The Annals of Statistics
2004, Vol. 32, No. 1, 13–29
© Institute of Mathematical Statistics, 2004

PROCESS CONSISTENCY FOR ADABOOST

BY WENXIN JIANG

Northwestern University

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 for sufficiently large sample size.

1. Introduction. Some recent experimental results [e.g., Friedman, Hastie and Tibshirani (2000), Grove and Schuurmans (1998) and Mason, Baxter, Bartlett and Frean (1999)] and theoretical examples [Jiang (2002)] suggest that the AdaBoost algorithm [e.g., Schapire (1999) and 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 (2002) 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, that is, $\lim_{n\to\infty} PE(AdaBoost_n^{t=\infty}) > Bayes\ Error$.

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 $\lim_{n\to\infty} \inf_{t\in\{1,2,3,\dots\}} PE(AdaBoost_n^t) = Bayes\ Error$? 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

Received December 2000; revised October 2002.

Supported in part by NSF Grant DMS-01-02636.

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

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

prediction error obtained before [e.g., Schapire, Freund, Bartlett and Lee (1998) and Breiman (1997)] are all semiempirical 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 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 for sufficiently large sample size.

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, $\lim_{t\to\infty} PE(AdaBoost^t_{n=\infty}) = Bayes\ Error$. 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 produces process-consistent predictions 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 i.i.d. (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) and let $C_\infty(F) = Ee^{-ZF(X)}$ (the population version). For each $n = 1, 2, \dots, \infty$, define the following sequential fits F_n^t. They describe the sample version and the population version (for $n=\infty$) of the AdaBoost algorithm.

1. Let $F_n^0 = 0$.
2. For $t = 1, 2, \dots,$
   (a) ("Weak learning" at step t). Let $f_n^t = \arg\max_{f\in H} |\varepsilon_n^t(f) - 0.5|$, then set

$$\alpha_n^t = 0.5 \ln\left\{\frac{1-\varepsilon_n^t(f_n^t)}{\varepsilon_n^t(f_n^t)}\right\}.$$

Here the "weighted training errors" are

$$\varepsilon_n^t(f) = n^{-1}\sum_{j=1}^n \left\{\frac{e^{-F_n^{t-1}(X_j)Z_j}}{n^{-1}\sum_{k=1}^n e^{-F_n^{t-1}(X_k)Z_k}}\right\} I\{f(X_j) \neq Z_j\}$$

and

$$\varepsilon_\infty^t(f) = E\left\{\frac{e^{-F_\infty^{t-1}(X)Z}}{Ee^{-F_\infty^{t-1}(X)Z}}\right\} I\{f(X) \neq Z\}.$$

(b) (Linear combination of "weak hypotheses"). Let $F_n^t = F_n^{t-1} + \alpha_n^t f_n^t$.

The resulting prediction at step t is $\operatorname{sgn}\circ F_n^t$. [Note that in the case of a negation-closed H ($f \in H$ whenever $-f \in H$), step 2(a) equivalently uses f_n^t to minimize the weighted training error $\varepsilon_n^t(f)$.]

In this paper, we will use the following notational simplification:

$$\Delta_n^{t-1}(f) = |\tilde{\Delta}_n^{t-1}(f)| \qquad \text{where } \tilde{\Delta}_n^{t-1}(f) = n^{-1}\sum_{j=1}^{n} e^{-Z_j F_n^{t-1}(X_j)} Z_j f(X_j),$$

$$\Delta_\infty^{t-1}(f) = |\tilde{\Delta}_\infty^{t-1}(f)| \qquad \text{where } \tilde{\Delta}_\infty^{t-1}(f) = Ee^{-ZF_\infty^{t-1}(X)} Zf(X),$$

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

Then it follows that $\Delta_n^{t-1}(f)/C_n(F_n^{t-1}) = 2|\varepsilon_n^t(f) - 0.5|$, so we can rewrite the "weak learning" step 2(a) at time t as 2(a′):

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

[We use $f_n^t = \arg\max_{f\in H} \Delta_n^{t-1}(f)$ to denote an approximate maximizer satisfying $\Delta_n^{t-1}(f_n^t) = \sup_{f\in H} \Delta_n^{t-1}(f) + o_p(1)$ and having a consistent limit. We will see that f_n^t consistently estimates f_∞^t, a maximizer of $\Delta_\infty^{t-1}(f)$.]

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

We will use the following regularity conditions. Conditions (Ia) and (Ib) are on the joint distribution of (X, Y), while (IIa)–(IIc) are on the base hypothesis space H.

(Ia) (Distribution of predictor). The distribution of X is assumed to be absolutely continuous with respect to Lebesgue measure on $[0, 1]^d$.

(Ib) (Continuity of log odds). The function $F_{\mathrm{B}}(\cdot)$ is continuous on $[0, 1]^d$.

(IIa) (Completeness of base hypothesis space). The linear span of H is complete in $L_2(P_X)$ on $[0, 1]^d$. That is, for any function g such that $\|g\|_{L^2(P_X)} \equiv \sqrt{\int_{[0,1]^d} g(x)^2 P_X(dx)} < \infty$ and for any $\varepsilon > 0$, there exists a linear combination

$\sum_{s=1}^{m} \alpha_s f_s(\cdot)$ for some $m \in \{1, 2, \dots\}$ such that $f_s \in H$ and $\alpha_s \in \Re$ and $\|g(x) - \sum_{s=1}^{m} \alpha_s f_s(x)\|_{L^2(P_X)} < \varepsilon$.

(IIb) (Finite VC dimension of base hypothesis space). The VC (Vapnik–Chervonenkis) dimension of H is finite, that is, $VC(H) < \infty$. [For the concept of VC dimension, see, e.g., Anthony and Biggs (1992), Chapter 7.]

(IIc) (Compactness of base hypothesis space). The base hypothesis space H is a compact set of ±1-valued functions on $[0, 1]^d$ in the $L_2(P_X)$ metric.

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

REMARKS.

1. Conditions (Ia), (Ib) and (IIa) ensure that for any sequence (allowing possible multiple solutions) F_∞^t of fits from the population version of AdaBoost, one has $\lim_{t\to\infty} \|F_\infty^t - F_B\|_{L^2(P_X)} = 0$, due to Theorem 3 of Breiman (2000). This guarantees that at least at $n = \infty$ AdaBoost does the right thing and gives the optimal Bayes prediction. More preliminary sufficient conditions for (IIa) are given in Breiman (2000), for example, ±1 trees with number of terminal nodes (leaves) exceeding the dimension of X satisfy (IIa).
2. Conditions (IIb) and (IIc) are used to prove that the large-n case is close to the population version at finite time t. Condition (IIb) typically holds. Condition (IIc) holds in common situations when H is "continuously parameterized on a compact space"—see Lemma 9. Examples include the set of step functions in one dimension $H_{\text{step}} = \{\operatorname{sgn}(\cdot - \theta) : \theta \in [0, 1]\}$ as well as the set of linear perceptrons in the case $d > 1$. The space $H_{\text{tree}}(T) =$ closure of the set of ±1 trees with T leaves (i.e., the set of ±1 trees with T leaves or less) also is compact for every finite T. This follows because $H_{\text{tree}}(T)$ is a closed subset of a compact set $H_{\text{rectangle}} = \{f : f = \sum_{k=1}^{K} \eta_k \prod_{j=1}^{d} I[x_j \le \xi_k^j],\ \eta_k\text{'s and } \xi_k^j\text{'s} \in [-1, 1]\}$ with some finite K. See Lemma 10 for the compactness of $H_{\text{rectangle}}$. Note that, for $H_{\text{tree}}(T)$, the completeness condition (IIa) [for $T > d$, see Breiman (2000)] and condition (IIb) (for any finite T) are also satisfied. Therefore, the base hypothesis space $H_{\text{tree}}(T)$ with $T > d$ satisfies all the conditions (IIa)–(IIc) and the consistency of AdaBoost holds whenever (Ia) and (Ib) hold.
3. Condition (Ib) also implies that the coefficients of the population AdaBoost are well defined; that is, $|\alpha_\infty^s| < \infty$ for all s, or, equivalently, $2\delta_\infty^s \ne \pm 1$, the singularities of
$$\alpha_\infty^s = \frac{1}{2}\log\left(\frac{1 + 2\delta_\infty^s}{1 - 2\delta_\infty^s}\right).$$
Therefore, the population fit $F_\infty^t(X)$ is well defined and $|F_\infty^t(X)| \le \sum_{s=1}^{t} |\alpha_\infty^s| < \infty$. See Preparatory Lemma 1. See also Preparatory Lemma 2

for the definition and existence of certain expectations used later in the proof. Another implication of (Ib) is that the population version of the criterion function is continuous at each step of "weak learning"; that is, for any $s = 1, 2, \dots,$ $\Delta_\infty^s(f)$ is a continuous mapping from the $L^2(P_X)$ functions on $[0, 1]^d$ to $\Re$. See Lemma 7.

4. The proof is nonconstructive and we do not know what rate t_n can take. Presumably some t_n that increases to ∞ very slowly will work. This is because, as $t_n \to \infty$, the "approximation error" is related to $\|F_\infty^{t_n} - F_{\mathrm{B}}\|_{L_2(P_X)}$, which goes to 0. 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.

2.1. *Some preparatory lemmas.*

PREPARATORY LEMMA 1. *Condition* (Ib) *implies that the coefficients of the population AdaBoost are well defined, that is,* $|\alpha_\infty^t| < \infty$ *for all* t, *or, equivalently,* $2\delta_\infty^t \notin \{\pm 1\}$, *the singularities of*

$$\alpha_\infty^t = \frac{1}{2}\log\left(\frac{1 + 2\delta_\infty^t}{1 - 2\delta_\infty^t}\right).$$

Therefore, the population fit $F_\infty^t(X)$ *is well defined and* $|F_\infty^t(X)| \le \sum_{s=1}^t |\alpha_\infty^s| < \infty$.

PROOF. Note that condition (Ib) or $F_{\mathrm{B}}(\cdot)$ being continuous on $[0, 1]^d$ implies that $F_{\mathrm{B}}(\cdot)$ is finite, and thus $P(Z = 1|X)$ is bounded away from $\{0, 1\}$ on $[0, 1]^d$.

For $t = 1$: Suppose $|2\delta_\infty^1| = 1$. Then $1 = |EZf_\infty^1(X)| = |E\mu(X)f_\infty^1(X)|$, where $\mu(X) = E(Z|X)$. Then $E\mu f_\infty^1 = -1$ or $+1$. So $0 = E(1 - \mu f_\infty^1) = E|1 - \mu f_\infty^1|$ or $0 = E(1 + \mu f_\infty^1) = E|1 + \mu f_\infty^1|$, noting that $(1 \pm \mu f_\infty^1) \ge 0$. So $\mu f_\infty^1 = 1$ with probability 1 (w.p.1) or $\mu f_\infty^1(X) = -1$ w.p.1. So $2P(Z = 1|X) - 1 = \mu(X) \in \{\pm 1\}$ with probability 1. So $P(Z = 1|X) \in \{0, 1\}$ with probability 1, conflicting with the implication of condition (Ib). So (Ib) implies that $|2\delta_\infty^1| < 1$.

Now perform induction. Suppose that, for all $s = 1, \dots, t - 1$, we have $|2\delta_\infty^s| < 1$. Then all $|\alpha_\infty^s| < \infty$ and all $|F_\infty^s(X)| \le \sum_{r=1}^s |\alpha_\infty^r| < \infty$ for $s = 1, \dots, t - 1$. Suppose that (I) holds but $|2\delta_\infty^t| = 1$. Then $1 = |EW^tZf_\infty^t(X)|$, implying that $EW^tZf_\infty^t = -1 = -EW^t$ or $= +1 = +EW^t$, where $W^t = e^{-ZF_\infty^{t-1}(X)}/Ee^{-ZF_\infty^{t-1}(X)}$ is well defined and $\in (0, 1]$ [due to the boundedness of $F_\infty^{t-1}(X)$] and has unit expectation. So $0 = EW^t(1 - Zf_\infty^t) = EW^t|1 - Zf_\infty^t|$ or $0 = EW^t(1 + Zf_\infty^t) = EW^t|1 + Zf_\infty^t|$, noting that $W^t(1 \pm Zf_\infty^t) \ge 0$. So $Zf_\infty^t(X) = 1$ w.p.1 or $Zf_\infty^t(X) = -1$ w.p.1, implying that $1 = |EZf_\infty^t(X)| = |E\mu(X)f_\infty^t(X)|$. Then similar to the proof for $t = 1$, we have $P(Z = 1|X) \in$

$\{0, 1\}$ with probability 1, conflicting with the implication of condition (Ib). So (Ib) implies that $|2\delta_\infty^s| < 1$ for $s = t$ as well. □

We denote $S = (X_i, Z_i)_{i=1}^n$ and denote $Eg(X, Z, S) \equiv E_{X,Z|S}g(X, Z, S) = E_{X,Z}g(X, Z, S)$ as the integration over (X, Z) fixing S. [Note that S and (X, Z) are independent.] Then we have the following existence theorem of the integrations that will be used later.

PREPARATORY LEMMA 2. *For all $t = 1, 2, \dots$ and $n = 1, 2, \dots, \infty$, we have*

$$\begin{aligned} E\exp(-ZF_n^{t-1}(X)) &\in \exp\left(\pm\sum_{s=1}^{t-1}|\alpha_n^s|\right) \\ &\subset \left(\exp\left(-\sum_{s=1}^{t-1}|\alpha_\infty^s|\right)\exp\left(-\sum_{s=1}^{t-1}|\alpha_n^s - \alpha_\infty^s|\right),\right. \\ &\qquad \left.\exp\left(\sum_{s=1}^{t-1}|\alpha_\infty^s|\right)\exp\left(\sum_{s=1}^{t-1}|\alpha_n^s - \alpha_\infty^s|\right)\right). \end{aligned}$$

Therefore, $E\exp(-ZF_\infty^{t-1}(X)) \in \exp(\pm\sum_{s=1}^{t-1}|\alpha_\infty^s|) \subset (0, \infty)$ under condition (Ib) due to Preparatory Lemma 1, and $E\exp(-ZF_n^{t-1}(X)) \in \exp(\pm\sum_{s=1}^{t-1}|\alpha_\infty^s|)\{1 + o_p(1)\}$ in the inductive proof of Lemma 2 later.

PROOF OF PREPARATORY LEMMA 2. Note that $|-ZF_n^{t-1}(X)| = |Z \times \sum_{s=1}^{t-1}\alpha_n^s f_n^s(X)| \le \sum_{s=1}^{t-1}|\alpha_n^s| \le \sum_{s=1}^{t-1}|\alpha_\infty^s| + \sum_{s=1}^{t-1}|\alpha_n^s - \alpha_\infty^s|$ leads to the proof. □

2.2. *Proof of 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 (Ia)–(IIc) in the case of AdaBoost.

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 ψ is strictly increasing and continuously differentiable with derivative ψ' bounded on $\Re^1$, and $\psi(0) = 0.5$ [such that $\mathrm{sgn}(F) = \mathrm{sgn}(p - 1/2)$]. Here F is to be estimated by a sequential (t-step) additive fit $F_n^t = \sum_{s=1}^t \alpha_n^s f_n^s$, $\alpha_n^s \in \Re$, $f_n^s \in H$ (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 $F_{\mathrm{B}} = \psi^{-1} \circ p$. As an example, in AdaBoost, $\psi = e^{2F}/(1 + e^{2F})$, $|\psi'| \le 0.5$, and $F_{\mathrm{B}}(x) = \frac{1}{2}\log\{p(x)/(1 - p(x))\}$.

LEMMA 1. $L_\infty(F_n^t) - L_\infty(F_{\mathrm{B}}) \le 2\|\psi'\|_\infty \|F_n^t - F_{\mathrm{B}}\|_{L^2(P_X)}$.

This is a variant of Corollary 6.2 of Devroye, Györfi and Lugosi (1996), obtained by a first-order Taylor expansion. Here $\|\psi'\|_\infty = \sup_{x\in\Re}|\psi'(x)|$. 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 nonstochastic sequence F_∞^t of functions on the domain of X, independent of n, such that*:

(i) $\|F_n^t - F_\infty^t\| \le b_n^t$ *with probability* $P_S \ge 1 - a_n^t$.

(ii) $\|F_\infty^t - F_{\mathrm{B}}\| \le c^t$, *where* a_n^t, b_n^t, c^t *are nonnegative*, $c^t \to 0$ *as* $t \to \infty$, *and* a_n^t, $b_n^t \to 0$ *as* $n \to \infty$ $\forall t$. *Then there is a sequence* $t(n)$ *such that* $E_S L_\infty(F_n^{t(n)}) - L_\infty(F_{\mathrm{B}}) = o_n(1)$.

In the case of AdaBoost, F_∞^t was taken to be a population fit and (ii) is guaranteed via conditions (Ia), (Ib) and (IIa) 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 OF PROPOSITION 1. Since the nonnegative sequences $a_n^t, b_n^t \to 0$ as $n \to \infty$ $\forall t$, there exists a sequence $t(n) \to \infty$ sufficiently slow, such that $a_n^{t(n)}, b_n^{t(n)} \to 0$ as $n \to \infty$. Also we have $c^{t(n)} \to 0$.

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

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

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

$$\begin{aligned} Y_n &\le (\|F_n^{t(n)} - F_\infty^{t(n)}\| + \|F_\infty^{t(n)} - F_{\mathrm{B}}\|) 2\|\psi'\|_\infty \\ &\le \varepsilon, \end{aligned}$$

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

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. Condition (i) of Proposition 1 in the previous section is established immediately via the following proposition.

PROPOSITION 2 (Convergence of the sequential fits). *Suppose conditions* (Ia)–(IIc) *hold. Then, for any* $t = 1, 2, \dots$ *and sample AdaBoost fit* F_n^t, *there exists population AdaBoost fit* F_∞^t *such that, with probability* $P_S \ge 1 - a_n^t$, $\|F_n^t - F_\infty^t\| \le b_n^t$, *where* $a_n^t, b_n^t \to 0$ *as* $n \to \infty$, *or, equivalently,* $\|F_n^t - F_\infty^t\| = o_p(1)$ *as* $n \to \infty$.

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

LEMMA 2 (Convergence step by step). *Suppose conditions* (Ia)–(IIc) *hold. Then there exists a population version of "weak learning"* f_∞^s *for each step* $s = 1, 2, \dots$, *such that* $\|f_n^s - f_\infty^s\| = o_p(1)$ *and* $|\alpha_n^s - \alpha_\infty^s| = o_p(1)$ *as* $n \to \infty$.

Now let us prove this lemma. This is done by a method of induction that uses the following secondary lemmas, where we suppose that conditions (Ia)–(IIc) hold. These secondary lemmas will be proved in the next section.

LEMMA 3. $D_{n,\infty}^t \le Q_{1n}^{t-1} + R_n^{t-1}$, *where* $D_{n,\infty}^t = \sup_{f \in H} |\Delta_n^{t-1}(f) - \Delta_\infty^{t-1}(f)|$, $Q_{1n}^{t-1} = \sup_{f \in H} |n^{-1} \sum_{i=1}^n e^{-Z_i F_n^{t-1}(X_i)} f(X_i) Z_i - E e^{-Z F_n^{t-1}(X)} \times f(X) Z|$ *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* $P_S[Q_{1n}^{t-1} \le U_n^{t-1}] \ge 1 - 8n^{-\beta}$ *and* $P_S[Q_{2n}^{t-1} \le U_n^{t-1}] \ge 1 - 8n^{-\beta}$. *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*

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

LEMMA 5. $R_n^t \le \exp(|\alpha_\infty^t| + |\alpha_n^t - \alpha_\infty^t|)(R_n^{t-1} + \exp(\sum_{s=1}^{t-1} |\alpha_\infty^s|)|\alpha_n^t - \alpha_\infty^t| + 0.5|\alpha_\infty^t| \exp(\sum_{s=1}^{t-1} |\alpha_\infty^s|) \|f_n^t - f_\infty^t\|^2)$.

LEMMA 6. $|2\delta_n^t - 2\delta_\infty^t| \le \{C_n(F_n^{t-1})\}^{-1}(Q_{1n}^{t-1} + R_n^{t-1} + 2\delta_\infty^t Q_{2n}^{t-1} + 2\delta_\infty^t R_n^{t-1} + 0.5 \exp(\sum_{s=1}^{t-1} |\alpha_\infty^s|) \|f_n^t - f_\infty^t\|^2)$.

LEMMA 7. *Another implication of* (Ib) *is that the population version of the criterion function at each step of "weak learning" is continuous*; *that is*, *for any* $t = 1, 2, \ldots$, $\Delta_\infty^t(f)$ *is a continuous mapping from the* $L^2(P_X)$*-functions on* $[0, 1]^d$ *to* $\Re$.

LEMMA 8. *Assume that conditions* (Ib) *and* (IIc) *hold so that* $\Delta_\infty^{t-1}(f)$ *is continuous* (*see Lemma* 7) *and* H *is compact. Consider any approximate maximizer* $f_n^t = \arg\max_{f \in H} \Delta_n^{t-1}(f)$ *such that* $\Delta_n^{t-1}(f_n^t) = \sup_{f \in H} \Delta_n^{t-1}(f) + o_p(1)$. *If* $D_{n,\infty}^t \equiv \sup_{f \in H} |\Delta_n^{t-1}(f) - \Delta_\infty^{t-1}(f)| = o_p(1)$ *as* $n \to \infty$, *then*, *for any* $\varepsilon > 0$, *with probability tending to* 1 *as* $n \to \infty$, *there exist maximizers* $f_\infty^t = \arg\max_{f \in H} \Delta_\infty^{t-1}(f)$ [*such that* $\Delta_\infty^{t-1}(f_\infty^t) = \sup_{f \in H} \Delta_\infty^{t-1}(f)$] *such that* $\|f_n^t - f_\infty^t\| < \varepsilon$.

PROOF OF LEMMA 2. For $s = 1$,

$$D_{n,\infty}^1 = \sup_{f \in H} \left| \left| n^{-1} \sum_{i=1}^n f(X_i) Z_i \right| - |Ef(X)Z| \right|$$
$$\le \sup_{f \in H} \left| n^{-1} \sum_{i=1}^n f(X_i) Z_i - Ef(X)Z \right|,$$

which is at most $\sqrt{(32 \log n/n)\{VC(H) + \beta\} + (32/n) VC(H) \log\{e/VC(H)\}}$ with probability at least $1 - 8n^{-\beta}$, for any $\beta > 0$, by the VC uniform bounding technique (Lemmas 3 and 4). Therefore, $D_{n,\infty}^1 \equiv \sup_{f \in H} |\Delta_n^0(f) - \Delta_\infty^0(f)| = o_p(1)$. Now conditions (Ib) and (IIc) imply the continuity of $\Delta_\infty(f)$ (see Lemma 7) and the compactness of H. Then $\sup_{f \in H} |\Delta_n^0(f) - \Delta_\infty^0(f)| = o_p(1)$ implies that for $f_n^1 = \arg\max_{f \in H} \Delta_n^0(f)$ we have $f_\infty^1 = \arg\max_{f \in H} \Delta_\infty^0(f)$ such that $\|f_n^1 - f_\infty^1\| = o_p(1)$. (See also Lemma 8.) Note also that $Q_{1n}^0 = o_p(1)$, $Q_{1n}^0 = 0$, $R_n^0 = 0$. Therefore, $|2\delta_n^1 - 2\delta_\infty^1| = o_p(1)$ by Lemma 6. Then, due to the continuity of the relationship between α_n^s and $2\delta_n^s$ for all n and s (note that singularities of the relationship are avoided due to Preparatory Lemma 1), we have $|\alpha_n^1 - \alpha_\infty^1| = o_p(1)$ also. Therefore, at $s = 1$, the results of the lemma hold and $R_n^0 = 0 = o_p(1)$.

Now suppose that, for all $s = 1, \ldots, t - 1$, the results of the lemma hold and $R^0, \ldots, R^{t-2}$ are all $o_p(1)$. Then we have $U_n^{t-1} = o_p(1)$, implying that Q_{1n}^{t-1}, Q_{2n}^{t-1} and R_n^{t-1} are all $o_p(1)$ due to Lemmas 4 and 5. Then $D_{n,\infty}^t = \sup_{f \in H} |\Delta_n^{t-1}(f) - \Delta_\infty^{t-1}(f)|$ is $o_p(1)$ also, due to Lemma 3. So for $f_n^t =$

$\arg\max_{f\in H} \Delta_n^{t-1}(f)$ we have $f_\infty^t = \arg\max_{f\in H} \Delta_\infty^{t-1}(f)$ such that $\|f_n^t - f_\infty^t\| = o_p(1)$. Then we apply Lemma 6. Note that

$$\begin{aligned}
\left|C_n(F_n^{t-1}) - Ee^{-ZF_\infty^{t-1}}\right| &= \left|n^{-1}\sum_1^n e^{-Z_iF_n^{t-1}(X_i)} - Ee^{-ZF_\infty^{t-1}}\right| \\
&\le \left|n^{-1}\sum_1^n e^{-Z_iF_n^{t-1}(X_i)} - Ee^{-ZF_n^{t-1}}\right| \\
&\quad + \left|Ee^{-ZF_n^{t-1}} - Ee^{-ZF_\infty^{t-1}}\right| \\
&\le \left|n^{-1}\sum_1^n e^{-Z_iF_n^{t-1}(X_i)} - Ee^{-ZF_n^{t-1}}\right| \\
&\quad + E\left|e^{-ZF_n^{t-1}} - e^{-ZF_\infty^{t-1}}\right| \\
&= Q_{2n}^{t-1} + R_n^{t-1} = o_p(1).
\end{aligned}$$

Then, on the right-hand side of Lemma 6, the factor $\{C_n(F_n^{t-1})\}^{-1} = \{Ee^{-ZF_\infty^{t-1}} + o_p(1)\}^{-1}$, where

$$E\exp(-ZF_\infty^{t-1}) = E\exp\left(-Z\sum_{s=1}^{t-1}\alpha_\infty^s f_\infty^s(X)\right) \in E\exp\left(\pm\sum_{s=1}^{t-1}|\alpha_\infty^s|\right) \in (0,\infty)$$

due to the finite boundedness of $|\alpha_\infty^s|$ for all s (Preparatory Lemma 1). The other factor on the right-hand side of Lemma 6 is $o_p(1)$. Therefore, $|2\delta_n^t - 2\delta_\infty^t| = o_p(1)$ and $|\alpha_n^t - \alpha_\infty^t| = o_p(1)$ follows by a continuity argument. This completes the induction proof for the lemma. □

5. Proof of secondary lemmas.

PROOF OF LEMMA 3. Note that

$$\begin{aligned}
D_{n,\infty}^t &= \sup_{f\in H}\left||\tilde\Delta_n^{t-1}(f)| - |\tilde\Delta_\infty^{t-1}(f)|\right| \\
&\le \sup_{f\in H}|\tilde\Delta_n^{t-1}(f) - \tilde\Delta_\infty^{t-1}(f)| \\
&= \sup_{f\in H}\left|\left\{n^{-1}\sum_{i=1}^n e^{-Z_iF_n^{t-1}(X_i)}f(X_i)Z_i - Ee^{-ZF_n^{t-1}(X)}f(X)Z\right\}\right. \\
&\qquad\qquad \left. + \left\{E\left(e^{-ZF_n^{t-1}(X)} - e^{-ZF_\infty^{t-1}(X)}\right)f(X)Z\right\}\right|,
\end{aligned}$$

apply the triangle inequality and note that

$$\sup_{f\in H}\left|E\left(e^{-ZF_n^{t-1}(X)}-e^{-ZF_\infty^{t-1}(X)}\right)f(X)Z\right|\le E\left|e^{-ZF_n^{t-1}(X)}-e^{-ZF_\infty^{t-1}(X)}\right|.$$

This shows the further upper bound $Q_{1n}^{t-1}+R_n^{t-1}$. □

PROOF OF LEMMA 4. By the triangle inequality,

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

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

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

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

and

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

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

$$\exp\left(\sum_{s=1}^{t-1}|\alpha_\infty^s|\right)\exp\left(\sum_{s=1}^{t-1}|\alpha_n^s-\alpha_\infty^s|\right)\sum_{s=1}^{t-1}|\alpha_n^s-\alpha_\infty^s|$$

by applying a first-order Taylor expansion. Note that T_1 is bounded above by

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

with probability at least $1-8n^{-\beta}$, for any $\beta>0$, by applying the following proposition, which is proved in the same way as the VC result, Theorem 12.5 of Devroye, Györfi and Lugosi (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 $s(\Phi, n) = \max_{W_1^n} \operatorname{card}\{\varphi(W_1^n) : \varphi \in \Phi\}$. Then*

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

for any $\varepsilon > 0$ and

$$P\left[\sup_{\varphi\in\Phi}\left|n^{-1}\sum_{i=1}^{n} g\bigl(\varphi(W_i), W_i\bigr) - Eg\bigl(\varphi(W), W\bigr)\right| \right.$$
$$\left. \le M\sqrt{(32/n)\{\log s(\Phi, n) + \beta \log n\}}\right] \ge 1 - 8n^{-\beta}$$

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

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

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

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

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. □

PROOF OF LEMMA 5. Note that

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

Note that

$$\bigl|e^{-Z\alpha_n^t f_n^t}\bigr| \le e^{|\alpha_n^t|} \le e^{|\alpha_\infty^t| + |\alpha_n^t - \alpha_\infty^t|},$$

and use a first-order Taylor expansion to obtain

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

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 $|f_n^t|=1$, $|f_n^t-f_\infty^t|=0.5(f_n^t-f_\infty^t)^2$,

$$
\exp\bigl(-Z(F_\infty^{t-1}+\tilde{\alpha}_n^t\tilde{f}_n^t)\bigr)\le\exp\left(\sum_{s=1}^{t-1}|\alpha_\infty^s|+|\alpha_\infty^t|+|\alpha_n^t-\alpha_\infty^t|\right).
$$

Combining these results, we then get

$$
\begin{aligned}
&E\bigl|\exp(-ZF_n^t)-\exp(-ZF_\infty^t)\bigr|\\
&\quad\le\exp\bigl(|\alpha_\infty^t|+|\alpha_n^t-\alpha_\infty^t|\bigr)E\bigl|\exp(-ZF_n^{t-1})-\exp(-ZF_\infty^{t-1})\bigr|\\
&\qquad+\exp\left(\sum_{s=1}^{t-1}|\alpha_\infty^s|+|\alpha_\infty^t|+|\alpha_n^t-\alpha_\infty^t|\right)\\
&\qquad\quad\times\bigl(|\alpha_n^t-\alpha_\infty^t|+0.5|\alpha_\infty^t|\|f_n^t-f_\infty^t\|_{L_2(P_X)}^2\bigr),
\end{aligned}
$$

which proves the lemma. □

PROOF OF LEMMA 6. Note that

$$
\begin{aligned}
|2\delta_n^t-2\delta_\infty^t|&=\bigl|\tilde{\Delta}_n^{t-1}(f_n^t)/C_n(F_n^{t-1})-\tilde{\Delta}_\infty^{t-1}(f_\infty^t)/C_\infty(F_\infty^{t-1})\bigr|\\
&=C_n(F_n^{t-1})^{-1}\\
&\qquad\times\bigl|\bigl(\tilde{\Delta}_n^{t-1}(f_n^t)-\tilde{\Delta}_\infty^{t-1}(f_\infty^t)\bigr)-2\delta_\infty^t\bigl(C_n(F_n^{t-1})-C_\infty(F_\infty^{t-1})\bigr)\bigr|.
\end{aligned}
$$

Now apply the triangle inequality. Note that

$$
|\tilde{\Delta}_n^{t-1}(f_n^t)-\tilde{\Delta}_\infty^{t-1}(f_\infty^t)|\le Q_{1n}^{t-1}+R_n^{t-1}+0.5\exp\left(\sum_{s=1}^{t-1}|\alpha_\infty^s|\right)\|f_n^t-f_\infty^t\|^2
$$

by Lemma 3 and by using the triangle inequality we also have

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

Combining these results, we have the proof of the lemma. □

PROOF OF LEMMA 7. For any two $L^2(P_X)$ functions f and f' on $[0,1]^d$, we have

$$\begin{aligned}|\Delta_\infty^t(f)-\Delta_\infty^t(f')| &= \big||\tilde{\Delta}_\infty^t(f)|-|\tilde{\Delta}_\infty^t(f')|\big|\\ &\le |\tilde{\Delta}_\infty^t(f)-\tilde{\Delta}_\infty^t(f')|\\ &= \big|Ee^{-ZF_\infty^t(X)}Z\{f(X)-f'(X)\}\big|\\ &\le \sqrt{Ee^{-2ZF_\infty^t(X)}}\sqrt{E\{f(X)-f'(X)\}^2}\\ &\le \exp\Bigg(\sum_{s=1}^t|\alpha_\infty^s|\Bigg)\|f-f'\|_{L^2(P_X)},\end{aligned}$$

which implies continuity when all the $|\alpha_\infty^s|$'s are finite under (Ib) due to Preparatory Lemma 1. □

PROOF OF LEMMA 8. We omit the time index since the statement above can be for any t.

For any $\varepsilon > 0$, define $\Phi(\varepsilon) = \{f \in H : \|f - f_\infty\|_{L^2(P_X)} \ge \varepsilon \ \forall f_\infty = \arg\max_{f\in H}\Delta_\infty(f)\}$, which is the set of elements in H that are at a distance of ε or more away from the set of maximizers. Then $\Phi(\varepsilon)$ is compact since H is compact. Note that $\delta(\varepsilon) \equiv \sup_{f\in H}\Delta_\infty(f) - \sup_{f\in\Phi(\varepsilon)}\Delta_\infty(f) > 0$. This is because $\Delta_\infty(\cdot)$ is continuous and $\Phi(\varepsilon)$ is compact, and thus $\sup_{f\in\Phi(\varepsilon)}\Delta_\infty(f)$ is attained somewhere in $\Phi(\varepsilon)$, resulting in a value strictly smaller than $\sup_{f\in H}\Delta_\infty(f)$.

Consider now any approximate maximizer $f_n = \arg\max_{f\in H}\Delta_n(f)$ such that $\Delta_n(f_n) = \sup_{f\in H}\Delta_n(f) + o_p(1)$. Let A_n be the event $\|f_n - f_\infty\| \ge \varepsilon$ for all f_∞. [Here a maximizer $f_\infty \in H$ is such that $\Delta_\infty(f_\infty) = \sup_{f\in H}\Delta_\infty(f)$, which exists due to the continuity of Δ_∞ and the compactness of H.]

Assuming event A_n and picking any maximizer f_∞ in the following argument, we have

$$\begin{aligned}\Delta_\infty(f_n) - \Delta_\infty(f_\infty) &\le \sup_{f\in\Phi(\varepsilon)}\Delta_\infty(f) - \Delta_\infty(f_\infty)\\ &= \sup_{f\in\Phi(\varepsilon)}\Delta_\infty(f) - \sup_{f\in H}\Delta_\infty(f) = -\delta(\varepsilon).\end{aligned}$$

Then we have

$$\begin{aligned}\Delta_n(f_n) - \sup_{f\in H}\Delta_n(f) &\le \Delta_n(f_n) - \Delta_n(f_\infty)\\ &= \{\Delta_n(f_n) - \Delta_\infty(f_n)\}\\ &\quad + \{\Delta_\infty(f_n) - \Delta_\infty(f_\infty)\} + \{\Delta_\infty(f_\infty) - \Delta_n(f_\infty)\}\\ &\le -\delta(\varepsilon) + 2\sup_{f\in H}|\Delta_n(f) - \Delta_\infty(f)|.\end{aligned}$$

Therefore, $2\sup_{f\in H}|\Delta_n(f)-\Delta_\infty(f)|+\sup_{f\in H}\Delta_n(f)-\Delta_n(f_n)\geq\delta(\varepsilon)>0$, which will be referred to as event B_n.

Therefore, we have shown that A_n implies B_n, which further implies that $P[A_n]\leq P[B_n]\to 0$. [Note that $\sup_{f\in H}|\Delta_n(f)-\Delta_\infty(f)|$ and $\sup_{f\in H}\Delta_n(f)-\Delta_n(f_n)$ are both $o_p(1)$ by assumptions of the lemma, so $P[B_n]\to 0$.] Therefore, with probability tending to 1, the complement of A_n is true, which proves the lemma. □

6. Lemmas related to condition (IIc).

LEMMA 9. *Suppose $H=\{f:f=\operatorname{sgn}\varphi(\cdot,\theta),\ \theta\in\Theta\}$, $\varphi(x,\theta)$ is continuous in θ for all $x\in[0,1]^d$, $P_X[\varphi(X,\theta)=0]=0$ for all $\theta\in\Theta$ and Θ is a compact set in a metric space. Then H is compact in the metric space of $L_2(P_X)$ functions on* $[0,1]^d$.

PROOF. First, we show that $F(\theta)=\operatorname{sgn}\varphi(\cdot,\theta)$ is a continuous mapping from Θ to $L_2(P_X)$-functions on $[0,1]^d$. Then the lemma follows since $H=F(\Theta)$ is a continuous image of a compact set Θ.

The mapping $F(\theta)$ is continuous for the following reasons. Consider a sequence θ_k in Θ converging to any $\theta\in\Theta$. Then

$$\begin{aligned}
&\|F(\theta_k)-F(\theta)\|^2_{L_2(P_X)}\\
&\quad=\int_{[0,1]^d}\{\operatorname{sgn}\varphi(x,\theta_k)-\operatorname{sgn}\varphi(x,\theta)\}^2P_X(dx)\\
&\quad=4P_X[\operatorname{sgn}\varphi(X,\theta_k)\neq\operatorname{sgn}\varphi(X,\theta)]\\
&\quad=4P_X[\operatorname{sgn}\varphi(X,\theta_k)\neq\operatorname{sgn}\varphi(X,\theta),\ \varphi(X,\theta)=0]\\
&\qquad+4P_X[\operatorname{sgn}\varphi(X,\theta_k)\neq\operatorname{sgn}\varphi(X,\theta),\ \varphi(X,\theta)>0]\\
&\qquad+4P_X[\operatorname{sgn}\varphi(X,\theta_k)\neq\operatorname{sgn}\varphi(X,\theta),\ \varphi(X,\theta)<0]\\
&\quad=4P_X[\operatorname{sgn}\varphi(X,\theta_k)\neq\operatorname{sgn}\varphi(X,\theta),\ \varphi(X,\theta)=0]\\
&\qquad+4P_X[\operatorname{sgn}\varphi(X,\theta_k)=-1,\ \varphi(X,\theta)>0]\\
&\qquad+4P_X[\operatorname{sgn}\varphi(X,\theta_k)=1,\ \varphi(X,\theta)<0]\\
&\quad\leq4P_X[\varphi(X,\theta)=0]\\
&\qquad+4P_X[\varphi(X,\theta_k)\leq0,\ \varphi(X,\theta)>0]+4P_X[\varphi(X,\theta_k)\geq0,\ \varphi(X,\theta)<0].
\end{aligned}$$

The first term has been assumed to be 0. In the second term, $P_X[\varphi(X,\theta_k)\leq 0,\ \varphi(X,\theta)>0]=0$ for all sufficiently large k since $\varphi(X,\theta_k)\to\varphi(X,\theta)>0$ due to the continuity. Similarly, the third term converges to 0 also. Therefore, $\|F(\theta_k)-F(\theta)\|^2_{L_2(P_X)}\to 0$ as a result of $\theta_k\to\theta$ showing the continuity. □

LEMMA 10. *The set of "rectangular" functions on* $[0,1]^d$, $H_{\text{rectangle}} = \{f : f = \sum_{k=1}^K \eta_k \prod_{j=1}^d I[x_j \le \xi_k^j]$, $\eta_k' s$ *and* $\xi_k^{i\prime} s \in [-1,1]\}$ *with some finite* K, *is compact in* $L_2(P_X)$ *under condition* (Ia).

PROOF. Denote by θ the parameter vector including all components of η_k's and ξ_k^j's and let Θ be the corresponding compact set of θ's when the components vary in $[-1,1]$. Denote by $F(\theta)$ and $F(\theta')$ two elements in $H_{\text{rectangle}}$, where θ' is the parameter vector including all components of (η_k')'s and $\{(\xi')_k^j\}$'s. Then, by recursive applications of the triangle inequality, one can show that

$$\begin{aligned}
\|F(\theta') - F(\theta)\| &= \left\| \sum_{k=1}^K \eta_k' \prod_{j=1}^d I[x_j \le (\xi')_k^j] - \sum_{k=1}^K \eta_k \prod_{j=1}^d I[x_j \le \xi_k^j] \right\| \\
&= \left\| \sum_{k=1}^K \eta_k' \prod_{j=1}^d I[x_j \le (\xi')_k^j] - \sum_{k=1}^K \eta_k \prod_{j=1}^d I[x_j \le (\xi')_k^j] \right. \\
&\qquad \left. + \sum_{k=1}^K \eta_k \prod_{j=1}^d I[x_j \le (\xi')_k^j] - \sum_{k=1}^K \eta_k \prod_{j=1}^d I[x_j \le \xi_k^j] \right\| \\
&\le \sum_{k=1}^K |\eta_k' - \eta_k| + \sum_{k=1}^K \left\| \prod_{j=1}^d I[x_j \le (\xi')_k^j] - \prod_{j=1}^d I[x_j \le \xi_k^j] \right\| \\
&= \sum_{k=1}^K |\eta_k' - \eta_k| \\
&\qquad + \sum_{k=1}^K \left\| (I[x_1 \le (\xi')_k^1] - I[x_1 \le \xi_k^1]) \prod_{j=2}^d I[x_j \le (\xi')_k^j] \right. \\
&\qquad\qquad \left. + I[x_1 \le \xi_k^1] \left(\prod_{j=2}^d I[x_j \le (\xi')_k^j] - \prod_{j=2}^d I[x_j \le \xi_k^j] \right) \right\| \\
&\le \sum_{k=1}^K |\eta_k' - \eta_k| + \sum_{k=1}^K \left\{ \| I[x_1 \le (\xi')_k^1] - I[x_1 \le \xi_k^1] \| \right. \\
&\qquad\qquad \left. + \left\| \prod_{j=2}^d I[x_j \le (\xi')_k^j] - \prod_{j=2}^d I[x_j \le \xi_k^j] \right\| \right\} \\
&\le \cdots \\
&\le \sum_{k=1}^K |\eta_k' - \eta_k| + \sum_{k=1}^K \sum_{j=1}^d \| I[x_j \le (\xi')_k^j] - I[x_j \le \xi_k^j] \|
\end{aligned}$$

$$= \sum_{k=1}^{K} |\eta'_k - \eta_k| + \sum_{k=1}^{K} \sum_{j=1}^{d} \sqrt{P[X_j \text{ between } (\xi')_k^j \text{ and } \xi_k^j]},$$

which converges to 0 if $\theta' \to \theta$ in Euclidean norm, under condition (Ia). Therefore, $F(\cdot)$ is continuous and $H_{\text{rectangle}} = F(\Theta)$ is compact due to the compactness of Θ. □

Acknowledgment. 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 Univ. Press.

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

BREIMAN, L. (2000). Some infinity theory for predictor ensembles. Technical Report 579, Dept. Statistics, Univ. California, 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. *J. Comput. System Sci.* **55** 119–139.

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

GROVE, A. J. and SCHUURMANS, D. (1998). Boosting in the limit: Maximizing the margin of learned ensembles. In *Proc. 15th National Conference on Artificial Intelligence* 692–699. AAAI Press, Menlo Park, CA.

JIANG, W. (2002). On weak base hypotheses and their implications for boosting regression and classification. *Ann. Statist.* **30** 51–73.

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

SCHAPIRE, R. E. (1999). Theoretical views of boosting. In *Computational Learning Theory*: *Proc. Fourth European Conference* 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. *Ann. Statist.* **26** 1651–1686.

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