Journal of Machine Learning Research ? (?) 1–20 Submitted 12/2006; Published ?/?

AdaBoost is Consistent

Peter L. Bartlett BARTLETT@STAT.BERKELEY.EDU
Department of Statistics and Computer Science Division
University of California
Berkeley, CA 94720-3860, USA

Mikhail Traskin MTRASKIN@STAT.BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720-3860, USA

Editor: ?

Abstract

The risk, or probability of error, of the classifier produced by the AdaBoost algorithm is investigated. In particular, we consider the stopping strategy to be used in AdaBoost to achieve universal consistency. We show that provided AdaBoost is stopped after $n^{1-\varepsilon}$ iterations—for sample size n and $\varepsilon \in (0,1)$—the sequence of risks of the classifiers it produces approaches the Bayes risk.

Keywords: boosting, adaboost, consistency

1. Introduction

Boosting algorithms are an important recent development in classification. These algorithms belong to a group of voting methods (see, for example, Schapire, 1990; Freund, 1995; Freund and Schapire, 1996, 1997; Breiman, 1996, 1998), that produce a classifier as a linear combination of *base* or *weak* classifiers. While empirical studies show that boosting is one of the best off the shelf classification algorithms (see Breiman, 1998) theoretical results do not give a complete explanation of their effectiveness.

The first formulations of boosting by Schapire (1990); Freund (1995); Freund and Schapire (1996, 1997) considered boosting as an iterative algorithm that is run for a fixed number of iterations and at every iteration it chooses one of the base classifiers, assigns a weight to it and eventually outputs the classifier that is the weighted majority vote of the chosen classifiers. Later Breiman (1997, 1998, 2000) pointed out that boosting is a gradient descent type algorithm (see also Friedman et al., 2000; Mason et al., 2000).

Experimental results by Drucker and Cortes (1996); Quinlan (1996); Breiman (1998); Bauer and Kohavi (1999); Dietterich (2000) showed that boosting is a very effective method, that often leads to a low test error. It was also noted that boosting continues to decrease test error long after the sample error becomes zero: though it keeps adding more weak classifiers to the linear combination of classifiers, the generalization error, perhaps surprisingly, usually does not increase. However some of the experiments suggested that there might be problems, since boosting performed worse than bagging in the presence of noise (Dietterich, 2000), and boosting concentrated not only on the "hard" areas, but also on outliers and noise (Bauer and Kohavi, 1999). And indeed, some more experiments, for example by Friedman et al. (2000); Grove and Schuurmans (1998); Mason et al. (2000), see also Bickel et al. (2006), as well as some theoretical results (for example Jiang, 2002) showed that boosting, ran for an arbitrary large number of steps, overfits, though it takes a very long time to do it.

Upper bounds on the risk of boosted classifiers were obtained, based on the fact that boosting tends to maximize the margin of the training examples (Schapire et al., 1998; Koltchinskii and

Panchenko, 2002), but Breiman (1999) pointed out that margin-based bounds do not completely explain the success of boosting methods. In particular, these results do not resolve the issue of consistency: they do not explain under which conditions we may expect the risk to converge to the Bayes risk. A recent work by Reyzin and Schapire (2006) shows that while maximization of the margin is useful, it should not be done at the expense of the classifier complexity.

Breiman (2000) showed that under some assumptions on the underlying distribution "population boosting" converges to the Bayes risk as the number of iterations goes to infinity. Since the population version assumes infinite sample size, this does not imply a similar result for AdaBoost, especially given results of Jiang (2002), that there are examples when AdaBoost has prediction error asymptotically suboptimal at $t = \infty$ (t is the number of iterations).

Several authors have shown that *modified* versions of AdaBoost are consistent. These modifications include restricting the l_1-norm of the combined classifier (Mannor et al., 2003; Blanchard et al., 2003; Lugosi and Vayatis, 2004; Zhang, 2004) , and restricting the step size of the algorithm (Zhang and Yu, 2005). Jiang (2004) analyses the unmodified boosting algorithm and proves a process consistency property, under certain assumptions. Process consistency means that there exists a sequence (t_n) such that if AdaBoost with sample size n is stopped after t_n iterations, its risk approaches the Bayes risk. However Jiang also imposes strong conditions on the underlying distribution: the distribution of X (the predictor) has to be absolutely continuous with respect to Lebesgue measure and the function $F_B(X) = (1/2)\ln(\mathbf{P}(Y = 1|X)/\mathbf{P}(Y = -1|X))$ has to be continuous on $\mathcal{X}$. Also Jiang's proof is not constructive and does not give any hint on when the algorithm should be stopped. Bickel et al. (2006) prove a consistency result for AdaBoost, under the assumption that the probability distribution is such that the steps taken by the algorithm are not too large. In this paper, we study stopping rules that guarantee consistency. In particular, we are interested in AdaBoost, not a modified version. Our main result (Corollary 9) demonstrates that a simple stopping rule suffices for consistency: the number of iterations is a fixed function of the sample size. We assume only that the class of base classifiers has finite VC-dimension, and that the span of this class is sufficiently rich. Both assumptions are clearly necessary.

2. Notation

Here we describe the AdaBoost procedure formulated as a coordinate descent algorithm and introduce definitions and notation. We consider a binary classification problem. We are given $\mathcal{X}$, the measurable (feature) space, and $\mathcal{Y} = \{-1, 1\}$, the set of (binary) labels. We are given a sample $S_n = \{(X_i, Y_i)\}_{i=1}^n$ of i.i.d. observations distributed as the random variable $(X, Y) \sim \mathcal{P}$, where $\mathcal{P}$ is an unknown distribution. Our goal is to construct a classifier $g_n : \mathcal{X} \to \mathcal{Y}$ based on this sample. The quality of the classifier g_n is given by the misclassification probability

$$L(g_n) = \mathbf{P}(g_n(X) \neq Y|S_n).$$

Of course we want this probability to be as small as possible and close to the Bayes risk

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

where the infimum is taken over all possible (measurable) classifiers and $\eta(\cdot)$ is a conditional probability

$$\eta(x) = \mathbf{P}(Y = 1|X = x).$$

The infimum above is achieved by the Bayes classifier $g^\star(x) = g(2\eta(x) - 1)$, where

$$g(x) = \begin{cases} 1 & , \quad x > 0, \\ -1 & , \quad x \leq 0. \end{cases}$$

We are going to produce a classifier as a linear combination of *base* classifiers in $\mathcal{H} = \{h | h : \mathcal{X} \to \mathcal{Y}\}$. We shall assume that class $\mathcal{H}$ has a finite VC (Vapnik-Chervonenkis) dimension $d_{VC}(\mathcal{H}) = \max\left\{|S| : S \subseteq \mathcal{X},\, \left|\mathcal{H}_{|S}\right| = 2^{|S|}\right\}$.

AdaBoost works to find a combination f that minimizes the convex criterion

$$\frac{1}{n}\sum_{i=1}^{n} \exp(-Y_i f(X_i)).$$

Many of our results are applicable to a broader family of such algorithms, where the function $\alpha \mapsto \exp(-\alpha)$ is replaced by another function φ. Thus, for a function $\varphi : \mathbb{R} \to \mathbb{R}^+$, we define the empirical φ-risk and the φ-risk,

$$R_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i f(X_i)) \qquad \text{and} \qquad R_\varphi(f) = \mathrm{E}\varphi(Y f(X)).$$

Clearly, the function φ needs to be appropriate for classification, in the sense that a measurable f that minimizes $R_\varphi(f)$ should have minimal risk. This is equivalent (see Bartlett et al., 2006) to φ satisfying the following condition (‘classification calibration’). For all $0 \le \eta \le 1$, $\eta \ne 1/2$,

$$\inf\{\eta\varphi(\alpha) + (1-\eta)\varphi(-\alpha) : \alpha(2\eta - 1) \le 0\} > \inf\{\eta\varphi(\alpha) + (1-\eta)\varphi(-\alpha) : \alpha \in \mathbb{R}\}. \tag{1}$$

We shall assume that φ satisfies (1).

Then the boosting procedure can be described as follows.

1. Set $f_0 \equiv 0$. Choose number of iterations t.

2. For $k = 1, \dots, t$, set
$$f_k = f_{k-1} + \alpha_{k-1} h_{k-1},$$
where the following holds for some fixed $\gamma \in (0, 1]$ independent of k.
$$R_{\varphi,n}(f_k) \le \gamma \inf_{h \in \mathcal{H}, \alpha \in \mathbb{R}} R_{\varphi,n}(f_{k-1} + \alpha h) + (1-\gamma) R_{\varphi,n}(f_{k-1}). \tag{2}$$
We call α_i the step size of the algorithm at step i.

3. Output $g \circ f_t$ as the final classifier.

The choice of $\gamma < 1$ in the above algorithm allows approximate minimization. Notice that the original formulation of AdaBoost assumed exact minimization in (2), which corresponds to $\gamma = 1$.

We shall also use the convex hull of $\mathcal{H}$ scaled by $\lambda \ge 0$,

$$\mathcal{F}_\lambda = \left\{ f \,\middle|\, f = \sum_{i=1}^{n} \lambda_i h_i, n \in \mathbb{N} \cup \{0\}, \lambda_i \ge 0, \sum_{i=1}^{n} \lambda_i = \lambda, h_i \in \mathcal{H} \right\}$$

as well as the set of k-combinations, $k \in \mathbb{N}$, of functions in $\mathcal{H}$

$$\mathcal{F}^k = \left\{ f \,\middle|\, f = \sum_{i=1}^{k} \lambda_i h_i, \lambda_i \in \mathbb{R}, h_i \in \mathcal{H} \right\}.$$

We also need to define the $l_\star$-norm: for any $f \in \mathcal{F}$

$$\|f\|_\star = \inf\left\{\sum |\alpha_i|, f = \sum \alpha_i h_i, h_i \in \mathcal{H}\right\}.$$

Define the squashing function $\pi_l(\cdot)$ to be

$$\pi_l(x) = \begin{cases} l & , \quad x > l, \\ x & , \quad x \in [-l, l], \\ -l & , \quad x < -l. \end{cases}$$

Then the set of truncated functions is

$$\pi_l \circ \mathcal{F} = \left\{ \tilde{f} | \tilde{f} = \pi_l(f), f \in \mathcal{F} \right\}.$$

The set of classifiers based on a class $\mathcal{F}$ is denoted by

$$g \circ \mathcal{F} = \{ \tilde{f} | \tilde{f} = g(f), f \in \mathcal{F} \}.$$

Define the derivative of an arbitrary function $Q(\cdot)$ in the direction of h as

$$Q'(f; h) = \left. \frac{\partial Q(f + \lambda h)}{\partial \lambda} \right|_{\lambda = 0}.$$

The second derivative $Q''(f; h)$ is defined similarly.

3. Consistency of Boosting Procedure

In this section, we present the proof of the consistency of AdaBoost. We begin with an overview.

The usual approach to proving consistency involves a few key steps (see, for example, Bartlett et al., 2004). The first is a comparison theorem, which shows that as the φ-risk $R_\varphi(f_n)$ approaches $R^\star_\varphi$ (the infimum over measurable functions of R_φ), $L(f_n)$ approaches $L^\star$. The classification calibration condition (1) suffices for this (Bartlett et al., 2006). The second step is to show that the class of functions is suitably rich so that there is some sequence of elements $\bar{f}_n$ for which $\lim_{n\to\infty} R_\varphi(\bar{f}_n) = R^\star_\varphi$. The third step is to show that the φ-risk of the estimate f_n approaches that of the reference sequence $\bar{f}_n$. For instance, for a method of sieves that minimizes the empirical φ-risk over a suitable set $\mathcal{F}_n$ (which increases with the sample size n), one could define the reference sequence $\bar{f}_n$ as the minimizer of the φ-risk in $\mathcal{F}_n$. Then, provided that the sets $\mathcal{F}_n$ grow suitably slowly with n, the maximal deviation over $\mathcal{F}_n$ between empirical φ-risk and φ-risk would converge to zero. Such a uniform convergence result would imply that the sequence f_n has φ-risk converging to $R^\star_\varphi$.

The key difficulty with this approach is that the concentration inequalities behind the uniform convergence results are valid only for a suitably small class of suitably bounded functions. However boosting in general and AdaBoost in particular may produce functions that cannot be appropriately bounded. To circumvent this difficulty, we rely on the observation that, for the purposes of classification, we can replace the function f returned by AdaBoost by any function f' that satisfies $\mathrm{sign}(f') = \mathrm{sign}(f)$. Therefore we consider the clipped version $\pi_\lambda \circ f_t$ of the function returned by AdaBoost after t iterations. This clipping ensures that the functions f_t are suitably bounded. Furthermore, the complexity of the clipped class (as measured by its pseudo-dimension—see Pollard, 1984) grows slowly with the stopping time t, so we can show that the φ-risk of a clipped function is not much larger than its empirical φ-risk. Lemma 4 provides the necessary details. In order to compare the empirical φ-risk of the clipped function to that of a suitable reference sequence $\bar{f}_n$, we first use the fact that the empirical φ-risk of a clipped function $\pi_\lambda \circ f_t$ is not much larger than the empirical φ-risk of f_t.

The next step is to relate $R_{\varphi,n}(f_t)$ to $R_{\varphi,n}(\bar{f}_n)$. The choice of a suitable sieve depends on what can be shown about the progress of the algorithm. We consider an increasing sequence of $l_\star$-balls, and define $\bar{f}_n$ as the (near) minimizer of the φ-risk in the appropriate $l_\star$-ball. Theorems 6 and 8 show that as the stopping time increases, the empirical φ-risk of the function returned by AdaBoost is not much larger than that of $\bar{f}_n$. Finally Hoeffding's inequality shows that the empirical φ-risks of the reference functions $\bar{f}_n$ are close to their φ-risks. Combining all the pieces, the φ-risk of $\pi_\lambda \circ f_t$ approaches $R_\varphi^\star$, provided the stopping time increases suitably slowly with the sample size. The consistency of AdaBoost follows.

We now describe our assumptions. First, we shall impose the following condition.

Condition 1 ***Denseness.*** *Let the distribution $\mathcal{P}$ and class $\mathcal{H}$ be such that*

$$\lim_{\lambda\to\infty} \inf_{f\in\mathcal{F}_\lambda} R_\varphi(f) = R_\varphi^\star,$$

where $R_\varphi^\star = \inf R_\varphi(f)$ over all measurable functions.

For many classes $\mathcal{H}$, the above condition is satisfied for all possible distributions $\mathcal{P}$. Lugosi and Vayatis (2004, Lemma 1) discuss sufficient conditions for Condition 1. As an example of such a class, we can take the class of indicators of all rectangles or the class of indicators of half-spaces defined by hyperplanes or the class of binary trees with the number of terminal nodes equal to $d+1$ (we consider trees with terminal nodes formed by successive univariate splits), where d is the dimensionality of $\mathcal{X}$ (see Breiman, 2000).

The following set of conditions deals with uniform convergence and convergence of the boosting algorithm. The main theorem (Theorem 1) shows that these, together with Condition 1, suffice for consistency of the boosting procedure. Later in this section we show that the conditions are satisfied by AdaBoost.

Condition 2 *Let n be sample size. Let there exist non-negative sequences $t_n \to \infty$, $\zeta_n \to \infty$ and a sequence $\{\bar{f}_n\}_{n=1}^\infty$ of reference functions such that*

$$R_\varphi(\bar{f}_n) \underset{n\to\infty}{\to} R^\star,$$

and suppose that the following conditions are satisfied.

a. ***Uniform convergence of t_n-combinations.***

$$\sup_{f\in\pi_{\zeta_n}\circ\mathcal{F}^{t_n}} |R_\varphi(f) - R_{\varphi,n}(f)| \overset{a.s.}{\underset{n\to\infty}{\to}} 0. \tag{3}$$

b. ***Convergence of empirical φ-risks for the sequence $\{\bar{f}_n\}_{n=1}^\infty$.***

$$\max\left\{0, R_{\varphi,n}(\bar{f}_n) - R_\varphi(\bar{f}_n)\right\} \overset{a.s.}{\underset{n\to\infty}{\to}} 0. \tag{4}$$

c. ***Algorithmic convergence of t_n-combinations.***

$$\max\left\{0, R_{\varphi,n}(f_{t_n}) - R_{\varphi,n}(\bar{f}_n)\right\} \overset{a.s.}{\underset{n\to\infty}{\to}} 0. \tag{5}$$

Now we state the main theorem.

Theorem 1 *Assume φ is classification calibrated and convex. Assume, without loss of generality, that for $\varphi_\lambda = \inf_{x\in[-\lambda,\lambda]}\varphi(x)$,*

$$\lim_{\lambda\to\infty}\varphi_\lambda = \inf_{x\in(-\infty,\infty)}\varphi(x) = 0. \tag{6}$$

Let Condition 2 be satisfied. Then the boosting procedure stopped at step t_n returns a sequence of classifiers f_{t_n} almost surely satisfying $L(g(f_{t_n})) \to L^\star$ as $n \to \infty$.

Remark 2 *Note that Condition (6) could be replaced by the mild condition that the function φ is bounded below.*

Proof For almost every outcome ω on the probability space $(\Omega, \mathcal{S}, \mathbf{P})$ we can define sequences $\epsilon_n^1(\omega) \to 0$, $\epsilon_n^2(\omega) \to 0$ and $\epsilon_n^3(\omega) \to 0$, such that for almost all ω the following inequalities are true.

$$\begin{aligned}
R_\varphi(\pi_{\zeta_n}(f_{t_n})) &\leq R_{\varphi,n}(\pi_{\zeta_n}(f_{t_n})) + \epsilon_n^1(\omega) \quad \text{by (3)} \\
&\leq R_{\varphi,n}(f_{t_n}) + \epsilon_n^1(\omega) + \varphi_{\zeta_n} && (7) \\
&\leq R_{\varphi,n}(\bar{f}_n) + \epsilon_n^1(\omega) + \varphi_{\zeta_n} + \epsilon_n^2(\omega) \quad \text{by (5)} \\
&\leq R_\varphi(\bar{f}_n) + \epsilon_n^1(\omega) + \varphi_{\zeta_n} + \epsilon_n^2(\omega) + \epsilon_n^3(\omega) \quad \text{by (4).} && (8)
\end{aligned}$$

Inequality (7) follows from the convexity of $\varphi(\cdot)$ (see Lemma 14 in Appendix E). By (6) and choice of the sequence $\{\bar{f}_n\}_{n=1}^\infty$ we have $R_\varphi(\bar{f}_n) \to R^\star$ and $\varphi_{\zeta_n} \to 0$. And from (8) follows $R_\varphi(\pi_{\zeta_n}(f_{t_n})) \to R^\star$ a.s. Eventually we can use the result by Bartlett et al. (2006, Theorem 3) to conclude that

$$L(g(\pi_{\zeta_n}(f_{t_n}))) \overset{a.s.}{\to} L^\star.$$

But for $\zeta_n > 0$ we have $g(\pi_{\zeta_n}(f_{t_n})) = g(f_{t_n})$, therefore

$$L(g(f_{t_n})) \overset{a.s.}{\to} L^\star.$$

Hence, the boosting procedure is consistent if stopped after t_n steps. ■

The almost sure formulation of Condition 2 does not provide explicit rates of convergence of $L(g(f_{t_n}))$ to $L^\star$. However, a slightly stricter form of Condition 2, which allows these rates to be calculated, is considered in Appendix A.

In the following sections, we show that Condition 2 can be satisfied for some choices of φ. We shall treat parts (a)–(c) separately.

3.1 Uniform Convergence of t_n-Combinations

Here we show that Condition 2 (a) is satisfied for a variety of functions φ, and in particular for exponential loss used in AdaBoost. We begin with a simple lemma (see Freund and Schapire, 1997, Theorem 8 or Anthony and Bartlett, 1999, Theorem 6.1):

Lemma 3 *For any $t \in \mathbb{N}$ if $d_{VC}(\mathcal{H}) \geq 2$ the following holds:*

$$d_P(\mathcal{F}^t) \leq 2(t+1)(d_{VC}(\mathcal{H})+1)\log_2[2(t+1)/\ln 2],$$

where $d_P(\mathcal{F}^t)$ is the pseudo-dimension of class $\mathcal{F}^t$.

The proof of consistency is based on the following result, which builds on the result by Koltchinskii and Panchenko (2002) and resembles a lemma due to Lugosi and Vayatis (2004, Lemma 2).

Lemma 4 *For a continuous function φ define the Lipschitz constant*

$$L_{\varphi,\zeta} = \inf\{L | L > 0, |\varphi(x) - \varphi(y)| \leq L|x-y|, -\zeta \leq x, y \leq \zeta\}$$

and maximum absolute value of $\varphi(\cdot)$ when argument is in $[-\zeta, \zeta]$

$$M_{\varphi,\zeta} = \max_{x \in [-\zeta,\zeta]} |\varphi(x)|.$$

Then for $V = d_{VC}(\mathcal{H})$, $c = 24\int_0^1 \sqrt{\ln \frac{8e}{\epsilon^2}}d\epsilon$ *and any* n, $\zeta > 0$ *and* $t > 0$,

$$\mathrm{E} \sup_{f \in \pi_\zeta \circ \mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| \leq c\zeta L_{\varphi,\zeta}\sqrt{\frac{(V+1)(t+1)\log_2[2(t+1)/\ln 2]}{n}}. \tag{9}$$

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

$$\begin{aligned}\sup_{f \in \pi_\zeta \circ \mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| \quad &\leq \quad c\zeta L_{\varphi,\zeta}\sqrt{\frac{(V+1)(t+1)\log_2[2(t+1)/\ln 2]}{n}} \\ &+ \quad M_{\varphi,\zeta}\sqrt{\frac{\ln(1/\delta)}{2n}}. \end{aligned} \tag{10}$$

Proof The proof is given in Appendix B. ■

Now, if we choose ζ and δ as functions of n, such that $\sum_{n=1}^\infty \delta^2(n) < \infty$ and right hand side of (10) converges to 0 as $n \to \infty$, we can appeal to Borel-Cantelli lemma and conclude, that for such choice of ζ_n and δ_n Condition 2 (a) holds.

Lemma 4, unlike Lemma 2 of Lugosi and Vayatis (2004), allows us to choose the number of steps t, which describes the complexity of the linear combination of base functions, and this is essential for the proof of the consistency. It is easy to see that for AdaBoost (i.e. $\varphi(x) = e^{-x}$) we can choose $\zeta = \kappa \ln n$ and $t = n^{1-\varepsilon}$ with $\kappa > 0$, $\varepsilon \in (0,1)$ and $2\kappa - \varepsilon < 0$.

3.2 Convergence of Empirical φ-Risks for the Sequence $\{\bar{f}_n\}_{n=1}^\infty$.

To show that Condition 2(b) is satisfied for a variety of loss functions we use Hoeffding's inequality.

Theorem 5 *Define the variation of a function* φ *on the interval* $[-a,a]$ *(for* $a > 0$*) as*

$$V_{\varphi,a} = \sup_{x\in[-a,a]} \varphi(x) - \inf_{x\in[-a,a]} \varphi(x).$$

If a sequence $\{\bar{f}_n\}_{n=1}^\infty$ *satisfies the condition* $\bar{f}_n(x) \in [-\lambda_n, \lambda_n], \forall x \in \mathcal{X}$, *where* $\lambda_n > 0$ *is chosen so that* $V_{\varphi,\lambda_n} = o(\sqrt{n/\ln n})$, *then*

$$\max\{0, R_{\varphi,n}(\bar{f}_n) - R_\varphi(\bar{f}_n)\} \underset{n\to\infty}{\overset{a.s.}{\to}} 0. \tag{11}$$

Proof Since we restricted the range of $\bar{f}_n$ to the interval $[-\lambda_n, \lambda_n]$, we have, almost surely, $|\varphi(Y\bar{f}_n(X))| \leq V_{\varphi,\lambda_n}$. Therefore Hoeffding's inequality guarantees that for all ϵ_n

$$\mathbf{P}\left(R_{\varphi,n}(\bar{f}_n) - R_\varphi(\bar{f}_n) \geq \epsilon_n\right) \leq \exp\left(-2n\epsilon_n^2 V_{\varphi,\lambda_n}^2\right) = \delta_n.$$

To prove the statement of the theorem we require $\epsilon_n = o(1)$ and $\sum \delta_n < \infty$. Then we appeal to the Borel-Cantelli lemma to conclude that (11) holds. These restrictions are satisfied if

$$V_{\varphi,\lambda_n}^2 = o\left(\frac{n}{\ln n}\right)$$

and the statement of the theorem follows. ■

The choice of λ_n in the above theorem depends on the loss function φ. In the case of the AdaBoost loss $\varphi(x) = e^{-x}$ we shall choose $\lambda_n = \kappa \ln n$, where $\kappa \in (0, 1/2)$. One way to guarantee that the functions $\bar{f}_n$ satisfy condition $\bar{f}_n(x) \in [-\lambda_n, \lambda_n], \forall x \in \mathcal{X}$, is to choose $\bar{f}_n \in \mathcal{F}_{\lambda_n}$.

3.3 Algorithmic Convergence of AdaBoost

So far we dealt with the statistical properties of the function we are minimizing; now we turn to the algorithmic part. Here we show that Condition 2(c) is satisfied for the AdaBoost algorithm. We need the following simple consequence of the proof of Bickel et al. (2006, Theorem 1).

Theorem 6 *Let the function $Q(f)$ be convex in f and twice differentiable in all directions $h \in \mathcal{H}$. Let $Q^\star = \lim_{\lambda\to\infty} \inf_{f\in\mathcal{F}_\lambda} Q(f)$. Assume that $\forall c_1, c_2$, such that $Q^\star < c_1 < c_2 < \infty$,*

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

Also assume the following approximate minimization scheme for $\gamma \in (0,1]$. Define $f_{k+1} = f_k + \alpha_{k+1}h_{k+1}$ such that

$$Q(f_{k+1}) \le \gamma \inf_{h\in\mathcal{H},\alpha\in\mathbb{R}} Q(f_k + \alpha h) + (1-\gamma)Q(f_k)$$

and

$$Q(f_{k+1}) = \inf_{\alpha\in\mathbb{R}} Q(f_k + \alpha h_{k+1}).$$

Then for any reference function $\bar{f}$ and the sequence of functions f_m, produced by the boosting algorithm, the following bound holds $\forall m > 0$ such that $Q(f_m) > Q(\bar{f})$.

$$Q(f_m) \le Q(\bar{f}) + \sqrt{\frac{8B^3(Q(f_0) - Q(\bar{f}))^2}{\gamma^2\beta^3}} \left(\ln \frac{\ell_0^2 + c_3 m}{\ell_0^2}\right)^{-\frac{1}{2}}, \tag{12}$$

where $\ell_k = \left\|\bar{f} - f_k\right\|_\star$, $c_3 = 2(Q(f_0) - Q(\bar{f}))/\beta$, $\beta = \inf\{Q''(f;h) : Q(\bar{f}) < Q(f) < Q(f_0), h \in \mathcal{H}\}$, $B = \sup\{Q''(f;h) : Q(f) < Q(f_0), h \in \mathcal{H}\}$.

Proof The statement of the theorem is a version of a result implicit in the proof of (Bickel et al., 2006, Theorem 1). The proof is given in Appendix C. ■

Remark 7 *Results in (Zhang and Yu, 2005, e.g., Lemma 4.1) provide similar bounds under either an assumption of a bounded step size of the boosting algorithm or a positive lower bound on $Q''(f;h)$ for all f, h. Since we consider boosting algorithms with unrestricted step size, the only option would be to assume a positive lower bound on the second derivative. While such an assumption is fine for the quadratic loss $\varphi(x) = x^2$, second derivative $R''_n(f;h)$ of the empirical risk for the exponential loss used by the AdaBoost algorithm can not be bounded from below by a positive constant in a general case. Theorem 6 makes a mild assumption that second derivative is positive for all f such that $R(f) > R^\star$ ($R_n(f) > R^\star_n$) .*

It is easy to see, that the theorem above applies to the AdaBoost algorithm, since there we first choose the direction (base classifier) h_i and then we compute the step size α_i as

$$\alpha_i = \frac{1}{2}\ln\frac{1-\epsilon_i}{\epsilon_i} = \frac{1}{2}\ln\frac{R(f_i) - R'(f_i;h_i)}{R(f_i) + R'(f_i;h_i)}.$$

Now we only have to recall that this value of α_i corresponds to exact minimization in the direction h_i.

From now on we are going to specialize to AdaBoost and use $\varphi(x) = e^{-x}$. Hence we drop the subscript φ in $R_{\varphi,n}$ and R_φ and use R_n and R respectively.

Theorem 6 allows us to get an upper bound on the difference between the *exp*-risk of the function output by AdaBoost and the *exp*-risk of the appropriate reference function. For brevity in the next theorem we make an assumption $R^\star > 0$, though a similar result can be stated for $R^\star = 0$. For completeness, the corresponding theorem is given in Appendix D.

Theorem 8 *Assume $R^\star > 0$. Let t_n be the number of steps we run AdaBoost. Let $\lambda_n = \kappa \ln n$, $\kappa \in (0, 1/2)$. Let $a > 1$ be an arbitrary fixed number. Let $\{\bar f_n\}_{n=1}^\infty$ be a sequence of functions such that $\bar f_n \in \mathcal{F}_{\lambda_n}$. Then with probability at least $1-\delta_n$, where $\delta_n = \exp\left(-2(R^\star)^2 n^{1-2\kappa}/a^2\right)$, the following holds*

$$R_n(f_{t_n}) \le R_n(\bar f_n) + \frac{2\sqrt{2}(1-R^\star(a-1)/a)}{\gamma\left(\frac{a-1}{a}R^\star\right)^{3/2}}\left(\ln\frac{\lambda_n^2 + 2t_n(a/(a-1)-R^\star)/R^\star}{\lambda_n^2}\right)^{-1/2}.$$

Proof This theorem follows directly from Theorem 6. Because in AdaBoost

$$R_n''(f;h) = \frac{1}{n}\sum_{i=1}^n (-Y_i h(X_i))^2 e^{-Y_i f(X_i)} = \frac{1}{n}\sum_{i=1}^n e^{-Y_i f(X_i)} = R_n(f),$$

then all the conditions in Theorem 6 are satisfied as long as $R_n(\bar f_n) > 0$ (with $Q(f)$ replaced by $R_n(f)$) and in the Equation (12) we have $B = R_n(f_0) = 1$, $\beta \ge R_n(\bar f_n)$, $\left\|f_0 - \bar f_n\right\|_\star \le \lambda_n$. Since for t such that $R_n(f_t) \le R_n(\bar f_n)$ the theorem is trivially true, we only have to notice that $\exp(Y_i \bar f_n(X_i)) \in [0, n^\kappa]$, hence Hoeffding's inequality guarantees that

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n e^{Y_i \bar f_n(X_i)} - \mathrm{E} e^{Y \bar f_n(X)} \le -\frac{R^\star}{a}\right) \le \exp\left(-2(R^\star)^2 n^{1-2\kappa}/a^2\right) = \delta_n,$$

where we choose and fix the constant $a > 1$ arbitrarily and independently of n and the sequence $\{\bar f_n\}_{n=1}^\infty$. Therefore with probability at least $1-\delta_n$ we bound empirical risk from below as $R_n(\bar f_n) \ge R(\bar f_n) - R^\star/a \ge R^\star - R^\star/a = R^\star(a-1)/a$, since $R(\bar f_n) \ge R^\star$. Therefore $\beta \ge R^\star(a-1)/a$ and the result follows immediately from Equation (12) if we use the fact that $R^\star > 0$. ∎

It is easy to see that choice of λ_n's in the above theorem ensures that $\sum_{n=1}^\infty \delta_n < \infty$, therefore Borel-Cantelli lemma guarantees that for $t_n \to \infty$ sufficiently fast (for example as $O(n^\alpha)$ for $\alpha \in (0,1)$)

$$\max\{0, R_n(f_{t_n}) - R_n(\bar f_n)\} \underset{n\to\infty}{\overset{a.s.}{\to}} 0.$$

If in addition to the conditions of Theorem 8 we shall require that

$$R(\bar f_n) \le \inf_{f \in \mathcal{F}_{\lambda_n}} R(f) + \epsilon_n,$$

for some $\epsilon_n \to 0$, then together with Condition 1 this will imply $R(\bar f_n) \to R^*$ as $n \to \infty$ and Condition 2 (c) follows.

3.4 Consistency of AdaBoost

Having all the ingredients at hand, consistency of AdaBoost is a simple corollary of Theorem 1.

Corollary 9 *Assume $V = d_{VC}(\mathcal{H}) < \infty$,*

$$\lim_{\lambda\to\infty}\inf_{f\in\mathcal{F}_\lambda} R(f) = R^\star$$

and $t_n = n^{1-\varepsilon}$ for $\varepsilon \in (0,1)$. Then AdaBoost stopped at step t_n returns a sequence of classifiers almost surely satisfying $L(g(f_{t_n})) \to L^\star$.

Proof First assume $L^\star > 0$. For the exponential loss function this implies $R^\star > 0$. As was suggested after the proof of Lemma 4 we may choose $\zeta_n = \kappa \ln n$ for $2\kappa - \epsilon < 0$ (which also implies $\kappa < 1/2$)

to satisfy Condition 2 (a). Recall that discussion after the proof of the Theorem 8 suggests choice of the sequence $\{\bar{f}_n\}_{n=1}^{\infty}$ of reference functions such that $\bar{f}_n \in \mathcal{F}_{\lambda_n}$ and

$$R(\bar{f}_n) \leq \inf_{f \in \mathcal{F}_{\lambda_n}} R(f) + \epsilon_n$$

for $\epsilon_n \to 0$ and $\lambda_n = \kappa \ln n$ with $\kappa \in (0, 1/2)$ to ensure that Condition 2 (c) holds. Eventually, as it follows from the discussion after the proof of the Theorem 5, choice of the sequence $\{\bar{f}_n\}_{n=1}^{\infty}$ to satisfy Condition 2(c) also ensures that Condition 2(b) holds. Since function $\varphi(x) = e^{-x}$ is clearly classification calibrated and conditions of this Corollary assume Condition 1 then all the conditions of Theorem 1 hold and consistency of the AdaBoost algorithm follows.

For $L^\star = 0$ the proof is similar, but we need to use Theorem 13 in Appendix D instead of Theorem 8. ■

4. Discussion

We showed that AdaBoost is consistent if stopped sufficiently early, after t_n iterations, for $t_n = O(n^{1-\varepsilon})$ with $\varepsilon \in (0, 1)$. We do not know whether this number can be increased. Results by Jiang (2002) imply that for some $\mathcal{X}$ and function class $\mathcal{H}$ the AdaBoost algorithm will achieve zero training error after t_n steps, where $n^2/t_n = o(1)$ (see also work by Mannor and Meir (2001, Lemma 1) for an example of $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H} = \{\text{linear classifiers}\}$, for which perfect separation on the training sample is guaranteed after $8n^2 \ln n$ iterations), hence if run for that many iterations, the AdaBoost algorithm does not produce a consistent classifier. We do not know what happens in between $O(n^{1-\varepsilon})$ and $O(n^2 \ln n)$. Lessening this gap is a subject of further research.

The AdaBoost algorithm, as well as other versions of the boosting procedure, replaces the $0-1$ loss with a convex function φ to overcome algorithmic difficulties associated with the non-convex optimization problem. In order to conclude that $R_\varphi(f_n) \to R_\varphi^\star$ implies $L(g(f_n)) \to L^\star$ we want φ to be classification calibrated and this requirement cannot be relaxed, as shown by Bartlett et al. (2006).

The statistical part of the analysis, summarized in Lemma 4 and Theorem 5, works for quite an arbitrary loss function φ. The only restriction imposed by Lemma 4 is that φ must be Lipschitz on any compact set. This requirement is an artifact of our proof and is caused by the use of the "contraction principle". It can be relaxed in some cases: Shen et al. (2003) use the classification calibrated loss function

$$\psi(x) = \begin{cases} 2 & , \quad x < 0, \\ 1 - x & , \quad 0 \leq x < 1, \\ 0 & , \quad x \geq 1, \end{cases}$$

which is non-Lipschitz on any interval $[-\lambda, \lambda]$, $\lambda > 0$.

The algorithmic part, presented by Theorems 6 and 8, concentrated on the analysis of the exponential (AdaBoost) loss $\varphi(x) = e^{-x}$. This approach also works for the quadratic loss $\varphi(x) = (1-x)^2$. Theorem 6 assumes that the second derivative $R''_\varphi(f; h)$ is bounded from below by a positive constant, possibly dependent on the value of $R_\varphi(f)$, as long as $R_\varphi(f) > R_\varphi^\star$. This condition is clearly satisfied for $\varphi(x) = (1-x)^2$: $R''_\varphi(f; h) \equiv 2$ and we do not need an analog of Theorem 8; Theorem 6 suffices. Lemma 4 can be applied for the quadratic loss with $L_{\varphi,\lambda} = 2(1+\lambda)$ and $M_{\varphi,\lambda} = (1+\lambda)^2$. We may choose t_n, λ_n, ζ_n the same as for the exponential loss or set $\lambda_n = n^{1/4-\vartheta_1}$, $\vartheta_1 \in (0, 1/4)$, $\zeta_n = n^{\varrho-\vartheta_2}$, $\vartheta_2 = (0, \varrho)$, $\varrho = \min(\varepsilon/2, 1/4)$ to get the following analog of Corollary 9.

Corollary 10 *Assume $\varphi(x) = (1-x)^2$. Assume $V = d_{VC}(\mathcal{H}) < \infty$,*

$$\lim_{\lambda \to \infty} \inf_{f \in \mathcal{F}_\lambda} R(f) = R^\star$$

and $t_n = n^{1-\varepsilon}$ for $\varepsilon \in (0,1)$. Then boosting procedure stopped at step t_n returns a sequence of classifiers almost surely satisfying $L(g(f_{t_n})) \to L^\star$.

We cannot make analogous conclusion about other loss functions. For example for logit loss $\varphi(x) = \ln(1+e^{-x})$, Lemma 4 and Theorem 5 work, since $L_{\varphi,\lambda} = 1$ and $M_{\varphi,\lambda} = \ln(1+e^{\lambda})$, hence choosing t_n, λ_n, ζ_n as for either the exponential or quadratic losses will work. The assumption of the Theorem 6 also holds with $R''_{\varphi,n}(f;h) \geq R_{\varphi,n}(f)/n$, though the resulting inequality is trivial: the factor $1/n$ in this bound precludes us from finding an analog of Theorem 8. A similar problem arises in the case of the modified quadratic loss $\varphi(x) = [\max(1-x,0)]^2$, for which $R''_{\varphi,n}(f;h) \geq 2/n$. Generally, any loss function with "really flat" regions may cause trouble. Another issue is the very slow rate of convergence in Theorems 6 and 8. Hence further research intended either to improve convergence rates or extend the applicability of these theorems to loss functions other than exponential and quadratic is desirable.

Acknowledgements

This work was supported by the NSF under award DMS-0434383. The authors would like to thank Peter Bickel for useful discussions, as well as Jean-Philippe Vert and two anonymous referees for their comments and suggestions.

Appendix A. Rate of Convergence of $L(g(f_{t_n}))$ to $L^\star$

Here we formulate Condition 2 in a stricter form and prove consistency along with a rate of convergence of the boosting procedure to the Bayes risk.

Condition 3 *Let n be sample size. Let there exist non-negative sequences $t_n \to \infty$, $\zeta_n \to \infty$, $\delta_n^j \to 0$ such that $\sum_{i=1}^{\infty} \delta_i^j < \infty$, $j = 1,2,3$, $\epsilon_n^k \to 0$, $k = 1,2,3$, a sequence $\{\bar{f}_n\}_{n=1}^{\infty}$ of reference functions such that*

$$R_\varphi(\bar{f}_n) \underset{n\to\infty}{\to} R^\star,$$

and the following conditions hold.

a. ***Uniform convergence of t_n-combinations.***

$$\mathbf{P}\left(\sup_{f \in \pi_{\zeta_n} \circ \mathcal{F}^{t_n}} |R_\varphi(f) - R_{\varphi,n}(f)| > \epsilon_n^1\right) < \delta_n^1. \tag{13}$$

b. ***Convergence of empirical φ-risks for the sequence $\{\bar{f}_n\}_{n=1}^{\infty}$.***

$$\mathbf{P}\left(R_{\varphi,n}(\bar{f}_n) - R_\varphi(\bar{f}_n) > \epsilon_n^2\right) < \delta_n^2. \tag{14}$$

c. ***Algorithmic convergence of t_n-combinations.***

$$\mathbf{P}\left(R_{\varphi,n}(f_{t_n}) - R_{\varphi,n}(\bar{f}_n) > \epsilon_n^3\right) < \delta_n^3. \tag{15}$$

Now we state the analog of Theorem 1.

Theorem 11 *Assume φ is classification calibrated and convex, and for $\varphi_\lambda = \inf_{x\in[-\lambda,\lambda]} \varphi(x)$ without loss of generality assume*

$$\lim_{\lambda\to\infty} \varphi_\lambda = \inf_{x\in(-\infty,\infty)} \varphi(x) = 0. \tag{16}$$

Let Condition 3 be satisfied. Then the boosting procedure stopped at step t_n returns a sequence of classifiers f_{t_n} almost surely satisfying $L(g(f_{t_n})) \to L^\star$ as $n \to \infty$.

Proof Consider the following sequence of inequalities.

$$\begin{aligned}R_\varphi(\pi_{\zeta_n}(f_{t_n})) &\leq R_{\varphi,n}(\pi_{\zeta_n}(f_{t_n})) + \epsilon_n^1 \quad \text{by (13)} &(17)\\ &\leq R_{\varphi,n}(f_{t_n}) + \epsilon_n^1 + \varphi_{\zeta_n} \\ &\leq R_{\varphi,n}(\bar{f}_n) + \epsilon_n^1 + \varphi_{\zeta_n} + \epsilon_n^3 \quad \text{by (15)} &(18)\\ &\leq R_\varphi(\bar{f}_n) + \epsilon_n^1 + \varphi_{\zeta_n} + \epsilon_n^3 + \epsilon_n^2 \quad \text{by (14).} &(19)\end{aligned}$$

Inequalities (17), (19) and (18) hold with probability at least $1-\delta_n^1$, $1-\delta_n^2$ and $1-\delta_n^3$ respectively. We assumed in Condition 3 that $R_\varphi(\bar{f}_n) \to R^\star$ and (16) implies that $\varphi_{\zeta_n} \to 0$ by the choice of the sequence ζ_n. Now we appeal to the Borel-Cantelli lemma and arrive at $R_\varphi(\pi_{\zeta_n}(f_{t_n})) \to R^\star$ a.s. Eventually we can use Theorem 3 by Bartlett et al. (2006) to conclude that

$$L(g(\pi_{\zeta_n}(f_{t_n}))) \xrightarrow{a.s.} L^\star.$$

But for $\zeta_n > 0$ we have $g(\pi_{\zeta_n}(f_{t_n})) = g(f_{t_n})$, therefore

$$L(g(f_{t_n})) \xrightarrow{a.s.} L^\star.$$

Hence the boosting procedure is consistent if stopped after t_n steps. ■

We could prove Theorem 11 by using the Borel-Cantelli lemma and appealing to Theorem 1, but the above proof allows the following corollary on the rate of convergence.

Corollary 12 *Let the conditions of Theorem 11 be satisfied. Then there exists a non-decreasing function ψ, such that $\psi(0) = 0$, and with probability at least $1-\delta_n^1-\delta_n^2-\delta_n^3$*

$$L(g(f_{t_n})) - L^\star \leq \psi^{-1}\left((\epsilon_n^1 + \epsilon_n^2 + \epsilon_n^3 + \varphi_{\zeta_n}) + \left(\inf_{f\in\mathcal{F}_{\lambda_n}} R_\varphi - R_\varphi^\star\right)\right), \tag{20}$$

where ψ^{-1} is the inverse of ψ.

Proof From Theorem 3 of Bartlett et al. (2006), if ϕ is convex we have that

$$\psi(\theta) = \phi(0) - \inf\left\{\frac{1+\theta}{2}\phi(\alpha) + \frac{1-\theta}{2}\phi(-\alpha) : \alpha \in \mathbb{R}\right\},$$

and for any distribution and any measurable function f

$$L(g(f)) - L^\star \leq \psi^{-1}\left(R_\varphi(f) - R_\varphi^\star\right).$$

On the other hand,

$$R_\varphi(f) - R_\varphi^\star = \left(R_\varphi(f) - \inf_{f\in\mathcal{F}_{\lambda_n}} R_\varphi\right) + \left(\inf_{f\in\mathcal{F}_{\lambda_n}} R_\varphi - R_\varphi^\star\right).$$

The proof of Theorem 11 shows that for function f_{t_n} with probability at least $1-\delta_n^1-\delta_n^2$

$$R_\varphi(f_{t_n}) - \inf_{f\in\mathcal{F}_{\lambda_n}} R_\varphi \leq \epsilon_n^1 + \epsilon_n^2 + \epsilon_n^3 + \varphi_{\zeta_n}.$$

Putting all the components together we obtain (20). ■

The second term under ψ^{-1} in (20) is an approximation error and, in a general case, it may decrease arbitrarily slowly. However, if it is known that it decreases sufficiently fast, the first term becomes an issue. For example Corollary 9, even if the approximation error decreases sufficiently fast, will give a convergence rate of the order $O\left((\ln n)^{-\frac{1}{4}}\right)$. This follows from Example 1 by Bartlett et al. (2006), where it is shown that for AdaBoost (exponential loss function) $\psi^{-1}(x) \leq \sqrt{2x}$, and the fact that both ϵ_n^1 and ϵ_n^2, as well as φ_{ζ_n}, in Corollary 9 decrease at the rate $O(n^{1-\alpha})$ (in fact, α's might be different for all three of them), hence everything is dominated by ϵ_n^3, which is $O\left((\ln n)^{-\frac{1}{2}}\right)$.

Appendix B. Proof of Lemma 4

For convenience, we state the lemma once again.
Lemma 4 *For a continuous function φ define the Lipschitz constant*

$$L_{\varphi,\zeta} = \inf\{L | L > 0, |\varphi(x) - \varphi(y)| \leq L|x-y|, -\zeta \leq x, y \leq \zeta\}$$

and maximum absolute value of $\varphi(\cdot)$ when argument is in $[-\zeta, \zeta]$

$$M_{\varphi,\zeta} = \max_{x\in[-\zeta,\zeta]} |\varphi(x)|.$$

Then for $V = d_{VC}(\mathcal{H})$, $c = 24\int_0^1 \sqrt{\ln \frac{8e}{\epsilon^2}} d\epsilon$ and any n, $\zeta > 0$ and $t > 0$,

$$\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| \leq c\zeta L_{\varphi,\zeta} \sqrt{\frac{(V+1)(t+1)\log_2[2(t+1)/\ln 2]}{n}}.$$

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

$$\begin{aligned}\sup_{f\in\pi_\zeta\circ\mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| \quad &\leq \quad c\zeta L_{\varphi,\zeta} \sqrt{\frac{(V+1)(t+1)\log_2[2(t+1)/\ln 2]}{n}} \\ &+ \quad M_{\varphi,\zeta}\sqrt{\frac{\ln(1/\delta)}{2n}}.\end{aligned}$$

Proof The proof of this lemma is similar to the proof of Lugosi and Vayatis (2004, Lemma 2) in that we begin with symmetrization followed by the application of the "contraction principle". We use symmetrization to get

$$\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| \quad \leq \quad 2\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i(\varphi(-Y_i f(X_i)) - \varphi(0)) \right|,$$

where σ_i are i.i.d. with $\mathbf{P}(\sigma_i = 1) = \mathbf{P}(\sigma_i = -1) = 1/2$. Then we use the "contraction principle" (see Ledoux and Talagrand, 1991, Theorem 4.12, pp. 112–113) with a function $\psi(x) = (\varphi(x) - \varphi(0))/L_{\varphi,\zeta}$ to get

$$\begin{aligned}\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| \quad &\leq \quad 4L_{\varphi,\zeta}\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} \left| \frac{1}{n}\sum_{i=1}^n -\sigma_i Y_i f(X_i) \right| \\ &= \quad 4L_{\varphi,\zeta}\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \right|.\end{aligned}$$

Next we proceed and find the supremum. Notice, that functions in $\pi_\zeta \circ \mathcal{F}^t$ are bounded and clipped to absolute value equal ζ, therefore we can rescale $\pi_\zeta \circ \mathcal{F}^t$ by $(2\zeta)^{-1}$ and get

$$\mathrm{E} \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \right| = 2\zeta\mathrm{E} \sup_{f\in(2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \right|.$$

Next, we use Dudley's entropy integral (Dudley, 1999) to bound the right hand side above

$$\mathrm{E} \sup_{f\in(2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \right| \leq \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\ln \mathcal{N}(\epsilon, (2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t, L_2(P_n))} d\epsilon.$$

Since, for $\epsilon > 1$, the covering number $\mathcal{N}$ is 1, the upper integration limit can be taken as 1, and we can use Pollard's bound (Pollard, 1990) for $F \subseteq [0,1]^{\mathcal{X}}$,

$$\mathcal{N}(\epsilon, F, L_2(P)) \leq 2\left(\frac{4e}{\epsilon^2}\right)^{d_P(F)},$$

where $d_P(F)$ is a pseudo-dimension, and obtain for $\tilde{c} = 12\int_0^1 \sqrt{\ln \frac{8e}{\epsilon^2}} d\epsilon$,

$$\mathrm{E} \sup_{f \in (2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right| \leq \tilde{c}\sqrt{\frac{d_P((2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t)}{n}}.$$

Also notice that constant $\tilde{c}$ does not depend on $\mathcal{F}^t$ or ζ. Next,

since $(2\zeta)^{-1}\circ\pi_\zeta$ is non-decreasing, we use the inequality $d_P((2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t) \leq d_P(\mathcal{F}^t)$ (for example, Anthony and Bartlett, 1999, Theorem 11.3) to obtain

$$\mathrm{E} \sup_{f \in (2\zeta)^{-1}\circ\pi_\zeta\circ\mathcal{F}^t} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right| \leq c\sqrt{\frac{d_P(\mathcal{F}^t)}{n}}.$$

And then, since Lemma 3 gives an upper-bound on the pseudo-dimension of the class $\mathcal{F}^t$, we have

$$\mathrm{E} \sup_{f \in \pi_\zeta\circ\mathcal{F}^t} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right| \leq c\zeta\sqrt{\frac{(V+1)(t+1)\log_2[2(t+1)/\ln 2]}{n}},$$

with the constant c above being independent of $\mathcal{H}$, t and ζ. To prove the second statement we use McDiarmid's bounded difference inequality (Devroye et al., 1996, Theorem 9.2, p. 136), since for all $i \in \{1,\ldots,n\}$

$$\sup_{(x_j,y_j)_{j=1}^n,(x_i',y_i')} \left|\sup_{f\in\pi_\zeta\circ\mathcal{F}^t} |R_\varphi(f) - R_{\varphi,n}(f)| - \sup_{f\in\pi_\zeta\circ\mathcal{F}^t} |R_\varphi(f) - R'_{\varphi,n}(f)|\right| \leq \frac{M_{\varphi,\zeta}}{n},$$

where $R'_{\varphi,n}(f)$ is obtained from $R_{\varphi,n}(f)$ by changing each pair (x_i, y_i) to an independent pair (x_i', y_i'). This completes the proof of the lemma. ∎

Appendix C. Proof of Theorem 6

For convenience, we state the theorem once again.

Theorem 6 *Let the function $Q(f)$ be convex in f and twice differentiable in all directions $h \in \mathcal{H}$. Let $Q^\star = \lim_{\lambda\to\infty}\inf_{f\in\mathcal{F}_\lambda} Q(f)$. Assume that $\forall c_1, c_2$, such that $Q^\star < c_1 < c_2 < \infty$,*

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

Also assume the following approximate minimization scheme for $\gamma \in (0,1]$. Define $f_{k+1} = f_k + \alpha_{k+1}h_{k+1}$such that

$$Q(f_{k+1}) \leq \gamma \inf_{h\in\mathcal{H},\alpha\in\mathbb{R}} Q(f_k + \alpha h) + (1-\gamma)Q(f_k)$$

and

$$Q(f_{k+1}) = \inf_{\alpha\in\mathbb{R}} Q(f_k + \alpha h_{k+1}).$$

Then for any reference function $\bar{f}$ and the sequence of functions f_m, produced by the boosting algorithm, the following bound holds $\forall m > 0$ such that $Q(f_m) > Q(\bar{f})$.

$$Q(f_m) \le Q(\bar{f}) + \sqrt{\frac{8B^3(Q(f_0) - Q(\bar{f}))^2}{\gamma^2\beta^3}} \left(\ln \frac{\ell_0^2 + c_3 m}{\ell_0^2}\right)^{-\frac{1}{2}},$$

where $\ell_k = \left\|\bar{f} - f_k\right\|_\star$, $c_3 = 2(Q(f_0) - Q(\bar{f}))/\beta$, $\beta = \inf\{Q''(f;h) : Q(\bar{f}) < Q(f) < Q(f_0), h \in \mathcal{H}\}$, $B = \sup\{Q''(f;h) : Q(f) < Q(f_0), h \in \mathcal{H}\}$.

Proof The statement of the theorem is a version of a result implicit in the proof of Theorem 1 by Bickel et al. (2006). If for some m we have $Q(f_m) \le Q(\bar{f})$, then the theorem is trivially true for all $m' \ge m$. Therefore, we are going to consider only the case when $Q(f_m) > Q(\bar{f})$. We shall also assume $Q(f_{m+1}) \ge Q(\bar{f})$ (the impact of this assumption will be discussed later). Define $\epsilon_m = Q(f_m) - Q(\bar{f})$. By convexity of $Q(\cdot)$,

$$|Q'(f_m; f_m - \bar{f})| \ge \epsilon_m. \tag{21}$$

Let $f_m - \bar{f} = \sum \tilde{\alpha}_i \tilde{h}_i$, where $\tilde{\alpha}_i$ and $\tilde{h}_i$ correspond to the best representation (with the l_1-norm of $\tilde{\alpha}$ equal the $l_\star$-norm). Then from (21) and linearity of the derivative we have

$$\epsilon_m \le \left|\sum \tilde{\alpha}_i Q'(f_m; \tilde{h}_i)\right| \le \sup_{h\in\mathcal{H}} |Q'(f_m;h)| \sum |\tilde{\alpha}_i|,$$

therefore

$$\sup_{h\in\mathcal{H}} Q'(f_m;h) \ge \frac{\epsilon_m}{\left\|f_m - \bar{f}\right\|_\star} = \frac{\epsilon_m}{\ell_m}. \tag{22}$$

Next,

$$Q(f_m + \alpha h_m) = Q(f_m) + \alpha Q'(f_m; h_m) + \frac{1}{2}\alpha^2 Q''(\tilde{f}_m; h_m),$$

where $\tilde{f}_m = f_m + \tilde{\alpha}_m h_m$, for $\tilde{\alpha}_m \in [0, \alpha_m]$. By assumption $\tilde{f}_m$ is on the path from f_m to f_{m+1}, and we have assumed exact minimization in the given direction, hence f_{m+1} is the lowest point in the direction h_m starting from f_m, so we have the following bounds

$$Q(\bar{f}) < Q(f_{m+1}) \le Q(\tilde{f}_m) \le Q(f_m) \le Q(f_0).$$

Then by the definition of β, which depends on $Q(\bar{f})$, we have

$$Q(f_{m+1}) \ge Q(f_m) + \inf_{\alpha\in\mathbb{R}} (\alpha Q'(f_m; h_m) + \frac{1}{2}\alpha^2\beta) = Q(f_m) - \frac{|Q'(f_m;h_m)|^2}{2\beta}. \tag{23}$$

On the other hand,

$$\begin{aligned} Q(f_m + \alpha_m h_m) &\le \gamma \inf_{h\in\mathcal{H},\alpha\in\mathbb{R}} Q(f_m + \alpha h) + (1-\gamma)Q(f_m) \\ &\le \gamma \inf_{h\in\mathcal{H},\alpha\in\mathbb{R}} \left(Q(f_m) + \alpha Q'(f_m;h) + \frac{1}{2}\alpha^2 B)\right) + (1-\gamma)Q(f_m) \\ &= Q(f_m) - \gamma \frac{\sup_{h\in\mathcal{H}} |Q'(f_m;h)|^2}{2B}. \end{aligned} \tag{24}$$

Therefore, combining (23) and (24), we get

$$|Q'(f_m;h_m)| \ge \sup_{h\in\mathcal{H}} |Q'(f_m;h)| \sqrt{\frac{\gamma\beta}{B}}. \tag{25}$$

Another Taylor expansion, this time around f_{m+1} (and we again use the fact that f_{m+1} is the minimum on the path from f_m), gives us

$$Q(f_m) = Q(f_{m+1}) + \frac{1}{2}\alpha_m^2 Q''(\tilde{\tilde{f}}_m; h_m), \tag{26}$$

where $\tilde{\tilde{f}}_m$ is some (other) function on the path from f_m to f_{m+1}. Therefore, if $|\alpha_m| < \sqrt{\gamma}|Q'(f_m; h_m)|/B$, then

$$Q(f_m) - Q(f_{m+1}) < \frac{\gamma|Q'(f_m; h_m)|^2}{2B},$$

but by (24)

$$Q(f_m) - Q(f_{m+1}) \geq \frac{\gamma \sup_{h\in\mathcal{H}} |Q'(f_m; h)|^2}{2B} \geq \frac{\gamma|Q'(f_m; h_m)|^2}{2B},$$

therefore we conclude, by combining (25) and (22), that

$$|\alpha_m| \geq \frac{\sqrt{\gamma}|Q'(f_m; h_m)|}{B} \geq \frac{\gamma\sqrt{\beta}\sup_{h\in\mathcal{H}} |Q'(f_m; h)|}{B^{3/2}} \geq \frac{\gamma\epsilon_m\sqrt{\beta}}{\ell_m B^{3/2}}. \tag{27}$$

Using (26) we have

$$\sum_{i=0}^{m} \alpha_i^2 \leq \frac{2}{\beta}\sum_{i=0}^{m}(Q(f_i) - Q(f_{i+1})) \leq \frac{2}{\beta}(Q(f_0) - Q(\bar{f})). \tag{28}$$

Recall that

$$\begin{aligned}
\left\|f_m - \bar{f}\right\|_\star &\leq \left\|f_{m-1} - \bar{f}\right\|_\star + |\alpha_{m-1}| \leq \left\|f_0 - \bar{f}\right\|_\star + \sum_{i=0}^{m-1} |\alpha_i| \\
&\leq \left\|f_0 - \bar{f}\right\|_\star + \sqrt{m}\left(\sum_{i=0}^{m-1}\alpha_i^2\right)^{1/2},
\end{aligned}$$

therefore, combining with (28) and (27), since the sequence ϵ_i is decreasing,

$$\begin{aligned}
\frac{2}{\beta}(Q(f_0) - Q(\bar{f})) &\geq \sum_{i=0}^{m}\alpha_i^2 \\
&\geq \frac{\gamma^2\beta}{B^3}\sum_{i=0}^{m}\frac{\epsilon_i^2}{\ell_i^2} \\
&\geq \frac{\gamma^2\beta}{B^3}\epsilon_m^2\sum_{i=0}^{m}\frac{1}{\left(\ell_0 + \sqrt{i}\left(\sum_{j=0}^{i-1}\alpha_j^2\right)^{1/2}\right)^2} \\
&\geq \frac{\gamma^2\beta}{B^3}\epsilon_m^2\sum_{i=0}^{m}\frac{1}{\left(\ell_0 + \sqrt{i}\left(\frac{2(Q(f_0)-Q(\bar{f}))}{\beta}\right)^{1/2}\right)^2} \\
&\geq \frac{\gamma^2\beta}{2B^3}\epsilon_m^2\sum_{i=0}^{m}\frac{1}{\ell_0^2 + \frac{2(Q(f_0)-Q(\bar{f}))}{\beta}i}.
\end{aligned}$$

Since

$$\sum_{i=0}^{m}\frac{1}{a+bi} \geq \int_0^{m+1}\frac{dx}{a+bx} = \frac{1}{b}\ln\frac{a+b(m+1)}{a},$$

then

$$\frac{2}{\beta}(Q(f_0) - Q(\bar{f})) \geq \frac{\gamma^2\beta^2}{4B^3(Q(f_0) - Q(\bar{f}))}\epsilon_m^2 \ln \frac{\ell_0^2 + \frac{2(Q(f_0)-Q(\bar{f}))}{\beta}(m+1)}{\ell_0^2}.$$

Therefore

$$\epsilon_m \leq \sqrt{\frac{8B^3(Q(f_0) - Q(\bar{f}))^2}{\gamma^2\beta^3}} \left(\ln \frac{\ell_0^2 + \frac{2(Q(f_0)-Q(\bar{f}))}{\beta}(m+1)}{\ell_0^2}\right)^{-\frac{1}{2}}. \tag{29}$$

The proof of the above inequality for index m works as long as $Q(f_{m+1}) \geq Q(\bar{f})$. If $\bar{f}$ is such that $Q(f_m) \geq Q(\bar{f})$ for all m, then we do not need to do anything else. However, if there exists m' such that $Q(f_{m'}) < Q(\bar{f})$ and $Q(f_{m'-1}) \geq Q(\bar{f})$, then the above proof is not valid for index $m' - 1$. To overcome this difficulty, we notice that $Q(f_{m'-1})$ is bounded from above by $Q(f_{m'-2})$, therefore to get a bound that holds for all m (except for $m = 0$) we may use a bound for ϵ_{m-1} to bound $Q(f_m) - Q(\bar{f}) = \epsilon_m$: shift (decrease) the index m on the right hand side of (29) by one. This completes the proof of the theorem. ■

Appendix D. Zero Bayes Risk

Here we consider a modification of Theorem 8. In this case our assumptions imply that $R^\star = 0$, and the proof presented above does not work. However for AdaBoost we can modify the proof appropriately to show an adequate convergence rate.

Theorem 13 *Assume $R^\star = 0$. Let t_n be a number of steps we run AdaBoost. Let $\lambda_n = \kappa \ln\ln n$ for $\kappa \in (0, 1/6)$. Let $\epsilon_n = n^{-\nu}$, for $\nu \in (0, 1/2)$. Then with probability at least $1 - \delta_n$, where $\delta_n = \exp\left(-2n^{1-2\nu}/(\ln n)^{2\kappa}\right)$, for some constant C that depends on $\mathcal{H}$ and $\mathcal{P}$ but does not depend on n, for n such that*

$$\frac{C}{(\ln n)^\kappa} > \frac{2}{n^\nu}$$

the following holds

$$\begin{aligned} R_n(f_{t_n}) &\leq R_n(\bar{f}_n) \\ &+ \sqrt{\frac{16R_n^3(f_0)|R_n(f_0) - R_n(\bar{f}_n)|^2(\ln n)^{3\kappa}}{C\gamma^2}} \\ &\times \left(\ln \frac{(\kappa\ln\ln n)^2 + 4|R_n(f_0) - R_n(\bar{f}_n)|(\ln n)^\kappa t_n/C}{(\kappa \ln\ln n)^2}\right)^{-1/2}. \end{aligned}$$

Proof For the exponential loss assumption $R^\star = 0$ is equivalent to $L^\star = 0$. It also implies that the fastest decrease rate of the function $\tau : \lambda \to \inf_{f\in\mathcal{F}_\lambda} R(f)$ is $O(e^{-\lambda})$. To see this, assume that for some λ there exists $f \in \mathcal{F}_\lambda$ such that $L(g(f)) = 0$ (i.e. we have achieved perfect classification). Clearly, for any $a > 0$

$$R(af) = \mathrm{E}e^{-Yaf(X)} = \mathrm{E}\left(e^{-Yf(X)}\right)^a \geq (\inf_{x,y} e^{-yf(x)})^a.$$

Therefore, choose $\lambda_n = \kappa \ln\ln n$. Then $\inf_{f\in\mathcal{F}_{\lambda_n}} R(f) \geq C(\ln n)^{-\kappa}$, where C depends on $\mathcal{H}$ and $\mathcal{P}$, but does not depend on n. On the other hand Hoeffding's inequality for $\bar{f}_n \in \mathcal{F}_{\lambda_n}$ guarantees that

$$\mathbf{P}\left(R(\bar{f}_n) - R_n(\bar{f}_n) \geq \epsilon_n\right) \leq \exp\left(-2n\epsilon_n^2/(\ln n)^{2\kappa}\right) = \delta_n.$$

Choice of $\epsilon_n = n^{-\nu}$ for $\nu \in (0, 1/2)$ ensures that $\delta_n \to 0$. This allows to conclude that with probability at least $1 - \delta_n$ empirical risk $R_n(\bar{f}_n)$ can be lower bounded as

$$R_n(\bar{f}_n) \geq R(\bar{f}_n) - \epsilon_n$$

and for n large enough for

$$\frac{C}{(\ln n)^\kappa} > \frac{2}{n^\nu}$$

to hold we get a lower bound on β in (12) of Theorem 6 as

$$\beta \geq \frac{C}{2(\ln n)^\kappa}.$$

Since for $\bar{f}_n$ such that $R_n(\bar{f}_n) > R_n(f_0)$ theorem trivially holds, we only have to plug $R_n(\bar{f}_n) = 0$, $B = R_n(f_0)$ and $\beta = C(\ln n)^\kappa/2$ into (12) to get the statement of the theorem. Obviously, this bound holds for $R^\star > 0$. ■

Appendix E.

Lemma 14 *Let the function $\varphi : \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$ be convex. Then for any $\lambda > 0$*

$$\varphi(\pi_\lambda(x)) \leq \varphi(x) + \inf_{z \in [-\lambda, \lambda]} \varphi(z). \tag{30}$$

Proof If $x \in [-\lambda, \lambda]$ then the statement of the lemma is clearly true. Without loss of generality assume $x > \lambda$; case $x < -\lambda$ is similar. Then we have two possibilities.

1. $\varphi(x) \geq \varphi(\lambda) = \varphi(\pi_\lambda(x))$ and (30) is obvious.
2. $\varphi(x) < \varphi(\lambda)$. Due to convexity, for any $z < \lambda$ we have $\varphi(z) > \varphi(\lambda)$, therefore

$$\varphi(\pi_\lambda(x)) = \varphi(\lambda) \leq \varphi(\lambda) + \varphi(x) = \inf_{z \in [-\lambda, \lambda]} \varphi(z) + \varphi(x).$$

The statement of the lemma is proven. ■

References

Martin Anthony and Peter L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, 1999.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Discussion of boosting papers. *The Annals of Statistics*, 32(1):85–91, 2004.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36:105–139, 1999.

Peter J. Bickel, Ya'acov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, May 2006.

Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. 4:861–894, 2003.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Arcing the edge. Technical Report 486, Department of Statistics, University of California, Berkeley, 1997.

Leo Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517, 1999. (Was Department of Statistics, U.C. Berkeley Technical Report 504, 1997).

Leo Breiman. Arcing classifiers (with discussion). *The Annals of Statistics*, 26(3):801–849, 1998. (Was Department of Statistics, U.C. Berkeley Technical Report 460, 1996).

Leo Breiman. Some infinite theory for predictor ensembles. Technical Report 579, Department of Statistics, University of California, Berkeley, 2000.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2):139–158, 2000.

Harris Drucker and Corinna Cortes. Boosting decision trees. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 479–485. M.I.T. Press, 1996.

Richard M. Dudley. *Uniform central limit theorems*. Cambridge University Press, Cambridge, MA, 1999.

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

Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *13th International Conference on Machine Learning*, pages 148–156, San Francisco, 1996. Morgan Kaufman.

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28:337–407, 2000.

Adam J. Grove and Dale Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 692–699, Menlo Park, CA, 1998. AAAI Press.

Wenxin Jiang. On weak base hypotheses and their implications for boosting regression and classification. *The Annals of Statistics*, 30:51–73, 2002.

Wenxin Jiang. Process consistency for AdaBoost. *The Annals of Statistics*, 32(1):13–29, 2004.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30:1–50, 2002.

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

Gábor Lugosi and Nicolas Vayatis. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32(1):30–55, 2004.

Shie Mannor and Ron Meir. Weak learners and improved rates of convergence in boosting. In *Advances in Neural Information Processing Systems*, 13, pages 280–286, 2001.

Shie Mannor, Ron Meir, and Tong Zhang. Greedy algorithms for classification – consistency, convergence rates, and adaptivity. 4:713–742, 2003.

Llew Mason, Jonathan Baxter, Peter L. Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In S.A. Solla, T.K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems*, 12, pages 512–518. MIT Press, 2000.

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

David Pollard. *Empirical Processes: Theory and Applications*. IMS, 1990.

J. Ross Quinlan. Bagging, boosting, and C4.5. In *13 AAAI Conference on Artificial Intelligence*, pages 725–730, Menlo Park, CA, 1996. AAAI Press.

Lev Reyzin and Robert E. Schapire. How boosting the margin can also boost classifier complexity. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 753–760, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-383-2. doi: http://doi.acm.org/10.1145/1143844.1143939.

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

Robert E. Schapire, Yoav Freund, Peter L. Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:1651–1686, 1998.

Xiaotong Shen, George C. Tseng, Xuegong Zhang, and Wing H. Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Tong Zhang and Bin Yu. Boosting with early stopping: convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.