

# PROCESS CONSISTENCY FOR ADABOOST

BY WENXIN JIANG

*Department of Statistics, Northwestern University*

Submitted on November 30, 2000 <sup>1</sup>

Department of Statistics Technical Report 00-05

<sup>2</sup>Recent experiments and theoretical studies show that AdaBoost can overfit in the limit of large time. If running the algorithm forever is suboptimal, a natural question is how low can the prediction error be *during the process* of AdaBoost? We show under general regularity conditions that during the process of AdaBoost a consistent prediction is generated, which has the prediction error approximating the optimal Bayes error as the sample size increases. This result suggests that, while running the algorithm forever can be suboptimal, it is reasonable to expect that some regularization method via truncation of the process may lead to a near-optimal performance.

**1. Introduction.** Some recent experimental results [e.g., Friedman, Hastie and Tibshirani (1999); Grove and Schuurmans (1998); Mason et al. (1998)] and theoretical examples [Jiang (1999)] suggest that the AdaBoost algorithm [e.g., Schapire (1999); Freund and Schapire (1997)] can overfit in the limit of (very) large *time* (or the number of rounds of AdaBoost), despite the observation that the algorithm is often found to be resistant to overfitting after running hundreds of rounds. Jiang (1999) provides examples where it can be shown that the prediction error of Adaboost [ $PE(AdaBoost_n^t)$ , depending on the sample size  $n$  and the time  $t$ ] is asymptotically suboptimal at  $t = \infty$ , in the sense that the prediction at  $t = \infty$  is not consistent. Here by *consistency* of a prediction we mean that as the sample size  $n$  increases, the prediction based on the sample has a prediction error that converges to the optimal Bayes error. When running the unmodified AdaBoost algorithm forever ( $t = \infty$ ), there are situations when the resulting prediction error converges to a suboptimal value larger than the optimal Bayes error as the sample size  $n$  increases, i.e.,  $\lim_{n \rightarrow \infty} PE(AdaBoost_n^{t=\infty}) > Bayes\ Error$ .

If running the algorithm forever can be suboptimal, a natural question is how low a prediction error  $PE(AdaBoost_n^t)$  can AdaBoost achieve *during the process of  $t$* ? Can AdaBoost generate a prediction during the process that can have a nearly optimal prediction error as  $n$  increases? Is it true that  $\lim_{n \rightarrow \infty} \inf_{t \in \{1, 2, 3, \dots\}} PE(AdaBoost_n^t) = Bayes\ Error$ ? If this last formula holds, then we say that AdaBoost is *process consistent*. As far as we know this problem has not been addressed in the previous literature. The bounds on the prediction error obtained before [e.g., Schapire et al. (1998) and Breiman (1997)] are all semi-empirical in the sense that they involve some sample quantities (related to the margin or the top) and are not compared to the optimal Bayes error. This problem of process consistency is also theoretically important since the process consistency would imply that even though running the AdaBoost algorithm *forever* may be suboptimal, the

---

<sup>1</sup>See also revised version submitted on November 25, 2000 at <http://meyman.stats.nwu.edu/jiang/boost/boost.process2.ps>

<sup>2</sup>AMS 1991 *subject classifications*. Primary 62G99; secondary 68T99.

*Key words and phrases*. AdaBoost, Bayes error, boosting, consistency, prediction error, VC dimension

algorithm does achieve a good asymptotic performance *at some time during the process*. Therefore it is reasonable to expect that some regularization method via truncation of the process may lead to a near-optimal performance.

In a recent work, Breiman (2000) considers the case  $n = \infty$  and shows that this *population* version of AdaBoost typically leads to a limiting prediction that achieves the optimal Bayes error as  $t$  increases. That is,  $\lim_{t \rightarrow \infty} PE(\text{AdaBoost}_{n=\infty}^t) = \text{Bayes Error}$ . We will utilize some of his results to study the asymptotic behavior of the *sample* version of AdaBoost, in particular, the problem of process consistency. We will show that Adaboost is process consistent under very general regularity conditions. Therefore even though running the algorithm forever is often suboptimal, *the algorithm does achieve a good asymptotic performance at some time during the process*; and a systematic study on regularization by truncating the process may be a reasonable direction for future research. Below we introduce the notation and the main results.

**2. Notation and main results.** Let  $(X_i, Z_i)_1^n$  and  $(X, Z)$  be iid (independent and identically distributed) random quantities valued in  $[0, 1]^d \times \{\pm 1\}$ . Let  $H$  be a *base hypothesis space*, which is a set of functions  $f : [0, 1]^d \mapsto \{\pm 1\}$ . Let  $C_n(F) = n^{-1} \sum_1^n e^{-Z_i F(X_i)}$  (the AdaBoost cost function),  $C_\infty(F) = E e^{-ZF(X)}$  (the population version). For each  $n = 1, 2, \dots, \infty$  define the following sequential fits. They define the sample version and the population version (for  $n = \infty$ ) of the AdaBoost algorithm.  $F_n^0 = 0$ , and for  $t = 1, 2, \dots$ ,  $F_n^t = F_n^{t-1} + \alpha_n^t f_n^t$ . The resulting prediction at step  $t$  is  $\text{sgn} \circ F_n^t$ .

Here  $f_n^t = \arg \max_{f \in H} \Delta_n^{t-1}(f)$  and  $\alpha_n^t = \frac{1}{2} \log \left( \frac{1+2\delta_n^t}{1-2\delta_n^t} \right)$ . We use the notation

$$\Delta_n^{t-1}(f) = n^{-1} \sum_{j=1}^n e^{-Z_j F_n^{t-1}(X_j)} Z_j f(X_j),$$

$$\Delta_\infty^{t-1}(f) = E e^{-ZF_\infty^{t-1}(X)} Z f(X),$$

$$2\delta_n^t = \Delta_n^{t-1}(f_n^t) / C_n(F_n^{t-1}).$$

The goodness of any prediction of the form  $\text{sgn} \circ F$  is measured by the misclassification probability  $L_\infty(F) = P[Z \neq \text{sgn} \circ F(X)]$ . The gold-standard is *the Bayes error*  $L_\infty(F_B) = P[Z \neq \text{sgn} \circ F_B(X)]$  where  $F_B(X) = \frac{1}{2} \log \left\{ \frac{P(Z=1|X)}{P(Z=-1|X)} \right\}$  corresponding to the optimal Bayes prediction. If a sequence of predictions  $\text{sgn} \circ F_n$ , possibly depending on the data  $S = (X_i, Z_i)_1^n$ , has a *prediction error*  $E_S L_\infty(F_n) \rightarrow L_\infty(F_B)$ , then we say that the prediction is *consistent*. We will show that there is a sequence  $t_n$  such that the prediction  $\text{sgn} \circ F_n^{t_n}$  generated by AdaBoost is consistent. *Therefore the lowest point of the prediction error during the process of AdaBoost is close to the gold-standard for large sample sizes.*

We will use the following regularity conditions:

- (I) [Population Solution]. For any sequence (allowing possible multiple solutions)  $F_\infty^t$  of fits from the population version of AdaBoost, one has  $\lim_{t \rightarrow \infty} \|F_\infty^t - F_B\|_{L^2(P_X)} = 0$ .
- (II) [Base Hypothesis Space]. The VC (Vapnik-Chervonenkis) dimension of  $H$  is finite, i.e.,  $VC(H) < \infty$ . [For the concept of the VC dimension, see, e.g., Anthony and Biggs (1992, Chapter 7).]
- (III) [Population Coefficients]. The coefficients in the population AdaBoost are all finite, i.e.,  $|\alpha_\infty^s| < \infty$  for all  $s$ .

- (IV) [*t*-Step ‘Consistency’]. Given any  $t$  and any sample realization  $S$  such that  $D_{n,\infty}^t = \sup_{f \in H} |\Delta_n^{t-1}(f) - \Delta_\infty^{t-1}(f)| = o_n(1)$ , given a sequence  $f_n^t = \arg \max_{f \in H} \Delta_n^{t-1}(f)$ ,  $\exists f_\infty^t = \arg \max_{f \in H} \Delta_\infty^{t-1}(f)$  such that  $\|f_n^t - f_\infty^t\|_{L_2(P_X)} = o_n(1)$ .
- (V) [*t*-Step ‘Nonflatness’]. For each  $t$ , for some constant  $C_t$  that does not depend on  $n$ , we have  $\|f - f_\infty^t\|_{L_2(P_X)}^{2q} \leq 0.5C_t \{\Delta_\infty^{t-1}(f_\infty^t) - \Delta_\infty^{t-1}(f)\} \{1 + o(1)\}$  for some  $q \geq 1$ , whenever  $\|f - f_\infty^t\|_{L_2(P_X)} = o(1)$ .

**THEOREM.** (*Process Consistency for AdaBoost*). *Under conditions (I) to (V), there exists a sequence  $t_n$  such that  $\lim_{n \rightarrow \infty} E_S L_\infty(F_n^{t_n}) = L_\infty(F_B)$  [and therefore  $\lim_{n \rightarrow \infty} \inf_t E_S L_\infty(F_n^t) = L_\infty(F_B)$  also].*

**REMARKS.**

1. Condition (I) holds due to Theorem 3 of Breiman (2000), if  $F_B(\cdot)$  is continuous and if the linear span of  $H$  is complete in  $L_2(P_X)$ .
2. Condition (II) typically holds.
3. Condition (III) is plausible and basically prohibits a perfect population fit by the base space  $H$  at each step of ‘base learning’.
4. Condition (IV) basically says that at each step of base learning in AdaBoost, whenever the sample and population criterion functions are uniformly close to each other, the sets of maximizers also need to be close to each other. This condition is satisfied in typical cases when  $f$  in  $H$  is the sign transformation of a function that is continuously parameterized in a compact parameter space.
5. Condition (V) essentially says something about the nonflatness of the population criterion function at its peak, for each step of base learning in AdaBoost. When the criterion function  $\Delta_\infty^{t-1}(f)$  gets close to its maximum, the corresponding function  $f$  needs to get close to the maximizer at some polynomial rate. This seems to be a very relaxed condition since the rate can be arbitrarily slow (by choosing some large  $q$ ). However the condition may be hard to check in practice. We did check that the condition is satisfied (with  $q = 2$ ) when  $H$  is the family of step functions on  $[0, 1]$  [ $H = \{s \cdot \text{sgn}(\cdot - a) : s \in \{-1, 1\}, a \in [0, 1]\}$ ], provided that  $X$  has a nonzero density function  $\pi$  such that  $\inf_{x \in [0, 1]} \pi(x) > 0$  and that  $\mu(x) = E(Z|x)$  is a continuously differentiable function with derivatives bounded away from zero at the ‘crossing points’ in  $A^t = \{x : x \in [0, 1], F_\infty^{t-1}(x) = F_B(x)\}$ , i.e.,  $\inf_{\psi \in A^t} |\mu'(\psi)| > 0$ , for each  $t$ .
6. The proof is nonconstructive and we do not know what a rate  $t_n$  can take. Presumably some  $t_n$  that increases to infinity very slowly will work. This is because as  $t_n \rightarrow \infty$ , the ‘approximation error’ is related to  $\|F_\infty^{t_n} - F_B\|_{L_2(P_X)}$  which goes to zero. On the other hand if the growth  $t_n$  is slow enough, the sample AdaBoost fit  $F_n^{t_n}$  is sufficiently close to the population version  $F_\infty^{t_n}$  for large  $n$ . This is actually the main intuition behind the proof, which uses a method of induction over  $t$ .

*Proof for the Theorem:*

To prove the theorem we first consider a slightly more general setup and formulate some general sufficient conditions for process consistency in Proposition 1 in the next section. Then we will check

that these conditions are satisfied given the more primitive conditions (I) to (V) in the case of AdaBoost.  $\square$

**3. A slightly more general setup: generalized additive model (GAM) with sequential fits.** Denote  $p(x) = P[Z = 1|X = x]$ . Assume that  $p(x) = \psi \circ F(x)$  where  $\psi$  is a strictly increasing, continuously differentiable cdf (cumulative distribution function) such that  $\psi(0) = 0.5$  [such that  $\text{sgn}(F) = \text{sgn}(p - 1/2)$ ] and  $\psi'$  is a continuous pdf (probability density function) bounded on  $\mathfrak{R}^1$ . Here  $F$  is to be estimated by a sequential ( $t$ -step) additive fit  $F_n^t = \sum_{s=1}^t \alpha_n^s f_n^s$ ,  $\alpha_n^s \in \mathfrak{R}$ ,  $f_n^s \in H$  ( $H$  is a base hypothesis space), which may depend on the data  $S = (X_i, Z_i)_1^n$ , similar to the previous section. Denote  $F_B = \psi^{-1} \circ p$ . As an example, in AdaBoost,  $\psi = \frac{e^{2F}}{1+e^{2F}}$ ,  $|\psi'| \leq 0.5$ , and  $F_B(x) = \frac{1}{2} \log \left\{ \frac{p(x)}{1-p(x)} \right\}$ .

LEMMA 1.  $L_\infty(F_n^t) - L_\infty(F_B) \leq 2\|\psi'\|_\infty \|F_n^t - F_B\|_{L^2(P_X)}$ .

This is a variant of Corollary 6, Devroye, Györfi and Lugosi (1996), obtained by a first order Taylor expansion. Here  $\|\psi'\|_\infty = \sup_{x \in \mathfrak{R}} |\psi'(x)|$ . Below we will use  $\|\cdot\|$  to denote  $\|\cdot\|_{L^2(P_X)}$  unless otherwise noted.

PROPOSITION 1. (*Process Consistency for Sequential GAM*). Suppose that there exists a non-stochastic sequence  $F_\infty^t$  of functions on the domain of  $X$ , independent of  $n$ , such that

$$(i) \|F_n^t - F_\infty^t\| \leq b_n^t \text{ with probability } P_S \geq 1 - a_n^t;$$

$$(ii) \|F_\infty^t - F_B\| \leq c^t;$$

where  $a_n^t, b_n^t, c^t$  are nonnegative,  $c^t \rightarrow 0$  as  $t \rightarrow \infty$ , and  $a_n^t, b_n^t \rightarrow 0$  as  $n \rightarrow \infty \forall t$ .

Then, there is a sequence  $t(n)$  such that  $E_S L(F_n^{t(n)}) - L(F_B) = o_n(1)$ .

In the case of AdaBoost,  $F_\infty^t$  was taken to be a population fit and (ii) is guaranteed via Condition (I) by Theorem 3 of Breiman (2000); what we only need for proving the main theorem on process consistency is to establish condition (i). This will be done in the next section.

Below we first prove the proposition itself in the GAM context.

*Proof for Proposition 1:*

Since the nonnegative sequences  $a_n^t, b_n^t \rightarrow 0$  as  $n \rightarrow \infty \forall t$ , there exists a sequence  $t(n) \rightarrow \infty$  sufficiently slow, such that  $a_n^{t(n)}, b_n^{t(n)} \rightarrow 0$  as  $n \rightarrow \infty$ . Also we have  $c^{t(n)} \rightarrow 0$ .

Denote  $Y_n = L_\infty(F_n^{t(n)}) - L_\infty(F_B)$ . Note that  $Y_n \in [0, 1]$  and for any  $\epsilon \in [0, 1]$ ,

$$\begin{aligned} 0 \leq E_S Y_n &= \int_0^\epsilon Y_n dP_S + \int_\epsilon^1 Y_n dP_S \\ &\leq \epsilon + P[Y_n > \epsilon]. \end{aligned}$$

Now take  $\epsilon = 2\|\psi'\|_\infty (b_n^{t(n)} + c^{t(n)})$ . Note that by the previous lemma, triangular inequality and conditions (i) and (ii), we have

$$\begin{aligned} Y_n &\leq (\|F_n^{t(n)} - F_\infty^{t(n)}\| + \|F_\infty^{t(n)} - F_B\|) \cdot 2\|\psi'\|_\infty \\ &\leq \epsilon \end{aligned}$$

with probability  $P_S \geq 1 - a_n^{t(n)}$ . Therefore  $E_S Y_n \leq 2\|\psi'\|_\infty (b_n^{t(n)} + c^{t(n)}) + a_n^{t(n)} = o_n(1)$ .  $\square$

Now we prove condition (i) (convergence of the sequential fits) of the proposition for the case of AdaBoost under more primitive conditions listed in the previous section.

**4. Convergence of the sequential fits.** The condition (i) of the proposition in the previous section is established immediately via the following proposition:

**PROPOSITION 2.** (*Convergence of the Sequential Fits*). *Suppose Conditions (I) to (V) hold with some  $q \geq 2$  (without loss of generality). Then, for any  $\beta > 0$ , for any  $t = 1, 2, \dots$ , for some nonnegative constant  $\Lambda_t$  that does not depend on  $n$ , there is a population AdaBoost fit  $F_\infty^t$  such that with probability  $P_S \geq 1 - 16tn^{-\beta}$ ,*

$$\|F_n^t - F_\infty^t\| \leq \Lambda_t [(\log n/n)\{VC(H) + \beta\}]^{1/4q^t} \{1 + o_n(1)\}.$$

So we only need to prove this proposition now, which is done by using the following lemma. If this following lemma is true, then the proposition on  $\|F_n^t - F_\infty^t\|$  is easily proved by applying the triangular inequalities to  $\|F_n^t - F_\infty^t\| = \|\sum_{s=1}^t (\alpha_n^s f_n^s - \alpha_\infty^s f_\infty^s)\|$ .  $\square$

**LEMMA 2.** (*Convergence Step-by-Step*). *Suppose all conditions (I) to (V) hold and  $q \geq 2$  (without loss of generality). Then for any  $\beta > 0$ , for some nonnegative constants  $\lambda_1^t$  that do not depend on  $n$ , we can have  $\|f_n^s - f_\infty^s\|^{2q}$  and  $|\alpha_n^s - \alpha_\infty^s|$  both bounded above by  $\lambda_s [(\log n/n)\{VC(H) + \beta\}]^{1/2q^{s-1}} \{1 + o_n(1)\}$  for all  $s = 1, 2, \dots, t$ , with probability  $P_S$  at least  $1 - 16tn^{-\beta}$ .*

Now let us prove this lemma ‘Convergence Step-by-Step’. This is done by a method of induction that uses the following secondary lemmas, where we suppose that all conditions (I) to (V) hold and  $q \geq 2$  (without loss of generality). These secondary lemmas will be proved in the next section.

**LEMMA 3.**  $|\Delta_n^{t-1}(f_n^t) - \Delta_\infty^{t-1}(f_\infty^t)| \leq D_{n,\infty}^t \leq Q_{1n}^{t-1} + R_n^{t-1}$ , where  $D_{n,\infty}^t = \sup_{f \in H} |\Delta_n^{t-1}(f) - \Delta_\infty^{t-1}(f)|$ ,  $Q_{1n}^{t-1} = \sup_{f \in H} |n^{-1} \sum_{i=1}^n e^{-Z_i F_n^{t-1}(X_i)} f(X_i) Z_i - E e^{-Z F_n^{t-1}(X)} f(X) Z|$ , and  $R_n^{t-1} = E |e^{-Z F_n^{t-1}(X)} - e^{-Z F_\infty^{t-1}(X)}|$ .

**LEMMA 4.** For any  $\beta > 0$ , we have  $P_S[Q_{1n}^{t-1} \leq U_n^{t-1}] \geq 1 - 8n^{-\beta}$ , and  $P_S[Q_{2n}^{t-1} \leq U_n^{t-1}] \geq 1 - 8n^{-\beta}$ . Here  $Q_{2n}^{t-1} = |n^{-1} \sum_{j=1}^n e^{-Z_j F_n^{t-1}(X_j)} - E e^{-Z F_n^{t-1}(X)}|$  and

$$\begin{aligned} U_n^{t-1} &= e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|} \sqrt{(32 \log n/n)\{VC(H)t + \beta\} + (32/n)VC(H)t \log\{e/VC(H)\}} \\ &\quad + 2e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|} e^{\sum_{s=1}^{t-1} |\alpha_n^s - \alpha_\infty^s|} \sum_{s=1}^{t-1} |\alpha_n^s - \alpha_\infty^s|. \end{aligned}$$

**LEMMA 5.**  $R_n^t \leq e^{|\alpha_\infty^t| + |\alpha_n^t - \alpha_\infty^t|} [R_n^{t-1} + e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|} |\alpha_n^t - \alpha_\infty^t| + 0.5 |\alpha_\infty^t| e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|} \{\|f_n^t - f_\infty^t\|_{L_2(P_X)}^{2q}\}^{1/q}]$ .

**LEMMA 6.**  $|\alpha_n^t - \alpha_\infty^t| \leq 0.5\{(1 - 4(\delta_\infty^t)^2)^{-1} + (1 - 4(\delta_n^t)^2)^{-1}\} |2\delta_n^t - 2\delta_\infty^t|$  and  $|2\delta_n^t - 2\delta_\infty^t| \leq \{C_n(F_n^{t-1})\}^{-1} (Q_{1n}^{t-1} + R_n^{t-1} + 2\delta_\infty^t Q_{2n}^{t-1} + 2\delta_\infty^t R_n^{t-1})$ .

LEMMA 7.  $0 \leq \Delta_\infty^{t-1}(f_\infty^t) - \Delta_\infty^{t-1}(f_n^t) \leq 2D_{n,\infty}^t$ .

LEMMA 8. *For the maximizers  $f_n^t$  and  $f_\infty^t$  in Condition (IV), we have  $\|f_n^t - f_\infty^t\|_{L_2(P_X)}^{2q} \leq C_t D_{n,\infty}^t \{1 + o_D(1)\}$ , where  $o_D(1)$  is a term that converges to zero whenever  $D_{n,\infty}^t$  converges to zero.*

Now we prove Lemma 2 by induction and applying the secondary lemmas above.

*Proof for Lemma 2:*

For  $t = 1$ ,  $D_{n,\infty}^1 = \sup_{f \in \mathcal{H}} |n^{-1} \sum_{i=1}^n f(X_i)Z_i - Ef(X)Z|$ , which is at most  $\sqrt{(32 \log n/n)\{VC(H) + \beta\}\{1 + o_n(1)\}}$  with probability at least  $1 - 8n^{-\beta}$ , for any  $\beta > 0$ , by the VC uniform bounding technique (Lemmas 3 and 4). Under this event of  $D_{n,\infty}^t$  (or equivalently under the event  $[Q_{1n}^0 \leq U_n^0]$ ), we have  $\|f_n^1 - f_\infty^1\|_{L_2(P_X)}^{2q} \leq C_1 \sqrt{(32 \log n/n)\{VC(H) + \beta\}\{1 + o_n(1)\}}$  (Lemma 8) and also  $|\alpha_n^1 - \alpha_\infty^1| \leq \{1 - 4(\delta_\infty^1)^2\}^{-1} \sqrt{(32 \log n/n)\{VC(H) + \beta\}\{1 + o_n(1)\}}$  (Lemma 6). The secondary lemmas then allow us to perform induction from time  $t$  to  $t+1$ . If  $q \geq 2$  (without loss of generality), then each step will result in an upperbound with a larger order [from  $(\log n/n)^{1/2q^{t-1}}$  to  $(\log n/n)^{1/2q^t}$ ] due to the power  $1/q$  in Lemma 5. These bounds for all  $s = 1, \dots, t$  will be valid under the event  $Q_{1,2n}^{0,1,2,\dots,t-1} \leq U_n^{0,1,2,\dots,t-1}$ , which has probability at least  $1 - (2t)(8n^{-\beta})$ . This leads to the proof.  $\square$

## 5. Proof for secondary lemmas. *Proof for Lemma 3:*

Note that

$$|\Delta_n^{t-1}(f_n^t) - \Delta_\infty^{t-1}(f_\infty^t)| = |\sup_{f \in \mathcal{H}} \Delta_n^{t-1}(f) - \sup_{f \in \mathcal{H}} \Delta_\infty^{t-1}(f)| \leq D_{n,\infty}^t.$$

Next, write

$$D_{n,\infty}^t = \sup_{f \in \mathcal{H}} \left| \left\{ n^{-1} \sum_{i=1}^n e^{-Z_i F_n^{t-1}(X_i)} f(X_i)Z_i - E e^{-Z F_n^{t-1}(X)} f(X)Z \right\} + \left\{ E(e^{-Z F_n^{t-1}(X)} - e^{-Z F_\infty^{t-1}(X)}) f(X)Z \right\} \right|,$$

apply the triangular inequality, and note that

$$\sup_{f \in \mathcal{H}} |E(e^{-Z F_n^{t-1}(X)} - e^{-Z F_\infty^{t-1}(X)}) f(X)Z| \leq E|e^{-Z F_n^{t-1}(X)} - e^{-Z F_\infty^{t-1}(X)}|.$$

This shows the further upperbound  $Q_{1n}^{t-1} + R_n^{t-1}$ .  $\square$

*Proof for Lemma 4:*

By the triangular inequalities,

$$Q_{1n}^{t-1} = \sup_{f \in \mathcal{H}} \left| n^{-1} \sum_{i=1}^n e^{-Z_i F_n^{t-1}(X_i)} f(X_i)Z_i - E e^{-Z F_n^{t-1}(X)} f(X)Z \right|$$

is bounded above by  $T_1 + T_2 + T_3$  where

$$T_1 = \sup_{f \in \mathcal{H}} \left| n^{-1} \sum_{i=1}^n e^{-Z_i} \sum_{s=1}^{t-1} \alpha_\infty^s f_n^s(X_i) f(X_i)Z_i - E e^{-Z} \sum_{s=1}^{t-1} \alpha_\infty^s f_n^s(X) f(X)Z \right|,$$

$$T_2 = \sup_{f \in H} |n^{-1} \sum_{i=1}^n (e^{-Z_i \sum_{s=1}^{t-1} \alpha_n^s f_n^s(X_i)} - e^{-Z_i \sum_{s=1}^{t-1} \alpha_\infty^s f_n^s(X_i)}) f(X_i) Z_i|,$$

and

$$T_3 = \sup_{f \in H} |E(e^{-Z \sum_{s=1}^{t-1} \alpha_n^s f_n^s(X)} - e^{-Z \sum_{s=1}^{t-1} \alpha_\infty^s f_n^s(X)}) f(X) Z|.$$

Both  $T_2$  and  $T_3$  can be shown to be bounded above by

$$e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|} e^{\sum_{s=1}^{t-1} |\alpha_n^s - \alpha_\infty^s|} \sum_{s=1}^{t-1} |\alpha_n^s - \alpha_\infty^s|$$

by applying a first order Taylor expansion.  $T_1$  is bounded above by

$$e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|} \sqrt{(32 \log n/n) \{VC(H)t + \beta\} + (32/n)VC(H)t \log\{e/VC(H)\}}$$

with probability at least  $1 - 8n^{-\beta}$ , for any  $\beta > 0$ , by applying the following proposition that is proved in the same way as the VC result Theorem 12.5 Devroye et al. (1996):

**PROPOSITION 3.** *let  $W_1^n$  and  $W$  be i.i.d. random vectors,  $\varphi \in \Phi$  a family of nonstochastic (possibly multivariate) functions defined on the domain of  $W$ ,  $g$  a nonstochastic function such that  $g(\varphi(W), W) \in [-M, M]$  for all  $W$  and all  $\varphi \in \Phi$ . let  $s(\Phi, n) = \max_{W_1^n} \text{card}\{\varphi(W_1^n) : \varphi \in \Phi\}$ . Then*

$$P[\sup_{\varphi \in \Phi} |n^{-1} \sum_{i=1}^n g(\varphi(W_i), W_i) - Eg(\varphi(W), W)| > \epsilon] \leq 8s(\Phi, n)e^{-n\epsilon^2/32M^2}$$

for any  $\epsilon > 0$  and

$$P[\sup_{\varphi \in \Phi} |n^{-1} \sum_{i=1}^n g(\varphi(W_i), W_i) - Eg(\varphi(W), W)| \leq M \sqrt{(32/n) \{\log s(\Phi, n) + \beta \log n\}}] \geq 1 - 8n^{-\beta}$$

for all  $\beta > 0$ .

In our case set  $\varphi = (f^1, \dots, f^t) \in H^t \equiv \Phi$ ,  $W_i = (Z_i, X_i)$ ,  $W = (Z, X)$ . Then  $T_1$  is bounded above by

$$\sup_{\varphi \in \Phi} |n^{-1} \sum_{i=1}^n e^{-Z_i \sum_{s=1}^{t-1} \alpha_\infty^s f_n^s(X_i)} f^t(X_i) Z_i - E e^{-Z \sum_{s=1}^{t-1} \alpha_\infty^s f_n^s(X)} f^t(X) Z|.$$

Here  $M$  can be taken as  $e^{\sum_{s=1}^{t-1} |\alpha_\infty^s|}$  for application of the proposition and

$$s(\Phi, n) \leq s(H, n)^t \equiv [\max_{X_1^n} \text{card}\{f(X_1^n) : f \in H\}]^t \leq \{en/VC(H)\}^{VC(H)t}.$$

Combining the resulting bounds for  $T_1$ ,  $T_2$  and  $T_3$  we obtain the lemma for the statement on  $Q_{1n}^{t-1}$ . The proof for the statement on  $Q_{2n}^{t-1}$  is similar.  $\square$

*Proof for Lemma 5:*

Note that

$$\begin{aligned} & E|e^{-ZF_n^t} - e^{-ZF_\infty^t}| \\ &= E|e^{-ZF_n^{t-1}} e^{-Z\alpha_n^t f_n^t} - e^{-ZF_\infty^{t-1}} e^{-Z\alpha_\infty^t f_\infty^t}| \\ &\leq E(|e^{-Z\alpha_n^t f_n^t}||e^{-ZF_n^{t-1}} - e^{-ZF_\infty^{t-1}}|) + E(|e^{-ZF_\infty^{t-1}}||e^{-Z\alpha_n^t f_n^t} - e^{-Z\alpha_\infty^t f_\infty^t}|). \end{aligned}$$

Note that

$$|e^{-Z\alpha_n^t f_n^t}| \leq e^{|\alpha_n^t|} \leq e^{|\alpha_\infty^t| + |\alpha_m^t - \alpha_\infty^t|},$$

and use a first order Taylor expansion to obtain

$$\begin{aligned} & |e^{-ZF_\infty^{t-1}}| |e^{-Z\alpha_n^t f_n^t} - e^{-Z\alpha_\infty^t f_\infty^t}| \\ & \leq |e^{-Z(F_\infty^{t-1} + \alpha_n^t \tilde{f}_n^t)}| |\alpha_n^t f_n^t - \alpha_\infty^t f_\infty^t| \\ & \leq |e^{-Z(F_\infty^{t-1} + \alpha_n^t \tilde{f}_n^t)}| (|f_n^t| |\alpha_n^t - \alpha_\infty^t| + |\alpha_\infty^t| |f_n^t - f_\infty^t|) \end{aligned}$$

where  $\tilde{\alpha}_n^t \tilde{f}_n^t$  is between  $\alpha_\infty^t f_\infty^t(X)$  and  $\alpha_n^t f_n^t(X)$ . Note that  $|f_n^t| = 1$ ,  $|f_n^t - f_\infty^t| = 0.5(f_n^t - f_\infty^t)^2$ ,

$$e^{-Z(F_\infty^{t-1} + \alpha_n^t \tilde{f}_n^t)} \leq e^{\sum_{s=1}^{t-1} |\alpha_\infty^s| + |\alpha_\infty^t| + |\alpha_n^t - \alpha_\infty^t|}.$$

Combining these results we then get

$$\begin{aligned} & E|e^{-ZF_n^t} - e^{-ZF_\infty^t}| \\ & \leq e^{|\alpha_\infty^t| + |\alpha_m^t - \alpha_\infty^t|} E|e^{-ZF_n^{t-1}} - e^{-ZF_\infty^{t-1}}| \\ & + e^{\sum_{s=1}^{t-1} |\alpha_\infty^s| + |\alpha_\infty^t| + |\alpha_n^t - \alpha_\infty^t|} (|\alpha_n^t - \alpha_\infty^t| + 0.5|\alpha_\infty^t| (\|f_n^t - f_\infty^t\|_{L^2(P_X)}^{2q})^{1/q}) \end{aligned}$$

which proves the lemma.  $\square$

*Proof for Lemma 6:*

First it is easy to prove by a first order Taylor expansion that

$$|\alpha_n^t - \alpha_\infty^t| \leq 0.5\{(1 - 4(\delta_\infty^t)^2)^{-1} + (1 - 4(\delta_n^t)^2)^{-1}\} |2\delta_n^t - 2\delta_\infty^t|.$$

Then note that

$$\begin{aligned} |2\delta_n^t - 2\delta_\infty^t| & = |\Delta_n^{t-1}(f_n^t)/C_n(F_n^{t-1}) - \Delta_\infty^{t-1}(f_\infty^t)/C_\infty(F_\infty^{t-1})| \\ & = C_n(F_n^{t-1})^{-1} |(\Delta_n^{t-1}(f_n^t) - \Delta_\infty^{t-1}(f_\infty^t)) - 2\delta_\infty^t (C_n(F_n^{t-1}) - C_\infty(F_\infty^{t-1}))|. \end{aligned}$$

Now apply the triangular inequality. Note that

$$|\Delta_n^{t-1}(f_n^t) - \Delta_\infty^{t-1}(f_\infty^t)| \leq Q_{1n}^{t-1} + R_n^{t-1}$$

by Lemma 3 and by using a triangular inequality we also have

$$|C_n(F_n^{t-1}) - C_\infty(F_\infty^{t-1})| \leq Q_{2n}^{t-1} + R_n^{t-1}.$$

Combining these results we have the proof of the lemma.  $\square$

*Proof for Lemma 7:*

Note that  $0 \leq \Delta_\infty^{t-1}(f_\infty^t) - \Delta_\infty^{t-1}(f_n^t)$  since  $\Delta_\infty^{t-1}(f_\infty^t) = \sup_{f \in H} \Delta_\infty^{t-1}(f)$ . Next note that

$$\begin{aligned} \Delta_\infty^{t-1}(f_\infty^t) - \Delta_\infty^{t-1}(f_n^t) & = \{\Delta_\infty^{t-1}(f_\infty^t) - \Delta_n^{t-1}(f_\infty^t)\} + \\ & \{\Delta_n^{t-1}(f_\infty^t) - \Delta_n^{t-1}(f_n^t)\} + \{\Delta_n^{t-1}(f_n^t) - \Delta_\infty^{t-1}(f_n^t)\}. \end{aligned}$$

The first and third terms on the right hand side are both bounded above by  $\sup_{f \in H} |\Delta_n^{t-1}(f) - \Delta_\infty^{t-1}(f)| = D_{n,\infty}^t$ , while the second term is at most zero since  $\Delta_n^{t-1}(f_n^t) = \sup_{f \in H} \Delta_n^{t-1}(f)$ . These leads to the proof of Lemma 7.  $\square$

*Proof for Lemma 8:*

Condition (V) and Lemma 7 leads to an upperbound of  $\|f_n^t - f_\infty^t\|_{L_2(P_X)}^{2q}$  as  $C_t D_{n,\infty}^t \{1 + o(1)\}$ , where the  $o(1)$  term converges to zero as  $\|f_n^t - f_\infty^t\|_{L_2(P_X)}^{2q}$  becomes small, which is guaranteed if  $D_{n,\infty}^t$  is small due to Condition (IV). These arguments prove the lemma.  $\square$

**Acknowledgments.** The author is grateful to Leo Breiman for his encouragement and for providing the technical report Breiman (2000).

## REFERENCES

- ANTHONY, M. AND BIGGS, N. (1992). *Computational Learning Theory: An Introduction*. Cambridge University Press, Cambridge.
- BREIMAN, L. (1997). Prediction games and arcing classifiers. *Technical Report 504, Statistics Department, University of California at Berkeley*.
- BREIMAN, L. (2000). Some infinity theory for predictor ensembles. *Technical Report 579, Statistics Department, University of California at Berkeley*.
- DEVROYE, L., GYÖRFI, L. AND LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. 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., 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 SCHURMANS, 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.
- 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>.)
- 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*.
- 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.

DEPARTMENT OF STATISTICS  
NORTHWESTERN UNIVERSITY  
EVANSTON, IL 60208

E-MAIL: [wjiang@northwestern.edu](mailto:wjiang@northwestern.edu)

<http://neyman.stats.nwu.edu/jiang/boost/boost.process.ps>