The Consistency of Greedy Algorithms for Classification

Shie Mannor[1], Ron Meir[1], and Tong Zhang[2]

[1] Department of Electrical Engineering
Technion, Haifa 32000, Israel
(shie,rmeir)@(tx,ee).technion.ac.il
[2] IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
USA
tzhang@watson.ibm.com

Abstract. We consider a class of algorithms for classification, which are based on sequential greedy minimization of a convex upper bound on the $0-1$ loss function. A large class of recently popular algorithms falls within the scope of this approach, including many variants of Boosting algorithms. The basic question addressed in this paper relates to the statistical consistency of such approaches. We provide precise conditions which guarantee that sequential greedy procedures are consistent, and establish rates of convergence under the assumption that the Bayes decision boundary belongs to a certain class of smooth functions. The results are established using a form of regularization which constrains the search space at each iteration of the algorithm. In addition to providing general consistency results, we provide rates of convergence for smooth decision boundaries. A particularly interesting conclusion of our work is that Logistic function based Boosting provides faster rates of convergence than Boosting based on the exponential function used in AdaBoost.

1 Introduction

The Boosting [16] algorithm for classification, particularly AdaBoost [5] and its many variants, provides the Pattern Recognition practitioner with a large repertoire of useful algorithms. Following the successful application of Boosting across a broad spectrum of problems, it was realized that in spite of initial expectations, Boosting algorithms may overfit, and often badly so, under noisy conditions. Several authors suggested various approaches aimed at regularizing Boosting algorithms in order to prevent overfitting. A particularly interesting algorithm, termed LogitBoost, based on the logistic loss rather than the exponential loss used in AdaBoost, was proposed by Hastie *et al.* [6], who argued for its advantage over AdaBoost under noisy circumstances. Moreover, they provided an interpretation of Boosting in terms of additive models widely studied within the Statistics literature.

The present work considers a large class of algorithms, based on greedily minimizing a smooth convex loss function which upper bounds the $0-1$ loss

function used in classification. The utilization of a convex loss function serves two purposes. First, it greatly reduces the computational burden, as minimization of the $0-1$ loss is known to be intractable even for very simple (e.g., linear) classifiers. Second, the smoothness of the minimized loss function helps in forming a natural regularization procedure, which prevents overfitting. Since many recent approaches to Boosting and its variants are based on greedy algorithms of the above type (e.g., [4, 6, 13]), our results provide rigorous performance bounds for a large class of widely used learning algorithms.

The learning framework considered here is the standard supervised learning setup, where a learner attempts to select a hypothesis h from some space of hypotheses $\mathcal{H}$, based on a finite data set, generated according to some unknown probability distribution. The original work within the PAC framework, assumed that the 'true' hypothesis space was known, the only objective being to identify the correct hypothesis. Later work, extending this framework to the so-called *agnostic setup*, attempted to select an optimal hypothesis from the class $\mathcal{H}$ without making any assumptions about the true target. In this work we make rather general regularity assumptions about the data generating mechanism, but require that the algorithm produces an optimal hypothesis, where optimality is measured with respect to a large class of functions, as opposed to the relatively small classes of hypotheses considered within the agnostic framework.

Let $\mathcal{P}$ be a class of probability distributions, and denote by $\hat{f}_m$ a hypothesis produced by a learning algorithm based on m examples. Denoting the expected $0-1$ loss of a hypothesis f by $L(f)$, and by L^* the minimal loss attainable (assuming knowledge of the underlying law generating the data), we investigate conditions under which $L(\hat{f}_m)$ converges to L^* (in a well defined probabilistic sense), with increasing sample size m. Algorithms for which this occurs, without any assumptions on $\mathcal{P}$, are termed *universally consistent*. In general, we refer to algorithms for which $L(\hat{f}_m) \to L^*$ for all distributions in a class $\mathcal{P}$, as *consistent with respect to* $\mathcal{P}$.

For the class of greedy algorithms mentioned above, we provide conditions for consistency for large classes of probability distributions $\mathcal{P}$. Moreover, we establish finite sample bounds under certain smoothness conditions on $\mathcal{P}$.

Before moving to the detailed derivation of the results, we discuss some previous work related to the issues studied here. The question of the consistency of Boosting algorithms has attracted some attention in recent years. Jiang raised the questions of whether AdaBoost is consistent and whether regularization is needed. It was shown in [8] that AdaBoost is consistent at some point in the process of boosting. Since no stopping conditions were provided, this result essentially does not determine whether boosting forever is consistent or not. A one dimensional example was provided in [7], where it was shown that AdaBoost is not consistent in general since it tends to a Nearest neighbor rule. Furthermore, it was shown in the example that for noiseless situations AdaBoost is in fact consistent. The conclusion from this series of papers is that boosting forever for AdaBoost is not consistent and that sometimes along the boosting process a good classifier may be found.

In a recent paper Lugosi and Vayatis [10] also presented an approach to establishing consistency based on the minimization of a convex upper bound on the $0-1$ loss. According to this approach the convex cost function, is modified depending on the sample size. By making the convex cost function sharper as the number of samples increases, it was shown that the solution to the convex optimization problem yields a consistent classifier. Finite sample bounds are also provided in [10]. The issue of greedy algorithms, which is one of the main foci of this paper, was not considered in [10].

A different kind of consistency result was established by Mannor and Meir ([11, 12]). In this work geometric conditions needed to establish the consistency of boosting with linear weak learners were established. It was shown that if the Bayes error is zero (and the oppositely labelled points are well separated) then AdaBoost is consistent.

We summarize our main findings. We consider general two-category classification problems, where the classes may overlap. We find that for rather general smooth decision boundaries, running the greedy algorithm until convergence (termed 'Boosting forever' in [8]) leads to the best results, leading to universal consistency. Regularization is achieved by restricting certain parameters during the optimization process. In particular, we show that AdaBoost-type algorithms, based on linear classifiers as weak learners, are consistent if the Bayes decision boundary is smooth. However, approaches based on loss functions other than the exponential loss used in AdaBoost lead to improved upper bounds on the rates of convergence. It remains an open question to see whether the upper bounds we provide are indicative of the true behavior. These results lend theoretical support to the experimentally observed superiority of algorithms such as LogitBoost over AdaBoost in noisy situations.

2 Background and Preliminary Results

We begin with the standard formal setup for supervised learning. Let $(\mathcal{Z}, \mathcal{A}, P)$ be a probability space and let $\mathcal{F}$ be a class of $\mathcal{A}$ measurable functions from $\mathcal{Z}$ to $\mathbb{R}$. In the context of learning one takes $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. We let $D_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ denote a sample generated independently at random according to the probability distribution $P = P_{X,Y}$; in the sequel we drop subscripts (such as X, Y) from P, as the argument of P will suffice to specify the particular probability. In this paper we consider the problem of classification where $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{X} = \mathbb{R}^d$, and where the decision is made by taking the sign of a real-valued function $f(x)$. Consider the $0-1$ loss function given by

$$\ell(y, f(x)) = I[yf(x) \leq 0], \tag{1}$$

where $I[E]$ is the indicator function of the event E. Using the notation $\eta(x) \stackrel{\triangle}{=} P(Y = 1 | X = x)$, it is well known that L^*, the minimum of $L(f) = \mathbf{E}\ell(Y, f(X))$, can be achieved by setting $f(x) = 2\eta(x) - 1$. Note that the decision choice at

the point $f(x) = 0$ is not essential in the analysis. In this paper we simply assume that $I[0] = 1/2$, so that the decision rule $2\eta(x) - 1$ is Bayes optimal at $\eta(x) = 1/2$.

We consider a learning algorithm which, based on the sample D_m, selects a hypothesis $\hat{f}_m$ from some class of functions $\mathcal{F}$. The following definition is the standard definition of strong consistency for this setup.

Definition 1. *A classification algorithm is* strongly consistent *with respect to a class of distributions $\mathcal{P}$ if*

$$\lim_{m\to\infty} L(\hat{f}_m) = L^* \ ,$$

with probability one for any $P \in \mathcal{P}$. If $\mathcal{P}$ contains all Borel probability measures, we say that the algorithm is universally consistent.

In this work we show that algorithms based on greedily minimizing a convex upper bound on the $0-1$ loss are consistent with respect to the class of distributions $\mathcal{P}$, where certain regularity assumptions will be made concerning the conditional distribution $\eta(x) = P(y = 1|x)$. Denote the class of functions to which $\eta(x)$ belongs by $\mathcal{T}$ (the 'target' class). We start by proving universal consistency under certain conditions. We then refine our results, and impose certain smoothness conditions on the class of conditional densities. In this case, bounds on the convergence rates will be established.

We begin with a few useful results. Let $\{\epsilon_i\}_{i=1}^m$ be a sequence of binary random variables such that $\epsilon_i = \pm 1$ with probability $1/2$. The *Rademacher complexity* of $\mathcal{F}$ is given by

$$R_m(\mathcal{F}) \triangleq \mathbf{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i f(X_i) \right| ,$$

where the expectation is over $\{\epsilon_i\}$ and $\{X_i\}$, see [3] for some properties of $R_m(\mathcal{F})$.

The following theorem can be obtained by a slight modification of the proof of Theorem 1 in [9].

Theorem 1. (Adapted from Theorem 1 in [9])
Let $\{X_1, X_2, \ldots, X_m\} \in \mathcal{X}$ be a sequence of points generated independently at random according to a probability distribution P, and let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Furthermore, let ϕ be a non-negative Lipschitz function with Lipschitz constant κ, such that $\phi \circ f$ is uniformly bounded by a constant M. Then with probability at least $1-\delta$

$$\mathbf{E}\phi(f(X)) - \frac{1}{m}\sum_{i=1}^{m} \phi(f(X_i)) \leq 4\kappa R_m(\mathcal{F}) + M\sqrt{\frac{\log(1/\delta)}{2m}}$$

for all $f \in \mathcal{F}$.

For many function classes, $R_m(\mathcal{F})$ can be estimated directly. Results summarized in [3] are useful for bounding this quantity for algebraic composition of

function classes. We can also relate the Rademacher complexity, $R_m(\mathcal{F})$, to the ℓ_2^m covering number of $\mathcal{F}$. Let ℓ_2^m be the empirical ℓ_2 norm with respect to the data $\{X_1, X_2, \ldots, X_m\}$, namely $\ell_2^m(f,g) = \left(\frac{1}{m}\sum_{i=1}^m |f(X_i) - g(X_i)|^2\right)^{1/2}$. If $\mathcal{F}$ contains 0, then there exists a constant C such that (see Corollary 2.2.8 in [17])

$$R_m(\mathcal{F}) \leq \left(\mathbf{E}\int_0^\infty \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \ell_2^m)}\, d\epsilon\right)\frac{C}{\sqrt{m}}, \tag{2}$$

where $\mathcal{N}(\epsilon, \mathcal{F}, \ell_2^m)$ is the covering number of the class $\mathcal{F}$ at a scale of ϵ with respect to the ℓ_2^m norm, and the expectation is taken with respect to the choice of m points. We note that the approach of using Rademacher complexity and the ℓ_2^m covering number of a function class can often result in tighter bounds than some of the earlier studies that employed the ℓ_1^m covering number (for example, in [15]).

3 Consistency of Methods Based on Greedy Minimization of a Convex Upper Bound

Consider a class of so-called *weak learners* $\mathcal{H}$, and assume that it is closed under negation. We define the order t convex hull of $\mathcal{H}$ as

$$\mathrm{CO}_t(\mathcal{H}) = \left\{ f : \; f(x) = \sum_{i=1}^t \alpha_i h_i(x), \; \alpha_i \geq 0, \; \sum_{i=1}^t \alpha_i \leq 1, \; h_i \in \mathcal{H} \right\}.$$

The convex hull of $\mathcal{H}$, denoted by $\mathrm{CO}(\mathcal{H})$, is given by taking the limit $t \to \infty$. We are interested in algorithms that sequentially select a hypothesis from $\beta\mathcal{H}$, where β is a constant which will be specified at a later stage. We use the notation $\beta\mathcal{H} = \{\beta f : \; f \in \mathcal{H}\}$.

We assume throughout that functions in $\mathcal{H}$ take values in $[-1, 1]$. This implies that functions in $\beta\mathrm{CO}(\mathcal{H})$ take values in $[-\beta, \beta]$. Since the space $\beta\mathrm{CO}(\mathcal{H})$ may be huge, we consider algorithms that sequentially and greedily select a hypothesis from $\beta\mathcal{H}$. Moreover, since minimizing the $0-1$ loss is often intractable, we consider approaches which are based on minimizing a convex upper bound on the $0-1$ loss. The main contribution of this work is the demonstration of the consistency of such a procedure.

To describe the algorithm, let $\phi(x)$ be a strictly convex function, which upper bounds the $0-1$ loss, namely $\phi(yf(x)) \geq I[yf(x) \leq 0]$. Specific examples for ϕ are given in Section 3.1. Consider the empirical and true losses incurred by a function f based on the loss ϕ,

$$\begin{aligned}
\hat{A}(f) &\triangleq \frac{1}{m}\sum_{i=1}^m \phi(y_i f(x_i)) \; , \\
A(f) &\triangleq \mathbf{E}_{X,Y}\phi(Yf(X)) \; , \\
&= \mathbf{E}_X \left\{\eta(X)\phi(f(X)) + (1-\eta(X))\phi(-f(X))\right\} \; .
\end{aligned}$$

Where $E_{X,Y}$ is the expectation operator with respect to the measure P and E_X is the expectation with respect to the marginal on X.

Based on a finite sample D_m, we cannot hope to minimize $A(f)$ directly, but rather minimize its empirical counterpart $\hat{A}(f)$. Instead of minimizing $\hat{A}(f)$ directly, we consider a sequential greedy algorithm, which is described in Figure 1.

Input: A sample D_m; a stopping time t; a constant β
Algorithm:
1. Set $\hat{f}^0_{\beta,m} = 0$
2. For $\tau = 1, 2, \ldots, t$

$$\hat{h}_\tau, \hat{\alpha}_\tau, \hat{\beta}_\tau = \operatorname*{argmin}_{h\in\mathcal{H}, 0\le\alpha\le 1, 0\le\beta'\le\beta} \hat{A}((1-\alpha)\hat{f}^{\tau-1}_{\beta,m} + \alpha\beta' h)$$

$$\hat{f}^\tau_{\beta,m} = (1-\hat{\alpha}_\tau)\hat{f}^{\tau-1}_{\beta,m} + \hat{\alpha}_\tau\hat{\beta}_\tau\hat{h}_\tau$$

Output: Classifier $\hat{f}^t_{\beta,m}$

Fig. 1. A sequential greedy algorithm based on the convex empirical loss function $\hat{A}$.

We observe that many recent approaches to Boosting-type algorithms (e.g., [4, 6, 13]) are based on algorithms similar to the one presented in Figure 1. Two points are worth noting. First, at each step τ, the value of the previous composite hypothesis $\hat{f}^{\tau-1}_{\beta,m}$ is multiplied by $(1-\alpha)$, a procedure which is usually not followed in other Boosting-type algorithms; this ensures that the composite function at every step is in $\beta\mathrm{CO}(\mathcal{H})$. Second, the parameters α and β are constrained at every stage; this serves as a regularization measure and prevents overfitting.

It is clear from the description of the algorithm that $\hat{f}^t_{\beta,m} \in \beta\mathrm{CO}_t(\mathcal{H})$ for every t. Note that for fixed α and β, the function $\hat{A}((1-\alpha)\hat{f}^{\tau-1}_{\beta,m} + \alpha\beta h)$ is convex in h.

In order to analyze the behavior of the algorithm, we need several definitions. For $\eta \in [0,1]$ and $f \in \mathbb{R}$ let

$$G(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f).$$

Let $\mathbb{R}^*$ denote the extended real line ($\mathbb{R}^* = \mathbb{R} \cup \{-\infty, +\infty\}$). We extend a convex function $g : \mathbb{R} \to \mathbb{R}$ to a function $g : \mathbb{R}^* \to \mathbb{R}^*$ by defining $g(\infty) = \lim_{x\to\infty} g(x)$ and $g(-\infty) = \lim_{x\to-\infty} g(x)$. Note that this extension is merely for notational convenience. It ensures that the optimal minimizer $f_G(\eta)$ given below is well-defined at $\eta = 0$ or 1 for certain loss functions. For every value of $\eta \in [0,1]$ let

$$f_G(\eta) \triangleq \operatorname*{argmin}_{f\in\mathbb{R}^*} G(\eta, f) \quad ; \quad G^*(\eta) \triangleq G(\eta, f_G(\eta)) = \inf_{f\in\mathbb{R}^*} G(\eta, f).$$

It can be shown [19] that in many cases of interest, including the examples given in Section 3.1, $f_G(\eta) > 0$ when $\eta > 1/2$. We begin with a result from [19]. For completeness, we duplicate the proof in the appendix. Let f_β^* minimize $A(f)$ over $\beta\mathrm{CO}(\mathcal{H})$, and denote by f_{opt} the minimizer of $A(f)$ over all Borel measurable functions f.

Theorem 2. (Theorem 2.1 from [19]) *Assume that $f_G(\eta) > 0$ when $\eta > 1/2$, and that there exist $c > 0$ and $s \geq 1$ such that for all $\eta \in [0,1]$,*

$$|\eta - 1/2|^s \leq c^s (G(\eta, 0) - G^*(\eta)).$$

Then for all Borel measurable functions $f(x)$

$$L(f) - L^* \leq 2c \left(A(f) - A(f_{\mathrm{opt}})\right)^{1/s}, \tag{3}$$

where the Bayes error is given by $L^* = L(2\eta(\cdot) - 1)$.

The condition that $f_G(\eta) > 0$ when $\eta > 1/2$ in Theorem 2 ensures that the optimal minimizer f_{opt} achieves the Bayes error. This condition can be satisfied by assuming that $\phi(f) < \phi(-f)$ for all $f > 0$. The parameters c and s depend only on the loss ϕ. In general, if ϕ is second order differentiable, then one can take $s = 2$. Examples of the values of c and s are given in Section 3.1. The bound (3) allows one to work directly with the function A rather than with the less wieldy $0-1$ loss L.

We are interested in bounding the loss $L(f)$ of the empirical estimator $\hat{f}_{\beta,m}^t$ obtained after t steps of the sequential greedy algorithm described in Figure 1. Substitute $\hat{f}_{\beta,m}^t$ in (3), and consider bounding the r.h.s. as follows (ignoring the $1/s$ exponent for the moment):

$$\begin{aligned} A(\hat{f}_{\beta,m}^t) - A(f_{\mathrm{opt}}) = &\left[A(\hat{f}_{\beta,m}^t) - \hat{A}(\hat{f}_{\beta,m}^t)\right] + \left[\hat{A}(\hat{f}_{\beta,m}^t) - \hat{A}(f_\beta^*)\right] \\ &+ \left[\hat{A}(f_\beta^*) - A(f_\beta^*)\right] + \left[A(f_\beta^*) - A(f_{\mathrm{opt}})\right]. \end{aligned} \tag{4}$$

Next, we bound each of the terms separately.

The first term can be bounded using Theorem 1. In particular, since $A(f) = \mathbf{E}\phi(Yf(X))$, where ϕ is assumed to be convex, and since $\hat{f}_{\beta,m}^t \in \beta\mathrm{CO}(H)$ then $f(x) \in [-\beta, \beta]$ for every x. It follows that on its (bounded) domain the Lipschitz constant of ϕ is finite and can be written as κ_β (see explicit examples in Section 3.1). We have that with probability at least $1 - \delta$,

$$A(\hat{f}_{\beta,m}^t) - \hat{A}(\hat{f}_{\beta,m}^t) \leq 4\beta\kappa_\beta R_m(\mathcal{H}) + \phi_\beta \sqrt{\frac{\log(1/\delta)}{2m}},$$

where $\phi_\beta \stackrel{\triangle}{=} \sup_{f \in [-\beta,\beta]} \phi(f)$. Recall that $\hat{f}_{\beta,m}^t \in \beta\mathrm{CO}(\mathcal{H})$, and note that we have used the fact that $R_m(\beta\mathrm{CO}(\mathcal{H})) = \beta R_m(\mathcal{H})$ (e.g., [3]). The third term on the r.h.s. of (4) can be estimated directly from the Chernoff bound. We have with probability at least $1 - \delta$:

$$\hat{A}(f_\beta^*) - A(f_\beta^*) \leq \phi_\beta \sqrt{\frac{\log(1/\delta)}{2m}}.$$

Note that f^* is fixed (independent on the sample), and therefore a simple Chernoff bound suffices here. In order to bound the second term in (4) we assume that

$$\sup_{v\in[-\beta,\beta]} \phi''(v) \leq M_\beta < \infty, \tag{5}$$

where $\phi''(u)$ is the second derivative of $\phi(u)$.

From Theorem 4.2 in [20] we know that for a fixed sample

$$\hat{A}(\hat{f}^t_{\beta,m}) - \hat{A}(f^*_\beta) \leq \frac{8\beta^2 M_\beta}{t}.$$

This result holds for every convex ϕ and fixed β.

The fourth term in (4) is a purely approximation theoretic term. An appropriate assumption will need to be made concerning the Bayes boundary for this term to vanish.

In summary, for every t, with probability at least $1-2\delta$,

$$A(\hat{f}^t_{\beta,m}) - A(f_{\text{opt}}) \leq 4\beta\kappa_\beta R_m(\mathcal{H}) + \frac{8\beta^2 M_\beta}{t} + \phi_\beta\sqrt{\frac{2\log(1/\delta)}{m}} + (A(f^*_\beta) - A(f_{\text{opt}})). \tag{6}$$

The final term in (6) can be bounded using the Lipschitz property of ϕ. In particular,

$$\begin{aligned} A(f^*_\beta) - A(f_{\text{opt}}) &= \mathbf{E}_X\left\{\eta(X)\phi(f^*_\beta(X)) + (1-\eta(X))\phi(-f^*_\beta(X))\right\} \\ &\quad - \mathbf{E}_X\left\{\eta(X)\phi(f_{\text{opt}}(X)) + (1-\eta(X))\phi(-f_{\text{opt}}(X))\right\} \\ &= \mathbf{E}_X\left\{\eta(X)[\phi(f^*_\beta(X)) - \phi(f_{\text{opt}}(X))]\right\} \\ &\quad + \mathbf{E}_X\left\{(1-\eta(X))[\phi(-f^*_\beta(X)) - \phi(-f_{\text{opt}}(X))]\right\} \\ &\leq \kappa_\beta \mathbf{E}_X|f^*_\beta(X) - f_{\beta,\text{opt}}(X)| + \Delta_\beta, \end{aligned} \tag{7}$$

where the triangle inequality and the Lipschitz property were used in the final inequality. Here $f_{\beta,\text{opt}}(X) = \max(-\beta, \min(\beta, f_{\text{opt}}(X)))$ is the projection of f_{opt} onto $[-\beta,\beta]$, and

$$\Delta_\beta \triangleq \sup_{\eta\in[1/2,1]} \left\{I(f_G(\eta) > \beta)[G(\eta,\beta) - G(\eta, f_G(\eta))]\right\}.$$

Note that $\Delta_\beta \to 0$ when $\beta \to \infty$ since Δ_β represents the tail behavior $G(\eta,\beta)$. See examples in the next section.

3.1 Examples for ϕ

We consider three commonly used choices for the convex function ϕ.

$\exp(-x)$	Exponential
$\log(1+\exp(-x))/\log 2$	Logistic loss
$(x-1)^2$	Least squares

In this paper, the natural logarithm is used in the definition of logistic loss. The division by $\log 2$ sets the scale so that the loss function equals 1 at $x = 0$. For each one of these cases we provide in Table 1 the values of the constants M_β, ϕ_β, κ_β, and Δ_β defined above. We also include the values of c and s from Theorem 2, as well as the optimal minimizer $f_G(\eta)$. Note that the values of Δ_β and κ_β listed in Table 1 are upper bounds (see [19]).

$\phi(x)$	$\exp(-x)$	$\log(1+\exp(-x))/\log 2$	$(x-1)^2$
M_β	$\exp(\beta)$	$1/(4\log 2)$	2
ϕ_β	$\exp(\beta)$	$\log(1+\exp(\beta))/\log 2$	$(\beta+1)^2$
κ_β	$\exp(\beta)$	$1/\log 2$	$2\beta+2$
Δ_β	$\exp(-\beta)$	$\exp(-\beta)/\log 2$	$\max(0,1-\beta)^2$
$f_G(\eta)$	$\frac{1}{2}\log\frac{\eta}{1-\eta}$	$\log\frac{\eta}{1-\eta}$	$2\eta-1$
c	$1/\sqrt{2}$	$\sqrt{\log 2/2}$	$1/2$
s	2	2	2

Table 1. Parameter values for several popular choices of ϕ.

3.2 Universal consistency

We assume that $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$, which in turn implies that $0 \in \textsc{co}(\mathcal{H})$. This implies that $\beta_1 \textsc{co}(\mathcal{H}) \subset \beta_2 \textsc{co}(\mathcal{H})$ when $\beta_1 \leq \beta_2$. Therefore, using a larger β always gives us more approximation power.

We define $\textsc{span}(\mathcal{H}) = \cup_{\beta>0} \beta \textsc{co}(\mathcal{H})$, which is the largest function class that can be reached in the greedy algorithm by increasing β.

In order to establish universal consistency, we may assume initially that the class of functions $\textsc{span}(\mathcal{H})$ is dense in $C(K)$ — the class of continuous functions over a domain $K \subseteq \mathbb{R}^d$ under the uniform norm topology. From Theorem 4.1 in [19], we know that for all ϕ considered in this paper, and all Borel measures, $\inf_{f \in \textsc{span}(\mathcal{H})} A(f) = A(f_{\text{opt}})$. Since $\textsc{span}(\mathcal{H}) = \cup_{\beta>0} \beta \textsc{co}(\mathcal{H})$, we obtain $\lim_{\beta\to\infty} A(f^*_\beta) - A(f_{\text{opt}}) = 0$, leading to the vanishing of the final term in (6) when $\beta \to \infty$.

Theorem 3. *Assume that the class of functions* $\textsc{span}(\mathcal{H})$ *is dense in* $C(K)$ *over a domain* $K \subseteq \mathbb{R}^d$. *Assume further that* ϕ *is convex and Lipschitz and that (5) holds. Choose* β *such that as* $m \to \infty$, *we have* $\beta \to \infty$, $\phi_\beta^2 \log m/m \to 0$, *and* $\beta\kappa_\beta R_m(\mathcal{H}) \to 0$. *Then the greedy algorithm of Figure 1, applied for* t *steps where* $(\beta^2 M_\beta)/t \to 0$ *as* $m \to \infty$, *is strongly universally consistent.*

Proof. Let $\delta_m = \frac{1}{m^2}$. It follows from (6) that with probability smaller than $2\delta_m$

$$A(\hat{f}^{t_m}_{\beta,m}) - A(f_{\text{opt}}) > 4\beta_m \kappa_{\beta_m} R_m(\mathcal{H}) + \frac{8\beta_m^2 M_{\beta_m}}{t_m} + 2\phi_\beta \sqrt{\frac{\log m}{m}} + \Delta A_\beta,$$

where $\Delta A_\beta = A(f^*_\beta) - A(f_{\text{opt}}) \to 0$ as $\beta \to \infty$. Using the Borel Cantelli Lemma this happens finitely many times, so there is a (random) number of samples m_1 after which the above inequality is always reversed. Since all terms (6) converge to 0, it follows that for every $\epsilon > 0$ from some time on $A(\hat{f}^t_{\beta,m}) - A(f_{\text{opt}}) < \epsilon$ with probability 1. Using (3) concludes the proof. □

Unfortunately, no convergence rate can be established in the general setting of universal consistency. Convergence rates for particular functional classes can be derived by applying appropriate assumptions on the classes $\mathcal{H}$ and $\mathcal{T}$.

3.3 Consistency and convergence rates for smooth decision boundaries

We consider a special case here, where $\mathcal{H}$ is the class of ridge functions

$$\mathcal{H} = \left\{ f : f(x) = \alpha\sigma(w^T x + w_0),\ w \in \mathbb{R}^d, w_0 \in \mathbb{R}, \alpha \in [-1,1] \right\}, \tag{8}$$

often used in Neural Networks research. We assume that $\sigma(\cdot)$ is a non-decreasing bounded function taking values in $[-1,+1]$. Since it is well known that $\text{SPAN}(\mathcal{H})$ is dense in $C(K)$ for any compact set K, Theorem 3 implies that the greedy algorithm of Figure 1 is strongly universally consistent, when β is selected as in Theorem 3.

We show that a bound on the rate of convergence can be obtained under certain assumptions. In order to define the target class, we define smoothness of functions in terms of the integrability of high order derivatives. In particular, for a d-dimensional function $f(x)$, define

$$D^{\boldsymbol{k}} f(x) = \frac{\partial^{|\boldsymbol{k}|} f(x)}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $|\boldsymbol{k}| = k_1 + \cdots + k_d$. The Sobolev class (e.g., [1]) over a domain K is then defined for any natural number r as

$$W_p^r(K) = \left\{ f \ :\ \|f\|_{W_p^r(K)} = \max_{0 \le |\boldsymbol{k}| \le r} \|D^{\boldsymbol{k}} f\|_{L_p(K)} < \infty,\ r \in \mathbb{N} \right\}, \tag{9}$$

where the maximum is over all partitions of $\boldsymbol{k}$, $|\boldsymbol{k}| = k_1 + \cdots + k_d$, $k_i \ge 0$. In this work we assume $f_G(\mathcal{T}) = \{f_G(\eta) : \eta \in \mathcal{T}\}$ belongs to the Sobolev ball $B_2^r(K)$ characterized by $\|f\|_{W_2^r(K)} \le 1$ (any constant replacing 1 would lead to similar results). Furthermore, let $(u)_+ = \max(0, u)$. We quote a slightly revised version of a result from [14]. Observe that this result requires that the domain K is compact.

Theorem 4. (Adapted from Theorem 4.2 of [14]) *There exists a positive number $\bar{\beta}$ such that for all $\beta \geq \bar{\beta}$ and $f \in B_2^r(K)$, where $K \subset \mathbb{R}^d$ is a compact domain, and for every $\eta > 0$,*

$$\inf_{g \in \beta \mathrm{CO}(\mathcal{H})} \|g - f\|_{L_2(K)} \leq c\beta^{-a_r}, \tag{10}$$

where $a_r = \frac{1}{(d/2r-1)_+} + \eta$ and c is a constant depending on r, d and K.

Observe that $a_r = \infty$ if $r \geq d/2$, implying that if the target space $B_2^r(K)$ is characterized by highly smooth functions, i.e. $r \geq d/2$, then

$$\inf_{g \in \bar{\beta} \mathrm{CO}(\mathcal{H})} \|g - f\|_{L_2(K)} = 0, \tag{11}$$

for some finite number $\bar{\beta}$. Thus, functions belonging to $B_2^r(K)$, $r \geq d/2$, can be represented *exactly* by functions from the closure of $\bar{\beta}\mathrm{CO}(\mathcal{H})$ for some *finite* value of $\bar{\beta}$. We note in passing (see [1]) that the condition $r \geq d/2$ implies boundedness of functions in $B_2^r(K)$ in the supremum norm. Finally, if we assume that the underlying probability distribution can be represented as a probability density $P(x, y) = P(y|x)\mu(x)$, then it follows from the Cauchy-Schwartz inequality, assuming that $\int_K \mu(x)^2 dx$ is finite, that $E_X|g - f_{\beta,\mathrm{opt}}| \leq c\beta^{-a_r}$.

In order to establish rates we need to bound the Rademacher complexity of the class $\mathcal{H}$. First, let $\mathcal{F}$ be a class of bounded functions such that $f(x) \in [-1, +1]$ for every $f \in \mathcal{F}$ and $x \in K$. Assume further that $d_p = \mathrm{Pdim}(\mathcal{F})$ is finite, where $\mathrm{Pdim}(\mathcal{F})$ is the pseudo-dimension of $\mathcal{F}$. Then Theorem 2.6.7 in [17] implies that there is a universal constant K such that

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \ell_2^m) \leq K + \log d_p + d_p \log(16e) + 2(d_p - 1) \log \frac{1}{\epsilon}, \tag{12}$$

where $0 < \epsilon < 1$. Consider the class of linear functions $\mathcal{L} = \{w^T x + w_0 : w \in \mathbb{R}^d, w_0 \in \mathbb{R}\}$, for which it is known that $\mathrm{Pdim}(\mathcal{L}) = d+1$, where d is the Euclidean dimension of x. Since σ is monotonic, it follows from Theorem 11.3 in [2] that $\mathrm{Pdim}(\sigma(\mathcal{L})) \leq \mathrm{Pdim}(\mathcal{L})$. Since $\sigma(\cdot)$ is bounded, we conclude that (12) applies to the class of functions $\mathcal{H}$ given in (8). Using (2) and (12), and the observation that $\int_0^1 \sqrt{\log(1/\epsilon)} d\epsilon$ is finite, it follows that

$$R_m(\mathcal{H}) \leq C\sqrt{\frac{d}{m}}\,, \tag{13}$$

for some constant C. Combining our results we have from (4),(6),(10) and (13) that for β large enough with probability at least $1 - 2\delta$

$$A(\hat{f}_{\beta,m}^t) - A(f_{\mathrm{opt}}) \leq 4C\beta\kappa_\beta\sqrt{\frac{d}{m}} + \frac{8\beta^2 M_\beta}{t} + \phi_\beta\sqrt{\frac{2\log(1/\delta)}{m}} + c\kappa_\beta\beta^{-a_r} + \Delta_\beta. \tag{14}$$

It is interesting to observe the trade-off implicit in (14). First, in the case where $r \geq d/2$, the final approximation terms are absent from the bound. In this case,

simply set $\beta_m = \bar{\beta}$, and the bound is minimized. In the general case where $r < d/2$, a trade-off in the choice of β_m is evident. In order for the first term to vanish, we require that β_m increase slowly, while the final, approximation terms, require that β_m increase rapidly. The balance between the two terms determines the optimal rates. An additional consequence of (14) is that it is always best to choose the number of iterations t to be as large as possible. Regularization is incorporated through the proper selection of the value of β_m, rather than through the number of iterations.

Let $\beta = \beta_m$ be a monotonically increasing function of m, and let the number of iterations of the algorithm in Figure 1 be given by $t = t_m$. Select β_m and t_m as in Theorem 3, keeping (13) in mind. From Theorem 3 we immediately conclude that strong consistency holds with respect to the class $B_2^r(K)$.

Finally, we provide an explicit selection of β_m and t_m for which strong consistency holds, and for which convergence rates can be established for the three examples given in Section 3.1. In order to obtain bounds on the $0-1$ loss, we compute an upper bound on $\mathbf{E}L(\hat{f}^t_{\beta,m}) - L^*$. To do so, let X be a random variable satisfying $\mathbf{P}\{X > A + B\sqrt{\log(1/\delta)}\} < \delta$ for every $\delta > 0$. It can be shown that in this case $\mathbf{E}X \leq A + cB$ where $c \leq 2\sqrt{\log 2}$. The proof can be found in the appendix. Substituting the value of a from (14), recalling (3) and using Jensen's inequality $\mathbf{E}X^{1/s} \leq (\mathbf{E}X)^{1/s}$, $s \geq 1$, we have that

$$\mathbf{E}L(\hat{f}^t_{\beta,m}) - L^* \leq c_1 \left(\beta\kappa_\beta\sqrt{\frac{d}{m}}\right)^{1/s} + c_2 \left(\frac{\beta^2 M_\beta}{t}\right)^{1/s} + c_3 \left(\kappa_\beta \beta^{-a_r}\right)^{1/s} + c_4 \Delta_\beta^{1/s}, \tag{15}$$

for some constants c_1, c_2, c_3 and c_4, where we have used the inequality $(x+y)^\rho \leq (1/2)^\rho (x^\rho + y^\rho)$ for $x, y, \rho \geq 0$.

First, observe that for the exponential loss, because of the exponential dependence of κ_β on β, the third term in (15) can only vanish if $r \geq d/2$, implying that a fast convergence rate for the exponential loss can only be obtained in our setting if the boundary is sufficiently smooth. In the least squares case, the factor of κ_β in the penultimate term in (15) can be eliminated by noting that Theorem 4 applies directly to the L_2 loss, and using the results from Section 3.1 in [19]. For the logistic loss, since $\kappa_\beta = O(1)$, we find that the approximation term vanishes for every r.

Theorem 5. *Let K be a compact subset of $\mathbb{R}^d$ and suppose that $\int_K \mu(x)^2 dx < \infty$. Then the algorithm of Figure 1 is strongly consistent for $f_G(\mathcal{T}) = \{f_G(\eta) : \eta \in \mathcal{T}\} \subseteq B_2^r(K)$ under the conditions specified in Table 2, where $\beta_m \to \infty$ and $t_m \to \infty$ is assumed. The asymptotic convergence rates obtained from (15) are also given in Table 2. We use the notation $\tilde{O}(\cdot)$ in order to disregard logarithmic factors in m.*

Proof. The proof is a direct application of (15) where we select β_m so as to minimize the bound, and recall that $s = 2$ for all the loss functions considered. □

Loss	Smoothness	β_m	$\mathbf{E}L(\hat{f}^t_{\beta,m}) - L^*$
Exponential	$r \geq d/2$	$\log m^{\alpha}$, $(\alpha < 1/4)$	$\tilde{O}(1/m^{1/4-\alpha})$
LS	$r \geq d/2$	$\log m$	$\tilde{O}(1/m^{1/4})$
LS	$r < d/2$	$m^{(d-2r)/2d}$	$\tilde{O}(1/m^{r/2d})$
Logistic	$r \geq d/2$	$\log m$	$\tilde{O}(1/m^{1/4})$
Logistic	$r < d/2$	$m^{(d-2r)/2d}$	$\tilde{O}(1/m^{r/2d})$

Table 2. Rate bounds for popular choices of ϕ and Sobolev Bayes boundary.

4 Discussion

We have considered a class of sequential greedy algorithms, similar to AdaBoost, but differing in some implementation details. Our conclusions concerning the consistency of these algorithms can be described as follows. In all cases we have shown that "boosting forever", namely letting the number of Boosting iterations be infinite, is advantageous and does not lead to overfitting. Overfitting is prevented by regularization through the proper choice of the value of β in Figure 1. The upper bounds on the rates of convergence provided in Theorem 5 indicate faster rates of convergence for the logistic and least squares losses over that of the exponential loss. Note that the choice of β is Table 2 requires knowledge of the smoothness parameter r. It would be nice to extend the algorithm so that it is fully adaptive, not requiring knowledge of r. A further open question at this point is whether the bounds derived are indicative of the true behavior, a conclusion which can only be drawn if the asymptotic tightness of the bounds is established. In this context it would be particularly interesting to see whether the recently derived minimax rates for nonparametric classification [18] can be established for the algorithms analyzed in this paper.

A Appendix

Proof of Theorem 2

The classification error of f can be expressed as

$$L(f(\cdot)) = \mathbf{E}(1-\eta(x))I(f(x) \geq 0) + \mathbf{E}(\eta(x))I(f(x) \leq 0),$$

where the expectation is with respect to the input variable x. We thus have

$$\begin{aligned} L(f(\cdot)) - L^* &= L(f(\cdot)) - L(2\eta(\cdot) - 1) \\ &= \mathbf{E}_{2\eta(x)-1)f(x)\leq 0}|2\eta(x) - 1| \\ &\leq 2(\mathbf{E}_{(\eta(x)-1)f(x)\leq 0}|\eta(x) - 0.5|^s)^{1/s}. \end{aligned}$$

The last inequality follows from Jensen's inequality. Using the assumption of the theorem, we obtain

$$L(f(\cdot)) - L^* \leq 2c[\mathbf{E}_{(2\eta(x)-1)f(x)\leq 0}(G(\eta,0) - G^*(\eta))]^{1/s}. \tag{16}$$

If we can further show that $(2\eta - 1)f \leq 0$ implies $G(\eta, 0) \leq G(\eta, f)$, then

$$\mathbf{E}_{(2\eta(x)-1)f(x)\leq 0}(G(\eta, 0) - G^*(\eta)) \leq \mathbf{E}(G(\eta(x), f(x)) - G^*(\eta)) = A(f) - A(f_{\text{opt}}).$$

Combining this inequality with (16), we obtain the theorem. Therefore in the following we only need to prove the fact that $(2\eta - 1)f \leq 0$ implies $G(\eta, 0) \leq G(\eta, f)$. To see this, we consider the following three cases:

- $\eta > 0.5$: By assumption, we have $f_G(\eta) > 0$. Now $(2\eta - 1)f \leq 0$ implies $f \leq 0$. Using $0 \in [f, f_G(\eta)]$, and the convexity of $G(\eta, f)$ with respect to f, we obtain $G(\eta, 0) \leq \max(G(\eta, f), G(\eta, f_G(\eta))) = G(\eta, f)$.
- $\eta < 0.5$: Due to the symmetry, we can set $f_G(\eta) = -f_G(1 - \eta)$. Therefore $f_G(\eta) < 0$. Since $(2\eta - 1)f \leq 0$ implies that $f \geq 0$, we have $0 \in [f_G(\eta), f]$, which implies that $G(\eta, 0) \leq \max(G(\eta, f), G(\eta, f_G(\eta))) = G(\eta, f)$.
- $\eta = 0.5$: Note that we can take $f_G(\eta) = 0$ which implies $G(\eta, 0) \leq G(\eta, f)$ for all values f.

This completes the proof.

An auxiliary lemma

Lemma 1. *Let X be a non-negative scalar random variable for which $\mathbf{P}\{X > A + B\sqrt{\log(1/\delta)}\} < \delta$ for every $\delta > 0$. Then $\mathbf{E}X \leq A + cB$ where $c \leq 2\sqrt{\log 2}$.*

Proof. Recall that $\mathbf{E}X = \int_0^\infty P(X > x)dx$. Since $P(X > x)$ is a decreasing function of x we can bound the expectation as follows. Let $x_0 = 0$, and $x_i = A + B\sqrt{\log 2^i}$ for $i = 1, 2, \ldots$. Obviously $x_i \nearrow \infty$. Bounding the integral one gets that

$$\begin{aligned}
\mathbf{E}X &\leq \sum_{i=0}^{\infty} P(X > x_i)(x_{i+1} - x_i) \\
&= A + B\sqrt{\log 2} + \sum_{i=1}^{\infty} P(X > x_i)(x_{i+1} - x_i) \\
&= A + B\sqrt{\log 2} + B\sqrt{\log 2}\sum_{i=1}^{\infty} \frac{1}{2^i}(\sqrt{i+1} - \sqrt{i}) \\
&\leq A + (2\sqrt{\log 2})B\,.
\end{aligned}$$

□

Acknowledgements The authors are grateful to Shahar Mendelson for helpful discussions, and to Gabor Lugosi for sending them a copy of his work prior to publication. The work of R.M. was partially supported by the Technion fund for promotion of research. Support from the Ollendorff foundation at the Technion is also gratefully acknowledged.

References

1. R.A. Adams. *Sobolev Spaces.* Academic Press, New York, 1975.
2. M. Anthony and P.L. Bartlett. *Neural Network Learning; Theoretical Foundations.* Cambridge University Press, 1999.
3. P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, pages 224–240, 2001.
4. L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998.
5. Y. Freund and R.E. Schapire. A decision theoretic generalization of on-line learning and application to boosting. *Comput. Syst. Sci.*, 55(1):119–139, 1997.
6. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 38(2):337–374, 2000.
7. W. Jiang. Does boosting overfit: Views from an exact solution. Technical Report 00-03, Department of Statistics, Northwestern University, 2000.
8. W. Jiang. Process consistency for adaboost. Technical Report 00-05, Department of Statistics, Northwestern University, 2000.
9. V. Koltchinksii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statis.*, 30(1), 2002.
10. G. Lugosi and N. Vayatis. On the bayes-risk consistency of bosting methods. Technical report, Pompeu Fabra University, 2001.
11. S. Mannor and R. Meir. Geometric bounds for generlization in boosting. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, pages 461–472, 2001.
12. S. Mannor and R. Meir. On the existence of weak learners and applications to boosting. *Machine Learning*, 2002. To appear.
13. L. Mason, P. Bartlett, J. Baxter, and M. Frean. Functional gradient techniques for combining hypotheses. In B. Schölkopf A. Smola, P. Bartlett and D. Schuurmans, editors, *Advances in Large Margin Classifiers.* MIT Press, 2000.
14. R. Meir and V. Maiorov. On the optimality of neural network approximation using incremental algorithms. *IEEE Trans. Neural Networks*, 11(2):323–337, 2000.
15. D. Pollard. *Convergence of Empirical Processes.* Springer Verlag, New York, 1984.
16. R. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.
17. A.W. van der Vaart and J.A. Wellner. *Weak Convergence and EmpiricalProcesses.* Springer Verlag, New York, 1996.
18. Y. Yang. Minimax nonparametric classification - patr i: rates of convergence. *IEEE Trans. Inf. Theory*, 45(7):2271–2284, 1999.
19. T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. Technical Report RC22155, IBM T.J. Watson Research Center, Yorktown Heights, 2001.
20. T. Zhang. Sequential greedy approximation for certain convex optimization problems. Technical Report RC22309, IBM T.J. Watson Research Center, Yorktown Heights, 2002.