Is Regularization Unnecessary for Boosting?

Wenxin Jiang
Department of Statistics, Northwestern University
Evanston, IL 60208, USA

1 Introduction

Boosting algorithms are often observed to be resistant to overfitting, to a degree that one may wonder whether it is harmless to run the algorithms forever, and whether regularization in on way or another is unnecessary [see, e.g., Schapire (1999); Friedman, Hastie and Tibshirani (1999); Grove and Schuurmans (1998); Mason, Baxter, Bartlett and Frean (1999)]. One may also wonder whether it is possible to adapt the boosting ideas to regression, and whether or not it is possible to avoid the need of regularization by just adopting the boosting device.

In this paper we present examples where 'boosting forever' leads to suboptimal predictions; while some regularization method, on the other hand, can achieve asymptotic optimality, at least in theory. We conjecture that this can be true in more general situations, and for some other regularization methods as well. Therefore the emerging literature on regularized variants of boosting is not unnecessary, but should be encouraged instead. The results of this paper are obtained from an analogy between some boosting algorithms that are used in regression and classification.

2 Framework for Boosting Regression and Classification

In statistical learning, we are faced with an observed data set $S = (X_i, Y_i)_1^n$, where X_1^n are *predictors*, which are assumed here to lie in $[0, 1]$ and take distinct values, (only) for convenience. We allow the *responses* Y_1^n to be random for potential noises of the data. It is noted that in the machine learning literature the Y_i's are usually fixed and the X_i's are random; while in statistics the Y_i's are invariably random, and the X_i's can be sometimes fixed and chosen by the researcher who collects the data. We call n the sample size. 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. Here, a hypothesis space $\mathcal{H}_{r,c}$ is a set of functions $f : [0, 1] \mapsto \Re$ or $f : [0, 1] \mapsto \{\pm 1\}$, respectively. A relatively simple 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 $\text{lin}^t(H_{r,c})$. Formally, $\text{lin}^t(H) = \{\sum_1^t \alpha_s f_s : (\alpha_s, f_s) \in \Re \times H\}$. A regression space $\mathcal{H}_r$ is said to *induce* a classifier space $\mathcal{H}_c$, if $\mathcal{H}_c = \text{sgn}(\mathcal{H}_r) = \{\text{sgn}(f) : f \in \mathcal{H}_r\}$.

In boosting, a cost function $C(F|S)$ is used, which depends on the sample S and is a functional of F, where $F \in \text{lin}^t(H_{r,c})$ for some t. A boosting algorithm acting on a base system $H_{r,c}$ minimizes $C(F|S)$ with respect to F sequentially and adaptively in a way similar to the following.

1. Let $\hat{F}_0 = 0$.
2. For all $t = 1, 2, ...,$
   - a. Let $\hat{\alpha}_t \hat{f}_t = \arg\min_{\alpha f \in \Re \times H_{r,c}} C(\hat{F}_{t-1} + \alpha f|S)$.
   - b. Let $\hat{F}_t = \hat{F}_{t-1} + \hat{\alpha}_t \hat{f}_t$.

For the regression case or the classification case, respectively, at step (or *time*) t, $\hat{F}_t(x)$ or $\text{sgn}(\hat{F}_t(x))$ form the prediction for a future response Y or Z when the predictor takes value x, $\forall x \in [0, 1]$.

This framework of boosting with a general cost function is similar to the ones used in Friedman et al. (1999) and Mason et al. (1999). Some examples we will consider include *LSBoost.Reg*, the least squares boosting algorithm for regression, where a square cost $C(F|S) = n^{-1} \sum_{i=1}^n \{Y_i - F(X_i)\}^2$ is used; and *AdaBoost*, the adaptive boosting algorithm for classification by Freund and Schapire (1997), where an exponential cost $C(F|S) = n^{-1} \sum_{i=1}^n e^{-Z_i F(X_i)}$ is used.

It is noted that LSBoost.Reg is essentially the same as the matching pursuit (MP) algorithm of Mallat and Zhang (1993) used in signal processing for sequential combination of waveforms, which is later recognized by Friedman et al. (1999) as an analog of AdaBoost for least squares regression. We will utilize this analogy in studying some theoretical properties of AdaBoost. Properties for LSBoost.Reg are often easier to derive due to the use of the square cost, which can be instructive for the study of AdaBoost. Other approaches of regression boosting include methods that involve discretization [see, for example, Freund and Schapire (1997); Ridgeway, Madigan and Richardson (1999)].

3 Orthogonal Boosting for Regression

First let us consider some very simple examples. Consider fixed predictors X_1^n chosen at a set of mutually distinct *design points* x_1^n. E.g., in the typical set up of nonparametric regression, $x_1^n = \{i/n\}_{i=1}^n$. In particular, we will first consider LSBoost.Reg operating on an *orthogonal base hypothesis space* $H_r = \{\phi_1, \phi_2, ..., \phi_n\}$, where the functions form an orthonormal basis of $\Re^n$ when evaluated at the design points x_1^n. Without confusion, we will use the functions to denote the vectors evaluated at the design points, as $\phi_1 = \phi_1(x_1^n)$, $\phi_2 = \phi_2(x_1^n)$, and so on. Then we have $\langle \phi_k, \phi_j \rangle = \delta_{kj}$ (using the Kronecker's δ). Here the inner product for two $n \times 1$ vectors $a = a_1^n$ and $b = b_1^n$ are defined to be $\langle a, b \rangle = n^{-1} \sum_{i=1}^n a_i b_i$ and the norm $||a|| = \langle a, a \rangle$.

The t-step boosted prediction of this system is exactly solvable. It basically retains the t largest sample Fourier coefficients. Let $y = y_1^n$ be the vector of observed responses. Then the *sample Fourier coefficients* are defined as $\tilde{\theta}_s = \langle \phi_s, y \rangle$, $s = 1, \ldots, n$. Denote $\tilde{\theta}_{\hat{j}_t}$ as the tth largest (in magnitude) of these n sample Fourier coefficients, so that $\tilde{\theta}^2_{\hat{j}_t} = \tilde{\theta}^2_{(t)}$, the t-th largest of $\{\tilde{\theta}^2_j\}_1^n$, and $\tilde{\theta}^2_{\hat{j}_1} > \tilde{\theta}^2_{\hat{j}_2} > \ldots > \tilde{\theta}^2_{\hat{j}_n}$. Assume there is not tie (this will happen with probability one if y is continuous).

It is straightforward to show that

Proposition 1 *(Expression of the Boosted Prediction). For $t \le n$, $\hat{F}_t = \sum_{k=1}^t \tilde{\theta}_{\hat{j}_k} \phi_{\hat{j}_k} \equiv \sum_{k=1}^t \langle y, \phi_{\hat{j}_k} \rangle \phi_{\hat{j}_k}$, where $\tilde{\theta}^2_{\hat{j}_1} > \tilde{\theta}^2_{\hat{j}_2} > \ldots > \tilde{\theta}^2_{\hat{j}_n}$. The prediction then stabilizes from that time on, i.e., $\hat{F}_t = \hat{F}_n$ for all $t > n$.*

How good is this prediction? Let us consider the following fixed-predictor regression problem $y_i = \mu(x_i) + \epsilon_i$, $i = 1, \ldots, n$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. To measure the goodness of a prediction $\hat{F}(\cdot)$, we define the *prediction error* for new responses as $L = n^{-1} \sum_{i=1}^n E\{y_i^{new} - \hat{F}(x_i)\}^2 \equiv E||y^{new} - \hat{F}||^2$. Note that $L = L^* + R(\hat{F}, \mu)$ where $L^* = \sigma^2$ is the prediction error of the optimal prediction μ pretending that the mean function is known, and $R(\hat{F}, \mu) = n^{-1} \sum_{i=1}^n E\{\hat{F}(x_i) - \mu(x_i)\}^2 \equiv E||\hat{F} - \mu||^2$ is the l_2 estimation *risk*. A good prediction should have $R \to 0$ or $L \to L^*$ as n increases, which is sometimes said to be *risk consistent.*

What happens to the prediction when letting the boosting algorithm run forever? In that case, the prediction becomes stabilized to be $\hat{F}_n$. By the completeness of the basis H_r, we get $\hat{F}_n(x_1^n) = y_1^n$. I.e., the limiting prediction uses the observations themselves! It is immediate to conclude that $L = 2\sigma^2$ and the *amount of overfit* $L - L^* = \sigma^2 \not\to 0$ however large the sample size n is. *The 'boosting-forever' predictor therefore is suboptimal asymptotically.* So in this case 'boosting-forever' without regularization overfits for the purpose nonparametric regression.

In this case we have a very special base hypothesis space that has a finite number of hypotheses which are mutually orthogonal. What happens to more general hypothesis spaces? Will 'boost-forever' be suboptimal also?

4 More General Base Hypothesis Spaces for Regression

Now consider a general base hypothesis space H_r, which may be nonorthogonal and may contain an uncountable number of hypotheses. Consider the typical situation of nonparametric regression with fixed x_1^n as before. Based on the concepts developed in matching pursuit, let us define a measure of capacity $\mathrm{asp}(H_r;\ x_1^n)$ called the *(regression) angular span*:

$$\mathrm{asp}(H_r;\ x_1^n) = \inf_{\epsilon \in \Re^n,\ ||\epsilon||=1} \ \sup_{f \in H_r, ||f||>0} \langle \epsilon,\ f/||f|| \rangle^2.$$

[For example for the orthogonal base space in the previous section, we have $\mathrm{asp}(H_r;\ x_1^n) = 1/n$.] Then as a special case of Jiang (1999, Proposition 2), we have the following result.

Proposition 2
(Prediction Error for Regression Boosting). Suppose $\mathrm{asp}(H_r) > 0$ for the base hypothesis space H_r used in LSBoost.reg. Consider the prediction error for the prediction $\hat{F}_t$ obtained from t rounds of LSBoost.Reg: $L_t = E||y^{new} - \hat{F}_t||^2$. Then we have

$$\left|\sqrt{L_t} - \sqrt{2\sigma^2}\right| \le \sqrt{n^{-1} \sum_{i=1}^n \{\mu(x_j)^2 + \sigma^2\}}\ \exp\{-\mathrm{asp}(H_r) t/2\}. \quad (1)$$

If $\mathrm{asp}(H_r) > 0$, the bounds of the proposition are nontrivial and suggest that running the unmodified LSBoost.Reg forever will still let the prediction error converge to the suboptimal limit $2\sigma^2$, and the amount of overfit will be σ^2.

It is noted that most of the commonly used base hypothesis spaces do have a nonzero angular span. This is due to the following lemma that combines Lemmas 1 and 2 of Jiang (1999).

Lemma 1 *(Approximation and Angular Span). 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 $\mathrm{sgn}(x-a)$ at all points $x \neq a$. Then, for any set of distinct design points x_1^n, we have $\mathrm{asp}(H_r;\ x_1^n) > 0$.*

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) $\{2F\{(\cdot - \mu)/\sigma\} - 1 : \sigma > 0, \mu \in \Re\}$. Examples include the case when F is the logistic cdf, when the q-combined system is the usual neural nets with q (tanh) nodes; the case when F is the Gaussian cdf; the 'stumps' 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 et al. (1984)]. By the consequences of the previous results, we see that all these base systems, even the ones as simple as the 'stumps', will unavoidably lead to suboptimal predictions when boosted forever, due to the nonzero angular span. The exponential decay suggests that a typical time used to approach the limit may be of order $1/\mathrm{asp}_c(H)$, which is of order n for the case of orthogonal base systems.

Similar results also exist for the classification case.

5 Boosting Forever in Classification

Consider a general classification base hypothesis space H_c and consider a situation with fixed x_1^n and independent random binary responses y_1^n that have mean μ_1^n. Analogous to the regression case, we define a measure of capacity $\mathrm{asp}_c(H_c;\ x_1^n)$ called the *(classification) angular span*:

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

where $P^n = \{w_1^n :\ w_j \geq 0, \sum_1^n w_j = 1\}$. [For example, for the 'stumps' $H_c = \{s \cdot \mathrm{sgn}_a :\ s \in \{\pm 1\},\ a \in \Re\}$, where $\mathrm{sgn}_a(x) = 2I\{x \geq a\} - 1$, we have $2/n \geq \mathrm{asp}_c(H_c) \geq 1/n$ for any set of n mutually distinct design points.]

Then as a special case of Jiang (1999, Proposition 5), we have the following result.

Proposition 3 *(Prediction Error for Classification Boosting). Denote $\mathrm{asp}_c(H_c)$ as the angular span of the base hypothesis space used in AdaBoost. Suppose $\mathrm{asp}_c(H_c) > 0$. Consider the prediction error L_t for the prediction $\hat{Y}_t \equiv (1 + \mathrm{sgn} \circ \hat{F}_t)/2$ obtained from t rounds of AdaBoost: $L_t \equiv n^{-1} E \sum_{i=1}^n \{\hat{Y}_t(x_i) - y_i^{new}\}^2 = n^{-1} \sum_{i=1}^n P\{\hat{Y}_t(x_i) \neq y_i^{new}\}$. Then we have*

$$L_t \leq L_\infty + A \text{ and } \sqrt{L_t} \geq \sqrt{L_\infty} - \sqrt{A}. \tag{2}$$

Here

$$\begin{aligned} A &= \exp\{-\mathrm{asp}_c(H_c)^2 t/2\}, \\ L_\infty &= n^{-1} \sum_{j=1}^{n} 2\mu_j(1-\mu_j) \equiv L^* + R, \end{aligned}$$

where $L^* = n^{-1} \sum_{i=1}^n \min(\mu_i,\ 1-\mu_i)$ is the (optimal standard) Bayes error, and $R = n^{-1} \sum_{i=1}^n 2|\mu_i - 1/2| \min(\mu_i,\ 1-\mu_i)$ measures the difference $L_\infty - L^*$.

If $\mathrm{asp}_c(H_c) > 0$, the bounds of the proposition are nontrivial and suggest that running the unmodified AdaBoost forever will still let the prediction error converge to the suboptimal limit L_∞ (the 'nearest neighbor' error), with the amount of overfit $R \in [0,\ \min(0.125, L^*)]$.

It is noted that most of the commonly used base systems do have a nonzero asp_c, even very simple ones such as the stumps. This is due to the following proposition [Jiang (1999), Proposition 3].

Proposition 4 *(Approximation and Angular Span). $H_c = \mathrm{sgn}(H_r)$ and H_r can approximate any sign function (see Lemma 1) imply that $\mathrm{asp}_c(H_c;\ x_1^n) > 0$ for any set of (mutually distinct) design points x_1^n.*

(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.)

Due to Remark 1 and Proposition 4, most of the commonly used base systems do have nonzero asp_c, and Proposition 3 above suggests that running the unmodified AdaBoost forever will still let the prediction error converge to the suboptimal limit L_∞ (the 'nearest neighbor' error), with the amount of overfit $R \in [0,\ \min(0.125, L^*)]$. Comparing to the case of regression boosting, one difference here is that the amount of overfit cannot be arbitrarily large and is small for data with a low noise level L^*. Another difference is that it may take longer to approach the overfitting limit here than in the regression case. The exponential decay suggests that a typical time used to approach the limit may be of order $1/\mathrm{asp}_c(H_c)^2$, which is of order n^2 when stumps are used in boosting.

The worst amount of overfit occurs when the probability μ_i's are 0.25 or 0.75. In this case the Bayes error L^* is 0.25, while the 'boost-forever' approach has prediction error converging to $L_\infty = 0.375$. The difference does not disappear as the sample size n increases. Therefore boosting forever will lead to a suboptimal prediction for noisy data.

These results are only for fixed x. What happens to more general cases with random x?

a. In fact, with random continuous predictors on $[0, 1]$, in a case of boosting the decision stumps or CART systems, the boosted solutions are easily found to be nonunique. One typical way to ensure a unique solution is to limit the cuts of the step functions to be located at the mid-data points. Jiang (2000) shows that in this case AdaBoost will also generate the nearest neighbor rule for all time $t \geq 2n^2 \log(n + 1)$. Therefore similar overfitting behavior can occur for noisy data.

b. The current method does not provide a general result for the most realistic case with high dimensional random continuous predictors. It is only in this case, where it is possible that the prediction error of AdaBoost continuous 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 in this paper, the prediction error stabilizes simultaneously with the training error.* The best explanations so far for this mystery seem to be the margin approach by Schapire et al. (1998) and the top approach by Breiman (1997), which are still semi-empirical in nature. It is, however, plausible to conjecture that even in the case of higher dimensional data running AdaBoost *forever* can still lead to a suboptimal prediction which does not 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); Friedman et al. (1999)].

It is therefore foreseeable that some kind of regularization may improve the performance in noisy situations. Below we will show some examples where regularization of some kind can be provably beneficial, at least theoretically. We first consider the regression case with orthogonal hypotheses — this is a pleasant situation where many analytic results can be derived which can be instructive.

6 Regularization in the Orthogonal Boosting for Regression

Regarding the discussions in the previous sections, we see that it may not be desirable to let the unmodified boosting algorithm run forever, because it can overfit. *Therefore regularization in one form or another may not be unnecessary.* An interesting question is: how low can the boosted prediction error be throughout the process of boosting? Also: how can we find the best number of steps that achieves the lowest prediction error? What are some regularization methods to prevent overfitting? To gain some experience with these questions, we will consider the very easy regression boosting system in Section 3 and establish the relation and utilize the extensive results of orthogonal series regression and minimax theory.

Consider the nonparametric regression set-up as in Section 3. Note that $\mu(x_i) = Ey_i$ can be expanded in the orthonormal basis $H_r = \{\phi_1, \ldots, \phi_n\}$ as $\mu = \sum_{j=1}^n \theta_j \phi_j$ (where $\langle \phi_j, \phi_k \rangle = \delta_{jk}$), where $\mu = \mu(x_1^n)$, $\phi_j = \phi_j(x_1^n)$, $\theta_j = \langle \phi_j, \mu \rangle$. Call the θ_j's as the *Fourier coefficients* of μ. Recall that the *sample Fourier coefficients* were defined as $\tilde{\theta}_t = \langle \phi_t, y \rangle$ where $y = y_1^n$. Then $y = \sum_{j=1}^n \langle \phi_j, y \rangle \phi_j$ and $||y||^2 = \sum_{j=1}^n \langle \phi_j, y \rangle^2$.

As before, define $\tilde{\theta}_{\hat{j}_t} \in \{\tilde{\theta}_1^n\}$ so that $\tilde{\theta}^2_{\hat{j}_t} = \tilde{\theta}^2_{(t)}$, the t-th largest of $\{\tilde{\theta}_j^2\}_1^n$, and $\tilde{\theta}^2_{\hat{j}_1} > \tilde{\theta}^2_{\hat{j}_2} > \ldots > \tilde{\theta}^2_{\hat{j}_n}$. Assume there is not tie (this will happen with probability one if y is continuous).

Consider LSBoost.Reg applied to H_r. Now change the notation $\hat{F}_t$ to be $\hat{\mu}_t$ for the boosted prediction at stage t, since it is also an estimator of the mean response μ. Let $\hat{\mathcal{J}}_t = \{\hat{j}_1, \ldots, \hat{j}_t\}$, then the expression of the boosted prediction in Section 3 can be expressed as $\hat{\mu}_t = \sum_{j \in \hat{\mathcal{J}}_t} \tilde{\theta}_j \phi_j = \sum_{j=1}^n \tilde{\theta}_j \phi_j I[j \in \hat{\mathcal{J}}_t]$. We are interested in studying the l_2 risk $R(\hat{\mu}_t, \mu) = E||\hat{\mu}_t - \mu||^2$, which measures the goodness of $\hat{\mu}_t$ as an estimator of μ, and also measures the 'amount of overfit' $L_t - L^*$ of $\hat{\mu}_t$, as a prediction of future responses when comparing to the optimal standard L^* (or σ^2).

We first consider a 'parametric'-family of signals. Suppose that there are a finite number of θ_j's being nonzero. I.e., let $\mathcal{J} = \{j : |\theta_j| > 0, j = 1, \ldots, n\}$ and $|\mathcal{J}|$ denote its cardinality; we suppose that $|\mathcal{J}| = J < n$ and is independent of n. Call the corresponding family of μ's ($\mu = \sum_{i=1}^{n} \theta_j \phi_j$) the *$J$-sparse family* of signals M_J. Also define the family $\mathcal{M}_{J_m} = \cup_{J=0}^{J_m} M_J$ of signals with at most J_m nonzero Fourier coefficients, called the *J_m-maximal sparse family.* Then we can prove the following two propositions, which indicate that one of the boosted fits is asymptotically (for large n) minimax-optimal for the family $\mathcal{M}_{J_m}$ of signals. [The second proposition follows from the results in Efromovich (1999, Sec. 7.1).]

Proposition 5 *(Risk Upperbound with Sparse Signals).*
$\sup_{\mu \in \mathcal{M}_{J_m}} R(\hat{\mu}_{|\mathcal{J}|}, \mu) = \left(\frac{J_m}{n}\right) \sigma^2 \{1 + o_n(1)\}$; *and* $\sup_{\mu \in \mathcal{M}_{J_m}} \inf_{t \in \{0,\ldots,n\}} R(\hat{\mu}_t, \mu) = \left(\frac{J_m}{n}\right) \sigma^2 \{1 + o_n(1)\}$.

Proposition 6 *(Minimax Lowerbound with Sparse Signals).*
$\inf_{\hat{\mu}} \sup_{\mu \in \mathcal{M}_{J_m}} R(\hat{\mu}, \mu) = \left(\frac{J_m}{n}\right) \sigma^2 \{1 + o_n(1)\}$.

Next, we relate the boosted fits to the fits obtained from hard thresholding, and derive risk bounds for arbitrary signals. We first note that one of the boosted estimator $\hat{\mu}_t$'s beats *any* hard thresholding estimator $\hat{\mu}^h(\delta) = \sum_{j=1}^{n} \tilde{\theta}_j \phi_j I[|\tilde{\theta}_j| > \delta]$, and in particular it beats the universal hard thresholding with $\delta = \sigma\sqrt{2 \log n / n}$. (Assume that the noise variance σ^2 is known for now.) This is because, for all $\delta \geq 0$, $\hat{\mu}^h(\delta) = \sum_{j=1}^{n} \tilde{\theta}_j \phi_j I[j \in \hat{\mathcal{J}}_t] = \hat{\mu}_t$ if $\delta^2 \in [\tilde{\theta}^2_{(t+1)}, \tilde{\theta}^2_{(t)})$, where $0 \equiv \tilde{\theta}^2_{(n+1)} < \tilde{\theta}^2_{(n)} = \tilde{\theta}^2_{\hat{j}_n} < \tilde{\theta}^2_{(n-1)} = \tilde{\theta}^2_{\hat{j}_{n-1}} < \ldots < \tilde{\theta}^2_{(1)} = \tilde{\theta}^2_{\hat{j}_1} < \tilde{\theta}^2_{(0)} \equiv \infty$; $\hat{\mathcal{J}}_0 = \{\mathit{empty}\}$, $\hat{\mathcal{J}}_t = \{\hat{j}_1, \ldots, \hat{j}_t\}$ for $t = 1, \ldots, n$. Therefore, for any $\delta > 0$, we have $\hat{\mu}^h(\delta) \in \{\hat{\mu}_t :\ t = 0, 1, ..., n\}$, and

Proposition 7 *(Boosting vs Hard Thresholding).*
$\inf_{t \in \{0,\ldots,n\}} R(\hat{\mu}_t, \mu) \leq R\{\hat{\mu}^h(\delta), \mu\}$, $\forall\ \delta > 0$, $\forall\ \mu$, $\forall\ R(\cdot,\cdot)$.

Take, now, $\delta = \sigma\sqrt{2 \log n / n}$, and apply Theorem 4 or Corollary 1 of Donoho and Johnstone (1994), we get

Proposition 8 *(Risk Bounds with Arbitrary Signals).*
(i) For all $\mu \in \Re^n$, we have

$$\frac{\inf_{t \in \{0,\ldots,n\}} R(\hat{\mu}_t, \mu)}{\sigma^2/n + \sum_{i=1}^{n} \min(\langle \phi_i, \mu \rangle^2, \sigma^2/n)} = 2 \log n \{1 + o_n(1)\};$$

(ii)

$$\inf_{\mathit{all}\ \hat{\mu}} \sup_{\mathit{all}\ \mu} \frac{R(\hat{\mu}, \mu)}{\sigma^2/n + \sum_{i=1}^{n} \min(\langle \phi_i, \mu \rangle^2, \sigma^2/n)} = 2 \log n \{1 + o_n(1)\}.$$

They are minimax results for guarding against *all* possible μ, not just the 'sparse' ones; together with a result that 'one of the boosted estimator' achieves asymptotic minimax optimality.

Now the question is: which $\hat{\mu}_t$ to use? Note that it is not wise to let the boosting algorithm run forever, since $\hat{\mu}_t = y$ for all $t \geq n$, and $R(y, \mu) = ||y - \mu||^2 = \sigma^2 \not\to 0$. Therefore $\hat{\mu}_\infty$ is not even risk consistent, for any μ.

Which $\hat{\mu}_t$ is a good one to choose? For sparse signals, we'd like $\hat{\mu}_t = \hat{\mu}_{|\mathcal{J}|}$, but usually we do not know what $|\mathcal{J}|$, the number of nonzero Fourier coefficients, is. One idea is to use the thresholding techniques to kill all $\tilde{\theta}_j$, $j \notin \mathcal{J}$. By the normal extremal theory, we can use the familiar *universal threshold* $\delta = \sigma\sqrt{2 \log n / n}$. Then with probability tending to one, we have $|\tilde{\theta}_j| > \delta$ if and only if $j \in \mathcal{J}$. If we threshold the coefficients during the boosting process and keep only those with magnitudes exceeding δ, then at the tth step we get a *thresholded* estimator $\hat{\mu}_t(\delta) = \sum_{j \in \hat{\mathcal{J}}_t} \tilde{\theta}_j \phi_j I[|\tilde{\theta}_j| > \delta]$. Then

Proposition 9 *(Risk Upperbound when Boosting with Universal Threshold). For $t \geq t_\infty \equiv \sum_{i=1}^{n} I[|\tilde{\theta}_j| > \delta]$, $\hat{\mu}_t(\delta) = \hat{\mu}_{t_\infty}(\delta) = \sum_{j \in \hat{\mathcal{J}}_t} \tilde{\theta}_j \phi_j I[|\tilde{\theta}_j| > \delta] = \hat{\mu}^h(\delta)$, and*

$$\frac{R(\hat{\mu}_t(\delta), \mu)}{\sigma^2/n + \sum_{i=1}^{n} \min(\langle \phi_i, \mu \rangle^2, \sigma^2/n)} = 2 \log n \{1 + o_n(1)\},$$

achieving asymptotic minimax optimality.

In practice, one can simply use $\hat{\mu}_t(\delta)$ for t being so large that $||y - \hat{\mu}_t(\delta)||^2$ becomes stable.

Remark 2 *(Estimating σ).* See Vidakovic (1999, Sec. 6.6.2).
Usually signal-to-noise ratio (SNR) is very small for $\tilde{\theta}_{\hat{j}_t}$'s with $t > n/2$. then we can let the estimator of σ be $\hat{\sigma} = \sqrt{(n/2 - 1)^{-1} \sum_{t=n/2+1}^{n} (\tilde{\theta}_{\hat{j}_t} - \bar{\tilde{\theta}}_{(n/2,n]})^2}$; where the appropriate integer part is taken for $n/2$; $\bar{\tilde{\theta}}_{(n/2,n]}$ represents the average of $\tilde{\theta}_{\hat{j}_t}$'s for the second half of t's; $\tilde{\theta}^2_{\hat{j}_t} = \tilde{\theta}^2_{(t)}$; $\tilde{\theta}_s = \langle y, \phi_s \rangle$ is a sample Fourier coefficient.

In summary we have shown the following for this system of orthogonal hypotheses used in LSBoost.Reg:

1. The boosted prediction at any time is exactly solvable and basically retains the largest sample Fourier coefficients. The prediction reaches the overfitting limit at a finite time equal to the sample size.

2. For a 'sparse family' of signals (mean responses) having at most a finite number of nonzero Fourier coefficients, the boosted prediction at some time

is asymptotically minimax in reducing the prediction error in this family. In other words, no measurable estimator can beat all boosted predictions simultaneously for all sparse signals.

3. The boosted prediction at some time is 'risk consistent', in the sense that the l_2 risk or the mean square error from the actual mean response converges to zero for large sample size, for sparse signals. The lowest prediction error converges at the parametric rate. On the other hand, the 'unboosted' prediction or the 'boost-forever' prediction are not 'risk consistent' in general.

4. For more general signals that may not be 'sparse', we note that there is a boosted prediction at some time which is at least as good as an 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. 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.* In practice, the boosted prediction at an optimal time can be effectively found by applying hard thresholding.

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

5. If the threshold is chosen to be the universal threshold [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 prediction is asymptotically (minimax-)optimal among all possible predictions when protecting against all possible signals.

The bottom line is that there are cases where it is provable that 'boosting forever' is not desirable and that regularization is still necessary and beneficial for regression boosting. What about the classification case? Is it possible to show the existence of some regularization method, at least in theory, which will lead to an improvement over the nearest neighbor-type performance that can be achieved by 'AdaBoost-forever'?

7 Quantization for AdaBoost

For general sparse data with random predictors on $[0,1]^d$, say, at least theoretically it is possible to 'quantize' the data to make them 'crowded' as a method of regularization and prevent overfitting in the large time limit of AdaBoost. The quantization involves grouping together the nearby data points in a cell of volume h^d to use a common predictor value or design point ξ_k, say. The large time prediction obtained by boosting overfits less since the limiting rule can be shown to be similar to the 'histogram rule', which uses the majority of the *labels* (or responses) at a design point as the prediction, which becomes less noisy when more labels / responses fall on the same design point. Let $m = 1/h^d$ which we assume to be an integer for convenience, and let $[0,1]^d = \cup_{k=1}^m A_1^k$ being partitioned into m cells A_1^m, and let $\xi_k \in A_k$ be any design point that represents the cell. Define $N(\pm|k) = \sum_{i=1}^n I[x_i \in A_k,\ z_i = \pm 1]$, as the number of positive (or respectively, negative) observations falling in A_k. In the quantized version, Adaboost now uses the new quantized data set $(\xi_{k(i)}, z_i)_{i=1}^n$, where $\xi_{k(i)}$ is the design point for the cell that contains x_i. Also, at the tth step of AdaBoost, the t-combined fit $\hat{F}_t(x)$ is now quantized to be $\hat{F}_t(\xi_k)$, if $x \in A_k$, say. Then we let $\hat{Z}_t = sgn \circ \hat{F}_t$ be the corresponding prediction.

Proposition 10 *('Quantization' vs Histogram Rule). Suppose* $\mathrm{asp}_c(H_c;\ \xi_1^m) > 0$ *for the base hypothesis space H_c used in AdaBoost. Then we have the following results for the quantized AdaBoost:*

(i). For any x in any nonempty cell A_k [where $N(+|k) + N(-|k) > 0$], we have

$$\lim_{t\to\infty} \hat{F}_t(x) = \frac{1}{2} \log\left\{\frac{N(+|k)}{N(-|k)}\right\},$$

which is proportional to the sample estimate of the log odds in A_k.

(ii). For any x in any 'unbalanced' cell A_k with an unequal number of positive and negative labels [where $N(+|k) \neq N(-|k)$], we have

$$\lim_{t\to\infty} \hat{Z}_t(x) = sgn\{N(+|k) - N(-|k)\},$$

which gives the histogram majority rule.

(iii). For any x in any 'unbalanced' cell A_k, the limiting prediction is realized at a finite time; that is, there exists a finite time τ such that $\hat{Z}_t(x) = \lim_{s\to\infty} \hat{Z}_s(x)$ for all $t \geq \tau$.

Define the prediction error (with expectation also running over the random predictor X) as $L_t \equiv E\{Y^{new} -$

$\hat{Y}_t(X)\}^2 = P\{Z^{new} \neq \hat{Z}_t(X)\}$, for the step-$t$ prediction from the quantized AdaBoost. Suppose $P(Z = 1|x)$ is a smooth function of x with bounded derivatives. Then when $h \sim n^{-1/(d+2)}$, the limiting prediction error is bounded by

$$\lim_{t\to\infty} L_t - L^* \leq (constant) * n^{-1/(d+2)}\{1 + o_n(1)\},$$

by adapting the known results on quantization [see, e.g., Devroye et al. (1996), Chapter 27]. This leads to a 'consistent' large time prediction that performs close to the Bayes rule for large sample sizes. The result holds for high dimensional sparse data and the $O(n^{-1/(d+2)})$-rate of convergence of the large time prediction error is actually minimax optimal, i.e., no other prediction can be uniformly superior over this family of smooth signals [see, e.g., Yang (1999)].

The point is that *with this regularization treatment*, boosting forever leads to a difference $L_\infty - L^*$ that decreases to zero for large sample size n, and the rate of decrease is optimal over this family of smooth signals. So asymptotically there will be no overfit. On the other hand, the prediction error obtained from running the unmodified AdaBoost forever is often comparable to the nearest neighbor rule, as we commented in earlier sections, which are suboptimal in the sense that the amount of overfit $L_\infty - L^*$ generally does not decrease to zero for any n.

Therefore we established the existence of an example where regularization of AdaBoost leads to provable asymptotic improvement, at least theoretically. We are not recommending the quantization method as a practical regularization tool. Again, the point is that even though AdaBoost seems to be resistant to overfitting, running it forever without regularization can still lead to a suboptimal solution; *there is still room for improvement achievable by regularization.* We suspect that superior performance of some other regularized versions of boosting may also be established as compared to the unmodified one, which may even be empirically supported as well—see, e.g., empirical work with shrinkage and randomization [e.g., Friedman (1999a, b)]; complexity penalty [e.g., Mason et al., (1999)]; bagged versions of boosting [e.g., Breiman (1996), (1999); Brühlmann and Yu (2000)].

8 Brief Conclusions

Despite their resistance to overfitting, it is often undesirable to run boosting algorithms forever without any kind of regularization— this could lead to asymptotically suboptimal predictions. There is still room for improvement achievable by regularization, at least asymptotically. New studies on regularization of boosting are not unnecessary, and should be encouraged instead, whether they use the method of stopping at an optimal time, or randomization, or shrinkage, or thresholding, or quantization, or complexity penalty.

Lastly, there have been efforts in the literature on finding analogs of boosting in regression. They are probably motivated by the good performance of AdaBoost in classification and these efforts can certainly be fruitful. However, we suspect that one still cannot hope to avoid regularization in one way or another in regression, by just adopting the boosting device— regularization probably is more important and necessary in regression.

Acknowledgments

The author wishes to thank Leo Breiman for very useful email exchanges.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Breiman, L. (1997). Prediction games and arcing classifiers. *Technical Report 504, Statistics Department, University of California at Berkeley.*

Breiman, L. (1999). Using adaptive bagging to debias regressions. *Technical Report 547, Statistics Department, University of California at Berkeley.*

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Brühlmann, P. and Yu, B. (2000). Explaining bagging.
Technical Report, Statistics Department, University of California at Berkeley. (Downloadable at http://www.stat.berkeley.edu/users/binyu /publications.html.)

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 via wavelet shrinkage. *Biometrika*, **81**, 425-455.

Efromovich, S. (1999). *Nonparametric Curve Estimation.* Springer, New York.

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

Friedman, J. H.. (1999a). Greedy function approximation: a gradient boosting machine. *Technical Report, Department of Statistics, Stanford University.*

Friedman, J. H.. (1999b). Stochastic gradient boosting. *Technical Report, Department of Statistics, Stanford University.*

Friedman, J., Hastie, T. and Tibshirani, R. (1999). Additive logistic regression: a statistical view of boosting. *Technical Report, Department of Statistics, Stanford University.*

Grove, A. J. and Schuurmans, D. (1998). Boosting in the limit: maximizing the margin of learned ensembles. *Proceedings of the Fifteenth National Conference on Artificial intelligence (AAAI-98)*, Madison, WI, July 1998.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp.* **3**, 79-87.

Jiang, W. (1999). On weak base hypotheses and their implications for boosting regression and classification. *Technical Report, Department of Statistics, Northwestern University.* (Revised on 10/31/2000, downloadable at http://neyman.stats.nwu.edu/jiang/boost /boost.largetime2.ps.)

Jiang, W. (2000). Does boosting overfit: views from an exact solution. *Technical Report, Department of Statistics, Northwestern University.* (Downloadable at http://neyman.stats.nwu.edu/jiang/boost /boost.onedim.ps.)

Jordan, M. I., and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.* **6**, 181-214.

Mallat, S. and Zhang, S. (1993). Matching pursuit in a time-frequency dictionary. *IEEE Transactions on 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, Department of Systems Engineering, Australian National University.*

Ridgeway, G., Madigan, D. and Richardson, T. (1999). Boosting Methodology for regression problems. *The Seventh International Workshop on Artificial Intelligence and Statistics (Uncertainty '99)*, January 1999, Fort Lauderdale, FL, 152-161.

Schapire, R. E. (1999). Theoretical views of boosting. *Computational Learning Theory: Fourth European Conference, EuroCOLT'99*, 1-10.

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, **26** 1651-1686.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley, New York.

Yang, Y. (1999). Minimax nonparametric classification—Part I: rates of convergence. *IEEE Trans. Info. Theory*, **45**, 2271-2284.