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 one way or another is unnecessary [see, e.g., Schapire (1999); Friedman, Hastie and Tibshirani (1999); Grover 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 $S = (X_i, Y_i)^n_{i=1}$, where $X_i^n$ are predictors, which are assumed here to lie in $[0, 1]$ and take distinct values, (only) for convenience. We allow the responses $Y_i^n$ 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 transformation $Z_i = 2Y_i - 1$ 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 $\{+1\}$ valued classification functions $\mathcal{H}_c$ to fit the data. Here, a hypothesis space $\mathcal{H}_{r,c}$ is a set of functions $f : [0, 1] \mapsto \mathbb{R}$ or $f : [0, 1] \mapsto \{\pm 1\}$, respectively. A relatively simple hypothesis space, called the base hypothesis space or base system $\mathcal{H}_{r,c}$, can be made more complex by linear combinations of $t$ members as the $t$-combined system or $t$-combined hypothesis space denoted as $\text{lin}^t(\mathcal{H}_{r,c})$. Formally, $\text{lin}^t(\mathcal{H}) = \{\sum_i^t \alpha_i f_i : (\alpha_i, f_i) \in \mathbb{R} \times \mathcal{H}\}$. 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 $F \in \text{lin}^t(\mathcal{H}_{r,c})$ for some $t$. A boosting algorithm acting on a base system $\mathcal{H}_{r,c}$ 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 $t = 1, 2, \ldots$,
   a. Let $\alpha_t \hat{f}_t = \arg \min_{\alpha f \in \mathcal{H}_{r,c}} C(\hat{F}_{t-1} + \alpha f|S)$.
   b. Let $\hat{F}_t = \hat{F}_{t-1} + \alpha_t \hat{f}_t$.

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 $\text{LSBoost-Reg}$, the least squares boosting algorithm for regression, where a square cost $C(F|S) = n^{-1} \sum^n_{i=1} \{Y_i - F(X_i)\}^2$ is used; and $\text{AdaBoost}$, the adaptive boosting algorithm for classification by Freund and Schapire (1997), where an exponential cost $C(F|S) = n^{-1} \sum^n_{i=1} e^{-Z_i F(X_i)}$ 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 $X_i^n$ chosen at a set of mutually distinct design points $x_i^n$. E.g., in the typical setup of nonparametric regression, $x_i^n = (i/n)^{n-1}$. 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 $\mathbb{R}^n$ when evaluated at the design points $x_i^n$. Without confusion, we will use the functions to denote the vectors evaluated at the design points, as $\phi_1 = \phi_1(x_i^n)$, $\phi_2 = \phi_2(x_i^n)$, and so on. Then we have $\langle \phi_k, \phi_j \rangle = \delta_{k,j}$ (using the Kronecker’s $\delta$). Here the inner product for two $n \times 1$ vectors $a = a_i^n$ and $b = b_j^n$ 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 $y = y_i^n$ be the vector of observed responses. Then the sample Fourier coefficients are defined as $\theta_i = \langle \phi_i, y \rangle$, $s = 1, ... n$. Denote $\theta_{i_t}$ as the $t$th largest (in magnitude) of these $n$ sample Fourier coefficients, so that $\hat{\theta}_{i_t} = \theta_{i_t}$, the $t$-th largest of $\{\hat{\theta}_i^2\}_{i=1}^{t}$, and $\hat{\theta}_{i_1}^2 > \hat{\theta}_{i_2}^2 > ... > \hat{\theta}_{i_t}^2$. 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 $t \leq n$, $\hat{F}_t = \sum_{k=1}^{t} \hat{\theta}_{i_k} \phi_{i_k}$ is a function of $x_i^n$ at the design points $x_i^n$. 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 $y_i = \mu(x_i) + \epsilon_i$, $i = 1, ... n$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. To measure the goodness of a prediction $\hat{F}(\cdot)$, we define the prediction error for new responses as $L = n^{-1} \sum_{i=1}^{n} E\{y_i^{\text{new}} - \hat{F}(x_i)^2\} \equiv E\|y - \hat{F}\|^2$. Note that $L = L^* + R(\hat{F}, \mu)$ where $L^* = \sigma^2$ is the prediction error of the optimal predictor $\mu$ 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 this case, the prediction becomes stabilized to be $\hat{F}_n$. By the completeness of the basis $H_r$, we get $\hat{F}_n(x_i^n) = y_i^n$. I.e., the limiting prediction uses the observations themselves! It is immediate to conclude that $L = 2\sigma^2$ and the amount of overfitting $L - L^* = \sigma^2 / \sigma^2$ 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 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 $x_i^n$ as before. Based on the concepts developed in matching pursuit, let us define a measure of capacity $\text{asp}(H_r, x_i^n)$ called the (regression) angular span:

$$\text{asp}(H_r, x_i^n) = \inf_{\epsilon \in \mathbb{R}^n} \sup_{f \in H_r, ||f|| = 1} \langle \epsilon, f \rangle / ||f||.$$

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

**Proposition 2 (Prediction Error for Regression Boosting).** Suppose $\text{asp}(H_r) > 0$ 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: $L_t = E\|y^{\text{new}} - \hat{F}_t\|^2$. Then we have

$$\sqrt{L_t - 2\sigma^2} \leq \sqrt{n^{-1} \sum_{i=1}^{n} (\mu(x_i)^2 + \sigma^2) \exp\{-\text{asp}(H_r)t/2\}.$$

(1)
If $\text{asp}(H_r) > 0$, the bounds of the proposition are nontrivial and suggest that running the unmodified LSBoost forever will still let the prediction error converge to the suboptimal limit $2\sigma^2$, and the amount of overfit will be $\sigma^2$.

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 $\text{sgn}(x - a)$ at all points $x \neq a$. Then, for any set of distinct design points $x_i^n$, we have $\text{asp}(H_r; x_i^n) > 0$.

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) $\{2F\{-\mu}\sigma - 1 : \sigma > 0, \mu \in \mathbb{R}\}$. 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 $1/\text{asp}(H_r)$, 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 $x_i^n$ and independent random binary responses $y_{it}$ that have mean $\mu_{it}$. Analogous to the regression case, we define a measure of capacity $\text{asp}_c(H_c; x_i^n)$ called the (classification) angular span:

$$\text{asp}_c(H_c; x_i^n) = \inf_{w_i^n \in \mathbb{R}^n, z_i^n \in \{\pm 1\}^n} \sup_{f \in H_c} \left| \sum_{j=1}^{n} w_j z_j f(x_j) \right|,$$

where $P^n = \{w_i^n : w_j \geq 0, \sum_{j=1}^{n} w_j = 1\}$. For example, for the ‘stumps’ $H_c = \{s \cdot \text{sgn}_a : s \in \{\pm 1\}, a \in \mathbb{R}\}$, where $\text{sgn}_a(x) = 2I\{x \geq a\} - 1$, we have $2/n \geq \text{asp}_c(H_c) \geq 1/n$ 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 $\text{asp}_c(H_c)$ as the angular span of the base hypothesis space used in AdaBoost. Suppose $\text{asp}_c(H_c) > 0$. Consider the prediction error $L_t$ for the prediction $\hat{Y}_t \equiv (1 + \text{sgn}(F_t))F_t/2$ obtained from $t$ rounds of AdaBoost: $L_t = \sum_{i=1}^{n} (\hat{Y}_t(x_i) - y_{i\text{new}})^2 = \sum_{i=1}^{n} P(\hat{Y}_t(x_i) \neq y_{i\text{new}})$. Then we have

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

Here

$$A = \exp\{-\text{asp}_c(H_c)^2 t/2\},$$

$$L_\infty = n^{-1} \sum_{i=1}^{n} 2\mu_j (1 - \mu_j) \equiv L^* + R,$$

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

If $\text{asp}_c(H_c) > 0$, 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 $L_\infty$ (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 $\text{asp}_c$, 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). $H_c = \text{sgn}(H_r)$ and $H_r$ can approximate any sign function (see Lemma 1) imply that $\text{asp}_c(H_c; x_i^n) > 0$ for any set of (mutually distinct) design points $x_i^n$.

(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 $\text{asp}_c$, and Proposition 3 above suggests that running the unmodified AdaBoost forever will still let the prediction error converge to the suboptimal limit $L_\infty$ (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 $1/\text{asp}_p(H_c)^2$, which
is of order $n^2$ when stumps are used in boosting.

The worst amount of overfit occurs when the proba-
bility $\mu_i$'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 dif-
ference 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 Ad-
aBoost 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 dimen-
sional 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 con-
sidered 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, plausi-
ble 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 neigh-
bor rule. Recent empirical studies also confirm
that even for high dimensional sparse data Ad-
aBoost may deteriorate after running for a very
long time [E.g., Grove and Schuurmans (1998);
Mason, Baxter, Bartlett and Frean (1999); Fried-
man 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 over-
fit. Therefore regularization in one form or another
may not be unnecessary. An interesting question is:
how low can the boosted prediction error be through-
out the process of boosting? Also: how can we find
the best number of steps that achieves the lowest pre-
diction error? What are some regularization methods
to prevent overfitting? To gain some experience with
these questions, we will consider the very easy regres-
sion 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 $\mu(x_t) = E_y$ can be expanded in
the orthonormal basis $H_r = \{\phi_0, \ldots, \phi_N\}$ as $\mu = \sum_{j=1}^{N} \theta_j \phi_j$ (where $\langle \phi_j, \phi_k \rangle = \delta_{jk}$), where $\mu = \mu(x_{t})$, $\phi_j = \phi_j(x_{t})$, $\theta_j = \langle \phi_j, \mu \rangle$. Call the $\theta_j$'s as the Fourier
coefficients of $\mu$. Recall that the sample Fourier co-
efficients were defined as $\bar{\theta}_k = \langle \phi_n, y \rangle$ where $y = y_{t}$.

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 $\hat{\theta}_j \in \{\hat{\theta}_j^N\}$ so that $\hat{\theta}_j^2 = \hat{\theta}_j^2(\hat{\theta}_j)$ is the $t$-th
largest of $\{\hat{\theta}_j^2(\hat{\theta}_j)\}$, and $\hat{\theta}_j^2 > \hat{\theta}_j^2 > \ldots > \hat{\theta}_j^2$. 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 $\hat{\theta}_j$ to be $\hat{\mu}_t$ for the boosted prediction at stage
$t$, since it is also an estimator of the mean response $\mu$. Let $\mathcal{J}_t = \{\hat{\theta}_j, \ldots, \hat{\theta}_{n_t}\}$, then the expression of the
boosted prediction in Section 3 can be expressed as
$\hat{\mu}_t = \sum_{j \in \mathcal{J}_t} \theta_j \phi_j = \sum_{j=1}^{n_t} \theta_j \phi_j \hat{\mathcal{J}}_t \hat{\mathcal{J}}_t \hat{\mathcal{J}}_t \hat{\mathcal{J}}_t$.

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 $\mu$, 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 $\sigma^2$).

We first consider a ‘parametric’-family of signals. Sup-
pose that there are a finite number of $\theta_j$’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 $\mu$’s ($\mu = \sum_{i=1}^{n} \theta_i \phi_j$) the $J$-sparse family of signals $M_J$. Also define the family $M_{\hat{J}} = \cup_{j=0}^{\hat{J}} M_J$ 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 $M_{\hat{J}}$ of signals. [The second proposition follows from the results in Efroymovich (1999, Sec. 7.1).]

**Proposition 5 (Risk Upperbound with Sparse Signals).**
\[
\sup_{\mu \in M_{\hat{J}}} R(\hat{\mu}_{\hat{J}}, \mu) = \left(\frac{\sigma}{n}\right)^2 \{1 + o_n(1)\};
\]
and
\[
\sup_{\mu \in M_{\hat{J}}} \inf_{\ell \in \{0, \ldots, n\}} R(\hat{\mu}_{\ell}, \mu) = \left(\frac{\sigma}{n}\right)^2 \{1 + o_n(1)\}.
\]

**Proposition 6 (Minimax Lowerbound with Sparse Signals).**
\[
\inf_{\ell} \sup_{\mu \in M_{\hat{J}}} R(\hat{\mu}_{\ell}, \mu) = \left(\frac{\sigma}{n}\right)^2 \{1 + o_n(1)\}.
\]

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 $\hat{\mu}_{\ell}$’s beats any hard thresholding estimator $\hat{\mu}^h(\delta) = \sum_{j=0}^{\hat{J}} \theta_j \phi_j I[|\theta_j| > \delta]$, and in particular it beats the universal hard thresholding with $\delta = \sigma \sqrt{2 \log n}/n$. (Assume that the noise variance $\sigma^2$ is known.) This is because, for all $\delta > 0$,
\[
\hat{\mu}^h(\delta) = \sum_{j=0}^{\hat{J}} \theta_j \phi_j I[j \in \hat{J}] = \hat{\mu}_{\ell} \text{ if } \delta^2 \leq \hat{\theta}_j^2 (n+1)^2
\]
where $0 \equiv \hat{\theta}_j^2 (n+1) < \hat{\theta}_j^2 (n) < \hat{\theta}_j^2 (n-1) < \cdots < \hat{\theta}_j^2 (1) = \hat{\theta}_0^2 \equiv \infty; \hat{J}_t = \{t\}$ for $t = 1, \ldots, n$. Therefore, for any $\delta > 0$, we have $\hat{\mu}^h(\delta) \in \hat{\mu}_{\ell}: \ell = 0, 1, \ldots, n$, and

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

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 $\mu \in \mathbb{R}^n$, we have
\[
\inf_{\ell \in \{0, \ldots, n\}} R(\hat{\mu}_{\ell}, \mu)/\sigma^2/n + \sum_{i=1}^{n} \min(|\phi_i|, |\theta|)^2, \sigma^2/n = 2 \log n \{1 + o_n(1)\};
\]

(ii)
\[
\inf_{\ell} \sup_{\mu \in \mathbb{R}^n} R(\hat{\mu}_{\ell}, \mu)/\sigma^2/n + \sum_{i=1}^{n} \min(|\phi_i|, |\theta|)^2, \sigma^2/n = 2 \log n \{1 + o_n(1)\}.
\]

They are minimax results for guarding against all possible $\mu$, not just the ‘sparse’ ones; together with a result that ‘one of the boosted estimator’ achieves asymptotic minimax optimality.

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

Which $\hat{\mu}_{\ell}$ is a good one to choose? For sparse signals, we’d like $\hat{\mu}_{\ell} = \hat{\mu}_{\ell}\hat{J}$, but usually we do not know what $|\hat{J}|$, the number of nonzero Fourier coefficients, is. One idea is to use the thresholding techniques to kill all $\theta_j$, $j \not\in \hat{J}$. 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 $|\theta_j| > \delta$ if and only if $j \in \hat{J}$. If we threshold the coefficients during the boosting process and keep only those with magnitudes exceeding $\delta$, then at the $t$th step we get a thresholded estimator $\hat{\mu}_t(\delta) = \sum_{j \in \hat{J}_t} \theta_j \phi_j I[|\theta_j| > \delta]$. Then

**Proposition 9 (Risk Upperbound when Boosting with Universal Threshold).** For $t \geq t_\infty \equiv \sum_{i=1}^{n} \min(|\phi_i|, |\theta|)^2, \sigma^2/n = 2 \log n \{1 + o_n(1)\}$, 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 $\sigma$).* See Vidakovic (1999, Sec. 6.6.2).

Usually signal-to-noise ratio (SNR) is very small for $\theta_j$’s with $t > n/2$. Then we can let the estimator of $\sigma$ be $\hat{\sigma} = \sqrt{(n/2 - 1) \sum_{i=n/2+1}^{n} \hat{\theta}_j^2 - \hat{\theta}_{(n/2)}^2}$; where the appropriate integer part is taken for $n/2$; $\hat{\theta}_{(n/2)}$ represents the average of $\hat{\theta}_j$’s for the second half of $t$’s; $\hat{\theta}_j = \hat{\theta}_j^2$; $\hat{\theta}_s = \langle y, \phi_s \rangle$ is a sample Fourier coefficient.

In summary we have shown the following for this system of orthogonal hypotheses used in LBoost.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 $\xi_k$, 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 $[0, 1]^d = \bigcup_{k=1}^m A_k$ being partitioned into $m$ cells $A_k$, and let $\xi_k \in A_k$ be any design point that represents the cell. Define $N(\pm |k|) = \sum_{i=1}^m I[z_i \in A_k, z_i = \pm 1]$, 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 $(\xi_{k(i)} z_i)_{i=1}^m$, where $\xi_{k(i)}$ is the design point for the cell that contains $x_i$. Also, at the $t$th 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 $\hat{Z}_t = \text{sgn} \circ \hat{F}_t$ be the corresponding prediction.

**Proposition 10** (‘Quantization’ vs Histogram Rule). Suppose $\text{asp}_{+}(H_c; \xi_{[n]}) > 0$ 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

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

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 $\tau$ such that $\hat{Z}_t(x) = \lim_{s \to \infty} Z_s(x)$ for all $t \geq \tau$.

Define the prediction error (with expectation also running over the random predictor $X$) as $L_t \equiv E[Y^{new} - \hat{Y}_t]$. In other words, this is the prediction error of the optimal quantized predictor $\hat{F}_t$.

$$L_t \equiv E[Y^{new} - \hat{F}_t],$$

and we are interested in the asymptotic minimax properties of $L_t$. We will also make use of the following result due to Donoho and Johnstone (1994):

**Theorem 1**. Suppose $\text{asp}_{+}(H_c; \xi_{[n]}) > 0$ 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

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

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 $\tau$ such that $\hat{Z}_t(x) = \lim_{s \to \infty} Z_s(x)$ 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

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

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)].

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


