The Annals of Statistics
2002, Vol. 30, No. 1, 51–73

ON WEAK BASE HYPOTHESES AND THEIR IMPLICATIONS FOR BOOSTING REGRESSION AND CLASSIFICATION

BY WENXIN JIANG[1]

Northwestern University

When studying the training error and the prediction error for boosting, it is often assumed that the hypotheses returned by the base learner are weakly accurate, or are able to beat a random guesser by a certain amount of difference. It has been an open question how much this difference can be, whether it will eventually disappear in the boosting process or be bounded by a positive amount. This question is crucial for the behavior of both the training error and the prediction error. In this paper we study this problem and show affirmatively that the amount of improvement over the random guesser will be at least a positive amount for almost all possible sample realizations and for most of the commonly used base hypotheses. This has a number of implications for the prediction error, including, for example, that boosting forever may not be good and regularization may be necessary. The problem is studied by first considering an analog of AdaBoost in regression, where we study similar properties and find that, for good performance, one cannot hope to avoid regularization by just adopting the boosting device to regression.

1. Introduction.

1.1. *Background and problems.* Boosting is a very useful tool for improving the performance of classification procedures and was originally developed in the field of machine learning [see, e.g., Schapire (1990) and Freund and Schapire (1997)]. In classification, the basic task is to predict a sign-valued "label" (or ± 1-valued response) Z based on the knowledge of a predictor X, with a "hypothesis" (or prediction rule) $\hat{Z}(\cdot)$ being a sign-valued function on the domain of X. The rule $\hat{Z}$ is often chosen from a "hypothesis space" H (a set of sign-valued functions), given the availability of "training data" [a set of (X, Z) pairs]. The performance of $\hat{Z}$ is often measured by the "training error" and the "test (or prediction) error," which are the misclassification probabilities on the training data and on new observations, respectively. Instead of just using a single member in H, a boosting algorithm, such as the widely used AdaBoost, uses a (sequential) linear combination of members in H and uses a combined hypothesis of the form $\hat{Z}_t = \operatorname{sgn}(\sum_{s=1}^{t} \alpha_s f_s)$ as the prediction rule at "round" or time t. Here

Received November 1999; revised July 2001.

[1]Supported in part by NSF Grant DMS-01-02636.

AMS 2000 subject classifications. Primary 62G99; secondary 68T99.

Key words and phrases. Angular span, boosting, classification, error bounds, least squares regression, matching pursuit, nearest neighbor rule, overfit, prediction error, regularization, training error, weak hypotheses.

the α_s's are coefficients and f_s's are "base hypotheses" in H chosen by some "base learning algorithm." In this context H is called a "base hypothesis space." The α_t's are determined by the "weighted training error" ε_t's at each round, which are misclassification probabilities on suitably reweighted training data. (For specific formulas see Section 6.)

It is observed that during AdaBoost the training error (on the original training data) of the combined rule $\hat{Z}_t$ decreases very quickly, while the prediction error (on new observations) sometimes does not significantly increase even after many rounds. This latter phenomenon of "resistance to overfitting" is so intriguing that it has become a serious question as to whether it is good to run boosting forever [see, e.g., Grove and Schuurmans (1998), Mason, Baxter, Bartlett and Frean (1999) and Friedman, Hastie and Tibshirani (2000)].

As Schapire (1999) pointed out, "The most basic theoretical property of AdaBoost concerns its ability to reduce training error." The training error was shown to decrease exponentially fast subject to the major *assumption of* "*weak hypotheses*" [Schapire (1999)], that the base hypotheses f_t's used in AdaBoost are "each...slightly better than random" guessing by a certain amount of difference, when evaluated by error ε_t on the weighted training data (see Section 6.1). *However, is this assumption usually valid or not? What are some implications of this assumption to the prediction error?* These will be the main focus of this paper.

It is noted that there has been much uncertainty and controversy related to this assumption of weak hypotheses. The assumption was originally justified under the assumption that the base learning algorithm is weak PAC [probably and approximately correct; see Freund and Schapire (1997)]. However, this PAC framework was found to be restrictive and inappropriate for noisy data when the Bayes error is nonzero [or when $Z|X$ is nondeterministic; see, e.g., Breiman (1998), Appendix and Discussions]. More recent work on AdaBoost therefore no longer assumes a weak PAC base learning algorithm, but instead that the base algorithm returns "weak hypotheses." As Freund and Schapire commented in the discussion of Breiman (1998), this is a very unsatisfactory characterization since it does not really tell when the assumption will be satisfied. In fact, Schapire, Freund, Bartlett and Lee (1998) were uncertain whether the weighted training error ε_t will eventually increase to 0.5 and how slow this increase would be as t, the *time* or number of rounds of boosting, increases. It was stated that "Characterizing the conditions under which the increase is slow is an open problem."

1.2. *Results and approaches.* This open problem of weak hypotheses will be studied in this paper. We will show (in Section 6.1) that for most base hypothesis spaces ε_t can be guaranteed to *not* deteriorate to 0.5 for almost all data realizations, and we will provide a bound on the difference $(\varepsilon_t - 0.5)$ based on a measure of capacity of the base hypothesis space. [For example, due to Corollary 1 and Lemma 8, if the base hypothesis space H is negation closed and contains the family of step functions and if the observed x-values (predictors) are untied, then

the weighted training error ε_t's, when optimized over H, can be made *better than random guessing for a positive amount throughout the process of AdaBoost*; i.e., $\varepsilon_t < 0.5 - \delta$ for some common $\delta > 0$ for all t.]

We will see that the wide validity of the assumption not only is relevant to the reduction of training error, but also has important consequences on the prediction error for the boosted predictions at large *time* (or number of steps of boosting). In particular, there will be important implications on whether or not it is good in general to let boosting run forever and whether or not boosting overfits eventually, which are very controversial topics [see, e.g., Grove and Schuurmans (1998), Mason, Baxter, Bartlett and Frean (1999) and Friedman, Hastie and Tibshirani (2000)]. Here overfitting refers to a prediction which is always perfect on the training sample but is poor on new test cases. We define the *amount of overfit* as the difference between the prediction error and the optimal Bayes error for large sample sizes.

We approach the problem by first studying an analog of AdaBoost in the context of regression. This was called matching pursuit by Mallat and Zhang (1993) in signal processing for sequential combination of waveforms and later recognized by Friedman, Hastie and Tibshirani (2000) as an analog of AdaBoost for least squares regression. We will reformulate this algorithm in the framework of boosting and introduce the concept of weak base hypotheses which will also imply an exponential decrease in the training error. In this case we found that the weak base hypotheses are guaranteed for most base hypothesis spaces, even for very simple ones such as the family of step functions.

Therefore, in traditional nonparametric regression with fixed x-design, the residual in fitting the y-values goes to 0 if boosting is run forever. This type of exact fitting is what is not wanted—it would suggest overfitting if the unmodified regression boosting were run forever; that is, the fit becomes perfect on the sample but poor for predicting a new observation. Regularization is needed in this case to enhance the performance. Therefore the assumption of weak hypotheses *does* hold in most cases, and that is *bad* for running the unmodified regression boosting forever—it overfits for traditional nonparametric regression. On the other hand, in an example of orthogonal base hypotheses, regularization can lead to provable improvement and avoid overfitting. *Therefore one cannot hope to avoid regularization* (*of some kind*) *just by adopting the boosting device to regression.*

What will happen in the classification case? Will the assumption of weak hypotheses be typically valid? What are some implications for the prediction error? These will be considered in an analogous treatment of AdaBoost similar to its regression analog and differences will also be discussed.

We will show that, for both regression and classification boosting, the assumption of weak hypotheses as well as how fast the training errors reduce depends on a measure of capacity of the base hypothesis space called the angular span. We will introduce the concept and the relevant properties. The assumption of weak hypotheses will be found to hold for all possible realizations of random responses

(and with all possible reweighting of the data points) if and only if the base hypothesis space has a nonzero angular span. Useful conditions for a nonzero angular span will be provided, based on a notion of completeness, and shown to actually hold for many commonly used base hypothesis spaces. We then derive bounds on the prediction error that are tight in the large time limit for fixed predictors. The bound can be easily adapted to allow for multiple response values and for random discrete predictors and for proving that "quantization" as a method of regularization can, at least theoretically, lead to optimal asymptotic performance, even for continuous or sparse data [see Jiang (2000a)].

These implications of the assumption of weak hypotheses on the prediction error are obtained by studying the assumption in a way to account for all possible realizations of the random labels (or responses). Previously, this assumption was studied by Goldmann, Hastad and Razborov (1992), Freund (1995), Freund and Schapire (1996) and Breiman (1997a, 1997b), for example, but the characterization depends on a given realization of the pattern or labels. In comparison, our approach clarifies that the assumption holds for most common hypothesis spaces (e.g., anything that contains the family of step functions as a subset) and that it holds for *all possible realizations of the random labels*, which leads to the implications regarding the prediction error. On the other hand, since our formalism protects against all possible labels, the rate we found for the training error reduction may be slower than the actual reduction that one experiences for a given data realization.

We will formulate the main results from regression (Sections 2–4) to classification (Sections 5–7). For both the regression and the classification sections, we first introduce the concept of angular span (Sections 2 and 5) as a capacity measure of the base hypothesis space. This concept is then used in Sections 3 and 6 to show the wide validity of the assumption of weak hypotheses. Section 6.1 contains some results related to the open problem of Schapire, Freund, Bartlett and Lee (1998). These imply that *even very simple base learners* can reduce the training error for a positive amount, and accumulating this over time in the unmodified boosting process will unavoidably lead to a perfect fit on the training sample. Then Sections 4 and 7 discuss the implications on the prediction errors with Propositions 2 and 5. Bounds (4.1) and (7.1) there show that the prediction errors in fixed design problems converge to suboptimal limits when boosting is run forever without regularization, under very common situations with a nonzero angular span (which can also guarantee weak hypotheses).

Below we first describe the setup of statistical learning with noisy data and define some relevant concepts and useful results. For convenience, we will formulate everything for predictors valued in $[0, 1]$, since it is obvious that most results can be easily extended to more general domains that may be multidimensional. We also assume the predictor x's to be untied. [The probability of observing ties is 0 if x is continuous. See also Jiang (2000a) for a formulation allowing ties.]

2. Angular span for regression. In statistical learning, we are faced with an observed data set $(x_i, y_i)_1^m$, where x_1^m are *predictors*, which we assume, for convenience, to be valued in $[0, 1]$ and untied. The locations of these m distinct values will be called the *predictor values*. The y_i's are real for regression problems and are $\{0, 1\}$ valued in the classification problem, where a useful transform $z_i = 2y_i - 1$ valued in $\{-1, +1\}$ is often used.

In learning, we usually have a *hypothesis space* of real regression functions $\mathcal{H}_r$ or a *hypothesis space* of $\{\pm 1\}$-valued classification functions $\mathcal{H}_c$ to fit the data. A hypothesis space, called the *base hypothesis space* or *base system* $H_{r,c}$, can be made more complex by linear combinations of t members as the *t-combined system* or *t-combined hypothesis space* denoted as $\operatorname{lin}^t(H_{r,c})$. Formally, $\operatorname{lin}^t(H) = \{\sum_1^t \alpha_s f_s : (\alpha_s, f_s) \in \Re \times H\}$. A regression space $\mathcal{H}_r$ is said to *induce* a classifier space $\mathcal{H}_c$ if $\mathcal{H}_c = \operatorname{sgn}(\mathcal{H}_r) = \{\operatorname{sgn}(f) : f \in \mathcal{H}_r\}$.

We now introduce a concept for describing the capacity of a hypothesis space $\mathcal{H}_r$, which we call the *angular span* or *a-span* and which is crucially related to the assumption of weak hypotheses and training error reduction in the context of regression boosting. We first define the angular span for a general set of nonzero vectors A in an inner product space with inner product $\langle\ ,\ \rangle_{\text{norm}}$ and squared norm $\|v\|^2 = \langle v, v\rangle_{\text{norm}}$, which is denoted as

$$\operatorname{asp}(A; \text{norm}) = \inf_{\varepsilon \neq 0} \sup_{v \in A} \langle \varepsilon/\|\varepsilon\|, v/\|v\|\rangle_{\text{norm}}^2$$

and is a quantity valued in $[0, 1]$. This is a measure of dispersion for the directions of the vectors in A. The smaller this quantity, the less well distributed the vectors in A. If A spans the vector space, then the asp is nonzero. Now consider a regression hypothesis space $\mathcal{H}_r$ and an inner product space associated with a set of distinct points x_1^m, with the inner product defined by $\langle f, g\rangle_{x_1^m} = m^{-1} \sum_1^m f(x_j) g(x_j)$ for $f, g \in \mathcal{H}_r$. The *regression a-span* for $\mathcal{H}_r$ with this particular norm is now defined as

$$\operatorname{asp}(\mathcal{H}_r; x_1^m) = \inf_{\varepsilon \in \Re^m,\, \|\varepsilon\|=1} \sup_{f \in \mathcal{H}_r,\, \|f\|>0} \langle \varepsilon, f/\|f\|\rangle_{x_1^m}^2,$$

with the obvious extension of the inner product acting on any two m-vectors a_1^m and b_1^m: $\langle a, b\rangle_{x_1^m} = m^{-1} \sum_1^m a_j b_j$, such that for a function f the corresponding m-vector is $f_1^m = f(x_j)_1^m$. By definition, the regression a-span has the following monotone properties with respect to the hypothesis space and with respect to the number of predictor values: (i) $\mathcal{H}_r \subset \mathcal{H}_r'$ implies that $\operatorname{asp}(\mathcal{H}_r; x_1^m) \leq \operatorname{asp}(\mathcal{H}_r'; x_1^m)$ and (ii) $\operatorname{asp}(\mathcal{H}_r; x_1^{m+1}) \leq \operatorname{asp}(\mathcal{H}_r; x_1^m)$.

Some examples of the regression a-span are given below.

1. If the hypothesis space is the $(p-1)$th-order regression $H = \{\sum_0^{p-1} a_k x^k : a_0^{p-1} \in \Re^p\}$, then $\operatorname{asp}(H; x_1^m) = I\{m \leq p\}$ (i.e., asp $= 1$ if $m \leq p$ and 0 if $m > p$).

2. If the hypothesis space contains m orthonormal basis vectors on x_1^m, that is, $H = \{\phi_k(\cdot)_1^m : [\phi_k(x_j)]_{1,1}^{m,m}\}$ is an orthogonal matrix, then asp $= 1/m$. This follows because in this case the asp is the squared cosine of the angle between the major diagonal of an m-dimensional cube and any of its edges.
3. If $H = \{x^k : k = 0, \ldots, m\}$, then $0 < \text{asp} \leq 1/m$.
4. If $H = \{\cos(ax) : a \in \Re\}$, $m = 2$ and $x_1^m = (0, 1)$, then the asp is $\cos^2(\pi/4)$, or 0.5.
5. If $H = \{\sin(ax) : a \in \Re\}$, $m = 2$ and $x_1^m = (0, 1)$, then the asp is 0.

The following two lemmas relate the condition of nonzero a-span to more primitive conditions that are easy to validate.

LEMMA 1 (Completeness versus nonzero a-span). *For any set of distinct predictor values x_1^m, $\text{asp}(H_r; x_1^m) > 0$ if and only if we can find m functions f_1^m from a hypothesis space H_r which produce a nonsingular matrix $[f_k(x_j)]_{1,1}^{m,m}$.*

PROOF. For the "only if" part, suppose no $f_1^m \in H_r$ can be found to produce a nonsingular matrix $[f_k(x_j)]_{1,1}^{m,m}$. Then the set $\{f(x_1^m) : f \in H_r\}$ does not span $\Re^m$, and we can find a nonzero vector $\psi = \psi_1^m$ such that $\sum_{j=1}^m f(x_j)\psi_j = 0$ for all f in H_r. Take ε to have components proportional to ψ_j. Then $\langle \varepsilon, f \rangle = 0$ for all f in H_r and the a-span must be 0. The "if" part follows from Mallat and Zhang (1993), Lemma 1. □

LEMMA 2 (Approximation and completeness). *Suppose the closure of H_r contains the set of all sign functions. More formally, suppose H_r contains, for any real number a, a sequence of functions $\{f^{(i),a}\}_{i=1}^\infty$ such that $f^{(i),a}$ converges to the function $\text{sgn}(x - a)$ at all points $x \neq a$. Then, for any set of distinct predictor values x_1^m, we can find m functions f_1^m from H_r or m functions f_1^m from $\text{sgn}(H_r)$ which produce a nonsingular matrix $[f_k(x_j)]_{1,1}^{m,m}$.*

PROOF. Consider any set of distinct predictor values $x_1^m \in [0, 1]^m$ and assume for convenience that they are ordered increasingly. Find $f_k^{(i)} \in H_r$ such that $\lim_{i\to\infty} f_k^{(i)}(x) = \text{sgn}\{x - (x_k + x_{k-1})/2\}$, $k = 1, \ldots, m$ and $x_0 \equiv -0.5$, for all x that are the continuous points of the limiting functions. Then the matrix $[f_k^{(i)}(x_j)]_{1,1}^{m,m}$ (as well as $[\text{sgn} \circ f_k^{(i)}(x_j)]_{1,1}^{m,m}$) as i increases converges to a matrix with $+1$'s in the diagonal as well as in the upper-right triangle, while with -1's in the lower-left triangle; and therefore has determinant 2^{m-1}. Therefore there must be a q large enough such that $[f_k^{(q)}(x_j)]_{1,1}^{m,m}$ (or $[\text{sgn} \circ f_k^{(q)}(x_j)]_{1,1}^{m,m}$) is nonsingular, where $f_k^{(q)} \in H_r$ for $k = 1, \ldots, m$. □

REMARK 1. The condition of this last lemma is satisfied by many base hypothesis spaces. They include all base systems that contain a family of "shifted"

cumulative distributing functions (cdf's) $\{2F\{(\cdot - \mu)/\sigma\} - 1 : \sigma > 0,\ \mu \in \Re\}$. Examples include the case when F is the logistic cdf, when the q-combined system is the usual neural net with q (tanh) nodes; the case when F is the Gaussian cdf; the threshold base system with a Heaviside cdf; the base system of mixtures of two experts [Jacobs, Jordan, Nowlan and Hinton (1991)]; and any more complicated base systems that include these base systems as submodels—for example, the base system of a neural net or the base system of a CART tree [Breiman, Friedman, Olshen and Stone (1984)]. By the consequences of the previous lemmas and the later ones, we will see that all these base systems accommodate "weak learners," which always return weak hypotheses in boosting, due to the nonzero angular span of the base systems.

Now we describe the setup for boosting least squares regression sequentially.

3. Boosting regression base learners. The *training error* for f in a regression hypothesis space $\mathcal{H}_r$ with respect to a data set $(x_i, y_i)_1^m$ is given by $m^{-1}\sum_{i=1}^m \{y_i - f(x_i)\}^2$ or $\|y - f\|^2_{x_1^m}$—we will suppress the subscripts of the norm or inner product here.

We now consider a hypothesis space H_r to be the base hypothesis space. We first build onto it by attaching a coefficient, $\alpha f \in \Re \times H_r$, and then later sequentially add up such terms to form $\sum_1^t \alpha_s f_s \in \operatorname{lin}^t(H_r)$. A *base learner* or *base learning algorithm* is defined to be an algorithm which is capable of mapping any set of responses (such as y_1^m) to $\Re \times H_r$, which can be written as $\hat{\alpha}\hat{f}\colon\ \Re^m \mapsto \Re \times H_r$. When the fit is obtained by the least squares procedure, it is typically assumed that $\hat{\alpha}\hat{f} = \arg\min_{\alpha f \in \Re \times H_r} \|y - \alpha f\|^2$ achieves the infimum of the objective function. We slightly relax this assumption and allow an approximate fit, by introducing a concept called the *precision* of $\hat{\alpha}\hat{f}$, denoted as

$$\operatorname{prec}(\hat{\alpha}\hat{f}) = \sup_{\varepsilon \in \Re^m,\, \varepsilon \neq 0} \Big(\|\varepsilon - \hat{\alpha}\hat{f}_\varepsilon\|^2/\|\varepsilon\|^2 - \inf_{\alpha f \in \Re \times H_r} \|\varepsilon - \alpha f\|^2/\|\varepsilon\|^2 \Big).$$

Similar to the tolerance level used in programming, this precision measures how complete the minimization one requires the base learner to achieve, relative to the best objective function achievable in $\Re \times H_r$. (The typical approaches assume $\operatorname{prec} = 0$ and that the minimizations are fully completed.)

Now we introduce the concept of "weak learner" similar to Schapire (1999), which will always return "weak hypotheses" that will reduce the training error by a positive amount. A base learner $\hat{\alpha}\hat{f}$ is *δ-weak* ($\delta > 0$) with respect to the set of predictor values x_1^m if the following condition holds for some $\delta > 0$:

$$\sup_{\varepsilon \in \Re^m,\, \varepsilon \neq 0} \|\varepsilon - \hat{\alpha}\hat{f}_\varepsilon\|^2/\|\varepsilon\|^2 \leq 1 - \delta.$$

This condition requires that the percentage reduction in the training error achieved by the base learner be bounded away from 0. We will see (in Lemma 4 and Remark 2) that the condition is widely valid (for some $\delta > 0$).

The following sequential algorithm LSBoost.Reg is essentially the matching pursuit algorithm of Mallat and Zhang (1993) and was recognized by Friedman, Hastie and Tibshirani (2000) as an analog to AdaBoost in the regression context. It performs sequential minimization of the square cost $C(F) = m^{-1} \sum_{j=1}^{m} \{y_j - F(x_j)\}^2$ over linear combination F of functions in H_r.

Algorithm LSBoost.Reg.

1. Let $\hat{F}_0 = 0$.
2. For all $t = 1, 2, \ldots$:
   a. Find $\hat{\alpha}_t \hat{f}_t$ which exactly or approximately minimizes $C(\hat{F}_{t-1} + \alpha f)$ over $\alpha f \in \Re \times H_r$. (Or, equivalently, let $\hat{\alpha}_t \hat{f}_t = \hat{\alpha} \hat{f}|_{\varepsilon_{t-1}}$ be a base hypothesis chosen by a base learner $\hat{\alpha} \hat{f}$ minimizing a cost function $\|\varepsilon_{t-1} - \alpha f\|^2$ over $\Re \times H_r$, with perhaps an imperfect precision, where $\varepsilon_{t-1} = y - \hat{F}_{t-1}$.)
   b. Let $\hat{F}_t = \hat{F}_{t-1} + \hat{\alpha}_t \hat{f}_t$.

The following lemmas and proposition reformulate the convergence properties of matching pursuit in the language of AdaBoost. The proofs for Lemma 4 and Proposition 1 are omitted since they are analogous to the corresponding results in classification boosting.

LEMMA 3 (Weakness and exponential rate). *If the base learner $\hat{\alpha} \hat{f}$ used in Step* 2a *of LSBoost.Reg is δ-weak, then, for any nonzero y, $\|y - \hat{F}_t\|^2 / \|y\|^2 \leq (1 - \delta)^t \leq e^{-\delta t}$.*

PROOF. The δ-weak base learner ensures that $\|\varepsilon_t\|^2 / \|\varepsilon_{t-1}\|^2 \leq 1 - \delta$ for all t. Then $\|\varepsilon_t\|^2 / \|\varepsilon_0\|^2 \leq (1 - \delta)^t$. □

LEMMA 4 ("Weakness" versus a-span, part I). *Suppose* $\operatorname{asp}(H_r) > \operatorname{prec}(\hat{\alpha} \hat{f}) \geq 0$. *Then $\hat{\alpha} \hat{f}$: $\Re^m \mapsto \Re \times H_r$ is δ-weak with $\delta = \operatorname{asp}(H_r) - \operatorname{prec}(\hat{\alpha} \hat{f}) > 0$.*

PROPOSITION 1 ("Weakness" versus a-span, part II). *Consider any specified set of predictor values x_1^m. A base learner $\hat{\alpha} \hat{f}$ valued in $\Re \times H_r$ can be made δ-weak for some positive δ, by achieving a sufficiently good precision, if and only if the base hypothesis space H_r has a nonzero a-span.*

Lemma 3 shows that the assumption of a weak base learner implies a combined learner that reduces the training error at an exponential rate. Lemma 4 implies that this exponential rate can be characterized by the a-span. Then Proposition 1 says that the weak learner assumption is in some sense equivalent to requiring a nonzero a-span (or, roughly speaking, that the hypotheses in the base hypothesis space "span a nonzero angle").

REMARK 2. More primitive conditions given in the previous section show that a large class of base systems (e.g., CART or even simple step functions) do have nonzero a-spans and accommodate δ-weak base learners. Then Lemma 3 implies that these base learners can generate a perfect fit by boosting in the large time limit, which may correspond to bad prediction errors. This implies that boosting forever may not be good in some situations and regularization may be needed—these will be discussed in later sections.

4. Overfitting behavior for regression boosting. It is interesting to see what happens to the regression boosting algorithm in the unmodified version, that is, without any regularization method. How does the prediction error behave? Is it resistant to overfitting when run forever?

We consider the case of traditional nonparametric regression when x_1^m are fixed design points [e.g., $(i/m)_{i=1}^m$]. Suppose the prediction is based on a data set $(x_i, Y_i)_1^m = \{x_i, Y(x_i)\}_1^m$, where the $Y(x_j)$'s are independent random variables with mean μ_j and variance σ^2. Consider a generic prediction $\{\hat{Y}(x) : x \in \Re\}$. The goodness is measured by the predictive mean square error or *prediction error*, defined by $L = m^{-1}\sum_{j=1}^m E\{Y_{\text{new}}(x_j) - \hat{Y}(x_j)\}^2 \equiv E\|Y_{\text{new}} - \hat{Y}\|^2$. Here the $Y_{\text{new}}(x_j)$'s are assumed to be independent new observations, which are also independent of the observed data $Y(x_i)$'s, with mean μ_j and variance σ^2 for each x_j.

The following is a bound for the prediction error for the prediction $\hat{F}_t$ obtained from t rounds of LSBoost.Reg. It is tight in the large time limit $t \to \infty$. We have seen in the previous sections that most commonly the assumption of weak hypotheses is valid and the base hypothesis space has a nonzero a-span $\text{asp}(H_r)$. In these cases, (4.1) of the proposition obviously suggests that running LSBoost.Reg forever will let the prediction error L_t converge to a (generally) suboptimal limit $L_\infty = 2\sigma^2$ [assume, e.g., $\text{prec}(\hat{\alpha}\hat{f}) = 0$ as in the usual approaches].

PROPOSITION 2 (Prediction error). *Suppose* $\text{asp}(H_r) > \text{prec}(\hat{\alpha}\hat{f}) \geq 0$. *Consider the prediction error for* $\hat{F}_t$ *obtained from* t *rounds of LSBoost.Reg*: $L_t = E\|Y_{\text{new}} - \hat{F}_t\|^2$. *Then we have*

$$|\sqrt{L_t} - \sqrt{2\sigma^2}| \leq \sqrt{m^{-1}\sum_{j=1}^m (\mu_j^2 + \sigma^2)}\exp\bigl(-\{\text{asp}(H_r) - \text{prec}(\hat{\alpha}\hat{f})\}t/2\bigr). \tag{4.1}$$

PROOF. Use the triangle inequality,

$$\bigl|\|Y_{\text{new}} - \hat{F}_t\|_E - \|Y_{\text{new}} - Y\|_E\bigr| \leq \|\hat{F}_t - Y\|_E,$$

where $\|F(\cdot) - G(\cdot)\|_E \equiv \sqrt{m^{-1}\sum_{j=1}^m E\{F(x_j) - G(x_j)\}^2}$. The proof is immediate by noting that the right-hand side of this inequality is bounded above by the right-hand side of the inequality in the proposition, due to the lemmas in the previous section which bound the training error. □

REMARK 3. (a) As a consequence of the previous section, the assumption of weak hypotheses is typically valid, and the boosting process eventually fits the data $Y(x_j)$ perfectly at each design point. The limiting prediction simply uses the data points themselves. This is clearly not what we want and severe overfit can occur: $L_\infty = \lim_{t\to\infty} L_t = 2\sigma^2$, while the optimal Bayes prediction $Y_B(x_j) = \mu_j$ has a prediction error $L^* = \sigma^2$. The difference $L_\infty - L^*$ is equal to σ^2, which can be large for noisy data and does not disappear as the sample size m increases. *Note that this can be guaranteed to happen even for very simple base hypothesis spaces as we have commented earlier, since base spaces as simple as the step functions can have nonzero a-span and can be boosted to give a perfect sample fit eventually.* Therefore we conclude that, in most cases, the unmodified regression boosting is not resistant to overfitting in the large time limit for traditional nonparametric regression.

(b) There is a very simple example to illustrate how these results work. Consider the situation $x_1^m = (j/m)_1^m$. The Y_j's are assumed to be independent Gaussian with mean μ_j and constant variance σ^2. The base learner is a set of m functions ϕ_1^m such that $\{\phi_k(1/m), \ldots, \phi_k(m/m)\}$, $k = 1, \ldots, m$, form an orthonormal basis for $\Re^m$. This system has a-span $1/m$. We use LSBoost.Reg to fit a linear combination of these base functions to the data Y_1^m. At each round, suppose the maximization is completed (prec $= 0$). Then the training error drops from $\|Y_1^m\|^2$ to 0 in exactly m steps, since LSBoost.Reg each time reduces the squared length $\|Y_1^m\|^2$ by iteratively removing the largest projection of Y_1^m on a base function. The prediction error L_t, however, reaches its limit point $2\sigma^2$ at step m and shows an overfit of amount σ^2 as compared to the best possible prediction error σ^2.

Therefore the assumption of weak hypotheses *does* hold in most cases, and that is *bad* for running regression boosting forever. However, we conjecture that the validity of the assumption is good for "boosting in the process." That is, the validity of the assumption implies a "complete spectrum of predictions" with varying degrees of complexity—it "traces the dots" at large time, while it uses the dumbest fit 0 at the beginning. One naturally guesses that there will be an optimal time at which the boosted prediction will have a good performance in prediction error.

Indeed, this can be rigorously stated and proved in the setup of item (b) of the remark preceding where orthogonal base hypotheses are used. For this system, it is straightforward to show the following:

1. The boosted prediction at any time t is exactly solvable and basically retains the t ($\leq m$) largest (in magnitude) sample Fourier coefficient $\langle Y, \phi_k\rangle$'s in the orthogonal expansion $\sum_{k=1}^m \langle Y, \phi_k\rangle \phi_k$.
2. There is a boosted prediction at some time which is at least as good as the orthogonal series estimator with any hard thresholding. As a corollary, and utilizing the results of Donoho and Johnstone (1994), we then see that the boosted

prediction at some time is asymptotically minimax in reducing the prediction error in the family of *all possible* signals μ_1^m. In other words, no measurable estimator can beat all boosted predictions simultaneously for all signals.

The main point of the above results is that *for this exactly solvable boosting system one boosting prediction at some time is essentially optimal* (*in the sense of asymptotic minimax*) *among* all possible *estimators*. But, in practice, how can one construct the boosted prediction at an optimal time, knowing that neither the "unboosted" nor the "boost-forever" predictions are good in general? This can be done by a method that is similar to hard thresholding, that is, retaining only the coefficients obtained in boosting that are larger (in magnitude) than a certain threshold.

Such an algorithm can be run forever without overfitting by using a suitably chosen threshold (the prediction will stabilize after a certain time):

3. If the threshold is chosen to estimate the universal threshold $\sigma\sqrt{2\log m/m}$ [see, e.g., Donoho and Johnstone (1994)], then the boosted prediction after a certain time becomes the same as the orthogonal series estimator with the universal hard thresholding. Consequently, the resulting estimator is asymptotically (minimax-)optimal among all possible estimators when protecting against all possible signals.

Presumably, some thresholding techniques could be applied to boosting with other base hypothesis spaces and be adapted to the classification context. However, analytic results would be harder to obtain and this is currently under investigation.

The main message here is that, in most cases in standard nonparametric regression with fixed x, the assumption of weak hypotheses *does* hold and implies overfit in the large time limit and that it is not good to run the unmodified regression boosting forever. Regularization is not unnecessary but potentially beneficial. A natural question is: what happens in the case of classification?

5. Angular span for classification. The response or *label* y_i's are $\{0, 1\}$ valued in the classification problem, where a useful transform $z_i = 2y_i - 1$ valued in $\{-1, +1\}$ is often used. A hypothesis space $\mathcal{H}_c$ is a set of functions $f\colon [0, 1] \mapsto \{\pm 1\}$. It can often be induced by a regression space $\mathcal{H}_r$ by $\mathcal{H}_c = \operatorname{sgn}(\mathcal{H}_r)$. The space $\mathcal{H}_c$ is said to be *negation closed* if $f \in \mathcal{H}_c$ whenever $-f \in \mathcal{H}_c$. For measuring the capacity of $\mathcal{H}_c$, we define the *classification angular span* related to a set of predictor values x_1^m. Denoting $P^m = \{w_1^m : w_j \geq 0, \sum_1^m w_j = 1\}$, we define

$$\operatorname{asp}_c(\mathcal{H}_c; x_1^m) = \inf_{w_1^m \in P^m,\, z_1^m \in \{\pm 1\}^m} \sup_{f \in \mathcal{H}_c} \left| \sum_{j=1}^m w_j z_j f(x_j) \right|.$$

This quantity obviously lies in $[0, 1]$ as the regression a-span. It also has similar monotone properties: (i) $\mathcal{H}_c \subset \mathcal{H}_c'$ implies that $\operatorname{asp}_c(\mathcal{H}_c; x_1^m) \leq \operatorname{asp}_c(\mathcal{H}_c'; x_1^m)$ and (ii) $\operatorname{asp}_c(\mathcal{H}_c; x_1^{m+1}) \leq \operatorname{asp}_c(\mathcal{H}_c; x_1^m)$.

Some simple examples are:

1. For the hypothesis space of delta-functions [Schapire, Freund, Bartlett and Lee (1998)] $H_c = \{s \cdot \delta_a : s \in \{\pm 1\},\ a \in \Re\}$, where $\delta_a(x) = 2I\{x = a\} - 1$, we have $3/m \geq \text{asp}_c(H_c) \geq 1/m$ for any set of m predictor values.
2. For the hypothesis space of threshold functions (or "*stumps*") $H_c = \{s \cdot \text{sgn}_a : s \in \{\pm 1\},\ a \in \Re\}$, where $\text{sgn}_a(x) = 2I\{x \geq a\} - 1$, we have $2/m \geq \text{asp}_c(H_c) \geq 1/m$ for any set of m predictor values.
3. Suppose $x_1^m = \{0, 1\}$, $H_c = \{\text{sgn}[\cos\{a(x - 1/2)\}] : a \in \Re\}$. Then $\text{asp}_c(H_c) = 0$, which is easily proved by applying the definition and taking $w_1^m = \{1/2, 1/2\}$ and $z_1^m = \{-1, 1\}$.

Some results that are useful in obtaining upper bounds for the classification a-span are included in the Appendix.

Sufficient conditions for $\text{asp}_c > 0$ are summarized in the following lemma and proposition.

LEMMA 5 (Completeness versus nonzero a-span). $\text{asp}_c(H_c; x_1^m) > 0$ *if and only if there exist* $f_1^m \in H_c$ *such that the matrix* $[f_k(x_j)]_{1,1}^{m,m}$ *is nonsingular.*

PROOF. First prove "if." Suppose there exist $f_1, \ldots, f_m$ in H_c such that the matrix $[f_k(x_j)]_{1,1}^{m,m}$ is nonsingular. If the a-span were 0, then, by its definition, there would exist $w_1^m \in P^m$ and $z_1^m \in \{\pm 1\}^m$ such that $\sum_{j=1}^m w_j z_j f(x_j) = 0$ for all $f \in H_c$. Then the set of linear equations $\sum_{j=1}^m w_j z_j f_k(x_j) = 0$ for all $k = 1, \ldots, m$ would imply that $(w_j z_j)_1^m$ is a zero vector. This contradicts the fact that $(w_j z_j)_1^m$ has to be a nonzero vector (otherwise, $\sum_{j=1}^m |w_j z_j| = \sum_1^m w_j = 1$ would be violated). Therefore the a-span cannot be 0.

Now the "only if" part. Suppose no $f_1^m \in H_c$ can be found to produce a nonsingular matrix $[f_k(x_j)]_{1,1}^{m,m}$. Then the set $\{f(x_1^m) : f \in H_c\}$ does not span $\Re^m$, and we can find a nonzero vector ψ_1^m such that $\sum_{j=1}^m f(x_j)\psi_j = 0$ for all f in H_c. Take $w_j = |\psi_j| / \sum_1^m |\psi_k|$ and $z_i = \text{sgn}(\psi_i)$. Then $\sum_{j=1}^m f(x_j) w_j z_j = 0$ for all f in H_c and the a-span must be 0. □

By Lemma 2, we therefore also have:

PROPOSITION 3 (Approximation and a-span). $H_c = \text{sgn}(H_r)$ *and* H_r *can approximate any sign function* (*see Lemma* 2) *imply that* $\text{asp}_c(H_c; x_1^m) > 0$ *for any set of* (*distinct*) *predictor values* x_1^m.

(That is, the classification a-span is nonzero if the classifier space H_c is induced by a regression space H_r which can approximate any sign function.)

Now we show that a nonzero a-span of the classification base system H_c implies that the training error can be made arbitrarily small by applying the base learners sequentially and that the usual assumption of weak hypotheses is valid. Due to Remark 1 and Proposition 3, the assumption is actually valid for most situations. We now introduce the setup.

6. Boosting classification base learners. Let $S = (x_i, z_i)_1^m$, $z_i \in \{-1, +1\}$, be the observed data. Let $\hat{z}(\cdot) \in \mathcal{H}_c$ be a prediction based on the observed data, also taking values from $\{-1, +1\}$. Then the *training error* can be denoted as $\hat{L} = m^{-1} \sum_{j=1}^m I\{z_i \neq \hat{z}(x_i)\}$.

Suppose the sign-valued prediction $\hat{z}$ is induced by a real hypothesis: $\hat{z} = \operatorname{sgn} \circ F$ for some $F \in \mathcal{H}_r$. Then we have the following inequality:

$$\hat{L} = m^{-1} \sum_{j=1}^m I\{z_i \neq \hat{z}(x_i)\} \leq D(F) \equiv m^{-1} \sum_{j=1}^m e^{-F(x_j) z_j}.$$

AdaBoost can be regarded as sequentially minimizing this upper bound $D(F)$ as a cost function [see, e.g., Breiman (1997a), Mason, Baxter, Bartlett and Frean (1999) and Friedman, Hastie and Tibshirani (2000)]. The hypothesis space $\mathcal{H}_r$ of the F's is the space of linear combinations of t base hypotheses: $\mathcal{H}_r = \operatorname{lin}^t(H_c)$ at round t.

Algorithm AdaBoost:

1. Set $\hat{F}_0 = 0$.
2. For $t = 1, 2, \ldots$:
   a. Find $\hat{\alpha}_t \hat{f}_t$ which exactly or approximately minimizes $D(\hat{F}_{t-1} + \alpha f)$ over $\alpha f \in \Re \times H_c$.
   b. Set $\hat{F}_t = \hat{F}_{t-1} + \hat{\alpha}_t \hat{f}_t$, $\hat{z}_t = \operatorname{sgn} \circ \hat{F}_t$.

This algorithm is obviously similar to the LSBoost.Reg algorithm, except that the cost function is the exponential cost $D(F)$ and a sign transform of $\hat{F}_t$ is applied to produce a sign-valued prediction $\hat{z}_t$. With a minimization partially completed on the coefficient of the linear combination, *Step* 2a *is equivalent to the more familiar formulation* 2a′ *of* "*training on a reweighted data set*":

2a′. Find some $f = \hat{f}_t \in H_c$ which exactly or approximately minimizes

$$\inf_{\alpha \in \Re} \{D(\hat{F}_{t-1} + \alpha f)/D(\hat{F}_{t-1})\} = 2\sqrt{\tfrac{1}{4} - (\varepsilon_t - \tfrac{1}{2})^2}.$$

Then set $\hat{\alpha}_t = \frac{1}{2} \ln((1 - \hat{\varepsilon}_t)/\hat{\varepsilon}_t)$. [When H_c is negation closed, the minimization step is equivalent to finding $\hat{f}_t$ to (approximately) minimize the weighted training error ε_t.] Here we denote the *weighted training errors* $\varepsilon_t = \sum_{j=1}^m w_j^{(t)} I\{f(x_j) \neq z_j)\}$ and $\hat{\varepsilon}_t = \sum_{j=1}^m w_j^{(t)} I\{\hat{f}_t(x_j) \neq z_j\}$, with *weights* $w_j^{(t)} = e^{-\hat{F}_{t-1}(x_j) z_j} / \sum_{k=1}^m e^{-\hat{F}_{t-1}(x_k) z_k}$.

In this case, $\hat{f}_t$ is generated by a classification base learner. A *classification base learner* is, in general, a mapping $\hat{f}$ from $P^M \times \{\pm 1\}^m$ to H_c; that is, when input with a set of weights and labels $(w_1^m, z_1^m) \in P^M \times \{\pm 1\}^m$, the base learner $\hat{f}$ outputs a base hypothesis in H_c. We sometimes also denote by $\hat{f}$ the

function in H_c that is selected by the base learner, which should be clear from the context. Note that the weighted training error of a base hypothesis $f \in H_c$ is $\varepsilon = \sum_{j=1}^m w_j I\{z_j \neq f(x_j)\}$, and $(\varepsilon - 1/2)^2 = (1/4)|\sum_{j=1}^m w_j z_j f(x_j)|^2$, which is the type of quantity that Step 2a$'$ of AdaBoost maximizes. It is therefore reasonable to define the *precision* for a base learner $\hat{f}$ to measure how completely it performs the maximization, by

$$\begin{aligned}&\tfrac{1}{4}\operatorname{prec}(\hat{f})^2\\ &\quad= \sup_{(w_1^m, z_1^m) \in P^m \times \{\pm 1\}^m} \left\{ \sup_{f \in H_c} \tfrac{1}{4} \left| \sum_{j=1}^m w_j z_j f(x_j) \right|^2 - \tfrac{1}{4} \left| \sum_{j=1}^m w_j z_j \hat{f}(x_j) \right|^2 \right\}.\end{aligned}$$

(Typically, optimization is assumed to be complete and $\operatorname{prec} = 0$.)

Regarding the open problem of Schapire, Freund, Bartlett and Lee (1998), we note that the weighted training error $\hat{\varepsilon}_t$ will be bounded away from 0.5 if the base learner $\hat{f}$ can differ from random guessing by a positive amount, *uniformly for all weights and labels*. That it, we will be guaranteed that $|\hat{\varepsilon}_t - 0.5| > \delta$ for all t for some $\delta > 0$ (or $\hat{\varepsilon}_t < 0.5 - \delta$ $\forall t$ when H_c is negation closed) if

$$\inf_{(w_1^m, z_1^m) \in P^m \times \{\pm 1\}^m} \left| \sum_{j=1}^m w_j I\{z_j \neq \hat{f}(x_j)\} - \tfrac{1}{2} \right| \geq \delta,$$

in which case we say that the base learner $\hat{f}$ is *δ-weak* ($\delta > 0$). We will comment on this more in Section 6.1.

Similar to the regression case, it is easy to prove that (i) AdaBoost reduces the reducible training error exponentially fast if the base learner is δ-weak, (ii) the base learner is δ-weak for $\delta = (1/2)\sqrt{\operatorname{asp}_c(H_c)^2 - \operatorname{prec}(\hat{f})^2}$ if $\operatorname{asp}_c(H_c)^2 > \operatorname{prec}(\hat{f})^2 \geq 0$ and (iii) the base learner $\hat{f}$ is δ-weak for some positive δ by choosing a sufficiently small $\operatorname{prec}(\hat{f})$ if and only if the base hypothesis space H_c has a nonzero a-span (which holds very commonly; see Proposition 3 or Section 6.1).

LEMMA 6 (Weakness and exponential rate). *Suppose $\hat{f}_t = \hat{f}|_{(w_j^{(t)})_{j=1}^m, z_1^m}$ used in Step* 2a *of AdaBoost is generated by a base learner $\hat{f}$ that is δ-weak. Then the training error of the boosted prediction satisfies, for all $t = 1, 2, \ldots,$*

$$\hat{L}_t \equiv m^{-1} \sum_{j=1}^m I\{\hat{z}_t(x_j) \neq z_j\} \leq (1 - 4\delta^2)^{t/2} \leq e^{-2\delta^2 t}.$$

PROOF. This basically follows from the techniques of, for example, Schapire (1999). □

LEMMA 7 ("Weakness" versus a-span, part I). *Suppose* $\operatorname{asp}_c(H_c)^2 > \operatorname{prec}(\hat{f})^2 \geq 0$. *Then the base learner* $\hat{f}$: $P^m \times \{\pm 1\}^m \mapsto H_c$ *is* δ*-weak for* $\delta = \frac{1}{2}\sqrt{\operatorname{asp}_c(H_c)^2 - \operatorname{prec}(\hat{f})^2}$, *and therefore we have, for all* t,

$$m^{-1}\sum_{j=1}^{m} I\{\hat{z}_t(x_j) \neq z_j\} \leq \{1 - \operatorname{asp}_c(H_c)^2 + \operatorname{prec}(\hat{f})^2\}^{t/2}$$
$$\leq \exp\bigl(-(t/2)\{\operatorname{asp}_c(H_c)^2 - \operatorname{prec}(\hat{f})^2\}\bigr).$$

PROOF. Fix x_1^m for all $(w_1^m, z_1^m) \in P^m \times \{\pm 1\}^m$. Due to the definition of $\operatorname{prec}(\hat{f})$, we have

$$\tfrac{1}{4}\left|\sum_{j=1}^{m} w_j z_j \hat{f}(x_j)\right|^2 \geq \sup_{f \in H_c} \tfrac{1}{4}\left|\sum_{j=1}^{m} w_j z_j f(x_j)\right|^2 - \tfrac{1}{4}\operatorname{prec}(\hat{f})^2.$$

Hence

$$\inf_{(w_1^m, z_1^m) \in P^m \times \{\pm 1\}^m} \tfrac{1}{4}\left|\sum_{j=1}^{m} w_j z_j \hat{f}(x_j)\right|^2 \geq \tfrac{1}{4}\operatorname{asp}_c(H_c)^2 - \tfrac{1}{4}\operatorname{prec}(\hat{f})^2.$$

Taking square roots of both sides shows the lemma. [Note that

$$(*) \qquad \tfrac{1}{2}\left|\sum_{1}^{m} w_j z_j \hat{f}(x_j)\right| = \left|\sum_{j=1}^{m} w_j I\{z_j \neq \hat{f}(x_j)\} - \tfrac{1}{2}\right|.] \qquad \square$$

Similar to the regression case (Proposition 1), we again have a two-sided relationship between the existence of δ-weak base learners and the nonzero capacity (as measured by a-span) of the base hypothesis space:

PROPOSITION 4 ("Weakness" versus a-span, part II). *The base learner* $\hat{f}$: $P^m \times \{\pm 1\}^m \mapsto H_c$ *is* δ*-weak for some positive* δ *by choosing a sufficiently small* $\operatorname{prec}(\hat{f})$ *if and only if the base hypothesis space* H_c *has a nonzero a-span.*

PROOF. "If" is obvious from the proof of Lemma 7. For proving "only if," fix x_1^m and consider any $(w_1^m, z_1^m) \in P^m \times \{\pm 1\}^m$. Note that $\sup_{f \in H_c} \frac{1}{4}\bigl|\sum_{j=1}^m w_j z_j f(x_j)\bigr|^2 \geq \frac{1}{4}\bigl|\sum_{j=1}^m w_j z_j \hat{f}(x_j)\bigr|^2$. Taking inf over $(w_1^m, z_1^m) \in P^m \times \{\pm 1\}^m$ for both sides of the inequality leads to the proof. Note that $\inf(\text{left-hand side}) = \frac{1}{4}\operatorname{asp}_c(H_c)^2$, and $\inf(\text{right-hand side}) \geq \delta^2 > 0$, which is implied by $\hat{f}$ being δ-weak and the equation $(*)$ in the proof of Lemma 7. $\square$

6.1. *Weak hypotheses and related problems.* As discussed in the Introduction, the assumption of weak hypotheses, useful for proving the exponential reduction of training error in AdaBoost, assumes that the weighted training error $\hat{\varepsilon}_t < 0.5 - \delta$ for all t for some common $\delta > 0$. However, Schapire, Freund, Bartlett and Lee (1998) were uncertain whether $\hat{\varepsilon}_t$ increases as a function of t, "possibly even converging to $1/2$," and raised the open problem on "conditions under which the increase is slow."

As a corollary to Lemmas 5 and 7, we obtain:

COROLLARY 1 ("Weak edges"). *The weighted training error $\hat{\varepsilon}_t$ in AdaBoost satisfies $|\hat{\varepsilon}_t - 0.5| > \delta$ for all t (or $\hat{\varepsilon}_t < 0.5 - \delta\ \forall t$ when H_c is negation closed), where the constant δ can be taken as $(1/2)\sqrt{\operatorname{asp}_c(H_c)^2 - \operatorname{prec}(\hat{f})^2}$ and can be made positive by achieving a relatively precise optimization in Step* 2a′, *if the base hypothesis space H_c has a nonzero a-span, or if $H_c(x_1^m) \equiv \{f(x_j)_1^m : f \in H_c\}$ spans $\Re^m$.*

For simplicity, from now on we suppose that the optimization steps are ideally carried out [$\operatorname{prec}(\hat{f}) = 0$]. The most important condition of the corollary is the "completeness condition" that $H_c(x_1^m)$ spans $\Re^m$, which is satisfied in most common situations due to the following lemma:

LEMMA 8 (Sufficient conditions for completeness). *$H_c(x_1^m)$ spans $\Re^m$ if x_1^m are mutually distinct and H_c contains one of the following sets of functions*:

(i) *the set of sign functions* $\mathrm{SGN} = \{\operatorname{sgn}(\cdot - a) : a \in \Re\}$; *or*
(ii) *the set of delta functions* $\mathrm{DLT} = \{I(\cdot = a) - I(\cdot \neq a) : a \in \Re\}$; *or*
(iii) *the set of disks with any radius $r \geq 0$*:

$$\mathrm{DSK}(r) \equiv \big\{I\big(\cdot \in [a - r, a + r]\big) - I\big(\cdot \notin [a - r, a + r]\big) : a \in \Re\big\}.$$

PROOF. Result (i) is a corollary to Lemma 2 and the other two results can be similarly proved by finding m functions f_1^m from H_c to form a nonsingular matrix $[f_k(x_j)]_{1,1}^{m,m}$. □

REMARK 4. (a) Very commonly, x_1^m are distinct with probability 1 (e.g., when x_1^m are realizations of continuous random variables) and $H_c \supset \mathrm{SGN}$ (e.g., when H_c is the set of CART trees). So, regarding the open problem of Schapire, Freund, Bartlett and Lee (1998), the corollary and the lemma guarantee that $\hat{\varepsilon}_t$ *will not deteriorate to* $1/2$ *for almost all data realizations and even for H_c as simple as the* "*stumps*" (or the set of threshold functions defined in Example 2 of Section 5). Therefore the assumption of weak base hypotheses is not restrictive but widely valid instead; that is, in most common situations the corresponding base hypotheses $\hat{f}_t$ will be "better than random guessing" by a positive amount and

the training error drops exponentially fast. [Breiman has also reached a similar conclusion independently (private communication).]

(b) Originally, the "weak edges" ($|\hat{\varepsilon}_t - 0.5| > \delta > 0\ \forall t$) required in the theory of training error reduction were guaranteed by a base learning algorithm that is *weak PAC* (probably and approximately correct); see Freund and Schapire (1997). Soon it was realized that the notion of weak PAC is restrictive and not appropriate for noisy data [see, e.g., Breiman (1998), Appendix and Discussions]. In the PAC framework, the label z is assumed to follow a deterministic function of the predictor x called a "concept." This framework is not suitable for noisy data where z given x is random and is no longer used in more recent papers on AdaBoost [e.g., Schapire, Freund, Bartlett and Lee (1998), Schapire (1999)].

(c) Even though the weak PAC framework is not appropriate for guaranteeing the nonzero weak edges for training error reduction when data are noisy, our results show that the differences $(\hat{\varepsilon}_t - 0.5)$ on the weighted training sets are still guaranteed to be bounded away from 0 in most common situations, which suffice for obtaining an exponential rate of training error reduction. These have been obtained under a different framework of "δ-weak" base learner to guarantee the weak edges: we require that the weighted training errors for the base learner be different from 0.5 for a positive amount uniformly for all weights and labels. This notion of "δ-weak" does not depend on the "underlying concept," is suitable for noisy data and is found to be valid for most common situations. [Due to Proposition 4 and Lemmas 5 and 8, a base learner $\hat{f}$ can be made δ-weak on almost all training sets (with x_1^m untied) with a relatively precise optimization if the corresponding base hypothesis space $H_c \supset \mathrm{SGN}$. Then, for example, the CART system (or even the stumps) are typically δ-weak.]

In summary, in this part of the paper we show that the assumption of weak hypotheses typically holds (even for very simple base systems such as the stumps). The training error will therefore be guaranteed to drop to 0. Now the question is, what do these results imply for the prediction error? An important implication is that boosting forever can eventually generate a perfect fit on almost all training samples (in fact, after some finite time, see Proposition 7 in the Appendix), which may not be good for the prediction error in some situations, as we will discuss in the later sections.

7. Overfitting behavior for classification boosting. Results on the large time behavior of the prediction error can be derived similar to the regression case. Suppose the prediction is based on a data set $(x_i, Z_i)_1^m = \{x_j, Z(x_j)\}_1^m$ with fixed x_1^m, where $Z(x_j)$'s are random and independent sign-valued ($\{-1, +1\}$-valued) variables with "signal" $P\{Z(x_j) = 1\} = \mu_j$. The prediction error is defined as $L_t = m^{-1} \sum_{j=1}^m I\{Z_{\mathrm{new}}(x_j) \neq \hat{Z}_t(x_j)\}$ for the prediction $\hat{Z}_t \equiv \mathrm{sgn} \circ \hat{F}_t$ obtained from t rounds of AdaBoost. Here $Z_{\mathrm{new}}(x_j)$'s are assumed to be random

and independent new observations that are also independent of the observed data, with "signal" μ_j for each x_j.

The following is a bound for the prediction error L_t which is tight in the large time limit $t \to \infty$. We have seen in the previous sections that most commonly the assumption of weak hypotheses is valid and the base hypothesis space has a nonzero a-span asp. In these cases, (7.1) of the proposition obviously suggests that running AdaBoost forever would let the prediction error L_t converge to a generally suboptimal limit L_∞ (assume, e.g., prec $= 0$ as in the usual approaches).

PROPOSITION 5 (Prediction error). *Denote* $\text{asp} = \text{asp}_c(H_c)$ *as the angular span of the base hypothesis space*, $\text{prec} = \text{prec}(\hat{f})$ *as the precision of the base learner. Suppose* $\text{asp} > \text{prec} \geq 0$. *Then we have*

$$L_t \leq L_\infty + e^{-t(\text{asp}^2 - \text{prec}^2)/2} \quad \textit{and} \quad \sqrt{L_t} \geq \sqrt{L_\infty} - \sqrt{e^{-t(\text{asp}^2 - \text{prec}^2)/2}}. \tag{7.1}$$

Here $L_\infty = m^{-1} \sum_{j=1}^m 2\mu_j(1 - \mu_j) \equiv L^* + \Delta$, *where* $L^* = m^{-1} \sum_{j=1}^m \min(\mu_j, 1 - \mu_j)$ *is the Bayes error and* $\Delta = m^{-1} \sum_{j=1}^m 2|\mu_j - 1/2| \min(\mu_j, 1 - \mu_j)$ *measures the difference* $L_\infty - L^*$.

See Jiang (2000a) for the proof and a more general formulation allowing multiple responses at the x-locations.

REMARK 5. (a) The large time limit L_∞ of the prediction error is the same as that of the nearest neighbor rule $L_{\text{NN}} = m^{-1} \sum_1^m 2\mu_j(1 - \mu_j)$, which exceeds the Bayes error L^* by at most $\min(0.125, L^*)$. So there is, in general, an overfit as compared to the Bayes rule. [It is noted, however, that the amount of overfit is usually small for data with little noise (i.e., when L^* is small) and cannot exceed 12.5%. This is in sharp contrast to the case of regression boosting. That is, although there can be a nonzero overfit, the amount of overfit $L_\infty - L^*$ cannot be arbitrarily large.]

(b) The typical time used to approach this overfitting limit may be of order $1/\text{asp}^2$, as suggested by the exponential rate of the bound when taking a perfect precision $\text{prec} = 0$. This typical time has the order of squared sample size and can be therefore quite long according to the example computations of asp in the earlier sections, for example, for decision stumps (step functions). It is unclear whether this is related to the empirical evidence that boosting often overfits only after tens of thousands of rounds [see, e.g., Grove and Schuurmans (1998)].

(c) What about the situation when the x's are random? In fact, with random continuous predictors on $[0, 1]$, in the case of boosting the decision stumps or CART systems, limiting the cuts of the step functions to be located at the mid-data points will also generate the nearest neighbor rule for all sufficiently large time. Therefore similar overfitting behavior can occur for noisy data.

(d) What about boosting forever with a higher dimensional random continuous predictor x with $\dim(x) > 1$? We do not have theoretical results on this so far. However, recent empirical studies also confirm that even for high dimensional data "boosting forever" is still suboptimal, when compared to predictions obtained from somewhere earlier in the boosting process [Grove and Schuurmans (1998), Mason, Baxter, Bartlett and Frean (1999) and Friedman, Hastie and Tibshirani (2000), among others].

Even though running the unmodified AdaBoost forever can lead to a suboptimal prediction error, we expect that [see Jiang (2000b)], as in the case of regression boosting, somewhere in the process of boosting a prediction rule is nearly asymptotically optimal, in the sense that the prediction error is close to the optimal Bayes error when the size of the training sample is large. Jiang (2000a), Remark 6b, also discussed a quantization method for regularizing AdaBoost to avoid overfitting.

8. Conclusions. This paper investigates when the assumption of weak base hypotheses used in boosting is valid and discusses its implications for the prediction error in the large time limit. Most of the commonly used base hypothesis spaces (even very simple ones as the decision stumps) are shown capable of generating weak hypotheses and can eventually generate a perfect fit when there are no ties in the data. An implication is that both the unmodified regression and the classification boosting algorithms will likely overfit when run forever. The amount of overfit is typically smaller for classification boosting and is related to the noise level (which may be small when the Bayes error is low); the time needed for approaching the limiting fit may also be longer and may have the order of the squared sample size. This may be part of the reason why overfitting has not been noticed until recently. However, we conjecture that regularization of the boosting processes, whether to stop at some finite time, or to shrink the coefficients or to quantize the predictor space, may still lead to better performance for noisy data (this is provable at least in some examples). Therefore the emerging literature on regularized variants of boosting may not be unnecessary, despite the fact that the unmodified AdaBoost is often resistant to overfitting after hundreds of runs. For some work on regularization, see, for example, Friedman (1999a, 1999b) (empirical work with shrinkage and randomization), Mason, Baxter, Bartlett and Frean (1999) (complexity penalty) and Breiman (1996, 1999) and Bühlmann and Yu (2000) (bagged versions of boosting).

The current method does not provide a general result for the most realistic case with sparse data (with high-dimensional random continuous predictors). It is only for the case of sparse data, where it is possible that the prediction error of AdaBoost continues to decrease after a perfect fit on the training sample. *It is important to note that the results of this paper cannot explain this observed mystery. In most of the cases considered, the prediction error stabilizes simultaneously with the*

training error. The best explanations so far for this mystery seem to be the margin approach of Schapire, Freund, Bartlett and Lee (1998) and the top approach of Breiman (1997a), which are still semiempirical in nature. It is, however, plausible to conjecture that even in the case of sparse data running AdaBoost *forever* can still lead to a suboptimal prediction in the sense defined in this paper, since our results imply that the fit will be perfect *for all sample realizations* and agree with the nearest neighbor rule at all the data points as well as in some of their neighborhoods. The limiting prediction presumably cannot perform much better than the nearest neighbor rule. Recent empirical studies also confirm that even for high-dimensional sparse data AdaBoost may deteriorate after running for a *very* long time [e.g., Grove and Schuurmans (1998), Mason, Baxter, Bartlett and Frean (1999) and Friedman, Hastie and Tibshirani (2000)].

One may wonder why the validity of the assumption of weak hypotheses was probably perceived as a positive thing. One possible reason is that boosting was originally derived in the PAC framework with data with noiseless labels, where a perfect sample fit implied by the assumption is typically good for the prediction error also. An algorithm that fits the data perfectly is said to be "consistent" in the PAC framework and is an important condition to prove good performance in the prediction error [see, e.g., Anthony and Biggs (1992), Chapter 4]. Another possible reason is that originally the inventors of AdaBoost may not have intended to let the algorithm run forever, but rather to truncate the process (which *is* a regularization method!) [see Freund and Schapire (1997)]. For such an approach of "boosting in the process" the validity of the assumption may not have negative implications.

Although we argue that the assumption of weak hypotheses typically holds and this can be problematic for the approach of "boosting forever" with noisy data, we suspect that *in the process* of boosting a prediction rule can still achieve a very good prediction error at some time. This is illustrated in an example at the end of the sections on regression boosting. See also Jiang (2000b) for some results on the performance of "AdaBoost in the process."

Our approach is based on an analog of AdaBoost in the regression context. The analogous treatment has been helpful in understanding the weak hypotheses and their implications on prediction error, whether or not the boosting algorithms will eventually overfit and by what amount, whether regularization is needed at all or potentially beneficial and what are some possible approaches. We believe that further studies of the analogy still have a lot more to tell.

APPENDIX

Some results related to the classification angular span. Upper bounds of the classification angular span may be obtained from the following two results.

LEMMA 9 (Sign change). *Suppose all the hypotheses in H_c change signs K times or less. More formally, let $\mathcal{K}_f$ be the number of connected components of*

the positive support $\{x : f(x) = 1\}$ *plus the number of connected components of the negative support* $\{x : f(x) = -1\}$ *and suppose that* $\sup_{f \in H_c} \mathcal{K}_f \le K$. *Then we have* $\operatorname{asp}_c(H_c : x_1^m) \le K/m$ *for any set of* (*distinct*) *predictor values* x_1^m.

PROOF. Without loss of generality, assume x_1^m to be ordered increasingly. Let $z_1^m = (-1, 1, -1, 1, \ldots)$ alternating the signs and $w_1^m = 1/m$. Then, for any $f \in H_c$, there exists $\bigcup_{k=1}^K R_k$ a partition of the set of integers $\{1, \ldots, m\}$ such that R_k either is empty or contains consecutive integers such that $\{f(x_j)\}_{j \in R_k}$ carry the same sign. Then (treating the empty summands to be 0) we have

$$\begin{aligned}\left|\sum_{j=1}^m w_j z_j f(x_j)\right| &= \left|\sum_{k=1}^K \sum_{j \in R_k} w_j z_j f(x_j)\right| \\ &\le \sum_{k=1}^K \left|\sum_{j \in R_k} w_j z_j\right| |f(x_j)| = \sum_{k=1}^K \left|\sum_{j \in R_k} w_j z_j\right| = (1/m) \sum_{k=1}^K \left|\sum_{j \in R_k} z_j\right| \\ &\le K/m,\end{aligned}$$

since $|\sum_{j \in R_k} z_j| \le 1$, due to the alternating signs of $(z_j)_{j \in R_k}$.

Therefore $\operatorname{asp}_c(H_c; x_1^m) \le \sup_{f \in H_c} |\sum_{j=1}^m w_j z_j f(x_j)| \le K/m$. □

The following relationship holds between the classification a-span and a more commonly used measure of capacity, the VC dimension [for the concept of the VC dimension, see, e.g., Anthony and Biggs (1992), Chapter 7]:

PROPOSITION 6 (A-span and VC dimension). $\operatorname{asp}_c(\mathcal{H}_c; x_1^m) \le \sqrt{(2\mathrm{VC}(\mathcal{H}_c) \log m + 4)/m}$ *if* $\mathrm{VC}(\mathcal{H}_c) > 2$, *or* $\le \sqrt{(2 \log \operatorname{card}(\mathcal{H}_c) + 4)/m}$ *if* $\mathcal{H}_c$ *is finite and nonempty*.

PROOF. Due to the definition of a-span,

$$\operatorname{asp}_c(\mathcal{H}_c; x_1^m) \le E(Q) \equiv E \sup_{f \in \mathcal{H}_c} \left| m^{-1} \sum_{j=1}^m Z_j f(x_j) \right|,$$

where the Z_j's are i.i.d. sign-valued zero-mean random variables. Apply the Hoeffding bound and the union bound on the probability of a large deviation of Q, we get the following bound for its expectation: $E(Q) \le \sqrt{2m^{-1} \log(2e|\mathcal{H}_c|)}$, where $|\mathcal{H}_c|$ is the number of distinct vectors $f(x_1^m)$ when f varies in $\mathcal{H}_c$. Apply the VC bound to this number and we get the proof. □

In Section 6 we showed that a nonzero a-span of the base hypothesis space implies an exponential reduction in the training error. In fact, we will also show that the training error is guaranteed to become *exactly* 0 *after some finite time*

for any data set, provided that the base hypothesis space has a nonzero a-span. [Assume $\text{prec}(\hat{f}) = 0$ for convenience.]

PROPOSITION 7 (Time needed for a perfect fit). *Let τ be the smallest time beyond which the training error is always 0 regardless of how the data set is labeled. Then there exists the following relationship between this time τ, and the VC dimension and the a-span of the base hypothesis space, given x_1^m:*

$$\frac{2m(L^*/8)^2 - 4}{\text{VC}(H)\log m} \le \tau \le \frac{2\log(m+1)}{\text{asp}_c(H; x_1^m)^2},$$

where $L^* = m^{-1}\sum_{j=1}^m \min\{p(Y=1|x_j), \text{and } p(Y=0|x_j)\}$ is the Bayes error.

PROOF. The lower bound is obtained by observing that, while the training error is 0, the prediction error is at least L^*. Then a VC bound over the combined hypothesis space for the difference of the training and prediction errors, which is $8\sqrt{(2m)^{-1}(4+\tau\text{VC}(H)\log m)}$, should be at least L^*. The upper bound is obtained by setting the upper bound of the training error to be $1/(m+1)$ and noting that at this time the training error actually needs to be exactly 0 since it values in $\{i/m\}_1^m$. □

Acknowledgments. The author thanks Martin Tanner for encouraging him to work on this research topic. He is also grateful to Leo Breiman for very informative e-mail exchange and to the referees and editors for their comments that have been very helpful in improving the presentation.

REFERENCES

ANTHONY, M. and BIGGS, N. (1992). *Computational Learning Theory: An Introduction*. Cambridge Univ. Press.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

BREIMAN, L. (1997a). Prediction games and arcing classifiers. Technical Report 504, Dept. Statistics, Univ. California, Berkeley.

BREIMAN, L. (1997b). Arcing the edge. Technical Report 486, Dept. Statistics, Univ. California, Berkeley.

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

BREIMAN, L. (1999). Using adaptive bagging to debias regressions. Technical Report 547, Dept. Statistics, Univ. California, Berkeley.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BÜHLMANN, P. and YU, B. (2000). Explaining bagging. Technical report, Dept. Statistics, Univ. California, Berkeley.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **121** 256–285.

FREUND, Y. and SCHAPIRE, R. E. (1996). Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual ACM Conference on Computational Learning Theory* 325–332. ACM Press, New York.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139.

FRIEDMAN, J. H. (1999a). Greedy function approximation: a gradient boosting machine. Technical report, Dept. Statistics, Stanford Univ.

FRIEDMAN, J. H. (1999b). Stochastic gradient boosting. Technical report, Dept. Statistics, Stanford Univ.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.* **28** 337–407.

GOLDMANN, M., HASTAD, J. and RAZBOROV, A. (1992). Majority gates vs. general weighted threshold gates. *Comput. Complexity* **2** 277–300.

GROVE, A. J. and SCHUURMANS, D. (1998). Boosting in the limit: maximizing the margin of learned ensembles. In *Proceedings of the 15th National Conference on Artificial Intelligence*. AAAI, Menlo Park, CA.

HAUSSLER, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. and Comput.* **100** 78–150.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87.

JIANG, W. (2000a). On weak base hypotheses and their implications for boosting regression and classification. Technical Report 00-01, Dept. Statistics, Northwestern Univ.

JIANG, W. (2000b). Process consistency for AdaBoost. Technical Report 00-05, Dept. Statistics, Northwestern Univ.

MALLAT, S. and ZHANG, S. (1993). Matching pursuit in a time-frequency dictionary. *IEEE Trans. Signal Processing* **41** 3397–3415.

MASON, L., BAXTER, J., BARTLETT, P. and FREAN, M. (1999). Boosting algorithms as gradient descent in function space. Technical report, Dept. Systems Engineering, Australian National Univ.

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

SCHAPIRE, R. E. (1999). Theoretical views of boosting. In *Computational Learning Theory. Lecture Notes in Comput. Sci.* **1572** 1–10. Springer, Berlin.

SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686.

VAPNIK, V. N. (1998). *Statistical Learning Theory.* Wiley, New York.

YANG, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284.

DEPARTMENT OF STATISTICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60208
E-MAIL: wjiang@northwestern.edu