On the Bayes-risk consistency of regularized boosting methods

By Gábor Lugosi [1] and Nicolas Vayatis [2]

Department of Economics

Pompeu Fabra University

Barcelona

The probability of error of classification methods based on convex combinations of simple base classifiers by "boosting" algorithms is investigated. The main result of the paper is that certain regularized boosting algorithms provide Bayes-risk consistent classifiers under the only assumption that the Bayes classifier may be approximated by a convex combination of the base classifiers. Non-asymptotic distribution-free bounds are also developed which offer interesting new insight into how boosting works and help explain its success in practical classification problems.

1. Introduction One of the most important recent developments in the practice of classification has been the introduction of *boosting* algorithms which have had a remarkable performance on a large variety of classification problems. These algorithms aim at producing a combined classifier from a given class of *weak* (or

Received

AMS 2000 subject classifications. Primary

Key words and phrases.

[1]The work of the first author was supported by DGI grant BMF2000-0807

[2]The work of the second author was supported by a Marie Curie Fellowship of the European Community "Improving Human Potential" programme under contract number HPMFCT-2000-00667.

base) classifiers. While the corpus of empirical studies has reached impressive proportions, there are still few theoretical results to explain their efficiency. Originally (see, e.g., Schapire [31], Freund [16], Freund and Schapire [18]), boosting was considered as an iterative procedure which, given the training data, would at each step: (i) select a classifier from a given class of base classifiers, (ii) evaluate a weight for this classifier, (iii) output the weighted majority vote of the selected classifiers up to this step. The main idea in the updating rule of this procedure is to put a probability distribution on the sample points, starting with the uniform distribution at the initial step, and then to change this distribution at every step according to some rule reinforcing the probability associated to misclassified points along the process. This heuristic generated numerous variations on the choice of the initial pool of weak learners and on the way to perform the weight update. Generally speaking, the derived algorithms perform very well on most of the usual benchmark data sets in the sense that the generalization error decreases as successive iterations are run. However, the resistance to overfitting of the family of boosting methods have not found a fully satisfactory explanation so far (see Breiman [10], Freund, Mansour, Schapire [17]).

One of the most interesting attempts to explain the success of boosting methods points out that they tend to maximize the margin of the correctly classified points. These arguments are based on margin-based bounds for the probability of misclassification, see Schapire, Freund, Bartlett, and Lee [32], Koltchinskii and Panchenko [25]. However, as it was pointed out by Breiman [8], these bounds alone do not completely explain the efficiency of these methods (see also Freund and Schapire [21]). Boosting algorithms have also been found explicitly related to the stagewise fitting of additive logistic regression by Friedman, Hastie, and Tibshirani [22] and

Bühlmann and Yu [11]. This connection points out that boosting methods effectively minimize an empirical loss functional (different from the probability of misclassification). This property has also been pointed out in slightly different contexts by Breiman [9], Mason, Baxter, Bartlett, and Frean [30], and Collins, Schapire, and Singer [13].

The last observation places boosting into a much wider perspective and this is the one we adopt along the present paper. The main idea is to leave aside the sequential nature of the original boosting algorithm and to focus on the underlying optimization procedure which is implemented. It is now common knowledge that unregularized minimization of the empirical probability of misclassification is subject to overfitting and is computationally unfeasible. Even though overfitting can be avoided with regularized strategies for the misclassification error, this does not solve the computational difficulty. The advantage of replacing the empirical probability of misclassification by an appropriate smooth loss functional is to simultaneously avoid overfitting and become computationally feasible in many cases.

However, this leaves open the issue of consistency because it is not clear to what extent solving the approximated problem of the empirical functional minimization is equivalent to minimizing the generalization error. This paper considers the theoretical issue of the consistency of regularized boosting methods, and proposes elements of explanation for their efficiency in practice. We combine known techniques for deriving margin-based bounds with some new results to explain under which conditions minimizing a cost functional conducts to the Bayes risk, at least asymptotically. Our main result shows the existence of consistent regularized boosting strategies for classification which can then be implemented sequentially.

As it is mentioned above, several versions of boosting algorithms perform a gradient descent minimization of a convex empirical functional over the class of linear combinations of the base classifiers. The original versions, such as AdaBoost, do not put any restriction on the sum of the weights of the combined classifiers. As a result, if the algorithm is run for a sufficiently long time, the resulting classifier will inevitably overfit the data, and consistency is not achieved for most distributions. (Interestingly, however, due to the slow convergence of the gradient descent algorithm, this overfitting typically does not occur until a very large number of iterations.) In this paper we focus on a regularized version of boosting, suggested, for example, in [30]. In this version, the sum of the weights of the combined base classifiers is restricted to a fixed value, and gradient descent optimization is performed over the restricted class. The sum of the weights (which we denote by λ) plays the role of a regularization parameter, controlling a kind of bias/variance tradeoff. The main results show that λ may be chosen, as a function of the sample size, in such a way that Bayes-risk consistency of the resulting classifier is guaranteed. Data-dependent choices of λ are also considered.

Our approach is different from those that recommend early stopping of the AdaBoost algorithm to achieve regularization (see, e.g., Jiang [23]). To make the arguments clear, we assume that for each chosen value of λ, a separate optimization is performed to minimize the value of the cost function over the class of linear combinations with the sum of weights restricted to be λ. On the other hand, our experiments reported in Section 7 reveal a close connection between the two aproaches.

For previous work on Bayes risk consistency, we refer to Breiman [10], Bühlmann and Yu [12], Jiang [23], [24], Mannor and Meir [27], Mannor, Meir and Mendelson

[28]. For recent advances we refer to Mannor, Meir and Zhang [29], Steinwart [33], and Zhang [34].

The rest of the paper is organized as follows. The next section introduces the formal setting and notation. In Section 3 the simplest version of the method is defined and the consistency results are presented. In Section 4 the effect of the cost function is studied. Section 5 discusses variants of the main results based on data-dependent regularized choices of the regularization parameter λ. In Section 6 we point out that some of the bounds developed here show that the proposed regularized boosting algorithm minimizes the best Chernoff bound on the probability of error. In Section 7 we describe some interesting phenomena observed in experiments. Some proofs are postponed to the Appendix.

2. Setup and notation Consider the binary classification problem described as follows. Let $\mathcal{X}$ be a measurable feature space and let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \{-1, 1\}$. The random variable X models some observation and Y its unknown binary label. In the standard classification problem, the statistician is asked to construct a classifier $g_n : \mathcal{X} \to \{-1, 1\}$ which assigns a label to each possible value of the observation. The statistician has access to training data consisting 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 quality of g_n is measured by the loss

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

Ideally, $L(g_n)$ should be close, with large probability, to the Bayes risk, that is, to the minimal possible probability of error

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

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. Note that the infimum is achieved by the Bayes classifier[3] $g^*(x) = \mathbb{I}_{[\eta(x)>1/2]} - \mathbb{I}_{[\eta(x)\le 1/2]}$ (where $\mathbb{I}$ denotes the indicator function).

In recent years, a large part of research has been focused on classifiers which base their decision on a certain combination of very simple rules. Such rules include different versions of "boosting", "bagging", and "arcing" methods, see, for example, Breiman [4, 5, 7, 9], Freund and Schapire [16, 19, 20, 31]. To describe such averaging methods, consider a class of classifiers $\mathcal{C}$, where elements $g : \mathcal{X} \to \{-1, 1\}$ of $\mathcal{C}$ are called the *base classifiers*. We usually think of $\mathcal{C}$ as a class of simple rules such as all "decision stumps" (i.e., rules which split $\mathcal{X} = \mathbb{R}^d$ along a hyperplane parallel to the coordinate axes), or all binary trees with $d+1$ terminal nodes, but our results hold for more general families. In the general framework defined here we only assume that the VC dimension of $\mathcal{C}$ is finite.

Define $\mathcal{F}$ as the class of functions $f : \mathcal{X} \to [-1, 1]$ obtained as convex combinations of the classifiers in $\mathcal{C}$:

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{N} w_j g_j(x) \ : \ N \in \mathbb{N},\ w_1, \ldots, w_N \ge 0\,,\ \sum_{j=1}^{N} w_j = 1 \right\}.$$

Each estimator $f \in \mathcal{F}$ defines a classifier g_f, in a natural way, by

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

---

[3]Note that, with this convention, ties are broken in favor of -1, and this has no consequence on the value of the Bayes error.

To simplify notation, we write $L(f)$ for the probability of error $L(g_f)$ of the corresponding classifier. Similarly, introduce the empirical loss by

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

Observe that, for any $f \in \mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[f(X_i)Y_i<0]} \leq \widehat{L}(f) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[f(X_i)Y_i\leq 0]}$$

and similarly,

$$\mathbb{E}\mathbb{I}_{[f(X)Y<0]} \leq L(f) \leq \mathbb{E}\mathbb{I}_{[f(X)Y\leq 0]}.$$

To shorten notation, we define $Z(f) = -f(X)Y$ and $Z_i(f) = -f(X_i)Y_i$. Thus, minimization of the probability of error $L(f)$ over $f \in \mathcal{F}$ is approximately equivalent to the minimization of the expected value of the "cost function" $\mathbb{I}_{[\cdot>0]}$ of $Z(f)$. Since the expected value cannot be evaluated in the absence of the knowledge of the joint distribution of (X, Y), minimization of the empirical cost $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[Z_i(f)>0]}$ may be attempted as an approximation. However, in typical cases the class $\mathcal{F}$ of convex combinations is too large in the sense that minimization of the empirical cost over the whole class overfits the data and may yield a classifier with large probability of error. Indeed, even in the simplest cases the class of all g_f is easily seen to have an infinite VC dimension.

The estimators studied in this paper minimize a criterion which replaces the natural cost function (i.e., the indicator of misclassification $\mathbb{I}_{[-f(X)Y>0]}$) by a smooth convex cost function of the random variable $Z = -f(X)Y$. Indeed, various versions of boosting algorithms have been shown to minimize such cost functions, see, for example, Mason, Baxter, Bartlett, and Frean [30], Friedman, Hastie, and Tibshirani [22], Collins, Schapire, and Singer [13]. The usual choice is the exponential cost function, though other choices have also been proposed. The advantage of introducing such cost functions is twofold. First, the empirical optimization problem becomes tractable, since the objective function becomes convex. The second advantage is that the probability of error of the resulting classifier may be bounded nontrivially, something that cannot be done for the direct minimizer of the empirical probability of error. The first such result was pointed out by Schapire, Freund, Bartlett, and Lee [32] (interesting extensions are given by Blanchard [3]) and more explicitly by Koltchinskii and Panchenko [25], who showed that meaningful confidence bounds may be derived for the probability of error of the classifier minimizing a smooth upper bound. Even though, as Breiman [9] argues, the bounds in [32] alone do not explain the spectacular practical success of these algorithms, in this paper we show how these bounds may be used to prove Bayes-risk consistency.

In this paper all we assume about the classification algorithm is that the method at hand minimizes a smooth convex cost function, and we do not investigate the specific ways such algorithms are realized. For example, minimization based on gradient-descent methods leads to some standard versions of boosting algorithms, see Mason, Baxter, Bartlett, and Frean [30].

More specifically, let $\phi : [-1,1] \to \mathbb{R}^+$ be a positive nondecreasing convex function such that $\phi(0) = 1$. Introduce the notation

$$A(f) = \mathbb{E}\phi(Z(f)) \quad \text{and} \quad A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i(f)) \,.$$

Then clearly, since $\mathbb{I}_{[x>0]} \le \phi(x)$, we have $L(f) \le A(f)$ for all $f \in \mathcal{F}$. It is also clear that the empirical cost $A_n(f)$ is a convex function of the parameters $w_1, w_2, \ldots$ and therefore efficient algorithms are available for minimizing $A_n(f)$ over $f \in \mathcal{F}$. In fact, some versions of boosting, for example, the so-called L_1-boosting method in [30] minimize $A_n(f)$ using gradient descent algorithms. (We remark here that since $\mathcal{F}$ is an infinite-dimensional class, exact minimization of $A_n(f)$ is typically impossible, since A_n does not achieve its minimum in $\mathcal{F}$. However, the infimum may be approximated with arbitrary precision, and such an approximate minimization is sufficient for our purposes.)

The purpose of this paper is to investigate the probability of error $L(g_n)$ of classifiers $g_n = g_{\widehat{f}_n}$, where $\widehat{f}_n$ is obtained by (approximately) minimizing a convex cost functional $A_n(f)$ over $f \in \mathcal{F}$. The performance, of course, depends on the choice of ϕ. The first main result of this paper (Theorem 1) shows that ϕ may be chosen such that the probability of error $L(\widehat{f}_n)$ of the corresponding minimizer $\widehat{f}_n$ converges, almost surely, as $n \to \infty$, to the Bayes risk L^* for all distributions under the sole assumption that the class of all constant multiples of elements in $\mathcal{F}$ is dense in the class of all measurable functions which is easily seen to hold for some simple choices of the base class $\mathcal{C}$. The cost function is chosen from a one-dimensional family parameterized by the scale parameter λ. We offer several possible choices. The simplest choice selects λ before seeing the data, as a function of the sample size. For better performance, data-dependent regularization may be performed, which is detailed in Section 5.

We also derive some distribution-free non-asymptotic upper bounds for the probability of error of the classifier obtained by a regularized choice of the cost function. These bounds offer a new and interesting interpretation of what regularized boosting methods do: instead of minimizing the probability of error, they minimize the best Chernoff bound on this probability (see Section 6).

3. Consistency As indicated in the introduction, the main focus of this paper is on classifiers $g_{\widehat{f}_n}$ where the estimator $\widehat{f}_n$ minimizes the empirical quantity $A_n(f)$ based on some convex cost function ϕ. We begin this section by defining a prototype regularization procedure for the choice of the cost function and prove that the resulting classifier has a probability of error converging to the Bayes risk, almost surely.

To this end, let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a differentiable strictly increasing strictly convex function such that $\phi(0) = 1$, $\lim_{x\to-\infty} \phi(x) = 0$, and introduce, for all $\lambda > 0$,

$$\phi_\lambda(x) = \phi(\lambda x) \ .$$

Denote the empirical and expected loss functional associated with the cost function ϕ_λ by A_n^λ and A^λ, that is,

$$A_n^\lambda(f) = \frac{1}{n}\sum_{i=1}^n \phi_\lambda(Z_i(f)) \quad \text{and} \quad A^\lambda(f) = \mathbb{E}\phi_\lambda(Z(f)) \ .$$

Note that on $[-1,1]$ the function ϕ_λ is Lipschitz with constant $\lambda\phi'(\lambda)$. If $\lambda = 1$, we simply write $A_n(f)$ and $A(f)$ instead of $A_n^\lambda(f)$ and $A^\lambda(f)$. Observe that

$$A_n^\lambda(f) = A_n(\lambda f) \qquad A^\lambda(f) = A(\lambda f) \ .$$

REMARK. This simple observation highlights the two different interpretations of the scale parameter λ. One way of viewing λ reveals it as a parameter of the cost function we minimize. However, minimizing $A^\lambda(f)$ over $\mathcal{F}$ is equivalent to minimizing $A(f)$ over the scaled class $\lambda \cdot \mathcal{F}$. Though scaling the estimator f has no effect on the corresponding classifier (performance is unchanged since $L(f) = L(\lambda f)$ for all $\lambda > 0$), the introduction of the parameter λ is indeed a decisive step in designing consistent strategies. To understand the role of the parameter λ, consider the simple one-dimensional example when the base classifiers have the form $g(x) = 2\mathbb{I}_{[x<c]} - 1$ or $g(x) = 2\mathbb{I}_{[x\geq c]} - 1$ for some $c \in \mathbb{R}$. In this case the closure of the class $\lambda \cdot \mathcal{F}$ is just the class of functions with total variation bounded by 2λ. As the target estimator can be wildly oscillating and even have unbounded total variation, large values of λ offer more flexibility of approximation at the price of making the estimation problem more difficult. It is worth noting at this point

that the original AdaBoost algorithm attempts to minimize the functional A in the linear span of $\mathcal{C}$. In contrast to this, here we consider a family of optimization problems (minimizing various functionals A^λ) over the convex hull of $\mathcal{C}$.

Now let $\widehat{f}_n^\lambda$ denote a function in $\mathcal{F}$ which minimizes the empirical loss

$$A_n^\lambda(f) = \frac{1}{n}\sum_{i=1}^n \phi_\lambda(Z_i(f))$$

over $f \in \mathcal{F}$.

REMARK. Very often the functional A_n^λ does not achieve its minimum in $\mathcal{F}$. For convenience, we ignore this slight complication here, and simply mention that all arguments below work for any approximate minimizer, that is, if $\widehat{f}_n^\lambda$ is such that

$$A_n^\lambda(\widehat{f}_n^\lambda) \le \inf_{f\in\mathcal{F}} A_n^\lambda(f) + \epsilon_n$$

if ϵ_n is a sequence of positive numbers converging to zero.

The simplest version of the main consistency result of the paper is the following. The only assumption for the class $\mathcal{C}$ of base classifiers (apart from having a finite VC dimension) is that the union, for all $\lambda > 0$, of the classes of functions $\lambda \cdot \mathcal{F} = \{\lambda f : f \in \mathcal{F}\}$ is sufficiently rich to approximate the target function minimizing $A(f)$.

For the cost function ϕ we need the following properties.

ASSUMPTION 1. *Let ϕ be a differentiable strictly convex, strictly increasing cost function such that $\phi(0) = 1$, $\lim_{x\to-\infty}\phi(x) = 0$.*

THEOREM 1. *Assume that the cost function ϕ satisfies Assumption 1 and that the distribution of (X,Y) and the class $\mathcal{C}$ are such that*

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

where $A^* = \inf A(f)$ *over all measurable functions* $f : \mathcal{X} \longrightarrow \mathbb{R}$. *Assume that* $\mathcal{C}$ *has a finite* VC *dimension.*

Let $\lambda_1, \lambda_2, \ldots$ *be a sequence of positive numbers satisfying*

$$\lambda_n \to \infty \quad \textit{and} \quad \lambda_n \phi'(\lambda_n) \sqrt{\frac{\ln n}{n}} \to 0 \quad \textit{as } n \to \infty$$

and define the estimator $f_n = \widehat{f}_n^{\lambda_n} \in \mathcal{F}$. *Then* g_{f_n} *is strongly Bayes-risk consistent, that is,*

$$\lim_{n\to\infty} L(g_{f_n}) = L^* \quad \textit{almost surely.}$$

REMARK. (UNIVERSAL CONSISTENCY.) For many base classes $\mathcal{C}$, the first condition of the theorem is satisfied for all possible distributions of (X, Y). In Lemma 1 below we provide a simple sufficient condition for such a universal approximation property. In such cases, the classifier g_{f_n} is *strongly universally consistent*.

REMARK. (ASSUMPTION ON THE COST FUNCTION.) The cost function perhaps most widely used in practice is the exponential function, which clearly satisfies Assumption 1. Another important cost function meeting the conditions is the logit function $\log_2(1 + e^x)$. For more discussion on the choice of ϕ, we refer to Section 4.

REMARK. (DENSENESS ASSUMPTION.) The assumption on the class $\mathcal{C}$ given by $\lim_{\lambda\to\infty} \inf_{f\in\lambda\cdot\mathcal{F}} A(f) = A^*$ may be replaced by a simpler completeness condition such as in the argument of Breiman in [10]. Examples satisfying this condition are the class of indicators of all rectangles, indicators of halfspaces defined by hyperplanes, or the class of binary trees with a number of terminal nodes equal to the dimension d plus one (see [10]). We offer the following lemma to emphasize this remark:

Lemma 1. *Let the class $\mathcal{C}$ be such that its convex hull 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*

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

The proof relies on some properties of the minimizer of the functional A described in Lemma 3. Details can be found in the Appendix.

remark. (λ as a smoothing parameter.) The theorem requires that the cost function ϕ_λ become steeper as the sample size grows, but this steepness should grow at a controlled fashion. One may think about λ as a smoothing parameter. Larger values of λ penalize misclassification in the training sample more severely, and therefore have a tendency of overfitting. At the same time, large values of λ allow more flexibility in approximating the target function minimizing $A(f)$. Small values of λ produce smoother estimators. In this sense, λ may be regarded as a smoothing parameter, responsible for controlling the tradeoff between "bias and variance". As is always the case in nonparametric curve estimation, a data-dependent choice of the smoothing parameter is desirable for better performance. This may be performed in several different ways, one of which is discussed in Section 5.

The proof of Theorem 1 is based on a few simple lemmas. One of the main ingredients is the following result. It summarizes the "probabilistic" part of the argument.

Lemma 2. *For any n and $\lambda > 0$,*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| A^\lambda(f) - A_n^\lambda(f) \right| \le 4\lambda\phi'(\lambda)\sqrt{\frac{2V\ln(4n+2)}{n}}.$$

Also, for any $\delta > 0$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \left|A^\lambda(f) - A_n^\lambda(f)\right| \le 4\lambda\phi'(\lambda)\sqrt{\frac{2V\ln(4n+2)}{n}} + \lambda\phi'(\lambda)\sqrt{\frac{\ln(1/\delta)}{2n}} \; .$$

This lemma is a variation of a result of Koltchinskii and Panchenko [25]. For completeness, the proof is given in the Appendix.

Some basic properties of the minimizer are summarized in the following lemma. Introduce

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

Lemma 3 below implies that f^* is well-defined for all x with $\eta(x) \in (0,1)$.

LEMMA 3. *Let ϕ be a cost function satisfying Assumption 1. Consider either one of the following two cases:*

- *If $\eta(X) \notin \{0,1\}$ almost surely then, for each λ, there exists a unique measurable function f^*_λ, such that*

  $$A^\lambda(f^*_\lambda) \le A^\lambda(f) \quad \textit{for all functions } f.$$

  Then the classifier

  $$\begin{cases} 1 \textit{ if } f^*_\lambda(x) > 0 \\ -1 \textit{ otherwise} \end{cases}$$

  is just the Bayes classifier g^*. Also, for all $\lambda > 0$, $\lambda f^*_\lambda = f^*$.

- *If $\eta(X) \in \{0,1\}$ almost surely then, for each λ, we have*

  $$\inf_f A^\lambda(f) = 0 \, .$$

PROOF. Just note that

$$A^\lambda(f) = \mathbb{E}\{\eta(X)\phi_\lambda(-f(X)) + (1-\eta(X))\phi_\lambda(f(X))\}$$

and therefore the function f_λ^* defined, for all x, by

$$f_\lambda^*(x) = \arg\min_{\alpha \in \mathbb{R}} \{\eta(x)\phi_\lambda(-\alpha) + (1-\eta(x))\phi_\lambda(\alpha)\}$$

minimizes A^λ.

- If $\eta(x) \notin \{0,1\}$ then, since ϕ is increasing and strictly convex, the function $h(\alpha) = \eta(x)\phi_\lambda(-\alpha) + (1-\eta(x))\phi_\lambda(\alpha)$ has a unique minimum, and therefore $f_\lambda^*(x)$ is well-defined.

  Since ϕ is differentiable, the derivative of h is zero at $f_\lambda^*(x)$, and so

  $$\frac{\phi_\lambda'(-f_\lambda^*(x))}{\phi_\lambda'(f_\lambda^*(x))} = \frac{1-\eta(x)}{\eta(x)} \, .$$

  Clearly, $\eta(x) < 1/2$ if and only if $\phi_\lambda'(-f_\lambda^*(x)) > \phi_\lambda'(f_\lambda^*(x))$ and thus $f_\lambda^*(x) < 0$, by the strict convexity of ϕ_λ, proving the second statement.

  Finally, the equality above is equivalent to

  $$\frac{\phi'(-\lambda f_\lambda^*(x))}{\phi'(\lambda f_\lambda^*(x))} = \frac{1-\eta(x)}{\eta(x)}$$

  Since $\nu(\cdot) = \phi'(-\cdot)/\phi'(\cdot)$ is strictly monotone, this implies that $\lambda f_\lambda^* = f_1^* = f^*$.

- Now assume that $\eta(X)$ is 0 or 1 with probability one. Then, obviously the minimizer f_λ^* no longer exists. We consider the sequence of measurable functions f_n taking values $+n$ on $[\eta(X) = 1]$ and $-n$ on $[\eta(X) = 0]$ and we obtain that $A^\lambda(f_n) = \phi(-n)$. Taking the limit in n leads to $\inf_f A^\lambda(f) = 0$. ∎

Now, before considering the fundamental Lemma 5, we need an auxiliary result which characterizes the pointwise minimum of the cost functional as a function of the posterior probability function η. This quantity can be understood as an entropy measure (see Section 4).

LEMMA 4. *Let ϕ be a cost function satisfying Assumption 1. Then the function*

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(-\alpha) + (1-\eta)\phi(\alpha) \right) ,$$

defined for $\eta \in [0,1]$, is strictly concave, symmetric around $1/2$, and $H(0) = H(1) = 0$, $H(1/2) = 1$.

PROOF. Concavity follows from the fact that the infimum of concave functions is concave. By the convexity of ϕ, we have $(1/2)\phi(-\alpha) + (1/2)\phi(\alpha) \geq \phi(0)$ with equality achieved at $\alpha = 0$, which implies $H(1/2) = 1$. The rest of the properties are obvious. ∎

Our last key lemma, before turning to the proof of Theorem 1, shows that near optimization of the functional A yields nearly optimal classifiers.

LEMMA 5. *Let ϕ be a cost function satisfying Assumption 1. Let f_n be an arbitrary sequence of functions such that*

$$\lim_{n\to\infty} A(f_n) = A^* .$$

Then the classifier

$$g_{f_n}(x) = \begin{cases} 1 & \textit{if } f_n(x) > 0 \\ -1 & \textit{otherwise} \end{cases}$$

has a probability of error converging to L^*.

PROOF. In the proof we use the notation $g_n = g_{f_n}$. Recall that

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

(see, e.g., [14]). Fix $\delta > 0$ and define $S_\delta = \{x : \eta(x) \in [1/2-\delta, 1/2+\delta]\}$. Then clearly,

$$L(g_n) - L^* \leq 2\delta + \mathbb{P}\left\{X \notin S_\delta \text{ and } g_n(X) \neq g^*(X)\right\},$$

and therefore it suffices to show that $\mathbb{P}\{X \notin S_\delta \text{ and } g_n(X) \neq g^*(X)\} \to 0$ as $n \to \infty$. Proceeding by contradiction, assume that this statement is false. Then there exists a sequence of sets $K_n \subset S_\delta$ with $\liminf_{n\to\infty} \mathbb{P}\{X \in K_n\} > 0$ such that the estimators f_n and f^* lead to a different prediction. By symmetry, we may assume that on K_n, $f_n(x) > 0$ and $f^*(x) < 0$. Note that $f^*(x) < 0$ implies that $\eta < 1/2 - \delta$.

The difference $A(f_n) - A^*$ can be written as the expectation of a positive function since, by definition, the optimal $f^*(x)$ is the pointwise minimizer of the quantity $h(\alpha) = \eta(x)\phi(-\alpha) + (1-\eta(x))\phi(\alpha)$. Now write, for any set $B \subset \mathcal{X}$,

$$A|_B(f_n) = \mathbb{E}\{\mathbb{I}_B(X)\,(\eta(X)\phi(-f_n(X)) + (1-\eta(X))\phi(f_n(X)))\} \ .$$

We then have, for any B, $A(f_n) - A^* \geq A|_B(f_n) - A|_B(f^*)$.

On the one hand, since the function H defined in Lemma 4 is increasing in $[0, 1/2]$,

$$A|_{K_n}(f^*) = \mathbb{E}\{\mathbb{I}_{K_n}(X)H(\eta(X))\} \leq H(1/2 - \delta)\,\mathbb{P}\{K_n\} \ .$$

On the other hand, we note that h is a strictly convex function which has its minimum at $f^*(X)$. Therefore, considering that $f_n(X) > 0$ and $f^*(X) < 0$ on K_n, we have that $h(f_n(X)) > h(0) = \phi(0) = 1$. We then obtain

$$A|_{K_n}(f_n) = \mathbb{E}\{\mathbb{I}_{K_n}(X)h(f_n(X))\} \geq \mathbb{P}\{K_n\} \ .$$

Thus,

$$A(f_n) - A^* \geq A|_{K_n}(f_n) - A|_{K_n}(f^*) \geq (1 - H(1/2 - \delta))\,\mathbb{P}\{K_n\} \ .$$

Then because of the strict concavity of H (see Lemma 4), we have $\liminf_{n\to\infty} A(f_n) - A^* > 0$, which is a contradiction. ∎

PROOF OF THEOREM 1. Denote by $\overline{f}_\lambda$ an element of $\mathcal{F}$ which minimizes A^λ. Then we may write

$$\begin{aligned} A(\lambda_n f_n) - A^* &= \left(A(\lambda_n f_n) - A(\lambda_n \overline{f}_{\lambda_n})\right) + \left(A(\lambda_n \overline{f}_{\lambda_n}) - A^*\right) \\ &= \left(A^{\lambda_n}(\widehat{f}_n^{\lambda_n}) - A^{\lambda_n}(\overline{f}_{\lambda_n}\right) + \left(\inf_{f\in\lambda_n\cdot\mathcal{F}} A(f) - A^*\right) . \end{aligned}$$

Clearly, the second term on the right-hand side converges to zero by the assumption on $\mathcal{F}$ and since $\lambda_n \to \infty$. To bound the first term, simply note that

$$A^{\lambda_n}(\widehat{f}_n^{\lambda_n}) - A^{\lambda_n}(\overline{f}_{\lambda_n}) \leq 2\sup_{f\in\mathcal{F}} \left|A^{\lambda_n}(f) - A_n^{\lambda_n}(f)\right| ,$$

see, for example, [14, Lemma 8.2]. But by Lemma 2 this converges to zero with probability one, by the choice of λ_n.

Thus, we have that $A(\lambda_n f_n) \to A^*$ with probability one. The consistency result now follows by Lemma 5. ∎

4. Cost functions The choice of the cost function $\phi : \mathbb{R} \to \mathbb{R}^+$ may influence the performance of the obtained classifier. The inequality of Lemma 2 offers a guide for choosing ϕ: the smaller the quantity $\phi'(\lambda)$, the tighter the bound is. On the other hand, the choice is restricted by the requirement that ϕ is convex. In addition, Assumption 1 indicates that there are very few and elementary conditions needed for the cost function to make a consistent algorithm.

We mention here that Assumption 1 on the cost function ϕ in the consistency theorem may be replaced by the following alternative assumption, proposed by Zhang [34]:

ASSUMPTION 2. *Let ϕ be a convex cost function such that*

(i) $f^*(x) > 0$ if and only if $\eta(x) > 1/2$,

(ii) there exists a constant c and $s \geq 1$ satisfying, for any $\eta \in [0,1]$,

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

Then Theorem 1 remains true under this assumption as well. This follows from the fact that in the proof of Theorem 1, our Lemma 5 may be replaced the following result of Zhang [34]. Even though Zhang's assumption may be more difficult to interpret than Assumption 1, it may be used to derive rates of convergence via the lemma below. In this paper we do not investigate such rates.

LEMMA 6. *(Zhang [34].) Under Assumption 2, for any estimator f,*

$$L(f) - L^* \leq 2c(A(f) - A^*)^{1/s} \, .$$

Below we consider some specific choices of the cost function.

EXAMPLE 1. The standard choice in most boosting algorithms is $\phi(x) = \exp(x)$. This function obviously satisfies Assumption 1, and therefore consistency holds without any restriction on the distribution. Note that in this case $f^*(x) = \frac{1}{2}\ln\left(\frac{\eta(x)}{1-\eta(x)}\right)$ and $H(\eta) = 2\sqrt{\eta(1-\eta)}$.

EXAMPLE 2: Another important example is

$$\phi(x) = \text{logit}(x) = \log_2(1 + \exp(x))$$

considered in [22]. This cost function also satisfies Assumption 1 with $s = 2$. Observe that $f^*(x) = \ln\left(\frac{\eta(x)}{1-\eta(x)}\right)$ and $H(\eta) = -\eta\log_2\eta - (1-\eta)\log_2(1-\eta)$ is the binary entropy function.

EXAMPLE 3: Another interesting alternative is

$$\phi(x) = \psi(x) = \begin{cases} \exp(x) \ , \text{if } x < 0 \\ \\ x + 1 \quad , \text{if } x \geq 0 \end{cases}$$

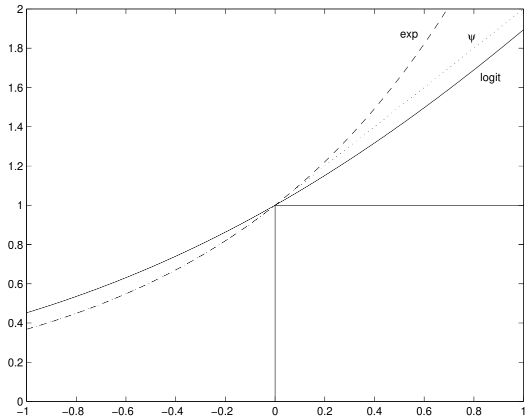

FIG. 1. *Bounding the step function with different cost functions* $\phi = \exp, \psi, logit$.

Here $f^*(x) = \ln\left(\frac{\eta(x)}{1-\eta(x)}\right)$ and $H(\eta) = \eta(2+\ln((1-\eta)/\eta))$ for $\eta \leq 1/2$. Even though the function ψ is not strictly convex, it is easy to see that the proof of Lemma 3 remains valid for this function as well, and consistency may be established.

Further investigation of the role of the cost function has been carried out very recently by Bartlett, Jordan, and McAuliffe [2].

The experiments of Section 7 show little influence of the specific form of the cost function on the performance of the algorithms for relatively small sample sizes. These cost functions emphasize differently data points according to the size of the margin $Yf(X)$ (see Figure 1).

5. Penalized model selection In Section 3 we established the existence of a consistent strategy based on regularized boosting methods. In these results, the sequence of estimators $f_n = \widehat{f}_n^{\lambda_n}$ requires, for each n, to minimize the functional $A_n^{\lambda_n}$ over $\mathcal{F}$, for a predetermined sequence $\lambda_1, \lambda_2, \ldots$. Of course, it is desirable to handle the choice of λ on the basis of the sample. The following theorem shows that consistency remains true for a data-dependent regularized choice of the smoothing parameter λ.

THEOREM 2. *Assume that the cost function ϕ satisfies Assumption 1 and that the distribution of (X,Y) and the class $\mathcal{C}$ are such that*

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

where $A^* = \inf A(f)$ over all measurable functions $f : \mathcal{X} \longrightarrow \mathbb{R}$. Let V denote the VC *dimension of the base class $\mathcal{C}$.*

For any divergent sequence $\lambda_1, \lambda_2, \ldots$ of positive numbers, let

$$f_n = \arg\min_{k\geq 1} \tilde{A}_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) ,$$

where

$$\tilde{A}_n^{\lambda_k}(f) = A_n^{\lambda_k}(f) + 4\lambda_k\phi'(\lambda_k)\sqrt{\frac{2V\ln(4n+2)}{n}} + \lambda_k\phi'(\lambda_k)\sqrt{\frac{\ln(6n^2k^2/\pi^2)}{n}} ,$$

and

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

Then f_n is strongly Bayes-risk consistent, that is,

$$\lim_{n\to\infty} L(f_n) = L^* \quad \textit{almost surely.}$$

The outline of the proof is the same as before. The next result replaces here Lemma 2 and provides an oracle inequality.

THEOREM 3. *For any sequence $\lambda_1, \lambda_2, \ldots$ of positive numbers, let*

$$f_n = \arg\min_{k\geq 1} \tilde{A}_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) ,$$

where

$$\tilde{A}_n^{\lambda_k}(f) = A_n^{\lambda_k}(f) + 4\lambda_k\phi'(\lambda_k)\sqrt{\frac{2V\ln(4n+2)}{n}} + \lambda_k\phi'(\lambda_k)\sqrt{\frac{\ln(6n^2k^2/\pi^2)}{n}} .$$

Then, with probability at least $1 - \frac{1}{n^2}$,

$$
\begin{aligned}
L(f_n) &\le \inf_{k \ge 1} A^{\lambda_k}(f_n) \\
&\le \inf_{k \ge 1} \left\{ A^{\lambda_k}(\overline{f}_{\lambda_k}) + 8\lambda_k \phi'(\lambda_k) \sqrt{\frac{2V \ln(4n+2)}{n}} + 2\lambda_k \phi'(\lambda_k) \sqrt{\frac{\ln(12n^2k^2/\pi^2)}{n}} \right\} .
\end{aligned}
$$

PROOF. For a given k, consider the minimizers

$$
\begin{aligned}
\widehat{f}_n^{\lambda_k} &= \operatorname*{arg\,min}_{f \in \mathcal{F}} A_n^{\lambda_k}(f) \\
\overline{f}_{\lambda_k} &= \operatorname*{arg\,min}_{f \in \mathcal{F}} A^{\lambda_k}(f) .
\end{aligned}
$$

By Lemma 2 we have, for all integers k, with probability at least $1 - \delta_k$, that for all $f \in \mathcal{F}$,

$$
A^{\lambda_k}(f) \le A_n^{\lambda_k}(f) + 4\lambda_k \phi'(\lambda_k) \sqrt{\frac{2V \ln(4n+2)}{n}} + \lambda_k \phi'(\lambda_k) \sqrt{\frac{\ln(1/\delta_k)}{2n}} .
$$

In particular, with probability at least $1 - \delta_k$,

$$
A^{\lambda_k}(\widehat{f}_n^{\lambda_k}) \le A_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) + 4\lambda_k \phi'(\lambda_k) \sqrt{\frac{2V \ln(4n+2)}{n}} + \lambda_k \phi'(\lambda_k) \sqrt{\frac{\ln(1/\delta_k)}{2n}} .
$$

Now take f_n to be

$$
f_n = \operatorname*{arg\,min}_{k \ge 1} \tilde{A}_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) ,
$$

where

$$
\tilde{A}_n^{\lambda_k}(f) = A_n^{\lambda_k}(f) + 4\lambda_k \phi'(\lambda_k) \sqrt{\frac{2V \ln(4n+2)}{n}} + \lambda_k \phi'(\lambda_k) \sqrt{\frac{\ln(1/\delta_k)}{n}} .
$$

Then, if we take $\delta_k = 6\delta/(\pi^2 k^2)$, we have, with probability $1 - \delta$,

$$
L(f_n) \le \inf_{k \ge 1} A^{\lambda_k}(f_n) \le \inf_{k \ge 1} \tilde{A}_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) .
$$

Since $A_n^{\lambda_k}(\widehat{f}_n^{\lambda_k}) \le A_n^{\lambda_k}(\overline{f}_{\lambda_k})$, by using again Lemma 2 as an upper bound on the empirical quantity $A_n^{\lambda_k}(\overline{f}_{\lambda_k})$, and the union bound, we get, for any k, with probability $1-\delta_k$,

$$A^{\lambda_k}(\widehat{f}_n^{\lambda_k}) \le A^{\lambda_k}(\overline{f}_k) + 8\lambda_k\phi'(\lambda_k)\sqrt{\frac{2V\ln(4n+2)}{n}} + 2\lambda_k\phi'(\lambda_k)\sqrt{\frac{\ln(2/\delta_k)}{2n}} .$$

Hence, we obtain that, with probability at least $1-\delta$,

$$\begin{aligned} L(f_n) &\le \inf_{k\ge 1} A^{\lambda_k}(f_n) \\ &\le \inf_{k\ge 1}\left\{A^{\lambda_k}(\overline{f}_{\lambda_k}) + 8\lambda_k\phi'(\lambda_k)\sqrt{\frac{2V\ln(4n+2)}{n}} + 2\lambda_k\phi'(\lambda_k)\sqrt{\frac{\ln((12k^2)/(\pi^2\delta))}{2n}}\right\} . \end{aligned}$$

Set $\delta = \frac{1}{n^2}$ to end the proof. ▮

PROOF OF THEOREM 2. We know by Theorem 3, that, for any integer k, we have

$$A^{\lambda_k}(f_n) \le A^{\lambda_k}(\overline{f}_{\lambda_k}) + 8\lambda_k\phi'(\lambda_k)\sqrt{\frac{2V\ln(4n+2)}{n}} + 2\lambda_k\phi'(\lambda_k)\sqrt{\frac{\ln(12n^2k^2/\pi^2)}{2n}} .$$

Moreover, by definition of $\overline{f}_{\lambda_k}$, we have that

$$A^{\lambda_k}(\overline{f}_{\lambda_k}) \le A^{\lambda_k}(f_n) .$$

As a consequence, taking the limit as n goes to infinity, in these two inequalities, by the Borel-Cantelli lemma we obtain

$$\lim_{n\to\infty}\inf_k A^{\lambda_k}(f_n) = \inf_k A^{\lambda_k}(\overline{f}_{\lambda_k}) \quad \text{almost surely.}$$

Since, by assumption,

$$\inf_k A^{\lambda_k}(\overline{f}_{\lambda_k}) = A^* ,$$

for some subsequence $\lambda_{k_1}, \lambda_{k_2}, \ldots$, we have that $A(\lambda_{k_n} f_n) \to A^*$ with probability one. The consistency result now follows by Lemma 5. ▮

REMARK. For a successful practical implementation of the regularized boosting procedure described above, the constants of the penalty need to be fine-tuned, an issue we do not address here. An alternative may be to hold out a small fraction of the data and choose the λ_k for which the estimate of $L(\widehat{f}_n^{\lambda_k})$ on the held-out sample is minimal. Consistency of this version is now a simple exercise. The experiments described in Section 7 may provide some insight.

REMARK. Independently of this work, Zhang (2001) also considers certain regularized boosting methods. However, Zhang is mostly concerned with the approximation error (except in some settings not considered here) and does not derive an oracle inequality as Theorem 3. For more recent results on the consistency of related methods, including support vector machines, the reader is referred to Bartlett, Jordan, and McAuliffe [2], Mannor, Meir and Zhang [29], Steinwart [33].

6. Regularized boosting minimizes Chernoff bounds The idea of combining simple classifiers via voting methods is often motivated by the argument that if one happens to find several "independent" classifiers such that all of them have a probability of error slightly better than random guessing, then taking a majority vote will radically decrease the probability of error. The improvement may easily be quantified by Chernoff bounds. The discussion we develop here was mainly inspired by Amit, Blanchard, and K. Wilder [1] and Blanchard [3].

We say that the classifiers $g_1, \ldots, g_N$ are independent if the events $[g(X) = Y]$ are independent. If a combined estimator f is obtained as the convex combination

$$f = \frac{1}{N} \sum_{j=1}^{N} g_j$$

where g_j are independent classifiers with error $L(g_j) \leq p < 1/2$, then by a classical Chernoff bound,

$$L(f) \leq \inf_{\lambda > 0} \left(p \exp(\lambda/N) + (1-p) \exp(-\lambda/N)\right)^N .$$

A straightforward computation shows that the optimal value of λ corresponds to

$$\lambda^* = \frac{N}{2} \log\left(\frac{1-p}{p}\right) .$$

yielding

$$L(f) \le \left(2\sqrt{p(1-p)}\right)^N .$$

The difficulty, of course, is to find, in a data-based manner, independent classifiers with a sufficiently small probability of error. The point of this section is that the regularized boosting method of Section 5 effectively finds such classifiers, provided that they exist in the convex hull of the base class $\mathcal{C}$.

In order to support this point, recall that, assuming that the exponential cost function is used, Theorem 3 shows that the probability of error of the classifier obtained with f_n minimizing the penalized cost function satisfies

$$L(f_n) \le \inf_{k \ge 1} \left\{ \inf_{f \in \mathcal{F}} \mathbb{E} \exp(-\lambda_k f(X)Y) + \epsilon_{n,k} \right\}$$

where for each k, $\lim_{n \to +\infty} \epsilon_{n,k} = 0$. Now it is clear that, if the sequence λ_k covers sufficiently densely the set of positive numbers, the upper bound is close to the best Chernoff bound on the generalization error which is given by

$$\inf_{\lambda > 0} \inf_{f \in \mathcal{F}} \mathbb{E} \exp(-\lambda f(X)Y) .$$

In particular, if the class $\mathcal{F}$ contains N independent classifiers with probability of error bounded by $p < 1/2$, then the quantity above is bounded by $\left(2\sqrt{p(1-p)}\right)^N$, an exponentially small quantity.

Note, however, that the assumption of the presence of such independent classifiers is very strong and it is more realistic to consider weaker concepts of independence as in [1].

Moreover, when $L(f) \ne 0$, there is no hope to obtain tight upper bounds with this method. For instance, assuming that $f(X) \in \{-1, +1\}$, we have, for a general cost function ϕ,

$$L(f) \le \mathbb{E}\phi(-\lambda f(X)Y) = (1 - L(f))\phi(-\lambda) + L(f)\phi(\lambda) .$$

Indeed, this upper bound is clearly suboptimal in the case when $L(f) \neq 0$, for all choices of λ.

7. Empirical study In this section, we propose an experimental study to understand the extent to which the theoretical analysis can be efficiently converted into practical strategies. Indeed, the results presented above show that there are two elements governing the consistency of regularized boosting methods: (i) the choice of the cost function ϕ, (ii) the tuning of the smoothing parameter λ. However, universal consistency (or particular non-consistency) can hardly be checked empirically. Therefore, we focus here on a rather qualitative analysis aiming at making clear that the performance of efficient model selection algorithms is rather sensitive to the tuning of the smoothing parameter λ depending on the noise level and on the difficulty of the classification problem. Note that a similar smoothing parameter is also studied in [12] and [30]. This leaves open the choice of the cost function though we provide some discussion below.

7.1. *Data sets* We have opted for a simple setting: we consider binary classification of artificial data and the weak learners are all decision stumps[4]. We used synthetic 6-dimensional data from the "twonorm", "threenorm" and "ringnorm" distributions (see in the Appendix for a description). These problems are expected to be of increasing difficulty for the class of convex combinations obtained from decision stumps. We considered relatively small sample sizes for the training set (n between 100 and 500).

[4]Decision stumps partition $\mathbb{R}^d$ along hyperplanes orthogonal to the axes.

7.2. *Algorithms* We have implemented the following algorithms described in [30] (to which we refer for detailed description and convergence properties):

- MarginBoost - Basically, the algorithm MarginBoost implements a gradient descent in the *linear span* of the class $\mathcal{C}$ to minimize a criterion of the form $\frac{1}{n}\sum_{i=1}^{n}\phi(-Y_i f(X_i))$, for $f = \sum_j w_j g_j$ with $g_j \in \mathcal{C}$. In this case, the parameter λ is interpreted as the sum of the unnormalized weights (their L_1-norm). Note that the original AdaBoost algorithm is a particular case of MarginBoost with exponential cost function.

- MarginBoost.L_1 - This algorithm implements a gradient descent in the *convex hull* of the class $\mathcal{C}$ to minimize $\frac{1}{n}\sum_{i=1}^{n}\phi(-\lambda Y_i f(X_i))$, for $f = \sum_j w_j g_j$ with $\sum_j w_j = 1$.

7.3. *Experiments* In the experiments we track generalization error and the optimal value of the cost functional as functions of the smoothing parameter λ, for fixed samples. More precisely, for each λ, the combined classifier $\widehat{f}_n^\lambda$ is constructed by the MarginBoost.L_1 algorithm after 300 iterations [5], on the basis of training samples of size n. We then estimate the expected cost $A^\lambda(\widehat{f}_n^\lambda)$ and the generalization error $L(\widehat{f}_n^\lambda)$ on a test set of size m.

A series of experiments were run to give some indications about:

- the influence of the choice of the cost function (see Figure 2)

  For a small sample size, the choice of a particular cost function appears to have a moderate impact on the generalization error performed by the corre-

[5] The number of iterations was chosen to be 300 because in our experiments convergence seemed to take place at such values. For larger sample sizes more iterations might be required to obtain a reliable output for MarginBoost.L_1.

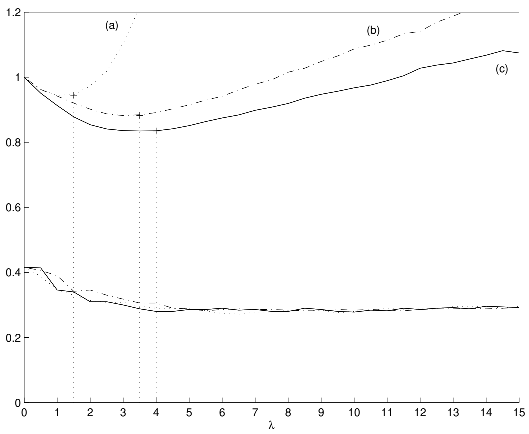

FIG. 2. *Threenorm.* $\eta = 0.1$. $n = 100$. $m = 500$. *Plots of the cost* $A^{\lambda}(\widehat{f}_n^{\lambda})$ *(upper curves) and test error (lower curves) for various cost functions (a)* exp, *(b) logit, (c)* ψ.

sponding boosting algorithm. In the sequel we report only experiments with cost function $\phi = \psi$ which seems to behave slightly better on this particular range of sample sizes.

- the comparison between test error and the oracle prediction (see Figures 2 and 3)

  We ran these experiments many times on artificial data sets, but also on real data, and we observed similar plots as the ones in Figure 3. All experiments show clearly the gap between the oracle inequality proved in Theorem 3 and the actual generalization error. This empirical evidence shows that it is not sufficient to rely on bounds in general. However, it seems that minimizing the cost functional gives a good estimate of the optimal generalization error even in the case where the sample size is small. Moreover, the fact that in many cases the generalization error remains quite flat independently of the values of λ gives some insight about why boosting algorithms behave so well in practice.

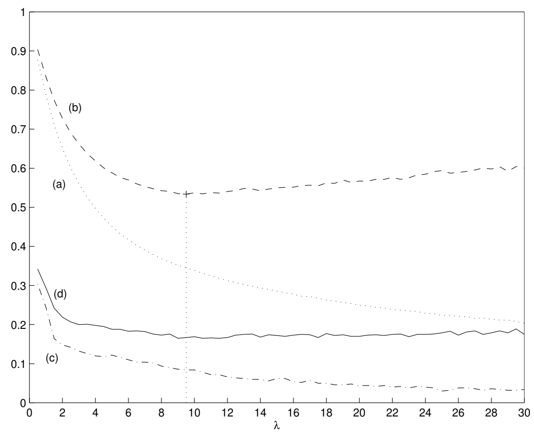

FIG. 3. *Threenorm. Cost $\phi = \psi$. $n = 500$. $m = 1000$. Plots of (a) $A_n^\lambda(\widehat{f}_n^\lambda)$. (b) $A^\lambda(\widehat{f}_n^\lambda)$. (c) training error. (d) test error.*

On the other hand, in almost all cases, for large values of λ an increase of the test error is present, making regularization necessary. This phenomenon tends to be more pronounced for complex and high-noise problems. Still, we feel that understanding the effect of the smoothing parameter λ in sample-based model selection needs further investigation.

- sensitivity to the level of label noise (see Figure 4)

  In these experiments, the observations vectors X_i are fixed and the labels Y_i are exposed to a constant level of noise η. The algorithms are run for different levels of label noise. The overfitting phenomenon can be observed even for small values of λ. The general effect is that the increase of the level of label noise η results in a decrease of the optimal λ. Moreover, the fact that the minimizer of the cost functional tracks so well the optimal classifier deserves to be mentioned.

- comparison with ADABOOST (see Figure 5)

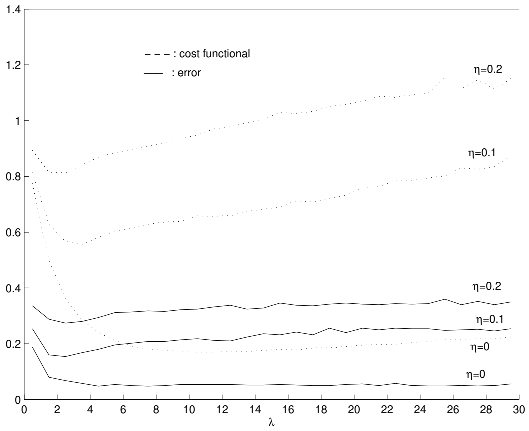

FIG. 4. *Twonorm. Cost $\phi = \psi$. $n = 100$. $m = 500$. Plots of $A^\lambda(\widehat{f}_n^\lambda)$ (dotted lines) and of the test error (solid lines) for levels of noise $\eta = 0, 0.1, 0.2$.*

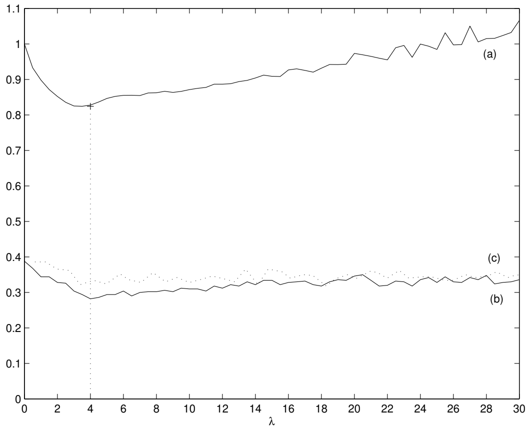

FIG. 5. *Threenorm. Cost $\phi = \psi$. $\eta = 0.1$. $n = 100$. $m = 500$. Plot of (a) $A^\lambda(\widehat{f}_n^\lambda)$ with* MARGINBOOST.L_1 *for various λ's, (b) test error with* MARGINBOOST.L_1 *for various λ's, (c) test error with one run of* MARGINBOOST *(unnormalized weights) where the test error is plotted as a function of the sum of the weights of the combined classifier denoted by λ_T.*

We think that these experiments provide some interesting insights on how the original AdaBoost algorithm works. Indeed, we can give a comparison by representing AdaBoost performance as a function of the norm of the weights in the combined classifier (instead of the number of iterations). Note that here, in order to make fair comparisons, we implemented AdaBoost using MarginBoost with cost function $\phi = \psi$. This algorithm constructs iteratively a combined classifier associated to the estimator $f_T = \sum_{t=1}^{T} w_t g_t$ with $g_t \in \mathcal{C}$ (step T) and w_t are positive weights (no normalization). Therefore, at each step T, MarginBoost outputs some element f_T of the class $\lambda_T \cdot \mathcal{F}$ where $\lambda_T = \sum_{t=1}^{T} w_t$. In Figure 5, we keep track of the test error of MarginBoost along the iterations with respect to λ_T. On this simple example, it turns out that AdaBoost constructs very quickly a classifier with the "optimal" complexity but that the intrinsic discretization of the method (at least in its original version) does not allow it to approximate the optimal generalization error too well.

Acknowledgments We would like to thank Leo Breiman for drawing our attention to the problems discussed here. We thank Peter Bickel for suggesting that our preliminary results might have lead to the proof of consistency. We also thank Peter Bartlett for stimulating discussions at an early stage of this work. We thank Gilles Blanchard who pointed out a simplification in the main proof of the paper and the referees for their helpful comments.

Appendix

A. Proof of Lemma 1 The lemma relies on the properties of the minimizer f^* of the functional A obtained in the proof of Lemma 3.

We introduce $H_\alpha = \{x : |\eta(x) - 1/2| > \alpha\}$ to denote the part of the domain where the noise level is low and $\overline{H}_\alpha$ its complementary set. We then consider the decomposition

$$A^* = A|_{H_\alpha}(f^*) + A|_{\overline{H}_\alpha}(f^*)$$

From the properties of f^* it follows that its restriction to $\overline{H}_\alpha$ is measurable, with range $[-\lambda_\alpha, \lambda_\alpha]$ where $\lambda_\alpha = f^*(1/2 + \alpha)$, and therefore can be approximated by a finite linear combination $\overline{f}_\alpha$ of indicator functions of disjoint sets from $\mathcal{B}_0$ such that

$$A|_{\overline{H}_\alpha}(\overline{f}_\alpha) - A|_{\overline{H}_\alpha}(f^*) < \epsilon$$

for any $\epsilon > 0$. Moreover, by Lemma 3, the term $A|_{H_\alpha}(f^*)$ is arbitrarily small when α goes to $1/2$. Note that λ_α goes to $+\infty$ as α goes to $1/2$ to complete the proof. ∎

B. Proof of Lemma 2 The proof of the first statement of Lemma 2 uses a standard symmetrization argument along with the contraction principle for Rademacher averages, just like Koltchinskii and Panchenko [25] did it in a similar setup.

- Step 1: Symmetrization.

  Define the i.i.d. pairs of random variables $(X_1', Y_1'), \ldots, (X_n', Y_n')$ where the (X_i', Y_i') are independent of the (X_i, Y_i) and have the same distribution as that of the pair (X, Y). Introduce also the i.i.d. symmetric sign variables $\sigma_1, \ldots, \sigma_n$ (i.e., $\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = 1\} = 1/2$), which are independent of all other random variables defined so far. Denote, for each $f \in \mathcal{F}$, $Z_i'(f) = -f(X_i')Y_i'$ and $A'^{\lambda}_n(f) = (1/n)\sum_{i=1}^n \phi_\lambda(Z_i'(f))$. Then, by a standard symmetrization argument,

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left|A^{\lambda}(f)-A_n^{\lambda}(f)\right| &\le \mathbb{E}\sup_{f\in\mathcal{F}}\left|A_n^{\lambda}(f)-A'^{\lambda}_n(f)\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\phi_\lambda\left(Z_i(f)\right)-\phi_\lambda\left(Z_i'(f)\right)\right)\right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\phi_k\left(Z_i(f)\right)-\phi_\lambda\left(Z_i'(f)\right)\right)\right| \\
&\le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\phi_\lambda\left(Z_i(f)\right)-1\right)\right| .
\end{aligned}$$

- Step 2: Contraction principle

The key of this step is the following version of the "contraction principle", see Ledoux and Talagrand [26], pages 112–113.

LEMMA 7. *If $\psi:\mathbb{R}\to\mathbb{R}$ is a Lipschitz function such that $|\psi(x)-\psi(y)|\le|x-y|$ with $\psi(0)=0$, then*

$$\mathbb{E}\sup_f\left|\sum_{i=1}^{n}\sigma_i\psi(Z_i(f))\right|\le 2\mathbb{E}\sup_f\left|\sum_{i=1}^{n}\sigma_i Z_i(f)\right| .$$

Applying Lemma 7 with the Lipschitz function $\psi(x)=(1/\lambda\phi'(\lambda))(\phi_\lambda(x)-1)$, we obtain

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left|A^{\lambda}(f)-A_n^{\lambda}(f)\right| &\le 4\lambda\phi'(\lambda)\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i Z_i(f)\right| \\
&= 4\lambda\phi'(\lambda)\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right|
\end{aligned}$$

where at the last step we used the fact that $-\sigma_i Y_i$ is a symmetric sign variable, independent of the X_i and therefore $-\sigma_i Y_i f(X_i)$ has the same distribution as that of $\sigma_i f(X_i)$.

- Step 3: Supremum computation

Note that the last expectation may be rewritten as

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right| = \frac{1}{n}\mathbb{E}\sup_{N\geq 1}\sup_{g_1,\ldots,g_N\in\mathcal{C}}\sup_{w_1,\ldots,w_N}\left|\sum_{i=1}^{n}\sum_{j=1}^{N} w_j\sigma_i g_j(X_i)\right|.$$

The key observation is that for any N and base classifiers $g_i,\ldots,g_N$, the supremum in

$$\sup_{w_1,\ldots,w_N}\left|\sum_{i=1}^{n}\sum_{j=1}^{N} w_j\sigma_i g_j(X_i)\right|$$

is achieved for a weight vector which puts all the mess in one index, that is, when $w_j = 1$ for some j. (This may be seen by observing that a linear function over a convex polygon achieves its maximum at one of the vertices of the polygon.) Thus,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right| = \frac{1}{n}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\sum_{i=1}^{n}\sigma_i g(X_i)\right| .$$

Finally, by a version of the Vapnik-Chervonenkis inequality (see [15, page 18])

$$\frac{1}{n}\mathbb{E}\sup_{g\in\mathcal{C}}\left|\sum_{i=1}^{n}\sigma_i g(X_i)\right| \leq \sqrt{\frac{2V\log(4n+2)}{n}}$$

which completes the proof of the first statement.

The second statement follows immediately by observing that, by McDiarmid's bounded difference inequality (see, e.g., Theorem 9.2 p.136 in [14]), for all $\epsilon > 0$,

(7.1)
$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}}\left|A^\lambda(f) - A_n^\lambda(f)\right| > \mathbb{E}\sup_{f\in\mathcal{F}}\left|A^\lambda(f) - A_n^\lambda(f)\right| + \epsilon\right\} \leq \exp\left\{-2n\left(\frac{\epsilon}{\lambda\phi'(\lambda)}\right)^2\right\}$$

which completes the proof of Lemma 2. ∎

C. Description of data sets The following data generators have now become a reference as benchmarks in classification. They were introduced by Breiman (see *e.g.* [6]).

- TWONORM: d-dimensional two-class data with $X \sim \mathcal{N}(m_1, \mathbf{I})$ for class $(+1)$, and $X \sim \mathcal{N}(m_2, \mathbf{I})$ for class (-1), where $m_1 = (a, \ldots, a)$ and $m_2 = (-a, \ldots, -a)$, and $a = 2/\sqrt{d}$.

- THREENORM: d-dimensional two-class data with $X \sim \frac{1}{2}\mathcal{N}(m_1, \mathbf{I}) + \frac{1}{2}\mathcal{N}(m_2, \mathbf{I})$ for class $(+1)$, and $X \sim \mathcal{N}(m_3, \mathbf{I})$ for class (-1), where $m_1 = (a, \ldots, a)$, $m_2 = (-a, \ldots, -a)$, $m_3 = (a, -a, a, -a, \ldots, -a)$ and $a = 2/\sqrt{d}$.

- RINGNORM: d-dimensional two-class data with $X \sim \mathcal{N}(0, 4\mathbf{I})$ for class $(+1)$, and $X \sim \mathcal{N}(m_1, \mathbf{I})$ for class (-1), where $m_1 = (a, \ldots, a)$, and $a = 1/\sqrt{d}$.

REFERENCES

[1] Y. Amit, G. Blanchard and K. Wilder. *Multiple Randomized Classifiers.* Submitted, 2001.

[2] P. Bartlett. Personal communication.

[3] G. Blanchard. *Méthodes de mélange et d'agrégation en reconnaissance de formes. Application aux arbres de décision.* PhD thesis, Université Paris XIII, 2001. In English.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[5] L. Breiman. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, April 1996.

[6] L. Breiman. Arcing the edge. Technical Report 486, Statistics Department, University of California, June 1997.

[7] L. Breiman. Pasting bites together for prediction in large data sets. Technical report, Statistics Department, University of California, July 1997.

[8] L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11:1493-1517, 1999.

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

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

[11] P. Bühlmann and B. Yu. Discussion of the paper "Additive Logistic Regression" by Jerome Friedman, Trevor Hastie and Robert Tibshirani. *The Annals of Statistics*, 28(2):377–

386, 2000.

[12] P. Bühlmann and B. Yu. Boosting with the L_2-Loss: Regression and Classification. Manuscript, August 2001.

[13] 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.

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

[15] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York, 2000.

[16] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, September 1995.

[17] Y. Freund, Y. Mansour, and R.E. Schapire. Why averaging classifiers can protect against overfitting. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, 2001.

[18] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148–146. Morgan Kaufmann, 1996.

[19] Y. Freund and R.E. Schapire. Game theory, on-line prediction and boosting. In *Proc. 9th Annu. Conf. on Comput. Learning Theory*, pages 325–332. ACM Press, New York, NY, 1996.

[20] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[21] Y. Freund and R.E. Schapire. Discussion of the paper "additive logistic regression: a statistical view of boosting" by J. Friedman, T. Hastie and R. Tibshirani. *The Annals of Statistics*, 28(2):391–393, 2000.

[22] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337-374, 2000.

[23] W. Jiang. Process consistency for adaboost. Technical Report 00-05, Department of Statistics, Northwestern University, November 2000.

[24] W. Jiang. Some theoretical aspects of boosting in the presence of noisy data. In Proceedings of The Eighteenth International Conference on Machine Learning (ICML-2001), June 2001, Morgan Kaufmann.

[25] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, Vol.30, No.1, 2002.

[26] M. Ledoux and M. Talagrand. *Probability in Banach Space.* Springer-Verlag, New York, 1991.

[27] 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.

[28] S. Mannor, R. Meir, and S. Mendelson. On the consistency of boosting algorithms. Manuscript, June 2001.

[29] 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*, Sidney, July 2002.

[30] 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.

[31] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[32] R.E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

[33] I. Steinwart. On the generalization ability of support vector machines. Manuscript, 2001.

[34] T. Zhang Statistical Behavior and Consistency of Classification Methods based on Convex Risk Minimization. Technical Report RC 22155, IBM Thomas Watson Research Center, Yorktown Heights, 2001.

Department of Economics

Pompeu Fabra University

Ramon Trias Fargas 25-27

08005 Barcelona, Spain.

{gabor.lugosi},{nicolas.vayatis}@econ.upf.es