Journal of Machine Learning Research 7 (2006) 705–732 Submitted 7/05; Revised 12/05; Published 5/06

Some Theory for Generalized Boosting Algorithms

Peter J. Bickel BICKEL@STAT.BERKELEY.EDU
Department of Statistics
University of California at Berkeley
Berkeley, CA 94720, USA

Ya'acov Ritov (corresponding author) YAACOV.RITOV@HUJI.AC.IL
Department of Statistics and The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
91905 Jerusalem, Israel

Alon Zakai ALONZAKA@POB.HUJI.AC.IL
The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

Editor: Bin Yu

Abstract

We give a review of various aspects of boosting, clarifying the issues through a few simple results, and relate our work and that of others to the minimax paradigm of statistics. We consider the population version of the boosting algorithm and prove its convergence to the Bayes classifier as a corollary of a general result about Gauss-Southwell optimization in Hilbert space. We then investigate the algorithmic convergence of the sample version, and give bounds to the time until perfect separation of the sample. We conclude by some results on the statistical optimality of the L_2 boosting.

Keywords: classification, Gauss-Southwell algorithm, AdaBoost, cross-validation, non-parametric convergence rate

1. Introduction

We consider a standard classification problem: Let $(X,Y),(X_1,Y_1),\dots,(X_n,Y_n)$ be an i.i.d. sample, where $Y_i \in \{-1,1\}$ and $X_i \in \mathcal{X}$. The goal is to find a good classification rule, $\mathcal{X} \to \{-1,1\}$.

The AdaBoost algorithm was originally defined, Schapire (1990), Freund (1995), and Freund and Schapire (1996) as an algorithm to construct a good classifier by a "weighted majority vote" of simple classifiers. To be more exact, let $\mathcal{H}$ be a set of simple classifiers. The AdaBoost classifier is given by $\operatorname{sgn}\bigl(\sum_{m=1}^{M}\lambda_m h_m(x)\bigr)$, where $\lambda_m \in \mathbb{R}$, $h_m \in \mathcal{H}$, are found sequentially by the following algorithm:

0. Let $c_1 = c_2 = \cdots = c_n = 1$, and set $m = 1$.

1. Find $h_m = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} c_i h(X_i)Y_i$. Set

$$\lambda_m = \frac{1}{2}\log\Bigl(\frac{\sum_{i=1}^{n} c_i + \sum_{i=1}^{n} c_i h_m(X_i)Y_i}{\sum_{i=1}^{n} c_i - \sum_{i=1}^{n} c_i h_m(X_i)Y_i}\Bigr) = \frac{1}{2}\log\Bigl(\frac{\sum_{h_m(X_i)=Y_i} c_i}{\sum_{h_m(X_i)\neq Y_i} c_i}\Bigr).$$

©2006 Peter J. Bickel, Ya'acov Ritov and Alon Zakai.

2. Set $c_i \leftarrow c_i \exp\bigl(-\lambda_m h_m(X_i)Y_i\bigr)$, and $m \leftarrow m+1$, If $m \le M$, return to step 1.

M is unspecified and can be arbitrarily large.

The success of these methods on many data sets and their "resistance to overfitting"—the test set error continues to decrease even after all the training set observations were classified correctly, has led to intensive investigation to which this paper contributes.

Let $\mathcal{F}_\infty$ be the linear span of $\mathcal{H}$. That is,

$$\mathcal{F}_\infty = \bigcup_{k=1}^{\infty} \mathcal{F}_k, \text{ where } \mathcal{F}_k = \Bigl\{ \sum_{j=1}^{k} \lambda_j h_j : \lambda_j \in \mathbb{R},\ h_j \in \mathcal{H},\ 1 \le j \le k \Bigr\} .$$

A number of workers have noted, Breiman (1998,1999), Friedman, Hastie and Tibshirani (2000), Mason, Bartlett, Baxter and Frean (2000), and Schapire and Singer (1999), that the AdaBoost classifier can be viewed as $\operatorname{sgn}\bigl(F(X)\bigr)$, where F is found by a greedy algorithm minimizing

$$n^{-1} \sum_{i=1}^{n} \exp\bigl(-Y_i F(X_i)\bigr)$$

over $\mathcal{F}_\infty$.

¿From this point of view, the algorithm appeared to be justifiable, since as was noted in Breiman (1999) and Friedman, Hastie, and Tibshirani (2000), the corresponding expression $E\exp\bigl(-YF(X)\bigr)$, obtained by replacing the sum by expectation, is minimized by

$$F(X) = \frac{1}{2} \log\Bigl(P(Y=1|X)/P(Y=-1|X)\Bigr),$$

provided the linear span $\mathcal{F}_\infty$ is dense in the space $\mathcal{F}$ of all functions in a suitable way. However, it was also noted that the empirical optimization problem necessarily led to rules which would classify every training set observation correctly and hence not approach the Bayes rule whatever be n, except in very special cases. Jiang (2003) established that, for observation centered stumps, the algorithm converged to nearest neighbor classification, a good but rarely optimal rule.

In another direction, the class of objective functions $W(\cdot)$ that can be considered was extended by Friedman, Hastie, and Tibshirani (2000) to other W, in particular, $W(t) = \log(1+e^{-2t})$, whose empirical version they identified with logistic regression in statistics, and $W(t) = -2t + t^2$, which they referred to as "L_2 Boosting" and has been studied, under the name "matching pursuit", in the signal processing community. For all these objective functions, the population optimization of $EW\bigl(YF(X)\bigr)$ over $\mathcal{F}$ leads to a solution such that $\operatorname{sgn}F(X)$ is the Bayes rule. Friedman et al. also introduced consideration of other algorithms for the empirical optimization problem. Lugosi and Vayatis (2004) added regularization, changing the function whose expectation (both empirically and in the population) is to be minimized from $W\bigl(YF(X)\bigr)$ to $W_n\bigl(YF(X)\bigr)$ where $W_n \to W$ as $n \to \infty$. Bühlmann and Yu (2003) considered L_2 boosting starting from very smooth functions. We shall elaborate on this later.

We consider the behavior of the algorithm as applied to the sample $(Y_1,X_1),\dots,(Y_n,X_n)$, as well to the "population", that is when means are replaced by expectations and sums by probabilities. The structure of, and the differences between, the population and sample versions of the optimization problem has been explored in various ways by Jiang (2003), Zhang and Yu (2003), Bühlmann (2003), Bartlett, Jordan, and McAuliffe (2003), Bickel and Ritov (2003).

Our goal in this paper is

1. To clarify the issues through a few simple results.
2. To relate our work and that of Bühlmann (2003), Bühlmann and Yu (2003), Lugosi and Vayatis (2004), Zhang (2004), Zhang and Yu (2003) and Bartlett, Jordan, and McAuliffe (2003) to the minimax results of Mammen and Tsybakov (1999), Baraud (2001) and Tsybakov (2001).

In Section 2 we will discuss the population version of the basic boosting algorithms and show how their convergence and that of more general greedy algorithms can be derived from a generalization of Theorem 3 of Mallat and Zhang (1993) with a simple proof. The result can, we believe, also be derived from the even more general theorem of Zhang and Yu (2003), but our method is simpler and the results are transparent.

In Section 3 we show how Bayes consistency of various sample algorithms when suitably stopped or of sample algorithms based on minimization of a regularized W follow readily from population convergence of the algorithms and indicate how test bed validation can be used to do this in a way leading to optimal rates (in Section 4).

In Section 5 we address the issue of bounding the time to perfect separation of the different boosting algorithm (including the standard AdaBoost).

Finally in Section 6 we show how minimax rate results for estimating $E(Y|X)$ may be attained for a "sieve" version of the L_2 boosting algorithm, and relate these to results of Baraud (2001), Lugosi and Vayatis (2004), Bühlmann and Yu (2003), Barron, Birgé, Massart(1999) and Bartlett, Jordan and McAuliffe (2003). We also discuss the relation of these results to classification theory.

2. Boosting "Population" Theorem

We begin with a general theorem on Gauss-Southwell optimization in vector space. It is, in part, a generalization of Theorem 1 of Mallat and Zhang (1993) with a simpler proof. A second part relates to procedures in which the step size is regularized cf. Zhang and Yu (2003) and Bartlett et al. (2003). We make the boosting connection after its statement.

Let w be a real, bounded from below, convex function on a vector space $\mathbb{H}$. Let $\mathcal{H} = \mathcal{H}' \cup (-\mathcal{H}')$, where $\mathcal{H}'$ is a subset of $\mathbb{H}$ whose members are linearly independent, with linear span $\mathcal{F}_\infty = \{\sum_{m=1}^k \lambda_m h_m : \lambda_j \in \mathbb{R},\ h_j \in \mathcal{H},\ 1 \le j \le k, 1 \le k < \infty\}$. We assume that $\mathcal{F}_\infty$ is dense in $\mathbb{H}$, at least in the sense that $\{w(f) : f \in \mathcal{F}_\infty\}$ is dense in the image of w. We define two relaxed Gauss-Southwell "algorithms".

Algorithm I: For $\alpha \in (0,1]$, and given $f_1 \in \mathbb{H}$, find inductively $f_2, f_3, \ldots, \ldots$ by, $f_{m+1} = f_m + \lambda_m h_m$, $\lambda_m \in \mathbb{R}$, $h_m \in \mathcal{H}$ and

$$w(f_m + \lambda_m h_m) \le \alpha \min_{\lambda \in \mathbb{R},\, h \in \mathcal{H}} w(f_m + \lambda h) + (1-\alpha) w(f_m)\ . \tag{1}$$

Generalize Algorithm I to :

Algorithm II: Like Algorithm I, but replace (1) by

$$w(f_m + \lambda_m h) + \gamma \lambda_m^2 \le \alpha \min_{\lambda \in \mathbb{R},\, h \in \mathcal{H}} (w(f_m + \lambda h) + \gamma \lambda^2) + (1-\alpha) w(f_m)\ .$$

There are not algorithms in the usual sense since they do not specify a unique sequence of iterations but our theorems will apply to any sequence generated in this way. Technically, this scheme is used in the proof of Theorem 3. The standard boosting algorithms theoretically correspond to $\alpha = 1$, although in practice, since numerical minimization is used, α may equal 1 only approximately. Our generalization makes for a simple proof and covers the possibility that the minimum of $w(f_m + \lambda h)$ over $\mathcal{H}$ and $\mathbb{R}$ is not assumed, or multiply assumed. Let $\omega_0 = \inf_{f \in \mathcal{F}_\infty} w(f) > -\infty$. Let $w'(f;h)$ the linear operator of the Gataux derivative at $f \in \mathcal{F}_\infty$ in the direction $h \in \mathcal{F}_\infty$: $w'(f;h) = \partial w(f + \lambda h)/\partial\lambda\big|_{\lambda=0}$, and let $w''(f;h)$ be the second derivative of w at f in the direction h: $w''(f,h) \equiv \partial^2 w(f+\lambda h)/\partial\lambda^2\big|_{\lambda=0}$ (both derivative are assumed to exist). We consider the following conditions.

GS1. For any c_1 and c_2 such that $\omega_0 < c_1 < c_2 < \infty$,

$$\begin{aligned} 0 < \ & \inf\{w''(f,h) : c_1 < w(f) < c_2,\ h \in \mathcal{H}\} \\ & \le \ \sup\{w''(f,h) : w(f) < c_2,\ h \in \mathcal{H}\} < \infty. \end{aligned}$$

GS2. For any $c_2 < \infty$,

$$\sup\{w''(f,h) :\ w(f) < c_2,\ h \in \mathcal{H}\} < \infty\,.$$

Theorem 1 *Under Assumption* GS1*, any sequence of functions generated according to Algorithm I satisfies:*

$$w(f_m) \le \omega_0 + c_m$$

and if $c_m > 0$*:*

$$w(f_m) - w(f_{m+1}) \ge \xi(w(f_m)) > 0$$

where the sequence $c_m \to 0$ *and the function* $\xi(\cdot)$ *depend only on* α*, the initial points of the iterates, and* $\mathcal{H}$*. The same conclusion holds under Condition* GS2 *for any sequence* f_m *generated according to algorithm II.*

The proof can be found in Appendix A.

Remark:

1. Condition GS2 of Theorem 1 guarantees that $\sum_{m=1}^{\infty} \lambda_m^2 < \infty$. It can be replaced by any other condition that guarantees the same, for example, limiting the step size, replacing the penalty by other penalties, etc.

2. It will be clear from the proof in Appendix A that if w'' is bounded away from 0 and ∞ then c_m is of order $(\log m)^{-\frac{1}{2}}$ so that we, in fact, have an approximation rate – but it is so slow as to be essentially useless. On the other hand, with strong conditions such as orthonormality of the elements of $\mathcal{H}$, and $\mathcal{H}$ a classical approximation class such as trigonometric functions we expect, with L_2 boosting, to obtain rates such as $m^{-1/2}$ or better.

Let $(X,Y) \sim P$, $X \in \mathcal{X}$, $Y \in \{-1,1\}$. Let $\mathcal{H} \subset \{h : \mathcal{X} \to [-1,1]\}$ be a symmetric set of functions. In particular, $\mathcal{H}$ can, but need not, be a set of classifiers such as trees with

$$\mathcal{H} = -\mathcal{H}\,. \tag{2}$$

Given a loss function $W:\mathbb{R}\to\mathbb{R}^+$, we consider a greedy sequential procedure for finding a function F that minimizes $EW\big(YF(X)\big)$. That is, given $F_0\in\mathcal{H}$ fixed, we define for $m\geq 0$:

$$\begin{aligned}
\lambda_m(h) &= \arg\min_{\lambda\in\mathbb{R}} EW\Big(Y\big(F_m(X)+\lambda h(X)\big)\Big)\\
h_m &= \arg\min_{h\in\mathcal{H}} EW\Big(Y\big(F_m(X)+\lambda_m(h)h(X)\big)\Big)\\
F_{m+1} &= F_m+\lambda_m(h_m)h_m.
\end{aligned}$$

Assume, wlog (without loss of generality), by shifting and rescaling, that $W(0)=-W'(0)=1$. Note that by Bartlett et al. (2003), $W'(0)<0$ is necessary and sufficient for population consistency defined below. We can suppose again wlog in view of (2), that $\lambda_m\geq 0$. Define $\mathcal{F}_k$ and $\mathcal{F}_\infty$ as in Section 1 and let $\mathcal{F}\equiv\overline{\mathcal{F}_\infty}$ be the closure of $\mathcal{F}_\infty$ in convergence in probability:

$$\begin{aligned}
\mathcal{F} &\equiv \big\{F:\exists F_m\in\mathcal{F}_m,\ F_m(X)\xrightarrow{p}F(X)\big\}\\
F_\infty &\equiv \arg\min_{F\in\mathcal{F}}\ EW\big(YF(X)\big)
\end{aligned}$$

If $\mathrm{sgn}F_\infty$ is the Bayes rule for 0-1 loss, we say that F_∞ is population consistent for classification, “calibrated” in the Bartlett et al. terminology. Let

$$\begin{aligned}
p(X) &\equiv P(Y=1|X)\\
\widetilde{W}(x,d) &\equiv p(x)W\big(d\big)+\big(1-p(x)\big)W\big(-d\big).\\
\widetilde{W}(F) &\equiv \widetilde{W}(X,F(X))
\end{aligned}$$

By the assumptions below F_∞ is the unique function such that $\widetilde{W}'(F_\infty)=0$ with probability 1, where $\widetilde{W}'(F)=\widetilde{W}'(X,F(X))$ and $\widetilde{W}'(x,d)=\partial W(x,d)/\partial d$. Define $\widetilde{W}''$ similarly.

Here are some conditions.

P1. $P[p(X)=0 \text{ or } 1]=0$.

P2. W is twice differentiable and convex on $\mathbb{R}$.

P3. $\mathcal{H}$ is closed and compact in the weak topology. $\mathcal{F}$ is the set of all measurable functions on $\mathcal{X}$.

P4. $\widetilde{W}''(F)$ is bounded above and below on $\{F:\ c_1<\widetilde{W}(F)<c_2\}$ for all c_1,c_2 such that

$$\inf_{F\in\mathcal{F}} E\widetilde{W}(F)<c_1<c_2<E\widetilde{W}(F_0).$$

P5. $F_\infty\in L_2(P)$.

Note that P1 and P2 imply that $\widetilde{W}(x,d)\to\infty$ as $|d|\to\infty$, which ensures that F_∞ is finite almost anywhere. Condition P1, which says that no point can be classified with absolute certainty, is only needed technically to ensure that $\widetilde{W}(x,d)\to\infty$ as $|d|\to\infty$, even if W itself is monotone. It is not needed for L_2 boosting.

Conditions P2 and P4 ensure that along the optimizing path W behaves locally like $W_0(t)=-2t+t^2$ corresponding to L_2 boosting. They are more stringent than we would like and, in particular, rule out W such as the "hinge" appearing in SVM. More elaborate arguments such as those of Zhang and Yu (2003) and Bartlett et al. (2003) can give somewhat better results.

The functions commonly appearing in boosting such as, $W_1(t) = e^{-t}$, $W_2(t) = -2t + t^2$, $W_3(t) = -\log(1+e^{-2t})$ satisfy condition P4 if P1 also holds. This is obvious for W_2. For W_1 and W_3, it is clear that P4 holds, if P1 does, since otherwise $E\widetilde{W}\big(YF_m(X)\big) \to \infty$. The conclusions of Theorem 2 continue to hold if $h \in \mathcal{H} \implies |h| \geq \delta > 0$ since then below $w''(F;h) = Eh^2(X)\widetilde{W}\big(F(X)\big) \geq \delta^2 E\widetilde{W}(F(X))$ and P4 follows. Note that if $|h| \not\equiv 1$ the λ optimization step requires multiplying λ^2 by $Eh^2(x)$.

We have,

Theorem 2 *If $\mathcal{H}$ is a set of classifiers,* $(h^2 \equiv 1)$ *and Assumptions* P2 – P5 *hold, then*

$$F_m(X) \xrightarrow{P} F_\infty(X) \ ,$$

and the misclassification error, $P\big(YF_m(X) \leq 0\big) \to P[YF_\infty(X) \leq 0]$, *the Bayes risk.*

Proof Identify $w(F) = EW\big(YF(X)\big) = E\widetilde{W}\big(F(X)\big)$. Then,

$$w''(F,h) = Eh^2(X)\widetilde{W}''\big(F(X)\big) = E\widetilde{W}''\big(F(X)\big)$$

and (P4) can be identified with condition GS1 of Theorem 1. Thus,

$$E\widetilde{W}\big(F_m(X)\big) \to E\widetilde{W}(F_\infty(X)) \ .$$

Since,

$$E\widetilde{W}\big(F_m(X)\big) - E\widetilde{W}\big(F_\infty(X)\big) = E\left((F_\infty - F_m)^2 \int_0^1 \widetilde{W}''\big((1-\lambda)(X)F_\infty(X) + \lambda F_m(X)\big)\lambda d\lambda\right) \to 0 \ ,$$

the conclusion of Theorem 2 follows from (P4). The second assertion is immediate. ■

3. Consistency of the Boosting Algorithm

In this section we study the Bayes consistency properties of the sample versions of the boosting algorithms we considered in Section 2. In particular, we shall

(i) Show that under mild additional conditions, there will exist a random sequence $m_n \to \infty$ such that $\hat{F}_{m_n} \xrightarrow{P} F_\infty$, where $\hat{F}_m$ is defined below as the mth sample iterate, and moreover, that such a sequence can be determined using the data.

(ii) Comment on the relationship of this result to optimization for penalized versions of W. The difference is that the penalty forces $m < \infty$ to be optimal while with us, cross-validation (or a test bed sample) determines the stopping point. We shall see that the same dichotomy applies later, when we "boost" using the method of sieves for nonparametric regression studied by Barron, Birge and Massart (1999) and Baraud (2001).

3.1 The Golden Chain Argument

Here is a very general framework. This section is largely based on Bickel and Ritov (2003).

Let $\Theta_1 \subset \Theta_2 \subset \dots$ be a sequence of sets contained in a separable metric space, $\Theta = \overline{\cup\Theta_m}$ where $\overline{}$ denotes closure. Let $\Pi_m : \Theta_m \to 2^{\Theta_{m+1}}$ be a sequence of point to set mappings. Let K be a target function, and $\vartheta_\infty = \arg\min_{\vartheta\in\Theta} K(\vartheta)$. Finally, let $\hat{K}_n$ be a sample based approximation of K. We assume:

G1. $K : \Theta \to \mathbb{R}$ is strictly convex, with a unique minimizer ϑ_∞.

Our result is applicable to loosely defined algorithms. In particular we want to be able to consider the result of the algorithm applied to the data as if it were generated by a random algorithm applied to the population. We need therefore, the following definitions. Let $\mathcal{S}(\vartheta_0,\alpha)$ be the set of all sequences $\bar{\vartheta}_m \in \Theta_m$, $m = 0,1,\dots$ with $\bar{\vartheta}_0 = \vartheta_0$ and satisfying:

$$\begin{aligned}
&\bar{\vartheta}_{m+1} \in \Pi_m(\bar{\vartheta}_m)\\
&K(\bar{\vartheta}_{m+1}) \le \alpha \inf_{\vartheta\in\Pi_m(\bar{\vartheta}_m)} K(\vartheta) + (1-\alpha)K(\bar{\vartheta}_m).
\end{aligned}$$

The resemblance to Gauss-Southwell Algorithm I and the boosting procedures is not accidental. Suppose the following uniform convergence criterion is satisifed:

G2. If $\{\bar{\vartheta}_m\} \in \mathcal{S}(\vartheta_0,\alpha)$ with any initial ϑ_0, then $K(\bar{\vartheta}_m) - K(\bar{\vartheta}_{m+1}) \ge \xi\big(K(\bar{\vartheta}_m) - K(\vartheta_\infty)\big)$, for $\xi(\cdot) > 0$ strictly increasing, and $K(\bar{\vartheta}_m) - K(\vartheta_\infty) \le c_m$ where $c_m \to 0$ uniformly over $\mathcal{S}(\vartheta_0,\alpha)$.

In boosting, given P, $\Theta = \{F(X), F \in \tilde{\mathcal{F}}\}$ with a metric of convergence in probability, $\Theta_m = \{\sum_{j=1}^{m} \lambda_j h_j, h_j \in \mathcal{H}\}$, $\Pi_m(F) = \Pi(F) = \{F + \lambda h, \lambda \in \mathbb{R}, h \in \mathcal{H}\}$, and $K(F) = \mathrm{E}W\big(YF(X)\big)$. Condition G2, follows from the conclusion of Theorem 1.

Now suppose $\hat{K}_n(\cdot)$ is a sequence of random functions on Θ, empirical entities that resemble the population K. Let $\hat{\mathcal{S}}_n(\vartheta_0,\alpha')$ be the set of all sequences $\hat{\vartheta}_{0,n}, \hat{\vartheta}_{1,n}\dots$, such that $\hat{\vartheta}_{0,n} = \vartheta_0$, and

$$\begin{aligned}
&\hat{\vartheta}_{m+1,n} \in \Pi_m(\hat{\vartheta}_{m,n})\\
&\hat{K}_n(\hat{\vartheta}_{m+1,n}) \le \alpha' \min\{\hat{K}_n(\vartheta) :\ \vartheta \in \Pi_m(\hat{\vartheta}_{m,n})\} + (1-\alpha')\hat{K}_n(\hat{\vartheta}_{m,n}).
\end{aligned}$$

We assume

G3. $\hat{K}_n$ is convex, and for all integer m, $\sup\{|\hat{K}_n(\vartheta) - K(\vartheta)| : \vartheta \in A_m\} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$, for a sequence $A_m \subset \Theta_m$ such that $P(\hat{\vartheta}_{m,n} \in A_m) \to 1$.

In boosting, $\hat{K}_n(F) = n^{-1}\sum_{i=1}^{n} W\big(Y_i F(X_i)\big)$, $K(F) = E_p\big(YF(X)\big)$

The sequence $\{\bar{\vartheta}_m\}$ is the golden chain we try to follow using the obscure information in the sample.

We now state and prove,

Theorem 3 *If assumptions* G1– G3 *hold, and* $\alpha' \in (0,1]$, *then for any sequence* $\{\hat{\vartheta}_{m,n}\} \in \hat{\mathcal{S}}(\vartheta_0,\alpha')$, *there exists a subsequence* $\{\hat{m}_n\}$ *such that* $K(\hat{\vartheta}_{\hat{m}_n,n}) \xrightarrow{\text{p}} K(\vartheta_\infty)$.

Proof

Fix ϑ_0 and α, $\alpha < \alpha'$. Let $M_n \to \infty$ be some sequence, and let $\hat{m}_n = \arg\min_{m \le M_n} K(\hat{\vartheta}_{m,n})$. We need to prove that $K(\hat{\vartheta}_{\hat{m}_n,n}) \xrightarrow{\text{p}} K(\vartheta_\infty)$. We will prove this by contradiction. Suppose otherwise:

$$\inf_{m \le M_n} K(\hat{\vartheta}_{m,n}) - K(\vartheta_\infty) \ge c_1 > 0, \quad n \in \mathcal{N} \tag{3}$$

where $\mathcal{N}$ is unbounded with positive probability. Let $\varepsilon_{m,n} \equiv \sup_{\vartheta \in A_m} |K(\vartheta) - \hat{K}_n(\vartheta)|$. For any fixed m, $\varepsilon_{m,n} \xrightarrow{\text{a.s.}} 0$ by G3. Let

$$m_n = \arg\max\Big\{m' \le M_n : \forall m \le m', \varepsilon_{m-1,n} + 2\varepsilon_{m,n} < (\alpha' - \alpha)\xi(c_1) \;\&\; \hat{\vartheta}_{m,n} \in A_m\Big\}.$$

Clearly, $m_n \xrightarrow{\text{p}} \infty$, and for any $m \le m_n$, assuming (3):

$$\begin{aligned}
K(\hat{\vartheta}_{m,n}) &\le \hat{K}_n(\hat{\vartheta}_{m,n}) + \varepsilon_{m,n} \\
&\le \alpha' \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} \hat{K}_n(\vartheta) + (1-\alpha')\hat{K}_n(\hat{\vartheta}_{m-1,n}) + \varepsilon_{m,n} \\
&\le \alpha' \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha')K(\hat{\vartheta}_{m-1,n}) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&= \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\qquad - (\alpha'-\alpha)\big(K(\hat{\vartheta}_{m-1,n}) - \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta)\big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\qquad - (\alpha'-\alpha)\xi\Big(K(\hat{\vartheta}_{m,n}) - K(\vartheta_\infty)\Big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \\
&\qquad - (\alpha'-\alpha)\xi\Big(c_1\Big) + \varepsilon_{m-1,n} + 2\varepsilon_{m,n} \\
&\le \alpha \inf_{\vartheta \in \Pi_{m-1}\hat{\vartheta}_{m-1}} K(\vartheta) + (1-\alpha)K(\hat{\vartheta}_{m-1,n}) \text{ for all } m \le m_n\,.
\end{aligned}$$

Thus, there is a sequence $\{\bar{\vartheta}_1^{(n)}, \bar{\vartheta}_2^{(n)}, \dots\} \in \mathcal{S}(\vartheta_0, \alpha)$, such that $\bar{\vartheta}_m^{(n)} = \hat{\vartheta}_{m,n}$, $m \le m_n$. Hence, by Assumption G2, $K(\vartheta_{m_n,n}) \le K(\vartheta_\infty) + c_{m_n}$, where $\{c_m\}$ is independent of n, and $c_m \to 0$. Therefore, since $m_n \to \infty$, $K(\hat{\vartheta}_{m_n,n}) \to K(\vartheta_\infty)$, contradicting (3). ■

In fact we have proved that sequences m_n can be chosen in the following way involving K.

Corollary 4 *Let M_n be* any *sequence tending to ∞. Let $\tilde{m}_n = \arg\min\{K(\hat{\vartheta}_{m,n}) :\ 1 \le m \le M_n\}$. Then, under* G1 – G3, $\hat{\vartheta}_{\tilde{m}_n} \xrightarrow{P} \vartheta_\infty$.

To find $\hat{\vartheta}_{\hat{m}_n,n}$ which are totally determined by the data determining $\hat{K}_n$, we need to add some information about the speed of convergence of $\hat{K}_n$ to K on the “sample” iterates. Specifically, suppose we can determine, in advance, $M_n^* \to \infty$, $\varepsilon_n \to 0$ such that,

$$P[\sup\{|\hat{K}_n(\hat{\vartheta}_{m,n}) - K(\hat{\vartheta}_{m,n})| :\ 1 \le m \le M_n^*\} \ge \varepsilon_n] \le \varepsilon_n\,.$$

Then $\hat{m}_n = \arg\min\{\hat{K}_n(\hat{\vartheta}_{m,n}):\ 1 \le m \le M_n^*\}$ yields an appropriate $\hat{\vartheta}_{\hat{m}_n}$ sequence. We consider this in Section 4. Before that we return to the application of the result of this section to boosting.

3.2 Back to Boosting

We return to boosting, where we consider $\Theta_m = \{\sum_{j=1}^m \lambda_j h_j : \lambda_j \in \mathbb{R}, h_j \in \mathcal{H}\}$, and therefore $\Pi_m \equiv \Pi$, $\Pi(\vartheta) = \{\vartheta + \lambda h, \lambda \in \mathbb{R}, h \in \mathcal{H}\}$. To simplify notation, for any function $a(X,Y)$, let $P_n a(X,Y) = n^{-1}\sum_{i=1}^n a(X_i,Y_i)$ and $Pa(X,Y) = Ea(X,Y)$. Finally, we identify $\hat{\vartheta}_{m,n} = \sum_{j=1}^m \hat{\lambda}_j \hat{h}_j = \sum_{j=1}^m \hat{\lambda}_{j,n}\hat{h}_{j,n}$.

We assume further

GA1. $W(\cdot)$ is of bounded variation on finite intervals.

GA2. $\mathcal{H}$ has finite L_1 bracketing entropy.

GA3. There are finite $a_1, a_2, \ldots$ such that $\sup_n \sum_{j=1}^m |\hat{\lambda}_{j,n}| \le a_m$ with probability 1.

Theorem 5 *Suppose the conclusion of Theorem 1 and Conditions* GA1–GA3 *are satisfied, then conditions* G2, G3 *are satisfied.*

Proof Condition G2 follows from Theorem 1. It remains to prove the uniform convergence in Condition G3. However, GA2 and GA3 imply that $\mathcal{F} \equiv \{F : F = \sum_{j=1}^m \lambda_j h_j, h_j \in \mathcal{H}, |\lambda_j| \le M\}$ has finite L_1 bracketing entropy. Since W can be written as the difference of two monotone functions $\{W(YF) : F \in \mathcal{F}\}$ inherits this property. The result follows from Bickel and Millar (1991), Proposition 2.1. ■

4. Test Bed Stopping

Again we face the issue of data dependent and in some way optimal selection of $\hat{m}_n$. We claim that this can be achieved over a wide range of possible rates of convergence of $EW\big(\hat{F}_{\hat{m}_n}(YX)\big)$ to $EW\big(F_\infty(YX)\big)$ by using a test bed sample to pick the estimator. The following general result plays a key role.

Let $B = B_n \to \infty$, and let $(X,Y), (X_1,Y_1), \ldots, (X_{n+B}, Y_{n+B})$ be i.i.d. P, $X \in \mathcal{X}$, $|Y| \le 1$. Let $\hat{\vartheta}_m : \mathcal{X} \to \mathbb{R}$, $1 \le m \le m_n$ be data dependent functions which depend only on $(X_1,Y_1), \ldots, (X_n,Y_n)$ which are predictors of Y. For $g, g_1, g_2 :\ \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, given P, define

$$\begin{aligned}
\langle g_1, g_2\rangle_* &\equiv \frac{1}{B_n}\sum_{b=1}^{B_n} g_1(X_{b+n}, Y_{b+n}) g_2(X_{b+n}, Y_{b+n}) \\
\langle g_1, g_2\rangle_P &\equiv P\big(g_1(X,Y) g_2(X,Y)\big) = \int g_1(x,y) g_2(x,y) dP(x,y) \\
\|g\|_*^2 &\equiv \langle g_1, g_2\rangle_* \\
\|g\|_P^2 &\equiv \langle g_1, g_2\rangle_P
\end{aligned}$$

Let,

$$\tau = \arg\min\{\|Y - \hat{\vartheta}_m(X)\|_*^2 :\ 1 \le m \le M_n\}$$

and $\hat{\vartheta}_\tau$ be the selected predictor. Similarly, let

$$o = \arg\min\{\|Y - \hat{\vartheta}_m(X)\|_P^2 :\ 1 \le m \le M_n\}$$

and $\hat{\vartheta}_o$ be the corresponding predictor.

That is, $\hat{\vartheta}_o(X,Y)$ is the predictor an "oracle" knowing P and (X_i, Y_i), $1 \le i \le n$ would pick from $\hat{\vartheta}_1, \dots, \hat{\vartheta}_{M_n}$ to minimize squared error loss. Let $\vartheta_O(X) \equiv E_P(Y|X)$, the Bayes predictor. Let $\mathcal{P}$ be a set of probabilities and $r_n \equiv \sup\{E_P\|\hat{\vartheta}_O - \vartheta_O\|_P^2 :\ P \in \mathcal{P}\}$.

The following result is due to Györfi et al. (2002) (Theorem 7.1), although there it is stated in the form of an oracle inequality. We need the following condition:

C. $B_n r_n / \log M_n \to \infty$.

Theorem 6 *(Györfi et al.) Suppose condition* **C** *is satisfied, and* $|Y| \le 1$, $\|\hat{\vartheta}_m\|_\infty \le 1$. *Then,*

$$\sup\{\left|E_P(Y - \hat{\vartheta}_\tau)^2 - E_P(Y - \hat{\vartheta}_O)^2\right| : P \in \mathcal{P}\} = o(r_n).$$

Condition **C** very simply asks that the test sample size B_n be large only: (i) In terms of r_n, the minimax rate of convergence; (ii) In terms of the logarithm of the number of procedures being studied. If $|Y| \le 1$, there is no loss in requiring $\|\hat{\vartheta}_m\|_\infty \le 1$, since we could also replace $\hat{\vartheta}_m$ by its truncation at ±1, minimizing the L_2 cross validated test set risk. Along similar lines, using $\text{sgn}(\hat{\vartheta}_m)$ is equivalent to cross validating the probability of misclassification for these rules, since if $\hat{\vartheta}_m, Y \in \{-1, 1\}$, $E(Y - \hat{\vartheta}_m)^2 = 4P(\hat{\vartheta}_m \ne Y)$.

As we shall see in Section 6, typically $r_n = n^{-1+\delta}$, and M_n is at most polynomial in n. If n/B_n is slowly varying, we can check that the conditions hold. Essentially we can only not deal with r_n of order $n^{-1}\log n$.

5. Algorithmic Speed of Convergence

We consider now the time it takes the sample algorithm to convergence. The fact that the algorithm converges follows from Theorem 1. We show in this section that in fact the algorithm perfectly separates the data (*perfect separation* is achieved when $Y_i F_m(x_i) > 0$ for all $i = 1, \dots, n$) after no more than $c_1 n^2$ steps. Perfect separation is equivalent to empirical misclassification error 0.

The randomness considered in this section comes only from the Y_i, while the design points are considered fixed. We denote them, therefore, by lower case $x_1, \dots, x_n$. We consider the following assumptions:

O1. W has regular growth in the sense that $W'' < \kappa(W + 1)$ for some $\kappa < \infty$. Assume, wlog, that $W(0) = -W'(0) = 1$.

O2. Suppose $x_1, \dots, x_n$ are all different Then the points can be finitely isolated by $\mathcal{H}$ in the sense that there is k and positive $\alpha_1, \dots, \alpha_k$ such that for every i there are $h_1, \dots, h_k \in \mathcal{H}$ such that $\sum_{j=1}^k \alpha_j h_j(x_s) = 1$ if $s = i$, and 0 otherwise. Assume further, as usual, that if $h \in \mathcal{H}$ then $h^2 \equiv 1$ and $-h \in \mathcal{H}$.

Condition O1 is satisfied by all the loss functions mentioned in the introduction. Condition O2 is satisfied, for example by stumps, trees, and any $\mathcal{H}$ whose span includes indicators of small sets with arbitrary location. In particular, if $x_i \in \mathbb{R}$, $x_1 < x_2 < \cdots < x_n$, and $\mathcal{H} = \{\text{sgn}(\cdot - x), x \in \mathbb{R}\}$, we can then take $\alpha_1 = \alpha_2 = 1$, $h_1(\cdot) = \text{sgn}\big(\cdot - (x_{i-1} + x_i)/2\big)$, and $h_2(\cdot) = -\text{sgn}\big(\cdot - (x_i + x_{i+1})/2\big)$

Theorem 7 *Suppose assumptions* O1 *and* O2 *are satisfied and the algorithm starts with* $F_0(0) = 0$. *If* $Y_iF_m(x_i) < 0$ *for at least one i, then*

$$\frac{1}{n}\sum_{i=1}^{n} W\big(Y_iF_m(x_i)\big) - \frac{1}{n}\sum_{i=1}^{n} W\big(Y_iF_{m+1}(x_i)\big) \geq \frac{1}{2\kappa\big(n\sum_{j=1}^{k}\alpha_j\big)^2}.$$

Hence, the boosting algorithm perfectly separates the data after at most $2\kappa(n\sum_{j=1}^{k}|\alpha_j|)^2$ *steps.*

Proof Let, for i such that $Y_iF_m(x_i) < 0$,

$$f_m(\lambda;h) = n^{-1}\sum_{s=1}^{n} W\Big(Y_i\big(F_m(x_s) + \lambda h(x_s)\big)\Big),$$

and $f'_m(0;h) = d\,f_m(\lambda;h)/d\lambda\big|_{\lambda=0}$. Consider $h_1,\ldots,h_k$ as in assumption O2. Replace h_j by $-h_j$ if necessary to ensure that $Y_i\sum_{j=1}^{k}\alpha_j h_j(x_s) = \delta_{si}$. Then

$$\begin{aligned}\sum_{j=1}^{k}\alpha_j f'_m(0;h_j) &= n^{-1}\sum_{j=1}^{k}\alpha_j\sum_{s=1}^{n} W'\big(Y_iF_m(x_s)\big)Y_ih_j(x_s)\\ &= n^{-1}W'\big(Y_iF_m(x_i)\big).\end{aligned}$$

Hence

$$\inf_{h\in\mathcal{H}} f'_m(0;h) \leq \frac{1}{n\sum_{j=1}^{k}\alpha_j}\min_i W'(Y_iF_m(x_i)) \leq \frac{W'(0)}{n\sum_{j=1}^{k}\alpha_j} = \frac{-1}{n\sum_{j=1}^{k}\alpha_j}, \tag{4}$$

since $Y_iF_m(x_i) < 0$ for at least one i.

Let $\bar{h}$ be the minimizer of $f'_m(0,h)$. Note that in particular $f'_m(0;\bar{h}) < 0$. The function $f_m(\cdot;\bar{h})$ is convex, hence it is decreasing in some neighborhood of 0. Denote by $\bar{\lambda}$ its minimizer. Consider the Taylor expansion:

$$\begin{aligned}f_m(\bar{\lambda};\bar{h}) &= f_m(0;\bar{h}) + \bar{\lambda} f'_m(0;\bar{h}) + \frac{\bar{\lambda}^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\\ &= f_m(0;\bar{h}) + \inf_{\lambda}\bigg\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\bigg\}\end{aligned}$$

where $\tilde{\lambda}(\lambda)$ lies between 0 and $\bar{\lambda}$. By condition 01,

$$\begin{aligned}&\inf_{\lambda}\bigg\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2}{2n}\sum_{s=1}^{n} W''\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big)\bigg\}\\ &\leq \inf_{\lambda}\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2\kappa}{4n}\sum_{s=1}^{n} W\Big(Y_i\big(F_m(x_s) + \tilde{\lambda}(\lambda)\bar{h}(x_s)\big)\Big) + \frac{\lambda^2\kappa}{4}\bigg\}\\ &\leq \inf_{\lambda}\{\lambda f'_m(0;\bar{h}) + \frac{\lambda^2\kappa}{2}\}\end{aligned} \tag{5}$$

because $\frac{1}{n}\sum_{s=1}^{n} W(Y_i(F_m(x_s)+\widetilde{\lambda}(\lambda)\overline{h}(x_s)) \le \frac{1}{n}\sum_{s=1}^{n} W(Y_iF_m(x_s)) \le W(0)=1$ since $\overline{\lambda}$ minimizes $f_m(\lambda;\overline{h})$ on $[0,\overline{\lambda}]$, $\widetilde{\lambda}$ is an intermediate point, and $F_0 \equiv 0$. Combining (4) and (5) and the minimizing property of $\overline{h}$,

$$\begin{aligned} f_m(\overline{\lambda};\overline{h}) &\le f_m(0;\overline{h}) - \frac{\left(f'_m(0;\overline{h})\right)^2}{2\kappa} \\ &\le f_m(0;\overline{h}) - \frac{1}{2\kappa(n\sum_{j=1}^{k}\alpha_j)^2} \quad . \end{aligned}$$

The second statement of the theorem follows because the initial value of $n^{-1}\sum_{i=1}^{n} W\big(Y_iF_0(x_i)\big)$ is 1, and the value would fall below 0 after at most $m = 2\kappa(n\sum_{j=1}^{k}\alpha_j)^2$ steps in which at least one observation is not classified correctly. Since the value is necessarily positive, we conclude that all observations would be classified correctly before the mth step.

■

6. Achieving Rates with Sieve Boosting

We propose a regularization of L_2 boosting which we view as being in the spirit of the original proposal, but, unlike it, can be shown for, suitable $\mathcal{H}$, to achieve minimax rates for estimation of $E(Y|X)$ under quadratic loss for $\mathcal{P}$ for which $E(Y|X)$ is assumed to belong to a compact set of functions such as a ball in Besov space if $X \in \mathbb{R}$ or to appropriate such subsets of spaces of smooth functions in $X \in \mathbb{R}^d$—see, for example, the classes $\mathcal{F}$ of Györfi et al. (2003). In fact, they are adaptive in the sense of Donoho et al (1995) for scales of such spaces. We note that Bühlmann and Yu (2003) have introduced a version of L_2 boosting which achieves minimax rates for Sobolev classes on $\mathbb{R}$ adaptively already. However, their construction is in a different spirit than that of most boosting papers. They start out with $\mathcal{H}$ consisting of one extremely smooth and complex function and show that boosting reduces bias (roughness of the function) while necessarily increasing variance. Early stopping is still necessary and they show it can achieve minimax rates.

It follows, using a result of Yang (1999) that our rule is adaptive minimax for classification loss for some of the classes we have mentioned as well. Unfortunately, as pointed out by Tsybakov (2001), the sets $\{x: \ |F_B(x)| \le \varepsilon\}$ can behave very badly as $\varepsilon \downarrow 0$, no matter how smooth F_B, the misclassification Bayes rule, is, so that these results are not as indicative as we would like them to be. In a recent paper, Bartlett, Jordan, and McAuliffe (2003) considered minimization of the W empirical risk $n^{-1}\sum_{i=1}^{n} W(Y_iF(X_i))$, for fairly general convex W, over sets of the form $\mathcal{F} = \{F = \sum_{j=1}^{m}\alpha_j h_j,\ h_j \in \mathcal{H},\ \sum_{j=1}^{m}|\alpha_j| \le \alpha_n,$ (for some representation of F)$\}$. They obtained oracle inequalities relating $EW(Y\hat{F}(X))$ for $\hat{F}_j$ the empirical minimizer over $\mathcal{F}_j$ to the empirical W risk minimum. They then proceeded to show using conditions related to Tsybakov's (A1) above how to relate the misclassification regret of $\hat{\mathcal{F}}_j$, given by $\langle P[Y\hat{F}_j(X)<0]-P[YF_B(X)<0]\rangle$ to $\langle E_pW(Y\hat{F}_j) - E_pW(YF_B^*)\rangle$, the W regret where F_B^* is the Bayes rule for W. Using these results (Theorems 3 and 10) they were able to establish oracle inequalities for $\hat{F}_j$ under misclassification loss. Manor, Meir, and Zhang (2004) considered the same problem, but focused their analysis mainly on L_2 boosting. They obtained an oracle inequality similar to that of Bartlett et al. regularizing by permitting step sizes which are only a fraction $\beta < 1$ of the step size declared optimal by Gauss-Southwell. They went further by obtaining near minimax results on suitable sets.

We also limit our results to L_2 boosting, although we believe this limitation is primarily due to the lack of minimax theorems for prediction when other losses than L_2 are considered. We use yet a different regularization method in what follows. We show in Theorem 8 our variant of L_2 boosting achieves minimax rates for estimating $E(Y|X)$ in a wide class of situations. Boosting up to a simple data-determined cutoff in each sieve level of a model, and then cross-validating to choose between sieve levels, we can obtain results equivalent to those in which full optimization using penalties are used, such as Theorem 2.1 of Baraud (2000) and results of Baron, Birgé, Massart (1999). Then, in Theorem 9, we show, using inequalities related to ones of Tsybakov (2001), Zhang (2004) and Bartlett et al. (2003), that the rules we propose are also minimax for 0–1 loss in suitable spaces.

6.1 The Rule

Our regularization requires that $\mathcal{H} \equiv \mathcal{H}^{(\infty)} = \overline{\cup_{m\geq 1}\mathcal{H}^{(m)}}$ where $\mathcal{H}^{(m)}$ are finite sets with certain properties. For instance, if $\mathcal{H}$ consists of the stumps in $[0,1]$, $\mathcal{H} = \{F_y(\cdot):\ F_y(x) = \mathrm{sgn}(x-y),\ x,y \in [0,1]\}$ we can take $\mathcal{H}^{(m)} = \{F_y(\cdot):\ y$ a dyadic number of order k, $y = \frac{j}{2^k},\ 0 \leq j \leq 2^k\}$. Essentially, we construct a sieve approximating $\mathcal{H}$. Let $\mathcal{F}^{(m)}$ be the linear span of $\mathcal{H}^{(m)}$. Evidently $\mathcal{F} = \overline{\cup_{m\geq 1}\mathcal{F}^{(m)}}$. Let $|\mathcal{H}^{(m)}| \equiv D_m$. Then, $\dim(\mathcal{F}^{(m)}) = D_m$. We now describe our proposed regularization of L_2 boosting.

We use the following notation of Section 4, and begin with a glossary and conditions. Let $(X_1,Y_1),\ldots,(X_n,Y_n),(X,Y)$ i.i.d. with

$$
\begin{aligned}
(X,Y) &\sim P << \mu, \quad P \in \mathcal{P}, \quad \mathbf{X} \equiv (X_1,\ldots,X_n),\ \mathbf{Y} \equiv (Y_1,\ldots,Y_n)\ . \\
Y &\in \{-1,1\} \\
\|f\|_\mu^2 &\equiv \int f^2\, d\mu \\
\|f\|_n^2 &\equiv \frac{1}{n}\sum_{i=1}^n f^2(X_i,Y_i) \\
\|f\|_\infty &= \sup_{x,y} |f(x,y)| \\
F_P(X) &\equiv E_P(Y|X) \\
\hat{F}_m(X) &= \arg\min\{\|t(X)-Y\|_n^2:\ t \in \mathcal{F}^{(m)}\} \\
F_m(X) &= \arg\min\{\|t(X)-Y\|_P^2:\ t \in \mathcal{F}^{(m)}\} \\
E_{\mathbf{X}} &\equiv \text{ Conditional expectation given } X_1,\ldots,X_n
\end{aligned}
$$

Note that we will often suppress $\mathbf{X},\mathbf{Y}$ in $v(\mathbf{X},\mathbf{Y},X,Y)$ and drop subscript to P.

Let $\hat{F}_{m,k}$, the kth iterate in $\mathcal{F}_m$, be defined as follows

$$
\begin{aligned}
\hat{F}_{1,0} &\equiv F_0 \\
\hat{F}_{m+1,0} &= \hat{F}_{m,\hat{k}(m)} \\
\hat{F}_{m,k+1} &= \hat{F}_{m,k} + \hat{\lambda}_{m,k}\hat{h}_{m,km}
\end{aligned}
$$

where

$$
\begin{aligned}
(\hat{\lambda}_{m,k},\hat{h}_{m,k}) &\equiv \arg\min_{\lambda\in\mathbb{R},h\in\mathcal{H}^{(m)}} \{-2\lambda P_n(Y-\hat{F}_{m,k})h + \lambda^2 P_n(h^2)\} \\
\hat{k}(m) &= \text{ First } k \text{ such that } \hat{\lambda}_{m,k}^2 \leq \Delta_{m,n},
\end{aligned}
$$

where $\Delta_{m,n}$ are constants. Let

$$\tilde{F}_m = H(\hat{F}_{m,\hat{k}(m)})$$

where

$$H(x) = \begin{cases} x & \text{if } |x| \le 1 \\ \operatorname{sgn}(x) & \text{if } |x| > 1 \end{cases} \tag{6}$$

Note that we have suppressed dependence on n here, indicating it only by the "hats". Let,

$$\hat{m} = \arg\min\{\|Y - \widetilde{F}_m(x)\|_* : \; m \le M_n\}$$

where

$$\|f\|_*^2 = \frac{1}{B}\sum_{i=n+1}^{n+B} f^2(X_i, Y_i), \text{ and we take } B = B_n = \frac{n}{\log n} \;.$$

The rule we propose is: $\hat{\hat{\delta}} = \operatorname{sgn}(\hat{\hat{F}})$, where

$$\hat{\hat{F}} \equiv H(F_{\hat{m},\hat{k}(\hat{m})}) \;. \tag{7}$$

Note: We show at the end of the Appendix (Proof of Lemma 10) that for wavelet $\mathcal{H}$ we take at most $Cn\log n$ steps total in this algorithm.

6.2 Conditions and Results

We use C as a generic constant throughout, possibly changing from line to line but not depending on m, n, or P. Lemma 6.3 and the condition we give are essentially due to Baraud (2001). Let μ be a sigma finite measure on $\mathcal{H}$ and $\|f\|_\mu$ be the $L_2(\mu)$ norm.

R1. If $\mathcal{H}^{(m)} = \{h_{m,1}, \ldots, h_{m,D_m}\}$ and $f_{m,j} \equiv h_{m,j}/\|h_{m,j}\|_\mu$, then $\{f_{m,j}\}, j \ge 1$ is an orthonormal basis of $\mathcal{F}^{(m)}$ in $L_2(\mu)$ such that:

(i) $\|f_{m,j}\|_\infty \le C_\infty D_m^{\frac{1}{2}}$ for all j, where $\|f\|_\infty = \sup_x |f(x)|$.

(ii) There exists an L such that for all m, j, j',
$f_{m,j} f_{m,j'} = 0$ if $|j - j'| \ge L$.

R2. There exists $\varepsilon = \varepsilon(P) > 0$ such that, $\varepsilon \le \frac{dP}{d\mu} \le \varepsilon^{-1}$ for all $P \in \mathcal{P}$.

R3. $\sup_{P \in \mathcal{P}} \|F_P - F_m\|_P^2 \le C D_m^{-\beta}$ for all m, $\beta > 1$.

R4. $M_n \le D_{M_n} \le \frac{n}{(\log n))^p}$ for some $p > 1$.

Condition R1 is needed to conclude that we can bound the behavior of the L_∞ norm on $\mathcal{F}^{(m)}$ by that of the L_2 norm for μ. Condition R2 simply ensures that we can do so for $P \in \mathcal{P}$ as well. The members $f_{m,j}$ of the basis of $\mathcal{F}^{(m)}$ must have compact support. It is well known that if $\mathcal{H}_m$ consists of scaled wavelets (in any dimension) then R1 holds. Clearly, if say μ is Lebesgue measure on an hypercube then to satisfy R2 $\mathcal{P}$ can consist only of densities bounded from above and away from 0. Condition R3 gives the minimum approximation error incurred by using an estimate F based on $\mathcal{F}^{(m)}$, and thus limits our choice of $\mathcal{H}$. Finally, R4 links the oracle error for these sequences of procedures to the number of candidate procedures.

Let

$$r_n(P) = \inf\{E_P\|\hat{F}_m - F_P\|_P^2 :\ 1 \le m \le M_n\}, \qquad r_n \equiv \sup_{P\in\mathcal{P}} r_n(P).$$

Thus, r_n is the minimax regret for an oracle knowing P but restricted to $\hat{F}_m$. We use the notation $a_n \asymp b_n$ for a shortcut for $a_n = O(b_n)$ and $b_n = O(a_n)$, We have

Theorem 8 *Suppose that $\mathcal{P}$ and $\mathcal{F}$ satisfy* R1–R4 *and that $\mathcal{H}$ is a VC class. If $\Delta_{m,n} = O(D_m/n)$, then,*

$$\sup_{\mathcal{P}} E_P\|\hat{\hat{F}}(X) - F_P(X)\|_P^2 \asymp r_n\ . \tag{8}$$

Thus, $\hat{\hat{F}}$ given by (7) is rate minimax.

Theorem 9 *Suppose the assumptions of Theorem 8 hold and $\mathcal{P}_0 = \mathcal{P}\cap\{P:\ P\big(\,|F_P(X)|\le t\big)\le ct^{\alpha}\}$, $\alpha \ge 0$. Let $\Delta_n(F,P)$ be the Bayes classification regret for P,*

$$\Delta_n(F,P) \equiv P\big(YF(X)<0\big) - P\big(YF_P(X)<0\big)\ . \tag{9}$$

Then,

$$\sup_{\mathcal{P}_0} \Delta_n(\hat{\hat{F}},P) \asymp r_n^{\frac{\alpha+1}{\alpha+2}}\ . \tag{10}$$

The condition $P[|F_P(x)|\le t] \le ct^{\alpha}$, some $\alpha \ge 0$, t sufficiently small appears in Proposition 1 of Tsybakov (2001) as sufficient for his condition (A1) which is studied by both Bartlett et al. (2003) and Mammen and Tsybakov (1999).

The proof of Theorem 9 uses 2 lemmas of interest which we now state. Their proofs are in the Appendix.

We study the algorithm on $\mathcal{F}_m$. For any positive definite matrix Σ define the condition number $\gamma(\Sigma) \equiv \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$, where $\lambda_{\max}$, $\lambda_{\min}$ are the largest and smallest eigenvalues of Σ. Let $G_m(P) = \|E_P f_{m,i}(X) f_{m,j}(X)\|$ be the $D_m \times D_m$ Gram matrix of the basis $\{f_{m,1},\ldots,f_{m,D_m}\}$.

Lemma 10 *Under* R1 *and* R2*,*

a) $\gamma(G_m(P)) \le \varepsilon^{-2}$, where ε is as in R2.

b) Let $G_m(P_n)$ be the empirical Gram matrix $\hat{\gamma}_m \equiv \gamma(G_m(P_n))$. Then, if in addition to R1 *and* R2*, $\mathcal{H}$ is a VC class, $P[\gamma(\hat{G}_m) \ge C_1] \le C_2 \exp\{-C_3 n/L^2 D_m\}$ for all $m \le M_n$ for such that $D_m \le n/(\log n)^p$ for $p > 1$.*

c) If $\mathcal{H}$ is a VC class, $P[\|\hat{F}_{m,\hat{k}(m)} - \hat{F}_m\|_k \le C\frac{D_m}{n}] = 1 - O(\frac{1}{n})$ The C and 0 terms are determed solely by the constants appearing in the R conditions.

Lemma 11 *Suppose* R1, R2, *and* R4 *hold. Then,*

$$E_P(\widetilde{F}_m - F_P)^2 \le C\{E_P(F_m - F_P)^2 + \frac{D_m}{n} + E_P(\tilde{F}_m - \hat{F}_m)^2\}.$$

This "oracle inequality" is key for what follows.

Proof of Theorem 9

$$P\big(YF(x) < 0\big) = \frac{1}{2}E_P\Big(1(F(X) > 0)\big(1 - F_P(X)\big)\Big) + \frac{1}{2}E_P\Big(1(F(X) < 0)\big(1 + F_P(X)\big)\Big).$$

Hence for all $\varepsilon > 0$,

$$\begin{aligned}\Delta_n(F,P) &= E_P\Big(1\big(F(X) < 0, F_P(X) > 0\big)F_P(X) - 1\big(F(X) > 0, F_P(X) < 0\big)F_P(X)\Big)\\ &= E_P\Big(|F_P(X)|1(F_P(X)F(X) < 0)\Big)\\ &\le E_P\Big(\big|F(X) - F_P(X)\big|1(F_PF(X) < 0, |F_P(X)| > \varepsilon)\Big) + \varepsilon P\big(|F_P(X)| \le \varepsilon\big)\\ &\le \frac{1}{\varepsilon}E_P\big(F(X) - F_P(X)\big)^2 + c\varepsilon^{\alpha+1}\end{aligned}$$

by assumption. The theorem follows. ■

6.3 Discussion

1) If $X \in \mathbb{R}$ and $\mathcal{H}^{(m)}$ consists of stumps with the discontinuity at a dyadic rational $j/2^m$, then $\mathcal{F}^{(m)}$ is the linear space of Haar wavelets of order m. This is also true if $\mathcal{H}_m$ is the space of differences of two such dyadic stumps. More generally, if $\mathcal{H}$ consists of suitably scaled wavelets, so that $|h| \le 1$, based on the dyadic rationals of order m, them $\mathcal{F}^{(m)}$ is the linear space spanned by the first 2^m elements of the wavelet series. A slight extension of results of Baraud (2001) yields that if we run the algorithm to the limit $k = \infty$ for each m rather than stopping as we indicate, the resulting $\hat{F}_m$ obey the oracle inequality of Lemma 11 with $\Delta_{m,n} = 0$.

   Suppose that $X \in \mathbb{R}$ and F_∞ ranges over a ball in an approximation space such as Sobolev or, more generally, Besov. Then, if $\mathcal{F}^{(m)}$ has the appropriate approximation properties, e.g., wavelets as smooth as the functions in the specified space, it follows from Baraud (2001) that we can use penalties not dependent on the data to pick $\hat{F}_{\hat{m}}$ such that,

$$\begin{aligned}\max_{\hat{F}} E_P\Big(\hat{F}_{\hat{m}}(X) - E_P(Y|X)\Big)^2 &\asymp \min_{\hat{F}}\max\Big\{E_P\big(\hat{F}(X) - E_P(Y|X)\big)^2 : E_P(Y|X) \in \mathcal{F}\Big\}\\ &\asymp n^{-1+\varepsilon}\Omega(n)\end{aligned}$$

   where $\Omega(n)$ is slowly varying and $0 < \varepsilon < 1$. Here $\hat{F}$ ranges over all estimators based only on the data and not on P. The same type of result has been established for more specialized models with $X \in \mathbb{R}^d$ by Baron, Birgé, Massart (1999), and others, see Györfi et al. (2003).

   The resulting minimax risk,

$$\min_{\hat{F}}\max\{E_P\big(\hat{F}(X) - E_P(Y|X)\big)^2 :\ E_P(Y|X) \in \mathcal{F}\}$$

   is always of order $n^{-1+\varepsilon}\Omega(n)$ where $\Omega(n)$ is typically constant and $0 < \varepsilon < 1$.

   What we show in Theorem 8 is that if, rather than optimizing all the way for each m, we stop in a natural fashion and cross validate as we have indicated, then we can achieve the optimal order as well.

2) "Stumps" unfortunately do not satisfy condition R1 with μ Lebesgue measure. Their Gram matrices are too close to being singular. But differences of stumps work.

3) It follows from the results of Yang (1999) that the rate of Theorem 9 for $\alpha = 0$, that is, if $\mathcal{P}_0 = \mathcal{P}$, is best possible for Sobolev balls and the other spaces we have mentioned.

   Tsybakov implicitly defines a class of F_P for which he is able to specify classification minimax rates. Specifically let $X \in [0,1]^d$ and let $b(x_1,\ldots,x_{d-1})$ be a function having continuous partial derivatives up to order ℓ. Let $p_{b,x}(\cdot)$ be the Taylor polynomial or order ℓ obtained from expanding b at x. Then, he defines $\Sigma(l,L)$ to be the class of all such b for which, $|b(y)-p_{b,x}(y)| \le L|y-x|^{\ell}$ for all $x,y \in [0,1]^{d-1}$. Evidently if b has bounded partial derivatives of order $\ell+1$, $b \in \Sigma(\ell,L)$, for some L. Now let

$$\begin{aligned}\mathcal{P}_\ell = \quad & \{P : F_P(x) = x_d - b(x_1,\ldots,x_{d-1}),\\ & P[|F_P(x)| \le t] \le Ct, \text{ for all } 0 \le t \le 1, b \in \Sigma(\ell,L)\}\end{aligned}$$

   Tsybakov following Mammen and Tsybakov (1999) shows that the classification minimax regret for $\mathcal{P}$ (Theorem 2 of Tsybakov (2001) for $K=2$) is $\frac{2\ell}{3\ell+(d-1)}$. On the other hand, if we assume that $Y = F_P(x) + \varepsilon$ where ε is independent of X, bounded and $E(\varepsilon) = 0$, then the L_2 minimax regret rate is $2\ell/(2\ell+(d-1))$ – see Birgé and Massart (1999) Sections 4.1.1 and Theorem 9. Our theorem 9 now yields a classification minimax regret rate of

$$\frac{2}{3}\cdot\frac{2\ell}{2\ell+(d-1)} = \frac{2\ell}{3\ell+\frac{3}{2}(d-1)}$$

   which is slightly worse than what can be achieved using Tsybakov's not as readily computable procedures. However, note that as $\ell \to \infty$ so that F_P and the boundary become arbitrarily smooth, L_2 boosting approaches the best possible rate for $\mathcal{P}_\ell$ of $\frac{2}{3}$. Similar remarks can be made about $0 < \alpha \le 1$.

7. Conclusions

In this paper we presented different mathematical aspects of boosting. We consider the observations as an i.i.d. sample from a population (i.e., a distribution). The boosting algorithm is a Gauss-Southwell minimization of a classification loss function (which typically dominates the 0-1 misclassification loss). We show that the output of the boosting algorithm follows the theoretical path as if it were applied to the true distribution of the population. Since early stopping is possible as argued, the algorithm, supplied with an appropriate stopping rule, is consistent.

However, there are no simple rate results other than those of Bühlmann and Yu (2003), which we discuss, for the convergence of the boosting classifier to the Bayes classifier. We showed that rate results can be obtained when the boosting algorithm is modified to a cautious version, in which at each step the boosting is done only over a small set of permitted directions.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant DMS-0104075), and the Israel Science Foundation (ISF grant 793/03).

Appendix A. Proof of Theorem 1:

Let $w_0 = \inf_{f\in\mathcal{F}_\infty} w(f)$. Let $f^*_k = \sum_m \alpha_{km}h_{km}$, $h_{k,m}\in\mathcal{H}$, $\sum_m|\alpha_{km}|<\infty$, $k=0,1,2,\dots$ be any member of $\mathcal{F}_\infty$ such that (i) $f^*_0 = f_0$; (ii) $w(f^*_k)\searrow w_0$ is strictly decreasing sequence; (iii) The following condition is satisfied:

$$w(f^*_k) \geq \alpha w_0 + (1-\alpha)w(f^*_{k-1}) + (1-\alpha)(\nu_{k-1}-\nu_k), \tag{11}$$

where $\nu_k\searrow 0$ is a strictly decreasing real sequence. The construction of the sequence $\{f^*_k\}$ is possible since, by assumption, $\mathcal{F}_\infty$ is dense in the image of $w(\cdot)$. That is, we can start with the sequence $\{w(f^*_k)\}$, and then look for suitable $\{f^*_k\}$. Here is a possible construction. Let c and η be suitable small number. Let $\gamma=(1-\alpha)(1+2\eta)/(1-\eta)$, $\nu_k = c\eta\gamma^k/(1-\gamma)$. Select now f^*_k such $w_0+c(1-\eta)\gamma^k \leq w(f^*_k)\leq w_0+c(1+\eta)\gamma^k$. ($\eta$ should be small enough such that $\gamma<1$ and c should selected such that $w(f^*_1)<w(f_0)$.) Our argument rests on the following,

Lemma 12 *There is a sequence $m_k\to\infty$ such that $w(f_m)\leq w(f^*_k)+\nu_k$ for $m\geq m_k$, $k=1,2,\dots$, and $m_k\leq\zeta_k(m_{k-1})<\infty$, where $\zeta_k(\cdot)$ is a monotone non-decreasing functions which depends only on the sequences $\{\nu_k\}$ and $\{f^*_k\}$.*

Proof of Lemma 12:

We will use the following notation. For $f\in\mathcal{F}_\infty$ let $\|f\|_* = \inf\{\sum|\gamma_i|,\ f=\sum\gamma_i h_i,\ h_i\in\mathcal{H}\}$.

Recall that by definition $w(f_0)=w(f^*_0)$. Our argument proceeds as follows, We will inductively define m_k satisfying the conclusion of the lemma, and make, if $\varepsilon_{k,m}\equiv w(f_m)-w(f^*_k)$,

$$\varepsilon_{k,m}\leq c_{k,m}\equiv\max\Big\{\nu_k,\ \frac{\sqrt{512}B}{\alpha^2\beta_k}\frac{w(f^*_{k-1})-w_0}{\Big(\log\Big(1+\frac{8(w(f^*_{k-1})-w_0)}{\alpha\beta_k(\tau_k+\rho_k m_{k-1})}(m-m_{k-1}+1)\Big)\Big)^{1/2}}\Big\}, \tag{12}$$

where

$$\begin{aligned}\beta_k &= \ \inf\{w''(f;h): w_0+\nu_k\leq w(f)\leq w(f_0),\ h\in\mathcal{H}\}\\ B &= \ \sup\{w''(f;h): w(f)\leq w(f_o),\ h\in\mathcal{H}\}<\infty.\end{aligned} \tag{13}$$

and

$$\begin{aligned}\tau_k &= 2\|f_0-f^*_k\|^2_*\\ \rho_k &= \frac{16}{\alpha\beta_k}\big(w(f_0)-w_0\big).\end{aligned} \tag{14}$$

Having defined m_k we establish (12) as part of our induction hypothesis for $m_{k-1}<m\leq m_k$. We begin by choosing $m=m_1=1$ so that (12) holds for $m=M-1=1$. We do do this by choosing $\nu_0>0$, sufficiently small. Having established the induction for $m\leq m_{k-1}$ we define m_k as follows. Write now the RHS of (12) as $g(m_{k-1})$, where

$$g(\nu)\equiv\max\Big\{\nu_k,\ \frac{\sqrt{512}B}{\alpha^2\beta_k}\frac{w(f^*_{k-1})-w_0}{\Big(\log\Big(1+\frac{8(w(f^*_{k-1})-w_0)}{\alpha\beta_k(\tau_k+\rho_k\nu}(m-\nu+1)\Big)\Big)^{1/2}}\Big\},$$

We can now pick $\zeta_k(\nu) \equiv \max\big\{\nu+1, \min\{m:\ g(\nu) \le \nu_k\}\big\}$, and define $m_k = \zeta_k(\nu_{k-1})$.

Note that $\{\beta_k\}$, $\{\tau_k\}$, $\{\rho_k\}$, and B depend only the sequences $\{f_k^*\}$ and $\{\nu_k\}$. We now proceed to establish (12). for $m_{k-1} < m \le m_k$. Note first that since $\varepsilon_{k,m}$ as a function of m is non-increasing, (12) holds trivially for $m' > m$ if $\varepsilon_{k,m} \le 0$. By induction (12) holds for $m \le m_{k-1}$, and my hold for some $m > mk-1$. Recall that the definition of the algorithm relates the actual gain at the mth to the maximal gain achieved in this step given the previous steps, see its definition (1). Suppose

$$\inf_{\lambda} w(f_m + \lambda h_m) \le w_0 + \nu_k. \tag{15}$$

Then

$$\begin{aligned} w(f_{m+1}) &\le \alpha \inf_{\lambda} w(f_m + \lambda h_m) + (1-\alpha) w(f_m), \quad \text{by (1)} \\ &\le \alpha(w_0 + \nu_k) + (1-\alpha) w(f_m), \quad \text{by (15)} \\ &\le \alpha(w_0 + \nu_k) + (1-\alpha)\big(w(f_{k-1}^*) + \nu_{k-1}\big), \quad \text{by the outer induction, since } m \ge m_{k-1} \\ &\le \alpha(w_0 + \nu_k) + \big(w(f_k^*) - \alpha w_0 + (1-\alpha)\nu_k\big), \quad \text{by (11)} \\ &= w(f_k^*) + \nu_k, \end{aligned}$$

so that $\varepsilon_{k,m+1} \le \nu_k$. Therefore, m_k' is not larger than $m+1$, that is $\varepsilon_{k,m'} \le \nu_k$ for $m' > m$ then (12) holds trivially for $m' > m$, and hence, by the second induction assumption for all m. We have established (12) save for m such that,

$$\inf_{\lambda} w(f_m + \lambda h_m) > w_0 + \nu_k \text{ and } \varepsilon_{k,m} \ge 0. \tag{16}$$

We now deal with this case.

Note first that by convexity,

$$|w'(f_m; f_m - f_k^*)| \ge w(f_m) - w(f_k^*) \equiv \varepsilon_{k,m}. \tag{17}$$

We obtain from (17) and the linearity of the derivative that, if $f_m - f_k^* = \sum \gamma_i \widetilde{h}_i \in \mathcal{F}_\infty$,

$$\varepsilon_{k,m} \le \left| \sum -\gamma_i w'(f_m; \tilde{h}_i) \right| \le \sup_{h \in \mathcal{H}} |w'(f_m; h)| \sum |\gamma_i| \ .$$

Hence

$$\sup_{h \in \mathcal{H}} |w'(f_m; h)| \ge \frac{\varepsilon_{k,m}}{\| f_m - f_k^* \|_*}. \tag{18}$$

Now, if $f_{m+1} = f_m + \lambda_m h_m$ then,

$$w(f_m + \lambda_m h_m) = w(f_m) + \lambda_m w'(f_m; h_m) + \frac{1}{2}\lambda_m^2 w''(\tilde{f}_m; h_m), \quad \lambda \in [0, \lambda_m]. \tag{19}$$

where $\tilde{f}_m = f_m + \tilde{\lambda}_m h_m$ and $0 \le \widetilde{\lambda}_m \le \lambda_m$. By convexity, for $0 \le \lambda \le \lambda_m$,

$$w(f_m + \lambda h_m) = w(f_m(1 - \frac{\lambda}{\lambda_m}) + \frac{\lambda}{\lambda_m} f_{m+1}) \le \max\{w(f_m), w(f_{m+1})\} = w(f_m) \le w(f_1).$$

We obtain from Assumption GS1 that $w''(\tilde{f}_m;h) \in (\beta_k, B)$ given in (13). But then we conclude from (19) that,

$$\begin{aligned} w(f_m+\lambda_m h_m) &\geq w(f_m) + \inf_{\lambda\in\mathbb{R}}\bigl(\lambda w'(f_m;h_m) + \frac{1}{2}\lambda^2\beta_k\bigr) \\ &= w(f_m) - \frac{|w'(f_m;h_m)|^2}{2\beta_k} \quad . \end{aligned} \tag{20}$$

Note that $w(f_m+\lambda h) = w(f_m) + \lambda w'(f_m,h) + \lambda^2 w''(f_m+\lambda' h, h)/2$ for some $\lambda' \in [0,\lambda]$, and if $w(f_m+\lambda h)$ is close to $\inf_{\lambda,h} w(f_m+\lambda,h)$ then by convexity, $w(f_m+\lambda' h) \leq w(f_m) \leq w(f_0)$. We obtain from the upper bound on w'' we obtain:

$$\begin{aligned} w(f_m+\lambda_m h_m) &\leq \alpha \inf_{\lambda\in\mathbb{R},\, h\in\mathcal{H}} w(f_m+\lambda h) + (1-\alpha)w(f_m), \qquad \text{by definition,} \\ &\leq \alpha \inf_{\lambda\in\mathbb{R},\, h\in\mathcal{H}} (w(f_m) + \lambda w'(f_m;h) + \frac{1}{2}\lambda^2 B) + (1-\alpha)w(f_m) \\ &= w(f_m) - \frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|^2}{2B}, \end{aligned} \tag{21}$$

by minimizing over λ. Hence combining (20) and (21) we obtain,

$$|w'(f_m;h_m)| \geq \alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)| \sqrt{\frac{\beta_k}{B}} \tag{22}$$

By (21) for the LHS and convexity for the RHS:

$$\frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|^2}{2B} \leq w(f_m) - w(f_{m+1}) \leq -\lambda_m w'(f_m;h_m)$$

Hence

$$|\lambda_m| \geq \frac{\alpha \sup_{h\in\mathcal{H}} |w'(f_m;h)|}{2B}.$$

Applying (18) we obtain:

$$|\lambda_m| \geq \frac{\alpha}{2B}\frac{\varepsilon_{k,m}}{l_{k,m}}, \tag{23}$$

where $l_{k.m} \equiv \| f_m - f_k^* \|_*$.

Let λ_m^0 be the minimal point of $w(f_m+\lambda h_m)$. Taylor expansion around that point and using the lower bound on the curvature:

$$w(f_m+\lambda h_m) \geq w(f_m+\lambda_m^0 h_m) + \frac{1}{2}\beta_k(\lambda-\lambda_m^0)^2 \tag{24}$$

Hence

$$\begin{aligned} {\lambda_m^0}^2 &\leq \frac{2}{\beta_k}\bigl(w(f_m) - w(f_m+\lambda_m^0 h_m)\bigr) \\ &\leq \frac{2}{\alpha\beta_k}\bigl(w(f_m) - w(f_{m+1})\bigr), \end{aligned} \tag{25}$$

where the RHS follows (1). Similarly

$$\begin{aligned}(\lambda_m-\lambda_m^0)^2 &\le \frac{2}{\beta_k}\big(w(f_{m+1})-w(f_m+\lambda_m^0 h_m)\big)\\ &\le \frac{2(1-\alpha)}{\alpha\beta_k}\big(w(f_m)-w(f_{m+1})\big)\end{aligned} \tag{26}$$

Combining (25) and (26):

$$\lambda_m^2 \le \frac{8}{\alpha\beta_k}\big(w(f_m)-w(f_{m+1})\big). \tag{27}$$

Since $\varepsilon_{k,m} \ge 0$ by assumption (16), we conclude from (27) that,

$$\sum_{i=m_{k-1}}^{m} \lambda_i^2 \le \frac{8}{\alpha\beta_k}(w(f_{k-1}^*)-w_0). \tag{28}$$

However, by definition,

$$\begin{aligned} l_{k,m+1} &\le l_{k,m}+|\lambda_m|\\ &\le l_k+\sum_{i=m_{k-1}}^{m}|\lambda_i|\\ &\le l_k+(m+1-m_{k-1})^{1/2}\Big(\sum_{i=m_{k-1}}^{m}\lambda_i^2\Big)^{1/2}\end{aligned} \tag{29}$$

by Cauchy-Schwarz, where, similarly,

$$\begin{aligned} l_k = l_{k,m_{k-1}} &= \|f_{m_{k-1}}-f_k^*\|_*\\ &\le \|f_0-f_k^*\|_*+\|f_{m_{k-1}}-f_0\|_*\\ &\le \|f_0-f_k^*\|_*+\sum_{m=0}^{m_{k-1}-1}|\lambda_m|\\ &\le \|f_0-f_k^*\|_*+m_{k-1}^{1/2}\sqrt{\sum_{m=0}^{m_{k-1}-1}\lambda_m^2}\\ &\le \|f_0-f_k^*\|_*+\sqrt{\frac{8m_{k-1}}{\alpha\beta_k}}\sqrt{w(f_0)-w(f_{m_{k-1}})}, \quad \text{by (27)}\\ &\le \|f_0-f_k^*\|_*+\sqrt{\frac{8m_{k-1}}{\alpha\beta_k}}\sqrt{w(f_0)-w_0}\\ &\le \sqrt{\tau_k+\rho_k m_{k-1}}, \quad \text{as defined in (14).}\end{aligned} \tag{30}$$

Together, (23), (28), and (29) yield:

$$\begin{aligned}\frac{8}{\alpha\beta_k}(w(f^*_{k-1}) - w_0) &\geq \sum_{i=m_{k-1}}^{m} \lambda_i^2 \\ &\geq \frac{\alpha^2}{4B^2}\sum_{i=m_{k-1}}^{m} \frac{\varepsilon^2_{k,i}}{l^2_{k,i}} \\ &\geq \frac{\alpha^2}{4B^2}\sum_{i=m_{k-1}}^{m} \frac{\varepsilon^2_{k,i}}{(l_k + (8(w(f^*_{k-1}) - w_0)/\alpha\beta_k)^{1/2}(i - m_{k-1})^{1/2})^2}\end{aligned} \tag{31}$$

Further, since $\varepsilon_{k,m}$ are decreasing by construction and positive by assumption (16), we can simplify the sum on the RHS of (31):

$$\begin{aligned}&\sum_{i=m_{k-1}}^{m} \frac{\varepsilon^2_{k,i}}{(l_k + (8(w(f^*_{k-1}) - w_0)/\alpha\beta_k)^{1/2}(i - m_{k-1})^{1/2})^2} \\ &\qquad\qquad \geq \frac{\varepsilon^2_{k,m}}{2}\sum_{i=0}^{m-m_{k-1}} \frac{1}{l_k^2 + 8i(w(f^*_{k-1}) - w_0)/\alpha\beta_k}.\end{aligned} \tag{32}$$

Using the inequality,

$$\sum_{i=0}^{m-m_{k-1}} \frac{1}{a+bi} \geq \int_0^{m-m_{k-1}+1} \frac{1}{a+bt}\,dt = \frac{1}{b}\log\Big(1 + \frac{b}{a}(m - m_{k-1} + 1)\Big)$$

on the RHS of (32), we obtain from (31) and (32) that (12) holds, for the case (16). This establishes (16) for all k and m.

■

Proof of Theorem 1: Since the lemma established the existence of monotone ζ_k's, it followed from the definition of these function that $w(f_m) \leq w(f^*_{k(m)})$ where $k(m) = \sup\{k : \zeta^{(k)}(f^*_0) \leq m\}$ and $\zeta^{(k)} = \zeta_k \circ \cdots \circ \zeta_1$ is the kth iterate of the ζs. Since $\zeta^{(k)}(f^*_0) < \infty$ for all k, we have established the uniform rate of convergence and can define the sequence $\{c_m\}$, where $c_m = w(f^*_{k(m)}) - w_0$.

We now prove the uniform step improvement claim of the theorem and identify a suitable function $\xi(\cdot)$. From (26) and (23) if $\varepsilon_{k,m} \geq 0$

$$w(f_m) - w(f_{m+1}) \geq \frac{\alpha\beta_k}{2}\lambda_m^2 \geq \frac{\alpha\beta_k}{2}\left(\frac{\alpha}{2B}\frac{\varepsilon_{k,m}}{l_{k,m}}\right)^2, \tag{33}$$

Bound $l_{k,m}$ similarly to (30) by

$$l_{k,m} \leq l_{k,1} + m^{1/2}\Big(\sum_{i=1}^{m}\lambda_i^2\Big)^2 \leq l_{k,1} + \sqrt{\frac{8m}{\alpha\beta_k}(w(f_0) - w_0)}. \tag{34}$$

Let $m^*(v) = \inf\{m' : c_{m'} \leq v - w_0\}$, which is well defined since $c_m \to 0$. Thus, any realization of the algorithm will cross the v line on or before step number $m^*(v)$. In particular, $m \leq m^*\big(w(f_m)\big)$ for any m and any realization of the algorithm. We obtain therefore by plugging-in (34) in (33), using the m^* as a bound on m and the identity $(a+b)^2 \leq 2a^2+2b^2$ that:

$$w(f_m)-w(f_{m+1}) \geq \frac{\alpha^3\beta_k}{16B^2}\frac{w(f_m)-w(f_k^*)}{l_{k,1}^2+8m^*\big(w(f_m)\big)\big(w(f_0)-w_o\big)/\alpha\beta_k},$$

as long as $\varepsilon_{k,m} \geq 0$. Taking the maximum of the RHS over the permitted range, yields a candidate for the ξ function:

$$\xi(w) \equiv \sup_{k:\, w(f_k^*)\leq w}\left\{\frac{\alpha^3\beta_k}{16B^2}\frac{w-w(f_k^*)}{l_{k,1}^2+m^*\big(w\big)\big(w(f_0)-w_o\big)/\alpha\beta_k}\right\}.$$

This proves the theorem under GS1. Under GS2, the only inequality which we need to replace is (20) since now $\beta_k=0$ is possible. However the definition of Algorithm 2 ensures that we have a coefficient of at least γ on λ^2 in (20). The theorem is proved.

■

Appendix B. Proof of Lemmas 10 and 11 and Theorem 8

Proof of Lemma 10 Since by (R2)

$$\begin{aligned}
\lambda_{\max}(G_m(P)) &= \sup_{\|x\|=1} x'G_m(P)x \\
&= \sup_{\|x\|=1}\sum\sum x_i x_j\int f_{m,i}f_{m,d}dP \\
&= \sup_{\|x\|=1}\int(\sum x_i f_{m,i})^2 dP \qquad (35)\\
&\leq \varepsilon^{-1}\sup_{\|x\|=1}\int(\sum x_i f_{m,i})^2 d\mu = \varepsilon^{-1} \\
\lambda_{\max}(G_m(P)) &\geq \varepsilon, \quad \text{similarly.}
\end{aligned}$$

Part a) follows.

For any symmetric matrix M define its operator norm $\|\cdot\|_T$ by $\lambda_{\max}(M)$. For simplicity let $G_m = G_m(P)$ and $\hat{G}_m = G_m(P_n)$. Recall that for any symmetric matrices A and and M:

$$\begin{aligned}
|\lambda_{\max}(A)-\lambda_{\max}(M)| &\leq \|A-M\|_T \\
|\lambda_{\min}(A)-\lambda_{\min}(M)| &\leq \|A-M\|_T.
\end{aligned}$$

Now,

$$\begin{aligned}
&P\left[\left|\frac{\lambda_{\max}(\hat{G}_m)}{\lambda_{\min}(\hat{G}_m)}-\frac{\lambda_{\max}(G_m)}{\lambda_{\min}(G_m)}\right|\geq t\right) \qquad (36)\\
&\quad\leq P\Big(\|\hat{G}_m-G_m\|_T>\frac{\varepsilon}{2}\Big)+P\Big(\|\hat{G}_m-G_m\|_T\geq t/(\frac{1}{\varepsilon}+\frac{2}{\varepsilon^3})\Big)
\end{aligned}$$

Recall that for a banded matrix M of with band of width $2L$,

$$\begin{aligned}
\|M\|_T^2 &= \sup_{\|x\|=1} \|Mx\|^2 \\
&= \sup_{\|x\|=1} \sum_a \big(\sum_b M_{ab} x_b\big)^2 \\
&\le \sup_{\|x\|=1} \sum_a \sum_{|b-a|<L} x_b^2 M_\infty^2 \\
&\le 2LM_\infty^2 \sup_{\|x\|=1} \sum_a x_a^2 = 2LM_\infty^2,
\end{aligned}$$

where $\|M\|_\infty \equiv \max_{a,b} |M_{ab}|$. Since both $\hat{G}_m$ and $G_m(P)$ are banded of width d, say,

$$\|\hat{G}_m - G_m\|_T \le 2L \max\Big\{ \Big| \frac{1}{n} \sum_{i=1}^{n} \big(f_{m,a} f_{m,b})(X_i) - E_P f_{m,a} f_{m,b}(X_i) \big) \Big| : \ |a-b| < L \Big\} . \tag{37}$$

If $\mathcal{H}$ is a VC class, we can conclude from (35)–(37) that,

$$P[\gamma(\hat{G}_m) \ge C_1] \le C_2 \exp\{-C_3 n / L^2 D_m\} \tag{38}$$

since by R1 (i), $\|f_m\|_\infty \le C_\infty D_m^{\frac{1}{2}}$. The constants ε, C_1, C_2 and C_3 depend on the constants of the R conditions only. This is a consequence of Theorem 2.14.16 p. 246 of van der Vaart and Wellner (1996). This complete the proof of part b).

By a standard result for the Gauss-Southwell method, Luenberger (1984), page 229:

$$\|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2 \le \Big(1 - \frac{1}{\hat{\gamma}_m D_m}\Big) \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \tag{39}$$

Hence

$$\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2 \ge \frac{1}{\hat{\gamma}_m D_m} \|\hat{F}_{m,k} - \hat{F}_m\|_n^2$$

Thus, if

$$\frac{1}{n} \ge \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k+1} - \hat{F}_m\|_n^2$$

we obtain

$$\|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \le D_m \hat{\gamma}_m / n. \tag{40}$$

¿From (40) part (c) follows. ■

Note: Since

$$\|\hat{F}_{m,k-1} - \hat{F}_m\|_n^2 - \|\hat{F}_{m,k} - \hat{F}_m\|_n^2 \quad \ge \quad \frac{C}{n}$$

(39) implies that

$$\Big(1 - \frac{1}{\hat{\gamma}_m D_m}\Big)^{\hat{k}(m)} \quad \ge \quad \frac{1}{n}.$$

Therefore:

$$\hat{k}(m) \le \log n \hat{\gamma}_m D_m \ .$$

If, for instance, as with wavelets $D_m = 2^m, m \le \log_2 n$ we take at most $Cn \log n$ steps total.

Lemma 13 *:*
If $E_{\mathbf{x}}$ denotes conditional expectation give n $X_1,\ldots,X_n$, under R1 *and $F \equiv F_p$,*

$$E_{\mathbf{x}}\|\hat{F}_m - F_m\|_n^2 \le C(\frac{D_m}{n} + \|F_m - F\|_P^2) \tag{41}$$

This is a standard type of result – see Barron, Birgé, Massart (1999). We include the proof for completeness. Note that,

$$\|\hat{F}_m(X) - Y\|_n^2 = \frac{1}{n}\mathbf{Y}^T(I-P)\mathbf{Y}$$

where $\mathbf{Y} \equiv (Y_1,\ldots,Y_n)^T$ and P is the projection matrix of dimension D_m onto the L space spanned by $\big(h_j(X_1),\ldots,h_j(X_n)\big)$, $1 \le j \le D_m$. Then, $(I-P)v = 0$ for all $v \in L$. Hence,

$$E_{\mathbf{X}}\|\hat{F}_m(X) - Y\|_n^2 = \frac{1}{n}E_{\mathbf{X}}\big(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\big)^T(I-P)\big(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\big)$$

where $\mathbf{F}_m(\mathbf{X}) = \big(F_m(X_1),\ldots,F_m(X_n)\big)^T$ is the projection of $(F(X_1),\ldots,F(X_n))^T$ onto L. Note also that,

$$\|\hat{F}_m - F_m\|_n^2 = \|\mathbf{Y} - \mathbf{F}_m(\mathbf{X})\|_n^2 - \|\mathbf{Y} - \hat{\mathbf{F}}_m(\mathbf{X})\|_n^2$$

where $\hat{\mathbf{F}}(X) = (\hat{F}_m(X_1),\ldots,\hat{F}_m(X_n))^T$. Hence,

$$\begin{aligned}
E_{\mathbf{X}}\|\hat{F}_m - F_m\|_n^2 &= \tfrac{1}{n}E_{\mathbf{X}}(\mathbf{Y} - \mathbf{F}_m(X))^T P(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})) \\
&= \tfrac{1}{n}E_{\mathbf{X}}(\mathbf{Y} - \mathbf{F}(\mathbf{X}))^T P(\mathbf{Y} - \mathbf{F}(\mathbf{X})) + \tfrac{2}{n}E_{\mathbf{X}}(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{Y} - \mathbf{F}_m(\mathbf{X})) \\
&= \tfrac{1}{n}E_{\mathbf{X}}\,\mathrm{trace}[P(\mathbf{Y} - \mathbf{F}(\mathbf{X}))(\mathbf{Y} - \mathbf{F}(\mathbf{X}))] \\
&\quad + \tfrac{2}{n}E_{\mathbf{X}}(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{F}_m - \mathbf{F})(\mathbf{X})
\end{aligned}$$

But

$$E_{\mathbf{X}}\ \mathrm{trace}[P(\mathbf{Y} - \mathbf{F}(\mathbf{X})))(\mathbf{Y} - \mathbf{F}(\mathbf{X}))^T] = \frac{1}{n}\sum_{i=1}^n Var(Y_i|X_i)p_{ii}(X) \le \max_i Var(Y_i|X_i)\frac{D_m}{n}$$

since

$$\sum_{i=1}^n p_{ii}(X) = \mathrm{trace}\,P = D_m$$

Also, since P is a projection matrix

$$(\mathbf{F}_m - \mathbf{F})^T P(\mathbf{F}_m - \mathbf{F})(\mathbf{X}) \le \|F_m - F\|_n^2$$

and (41) *follows.*

Proof of Lemma 11:
Take $\Delta_{m,n} = 0$. Let $\tilde{\rho}_m = \sup\left\{\frac{\|t(X)\|_P}{\|t(X)\|_n}\ :\ t \in \mathcal{F}_m\right\}$. By Proposition 5.2 of Baraud (2001), if $\rho_0 > h_0^{-1}$,

$$P[\tilde{\rho}_m > \rho_0] \le D_m^2 \exp\{-\frac{(h_0 - \rho_0^{-1})^2}{4h_1}\ c_n \log n\}$$

where $c_n = \frac{n}{CD_m \log n}$. Here $h_0, h_1 C$ are generic constants. Baraud gives a proof for the case $Var(Y|X) =$ constant, but this is immaterial since only functions of $\underset{\sim}{X}$ are involved in $\tilde{\rho_m}$. Therefore,

$$\begin{aligned} & E_P(\hat{F}_m - F_P)^2 1(\rho_m \leq \rho_0) \\ \leq \quad & 2\rho_0^2 E_P\{E_n(\hat{F}_m - F_m)^2 + E_n(F_m - F_P)^2\} \\ \leq \quad & C(\frac{D_m}{n} + \|F_m - F_P\|^2) \end{aligned} \tag{42}$$

On the other hand,

$$\begin{aligned} E_P(\hat{F}_m - F_P)^2 1(\rho_m > \rho_0) &\leq 2P[\rho_m > \rho_0] \\ &= CD_m^2 \exp\{-AC_n \log n\} \end{aligned} \tag{43}$$

Combining (42) and (43) we obtain Lemma 11 for $\Delta_{m,n} = 0$, $\hat{F}_m = \widetilde{F}_m$. Putting in $\widetilde{F}_m$ we add a term $CE_P(\hat{F}_m - \widetilde{F}_m)^2$. We now apply Lemma 10 c) and the argument we used to obtain (42) and (43). ■

Proof of Theorem 8: Note that we are limited to rates of convergence which are slower than $n^{-\frac{1}{2}}$. This comes from the combination of R1(i) and bounding the operator by the l_∞ norm of the Gram matrix. It is not clear how either of these conditions can be relaxed.

We need only check that if the $\{\widetilde{F}_m\}$ are the θ_m of Theorem 6 then the conditions of that theorem are satisfied. By construction, $\|\widetilde{F}_m\|_\infty \leq 1$, $B_n = \frac{n}{\log n}$. By Lemma 11 and (R3),

$$r_n \leq C_1 \frac{D_m}{n} + C_2 D_m^{-B} \tag{44}$$

and the right hand side of (44) is bounded by $n^{-(\frac{\beta}{\beta+1})}$. ■

References

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *Ann. Stat.* 10:1100–1120, 1982.

Y. Baraud. Model selection for regression on a random design.*Tech. Report*, U. Paris Sud, 2001.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection under penalization. *Prob. Theory and Related Fields*, 113:301–413, 1999.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Tech. Report* 638, Department of Statistics, University of California at Berkeley, 2003.

P. J. Bickel and P. W. Millar. Uniform convergence of probability measures on classes of functions. Statistica Sinica 2:1-15, 1992.

P. J. Bickel and Y. Ritov. The golden chain. A comment. *Ann. Statist.*, 32:91–96, 2003.

L. Breiman. Arcing classifiers (with discussion). *Ann. Statist.* 26:801–849, 1998.

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

L. Breiman. Some infinity theory for predictor ensembles *Technical Report* U.C. Berkeley, 2000.

P. Bühlmann. Consistency for L_2 boosting and matching pursuit with trees and tree type base functions. *Technical Report* ETH Zürich, 2002.

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

D. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia (with discussion). *J. Roy. Statist. Soc. Ser. B* 57:371–394, 1995.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.* 28:337–407, 2000.

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

Y, Freund and R. E. Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proc. 13th International Conference*, 148–156. Morgan Kauffman, San Francisco, 1996.

G. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer, New York, 2002.

W. Jiang. Process consistency for ADABOOST. Technical Report 00-05, Dept. of Statistics, Northwestern University, 2002.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data . *J. of Amer. Statist. Assoc.*, 99:67–81, 2002.

D. G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading, 1984.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of boosting methods. *Ann. Statist.* 32:30–55, 2004.

S. Mallat and Z. Zhang. Matching pursuit with time frequency dictionaries. *IEEE Transactions on Signal Processing* 41:3397–3415, 1993.

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

S. Mannor, R. Meir, and T. Zhang. Greedy algorithms for classification—consistency, convergence rates and adaptivity. J. of Machine Learning Research 4:713–742, 2004.

L. Mason, P. Bartlett, J. Baxter, and M. Frean. Functional gradient techniques for combining hypotheses. In Schölkopg, Smola, A., Bartlett, P., and Schurmans, D. (edts.) *Advances in Large Margin Classifiers*, MIT Press, Boston, 2000.

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

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence related predictions. *Machine Learning*, 37:297–336, 1999.

A, Tsybakov. Optimal aggregation of classifiers in statistical learning. *Technical Report*, U. of Paris IV, 2001.

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

Y. Yang. Minimax nonparametric classification – Part I Rates of convergence, Part II Model selection, *IEEE Trans. Inf. Theory* 45:2271–2292, 1999.

T. Zhang and B. Yu. Boosting with early stopping: convergence and consistency. Tech Report 635, Stat Dept, UCB, 2003.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32:56–134, 2004.