Journal of Machine Learning Research 4 (2003) 713-741 Submitted 11/02; Published 10/03

Greedy Algorithms for Classification – Consistency, Convergence Rates, and Adaptivity

Shie Mannor SHIE@MIT.EDU
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion, Haifa 32000, Israel

Tong Zhang TZHANG@WATSON.IBM.COM
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

Editor: Yoram Singer

Abstract

Many regression and classification algorithms proposed over the years can be described as greedy procedures for the stagewise minimization of an appropriate cost function. Some examples include additive models, matching pursuit, and boosting. In this work we focus on the classification problem, for which many recent algorithms have been proposed and applied successfully. For a specific regularized form of greedy stagewise optimization, we prove consistency of the approach under rather general conditions. Focusing on specific classes of problems we provide conditions under which our greedy procedure achieves the (nearly) minimax rate of convergence, implying that the procedure cannot be improved in a worst case setting. We also construct a fully adaptive procedure, which, without knowing the smoothness parameter of the decision boundary, converges at the same rate as if the smoothness parameter were known.

1. Introduction

The problem of binary classification plays an important role in the general theory of learning and estimation. While this problem is the simplest supervised learning problem one may envisage, there are still many open issues related to the best approach to solving it. In this paper we consider a family of algorithms based on a greedy stagewise minimization of an appropriate smooth loss function, and the construction of a composite classifier by combining simple base classifiers obtained by the stagewise procedure. Such procedures have been known for many years in the statistics literature as *additive models* (Hastie and Tibshirani, 1990, Hastie et al., 2001) and have also been used in the signal processing community under the title of *matching pursuit* (Mallat and Zhang, 1993). More recently, it has transpired that the boosting algorithm proposed in the machine learning community (Schapire, 1990, Freund and Schapire, 1997), which was based on a very different motivation, can

©2003 Shie Mannor, Ron Meir, and Tong Zhang.

also be thought of as a stagewise greedy algorithm (e.g Breiman, 1998, Friedman et al., 2000, Schapire and Singer, 1999, Mason et al., 2000, Meir and Rätsch, 2003). In spite of the connections of these algorithms to earlier work in the field of statistics, it is only recently that certain question have been addressed. For example, the notion of the margin and its impact on performance (Vapnik, 1998, Schapire et al., 1998), the derivation of sophisticated finite sample bounds (e.g., Bartlett et al., 2002, Bousquet and Chapelle, 2002, Koltchinksii and Panchenko, 2002, Zhang, 2002, Antos et al., 2002), the utilization of a range of different cost functions (Mason et al., 2000, Friedman et al., 2000, Lugosi and Vayatis, 2001, Zhang, 2002, Mannor et al., 2002a) are but a few of the recent contributions to this field.

Boosting algorithms have been demonstrated to be very effective in many applications, a success which led to some initial hopes that boosting does not overfit. However, it became clear very quickly that boosting may in fact overfit badly (e.g., Dietterich, 1999, Schapire and Singer, 1999) if applied without regularization. In order to address the issue of overfitting, several authors have recently addressed the question of statistical consistency. Roughly speaking, consistency of an algorithm with respect to a class of distributions implies that the loss incurred by the procedure ultimately converges to the lowest loss possible as the size of the sample increases without limit (a precise definition is provided in Section 2.1). Given that an algorithm is consistent, a question arises as to the rates of convergence to the minimal loss. In this context, a classic approach looks at the so-called minimax criterion, which essentially measures the performance of the *best* estimator for the *worst* possible distribution in a class. Ideally, we would like to show that an algorithm achieves the minimax (or close to minimax) rate. Finally, we address the issue of adaptivity. In computing minimax rates one usually assumes that there is a certain parameter θ characterizing the smoothness of the target distribution. This parameter is usually assumed to be known in order to compute the minimax rates. For example, the parameter θ may correspond to the Lipschitz constant of a decision boundary. In practice, however, one usually does not know the value of θ. In this context one would like to construct algorithms which are able to achieve the minimax rates *without* knowing the value of θ in advance. Such procedures have been termed *adaptive in the minimax sense* by Barron et al. (1999). Using a boosting model that is somewhat different from ours, it was shown by Bühlmann and Yu (2003) that boosting with early stopping achieves the exact minimax rates of Sobolev classes with a linear smoothing spline weak learner, and the procedure adapts to the unknown smoothness of the Sobolev class.

The stagewise greedy minimization algorithm that is considered in this work is natural and closer to algorithms that are used in practice. This is in contrast to the standard approach of selecting a hypothesis from a particular hypothesis class. Thus, our approach provides a bridge between theory and practice since we use theoretical tools to analyze a widely used practical approach.

The remainder of this paper is organized as follows. We begin in Section 2 with some formal definitions of consistency, minimaxity and adaptivity, and recall some recent tools from the theory of empirical processes. In Section 3 we introduce a greedy stagewise algorithm for classification, based on rather general loss functions, and prove the universal consistency of the algorithm. In Section 4 we then specialize to the case of the squared loss, for which recent results from the theory of empirical processes enable the establishment of fast rates of convergence. We also introduce an adaptive regularization algorithm which is shown to lead to nearly minimax rates even if we do not assume a-priori knowledge of θ. We then present some numerical results in Section 5, which demonstrate the importance of regularization. We conclude the paper in Section 6 and present some open questions.

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 $S = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ 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, and the expected loss is given by

$$L(f) = \mathbf{E}\ell(Y, f(X)). \tag{2}$$

Using the notation $\eta(x) \stackrel{\triangle}{=} P(Y = 1|X = x)$, it is well known that L^*, the minimum of $L(f)$, can be achieved by setting $f(x) = 2\eta(x) - 1$ (e.g., Devroye et al., 1996). Note that the decision choice at the point $f(x) = 0$ is not essential in the analysis. In this paper we simply assume that $\ell(y, 0) = 1/2$, so that the decision rule $2\eta(x) - 1$ is Bayes optimal at $\eta(x) = 1/2$.

2.1 Consistency, Minimaxity and Adaptivity

Based on a sample S, we wish to construct a rule f which assigns to each new input x a (soft) label $f(S, x)$, for which the expected loss $L(f, S) = \mathbf{E}\ell(Y, f(S, X))$ is minimal. Since S is a random variable so is $L(f, S)$, so that one can only expect to make probabilistic statements about this random variable. In this paper, we follow standard notation within the statistics literature, and denote sample-dependent quantities by a hat above the variable. Thus, we replace $f(S, x)$ by $\hat{f}(x)$. In general, one has at one's disposal only the sample S, and perhaps some very general knowledge about the problem at hand, often in the form of some regularity assumptions about the probability distribution P. Within the PAC setting (e.g., Kearns and Vazirani, 1994), one makes the very stringent assumption that $Y_i = g(X_i)$ and that g belongs to some *known* function class. Later work considered the so-called *agnostic* setting (e.g., Anthony and Bartlett, 1999), where nothing is assumed about g, and one compares the performance of $\hat{f}$ to that of the *best* hypothesis f^* within a given model class $\mathcal{F}$, namely $f^* = \operatorname{argmin}_{f \in \mathcal{F}} L(f)$ (in order to avoid unnecessary complications, we assume f^* exists). However, in general one is interested in comparing the behavior of the empirical estimator $\hat{f}$ to that of the optimal Bayes estimator, which minimizes the probability of error. The difficulty, of course, is that the determination of the Bayes classifier $g_B(x) = 2\eta(x) - 1$, requires knowledge of the underlying probability distribution. In many situations, one possesses some general knowledge about the underlying class of distributions P, usually in the form of some kind of smoothness assumption. For example, one may assume that $\eta(x) = P(Y = 1|x)$ is a Lipschitz function, namely $|\eta(x) - \eta(x')| \leq K\|x - x'\|$ for all x and x'. Let us denote the class of possible distributions by $\mathcal{P}$, and an empirical estimator based on a sample of size m by $\hat{f}_m$. Next, we introduce the notion of consistency. Roughly, a classification procedure leading to a classifier $\hat{f}_n$ is consistent with respect to class of distributions $\mathcal{P}$, if the loss $L(\hat{f}_n)$ converges, for increasing sample sizes, to the minimal loss possible for this class. More formally, the following definition is the standard definition of strong consistency (e.g., Devroye et al., 1996).

Definition 1 *A classification algorithm leading to a classifier $\hat{f}_m$ is* strongly consistent *with respect to a class of distributions $\mathcal{P}$ if for every $P \in \mathcal{P}$*

$$\lim_{m\to\infty} L(\hat{f}_m) = L^* \,, \quad P \text{ almost surely.}$$

If $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{P}$ contains all Borel probability measures, we say that the algorithm is universally consistent.

In this work we show that algorithms based on stagewise greedy minimization of 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 class conditional distribution $\eta(x)$.

Consistency is clearly an important property for any learning algorithm, as it guarantees that the algorithm ultimately performs well, in the sense of asymptotically achieving the minimal loss possible. One should keep in mind though, that consistent algorithms are not necessarily optimal when only a finite amount of data is available. A classic example of the lack of finite-sample optimality of consistent algorithms is the James-Stein effect (see, for example, Robert, 2001, Section 2.8.2).

In order to quantify the performance more precisely, we need to be able to say something about the speed at which convergence to L^* takes place. In order to do so, we need to determine a yardstick by which to measure distance. A classic measure which we use here is the so-called *minimax* rate of convergence, which essentially measures the performance of the best empirical estimator on the most difficult distribution in $\mathcal{P}$. Let the class of possible distributions be characterized by a parameter θ, namely $\mathcal{P} = \mathcal{P}_\theta$. For example, assuming that $\eta(x)$ is Lipschitz, θ could represent the Lipschitz constant. Formally, the minimax risk is given by

$$r_m(\theta) = \inf_{\hat{f}_m} \sup_{P\in\mathcal{P}_\theta} \mathbf{E}\ell(Y, \hat{f}_m(X)) - L^*,$$

where $\hat{f}_m$ is *any* estimator based on a sample S of size m, and the expectation is taken with respect to X,Y and the m-sample S. The rate at which the minimax risk converges to zero has been computed in the context of binary classification for several classes of distributions by Yang (1999).

So far we have characterized the smoothness of the distribution P by a parameter θ. However, in general one does not possess any prior information about θ, except perhaps that it is finite. The question then arises as to whether one can design an *adaptive* scheme which constructs an estimator $\hat{f}_m$ *without* any knowledge of θ, and for which convergence to L^* at the minimax rates (which assumes knowledge of θ) can be guaranteed. Following Barron et al. (1999) we refer to such a procedure as *adaptive in the minimax sense*.

2.2 Some Technical Tools

We begin with a few useful results. Let $\{\sigma_i\}_{i=1}^m$ be a sequence of binary random variables such that $\sigma_i = \pm 1$ with probability $1/2$. The *Rademacher complexity* of $\mathcal{F}$ (e.g., van der Vaart and Wellner, 1996) is given by

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

where the expectation is over $\{\sigma_i\}$ and $\{X_i\}$. See Bartlett and Mendelson (2002) for some properties of $R_m(\mathcal{F})$.

The following theorem can be obtained by a slight modification of the proof of Theorem 1 of Koltchinksii and Panchenko (2002).

Theorem 2 (Adapted from Theorem 1 in Koltchinksii and Panchenko, 2002)
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 $\sup_{x\in\mathcal{X}} |\phi(f(x))| \leq M$ for all $f \in \mathcal{F}$. 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, the Rademacher complexity can be estimated directly. Results summarized by Bartlett and Mendelson (2002) are useful for bounding this quantity for algebraic composition of function classes. We recall the relation between Rademacher complexity and covering numbers. For completeness we repeat the standard definition of covering numbers and entropy (e.g., van der Vaart and Wellner, 1996), which are related to the Rademacher complexity.

Definition 3 *Let $\mathcal{F}$ be a class of functions, and let ρ be a distance measure between functions in $\mathcal{F}$. The covering number $\mathcal{N}(\varepsilon, \mathcal{F}, \rho)$ is the minimal number of balls $\{g : \rho(g, f) \leq \varepsilon\}$ of radius ε needed to cover the set. The entropy of $\mathcal{F}$ is the logarithm of the covering number.*

Let $X = \{X_1, \ldots, X_m\}$ be a set of points and let Q_m be a probability measure over these points. We define the $\ell_p(Q_m)$ distance between any two functions f and g as

$$\ell_p(Q_m)(f,g) = \left(\sum_{i=1}^{m} Q_m |f(x_i) - g(x_i)|^p\right)^{1/p}.$$

In this case we denote the (empirical) covering number of $\mathcal{F}$ by $\mathcal{N}(\varepsilon, \mathcal{F}, \ell_p(Q_m))$. The *uniform* ℓ_p covering number and the *uniform* entropy are given respectively by

$$\mathcal{N}_p(\varepsilon, \mathcal{F}, m) = \sup_{Q_m} \mathcal{N}(\varepsilon, \mathcal{F}, \ell_p(Q_m)) \qquad ; \qquad H_p(\varepsilon, \mathcal{F}, m) = \log \mathcal{N}_p(\varepsilon, \mathcal{F}, m),$$

where the supremum is over all probability distributions Q_m over sets of m points sampled from $\mathcal{X}$. In the special case $p = 2$, we will abbreviate the notation, setting $H(\varepsilon, \mathcal{F}, m) \equiv H_2(\varepsilon, \mathcal{F}, m)$.

Let ℓ_2^m denote the empirical ℓ_2 norm with respect to the uniform measure on the points $\{X_1, X_2, \ldots, X_m\}$, 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 van der Vaart and Wellner, 1996)

$$R_m(\mathcal{F}) \leq \left(\mathbf{E}\int_0^\infty \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, \ell_2^m)}\, d\varepsilon\right)\frac{C}{\sqrt{m}}, \tag{3}$$

where 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 Pollard, 1984). Moreover, the ℓ_2 covering numbers are directly related to the minimax rates of convergence (Yang, 1999).

2.3 Related Results

We discuss some previous work related to the issues studied in this work. The question of the consistency of boosting algorithms has attracted some attention in recent years. Jiang, following Breiman (2000), raised the questions of whether AdaBoost is consistent and whether regularization is needed. It was shown in Jiang (2000b) 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 by Jiang (2000a), 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 (2001) 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 by Lugosi and Vayatis (2001, 2002). The major differences between our work and (Lugosi and Vayatis, 2001, 2002) are the following: (i) The precise nature of the algorithms used is different; in particular the approach to regularization is different. (ii) We establish convergence rates and provide conditions for establishing adaptive minimaxity. (iii) We consider stagewise procedures based on greedily adding on a single base hypothesis at a time. The work of Lugosi and Vayatis (2002) focused on the effect of using a convex upper bound on the $0-1$ loss.

A different kind of consistency result was established by Mannor and Meir (2001, 2002). 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.

Zhang (2002) studied an approximation-estimation decomposition of binary classification methods based on minimizing some convex cost functions. The focus there was on approximation error analysis as well as behaviors of different convex cost functions. The author also studied estimation errors for kernel methods including support vector machines, and established universal consistency results. However, the paper does not contain any specific result for boosting algorithms.

All of the work discussed above deals with the issue of consistency. This paper extends our earlier results (Mannor et al., 2002a) where we proved consistency for certain regularized greedy boosting algorithms. Here we go beyond consistency and consider rates of convergence and investigate the adaptivity of the approach.

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

Consider a class of so-called *base hypotheses* $\mathcal{H}$, 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 $t \to \infty$. The algorithms considered in this paper construct a composite hypothesis by choosing a function f from $\beta\mathrm{CO}(\mathcal{H})$, where for any class $\mathcal{G}$, $\beta\mathcal{G} = \{f : f = \beta g,\ g \in \mathcal{G}\}$. The parameter β will be specified at a later stage.

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 section is the demonstration of the consistency of such a procedure.

To describe the algorithm, let $\phi(x)$ be a convex function, which upper bounds the $0-1$ loss, namely

$$\phi(yf(x)) \geq I[yf(x) \leq 0], \qquad \phi(u) \text{ convex}.$$

Specific examples for ϕ are given in Section 3.3. Consider the empirical and true losses incurred by a function f based on the loss ϕ,

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

Here $\mathbf{E}_{X,Y}$ is the expectation operator with respect to the measure P and $\mathbf{E}_X$ is the expectation with respect to the marginal on X.

3.1 Approximation by Convex Hulls of Small Classes

In order to achieve consistency with respect to a large class of distributions, one must demand that the class $\beta\mathrm{CO}(\mathcal{H})$ is 'large' in some well-defined sense. For example, if the class $\mathcal{H}$ consists only of polynomials of a fixed order, then we cannot hope to approximate arbitrary continuous functions, since $\mathrm{CO}(\mathcal{H})$ also consists solely of polynomials of a fixed order. However, there are classes of non-polynomial functions for which $\beta\mathrm{CO}(\mathcal{H})$ is large.

As an example, consider a univariate (i.e., one-dimensional) function $\sigma : \mathbb{R} \to [0,1]$. The class of symmetric ridge functions over $\mathbb{R}^d$ is defined as:

$$\mathcal{H}_\sigma \stackrel{\triangle}{=} \{\pm\sigma(a^\top x + b), a \in \mathbb{R}^d, b \in \mathbb{R}\}\ .$$

Recall that for a class of functions $\mathcal{F}$, $\mathrm{SPAN}(\mathcal{F})$ consists of all linear combinations of functions from $\mathcal{F}$. It is known from Leshno et al. (1993) that the span of $\mathcal{H}_\sigma$ is dense in the set of continuous functions over a compact set. Since $\mathrm{SPAN}(\mathcal{H}_\sigma) = \cup_{\beta\geq 0}\beta\mathrm{CO}(\mathcal{H})$, it follows that every continuous function mapping from a compact set Ω to $\mathbb{R}$ can be approximated with arbitrary precision by some g in $\beta\mathrm{CO}(\mathcal{H})$ for a large enough β.

For the case where $h(x) = \mathrm{sgn}(w^\top x + b)$ Barron (1992) defines the class

$$\mathrm{SPAN}_C(\mathcal{H}) = \left\{ f :\ f(x) = \sum_i c_i \mathrm{sgn}(w_i^\top x + b_i),\ c_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^d,\ \sum_i |c_i| \leq C \right\}.$$

Input: A sample S_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 $\hat{f}^t_{\beta,m}$

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

and refers to it as the class of functions with bounded variation with respect to half-spaces. In one dimension, this is simply the class of functions with bounded variation. Note that there are several extensions to the notion of bounded variation to multiple dimensions. We return to this class of functions in Section 4.2. Other classes of base functions which generate rich nonparametric sets of functions are free-knot splines (see Agarwal and Studden, 1980, for asymototic properties) and radial basis functions (e.g., Schaback, 2000).

3.2 A Greedy Stagewise Algorithm and Finite Sample Bounds

Based on a finite sample S_m, we cannot hope to minimize $A(f)$, but rather minimize its empirical counterpart $\hat{A}(f)$. Instead of minimizing $\hat{A}(f)$ directly, we consider a stagewise greedy algorithm, which is described in Figure 1. The algorithm proposed is related to the AdaBoost algorithm in incrementally minimizing a given convex loss function. In opposition to AdaBoost, we restrict the size of the weights α and β, which serves to regularize the algorithm, a procedure that will play an important role in the sequel. We also observe that many of the additive models introduced in the statistical literature (e.g., Hastie et al., 2001), operate very similarly to Figure 1. It is clear from the description of the algorithm that $\hat{f}^t_{\beta,m}$, the hypothesis generated by the procedure, belongs to $\beta\mathrm{CO}_t(\mathcal{H})$ for every t. Note also that, by the definition of ϕ, for fixed α and β the function $\hat{A}((1-\alpha)\hat{f}^{\tau-1}_{\beta,m} + \alpha\beta h)$ is convex in h.

We observe that many recent approaches to boosting-type algorithms (e.g., Breiman, 1998, Hastie et al., 2001, Mason et al., 2000, Schapire and Singer, 1999) 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 remains 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.

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, $f_G(\eta)$, the minimizer of $G(\eta, f)$, is well-defined at $\eta = 0$ or 1 for appropriate loss functions. For every value of $\eta \in [0,1]$ let

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

It can be shown (Zhang, 2002) that for many choices of ϕ, including the examples given in Section 3.3, $f_G(\eta) > 0$ when $\eta > 1/2$. We begin with a result from Zhang (2002). 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. For simplicity we assume that f_{opt} exists. In other words

$$A(f_{\mathrm{opt}}) \leq A(f) \qquad \text{(for all measurable } f\text{)}.$$

Our definition implies that $f_{\mathrm{opt}}(x) = f_G(\eta(x))$.

Theorem 4 (Zhang, 2002, Theorem 2.1) *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 $f(x)$

$$L(f) - L^* \leq 2c \left(A(f) - A(f_{\mathrm{opt}})\right)^{1/s}, \tag{4}$$

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 4 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 in Theorem 4 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 3.3. The bound (4) allows one to work directly with the function $A(\cdot)$ rather than with the less wieldy $0-1$ loss $L(\cdot)$.

We are interested in bounding the loss $L(f)$ of the empirical estimator $\hat{f}^t_{\beta,m}$ obtained after t steps of the stagewise greedy algorithm described in Figure 1. Substitute $\hat{f}^t_{\beta,m}$ in (4), and consider bounding the r.h.s. as follows (ignoring the $1/s$ exponent for the moment):

$$\begin{aligned} A(\hat{f}^t_{\beta,m}) - A(f_{\mathrm{opt}}) = &\left[A(\hat{f}^t_{\beta,m}) - \hat{A}(\hat{f}^t_{\beta,m})\right] + \left[\hat{A}(\hat{f}^t_{\beta,m}) - \hat{A}(f^*_\beta)\right] \\ &+ \left[\hat{A}(f^*_\beta) - A(f^*_\beta)\right] + \left[A(f^*_\beta) - A(f_{\mathrm{opt}})\right]. \end{aligned} \tag{5}$$

Next, we bound each of the terms separately.

The first term can be bounded using Theorem 2. In particular, since $A(f) = \mathbf{E}\phi(Yf(X))$, where ϕ is assumed to be convex, and since $\hat{f}^t_{\beta,m} \in \beta\mathrm{CO}(H)$ 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 3.3). From Theorem 2 we have that with probability at least $1 - \delta$,

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

where $\phi_\beta \stackrel{\triangle}{=} \sup_{f\in[-\beta,\beta]}\phi(f)$. Recall that $\hat{f}^t_{\beta,m} \in \beta\mathrm{CO}(\mathcal{H})$, and note that we have used the fact that $R_m(\beta\mathrm{CO}(\mathcal{H})) = \beta R_m(\mathcal{H})$ (e.g., Bartlett and Mendelson, 2002). The third term on the r.h.s. of (5) can be estimated directly from the Chernoff bound. We have with probability at least $1-\delta$:

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

Note that f^* is fixed (independent of the sample), and therefore a simple Chernoff bound suffices here. In order to bound the second term in (5) we assume that

$$\sup_{v\in[-\beta,\beta]} \phi''(v) \le M_\beta < \infty, \tag{6}$$

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

From Theorem 4.2 by Zhang (2003) we know that for a fixed sample

$$\hat{A}(\hat{f}^t_{\beta,m}) - \hat{A}(f^*_\beta) \le \frac{8\beta^2 M_\beta}{t}. \tag{7}$$

This result holds for every convex ϕ and fixed β.

The fourth term in (5) 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 $1-2\delta$,

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

The final term in (8) can be bounded using the Lipschitz property of ϕ. In particular,

$$\begin{aligned}
A(f^*_\beta) - A(f_{\mathrm{opt}}) &= \mathbf{E}_X\left\{\eta(X)\phi(f^*_\beta(X)) + (1-\eta(X))\phi(-f^*_\beta(X))\right\} \\
&\quad - \mathbf{E}_X\left\{\eta(X)\phi(f_{\mathrm{opt}}(X)) + (1-\eta(X))\phi(-f_{\mathrm{opt}}(X))\right\} \\
&= \mathbf{E}_X\left\{\eta(X)[\phi(f^*_\beta(X)) - \phi(f_{\mathrm{opt}}(X))]\right\} \\
&\quad + \mathbf{E}_X\left\{(1-\eta(X))[\phi(-f^*_\beta(X)) - \phi(-f_{\mathrm{opt}}(X))]\right\} \\
&\le \kappa_\beta \mathbf{E}_X\left\{\eta(X)|f^*_\beta(X) - f_{\mathrm{opt}}(X)| + (1-\eta(X))|f^*_\beta(X) - f_{\mathrm{opt}}(X)|\right\} \\
&\le \kappa_\beta \mathbf{E}_X|f^*_\beta(X) - f_{\beta,\mathrm{opt}}(X)| + \Delta_\beta,
\end{aligned} \tag{9}$$

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

$$\Delta_\beta \stackrel{\triangle}{=} \sup_{\eta\in[1/2,1]} \{I(f_G(\eta) > \beta)[G(\eta,\beta) - G(\eta, f_G(\eta))]\}.$$

Note that $\Delta_\beta \to 0$ when $\beta\to\infty$ since Δ_β represents the tail behavior $G(\eta,\beta)$. Several examples are provided in Section 3.3.

3.3 Examples for ϕ

We consider three commonly used choices for the convex function ϕ. Other examples are presented by Zhang (2002).

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

It is easy to see that all losses are non-negative and upper bound the $0-1$ loss $I(x \leq 0)$, where $I(\cdot)$ is the indicator function. The exponential loss function was previously shown to lead to the AdaBoost algorithm (Schapire and Singer, 1999), while the other losses were proposed by Friedman et al. (2000), and shown to lead to other interesting stagewise algorithms. The essential differences between the loss functions relate to their behavior for $x \to \pm\infty$.

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 4, 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 Zhang, 2002).

$\phi(x)$	$\exp(-x)$	$\log(1+\exp(-x))/\log(2)$	$(x-1)^2$
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.

3.4 Universal Consistency

We assume that $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$, which in turn implies that $0 \in \text{CO}(\mathcal{H})$. This implies that $\beta_1\text{CO}(\mathcal{H}) \subseteq \beta_2\text{CO}(\mathcal{H})$ when $\beta_1 \leq \beta_2$. Therefore, using a larger value of β implies searching within a larger space. We define $\text{SPAN}(\mathcal{H}) = \cup_{\beta>0}\beta\text{CO}(\mathcal{H})$, 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 $\text{SPAN}(\mathcal{H})$ 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 by Zhang (2002), we know that for all ϕ considered in this paper, and all Borel measures, $\inf_{f\in\text{SPAN}(\mathcal{H})} A(f) = A(f_{\text{opt}})$. Since $\text{SPAN}(\mathcal{H}) = \cup_{\beta>0}\beta\text{CO}(\mathcal{H})$, we obtain $\lim_{\beta\to\infty} A(f_\beta^*) - A(f_{\text{opt}}) = 0$, leading to the vanishing of the final term in (8) when $\beta \to \infty$. Using this observation we are able to establish sufficient conditions for consistency.

Theorem 5 *Assume that the class of functions* $\text{SPAN}(\mathcal{H})$ *is dense in* $C(K)$ *over a domain* $K \subseteq \mathbb{R}^d$. *Assume further that* ϕ *is convex and Lipschitz and that (6) holds. Choose* $\beta = \beta(m)$ *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 The basic idea of the proof is the selection of $\beta = \beta(m)$ in such a way that it balances the estimation and approximation error terms. In particular, β should increase to infinity so that the approximation error vanishes. However, the rate at which β increases should be sufficiently slow to guarantee convergence of the estimation error to zero as $m \to \infty$. Let $\delta_m = \frac{1}{m^2}$. It follows from (8) 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 in (8) converge to 0, it follows that for every $\varepsilon > 0$ from some time on $A(\hat{f}^t_{\beta,m}) - A(f_{\text{opt}}) < \varepsilon$ with probability 1. Using (4) concludes the proof. ■

As a simple example for the choice of $\beta = \beta(m)$, consider the logistic loss. From Table 1 we conclude that selecting $\beta = o\left(\sqrt{m/\log m}\right)$ suffices to guarantee consistency.

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 class $\mathcal{H}$ and the posterior probability $\eta(x)$. We noted elsewhere (Mannor et al., 2002a) we used (8) in order to establish convergence rates for the three loss functions described above, when certain smoothness conditions were assumed concerning the class conditional distribution $\eta(x)$. The procedure described in Mannor et al. (2002a) also established appropriate (non-adaptive) choices for β as a function of the sample size m. In the next section we use a different approach for the squared loss in order to derive faster, nearly optimal, convergence rates.

4. Rates of Convergence and Adaptivity – the Case of Squared Loss

We have shown that under reasonable conditions on the function ϕ, universal consistency can be established as long as the base class $\mathcal{H}$ is sufficiently rich. We now move on to discuss rates of convergence and the issue of adaptivity, as described in Section 2. In this section we focus on the squared loss, as particularly tight bounds are available for this case, using techniques from the empirical process literature (e.g., van de Geer, 2000). This allows us to demonstrate nearly minimax rates of convergence. Since we are concerned with establishing convergence rates in a nonparametric setting, we will not be concerned with constants which do not affect rates of convergence. We will denote generic constants by c, c', c_1, c_1', etc.

We begin by bounding the difference between $A(f)$ and $A(f_{\text{opt}})$ in the non adaptive setting, where we consider the case of a fixed value of β which defines the class $\beta\text{CO}(\mathcal{H})$. In Section 4.2 we use the multiple testing Lemma to derive an adaptive procedure that leads to a uniform bound over $\text{SPAN}(\mathcal{H})$. We finally apply those results for attaining bounds on the classification $(0-1)$ loss in Section 4.3. Observe that from the results of Section 3, for each fixed value of β, we may take the number of boosting iterations t to infinity. We assume throughout this section that this procedure has been adhered to.

4.1 Empirical Ratio Bounds for the Squared Loss

In this section we restrict attention to the squared loss function,

$$A(f) = \mathbf{E}\,(f(X) - Y)^2.$$

Since in this case

$$f_{\text{opt}}(x) = \mathbf{E}(Y|x),$$

we have the following identity for any function f:

$$\mathbf{E}_{Y|x}(f(x) - Y)^2 - \mathbf{E}_{Y|x}(f_{\text{opt}}(x) - Y)^2 = (f(x) - f_{\text{opt}}(x))^2.$$

Therefore

$$A(f) - A(f_{\text{opt}}) = \mathbf{E}\,(f(X) - f_{\text{opt}}(X))^2. \tag{10}$$

We assume that f belongs to some function class $\mathcal{F}$, but we do *not* assume that f_{opt} belongs to $\mathcal{F}$. Furthermore, since for any real numbers a, b, c, we have that $(a-b)^2 - (c-b)^2 = (a-c)^2 + 2(a-c)(c-b)$ the following is true:

$$\begin{aligned}\hat{A}(f) - \hat{A}(f_{\text{opt}}) &= \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 - \frac{1}{m}\sum_{i=1}^{m}(f_{\text{opt}}(x_i) - y_i)^2 \\ &= \frac{2}{m}\sum_{i=1}^{m}(f_{\text{opt}}(x_i) - y_i)(f(x_i) - f_{\text{opt}}(x_i)) + \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - f_{\text{opt}}(x_i))^2. \end{aligned} \tag{11}$$

Our goal at this point is to assess the expected deviation of $[A(f) - A(f_{\text{opt}})]$ from its empirical counterpart, $[\hat{A}(f) - \hat{A}(f_{\text{opt}})]$. In particular, we want to show that with probability at least $1-\delta$, $\delta \in (0,1)$, for all $f \in \mathcal{F}$ we have

$$A(f) - A(f_{\text{opt}}) \le c(\hat{A}(f) - \hat{A}(f_{\text{opt}})) + \rho_m(\delta)\,,$$

for appropriately chosen c and $\rho_m(\delta)$.

For any f it will be convenient to use the notation $\hat{\mathbf{E}} f \stackrel{\triangle}{=} \frac{1}{m}\sum_{i=1}^{m} f(x_i)$.

We now relate the expected and empirical values of the deviation terms $A(f) - A(f_{\text{opt}})$. The following result is based on the symmetrization technique and the so-called peeling method in statistics (e.g., Section 5.3 in van de Geer, 2000). The peeling method is a general method for bounding suprema of stochastic processes over some class of functions. The basic idea is to transform the task into a sequence of simpler bounds, each defined over an element in a nested sequence of subsets of the class (see (5.17) in van de Geer, 2000). Since the proof is rather technical, it is presented in the appendix.

Lemma 6 *Let $\mathcal{F}$ be a class of uniformly bounded functions, and let $X = \{X_1, \ldots, X_m\}$ be a set of points drawn independently at random according to some law P. Assume that for all $f \in \mathcal{F}$, $\sup_x |f(x) - f_{\text{opt}}(x)| \le M$. Then there exists a positive constant c such that for all $q \ge c$, with probability at least $1 - \exp(-q)$, for all $f \in \mathcal{F}$*

$$\mathbf{E}\left(f(X) - f_{\text{opt}}(X)\right)^2 \le 4\hat{\mathbf{E}}\left(f(X) - f_{\text{opt}}(X)\right)^2 + \left\{\frac{100qM^2}{m} + \frac{\Delta_m^2}{6}\right\},$$

where Δ_m is any number such that

$$m\Delta_m^2 \geq 32M^2 \max(H(\Delta_m, \mathcal{F}, m), 1). \tag{12}$$

Observe that Δ_m is well-defined since the l.h.s. is monotonically increasing and unbounded, while the r.h.s. is monotonically decreasing.

We use the following bound from van de Geer (2000):

Lemma 7 (van de Geer, 2000, Lemma 8.4) *Let $\mathcal{F}$ be a class of functions such that for all positive δ, $H(\delta, \mathcal{F}, m) \leq K\delta^{-2\xi}$, for some constants $0 < \xi < 1$ and K. Let X, Y be random variables defined over some domain. Let $W(x,y)$ be a real-valued function such that $|W(x,y)| \leq M$ for all x,y, and $\mathbf{E}_{Y|x} W(x,Y) = 0$ for all x. Then there exists a constant c, depending on ξ, K and M only, such that for all $\varepsilon \geq c/\sqrt{m}$:*

$$\mathbf{P}\left\{ \sup_{g \in \mathcal{F}} \frac{|\hat{\mathbf{E}}\{W(X,Y)g(X)\}|}{(\hat{\mathbf{E}}g(X)^2)^{(1-\xi)/2}} \geq \varepsilon \right\} \leq c \exp(-m\varepsilon^2/c^2).$$

In order to apply this bound, it is useful to introduce the following assumption.

Assumption 1 *Assume that $\exists M \geq 1$ such that $\sup_x |f(x) - f_{\text{opt}}(x)| \leq M$ for all $f \in \mathcal{F}$. Moreover, for all positive ε, $H(\varepsilon, \mathcal{F}, m) \leq K(\varepsilon/M)^{-2\xi}$ where $0 < \xi < 1$.*

We will now rewrite Lemma 7 in a somewhat different form using the notation of this section.

Lemma 8 *Let Assumption 1 hold. Then there exist positive constants c_0 and c_1 that depend on ξ and K only, such that $\forall q \geq c_0$, with probability at least $1 - \exp(-q)$, for all $f \in \mathcal{F}$*

$$\left|\hat{\mathbf{E}}\left\{(f_{\text{opt}}(X) - Y)(f(X) - f_{\text{opt}}(X))\right\}\right| \leq \frac{1-\xi}{2}\hat{\mathbf{E}}(f - f_{\text{opt}})^2 + c_1 M^2 \left(\frac{q}{m}\right)^{1/(1+\xi)}.$$

Proof Let

$$W(X,Y) = (f_{\text{opt}}(X) - Y)/M \qquad ; \qquad g(X) = (f(X) - f_{\text{opt}}(X))/M.$$

Using Lemma 7 we find that there exist constants c and c' that depend on ξ and K only, such that $\forall \varepsilon \geq c/\sqrt{m}$

$$\begin{aligned}
&\mathbf{P}\left\{ \exists g \in \mathcal{G} : |\hat{\mathbf{E}}\{W(X,Y)g(X)\}| > \frac{1-\xi}{2}\hat{\mathbf{E}}g(X)^2 + \frac{1+\xi}{2}\varepsilon^{2/(1+\xi)} \right\} \\
&\overset{(a)}{\leq} \mathbf{P}\left\{ \exists g \in \mathcal{G} : |\hat{\mathbf{E}}\{W(X,Y)g(X)\}| > (\hat{\mathbf{E}}g(X)^2)^{(1-\xi)/2}\varepsilon \right\} \\
&= \mathbf{P}\left\{ \sup_{g \in \mathcal{G}} \frac{|\hat{\mathbf{E}}\{W(X,Y)g(X)\}|}{(\hat{\mathbf{E}}g(X)^2)^{(1-\xi)/2}} > \varepsilon \right\} \\
&\overset{(b)}{\leq} c\exp(-m\varepsilon^2/c^2).
\end{aligned}$$

where (a) used the inequality $|ab| \leq \frac{1-\xi}{2}|a|^{2/(1-\xi)} + \frac{1+\xi}{2}|b|^{2/(1+\xi)}$, and (b) follows using Lemma 7. The claim follows by setting $\varepsilon = \sqrt{q/m}$ and choosing c_0 and c_1 appropriately. □

Combining Lemma 6 and Lemma 8, we obtain the main result of this section.

Theorem 9 *Suppose Assumption 1 holds. Then there exist constants $c_0, c_1 > 0$ that depend on ξ and K only, such that $\forall q \geq c_0$, with probability at least $1-\exp(-q)$, for all $f \in \mathcal{F}$*

$$A(f) - A(f_{\rm opt}) \leq \frac{4}{\xi}\left[\hat{A}(f) - \hat{A}(f_{\rm opt})\right] + \frac{c_1 M^2}{\xi}\left(\frac{q}{m}\right)^{1/(1+\xi)}.$$

Proof By (10) it follows that $A(f) - A(f_{\rm opt}) = \mathbf{E}(f - f_{\rm opt})^2$. There exists a constant c_0' depending on K only such that in Lemma 6, we can let $\Delta_m^2 = c_0' M^2 m^{-1/(1+\xi)}$ to obtain

$$A(f) - A(f_{\rm opt}) = \mathbf{E}(f - f_{\rm opt})^2 \leq 4\hat{\mathbf{E}}(f - f_{\rm opt})^2 + M^2(100q/m + c_0' m^{-1/(1+\xi)}) \tag{13}$$

with probability at least $1-\exp(-q)$ where $q \geq 1$. By (11) we have that

$$[\hat{A}(f) - \hat{A}(f_{\rm opt})] = 2\hat{\mathbf{E}}\left\{(f_{\rm opt}(X) - Y)(f(X) - f_{\rm opt}(X))\right\} + \hat{\mathbf{E}}(f(X) - f_{\rm opt}(X))^2.$$

Using Lemma 8 we have that there exist constants $c_1' \geq 1$ and c_2' that depend on K and ξ only, such that for all $q \geq c_1'$, with probability at least $1-\exp(-q)$:

$$\left|\hat{\mathbf{E}}\left\{(f_{\rm opt}(X) - Y)(f(X) - f_{\rm opt}(X))\right\}\right| \leq \frac{1-\xi}{2}\hat{\mathbf{E}}(f - f_{\rm opt})^2 + c_2' M^2\left(\frac{q}{m}\right)^{1/(1+\xi)}.$$

Combining these results we have that with probability at least $1 - e^{-q}$:

$$\begin{aligned}[\hat{A}(f) - \hat{A}(f_{\rm opt})] &= 2\hat{\mathbf{E}}\left\{(f_{\rm opt}(X) - Y)(f(X) - f_{\rm opt}(X))\right\} + \hat{\mathbf{E}}(f(X) - f_{\rm opt}(X))^2 \\ &\geq \xi\hat{\mathbf{E}}(f - f_{\rm opt})^2 - 2c_2' M^2\left(\frac{q}{m}\right)^{1/(1+\xi)}. \end{aligned} \tag{14}$$

From (13) and (14) we obtain with probability at least $1-2\exp(-q)$:

$$A(f) - A(f_{\rm opt}) \leq \frac{4}{\xi}[\hat{A}(f) - \hat{A}(f_{\rm opt})] + \frac{8}{\xi}c_2' M^2\left(\frac{q}{m}\right)^{1/(1+\xi)} + M^2(100q/m + c_0' m^{-1/(1+\xi)}).$$

Note that Assumption 1 was used when invoking Lemma 6. The theorem follows from this inequality with appropriately chosen c_0 and c_1. □

4.2 Adaptivity

In this section we let f be chosen from $\beta{\rm CO}(\mathcal{H}) \equiv \beta\mathcal{F}$, where β will be determined adaptively based on the data in order to achieve an optimal balance between approximation and estimation errors. In this case, $\sup_x |f(x)| \leq \beta M$ where $h \in \mathcal{H}$ are assumed to obey $\sup_x |h(x)| \leq M$. We first need to determine the precise β-dependence of the bound of Theorem 9. We begin with a definition followed by a simple Lemma, the so-called multiple testing Lemma (e.g., Lemma 4.14 in Herbrich, 2002).

Definition 10 *A test Γ is a mapping from the sample S and a confidence level δ to the logical values $\{\text{True}, \text{False}\}$. We denote the logical value of activating a test Γ on a sample S with confidence δ by $\Gamma(S,\delta)$.*

Lemma 11 *Suppose we are given a set of tests $\Gamma = \{\Gamma_1, \ldots, \Gamma_r\}$. Assume further that a discrete probability measure $P = \{p_i\}_{i=1}^r$ over Γ is given. If for every $i \in \{1,2,\ldots,r\}$ and $\delta \in (0,1)$, $\mathbf{P}\{\Gamma_i(S,\delta)\} \geq 1-\delta$, then*

$$\mathbf{P}\left\{\Gamma_1(S,\delta p_1) \wedge \cdots \wedge \Gamma_r(S,\delta p_r)\right\} \geq 1-\delta.$$

We use Lemma 11 in order to extend Theorem 9 so that it holds for all β. The proof again relies on the *peeling* technique.

Theorem 12 *Let Assumption 1 hold. Then there exist constants $c_0, c_1 > 0$ that depend on ξ and K only, such that $\forall q \geq c_0$, with probability at least $1-\exp(-q)$, for all $\beta \geq 1$ and for all $f \in \beta\mathcal{F}$ we have*

$$A(f) - A(f_{\text{opt}}) \leq \frac{4}{\xi}\left[\hat{A}(f) - \hat{A}(f_{\text{opt}})\right] + c_1 \beta^2 M^2 \left(\frac{q + \log\log(3\beta)}{m}\right)^{1/(1+\xi)}.$$

Proof For all $s = 1, 2, 3, \ldots$, let $\mathcal{F}_s = 2^s \mathcal{F}$. Let us define the test $\Gamma_s(S, \delta)$ to be TRUE if

$$A(f) - A(f_{\text{opt}}) \leq \frac{4}{\xi}\left[\hat{A}(f) - \hat{A}(f_{\text{opt}})\right] + \frac{c}{\xi} 2^{2s} M^2 \left(\frac{\log(1/\delta)}{m}\right)^{\frac{1}{1+\xi}}$$

for all $f \in \mathcal{F}_s$ and FALSE otherwise. Using Theorem 9 we have that $\mathbf{P}(\Gamma_s(S,\delta)) \geq 1-\delta$. Let $p_s = \frac{1}{s(s+1)}$, noting that $\sum_{s=1}^{\infty} p_s = 1$ and by Lemma 11 we have that

$$\mathbf{P}\left\{\Gamma_s(S, \delta p_s) \text{ for all } s\right\} \geq 1 - \delta.$$

Consider $f \in \beta\mathcal{F}$ for some $\beta \geq 1$. Let $s = \lfloor \log_2 \beta \rfloor + 1$, we have that $\mathbf{P}\left\{\forall s :\ \Gamma_s(S, \frac{\delta}{s(s+1)})\right\} \geq 1-\delta$ so that with probability at least $1-\delta$ we have that:

$$\begin{aligned} A(f) - A(f_{\text{opt}}) &\leq \frac{4}{\xi}\left[\hat{A}(f) - \hat{A}(f_{\text{opt}})\right] + \frac{c'}{\xi} 2^{2s} M^2 \left(\frac{\log(\frac{s^2+s}{\delta})}{m}\right)^{1/(1+\xi)} \\ &\leq \frac{4}{\xi}\left[\hat{A}(f) - \hat{A}(f_{\text{opt}})\right] + \frac{c_1}{\xi} \beta^2 M^2 \left(\frac{\log\log(3\beta) + q}{m}\right)^{1/(1+\xi)}, \end{aligned}$$

where we set $q = \log(1/\delta)$ and used the fact that $2^{s-1} \leq \beta \leq 2^s$. □

Theorem 12 bounds $A(f) - A(f_{\text{opt}})$ in terms of $\hat{A}(f) - \hat{A}(f_{\text{opt}})$. However, in order to determine overall convergence rates of $A(f)$ to $A(f_{\text{opt}})$ we need to eliminate the empirical term $\hat{A}(f) - \hat{A}(f_{\text{opt}})$. To do so, we first recall a simple version of the Bernstein inequality (e.g., Devroye et al., 1996) together with a straightforward consequence.

Lemma 13 *Let $\{X_1, X_2, \ldots, X_m\}$ be real-valued i.i.d. random variables such that $|X_i| \leq b$ with probability one. Let $\sigma^2 = \text{Var}[X_1]$. Then, for any $\varepsilon > 0$*

$$\mathbf{P}\left\{\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbf{E}[X_1] > \varepsilon\right\} \leq \exp\left(-\frac{m\varepsilon^2}{2\sigma^2 + 2b\varepsilon/3}\right).$$

Moreover, if $\sigma^2 \leq c' b \mathbf{E}[X_1]$, then for all positive q, there exists a constant c that depends only on c' such that with probability at least $1-\exp(-q)$

$$\frac{1}{m}\sum_{i=1}^{m} X_i \leq c\mathbf{E}[X_1] + \frac{bq}{m},$$

where c is independent of b.

Proof The first part of the Lemma is just the Bernstein inequality (e.g., Devroye et al., 1996). To show the second part we need to bound from above the probability that $(1/m)\sum_{i=1}^{m} X_i > c\mathbf{E}[X_1] + bq/m$. Set $\varepsilon = (c-1)\mathbf{E}[X_1] + bq/m$. Using Bernstein's inequality we have that

$$\begin{aligned}
\mathbf{P}\left\{\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbf{E}[X_1] > (c-1)\mathbf{E}[X_1] + \frac{bq}{m}\right\} &\leq \exp\left(-\frac{m\varepsilon^2}{2\sigma^2 + 2b\varepsilon/3}\right) \\
&\leq \exp\left(-\frac{m\varepsilon^2}{2c'b\mathbf{E}[X_1] + 2b\varepsilon/3}\right) \\
&\overset{(a)}{\leq} \exp\left(-\frac{m\varepsilon}{b}\right) \\
&\leq \exp(-q)\,,
\end{aligned}$$

where (a) follows by choosing c large enough so that $2c' < \frac{1}{3}(c-1)$, implying that $2c'b\mathbf{E}[X_1] < b\varepsilon/3$. □

Next, we use Bernstein's inequality in order to bound $\hat{A}(f) - \hat{A}(f_{\text{opt}})$.

Lemma 14 *Let Assumption 1 hold. Given any $\beta \geq 1$ and $f \in \beta\mathcal{F}$, there exists a constant $c_0 > 0$ (independent of β) such that $\forall q$, with probability at least $1 - \exp(-q)$:*

$$\hat{A}(f) - \hat{A}(f_{\text{opt}}) \leq c_0\left[(A(f) - A(f_{\text{opt}})) + \frac{(\beta M)^2 q}{m}\right].$$

Proof Fix $f \in \beta\mathcal{F}$. We will use Lemma 13 to bound the probability of a large difference between $\hat{A}(f)$ and $\hat{A}(f_{\text{opt}})$. Instead of working with $\hat{A}(f)$ we will use $Z \overset{\triangle}{=} 2[(f_{\text{opt}}(X) - Y)(f(X) - f_{\text{opt}}(X))] + (f(X) - f_{\text{opt}}(X))^2$. According to (11), $\hat{A}(f) - \hat{A}(f_{\text{opt}}) = \hat{\mathbf{E}}[Z]$. The expectation of Z satisfies that $\mathbf{E}[Z] = \mathbf{E}[\hat{A}(f) - \hat{A}(f_{\text{opt}})]$, so using (10) we have that $\mathbf{E}[Z] = A(f) - A(f_{\text{opt}}) = \mathbf{E}(f(X) - f_{\text{opt}}(X))^2$. Bounding the variance we obtain that

$$\begin{aligned}
\text{Var}[Z] \leq \mathbf{E}[Z^2] \quad &\leq \quad \mathbf{E}\big[4(f_{\text{opt}}(X) - Y)^2(f(X) - f_{\text{opt}}(X))^2 + \\
&\qquad 4(f_{\text{opt}}(X) - Y)(f(X) - f_{\text{opt}}(X))^3 + (f(X) - f_{\text{opt}}(X))^4\big] \\
&\leq \quad \sup_{x,y}\big|4(f_{\text{opt}}(x) - y)^2 + 4(f_{\text{opt}}(x) - y)(f(x) - f_{\text{opt}}(x)) + \\
&\qquad (f(x) - f_{\text{opt}}(x))^2\big|\,\mathbf{E}[Z]. \qquad (15)
\end{aligned}$$

By Assumption 1 for every $f \in \mathcal{F}$ we have that $\sup_x |f(x) - f_{\text{opt}}(x)| \leq M$, which implies that for $f \in \beta\mathcal{F}$ we have that

$$\sup_x |f(x) - f_{\text{opt}}(x)| = \sup_x |f(x) - \beta f_{\text{opt}}(x) + (\beta - 1) f_{\text{opt}}(x)| \leq \beta M + (\beta - 1).$$

We conclude that $\sup_x |f(x) - f_{\text{opt}}(x)| \leq 2\beta M$. Recall that $f_{\text{opt}}(x) = \mathbf{E}(Y|x)$, $Y \in \{-1, 1\}$, so we can bound $|(f_{\text{opt}}(X) - Y)| \leq 2$. Since $\beta \geq 1$ and by the assumption on M we have that $|f_{\text{opt}}(X) - Y| \leq 2\beta M$. Plugging these upper bounds into (15) we obtain $\text{Var}[Z] \leq c'\beta^2 M^2 \mathbf{E}[Z]$, with $c' = 36$. A similar argument shows that $|Z|$ is not larger than $c''\beta M$ (with probability 1, and $c'' = 12$). The claim then follows from a direct application of Lemma 13. □

We now consider a procedure for determining β adaptively from the data. Define a penalty term

$$\gamma_q(\beta) = \beta^2 M^2 \left(\frac{\log\log(3\beta) + q}{m}\right)^{1/(1+\xi)},$$

which penalizes large values of β, corresponding to large classes with good approximation properties.

The procedure then is to find $\hat{\beta}_q$ and $\hat{f}_q \in \hat{\beta}_q \mathcal{F}$ such that

$$\hat{A}(\hat{f}_q) + \gamma_q(\hat{\beta}_q) \leq \inf_{\beta \geq 1} \left[\inf_{f \in \beta \mathcal{F}} \hat{A}(f) + 2\gamma_q(\beta) \right]. \tag{16}$$

This procedure is similar to the so-called structural risk minimization method (Vapnik, 1982), except that the minimization is performed over the continuous parameter β rather than a discrete hypothesis class counter. Observe that $\hat{\beta}_q$ and $\hat{f}_q$ are non unique, but this poses no problem.

We can now establish a bound on the loss incurred by this procedure.

Theorem 15 *Let Assumption 1 hold. Choose $q_0 > 0$ and assume we compute $\hat{f}_{q_0}$ using (16). Then there exist constants $c_0, q_0 > 0$ that depend on ξ and K only, such that $\forall m \geq q \geq \max(q_0, c_0)$, with probability at least $1 - \exp(-q)$,*

$$A(\hat{f}_{q_0}) \leq A(f_{\text{opt}}) + c_1 \left(\frac{q}{q_0} \right)^{1/(1+\xi)} \inf_{\beta \geq 1} \left[\inf_{f \in \beta \mathcal{F}} A(f) - A(f_{\text{opt}}) + \gamma_q(\beta) \right].$$

Note that since for any q, $\gamma_q(\beta) = O((1/m)^{1/(1+\xi)})$, Theorem 15 provides rates of convergence in terms of the sample size m. Observe also that the main distinction between Theorem 15 and Theorem 12 is that the latter provides a data-dependent bound, while the former establishes a so-called *oracle* inequality, which compares the performance of the empirical estimator $\hat{f}_{q_0}$ to that of the optimal estimator within a (continuously parameterized) hierarchy of classes. This optimal estimator cannot be computed since the underlying probability distribution is unknown, but serves as a performance yard-stick.

Proof (of Theorem 15) Consider $\beta_q \geq 1$ and $f_q \in \beta_q \mathcal{F}$ such that

$$A(f_q) - A(f_{\text{opt}}) + 2\gamma_q(\beta_q) \leq \inf_{\beta \geq 1} \left[\inf_{f \in \beta \mathcal{F}} A(f) - A(f_{\text{opt}}) + 4\gamma_q(\beta) \right]. \tag{17}$$

Note that β_q and f_q determined by (17), as opposed to $\hat{\beta}_q$ and $\hat{f}_q$ in (16), are *independent* of the data. Using Lemma 14, we know that there exists a constant c_2' such that with probability at least $1 - \exp(-q)$:

$$\hat{A}(f_q) - \hat{A}(f_{\text{opt}}) + 2\gamma_q(\beta_q) \leq c_2' \inf_{\beta \geq 1} \left[\inf_{f \in \beta \mathcal{F}} A(f) - A(f_{\text{opt}}) + \gamma_q(\beta) \right]. \tag{18}$$

From (16) we have

$$\begin{aligned} \hat{A}(\hat{f}_{q_0}) - \hat{A}(f_{\text{opt}}) + \gamma_{q_0}(\hat{\beta}_{q_0}) &\overset{(a)}{\leq} \hat{A}(f_q) - \hat{A}(f_{\text{opt}}) + 2\gamma_{q_0}(\beta_q) \\ &\overset{(b)}{\leq} \hat{A}(f_q) - \hat{A}(f_{\text{opt}}) + 2\gamma_q(\beta_q) \\ &\overset{(c)}{\leq} c_2' \inf_{\beta \geq 1} \left[\inf_{f \in \beta \mathcal{F}} A(f) - A(f_{\text{opt}}) + \gamma_q(\beta) \right]. \end{aligned} \tag{19}$$

Here (a) results from the definition of $\hat{f}_{q_0}$, (b) uses $q \geq q_0$, and (c) is based on (18). We then conclude that there exist constants $c_0', c_1' > 0$ that depend on ξ and K only, such that $\forall q \geq c_0'$, with

probability at least $1-\exp(-q)$:

$$\begin{aligned} A(\hat{f}_{q_0}) - A(f_{\text{opt}}) &\overset{(a)}{\leq} c_1'[\hat{A}(\hat{f}_{q_0}) - \hat{A}(f_{\text{opt}}) + \gamma_q(\hat{\beta}_{q_0})] \\ &\overset{(b)}{\leq} c_1' \left(\frac{q}{q_0}\right)^{1/(1+\xi)} [\hat{A}(\hat{f}_{q_0}) - \hat{A}(f_{\text{opt}}) + \gamma_{q_0}(\hat{\beta}_{q_0})] \\ &\overset{(c)}{\leq} c_1' c_2' \left(\frac{q}{q_0}\right)^{1/(1+\xi)} \inf_{\beta \geq 1} [\inf_{f \in \beta\mathcal{F}} A(f) - A(f_{\text{opt}}) + \gamma_q(\beta)]. \end{aligned}$$

Here (a) is based on Theorem 12, (b) follows from the definition of $\gamma_q(\beta)$ and (c) follows from (19). □

4.3 Classification Error Bounds

Theorem 15 established rates of convergence of $A(\hat{f})$ to $A(f_{\text{opt}})$. However, for binary classification problems, the main focus of this work, we wish to determine the rate at which $L(\hat{f})$ converges to the Bayes error L^*. However, from the work of Zhang (2002), reproduced as Theorem 4 above, we immediately obtain a bound on the classification error.

Corollary 16 *Let Assumption 1 holds. Then there exist constants $c_0, c_1 > 0$ that depend on ξ and K only, such that $\forall m \geq q \geq \max(q_0, c_0)$, with probability at least $1-\exp(-q)$,*

$$L(\hat{f}_{q_0}) \leq L^* + c_0 \left(\frac{q}{q_0}\right)^{1/2(1+\xi)} \inf_{\beta \geq 1} \left[\inf_{f \in \beta\mathcal{F}} \left(A(f) - A(f_{\text{opt}})\right) + \gamma_q(\beta)\right]^{1/2}. \tag{20}$$

Moreover, if the conditional probability $\eta(x)$ is uniformly bounded away from 0.5, namely $|\eta(x) - 1/2| \geq \delta > 0$ for all x, then with probability at least $1-\exp(-q)$,

$$L(\hat{f}_{q_0}) \leq L^* + c_1 \left(\frac{q}{q_0}\right)^{1/(1+\xi)} \inf_{\beta \geq 1} \left[\inf_{f \in \beta\mathcal{F}} (A(f) - A(f_{\text{opt}})) + \gamma_q(\beta)\right].$$

Proof The first inequality follows directly from Theorems 4 and 15, noticing the $s = 2$ for the least squares loss. The second inequality follows from Corollary 2.1 of Zhang (2002). According to this corollary

$$L(\hat{f}_{q_0}) \leq L^* + 2c \inf_{\delta > 0} \left[\left(\mathbf{E}_{|\eta(x) - \frac{1}{2}| < \delta} (\hat{f}_{q_0} - f_{\text{opt}})^2\right)^{1/2} + c' \frac{1}{\delta} (A(\hat{f}_{q_0}) - A(f_{\text{opt}}))\right].$$

The claim follows since by assumption the first term inside the infimum on the r.h.s. vanishes. □

In order to proceed to the derivation of complete convergence rates we need to assess the parameter ξ and the approximation theoretic term $\inf_{f \in \beta\mathcal{F}} A(f) - A(f_{\text{opt}})$, where we assume that $\mathcal{F} = \text{CO}(\mathcal{H})$. In order to do so we make the following assumption.

Assumption 2 *For all $h \in \mathcal{H}$, $\sup_x |h(x)| \leq M$. Moreover, $\mathcal{N}_2(\varepsilon, \mathcal{H}, m) \leq C\,(M/\varepsilon)^V$, for some constants C and V.*

Note that Assumption 2 holds for VC classes (e.g., van der Vaart and Wellner, 1996). The entropy of the class $\beta\text{CO}(\mathcal{H})$ can be estimated using the following result.

Lemma 17 (van der Vaart and Wellner, 1996, Theorem 2.6.9) *Let Assumption 2 hold for $\mathcal{H}$. Then there exists a constant K that depends on C and V only such that*

$$\log \mathcal{N}_2(\varepsilon, \beta \mathrm{CO}(\mathcal{H}), m) \leq K \left(\frac{\beta M}{\varepsilon}\right)^{\frac{2V}{V+2}}. \tag{21}$$

We use Lemma 17 to establish precise convergence rates for the classification error. In particular, Lemma 17 implies that ξ in Assumption 1 is equal to $V/(V+2)$, and indeed obeys the required conditions. We consider two situations, namely the non-adaptive and the adaptive settings. First, assume that $f_{\mathrm{opt}} \in \beta\mathcal{F} = \beta \mathrm{CO}(\mathcal{H})$ where $\beta < \infty$ is *known*.

In this case, $\inf_{f \in \beta\mathcal{F}} A(f) - A(f_{\mathrm{opt}}) = 0$, so that from (20) we find that for sufficiently large m, with high probability

$$L(\hat{f}_{q_0}) - L^* \leq O\left(m^{-(V+2)/(4V+4)}\right).$$

where we selected $\hat{f}_{q_0}$ based on (16) with $q = q_0$.

In general, we assume that $f_{\mathrm{opt}} \in B\mathrm{CO}(\mathcal{H})$ for some *unknown* but finite B. In view of the discussion in Section 2, this is a rather generic situation for sufficiently rich base classes $\mathcal{H}$ (e.g., non-polynomial ridge functions). Consider the adaptive procedure (16). In this case we may simply replace the infimum over β in (20) by the choice $\beta = B$. The approximation error term $\inf_{f \in \beta\mathcal{F}} A(f) - A(f_{\mathrm{opt}})$ vanishes, and we are left with the term $\gamma(B)$, which yields the rate

$$L(\hat{f}_{q_0}) - L^* \leq O\left(m^{-(V+2)/(4V+4)}\right). \tag{22}$$

We thus conclude that the adaptive procedure described above yields the same rates of convergence as the non-adaptive case, which uses prior knowledge about the value of β.

In order to assess the quality of the rates obtained, we need to consider specific classes of functions $\mathcal{H}$. For any function $f(x)$, denote by $\tilde{f}(\omega)$ its Fourier transform. Consider the class of functions introduced by Barron (1993) and defined as,

$$N(B) = \left\{ f : \int_{\mathbb{R}^d} \|\omega\|_1 |\tilde{f}(\omega)| d\omega \leq B \right\},$$

consisting of all functions with a Fourier transform which decays sufficiently rapidly. Define the approximating class composed of neural networks with a single hidden layer,

$$\mathcal{H}_n = \left\{ f : \ f(x) = c_0 + \sum_{i=1}^{n} c_i \phi(v_i^\top x + b_i), \ |c_0| + \sum_{i=1}^{n} |c_i| \leq B \right\},$$

where ϕ is a (non-polynomial) sigmoidal Lipschitz function. Barron (1993) showed that the class $\mathcal{H}_n$ is dense in $N(B)$.

For the class $N(B)$ we have the following worst case lower bound from Yang (1999)

$$\inf_{\hat{f}_m} \sup_{\eta \in N(B)} \mathbf{E} L(\hat{f}) - L^* \geq \Omega(m^{-(d+2)/(4d+4)}), \tag{23}$$

where $\hat{f}_m$ is any estimator based on a sample of size m, and by writing $h(m) = \Omega(g(m))$ we mean that there exist m_0 and C such that $h(m) \geq Cg(m)$ for $m \geq m_0$. As a specific example for a class $\mathcal{H}$, assume $\mathcal{H}$ is composed of monotonically increasing sigmoidal ridge functions. In this case one can show (e.g., Anthony and Bartlett, 1999) that $V = 2(d+1)$. Substituting in (22) we find a

rate of the order $O(m^{-(d+2)/(4d+6)})$, which is slightly worse than the minimax lower bound (23). Previously Mannor et al. (2002a), we also established convergence rates for the classification error. For the particular case of the squared loss and the class $N(B)$ we obtained the (non-adaptive) rate of convergence $O(m^{-1/4})$, which does not depend on the dimension d, as is required in the minimax bound. The necessary dependence on the dimension that comes out of the analysis in the present section, hinges on the utilization of the more refined bounding techniques used here.

5. Numerical Experiments

The algorithm presented in Figure 1 was implemented and tested for an artificial data set. The algorithm and the scripts that were used to generate the graphs that appear in this paper are provided in the online appendix (Mannor et al., 2002b) for completeness.

5.1 Algorithmic Details

The optimization step in the algorithm of Figure 1 is computationally expensive. Unfortunately, while the cost function $A((1-\alpha)f+\alpha h)$ is convex in α for a fixed h, it is not necessarily convex in the parameters that define h. The weak learners we used were sigmoidal $\mathcal{H}=\{h(x)=\sigma(\theta^\top \mathbf{x}+\theta_0)\}$. Given a choice of h it should be noted that

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

is convex in α. We therefore used a coordinate search approach where we search on α and h alternately. The search over α was performed using a highly efficient line search algorithm based on the convexity. The search over the parameters of h was performed using the Matlab optimization toolbox function *fminsearch*, which implements the Nedler-Mead algorithm (Nedler and Mead, 1965). Due to the occurrence of local minima (The number of minima may be exponentially large as shown in Auer et al., 1996), we ran several instances until convergence, starting each run with different initial conditions. The best solution was then selected.

5.2 Experimental Results

The two-dimensional data set that was used for the experiments was generated randomly in the following way. Points with positive labels were generated at random in a the unit circle (the radius and angle were chosen uniformly from $[0,1]$ and $[0,2\pi]$, respectively.) Points with negative labels were generated at random in the ring (in spherical coordinates) $\{(r,\theta): 2\le r\le 3, 0\le\theta<2\pi\}$ (the radius and angle were chosen uniformly from $[2,3]$ and $[0,2\pi]$, respectively.) The sign of each point was flipped with probability 0.05. A sample data set is plotted Figure 2a. The Bayes error of this data set is 0.05 ($\log_{10}(0.05)\approx -1.3$).

In order to investigate overfitting as a function of the regularization parameter β, we ran the following experiment. We fixed the number of samples $m=400$ and varied β over a wide range. We ran the greedy algorithm with the squared loss. As expected, the squared loss per sample decreases as β increases. It can also be seen that the empirical classification training error decreases when β increases, as can be seen in Figure 2b. Every experiment was repeated fifteen times and the error bars represent one standard deviation of the sample.

The generalization error is plotted in Figure 3a. It seems that for β that is too small the approximation power does not suffice, while for large values of β overfitting occurs. We note that for large values of β the optimization process may fail with non negligible probability.

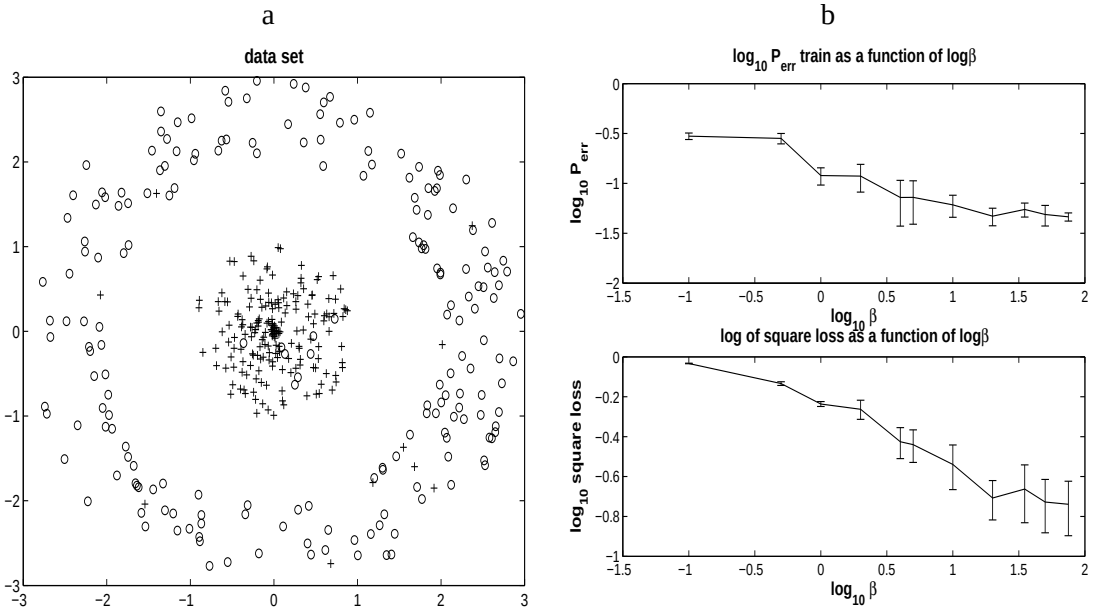

Figure 2: (a) An artificial data set, (b) Square loss and error probability for the artificial data set.

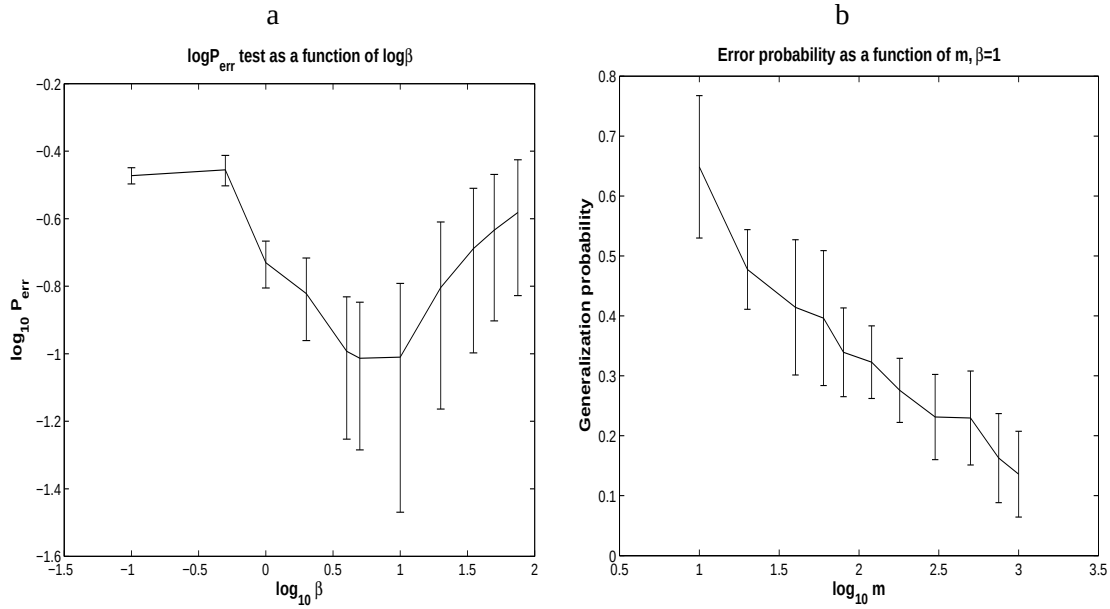

Figure 3: (a) Generalization error: as a function of β, sampled using 400 points; (b) Generalization error: plotted as a function of m for a fixed $\beta = 1$, sampled using 10-1000 points.

In spite of the overfitting phenomenon observed in Figure 3a we note that for a given value of β the performance improves with increasing sample size. For a fixed value of $\beta = 1$ we varied m from 10 to 1000 and ran the algorithm. The generalization error is plotted in Figure 3b (results are averaged over 15 runs and the error bars represent one standard deviation of the sample). We note that comparable behavior was observed for other data sets. Specifically, similar results were obtained for points in a noisy XOR configuration in two dimensions. We also ran experiments on the Indian Pima dataset. The results were comparable to state-of-the-art algorithms (the error for the Pima dataset using 15 fold cross validation was $29\% \pm 3\%$). The results are provided in detail, along with implementation sources, in an online appendix (Mannor et al., 2002b). The phenomenon that we would like to emphasize is the fact that for a fixed value of β, larger values of m lead to better prediction, as expected. However, the effect of regularization is revealed when m is fixed and β varies. In this case choosing a value of β which is too small leads to insufficient approximation power, while choosing β too large leads to overfitting.

6. Discussion

In this paper we have studied a class of greedy algorithms for binary classification, based on minimizing an upper bound on the $0-1$ loss. The approach followed bears strong affinities to boosting algorithms introduced in the field of machine learning, and additive models studies within the statistics community. While boosting algorithms were originally incorrectly believed to elude the problem of overfitting, it is only recently that careful statistical studies have been performed in an attempt to understand their statistical properties. The work of Jiang (2000b,a), motivated by Breiman (2000), was the first to address the statistical consistency of boosting, focusing mainly on the question of whether boosting should be iterated indefinitely, as had been suggested in earlier studies, or whether some early stopping criterion should be introduced. Lugosi and Vayatis (2001) and Zhang (2002) then developed a framework for the analysis of algorithms based on minimizing a continuous convex upper bound on the $0-1$ loss, and established universal consistency under appropriate conditions. The earlier version of this work (Mannor et al., 2002a) considered a stagewise greedy algorithm, thus extending the proof of universal consistency to this class of algorithms, and showing that consistency can be achieved by boosting forever, as long as some regularization is performed by limiting the size of a certain parameter. In Mannor et al. (2002a) we required prior knowledge of a smoothness parameter in order to derive convergence rates. Moreover, the convergence rates were worse than the minimax rates. In the current version, we have focused on the establishment of rates of convergence and the development of adaptive procedures, which assume nothing about the data, and yet converge to the optimal solution at nearly the minimax rate, which assumes knowledge of some smoothness properties.

While we have established nearly minimax rates of convergence and adaptivity for a certain class of base learners (namely ridge functions) and target distributions, these results have been restricted to the case of the squared loss where particularly tight rates of convergence are available. In many practical applications other loss functions are used, such as the logistic loss, which seem to lead to excellent practical performance. It would be interesting to see whether the rates of convergence established for the squared loss apply to a broader class of loss functions. Moreover, we have established minimaxity and adaptivity for a rather simple class of target functions. In future work it should be possible to extend these results to more standard smoothness classes (e.g., Sobolev and Besov spaces). Some initial results along these lines were provided in a previous paper (Mannor et al., 2002a), although the rates established in that work are not minimax. Another issue which

warrants further investigation is the extension of these results to multi-category classification problems.

Finally, we comment on the optimality of the procedures discussed in this paper. As pointed out in Section 4, near optimality for the adaptive scheme introduced in that section was established. On the other hand, it is well known that under very reasonable conditions Bayesian procedures (e.g., Robert, 2001) are optimal from a minimax point of view. In fact, it can be shown that Bayes estimators are essentially the only estimators which can achieve optimality in the minimax sense (Robert, 2001). This optimality feature provides strong motivation for the study of Bayesian type approaches in a *frequentist* setting (Meir and Zhang, 2003). In many cases Bayesian procedures can be expressed as a mixture of estimators, where the mixture is weighted by an appropriate prior distribution. The procedure described in this paper, as many others in the boosting literature, also generates an estimator which is formed as a mixture of base estimators. An interesting open question is to relate these types of algorithms to formal Bayes procedures, with their known optimality properties.

Acknowledgments We thank the three anonymous reviewers for their very helpful suggestions. The work of R.M. was partially supported by the Technion V.P.R. fund for the promotion of sponsored research. Support from the Ollendorff center of the department of Electrical Engineering at the Technion is also acknowledged. The work of S.M. was partially supported by the Fulbright postdoctoral grant and by the ARO under grant DAAD10-00-1-0466.

Appendix A

Proof of Lemma 6

In the following, we use the notation $g(x) = (f(x) - f_{\text{opt}}(x))^2$, and let $\mathcal{G} = \{g : g(x) = (f(x) - f_{\text{opt}}(x))^2, f \in \mathcal{F}\}$. Consider any $g \in \mathcal{G}$. Suppose we independently sample m points twice. We denote the empirical expectation with respect to the first m points by $\hat{\mathbf{E}}$ and the empirical expectation with respect to the second m points by $\hat{\mathbf{E}}'$. We note that the two sets of random variables are independent. We have from Chebyshev inequality ($\forall \gamma \in (0,1)$):

$$\mathbf{P}\left\{\left|\hat{\mathbf{E}}' g(X) - \mathbf{E}g(X)\right| \geq \gamma \mathbf{E}g(X) + \frac{M^2}{\gamma m}\right\} \leq \frac{\mathbf{Var}g(X)}{m} \frac{1}{(\gamma \mathbf{E}g(X) + \frac{M^2}{\gamma m})^2}.$$

Rearranging and taking the complement one gets that:

$$\mathbf{P}\left\{\hat{\mathbf{E}}' g(X) \geq (1-\gamma)\mathbf{E}g(X) - \frac{M^2}{\gamma m}\right\} \geq 1 - \frac{\mathbf{Var}g(X)}{m(\gamma \mathbf{E}g(X) + \frac{M^2}{\gamma m})^2}.$$

Since $0 \leq g(X) \leq M^2$ it follows that $\mathbf{Var}g(X) \leq \mathbf{E}g(X)^2 \leq M^2 \mathbf{E}g(X)$ so that:

$$\mathbf{P}\left\{\hat{\mathbf{E}}' g(X) \geq (1-\gamma)\mathbf{E}g(X) - \frac{M^2}{\gamma m}\right\} \geq 1 - \frac{\mathbf{E}g(X) M^2}{m(\gamma \mathbf{E}g(X) + \frac{M^2}{\gamma m})^2}.$$

Observe that for all positive numbers a, b, m, γ one has that

$$\frac{ab}{m(\gamma a + \frac{b}{\gamma m})^2} = \frac{1}{2 + m\gamma^2 \frac{a}{b} + \frac{b}{\gamma^2 m a}} \leq \frac{1}{4},$$

where the inequality follows since $a + \frac{1}{a} \geq 2$ for every positive number a. We thus have

$$\mathbf{P}\left\{\hat{\mathbf{E}}' g(X) \geq (1-\gamma)\mathbf{E}g(X) - \frac{M^2}{\gamma m}\right\} \geq \frac{3}{4}.$$

It follows (by setting $\gamma = 1/4$) that $\forall \varepsilon > 8\Delta_m^2$:

$$\begin{aligned}
&\frac{3}{4}\mathbf{P}\left\{\exists g \in \mathcal{G} : \mathbf{E}g(X) > 4\hat{\mathbf{E}}g(X) + \varepsilon + \frac{16M^2}{3m}\right\} \\
&\overset{(a)}{\leq} \mathbf{P}\left\{\exists g \in \mathcal{G} : \mathbf{E}g(X) > 4\hat{\mathbf{E}}g(X) + \varepsilon + \frac{16M^2}{3m} \;\&\; \hat{\mathbf{E}}'g(X) \geq \frac{3}{4}\mathbf{E}g(X) - \frac{4M^2}{m}\right\} \\
&\leq \mathbf{P}\left\{\exists g \in \mathcal{G} : \hat{\mathbf{E}}'g(X) > 3\hat{\mathbf{E}}g(X) + \frac{3\varepsilon}{4}\right\} \\
&\leq \mathbf{P}\left\{\exists g \in \mathcal{G} : 2|\hat{\mathbf{E}}'g(X) - \hat{\mathbf{E}}g(X)| > \hat{\mathbf{E}}g(X) + \hat{\mathbf{E}}'g(X) + \frac{3\varepsilon}{4}\right\},
\end{aligned}$$

where (a) follows by the independence of $\hat{\mathbf{E}}$ and $\hat{\mathbf{E}}'$ (Note that $\hat{\mathbf{E}}$ and $\hat{\mathbf{E}}'$ are random variables rather than expectations). Let $\{\sigma_i\}_{i=1}^m$ denote a set of independent identically distributed ± 1-valued random variable such $P\{\sigma_i = 1\} = 1/2$ for all i. We abuse notation somewhat and let $\hat{\mathbf{E}}\sigma g(X) = (1/m)\sum_{i=1}^m \sigma_i g(X_i)$, and similarly for $\hat{\mathbf{E}}'$. It follows that

$$\begin{aligned}
&\frac{3}{4}\mathbf{P}\left\{\exists g \in \mathcal{G} : \mathbf{E}g(X) > 4\hat{\mathbf{E}}g(X) + \varepsilon + \frac{16M^2}{3m}\right\} \\
&\leq \mathbf{P}\left\{\exists g \in \mathcal{G} : 2|\hat{\mathbf{E}}'\sigma g(X) - \hat{\mathbf{E}}\sigma g(X)| > \hat{\mathbf{E}}g(X) + \hat{\mathbf{E}}'g(X) + \frac{3\varepsilon}{4}\right\} \\
&\leq \mathbf{P}\left\{\exists g \in \mathcal{G} : 2(|\hat{\mathbf{E}}'\sigma g(X)| + |\hat{\mathbf{E}}\sigma g(X)|) > \hat{\mathbf{E}}g(X) + \hat{\mathbf{E}}'g(X) + \frac{3\varepsilon}{4}\right\} \\
&\overset{(a)}{\leq} 2\mathbf{P}\left\{\exists g \in \mathcal{G} : 2|\hat{\mathbf{E}}\sigma g(X)| > \hat{\mathbf{E}}g(X) + \frac{3\varepsilon}{8}\right\},
\end{aligned}$$

where (a) uses the union bound and the observation that $\hat{\mathbf{E}}$ and $\hat{\mathbf{E}}'$ satisfy the same probability law. For a fixed sample X let

$$\hat{\mathcal{G}}_s \triangleq \left\{g \in \mathcal{G} : 2^{s-1}\Delta_m^2 \leq \hat{\mathbf{E}}g(X) \leq 2^s\Delta_m^2\right\}.$$

We define the class $\sigma\hat{\mathcal{G}}_s = \{f : \; f(X_i) = \sigma_i g(X_i),\, g \in \hat{\mathcal{G}}_s,\, i = 1,2,\ldots,m\}$. Let $\hat{\mathcal{G}}_{s,\varepsilon/2}$ be an $\varepsilon/2$-cover of $\hat{\mathcal{G}}_s$, with respect to the ℓ_1^m norm, such that $\hat{\mathbf{E}}g \leq 2^s\Delta_m^2$ for all $g \in \hat{\mathcal{G}}_{s,\varepsilon/2}$. It is then easy to see that $\sigma\hat{\mathcal{G}}_{s,\varepsilon/2}$ is also an $\varepsilon/2$-cover of the class $\sigma\hat{\mathcal{G}}_s$. For each s we have

$$\begin{aligned}
\mathbf{P}_{X,\sigma}\left\{\exists g \in \hat{\mathcal{G}}_s : |\hat{\mathbf{E}}\sigma g(X)| > \varepsilon\right\} &= \mathbf{E}_X\mathbf{P}_\sigma\left(\exists g \in \hat{\mathcal{G}}_s : |\hat{\mathbf{E}}\sigma g(X)| > \varepsilon\right) \\
&\leq \mathbf{E}_X\mathbf{P}_\sigma\left(\exists g \in \hat{\mathcal{G}}_{s,\varepsilon/2} : |\hat{\mathbf{E}}\sigma g(X)| > \varepsilon/2\right) \\
&\overset{(a)}{\leq} 2\mathbf{E}_X\left|\hat{\mathcal{G}}_{s,\varepsilon/2}\right|\exp\left(-\frac{m\varepsilon^2}{2\hat{\mathbf{E}}g^2}\right) \\
&\leq 2\mathbf{E}\mathcal{N}_1(\varepsilon/2, \hat{\mathcal{G}}_s, \ell_1^m)\exp\left(-\frac{m\varepsilon^2}{2\hat{\mathbf{E}}g^2}\right) \\
&\leq 2\mathbf{E}\mathcal{N}_1(\varepsilon/2, \hat{\mathcal{G}}_s, \ell_1^m)\exp\left(-\frac{m\varepsilon^2}{M^2 2^{s+1}\Delta_m^2}\right), \qquad (24)
\end{aligned}$$

where in step (a) we used the union bound and Chernoff's inequality $\mathbf{P}(|\hat{\mathbf{E}}(\sigma g)| \geq \varepsilon) \leq 2\exp(-2m\varepsilon^2/\hat{\mathbf{E}}g^2)$. Using the union bound and noting that $\varepsilon > 8\Delta_m^2$, we have that:

$$\begin{aligned}
&\frac{3}{4}\mathbf{P}\left\{\exists g \in \mathcal{G} : \mathbf{E}g(x) > 4\hat{\mathbf{E}}g(x) + \varepsilon + \frac{16M^2}{3m}\right\} \\
&\leq 2\sum_{s=1}^{\infty}\mathbf{P}\left\{\exists g \in \hat{\mathcal{G}}_s : 2|\hat{\mathbf{E}}\sigma g(X)| > 2^{s-1}\Delta_m^2 + \frac{3\varepsilon}{8}\right\} \\
&\overset{(a)}{\leq} 4\sum_{s=1}^{\infty}\mathbf{E}\mathcal{N}_1(\varepsilon/11 + 2^{s-3}\Delta_m^2, \hat{\mathcal{G}}_s, \ell_1^m)\exp\left(-\frac{m(2^{s-2}\Delta_m^2 + \frac{3\varepsilon}{16})^2}{2^{s+1}\Delta_m^2 M^2}\right) \\
&\leq 4\sum_{s=1}^{\infty}\mathbf{E}\mathcal{N}_1(\varepsilon/11 + 2^{s-3}\Delta_m^2, \hat{\mathcal{G}}_s, \ell_1^m)\exp\left(-\frac{m2^s\Delta_m^2}{32M^2} - \frac{m\varepsilon}{32M^2}\right).
\end{aligned}$$

Inequality (a) follows from (24).

We now relate the ℓ_2 covering number of $\mathcal{F}$ to the ℓ_1 covering number of $\mathcal{G}$. Suppose that $\hat{\mathbf{E}}|f_1 - f_2|^2 \leq \varepsilon^2$. Using (11) this implies that

$$\begin{aligned}
\hat{\mathbf{E}}|(f_1 - f_{\text{opt}})^2 - (f_2 - f_{\text{opt}})^2| &\leq \varepsilon^2 + 2\hat{\mathbf{E}}|(f_2 - f_{\text{opt}})(f_1 - f_2)| \\
&\overset{(a)}{\leq} \varepsilon^2 + 2\sqrt{\hat{\mathbf{E}}(f_2 - f_{\text{opt}})^2}\sqrt{\hat{\mathbf{E}}(f_1 - f_2)^2} \\
&\overset{(b)}{\leq} \varepsilon^2 + 8\varepsilon^2 + \frac{1}{8}\hat{\mathbf{E}}(f_2 - f_{\text{opt}})^2,
\end{aligned}$$

where (a) follows from the Cauchy-Schwartz inequality, and (b) follows from the inequality $4a^2 + \frac{b}{16} \geq a\sqrt{b}$ (which holds for every a and b). Recalling that for $f_2 \in \hat{\mathcal{G}}_s$, $\hat{\mathbf{E}}(f_2 - f_{\text{opt}})^2 \leq 2^s\Delta_m^2$, we conclude that for all positive ε,

$$\mathcal{N}_1(9\varepsilon + 2^{s-3}\Delta_m^2, \hat{\mathcal{G}}_s, \ell_1^m) \leq e^{H(\sqrt{\varepsilon}, \mathcal{F}, m)}.$$

Note that we can choose ℓ_2-covers of $\hat{\mathcal{G}}_s$ so that their elements g satisfy $\hat{\mathbf{E}}g \leq 2^s\Delta_m^2$. Combining the above we have that $\forall \varepsilon > \Delta_m^2$:

$$\begin{aligned}
&\frac{3}{4}\mathbf{P}\left\{\exists g \in \mathcal{G} : \mathbf{E}g(x) > 4\hat{\mathbf{E}}g(x) + 100\varepsilon + \frac{16M^2}{3m}\right\} \\
&\leq 4\sum_{s=1}^{\infty} e^{H(\Delta_m, \mathcal{F}, m)}\exp\left(-\frac{m2^s\Delta_m^2}{32M^2} - \frac{100m\varepsilon}{32M^2}\right) \\
&\overset{(a)}{\leq} 4\sum_{s=1}^{\infty}\exp\left(\frac{m\Delta_m^2}{32M^2}\right)\exp\left(-\frac{m2^s\Delta_m^2}{32M^2} - \frac{3m\varepsilon}{M^2}\right) \\
&= 4\sum_{s=1}^{\infty}\exp\left(\frac{m\Delta_m^2}{32M^2}(1 - 2^s)\right)\exp\left(-\frac{3m\varepsilon}{M^2}\right) \\
&\leq 4\sum_{s=1}^{\infty}\exp\left(-\frac{m\Delta_m^2 2^{s-1}}{32M^2}\right)\exp\left(-\frac{3m\varepsilon}{M^2}\right) \\
&\overset{(b)}{\leq} 4\sum_{s=1}^{\infty}\exp\left(-2^{s-1}\right)\exp\left(-\frac{3m\varepsilon}{M^2}\right) \\
&\leq \frac{4e^{-1}}{1 - e^{-1}}e^{-3m\varepsilon/M^2} \\
&\leq 3e^{-3m\varepsilon/M^2}.
\end{aligned}$$

Here we used (12) in steps (a) and (b).

Set $q = -2 + 3m\varepsilon/M^2$, it follows that with probability at least $1 - \exp(-q)$ for all $g \in \mathcal{G}$

$$\mathbf{E}g(x) \leq 4\hat{\mathbf{E}}g(x) + 100\varepsilon + \frac{16M^2}{3m}.$$

By (12), $\frac{\Delta_m^2}{6} \geq \frac{16M^2}{3m}$. By our definition of q it follows that if $q \geq 3$ then $q+2 \leq 3q$ so that $\varepsilon \leq qM^2/m$. We conclude that with probability at least $1 - \exp(-q)$ for all $g \in \mathcal{G}$

$$\mathbf{E}g(x) \leq 4\hat{\mathbf{E}}g(x) + \frac{100qM^2}{m} + \frac{\Delta_m^2}{6}.$$

□

References

G. G. Agarwal and W. J. Studden. Asymptotic integrated mean square error using least squares and bias minimizing splines. *The Annals of Statistics*, 8:1307–1325, 1980.

M. Anthony and P.L. Bartlett. *Neural Network Learning; Theoretical Foundations*. Cambridge University Press, 1999.

A. Antos, B. Kégl, T. Lindet, and G. Lugosi. Date-dependent margin-based bounds for classification. *Journal of Machine Learning Research*, 3:73–98, 2002.

P. Auer, M. Herbster, and M. Warmuth. Exponentially many local minima for single neurons. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 316–322. MIT Press, 1996.

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

A.R. Barron. Universal approximation bound for superpositions of a sigmoidal function. *IEEE Trans. Inf. Th.*, 39:930–945, 1993.

A.R. Barron, L. Birgé, and P. Massart. Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

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

O. Bousquet and A. Chapelle. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998.

L. Breiman. Some infinity theory for predictor ensembles. Technical Report 577, Berkeley, August 2000.

P. Bühlmann and B. Yu. Boosting with the L_2 loss: regression and classification. *J. Amer. Statist. Assoc.*, 98:324–339, 2003.

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

T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 1999.

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.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 38(2):337–374, 2000.

T. Hastie and R. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, Berlin, 2001.

R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Boston, 2002.

W. Jiang. Does boosting overfit: Views from an exact solution. Technical Report 00-03, Department of Statistics, Northwestern University, 2000a.

W. Jiang. Process consistency for adaboost. Technical Report 00-05, Department of Statistics, Northwestern University, 2000b.

M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

V. Koltchinksii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statis.*, 30(1), 2002.

M. Leshno, V. Lin, A. Pinkus, and S. Schocken. Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate any Function. *Neural Networks*, 6:861–867, 1993.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of boosting methods. Technical report, Pompeu Fabra University, 2001.

G. Lugosi and N. Vayatis. A consistent strategy for boosting algorithms. In *Proceedings of the Fifteenth Annual Conference on Computational Learning Theory*, volume 2375 of *LNAI*, pages 303–318. Springer, 2002.

S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Processing*, 41(12):3397–3415, December 1993.

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.

S. Mannor and R. Meir. On the existence of weak learners and applications to boosting. *Machine Learning*, 48:219–251, 2002.

S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the fifteenth Annual conference on Computational learning theory*, volume 2375 of *LNAI*, pages 319–333, Sydney, 2002a. Springer.

S. Mannor, R. Meir, and T. Zhang. On-line appendix, 2002b. Available from http://www-ee.technion.ac.il/˜rmeir/adaptivityonlineappendix.zip.

L. Mason, P.L. Bartlett, J. Baxter, and M. Frean. Functional gradient techniques for combining hypotheses. In B. Schölkopf A. Smola, P.L. Bartlett and D. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 2000.

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

R. Meir and T. Zhang. Data-dependent bounds for Bayesian mixture methods. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 319–326. MIT Press, Cambridge, MA, 2003.

J.A. Nedler and R. Mead. A simplex method for function minimization. *Computer Journal*, 7: 308–313, 1965.

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

C. P. Robert. *The Bayesian Choice: A Decision Theoretic Motivation*. Springer Verlag, New York, second edition, 2001.

R. Schaback. A unified theory of radial basis functions. *J. of Computational and Applied Mathematics*, 121:165–177, 2000.

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

R.E. Schapire, Y. Freund, P.L. 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, 1998.

R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, U.K., 2000.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Verlag, New York, 1996.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, New York, 1982.

V. N. Vapnik. *Statistical Learning Theory*. Wiley Interscience, New York, 1998.

Y. Yang. Minimax nonparametric classification - part I: rates of convergence. *IEEE Trans. Inf. Theory*, 45(7):2271–2284, 1999.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statis.*, 2002. Accepted for publication.

T. Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Tran. Inf. Theory*, 49(3):682–691, 2003.