcal{F}_\lambda}A(f)=A(f^*)$) for some finite value of λ. In the next sections we consider several concrete important examples which help understand the real meaning of Corollary 7.

Remark. (APPROXIMATION ERROR). In Corollary 7 we only consider the case when $\inf_{f\in\mathcal{F}_\lambda}A(f)=A(f^*)$ for some finite value of λ. In this paper we focus on this simplest situation and try to understand the nature of the distributions satisfying such conditions. On the other hand, under general conditions it can be guaranteed that the approximation error $\inf_{f\in\mathcal{F}_\lambda}A(f)-A(f^*)$ converges to zero as $\lambda\to\infty$, see, for example, Lugosi and Vayatis (2003), and Section 6 of the present paper. In this case Theorem 1 implies that $A(\widehat{f}_n)\to A(f^*)$ with probability one, so that the procedure is always consistent (thus improving the results of Lugosi and Vayatis (2003) since the penalty we consider in the present paper is of strictly smaller order in n). Furthermore, Theorem 1 tells us more: the penalized procedure effectively finds a tradeoff between the approximation properties of the sets $\mathcal{F}_\lambda$ and the estimation error. A precise study of these approximation properties and of the corresponding rates of convergence is a complex, important, and largely unexplored problem.

4. Decision stumps on the real line

In this section we consider the simple one-dimensional case when $\mathcal{X}=[0,1]$ and when the base class contains all classifiers g of the form $g(x)=s_t^+(x)=\mathbb{I}_{[x\ge t]}-\mathbb{I}_{[x<t]}$ and of the

form $g(x) = s_t^-(x) = \mathbb{I}_{[x<t]} - \mathbb{I}_{[x\geq t]}$ where $t \in [0,1]$ can take any value. (We note here that all results of this section may be extended, in a straightforward way, to the case when $\mathcal{X} = \mathbb{R}$ by the scale invariance of the estimates we consider.) Clearly, the VC dimension of $\mathcal{C}$ is $V = 2$. In order to apply Corollary 7 it remains to describe the class of distributions satisfying its conditions. The next lemma states a simple sufficient condition.

Lemma 8 *Assume that the cost function and the distribution of (X,Y) are such that the function f^* is of bounded variation. If $|\cdot|_{BV}$ denotes the total variation, define $|f|_{BV,0,1} = \frac{1}{2}(f^*(0) + f^*(1) + |f^*|_{BV})$. Then $\inf_{f\in\mathcal{F}_\lambda} A(f) = A(f^*)$ whenever $\lambda \geq |f^*|_{BV,0,1}$.*

PROOF. Assume that f^* has a bounded variation. Then f^* may be written as a sum of a nondecreasing and a nonincreasing function. A nondecreasing function h on $[0,1]$ may be approximated by a finite mixture of stumps as follows. Denote $C = h(1) - h(0)$. Let N be a positive integer and let $t_1, \ldots, t_N$ be $1/N, \ldots, N/N$-quantiles of h, that is, $t_i = \sup\{x : h(x) < h(1)i/N\}$, $i = 1, \ldots, N$. Then the function

$$\widetilde{h}(x) = h(0) + \sum_{i=1}^{N} \frac{C}{N}\mathbb{I}_{[x\geq t_i]} = \frac{h(1)+h(0)}{2}s_0^+(x) + \sum_{i=1}^{N} \frac{C}{2N}s_{t_i}^+(x)$$

is at most C/N away from h in the supremum norm. Note also that $\widetilde{h} \in \mathcal{F}_{|h|_{BV,0,1}}$. Similarly, a nonincreasing function g may be approximated by a function $\widetilde{g} \in \mathcal{F}_{|g|_{BV,0,1}}$ such that $\sup_{x\in[0,1]} |g(x) - \widetilde{g}(x)| \leq (g(0) - g(1))/N$. Thus, the function $\widetilde{f} = \widetilde{h} + \widetilde{g}$ is such that

$$\sup_{x\in[0,1]} |f^*(x) - \widetilde{f}(x)| \leq \frac{h(1) - h(0) + g(0) - g(1)}{N} = \frac{|f^*|_{BV}}{N}$$

and moreover $\widetilde{f} \in \mathcal{F}_{|f^*|_{BV,0,1}}$ since $|h|_{BV} + |g|_{BV} = |f^*|_{BV}$. Thus, since N is arbitrary, f^* is in the closure of $\mathcal{F}_{|f^*|_{BV,0,1}}$ with respect to the supremum norm. The statement now follows by the continuity of ϕ and the boundedness of the functions in the closure of $\mathcal{F}_{|f^*|_{BV,0,1}}$ with respect to the supremum norm. ∎

Thus, the fast rates of convergence stated in Corollary 7 can be guaranteed whenever f^* is everywhere finite and has a bounded variation. Recall that for the exponential cost function $f^* = (1/2)\log(\eta/(1-\eta))$ and for the logit cost function $f^* = \log(\eta/(1-\eta))$. In both cases, it is easy to see that f^* has a bounded variation if and only if η is bounded away from zero and one and has a bounded variation. In particular, we obtain the following corollary matching the minimax rate of convergence for the probability of error obtained with a different method by Yang (1999a).

Corollary 9 *Let $X \in [0,1]$. Let ϕ be either the exponential or the logit cost function and consider the penalized estimate $\widehat{f}_n$ of Corollary 3 based on decision stumps on the real line. If there exists a constant $b > 0$ such that $b \leq \eta(X) \leq 1-b$ with probability one and η has a bounded variation, then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq Cn^{-\frac{1}{3}}$$

where the constant C depends on b and $|\eta|_{BV}$. Also, with probability one,

$$\lim_{n\to\infty} n^{\frac{1}{3}} \left(L(\widehat{f}_n) - L^* \right) = 0 \ .$$

If, in addition, condition (3) holds for some $\alpha \in [0,1]$ and $\beta > 0$, then for every n, with probability at least $1 - 1/n^2$,

$$L(\widehat{f}_n) - L^* \le C n^{-\frac{2}{3(2-\alpha)}} \ .$$

The dependence of the value of the constant C on b and $|\eta|_{BV}$ may be determined in a straightforward way from Theorem 1. If λ_k is the smallest value for which $\inf_{f\in\mathcal{F}_{\lambda_k}} A(f) = A^*$, then the constant C in the first inequality is proportional to $\left((L_\phi+2)\phi(\lambda_k)\right)^{1/6} \left(\lambda_k\phi'(\lambda_k)\right)^{1/3}$. Clearly, λ_k can be bounded as a function of b and $|\eta|_{BV}$ as shown in Lemma 8. Concrete values are given in Corollary 12 below in the more general multivariate case.

The condition that $\eta(x)$ is bounded away from zero and one may seem to be quite unnatural at first sight. Indeed, values of $\eta(x)$ close to zero and one mean that the distribution has little noise and should make the classification problem easier. However, regularized boosting methods suffer when faced with a low-noise distribution since very large values of λ are required to drive the approximation error $\inf_{f\in\mathcal{F}_\lambda} A(f) - A^*$ close to zero. (Note, however, that even when η does not satisfy the conditions of Corollary 9, $\lim_{n\to\infty} L(\widehat{f}_n) = L^*$ almost surely, under a denseness assumption, by Corollary 7.) The next simple example illustrates in part that phenomenon: indeed, if λ is not sufficiently large to make $\mathcal{F}_\lambda$ contain f^*, then the classifier minimizing $A(f)$ over $\mathcal{F}_\lambda$ may indeed have a very large probability of error because the function minimizing the A-risk puts all its mass on points for which η is close to 0 or 1, while "neglecting" other points.

Example 1. (MINIMIZING A COST FUNCTION FOR A FIXED λ MAY BE BAD.) This example shows a situation in which if λ is not large enough, even though the class $\mathcal{F}_\lambda$ contains a function f such that the corresponding classifier g_f equals the Bayes classifier g^*, the function $\overline{f}_\lambda$ minimizing the expected cost $A(f)$ over $\mathcal{F}_\lambda$ induces a classifier with a significantly larger probability of error.

Consider a simple problem where the distribution of X is atomic, distributed uniformly on the four points $x_1, \ldots, x_4$. The base class $\mathcal{C}$ contains five classifiers: for each $i = 1, \ldots, 4$ there is a $g_i(x) = 2\mathbb{I}_{[x=x_i]} - 1$ and also $\mathcal{C}$ contains the trivial classifier $g_0(x) \equiv 1$. Obviously, for any $\lambda > 0$, the functions in $\mathcal{F}_\lambda$ induce all possible 16 classifiers on the four-point set $\mathcal{X} = \{x_1, \ldots, x_4\}$. Now consider the distribution defined by $\eta(x_1) = 1/2+\delta$, $\eta(x_2) = 1/2-\delta$, $\eta(x_3) = 1$, and $\eta(x_4) = 0$. Then it is easy to show that if ϕ is a convex strictly increasing differentiable cost function and λ_0 is such that $\phi'(-\lambda_0) = 2\delta$, then for any $\lambda \le \lambda_0$, the optimizer of the cost function $\overline{f}_\lambda$ puts positive weight on x_3 and x_4 and zero weight on x_1 and x_2 and thus has a probability of error $L(g_{\overline{f}_\lambda}) = 1/4$ while the Bayes error is $L^* = 1/4 - \delta/2$. The details of the proof are given in Appendix B. Note that the fact that η is 1 and 0 on x_3 and x_4 is only to make the example simpler; we could assume $\eta(x_3) = 1/2 + \Delta$, $\eta(x_4) = 1/2 - \Delta$ with $\Delta > \delta$ and observe a comparable behavior.

If η can be arbitrarily close to 0 and 1, then f^* takes arbitrarily large positive or negative values and thus cannot be in any $\mathcal{F}_\lambda$ (since functions in this set take values in

$[-\lambda, \lambda]$). However, one may easily force the condition of Corollary 9 to hold by adding some random noise to the data. Indeed, if, for example, we define the random variable Y' such that it equals Y with probability $3/4$ and $-Y$ with probability $1/4$, then the function $\eta'(x) = \mathbb{P}[Y' = 1|X = x] = 1/4 + \eta(x)/2$ takes its values in the interval $[1/4, 3/4]$ (a similar transformation was also proposed by Yang (1999a), Yang (1999b)). More importantly, the Bayes classifier g' for the distribution (X, Y') coincides with the Bayes classifier g^* of the original problem. Also, recalling from Devroye, Györfi, and Lugosi (1996) that for any classifier g,

$$L(g) - L^* = \mathbb{E}\mathbb{I}_{[g(X) \neq g^*(X)]}|2\eta(X) - 1|$$

and denoting the probability of error of g under the distribution of (X, Y') by $L'(g)$ and the corresponding Bayes error by L'^*, we see that for any classifier g,

$$L(g) - L^* = 2(L'(g) - L'^*) \ . \tag{4}$$

This means that if one can design a classifier which performs well for the "noisy" problem (X, Y'), then the same classifier will also work well for the original problem (X, Y). Thus, in order to enlarge the class of distributions for which the fast rates of convergence guaranteed by Corollary 9 holds, one may artificially corrupt the data by a random noise, replacing each label Y_i by a noisy version Y_i' as described above. Then the distribution of the noisy data is such that $\eta'(x)$ is bounded away from zero. If we also observe that $|\eta'|_{BV} = (1/2)|\eta|_{BV}$ and that if $\eta(x)$ satisfies condition (3) for some $\alpha \in [0, 1]$ and $\beta > 0$ then $\eta'(x)$ also satisfies condition (3) with the same $\alpha \in [0, 1]$ but with $\beta' = 2^\alpha \beta$, we obtain the following corollary.

Corollary 10 *Let $X \in [0, 1]$. Let ϕ be either the exponential or the logit cost function and consider the penalized estimate $\widehat{f}_n$ based on decision stumps, calculated based on the noise-corrupted data set described above. If $\eta(x)$ has a bounded variation, then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq Cn^{-\frac{1}{3}}$$

where the constant C depends only on $|\eta|_{BV}$. If, in addition, condition (3) holds for some $\alpha \in [0, 1]$ and $\beta > 0$, then

$$L(\widehat{f}_n) - L^* \leq Cn^{-\frac{2}{3(2-\alpha)}} \ .$$

Of course, by corrupting the data deliberately with noise one loses information, but it is a curious property of the regularized boosting methods studied here that the rate of convergence may be speeded up considerably for some distributions. (Indeed, this fact was already pointed out by Yang in establishing general minimax rates of convergence in various settings (see Yang (1999a), Yang (1999b)).) Besides, recall that, in the case we consider a cost function ϕ such that the constant L_ϕ is infinite in equation (2), Theorem 1 cannot be applied in general; however since the noise-degraded η' is bounded away from 0 and 1, L_ϕ can be replaced by some finite constant (see the remark about cost functions following Theorem 1), and hence Theorem 1 can be applied for the noisy distribution. For many distributions, the performance deteriorates by adding noise, but at least the rate of convergence is guaranteed to stay the same, and only the value of the constant C will be affected. Unfortunately, it is impossible to test whether η is bounded away from zero or not, and it may be safe to add a little noise. Of course, the level of the added noise (*i.e.*, the probability of flipping the labels in the training set) does not need to be the 1/4 described above. Any strictly positive value may be used and Corollary 10 remains true. While a more precise study is out of the scope of this paper, let us just remark that a sensible choice of the noise level based on the present bounds should be able to find a tradeoff between the improvement of the bias in the A-risk and the performance degradation as appearing in equation (4).

Finally, a natural question is whether the improved convergence rate that could be obtained by adding a small labelling noise to the training data really is a practical consequence of using a "surrogate" convex loss (the function ϕ) instead of the $0-1$ loss, or if it is just an artefact of the analysis. Namely, consider a case where the data is completely separable with some margin $\theta > 0$ by some function $f \in \mathcal{F}_1$. In this situation the margin bounds of Koltchinskii and Panchenko (2002) ensure that the convergence rates are as fast as in our analysis, and no labelling noise is needed. However, in a generic situation the problem with using the surrogate A-risk is the disequilibrium between regions where the target function f^* is very large or even infinite, and other regions where it is relatively small (of course in such a situation the data is not separable). In this situation, it may very well happen that the estimator will tend to concentrate all of its efforts on the former regions while neglecting the latter, as was shown prototypically in Example 1. Then, adding a small amount of noise could effectively bring the estimator to improve on the latter regions, which would have a definite effect on generalization error. Whether adding noise artificially is helpful in practice should be investigated by an adequate experimental study.

5. Decision stumps in higher dimensions

5.1 Stumps and generalized additive models

In this section we investigate the case when $\mathcal{X} = [0,1]^d$ and the base class $\mathcal{C}$ contains all "decision stumps", that is, all classifiers of the form $s^+_{i,t}(x) = \mathbb{I}_{[x^{(i)} \geq t]} - \mathbb{I}_{[x^{(i)} < t]}$ and $s^-_{i,t}(x) = \mathbb{I}_{[x^{(i)} < t]} - \mathbb{I}_{[x^{(i)} \geq t]}$, $t \in [0,1]$, $i = 1, \ldots, d$, where $x^{(i)}$ denotes the i-th coordinate of x.

An important property of boosting using decision stumps is that of scale invariance. Indeed, if each component of the observation vectors X_i is transformed by a (possibly different) strictly monotone transformation then the resulting classifier does not change. This remark also implies that the assumption that the observations take their values from the bounded set $[0,1]^d$ is not essential, we use it for convenience.

A straightforward extension of the proof of Lemma 8 in the previous section shows that the closure of $\mathcal{F}_\lambda$ with respect to the supremum norm contains all functions f of the form

$$f(x) = f_1(x^{(1)}) + \cdots + f_d(x^{(d)})$$

where the functions $f_i : [0,1] \to \mathbb{R}$ are such that $|f_1|_{BV,0,1} + \cdots + |f_d|_{BV,0,1} \leq \lambda$. Therefore, if f^* has the above form, we have $\inf_{f \in \mathcal{F}_\lambda} A(f) = A(f^*)$.

Recalling that the function f^* optimizing the cost $A(f)$ has the form

$$f^*(x) = \frac{1}{2} \log \frac{\eta(x)}{1 - \eta(x)}$$

in the case of the exponential cost function and

$$f^*(x) = \log \frac{\eta(x)}{1-\eta(x)}$$

in the case of the logit cost function, we see that boosting using decision stumps is especially well fitted to the so-called additive logistic model in which η is assumed to be such that $\log(\eta/(1-\eta))$ is an additive function (i.e., it can be written as a sum of univariate functions of the components of x), see Hastie and Tibshirani (1990). The fact that boosting is intimately connected with additive logistic models of classification has already been pointed out by Friedman, Hastie, and Tibshirani (2000). The next result shows that indeed, when η permits an additive logistic representation then the rate of convergence of the regularized boosting classifier is fast and has a very mild dependence on the distribution.

Corollary 11 *Let $X \in [0,1]^d$ with $d \geq 2$. Let ϕ be either the exponential or the logit cost function and consider the penalized estimate $\widehat{f}_n$ of Corollary 3 based on decision stumps. Let $V_2 = 3$, $V_3 = 4, V_4 = 5$, and for $d \geq 5$, $V_d = \lfloor 2\log_2(2d) \rfloor$. If there exist functions $f_1, \ldots, f_n : [0,1] \to \mathbb{R}$ of bounded variation such that $\log \frac{\eta(x)}{1-\eta(x)} = \sum_{i=1}^d f_i(x^{(i)})$ then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq C n^{-\frac{1}{4}\left(\frac{V_d+2}{V_d+1}\right)}$$

where the constant C depends on $\sum_{i=1}^d |f_i|_{BV,0,1}$. If, in addition, condition (3) holds for some $\alpha \in [0,1]$ and $\beta > 0$, then

$$L(\widehat{f}_n) - L^* \leq C n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)} .$$

PROOF. The statements follow from Corollary 7. The only detail that remains to be checked is the VC dimension V_d of the class $\mathcal{C}$ of decision stumps. This may be bounded by observing that the shatter coefficient (*i.e.*, the maximum number of different ways n points in $[0,1]^d$ can be classified using decision stumps) is at most $\min(2d(n+1), 2^n)$. Thus, for $d \geq 5$, $2d(n+1) < 2^n$ if and only if $n > \log_2(2d) + \log_2(n+1)$ which is implied by $n > 2\log_2(2d)$. For $d \leq 4$, just notice that decisions stumps are linear splits and the VC dimension of the class of all linear splits in $\mathbb{R}^d$ equals $d+1$. ∎

Remark. (DEPENDENCE ON THE DIMENSION.) Under the assumption of the additive logistic model, the rate of convergence is of the order of $n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)}$ where V_d depends on d in a logarithmic fashion. Even for large values of d, the rate is always faster than $n^{-1/2(2-\alpha)}$. It is also useful to examine the dependence of the constant C on the dimension. A quick look at Theorem 1 reveals that C in the first inequality of Corollary 11 may be bounded by a universal constant times $\sqrt{V_d \phi(\lambda)^{1/V_d} \lambda \phi'(\lambda)}$ where λ is the smallest number such that $\inf_{f \in \mathcal{F}_\lambda} A(f) = A^*$. Thus, we may take $\lambda = \sum_{i=1}^d |f_i|_{BV,0,1}$. Since $V_d = \lfloor 2\log_2(2d) \rfloor$, the dependence on the dimension is primarily determined by the growth of the cost function ϕ. Here there is a significant difference between the behavior of the exponential and the logistic cost functions in high dimensions. For the purpose of comparison, it is reasonable to consider distributions such that $\lambda = \sum_{i=1}^{d} |f_i|_{BV,0,1}$ is bounded by a linear function of d. In that case the constant C depends on d as $O(\sqrt{de^d \log d})$ in the case of the exponential cost function, but only as $O(\sqrt{d \log d})$ in the case of the logistic cost function (using directly Theorem 1 instead of the upper bound mentioned above). In summary, regularized boosting using the logistic cost function and decision stumps has a remarkably good behavior under the additive logistic model in high dimensional problems, as stated in the next corollary.

Corollary 12 *Let $X \in [0,1]^d$ with $d \geq 2$. Let ϕ be the logit cost function and consider the penalized estimate $\widehat{f}_n$ of Corollary 3 based on decision stumps. Let B be a positive constant. If there exist functions $f_1, \ldots, f_n : [0,1] \to \mathbb{R}$ with $\lambda = \sum_{i=1}^{d} |f_i|_{BV,0,1} \leq Bd$ such that $\log \frac{\eta(x)}{1-\eta(x)} = \sum_{i=1}^{d} f_i(x^{(i)})$ then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ of the associated classifier satisfies*

$$L(\widehat{f}_n) - L^* \leq C\sqrt{d \log d}\; n^{-\frac{1}{4}\left(\frac{V_d+2}{V_d+1}\right)}$$

where C is a universal constant and V_d is as in Corollary 11. If, in addition, condition (3) holds for some $\alpha \in [0,1]$ and $\beta > 0$, then

$$L(\widehat{f}_n) - L^* \leq C(d \log d)^{\frac{1}{2-\alpha}}\; n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)}\ .$$

Remark 1. (ADDING NOISE.) Just like in the one-dimensional case, the conditions of Corollary 11 require that η be bounded away from zero and one. To relax this assumption, one may try to add a random noise to the data, just like in the one-dimensional case. However, this may not work in the higher-dimensional problem because even if f^* is an additive function, it may not have this property any longer after the noise is added.

Remark 2. (CONSISTENCY.) The results obtained in this paper (for instance, Corollary 7) imply the consistency of the classifier $\widehat{f}_n$ under the only assumption that f^* may be written as a sum of functions of the components, that is, that $L(\widehat{f}) \to L^*$ almost surely. The additional assumption on the bounded variation of the components guarantees the fast rates of convergence. However, if f^* is not an additive function, consistency cannot be guaranteed, and the example of the previous section shows that boosting is not robust in the sense that it is not even guaranteed to perform nearly as well as the best classifier contained in the class. Still, it is important to understand the structure of the classifiers that can be realized by aggregating decision stumps. The rest of this section is dedicated to this problem.

5.2 Set approximation properties of mixtures of stumps

In what follows we investigate what kind of sets $A \subset [0,1]^d$ can be well approximated by sets of the form $A_f = \{x | f(x) > 0\}$, where $f \in \mathcal{F}_\lambda$ is a linear combination of decision stumps.

It helps understand the main properties of these sets if we first consider the discrete case, that is, when $\mathcal{X}$ is a grid of the form $\mathcal{X} = \{0, 1/k, \ldots, k/k\}^d$. If $d = 1$, obviously any function can be written as a mixture of stumps since it is always of finite variation in this discrete setting.

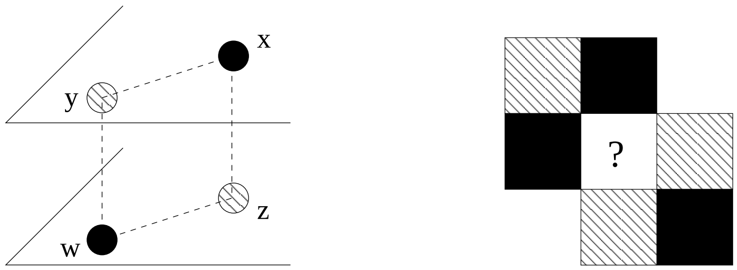

Figure 1: Points or regions belonging to the set A are in black. Left: four points in XOR position. Right: a counterexample to theorem 14 when $\mathcal{X}$ is not a cube: if the center square is not part of $\mathcal{X}$, the non-XOR requirement is satisfied, but any way to "extend" $\mathcal{X}$ and A to the center square will lead to a creation of an XOR position.

Next consider the case $d = 2$. It is then easy to see that if a set A is obtained as the support of the positive part of an additive function of the form $f(x) = f_1(x^{(1)}) + f_2(x^{(1)})$ then there cannot exist four points x, y, z, w, such that these points are the corners of a rectangle aligned with the axes, the two corners on one diagonal are elements of A, and the two points of the other diagonal are not in A. We call this the "XOR" position. It turns out that this simple property, which we call for brevity the "non-XOR requirement", is actually a necessary and sufficient condition for a set to be of the form A_f for $f \in \mathcal{F}_\lambda$ for any $\lambda > 0$.

Next we generalize this idea to d-dimensions and characterize completely the sets one can obtain with the additive models in the discrete setting. For this we need a more general definition of the XOR position (see also Figure 1).

Definition 13 *Let $\mathcal{X} = \{0, 1/k, \ldots, k/k\}^d$ and $A \subset \mathcal{X}$. We say that four points x, y, z, w are in* XOR *position with respect to A if there exists an $1 \le i_0 \le d$ such that*

$$\begin{cases} x^{(i_0)} = y^{(i_0)}, & z^{(i_0)} = w^{(i_0)}; \\ x^{(i)} = z^{(i)}, & y^{(i)} = w^{(i)}, \text{ for } i \neq i_0; \end{cases} \tag{5}$$

and $x, w \in A$ but $y, z \notin A$.

For a discrete grid we have the following characterization of sets realizable as the positive part of the mixture of stumps. Recall that a set S is called a monotone layer in $\mathbb{R}^d$ if it has one of the following properties: either (1) for any $x \in S$ all points $y \le x$ are also in S, or (2) for any $x \in S$ all points $y \ge x$ are also in S. (We say that $y \le x$ if the inequality holds componentwise.)

Theorem 14 *Let $\mathcal{X} = \{0, 1/k, \ldots, k/k\}^d$ and $A \subset \mathcal{X}$. The following properties are equivalent:*

(i) There exists f such that $A = \{x | f(x) > 0\}$ where $f(x) = f_1(x^{(1)}) + \ldots + f_d(x^{(d)})$;

(ii) There does not exist any $x, y, z, w \in \mathcal{X}$ in XOR *position with respect to A;*

(iii) A can be transformed into a monotone layer by a permutation of the order along each axis, that is, there exist permutations $\sigma_1, \ldots, \sigma_d$ of $\{0, \ldots, k\}$ such that the image of A by the function $s : x = (i_1/k, \ldots, i_d/k) \mapsto s(x) = (\sigma_1(i_1)/k, \ldots, \sigma_d(i_d)/k)$ is a monotone layer.

PROOF. (i) $\Rightarrow$ (ii): consider four points x, y, z, w satisfying (5). Suppose that $x, w \in A$ and $y \notin A$, which means $f(x), f(w) > 0$, $f(y) \leq 0$. Note that condition (i) and (5) imply that $f(x) + f(w) = f(y) + f(z)$. Hence we must have $f(z) > 0$ and the points cannot be in XOR position.

(ii) $\Rightarrow$ (iii): consider "slices" of X perpendicular to the first coordinate axis, that is, $S_i^1 = \left\{x \in \mathcal{X} | x^{(1)} = i/k\right\}$. Define an order on the slices by saying that $S_i^1 \preceq S_j^1$ if and only if for any $x = (i/k, x_2, \ldots, x_d) \in S_i^1$, if we denote $y = (j/k, x_2, \ldots, x_d) \in S_j^1$, then $\mathbb{I}_{[A(x)]} \leq \mathbb{I}_{[A(y)]}$. Now, note that (ii) implies that this order is total, that is, for any i, j either $S_i^1 \preceq S_j^1$ or $S_j^1 \preceq S_i^1$. As a consequence, we can rearrange the order along the first coordinate using a permutation σ_1, so that the slices are sorted in increasing order. By doing this we do not alter the non-XOR property, hence we can repeat the corresponding procedure along all the other coordinates. It is then easy to see that the image of A by these successive reorderings is now a monotone layer.

(iii) $\Rightarrow$ (i): first note that any monotone layer can be represented as a set of the form described in (i). Therefore, any set obtained from a monotone layer by permutations of the order along each of the axes can also be represented under this form, since it is just a matter of accordingly rearranging, separately, the values of $f_1, \ldots, f_d$. ∎

Note that it is essential in the last theorem that $\mathcal{X}$ is an hypercube $[0,1]^d$. In Figure 1 we show a contrived counterexample where $\mathcal{X}$ is not a cube and satisfies condition (ii) of the above theorem; yet it is not possible in this case to find a function f satisfying (i), because there is no way to "complete" the middle square so that the non-XOR requirement is still satisfied.

In the general case when $\mathcal{X} = [0,1]^d$, we can derive, based on the discrete case, an approximation result for sets whose boundary is of measure zero. The approximation is understood in the sense of L^1 distance between indicators of sets with respect to the probability measure of X on $\mathcal{X}$ (or, equivalently, the measure of the symmetric difference of the sets). Note that this distance is always at least as large as the excess classification error.

Theorem 15 *Let $A \subset \mathcal{X}$ be a set whose boundary δA is of measure zero. Suppose there do not exist four points $x, y, z, w \in \mathcal{X}$ in* XOR *position with respect to A. Then there exists a sequence (f_n) of linear combinations of decision stumps such that*

$$\lim_{n\to\infty} \mathbb{P}[|\mathbb{I}_{[f_n(X)>0]} - \mathbb{I}_{[X\in A]}|] \to 0 \ .$$

PROOF. We approximate $\mathcal{X}$ by discrete grids. Fix some $n \in \mathbb{N}$ and for $I = (i_1, \ldots, i_d) \in \{0, \ldots, n-1\}^d$ denote $x_I = (i_1/n, \ldots, i_d/n)$ and let $B(I)$ be the closed box $x_I + [0,1]^d$. Let Δ_n be the set of indices I such that $B(I)$ contains at least a point of the boundary of A, and $B_n = \cup_{I\in\Delta_n} B(I)$.

Now consider the discrete set $\mathcal{X}_n = \left\{x_I, I \in \{0, \ldots, n-1\}^d\right\}$ and the projection $A_n = A \cap \mathcal{X}_n$. Now in $\mathcal{X}_n$, A_n satisfies the hypothesis (ii) of Theorem 14, and hence (i) is satisfied as well and there exists a function $f_n(x) = f_{n,1}(x^{(1)}) + \ldots + f_{n,d}(x^{(d)})$ defined for $x \in \mathcal{X}_n$ with $A_n = \{x \in \mathcal{X}_n | f(x) > 0\}$. Extend the functions $f_{n,j}$ on $[0,1]$ by defining (with some abuse of notation) $f_{n,j}(i/n + \varepsilon) = f_{n,j}(i/n)$ for $\varepsilon \in (0, 1/n)$. Obviously, the extended functions $f_{n,j}$ are still mixtures of stumps.

Let now $g_n(x) = \mathbb{I}_{[f_n(x)>0]}, x \in \mathcal{X}$. We have $g_n(x) = \mathbb{I}_{[A(x)]}$ for $x \notin B_n$, by construction, and therefore

$$\mathbb{P}[|\mathbb{I}_{[f_n(X)>0]} - \mathbb{I}_{[X \in A]}|] \leq \mathbb{P}[X \in B_n] \ ,$$

which converges to zero as $n \to \infty$, since $\mathbb{I}_{[B_n]} \to \mathbb{I}_{[\delta A]}$ pointwise. ∎

Remark. (DISREGARDING THE BOUNDARY.) Since we concentrate on sets A with boundary of measure 0, it is equivalent in the sense of the L^1 distance between sets to consider A, its closure $\overline{A}$ or its interior $\mathrm{int}(A)$. One could therefore change the above theorem by stating that it is sufficient that the "non-XOR requirement" be satisfied by some set C such that $\mathrm{int}(A) \subset C \subset \overline{A}$. It would be even nicer, if only of side interest, only to take into account quadruples of points not on the boundary of A to satisfy the non-XOR requirement, so that any problem arising with the boundary may be disregarded. In Appendix C we show that this is actually the case whenever $P(\delta A) = 0$ for some measure P having full support e.g., the Lebesgue measure).

The theorems above help understand the structure of classifiers that can be realized by a linear combination of decision stumps. However, for boosting to be successful it is not enough that the Bayes classifier g^* can be written in such a form. It may happen that even though g^* is in the class of classifiers induced by functions in $\mathcal{F}_\lambda$, the classifier corresponding to $\overline{f}_\lambda$ minimizing the cost $A(f)$ in $\mathcal{F}_\lambda$ is very different. This is the message of Example 1 above. The next example shows a similar situation in which *for any* $\lambda > 0$ there exists an $f \in \mathcal{F}_\lambda$ such that $g_f = g^*$.

Example 2. (BAYES CLASSIFIER MAY BE DIFFERENT FROM THE ONE CHOSEN BY BOOSTING.) Consider a two-dimensional problem with only two non-trivial classifiers in $\mathcal{C}$ given by two linear separators, one vertical and one horizontal, and the trivial classifier assigning -1 to everything. We have four regions (denoted $\begin{pmatrix} D_1 & D_2 \\ D_3 & D_4 \end{pmatrix}$) and only three parameters (only one parameter per classifier including the trivial one). By considering only symmetric situations where η is the same on D_1 and D_4, we see that $\overline{f}$, the function minimizing $A(f)$ over $\bigcup_{\lambda>0} \mathcal{F}_\lambda$, must also be symmetric and hence we reduce (after re-parameterization) to two parameters a, b. The minimizer $\overline{f}$ is then of the form $\overline{f} = \begin{pmatrix} (a+b)/2 & a \\ b & (a+b)/2 \end{pmatrix}$.

First consider a situation in which X falls in D_1 or D_4 with probability zero. Then in this case $\overline{f} = f^*$ on D_2 and D_3. Furthermore, by choosing suitably η in these regions, one may assume that $a > 0 > b$ but $a + b > 0$. Now suppose that we put a tiny positive weight ε on regions D_1 and D_4, with the Bayes classifier on these regions being class -1. But by continuity, the associated $\overline{f}_\varepsilon$ will stay positive on these regions if ε is small enough. Then $g_{\overline{f}_\varepsilon} \neq g^*$ on these regions, while obviously for any $\lambda > 0$ we can find an appropriate function $f \in M_\lambda$ such that $g_f = g^* = \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}$ in this case.

6. Examples of consistent base classes

The results of the previous section show that using decision stumps as base classifiers may work very well under certain distributions such as additive logistic models but may fail if the distribution is not of the desired form. Thus, it may be desirable to use larger classes of base classifiers in order to widen the class of distributions for which good performance is guaranteed. Recent results on the consistency of boosting methods (see, e.g., Breiman (2000), Bühlmann and Yu (2003), Jiang (2003), Lugosi and Vayatis (2003), Mannor and Meir (2001), Mannor, Meir, and Zhang (2002), Zhang (2003)) show that universal consistency of regularized boosting methods may be guaranteed whenever the base class is so that the class of linear combinations of base classifiers is rich enough so that every measurable function can be approximated. In this section we consider a few simple choices of base classes satisfying this richness property. In particular, we recall here the following result (Lugosi and Vayatis, 2003, Lemma 1):

Lemma 16 (LUGOSI AND VAYATIS) *Let the class $\mathcal{C}$ be such that its convex hull $\mathcal{F}_1$ contains all the indicators of elements of $\mathcal{B}_0$, a subalgebra of the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ of $\mathbb{R}^d$, such that $\mathcal{B}_0$ generates $\mathcal{B}(\mathbb{R}^d)$. Then*

$$\lim_{\lambda\to\infty} \inf_{f\in\lambda\cdot\mathcal{F}_1} A(f) = A^* .$$

More generally, a straightforward modification of this Lemma shows that whenever $\mathcal{F} = \bigcup_{\lambda>0}\mathcal{F}_\lambda$ is dense in $L_1(\mu)$, then it is true that $\inf_{f\in\mathcal{F}} A(f) = A(f^*)$.

We consider the following examples; in all cases we assume that $\mathcal{X} = \mathbb{R}^d$.

(1) $\mathcal{C}_{lin}$ contains all linear classifiers, that is, functions of the form $g(x) = 2\mathbb{I}_{[a\cdot x\le b]} - 1$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$.

(2) $\mathcal{C}_{rect}$ contains classifiers of the form $g(x) = 2\mathbb{I}_{[x\in R]} - 1$ where R is either a closed rectangle or its complement in $\mathbb{R}^d$.

(3) $\mathcal{C}_{ball}$ contains classifiers of the form $g(x) = 2\mathbb{I}_{[x\in B]} - 1$ where B is either a closed ball or its complement in $\mathbb{R}^d$.

(4) $\mathcal{C}_{ell}$ contains classifiers of the form $g(x) = 2\mathbb{I}_{[x\in E]} - 1$ where either E a closed ellipsoid or its complement in $\mathbb{R}^d$.

(5) $\mathcal{C}_{tree}$ contains decision tree classifiers using axis parallel cuts with $d+1$ terminal nodes.

Clearly, the list of possibilities is endless, and these five examples are just some of the most natural choices. All five examples are such that $\bigcup_{\lambda>0}\mathcal{F}_\lambda$ is dense in $L_1(\mu)$ for any probability distribution μ (In the cases of $\mathcal{C}_{rect}$, $\mathcal{C}_{ball}$, and $\mathcal{C}_{ell}$ this statement is obvious. For $\mathcal{C}_{lin}$ this follows from denseness results of neural networks, see Cybenko (1989), Hornik, Stinchcombe, and White (1989). For $\mathcal{C}_{tree}$, see Breiman (2000).) (We also refer to the general statement given as a universal approximation theorem by Zhang (2003) and which shows that, for the classical choices of the cost function ϕ, we have, for any distribution, $\inf_{f\in\bigcup_{\lambda>0}\mathcal{F}_\lambda} A(f) = A^*$ as soon as $\bigcup_{\lambda>0}\mathcal{F}_\lambda$ is dense in the space of continuous functions under the supremum norm.) In particular, the results in the present paper imply that in all cases, the penalized estimate $\widehat{f}_n$ of Corollary 3 is universally consistent, that is, $L(\widehat{f}_n) \to L^*$ almost surely as $n \to \infty$.

Recall that the rates of convergence established in Corollary 7 depend primarily on the VC dimension of the base class. The VC dimension equals $V = d+1$ in the case of $\mathcal{C}_{lin}$, $V = 2d+1$ for $\mathcal{C}_{rect}$, $V = d+2$ for $\mathcal{C}_{ball}$, and is bounded by $V = d(d+1)/2+2$ for $\mathcal{C}_{ell}$

and by $V = d\log_2(2d)$ for $\mathcal{C}_{tree}$ (see, e.g., Devroye, Györfi, and Lugosi, 1996). Clearly, the lower the VC dimension is, the faster the rate (estimation is easier). The following question arises naturally: find a class with VC dimension as small as possible whose convex hull is sufficiently rich in $L_1(\mu)$. A recent result by Lugosi and Mendelson (2003) establishes the existence of such a class with VC dimension at most 2. This fact reveals that the combinatorial complexity of a class is not always a reliable measure of the approximation capacity of its convex hull. However, the construction by Lugosi and Mendelson is theoretical and there is probably more to say if one is concerned with practical implementations of boosting methods (see also Remark 1 below). In all cases, for even moderately large values of d, the rate of convergence stated in Corollary 7 is just slightly faster than $n^{-1/(2(2-\alpha))}$, and the most interesting problem is to determine the class of distributions for which $\inf_{f\in\mathcal{F}_\lambda} A(f) = A^*$ for some finite value of λ. In all the above-mentioned special cases this class is quite large, giving rise to a remarkably rich class of distributions for which the dimension-independent rates of convergence holds. The characterization of these classes of distributions similar to the one given in the one-dimensional case is far from being well understood. In the case of $\mathcal{C}_{lin}$ the problem is closely related to the approximation properties of neural networks. We merely refer to Barron (1992, 1993), Darken, Donahue, Gurvits, and Sontag (1997), Girosi and Anzelloti (1993), Maiorov, Meir, and Ratsaby (1999), Meir and Maiorov (2000), Pinkus (1999), Sontag (1992) for related results. Most of these references provide quantitative results relating the approximation error to the smoothness of the target function. However, there are very few attempts to characterize the functions that can actually be reconstructed with given dictionaries. In one dimension, the problem is well-understood: the closure under the uniform norm of the class of piecewise constant functions is the class of regulated functions (for which both left and right limits exist at each point). Hence, by limiting the bounded variation, we lose the ability to approximate these regulated functions with linear combinations of decision stumps. In $\mathbb{R}^d$, there is no straightforward generalization of regulated functions. Another interesting question is to investigate the approximation rates in terms of the smoothing parameter λ for universal base classes when the approximating function is taken in $\mathcal{F}_\lambda$, and the work by Meir and Maiorov (2000), Mannor, Meir, and Zhang (2002), may provide some hints for a systematic approach.

Remark 1. (computational problems.) Using the above-mentioned classes as base classifiers may cause computational problems in high-dimensional problems. Typical boosting algorithms perform an iterative gradient descent optimization to minimize the empirical cost $A_n(f)$ and each iteration step involves optimization over the class $\mathcal{C}$. This may be efficiently computed when $\mathcal{C}$ is the class of decision stumps but in any of the cases considered in this section, optimization may be problematic. There seems to exist a tradeoff between the richness of the base class and computational feasibility of the optimization. In practice one may try to find classes "in between", that is, base classes larger than decision stumps which may not give rise to universally consistent classifiers but still allow efficient optimization. Here we do not pursue this issue further.

Remark 2. (invariance.) In the previous section we already emphasized that the classifier $\widehat{f}_n$ is invariant under monotone transformations of the coordinate axes, when $\mathcal{C}$ is the class of decision stumps. This invariance property is important in situations when the different components of the feature vector X belong to incomparable physical quantities. Scale invariance shared by the method based on the classes $\mathcal{C}_{rect}$ and $\mathcal{C}_{tree}$ but not with the rest. On the other hand, the rest of the examples have different important invariance properties. For example, boosting based on $\mathcal{C}_{lin}$, $\mathcal{C}_{ball}$, and $\mathcal{C}_{ell}$ are rotation invariant, and $\mathcal{C}_{lin}$ and $\mathcal{C}_{ell}$ are invariant under arbitrary invertible linear transformations of the feature space. The choice of the base class should be influenced by the desirable invariance property in practice.

7. Proof of Theorem 1 and related results

In this section we apply general abstract single-model and model selection theorems appearing in Blanchard, Bousquet, and Massart (2003) (recalled in Appendix A for completeness) in the regularized boosting setting to derive Theorem 1. We state here single-model convergence rate theorems as well since the hypotheses to satisfy are essentially the same. This way we can recover a theorem that is similar to results appearing in Bartlett, Jordan, and McAuliffe (2003) (see a short discussion below). The theorems cited in Appendix are extensions of model selection methods by penalization originating in works by Birgé and Massart (1998), Massart (2000). We also use the technique of localized Rademacher averages for fine-scale estimates of the capacity of function classes, a principle that has been put forward in Bartlett and Mendelson (2002), Bartlett, Bousquet, and Mendelson (2002), and Bousquet (2003).

7.1 Rates of convergence in a fixed model

In this section we first restrict our attention to the empirical risk minimization estimator on a fixed model $\mathcal{F}_\lambda$. Define $\widehat{f}_n^\lambda = \arg\min_{f\in\mathcal{F}_\lambda} A_n(f)$. We then have the following theorems.

Theorem 17 *Assume that the base class $\mathcal{C}$ has VC dimension V. Then, for any $C>1$, with probability at least $1-\exp(-\delta)$, we have:*

$$A(\widehat{f}_n^\lambda) - A(f^*) \le \frac{C+1}{C-1}\left(\inf_{f\in\mathcal{F}_\lambda}(A(f)-A(f^*)) + CR_1(\lambda,n) + \frac{Cb_1(\lambda)\delta}{n}\right),$$

where

$$R_1(\lambda,n) = c_1(V+2)^{\frac{V+2}{V+1}}\left((L_\phi+2)\phi(\lambda)\right)^{\frac{1}{V+1}}\left(\lambda\phi'(\lambda)\right)^{\frac{V}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}$$

and

$$b_1(\lambda) = c_2(L_\phi+2)\phi(\lambda),$$

where c_1, c_2 are numerical constants, and L_ϕ is defined by (2).

Theorem 18 (exact bias; Bartlett, Jordan, and McAuliffe.) *Assume that the base class $\mathcal{C}$ has VC dimension V. Then, with probability at least $1-\exp(-\delta)$, we have:*

$$A(\widehat{f}_n^\lambda) - A(f^*) \le \inf_{f\in\mathcal{F}_\lambda}(A(f)-A(f^*)) + R_2(\lambda,n) + \frac{b_2(\lambda)\delta}{n},$$

where

$$R_2(\lambda,n) = c_1(V+2)^{\frac{V+2}{V+1}} \max\left(M(\lambda)^{-1}\phi'(\lambda)^2, \phi(\lambda)\right)^{\frac{1}{V+1}}\left(\lambda\phi'(\lambda)\right)^{\frac{V}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}$$

and

$$b_2(\lambda) = c_2(\phi'(\lambda)^2 M(\lambda)^{-1} + \phi(\lambda)),$$

where c_1, c_2 *are numerical constants, and* $M(\lambda) = \inf_{x\in[-\lambda,\lambda]} \phi''(x)$.

Remark. These two theorems are also consequences of a general theorem recalled in Appendix A–the difference comes from a slight difference in the application of the latter. We mention these two statements here to draw a short comparison. Theorem 18 is almost identical to Theorem 17 of Bartlett, Jordan, and McAuliffe (2003) (more precisely it is a special case of the latter, since, as already pointed out earlier, our general assumptions about ϕ are stronger in this paper). Note also that the proofs use very similar tools, although in the present paper a good part of them is wrapped up into the general theorem quoted in Appendix A. Theorem 17 on the other hand, is really the single-model counterpart of the penalized procedure of Theorem 1.

An advantage of Theorem 18 is the exact bias term, that is, the absence of the factor of $(C+1)/(C-1)$ in front of the approximation error. Note, however, that this exact bias is lost anyway when one turns to the true classification risk using Lemma 6. Also, since in our corollaries we assume the bias to be zero, this improvement becomes irrelevant. On the other hand, the dependence in V of the multiplicative constant is slightly better in Theorem 17 (note that a factor of order $\phi(\lambda)^{\frac{1}{V+1}}$ is replaced by $(\phi'(\lambda)^2 M(\lambda)^{-1})^{\frac{1}{V+1}}$ in Theorem 18: for instance taking ϕ as the exponential loss, the latter expression is the third power of the former; for the logit cost, this even more noticeable: the latter expression is of order $\exp(\lambda/(V+1))$ while the former is only of order $\lambda^{\frac{1}{V+1}}$).

Finally, note that neither of these theorems can be used directly (at least up to our knowledge) to derive an oracle bound for a penalized procedure. For the proof of Theorem 1 we need additional model selection machinery which in particular only works under the hypotheses of Theorem 17.

7.2 Proofs

Proof of Theorem 1. Theorem 1 will be derived as a consequence of Theorem 22 in Appendix A. According to the notations used in the Appendix, we define the loss function $\ell(x, y) = \phi(-xy)$ and write $\ell(f)$ as a shorthand notation for the function $(x, y) \mapsto \ell(f(x), y)$, so that $A(f) = \mathbb{E}[\ell(f)]$.

We define the reference space $\mathfrak{G}$ as the set of functions f from $\mathcal{X}$ into $\mathbb{R} \cup \{-\infty, +\infty\}$ such that $\ell(f) \in L^2(P)$ where P denotes the probability measure induced by (X, Y). Note that $f^* \in \mathfrak{G}$ (even if f^* is infinite at some points, because for any fixed point $x \in \mathcal{X}$, the average loss $\mathbb{E}[\ell(f^*(X), Y)|X = x]$ is always bounded by 1). We consider the countable family of models $(\mathcal{F}_{\lambda_k}), k \in \mathbb{N}$.

Next we verify assumptions $(i) - (iv)$ of Theorem 22. In the sequel, c will denote a numerical constant whose value is not necessarily the same in different lines. We first need to choose a pseudo-distance d on $\mathfrak{G}$. We use

$$d^2(f, f') = \mathbb{E}[(\ell(f) - \ell(f'))^2].$$

This makes assumption (i) trivially satisfied. Hypothesis (iii) (model-wise boundedness assumption) is also straightforward: for any $f \in \mathcal{F}_\lambda$,

$$|\ell(f)(x,y)| = |\phi(-yf(x))| \le \phi(\lambda),$$

so that hypothesis (iii) is satisfied with $b_k = \phi(\lambda_k)$.

The verification of hypothesis (ii) is summarized in the following Lemma.

Lemma 19 *Assume $\phi : \mathbb{R} \to \mathbb{R}_+$ is a twice differentiable, strictly increasing and strictly convex function. Denote*

$$L_\phi = 0 \vee \max_{x\in\mathbb{R}} \left(\frac{2(\phi'(x)+\phi'(-x))}{\frac{\phi''}{\phi'}(x) + \frac{\phi''}{\phi'}(-x)} - (\phi(x)+\phi(-x)) \right) .$$

If $L_\phi < \infty$, then for any function $f \in F_\lambda$, we have

$$\mathbb{E}[(\ell(f) - \ell(f^*))^2] \le (\phi(\lambda) + \phi(-\lambda) + L_\phi)\mathbb{E}[\ell(f) - \ell(f^*)] .$$

Thus, hypothesis (ii) holds with $C_k = (L_\phi + 2)\phi(\lambda_k)$.

Finally, we turn to hypothesis (iv) which contains the most information about the models. The goal is first to bound, for any $f_0 \in \mathcal{F}_\lambda$,

$$F_\lambda(r) = \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}_\lambda \\ d^2(f,f_0)\le r}} |(P-P_n)(\ell(f) - \ell(f_0))| \right],$$

where Pf and $P_n f$ denote the expectation of f under P and under the empirical probability distribution P_n, respectively. If we define the set of functions

$$\mathcal{G}_{\lambda,f_0} = \{\ell(f) - \ell(f_0) | f \in \mathcal{F}_\lambda\},$$

then

$$\begin{aligned} F_\lambda(r) &= \mathbb{E}\left[\sup_{\substack{g\in\mathcal{G}_{\lambda,f_0} \\ Pg^2\le r}} |(P-P_n)g| \right] \\ &\le \frac{2}{n}\mathbb{E}_P \mathbb{E}_\varepsilon \sup_{\substack{g\in\mathcal{G}_{\lambda,f_0} \\ Pg^2\le r}} \left| \sum_{i=1}^n \varepsilon_i g(X_i,Y_i) \right|, \end{aligned}$$

where the ε_i are i.i.d. Rademacher variables, by a standard symmetrization argument.

We use the following lemma (which is essentially the same as Lemma 2.5 in Mendelson (2002), except that we need to make some multiplicative factors explicit).

Lemma 20 (Mendelson). *Let $\mathcal{F}$ be a class of functions such that $\|f\|_\infty \le T$ for all $f \in \mathcal{F}$. Set $\tau^2 = \sup_{f\in\mathcal{F}} \mathbb{E}_P f^2$ and assume that for some $\gamma > 0$ and $p < 2$, for any empirical measure P_n,*

$$\log \mathcal{N}(\varepsilon, \mathcal{F}, L_2(P_n)) \le \gamma\varepsilon^{-p}.$$

(where $\mathcal{N}(\varepsilon, \mathcal{F}, L_2(P_n))$ denotes the ε-covering number of $\mathcal{F}$ with respect to the distance $L_2(P_n)$). Then, putting $B = \gamma^{\frac{1}{2}}(2-p)^{-1}$, we have

$$\frac{1}{\sqrt{n}}\mathbb{E}_P\,\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \leq c\max\left(B\tau^{\frac{2-p}{2}}, B^{\frac{4}{2+p}}T^{\frac{2-p}{2+p}}n^{-\frac{1}{2}\frac{2-p}{2+p}}\right)$$

To apply the lemma we need to estimate the entropy numbers of class $\mathcal{G}_{\lambda,f_0}$. First, since $\mathcal{C}$ is of finite dimension V, we have that for any empirical measure P_n,

$$\log\mathcal{N}(\varepsilon, \mathrm{conv}(\mathcal{C}), L^2(P_n)) \leq c\varepsilon^{-p},$$

where $p = \frac{2V}{V+2}$, as a consequence of Theorem 2.6.9 of van der Vaart and Wellner (1996, p.142). Now note that for a class of real functions $\mathcal{F}$ over $\mathcal{X}$, if we define $\mathcal{G}$ as the set of functions over $\mathcal{X}\times\{-1;1\}$ that can be written as $(x,y)\mapsto yf(x)$ for some $f\in\mathcal{F}$, then the covering numbers of $\mathcal{F}$ for $L^2(P_n)$ are the same as the covering numbers of set $\mathcal{G}$ for $L^2(Q_n)$ provided the marginal of Q_n on X is P_n.

Furthermore, functions in $\mathcal{F}_\lambda$ take values in $[-\lambda,\lambda]$, and ϕ has Lipschitz constant $\phi'(\lambda)$ on this interval. Therefore, by standard arguments (translation by a fixed function, dilation, application of a Lipschitz function, see, e.g., Pollard (1984) for the necessary tools), we have

$$\log\mathcal{N}(\varepsilon, \mathcal{G}_{\lambda,f_0}, L_2(P_n)) \leq \log\mathcal{N}\left(\frac{\varepsilon}{\lambda\phi'(\lambda)}, \mathcal{F}, L_2(P_n)\right) \leq c(\lambda\phi'(\lambda))^p\varepsilon^{-p}.$$

We can now apply Lemma 20 to the class $\mathcal{G}_{\lambda,f_0}$, with

$$\tau^2 = \sup_{\substack{g\in\mathcal{G}_{\lambda,f_0}\\ Pg^2\leq r}} Pg^2 \leq r\ ,$$

$T_\lambda = \phi(\lambda)$, and $\gamma_\lambda = c(\lambda\phi'(\lambda))^p$, so that we obtain, putting $B_\lambda = (\lambda\phi'(\lambda))^{p/2}(2-p)^{-1}$,

$$\frac{1}{\sqrt{n}}\mathbb{E}\sup_{\substack{g\in\mathcal{G}_{\lambda,f_0}\\ Pg^2\leq r}}\left|\sum_{i=1}^n\varepsilon_i g(X_i,Y_i)\right| \leq c\max\left(B_\lambda r^{\frac{2-p}{4}}, B_\lambda^{\frac{4}{2+p}}T_\lambda^{\frac{2-p}{2+p}}n^{-\frac{1}{2}\frac{2-p}{2+p}}\right). \tag{6}$$

To study the behavior of the last upper bound, we determine when then first term is dominant in the above max. This is the case when

$$r \geq (T_\lambda B_\lambda)^{\frac{4}{2+p}}n^{-\frac{2}{2+p}}\ . \tag{7}$$

Thus, if the above condition over r is satisfied, we have

$$F_\lambda(r) \leq \psi_\lambda(r) = \frac{A}{\sqrt{n}}B_\lambda r^{\frac{2-p}{4}}$$

for some numerical constant A that we can assume to be greater than 1, and ψ_λ is a sub-root function as requested.

Finally, the solution r^*_λ of the equation $\psi_\lambda(r) = r/C_\lambda$ is given by

$$r^*_\lambda = (AB_\lambda C_\lambda)^{\frac{4}{2+p}}n^{-\frac{2}{2+p}}.$$

For $\lambda=\lambda_k$, we take $C_{\lambda_k} = C_k = (2+L_\phi)\phi(\lambda_k)$ so that, since $A\geq 1$ and $C_k \geq T_{\lambda_k}$, condition (7) is ensured whenever $r\geq r^*_k = r^*_{\lambda_k}$. This concludes the check for hypothesis (iv).

To wrap up, hypotheses $(i)-(iv)$ of Theorem 22 are satisfied with the following choices

- $b_k = \phi(\lambda_k)$;
- $C_k = (L_\phi + 2)\phi(\lambda_k)$;
- $r_k^* = c(B_\lambda C_\lambda)^{\frac{4}{2+p}} n^{-\frac{2}{2+p}} = c((V+2)(L_\phi+2)\phi(\lambda))^{\frac{V+2}{V+1}}(\lambda\phi'(\lambda))^{\frac{V}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}$.

Eventually, set $x_k = \alpha \log \lambda_k$ which concludes the proof. ∎

PROOF OF LEMMA 19. It suffices to look at a fixed point x and to take the expectation as a final step. We therefore first omit the dependence on x to simplify the notation. Recall that if we denote $\eta = P(Y=1)$, then

$$f^*(\eta) = \arg\min_{\alpha\in\mathbb{R}} \{\eta\phi(-\alpha) + (1-\eta)\phi(\alpha)\}$$

is defined implicitly as the solution of

$$\eta\phi'(-f^*) = (1-\eta)\phi'(f^*) \ . \tag{8}$$

Since ϕ is strictly convex and increasing, $\phi'(x)/\phi'(-x)$ is increasing from $\mathbb{R}$ onto $\mathbb{R}_+$. It is then easy to deduce that f^* is an increasing function of η and that $f^*([0,1]) = \mathbb{R}$, so that f^* is invertible. Furthermore, by the implicit function theorem f^* is a differentiable function of η.

Consider some function $f \in \mathcal{F}_\lambda$ and put $\alpha = f(x)$ at the point x considered. Note that $|\alpha| \le \lambda$. Define

$$\begin{aligned} N(\eta,\alpha) &= \mathbb{E}_Y\left[(\ell(f) - \ell(f^*))^2\right] \\ &= \eta\left(\phi(-\alpha) - \phi(-f^*(\eta))\right)^2 + (1-\eta)\left(\phi(\alpha) - \phi(f^*(\eta))\right)^2 \end{aligned}$$

and

$$\begin{aligned} D(\eta,\alpha) &= \mathbb{E}_Y\left[\ell(f) - \ell(f^*)\right] \\ &= \eta\left(\phi(-\alpha) - \phi(-f^*(\eta))\right) + (1-\eta)\left(\phi(\alpha) - \phi(f^*(\eta))\right) . \end{aligned}$$

The goal is to show that $N \le C(\alpha)D$. To this end, first note that $N((f^*)^{-1}(\alpha),\alpha) = D((f^*)^{-1}(\alpha),\alpha) = 0$. We then compare the derivatives of N and D with respect to η. We have

$$\begin{aligned} \frac{\partial D}{\partial \eta} &= (\phi(-\alpha) - \phi(-f^*)) - (\phi(\alpha) - \phi(f^*)) + \left(\eta\phi'(-f^*) - (1-\eta)\phi'(f^*)\right)\frac{df^*}{d\eta} \\ &= (\phi(-\alpha) - \phi(-f^*)) - (\phi(\alpha) - \phi(f^*)) , \end{aligned}$$

using (8). Note that $\partial D/\partial\eta$ is therefore positive for $f^* \ge \alpha$ (or, equivalently, for $\eta \ge (f^*)^{-1}(\alpha)$) and negative otherwise. We now turn to the derivative of N:

$$\begin{aligned} \frac{\partial N}{\partial \eta} &= (\phi(-\alpha) - \phi(-f^*))^2 - (\phi(\alpha) - \phi(f^*))^2 \\ &\quad + 2\left(\eta\phi'(-f^*)(\phi(-\alpha) - \phi(-f^*)) - (1-\eta)\phi'(f^*)(\phi(\alpha) - \phi(f^*))\right)\frac{df^*}{d\eta} \\ &= (\phi(-\alpha) - \phi(-f^*) + \phi(\alpha) - \phi(f^*))\left(\phi(-\alpha) - \phi(-f^*) - (\phi(\alpha) - \phi(f^*))\right) \end{aligned}$$

$$
\begin{aligned}
&+ \left(\eta\phi'(-f^*) + (1-\eta)\phi'(f^*)\right)\left(\phi(-\alpha) - \phi(-f^*) - (\phi(\alpha) - \phi(f^*))\right)\frac{df^*}{d\eta} \\
&= \frac{\partial D}{\partial \eta}\left(\phi(\alpha) + \phi(-\alpha) + (\eta\phi'(-f^*) + (1-\eta)\phi'(f^*))\frac{df^*}{d\eta} - (\phi(f^*) + \phi(-f^*))\right),
\end{aligned}
$$

where the second equality follows from (8) again. If we now denote

$$
L_\phi = 0 \vee \max_{\eta\in[0,1]}\left((\eta\phi'(-f^*) + (1-\eta)\phi'(f^*))\frac{df^*}{d\eta} - (\phi(f^*) + \phi(-f^*))\right),
$$

we have, for all $\eta \geq (f^*)^{-1}(\alpha)$,

$$
\frac{\partial N}{\partial \eta} \leq (\phi(\alpha) + \phi(-\alpha) + L_\phi)\frac{\partial D}{\partial \eta},
$$

and the opposite inequality for $\eta \leq (f^*)^{-1}(\alpha)$. By integrating over η to the left or to the right of $(f^*)^{-1}(\alpha)$, we deduce that for any $\eta \in [0,1]$,

$$
N \leq (\phi(\alpha) + \phi(-\alpha) + L_\phi)D \leq (\phi(\lambda) + \phi(-\lambda) + L_\phi)D,
$$

where the second inequality follows from the convexity of ϕ. Integration over x leads to the desired inequality.

For a slightly more explicit expression of L_ϕ, note that by differentiating (8) we obtain

$$
\frac{df^*}{d\eta} = \frac{\phi'(-f^*) + \phi'(f^*)}{\eta\phi''(-f^*) + (1-\eta)\phi''(f^*)}.
$$

Then we can rewrite the ratio

$$
\begin{aligned}
\frac{\eta\phi'(-f^*) + (1-\eta)\phi'(f^*)}{\eta\phi''(-f^*) + (1-\eta)\phi''(f^*)} &= \frac{\frac{\eta}{1-\eta}\phi'(-f^*) + \phi'(f^*)}{\frac{\eta}{1-\eta}\phi''(-f^*) + \phi''(f^*)} \\
&= \frac{2}{\frac{\phi''}{\phi'}(-f^*) + \frac{\phi''}{\phi'}(f^*)},
\end{aligned}
$$

where we have used (8) again at the last line. This yields the expression for L_ϕ given in the statement of the Lemma.

sketch of the proof of lemma 20. Putting

$$
R_{n,P} = \frac{1}{\sqrt{n}}\mathbb{E}_P\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|,
$$

we have, following the proof of Lemma 2.5 in Mendelson (2002), and after applying standard chaining techniques (see Dudley, 1978) and contraction inequalities (see Ledoux and Talagrand, 1991),

$$
R_{n,P} \leq c\frac{\gamma^{\frac{1}{2}}}{2-p}(\tau^2 + Tn^{-\frac{1}{2}}R_{n,P})^{\frac{1}{2}(1-\frac{p}{2})}.
$$

(Note the slight difference as compared to Mendelson (2002) here as in this reference the author assumed $T = 1$). Now putting $B = \gamma^{\frac{1}{2}}(2-p)^{-1}$, we have

$$R_{n,P} \leq cB \max(\tau^2, Tn^{-\frac{1}{2}} R_{n,P})^{\frac{1}{2}(1-\frac{p}{2})}.$$

Now solving separately for the two terms of the above maximum, we obtain the conclusion. ∎

PROOF OF THEOREM 17. The theorem is a consequence of Theorem 21 quoted in Appendix A. The hypotheses to satisfy are exactly the same as for Theorem 1 (but for one single model $\mathcal{F}_\lambda$), so one can just recycle the previous proof. ∎

PROOF OF THEOREM 18. This theorem is again a consequence of Theorem 21 but this time we pick a different reference space $\mathfrak{G}$. We choose $\mathfrak{G} = \mathcal{F}_\lambda$ and denote $\overline{f}_\lambda = \arg\min_{f \in \mathcal{F}_\lambda} \mathbb{E}[\ell(f)]$. (Again, we suppose here that the above minimum is attained to simplify the argument; the proof may easily be adjusted accordingly if this is not the case.)

In this case hypothesis (iii) is changed as compared to the previous theorem. This, in turn, changes the definition of the factor C_λ and hence of r^*_λ. To check hypothesis (iii) we may apply directly Lemmas 15 and 16 from Bartlett, Jordan, and McAuliffe (2003). These imply that if ϕ has Lipschitz constant L on $[-\lambda, \lambda]$ and satisfies the uniform convexity assumption

$$\forall x, y \in [-\lambda, \lambda] \qquad \frac{\phi(x) + \phi(y)}{2} - \phi\left(\frac{x+y}{2}\right) \geq \delta(x-y)^2 \ ,$$

then for any $f \in \mathcal{F}_\lambda$,

$$\mathbb{E}[(\ell(f) - \ell(\overline{f}_\lambda))^2] \leq \frac{L^2}{2\delta} \mathbb{E}[\ell(f) - \ell(\overline{f}_\lambda)] \ .$$

In our setting we can take $L = \phi'(\lambda)$ and $\delta = cM(\lambda)$ (by second-order Taylor expansion). Thus we can take $\widetilde{C}_\lambda = c.\phi'(\lambda)^2 M(\lambda)^{-1}$. To satisfy hypothesis (iii) we use equation (6) again so that we can use the sub-root function

$$\widetilde{\psi}_\lambda(r) = c \max\left(B_\lambda r^{\frac{2-p}{4}} n^{-\frac{1}{2}}, B_\lambda^{\frac{4}{2+p}} T_\lambda^{\frac{2-p}{2+p}} n^{-\frac{2}{2+p}}\right),$$

with the same notation as in the proof of Theorem 1. Solving the equation $\widetilde{\psi}_\lambda(r) = r/\widetilde{C}_\lambda$, we then apply Theorem 21. The constant $C > 1$ appearing in that theorem can be taken arbitrarily close to 1, so that with probability $1 - \exp(-\delta)$ the following bound holds:

$$A(\widehat{f}_n^\lambda) - A(\overline{f}_\lambda) \leq R_2(\lambda, n) + \frac{b_2(\lambda)\delta}{n} \ ,$$

(where R_2 and b_2 are defined in the statement of the theorem). Adding $A(\overline{f}_\lambda) - A(f^*)$ on each side finishes the proof. ∎

Appendix A: general theorems for single-model and model selection estimator convergence

This section is devoted to recalling, in a compact version, the statements of abstract theorems appearing in Blanchard, Bousquet, and Massart (2003) (respectively: Proposition 1 and Theorem 7 in the latter reference).

Setup

We recall that $\mathcal{X}$ denotes a measurable feature space. Let $\ell(x,y) : \mathbb{R} \times \{-1,1\} \to \mathbb{R}$ be a loss function. Given a function $g : \mathcal{X} \to \mathbb{R}$, the notation $\ell(g)$ is used for the function $(x,y) \in \mathcal{X} \times \{-1,1\} \mapsto \ell(g(x),y)$. Let P be a probability distribution on $\mathcal{X} \times \{-1,1\}$ and $\mathfrak{G}$ a set of extended-real functions on $\mathcal{X}$ such that $\ell(\mathfrak{G}) \subset L_2(P)$. The target function g^* is defined as

$$g^* = \arg\min_{g \in \mathfrak{G}} P\ell(g)$$

and for any $g \in \mathfrak{G}$ we denote

$$L(g,g^*) = E[\ell(g)] - E[\ell(g^*)] \ .$$

Let $((X_i,Y_i))_{i=1,\dots,n}$ be an i.i.d. n-sample drawn from the probability distribution P and let P_n denote the associated empirical measure. For a real function f on $\mathcal{X} \times \{-1,1\}$, Pf is an alternative notation for $\mathbb{E}_P[f]$ (so that also $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i,Y_i)$). We say that a function $\psi : [0,\infty) \to [0,\infty)$ is *sub-root* if it is non-negative, non-decreasing, and if $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r > 0$.

Rate of convergence in a single model

Let $\mathcal{G}$ be a subset of $\mathfrak{G}$. The empirical risk minimization estimator over the model $\mathcal{G}$ is defined by

$$\widehat{g} = \arg\min_{g \in \mathcal{G}} P_n \ell(g).$$

Theorem 21 *Assume that there exists*

- *a pseudo-distance d on $\mathfrak{G}$*
- *a sub-root function ψ*
- *constants b and C*

such that

$$\begin{array}{lll} (i) & \forall g,g' \in \mathfrak{G}, & P(\ell(g)-\ell(g'))^2 \le d^2(g,g'); \\ (ii) & \forall g \in \mathcal{G}, & d^2(g,g^*) \le CL(g,g^*); \\ (iii) & \forall (x,y),\ \forall g \in \mathcal{G}, & |\ell(g(x),y)| \le b; \end{array}$$

and, if r^* denotes the solution of $\psi(r) = r/C$,

$$(iv) \qquad \forall g_0 \in \mathcal{G},\ \forall r \ge r^* \quad \mathbb{E}\left[\sup_{g \in \mathcal{G} : d^2(g,g_0) \le r} |(P-P_n)(\ell(g)-\ell(g_0))|\right] \le \psi(r).$$

Then for all $x > 0$ and all $K > 1$ the following inequality holds with probability at least $1 - e^{-x}$:

$$L(\widehat{g}, g^*) \leq \frac{K+1}{K-1}\left(\inf_{g \in \mathcal{G}} L(g, g^*) + 100K\frac{r^*}{C} + \frac{(2CK + 18b)x}{n}\right).$$

Model Selection Theorem (Deviation Bound)

Let $(\mathcal{G}_k)_{k\in\mathbb{N}}$ be a countable family of models with $\mathcal{G}_k \subset \mathfrak{G}$ for all $k \in \mathbb{N}$. If *pen* is a real function on $\mathbb{N}$, then the *penalized minimum empirical risk estimator* over the family of models is defined as

$$\widetilde{g} = \underset{\substack{k\in\mathbb{N},\\ g\in\mathcal{G}_k}}{\arg\min}\left(P_n \ell(g) + pen(k)\right).$$

Theorem 22 *Assume that there exist*

- *a pseudo-distance d on $\mathfrak{G}$;*
- *a sequence of sub-root functions (ψ_k);*
- *two real, nondecreasing sequences (b_k) and (C_k);*

such that

$$\begin{array}{lll}
(i) & \forall g, g' \in \mathfrak{G}, & P(\ell(g) - \ell(g'))^2 \leq d^2(g, g'); \\
(ii) & \forall k \in \mathbb{N},\ \forall g \in \mathcal{G}_k, & d^2(g, g^*) \leq C_k L(g, g^*); \\
(iii) & \forall k \in \mathbb{N},\ \forall g \in \mathcal{G}_k,\ \forall (x, y), & |\ell(g(x), y)| \leq b_k;
\end{array}$$

and, if r_k^* denotes the solution of $\psi_k(r) = r/C_k$,

$$(iv)\quad \forall k \in \mathbb{N},\ \forall g_0 \in \mathcal{G}_k,\ \forall r \geq r_k^* \quad \mathbb{E}\left[\sup_{\substack{g\in\mathcal{G}_k,\\ d^2(g,g_0)\leq r}} |(P - P_n)(\ell(g) - \ell(g_0))|\right] \leq \psi_k(r).$$

Let (x_k) be a nondecreasing sequence of real numbers such that $\sum_{k\in\mathbb{N}} e^{-x_k} \leq 1$. Let $\xi > 0, K > 1$ be some real numbers to be fixed in advance. If we define a penalty function $pen(k)$ such that

$$\forall k \in \mathbb{N} \qquad pen(k) \geq 250K\frac{r_k^*}{C_k} + \frac{(65KC_k + 56b_k)(x_k + \xi + \log(2))}{3n},$$

then, for the corresponding penalized minimum empirical risk estimator $\widetilde{g}$, the following inequality holds with probability greater than $1 - \exp(-\xi)$:

$$L(\widetilde{g}, g^*) \leq \frac{K + \frac{1}{5}}{K-1}\inf_{k\in\mathbb{N}}\left(\inf_{g\in\mathcal{G}_k} L(g, g^*) + 2\,pen(k)\right).$$

Appendix B: Details of Example 1.

First, we prove the following statements:

(i) If f is such that $\sum_{i=1}^4 f(x_i) = 0$, then $f \in \mathcal{F}_\lambda$ if and only if $\lambda \geq \frac{1}{2}\sum_i |f(x_i)|$.

(ii) If f^* (defined as in the rest of the paper) is such that $f^*(x_1) + f^*(x_2) = f^*(x_3) + f^*(x_4) = 0$, then so is $\overline{f}_\lambda$ for all λ.

PROOF. Denoting $z_i = f(x_i)$, we have $f = \sum_i z_i \mathbb{I}_{[x_i]}$. For all i, $\mathbb{I}_{[x_i]} = \frac{1}{2}(g_0 + g_i)$, and the linear relation $\sum_{i=1}^4 g_i + 2g_0 = 0$ holds, so that the only ways to write f as a combination of the base functions are exhaustively given by

$$f = \frac{1}{2}\left(\sum_{i=1}^4 (z_i + \mu) g_i + \left(\sum_i z_i + 2\mu\right) g_0\right), \ \mu \in \mathbb{R} .$$

If in addition we assume $\sum_i z_i = 0$, then the above combination is in $\mathcal{F}_\lambda$ for any $\lambda \geq \frac{1}{2}\sum_i |z_i + \mu| + |\mu|$. It is easily seen that the minimum value of this upper bound is obtained for $\mu = 0$. This proves (i).

For (ii), let $f \in \mathcal{F}_\lambda$. Consider f' obtained from f by switching its values on x_1, x_2 and x_3, x_4 respectively. Then $f' \in \mathcal{F}_\lambda$ by symmetry of $\mathcal{F}_\lambda$ and $A(f') = A(f)$ by the symmetry assumption on f^*. So $f'' = \frac{1}{2}(f + f') \in \mathcal{F}_\lambda$ by convexity of $\mathcal{F}_\lambda$ and $A(f'') \leq A(f)$ by convexity of ϕ; furthermore f'' satisfies the same symmetry relations as f. This proves (ii). ∎

Now for any $\lambda > 0$, put $x(\lambda) = \overline{f}_\lambda(x_1) \geq 0$ and $y(\lambda) = \overline{f}_\lambda(x_3) \geq 0$. Clearly these functions are increasing and hence almost everywhere differentiable functions. From (i) and (ii) we deduce that

$$\lambda = x(\lambda) + y(\lambda)$$

and that

$$A(\overline{f}_\lambda) = 2\left((0.5+\delta)\phi(-x(\lambda)) + (0.5-\delta)\phi(x(\lambda)) + \phi(-y(\lambda)\right) .$$

Differentiating these two equalities we get

$$y'(\lambda) + x'(\lambda) = 1,$$

and

$$\frac{dA(\overline{f}_\lambda)}{d\lambda} = 2\left(x'(\lambda)[(0.5-\delta)\phi'(x(\lambda)) - (0.5+\delta)\phi'(-x(\lambda))] - y'(\lambda)\phi'(-y(\lambda)\right) .$$

Clearly x' and y' must be such that $\frac{dA(\overline{f}_\lambda)}{d\lambda}$ is the lowest possible given the constraint $x' + y' = 1$. Therefore as long as $(0.5-\delta)\phi'(x(\lambda)) - (0.5+\delta)\phi'(-x(\lambda)) \geq -\phi'(-y(\lambda))$ we must have $x'(\lambda) = 0$ and $y'(\lambda) = 1$. Since $x(0) = y(0) = 0$ and $\phi'(0) = 1$, one deduces that as long as $\phi'(-\lambda) \geq 2\delta$, we have $y(\lambda) = \lambda, x(\lambda) = 0$.

Appendix C: refinement of the non-XOR condition in a continuous setup

In this section we give a slight refinement concerning Theorem 15. In this theorem the assumption is that no quadruple of points is in an XOR condition with respect to A and that $P(\delta A) = 0$. In that case the theorem says that we can approximate the indicator of A in the $L^1(P)$ sense by taking the sign of mixtures of stumps. We noticed that the result is unchanged if we replaced A by any set C such that $\mathrm{int}(A) \subset C \subset \overline{A}$ (where $\mathrm{int}(A)$ and $\overline{A}$ denote interior and closure for the usual topology on $[0,1]^d$). A natural idea would then be that the "non-XOR condition" should only be required for points not on the boundary of A, so that any problem arising with points on the boundary could be disregarded.

If such a results holds, it means that, assuming that any four points not on the boundary of A cannot be in a XOR position, there is a way of choosing a set C such that $\mathrm{int}(A) \subset C \subset \overline{A}$ and satisfying the full "non-XOR" requirement. The counterexample shown on the right-hand side of Figure 1 shows that we cannot expect such a result to hold in all generality, even if the boundary of A is of P-measure zero (consider the case where the center square is of P-measure 0, and the boundary of A is dense in this square).

Nevertheless, the following elementary topological lemma states that this result holds if we assume that δA is of Lebesgue measure zero.

Lemma 23 *Suppose that δA is of P-measure zero for some measure P having full support $\mathcal{X}$. Assume that there do not exist any four points $x, y, z, w \in \mathcal{X} \setminus \delta A$ in* XOR *position with respect to A. Then any four points $x, y, z, w \in \mathcal{X}$ cannot be in* XOR *position with respect to $C = \overline{\mathrm{int}(A)}$ (the closure of the interior of A).*

PROOF. Suppose that x_0, y_0, z_0, w_0 are in XOR position with respect to C, so that $x_0, w_0 \in C; y_0, z_0 \notin C$. We show this leads to a contradiction. Note that, if x, y, z, w satisfy (5), then knowing x, w and i_0 entirely determines y, z. Consider i_0 as fixed and denote the associated application (exchanging the i_0-th coordinates) $F : (x, w) \to (z, y) = F(x, w)$ from $\mathcal{X} \times \mathcal{X}$ into itself.

Let ε be a positive real and denote $B(u, \varepsilon)$ the open ε-ball centered in u. Denote $D_{x_0} = (B(x_0, \varepsilon) \cap \mathrm{int}(A))$ and $D'_{x_0} = D_{x_0} \setminus \delta A$. Since $x_0 \in C = \overline{\mathrm{int}(A)}$, D_{x_0} is a nonempty open set and thus $P(D_{x_0}) > 0$ since P has full support. Hence, $P(D'_{x_0}) = P(D_{x_0}) > 0$ since $P(\delta A) = 0$ and D'_{x_0} is also a nonempty open set. Define similarly D'_{w_0} and consider $H = F(D'_{x_0} \times D'_{w_0})$ and $H' = H \setminus (\delta A \times \delta A)$. H is a nonempty open set of $\mathcal{X} \times \mathcal{X}$ because F is a bicontinuous bijection, so $P \otimes P(H') = P \otimes P(H) > 0$, and therefore H' is non-void.

From this we deduce that there exist $(x, w) \in (B(x_0, \varepsilon) \cap \mathrm{int}(A)) \times (B(w_0, \varepsilon) \cap \mathrm{int}(A))$, and $(z, y) = F(x, w)$ such that x, y, z, w satisfy (5) and that none of these four points is in δA. This way we construct a sequence (x_n, y_n, z_n, w_n) of quadruples satisfying (5), and converging to (x, y, z, w) while staying outside of δA, with $x_n, w_n \in \mathrm{int}(A)$. By hypothesis (x_n, y_n, z_n, w_n) are not in a XOR position with respect to A, hence y_n or z_n must belong to $\mathrm{int}(A)$. Therefore either infinity many y_n's or infinitely many z_n's belong to $\mathrm{int}(A)$. Thus, y_0 or z_0 belongs to $\overline{\mathrm{int}(A)} = C$, in contradiction with the initial hypothesis. ∎

Acknowledgments

This paper was finished while the first author was a guest at the IDA group, Fraunhofer FIRST, Berlin. The work of the second author was supported by DGI grant BMF2000-0807.

References

A.R. Barron. Neural net approximation. In *Yale Workshop on Adaptive and Learning Systems*, 1992.

A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–944, 1993.

P.L. Bartlett, O. Bousquet and S. Mendelson. Localized Rademacher Complexities. *Proceedings of the 15th annual conference on Computational Learning Theory*, 44-58, 2002.

P.L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Manuscript*, 2003.

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research* 3, 463-482, 2002.

K. Bennett, A. Demiriz, and G. Rätsch. Sparse regression ensembles in infinite and finite hypothesis spaces. *Machine Learning*, 48:193–221, 2002.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* 4, 329–375, 1998.

O. Bousquet. New Approaches to Statistical Learning Theory. *Annals of the Institute of Statistical Mathematics*, 2003, to appear.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of Support Vector Machines. *Manuscript*, 2003.

L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–849, 1998.

L. Breiman. Some infinite theory for predictor ensembles. Technical Report 577, Statistics Department, UC Berkeley, 2000.

P. Bühlmann and B. Yu. Boosting with the l_2-loss: Regression and classification. *J. Amer. Statist. Assoc.*, 2003, to appear.

M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

G. Cybenko. Approximations by superpositions of sigmoidal functions. *Math. Control, Signals, Systems*, 2:303–314, 1989.

C. Darken, M. Donahue, L. Gurvits, and E. Sontag. Rates of Convex Approximation in Non-Hilbert Spaces. *Constructive Approximation*, 13(2):187-220, 1997.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

N. Duffy and D. Helmbold. Potential boosters? In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems* 12, pages 258-264. MIT Press, 2000.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.

F. Girosi and G. Anzelloti. Rates of convergence for radial basis functions and neural networks. *Artificial Neural Networks for Speech and Vision*, p.169-176, Chapman and Hall, 1993.

T. Hastie and R.J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, U. K., 1990.

K. Hornik, M. Stinchcombe, and H. White. Multi-layer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

W. Jiang. Process consistency for AdaBoost. *Annals of Statistics*, 2003, to appear (with discussion).

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

G. Lugosi and S. Mendelson. A note on the richness of convex hulls of VC classes. *Manuscript*, 2003.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, 2003, to appear (with discussion).

V. Maiorov, R. Meir, and J. Ratsaby. On the Approximation of Functional Classes Equipped with a Uniform Measure using Ridge Functions. *Jour. of Approximation Theory*, 99:95-111, 1999.

S. Mannor and R. Meir. Weak learners and improved convergence rate in boosting. In *Advances in Neural Information Processing Systems 13: Proc. NIPS'2000*, 2001.

S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, 2002.

P. Massart. Some applications of concentration inequalities in statistics. *Annales de la Faculté des Sciences de Toulouse, Mathématiques*, 9(2):245-303, 2000.

L. Mason, J. Baxter, P.L. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221–247. MIT Press, Cambridge, MA, 1999.

R. Meir and V. Maiorov. On the Optimality of neural network approximation using incremental algorithms. *IEEE Trans. Neural Network*, 11(2):323-337, 2000.

R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning*, LNCS, pages 119-184. Springer, 2003. (In press).

S. Mendelson. Improving the sample complexity using global data, *IEEE Transactions on Information Theory* 48(7), 1977-1991, 2002.

E. Nedelec and P. Massart. Risk bounds for statistical learning. *Manuscript*, 2003.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8;143-196, 1999.

D. Pollard. *Convergence of Stochastic Processes.* Springer-Verlag, New York, 1984.

R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

E. Sontag. Feedback Stabilization Using Two-Hidden-Layer Nets. *IEEE Trans. Neural Networks*, 3:981-990, 1992.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, to appear, 2003.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996.

Y. Yang. Minimax Nonparametric Classification-Part I: Rates of Convergence. *IEEE Transaction on Information Theory*, vol. 45, pp. 2271-2284, 1999.

Y. Yang. Minimax Nonparametric Classification-Part II: Model Selection for Adaptation. *IEEE Transaction on Information Theory*, vol. 45, pp. 2285-2292, 1999.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 2003, to appear (with discussion).