TAIWANESE JOURNAL OF MATHEMATICS
Vol. 14, No. 6, pp. 2125-2136, December 2010
This paper is available online at http://www.tjm.nsysu.edu.tw/

CONSISTENCY OF BOOSTING UNDER NORMALITY

W. Drago Chen and C. Andy Tsao*

Abstract. Boosting is one of the important ensemble classifiers emerging in the past decade. [10] provides a statistical insight: AdaBoost can be viewed as a Newton-like updates minimizing exponential criterion. This powerful insight, however, does not address (1) whether the Newton update converges (2) whether the update procedure converge to the Bayes procedure if it does converge. Under a normal-normal setting, we cast the learning problem as a Bayesian minimization problem. It is shown that the Bayes procedure can be obtained via an iterative Newton update minimizing exponential criterion. In addition, the step sizes of AdaBoost are shown to be highly effective and lead to a one-step convergence. While our results based on strong distributional assumption, they require little conditions on the complexity of base learners nor regularization on step sizes or number of boosting iterations.

1. INTRODUCTION

Boosting is a method to construct accurate outputs by combining some simple classifiers. Roughly speaking, it is a learning procedure starts with a weak learner and re-weights the data by giving the misclassified data higher weights each time, then it takes a weighted majority vote to make final predictions.

Earlier studies show that boosting reduces the training error under the *weak base hypothesis assumption* while upper bounds of its testing errors can be obtained in some PAC (Probably Approximately Correct) sense, see, for example, [9]. More recent advancements are referred to [17] and references therein. Empirically, it has been observed that boosting is relatively resistant to overfitting in many practical applications with less noisy data. These empirical successes motivate theoretical

Received October 31, 2008, accepted March 23, 2009.
Communicated by J. C. Yao.
2000 *Mathematics Subject Classification*: 62G05, 68T05, 62C10.
Key words and phrases: Boosting, Bayesian optimization, Loss approximation, Statistical machine learning.
The research is supported by NSC-95-2118-M-259-002-MY2, Taiwan.
*Corresponding author.

investigations, see, for example, [18] and [11]. Recent studies provide a much clearer picture of Bayes consistency of boosting. [5] shows that population version boosting is Bayes consistent and [12] shows boosting is process consistent for the sample version. Other asymptotic aspects are referred to, for example, [6, 20] and [2].
[10] shows that boosting can be viewed as a stagewise Newton update minimizing exponential criterion. This interpretation catches much attention, specially from statistical community. It also inspires ensuing investigations. For example, the exponential loss approximation motivates other loss approximations and in turn lead to new variations of boosting-like algorithms. However, that study fails to address two important questions

- whether the Newton update converges?
- does the update procedure converge to the (optimal) Bayes procedure if it does converge?

In the literature, the Bayes consistency refers to the convergence, to the optimal Bayes procedure, of boosting algorithm as it iterates forever. In this study, we investigate the Bayes consistency of boosting along the line of [10] using a population version approach. The problem of consistency of population version boosting has been studied actively by, for example, [5, 1, 13, 3, 15, 14] and [19]. Recently, [2] shows the regularized AdaBoost (condition on the stopping strategy) is consistent under mild conditions on the complexity of base learners and their span. However, our approach differs from the forementioned studies in many aspects. This approach is inspired by the observation that the performances of boosting depend greatly on the data. That is, the problems of consistency, such as whether conditions on the base learners should be imposed or the necessity of early stopping, might depend on the underlying distribution. So far as we know, this aspect has not been fully explored. In this study, we impose normal-normal distributional assumption on the sample and recast the classification problem as a Bayesian optimization problem. This framework is well within statistical domain and many concepts, tools and procedures are readily available. In fact, we have shown that F_{FHT} ((6), population AdaBoost given in [10]) converges to the Bayes estimate and no regularization nor the conditions on the base learners are required.

The rest of the paper is organized as follows. Section 2 briefly reviews [10], particularly, the interpretation of boosting and highlight some advantages of this approach. Under a normal-normal setting, Section 3 studies the convergence of population version algorithms by minimizing the conditional approximate risk. The convergence of two iterative algorithms, F_{PIB} ((5), population iterative Bayes procedure) and F_{FHT}, are established and contrasted. We then generalize the results for high dimensional explanatory variables under uncorrelated multivariate normal-normal setting. Concluding comments and discussion are summarized in Section 4.

2. Boosting as Iterative Newton Update

AdaBoost, see for example, [9], is one of the "mother" boosting algorithm which inspires many variants and modifications. It captures the main structure and features of boosting algorithm. In this study, we will focus on AdaBoost for its easy exposition and simplicity for theoretical analysis. Consider the training data $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in X$ and $y_i \in \mathcal{Y} = \{\pm 1\}$.

AdaBoost

(1) Initialize $D_1(i) = w_1(i) = n^{-1}$ for $i = 1, 2, \cdots, n$.

(2) Repeat for $t = 1, 2, \cdots, T$

- Train weak base learner using the weight D_t on the data.
- Obtain the (trained) learner $f_t : \mathcal{X} \to \mathcal{Y}$ and calculate the loss

$$L[f_t(x_i), y_i] = \mathbf{1}_{[f_t(x_i) \neq y_i]}.$$

- Compute the error

$$\epsilon_t = E_{D_t} L[f_t(x_i), y_i] = \sum_{i=1}^n D_t(i) \mathbf{1}_{[f_t(x_i) \neq y_i]}$$

and

$$\beta_t = \frac{1 - \epsilon_t}{\epsilon_t}, \qquad \alpha_t = \ln \beta_t.$$

- Update the weights for $i = 1, 2, \cdots, n$.

$$w_{t+1}(i) = w_t(i) e^{\alpha_t L[f_t(x_i), y_i]} = w_t(i) \beta_t^{\mathbf{1}_{[f_t(x_i) \neq y_i]}}$$

- Normalization for $i = 1, 2, \cdots, n$.

$$D_{t+1}(i) = w_{t+1}(i) \left(\sum_{i=1}^n w_{t+1}(i) \right)^{-1}$$

(3) Output the final hypothesis

$$F(x) = \text{sign} \left[\sum_{t=1}^T \alpha_t f_t(x) \right].$$

Typically, the iteration number T is fixed beforehand. [10] provides a powerful insight to AdaBoost. Specifically, the classifier F, without loss of generality, can be assumed as a real-valued function mapping from $\mathcal{X}$ to $\Re$. Then the task of binary classification can be viewed as determining the F with the right sign minimizing

$E_{X,Y}\{\mathbf{1}_{[YF(X)<0]}\}$. So far as the sign is concerned, the final hypothesis in step (3) of the boosting algorithm can be replaced by

$$F(x) = \sum_{t=1}^{T} \alpha_t f_t(x).$$

[10] renders AdaBoost as a Newton update minimizing an approximate risk $J(F) \equiv E_{X,Y}\{e^{-YF(X)}\} \geq E_{X,Y}\{\mathbf{1}_{[YF(X)<0]}\}$ and obtain the update formula

$$F(x) \leftarrow F(x) + \frac{1}{2}\ln\left(\frac{1-\text{err}}{\text{err}}\right) f(x)$$
$$w(x,y) \leftarrow w(x,y)\exp\left[\ln\left(\frac{1-\text{err}}{\text{err}}\right)\mathbf{1}_{[y\neq f(x)]})\right]$$

where $f(x) = \text{sign}[E_w\{Y|x\}]$ and $\text{err} = E_w\{\mathbf{1}_{[Yf(x)<0]}|x\}$.
Alternatively,

$$F_{t+1}(x) = F_t(x) + \alpha_t f(x) \tag{1}$$

where

$$\alpha_t = \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right), w_t(x,y) = \exp[-yF_t(x)],$$

and

$$\epsilon_t = E_{w_t}\{\mathbf{1}_{[Yf(x)<0]}\} = \frac{E_{Y|x}\{w_t(x,Y)\mathbf{1}_{[Yf(x)<0]}\}}{E_{Y|x}\{w_t(x,Y)\}}.$$

[10] provides a convenient framework for further investigation. For example, variants of boosting algorithms can be constructed by replacing the exponential criterion by other approximate loss functions. However, the convergence issue of the (population version) of Newton update with respect to the exponential criterion is not addressed. This question of convergence motivates the current study.

3. Results

We now introduce the Bayesian optimization formulation for analyzing boosting algorithm. For easy exposition, we will assume X is a continuous (univariate) random variable with sampling probability density function $f_X(x|\theta)$ with the parameter θ and the prior distribution of θ is $\pi(\theta)$. Let the posterior distribution of θ given x is $\pi(\theta|x)$ and $y = g(\theta)$ for any function g with range $\mathcal{Y}$. Then the objective is to find a classifier F minimizing

$$J(F) = E_{\pi(\theta|x)}\left\{e^{-g(\theta)F(x)}\right\}. \tag{2}$$

This problem can be considered as a Bayesian estimation problem or more precisely a Bayesian minimization problem where the uncertainty of θ affecting the objective function through g. As a starting point, we impose the well-studied normal-normal distributional assumption. Precisely, assume $X \sim N(\theta, \sigma^2)$, where θ and σ^2 are the normal parameters. Let $\pi(\theta)$ be the conjugate prior and $\pi(\theta) \sim N(\mu, \tau^2)$, where μ and τ^2 are known. Standard calculation shows the posterior of θ given x is $\pi(\theta|x) \sim N(\mu_x, \rho^{-1})$, where

$$\mu_x = \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \text{ and } \rho = \frac{1}{\tau^2} + \frac{1}{\sigma^2} = \frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}.$$

And the marginal density of X is

$$m(x) = \frac{1}{\sqrt{2\pi}\rho\sigma\tau} \exp\left\{-\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)}\right\}.$$

Consider the estimation problem of

$$g(\theta) = \text{sign}(\theta) = 2(\mathbf{1}_{[\theta>0]}) - 1 \in \mathcal{Y}. \tag{3}$$

where $\mathbf{1}_A$ denotes the indicator function of set A. This choice of g aims to represent the binary response variable in the binary classification problem.

Now we follow the steps similar to [10]. Firstly,

$$\begin{aligned} J(F+f) &= E_{\pi(\theta|x)}\left\{e^{-g(\theta)[F(x)+f(x)]}\right\} \\ \approx \tilde{J}(F+f) &= E_{\pi(\theta|x)}\left\{e^{-g(\theta)F(x)}[1 - g(\theta)f(x) + g^2(\theta)f^2(x)/2]\right\} \\ &= E_{\pi(\theta|x)}\left\{e^{-g(\theta)F(x)}[1 - g(\theta)f(x) + f^2(x)/2]\right\}. \end{aligned}$$

The minimizer f can then be found by differentiation

$$\begin{aligned} f(x) &= \frac{E_{\pi(\theta|x)}\left\{g(\theta)e^{-g(\theta)F(x)}\right\}}{E_{\pi(\theta|x)}\left\{e^{-g(\theta)F(x)}\right\}} \\ &= \frac{e^{-F(x)}\Phi(\sqrt{\rho}\mu_x) - e^{F(x)}[1 - \Phi(\sqrt{\rho}\mu_x)]}{e^{-F(x)}\Phi(\sqrt{\rho}\mu_x) + e^{F(x)}[1 - \Phi(\sqrt{\rho}\mu_x)]}. \end{aligned} \tag{4}$$

Here $f(x)$ is the greedy update step for Bayesian optimization of $g(\theta)$ with respect to the approximate conditional risk $\tilde{J}$. Hence

$$\begin{aligned} F(x) &\leftarrow F(x) + f(x) \\ &= F(x) + \frac{\Phi(\sqrt{\rho}\mu_x) - e^{2F(x)}\left[1 - \Phi(\sqrt{\rho}\mu_x)\right]}{\Phi(\sqrt{\rho}\mu_x) + e^{2F(x)}\left[1 - \Phi(\sqrt{\rho}\mu_x)\right]}. \end{aligned}$$

The procedure can be expressed as an iterative procedure, for $t = 1, 2, \cdots$

$$\begin{aligned} F_{PIB,t+1}(x) &= F_{PIB,t}(x) + f_t(x) \\ &= F_{PIB,t}(x) + \frac{\Phi(\sqrt{\rho}\mu_x) - e^{2F_{PIB,t}(x)}\left[1 - \Phi(\sqrt{\rho}\mu_x)\right]}{\Phi(\sqrt{\rho}\mu_x) + e^{2F_{PIB,t}(x)}\left[1 - \Phi(\sqrt{\rho}\mu_x)\right]}. \end{aligned} \tag{5}$$

The subscript `PIB` denotes this algorithm is an (population) iterative Bayesian procedure. Immediately questions arise

- Does $F_{PIB,t}(x)$ in (5) converge as t goes to infinity?
- If $F_{PIB,t}(x)$ does converge, does it converge to the optimal Bayes procedure with respect to $0-1$ loss?

To answer the questions, we need the following lemmas. Their proofs are standard or straightforward thus omitted.

Fixed Point Theorem. If φ is a contraction of $\Re \to \Re$, that is, there exists $\alpha \in (0,1)$ such that $|\varphi(x) - \varphi(y)| < \alpha|x - y|$ for all $x, y \in \Re$, then there exists one and only one $x \in \Re$ such that $\varphi(x) = x$.

Cauchy-Schwarz Inequality. For any real $a_i, b_i, i = 1, 2, \cdots, n$, we have

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2.$$

Lemma 1. *For all* $x \neq 0$, $\frac{2(e^x-1)}{x(e^x+1)} < 1$.

Theorem 1. *For any initial* $F_{PIB,1}(x)$, *as* t *goes to infinity*

$$F_{PIB,t}(x) \to F_{\pi}(x) = \frac{1}{2}\ln\left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1 - \Phi(\sqrt{\rho}\mu_x)}\right).$$

Proof. Substitute $F_{PIB,t}(x) = u_t$ and define

$$\varphi(u) = u + \frac{a - be^{2u}}{a + be^{2u}}.$$

Now (5) can be expressed in iterative form

$$u_{t+1} = \varphi(u_t)$$

where $a = \Phi(\sqrt{\rho}\mu_x), b = 1 - \Phi(\sqrt{\rho}\mu_x) \in (0,1)$. We will first show that φ is a contraction and then find its fixed point. Since

$$|\varphi(u) - \varphi(v)| = \left| u + \frac{a - be^{2u}}{a + be^{2u}} - v - \frac{a - be^{2v}}{a + be^{2v}} \right|$$
$$= \left| 1 - \frac{2ab(e^{2u} - e^{2v})}{(u - v)(a + be^{2u})(a + be^{2v})} \right| |u - v|.$$

By Cauchy-Schwarz Inequality

$$(a + be^{2u})(a + be^{2v}) = \left[(\sqrt{b}e^{u})^2 + \sqrt{a}^2 \right] \left[\sqrt{a}^2 + (\sqrt{b}e^{v})^2 \right]$$
$$\geq \left(\sqrt{ab}e^{u} + \sqrt{ab}e^{v} \right)^2 = ab(e^{u} + e^{v})^2.$$

Thus for any $u \neq v$

$$0 < \frac{2ab(e^{2u} - e^{2v})}{(u - v)(a + be^{2u})(a + be^{2v})} \leq \frac{2ab(e^{u} + e^{v})(e^{u} - e^{v})}{(u - v)ab(e^{u} + e^{v})^2}$$
$$= \frac{2(e^{u} - e^{v})}{(u - v)(e^{u} + e^{v})} = \frac{2(e^{u-v} - 1)}{(u - v)(e^{u-v} + 1)} = \frac{2(e^{x} - 1)}{x(e^{x} + 1)} < 1. \qquad (x = u - v)$$

Hence by fixed point theorem, the function φ is a contraction of $\Re \to \Re$ and there exists one and only one fixed point. Straightforward calculation shows that the fixed point is

$$F_{\pi}(x) = \frac{1}{2} \ln \left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1 - \Phi(\sqrt{\rho}\mu_x)} \right). \qquad \blacksquare$$

Note that $F_{\pi}(x)$ is a function of the ratio of posterior probabilities of $g(\theta)$ and $1 - g(\theta)$ thus it achieves the optimal Bayes risk. The proof is straightforward and refers to, for example, [5]. In other words, so far as the sign is concerned, $F_{\pi}(x)$ is essentially the optimal Bayes procedure. Theorem 1 shows that the greedy descent (5) converges to the optimal Bayes procedure.

Although our conditional approximate risk minimization approach is similar to [10], we impose a strong distributional assumption which allows detailed analysis and addresses the convergent questions. A question immediately arises: whether the population version AdaBoost as in [10] converges to the optimal Bayes procedure if it does converge at all? Recall the greedy descent in [10] in our notations is

$$F_{FHT}(x) \leftarrow F_{FHT}(x) + \frac{1}{2} \ln \left(\frac{1 - \text{err}}{\text{err}} \right) s(x) \tag{6}$$

where $s(x)$ is the sign of the greediest update step f in (4) and

$$\begin{aligned}\text{err} &= \frac{E_{\pi(\theta|x)}\{\mathbf{1}_{[s(x)\neq g(\theta)]}e^{-g(\theta)F_{FHT}(x)}\}}{E_{\pi(\theta|x)}\{e^{-g(\theta)F_{FHT}(x)}\}} \\ &= \begin{cases} \dfrac{e^{F_{FHT}(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}{e^{-F_{FHT}(x)}\Phi(\sqrt{\rho}\mu_x)+e^{F_{FHT}(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}, & \text{if } s(x)=+1, \\ \dfrac{e^{-F_{FHT}(x)}\Phi(\sqrt{\rho}\mu_x)}{e^{-F_{FHT}(x)}\Phi(\sqrt{\rho}\mu_x)+e^{F_{FHT}(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}, & \text{if } s(x)=-1. \end{cases}\end{aligned}$$

So

$$\frac{1-\text{err}}{\text{err}} = \left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1-\Phi(\sqrt{\rho}\mu_x)}e^{-2F_{FHT}(x)}\right)^{s(x)}.$$

And

$$\frac{1}{2}\ln\left(\frac{1-\text{err}}{\text{err}}\right) = \frac{s(x)}{2}\left[\ln\left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1-\Phi(\sqrt{\rho}\mu_x)}\right) - 2F_{FHT}(x)\right].$$

Thus (6) becomes

$$\begin{aligned}F_{FHT}(x) &\leftarrow F_{FHT}(x) + \frac{s^2(x)}{2}\left[\ln\left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1-\Phi(\sqrt{\rho}\mu_x)}\right) - 2F_{FHT}(x)\right] \\ &= \frac{1}{2}\ln\left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1-\Phi(\sqrt{\rho}\mu_x)}\right).\end{aligned}$$

That is, AdaBoost in the sense of [10] is a one-step convergent algorithm and converges to the optimal Bayes procedure. In other words, AdaBoost is very effective and performs even better than the greedy descent (5) under our distributional settings. One caveat: this comparison is a population-version comparison and does not necessarily transcribe to the sample-version superiority of AdaBoost.

The normal-normal setting allows us to study the convergence and Bayes risk of population version of boosting. Along this line, we generalize the results to high dimensional cases. Precisely, let X be a multivariate normal random vector denoted as $X = (X_1, \cdots, X_p)^t \sim MN(\theta, \Sigma)$, where $\theta = (\theta_1, \cdots, \theta_p)^t$ and $\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2)$ are multivariate normal parameters. Let $\pi(\theta)$ be the conjugate prior and $\pi(\theta) \sim MN(\mu, T)$, where $\mu = (\mu_1, \cdots, \mu_p)^t$ and $T = \text{diag}(\tau_1^2, \cdots, \tau_p^2)$ are known. Then the marginal density of X is

$$m(x) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi_j\rho_j}\sigma_j\tau_j} \exp\left\{-\frac{(x_j-\mu_j)^2}{2(\sigma_j^2+\tau_j^2)}\right\},$$

and the posterior of θ given x is

$$\pi(\theta|x) = \prod_{j=1}^{p} \frac{\sqrt{\rho_j}}{\sqrt{2\pi}} \exp\left\{-\frac{\rho_j}{2}\left[\theta_j - \frac{1}{\rho_j}\left(\frac{\mu_j}{\tau_j^2}+\frac{x_j}{\sigma_j^2}\right)\right]^2\right\}.$$

Note that $\pi(\theta|x) \sim MN(\mu_x, R)$, where

$$\mu_x = (\mu_{x,1}, \cdots, \mu_{x,p})^t,\ \mu_{x,j} = \frac{1}{\rho_j}\left(\frac{\mu_j}{\tau_j^2} + \frac{x_j}{\sigma_j^2}\right) = \frac{\sigma_j^2\mu_j + \tau_j^2 x_j}{\sigma_j^2 + \tau_j^2},$$

and

$$R = \text{diag}(\rho_1^{-1}, \cdots, \rho_p^{-1}),\ \rho_j = \frac{1}{\tau_j^2} + \frac{1}{\sigma_j^2} = \frac{\sigma_j^2 + \tau_j^2}{\sigma_j^2\tau_j^2}.$$

While we generalize the setting from univariate explanatory variable to high dimensional explanatory variables, our problem remains the binary classification problem. Define

$$g(\theta) = \text{sign}(\prod_{j=1}^{p}\theta_j) = 2(\mathbf{1}_{[\prod_{j=1}^{p}\theta_j > 0]}) - 1 \in \{\pm 1\}. \tag{7}$$

Note that (7) is a natural extension to (3). Admittedly, we do not claim this choice necessarily reflects the classification problem arised in practice but rather a choice of analytical convenience. Nonetheless, as a starting point, it provides a clearer picture of how the components of the boosting-like algorithm knitted together.

Following similar derivations as in one-dimensional case, it can be shown that the population iterative Bayesian procedure can be expressed as an iterative procedure, for $t = 1, 2, \cdots$

$$F_{PIB,t+1}(x) = F_{PIB,t}(x) + \frac{(1+\Phi_S) - e^{2F_{PIB,t}(x)}(1-\Phi_S)}{(1+\Phi_S) + e^{2F_{PIB,t}(x)}(1-\Phi_S)}.$$

Furthermore

Theorem 2. For any initial $F_{PIB,1}(x)$, as t goes to infinity

$$F_{PIB,t}(x) \to F_\pi(x) = \frac{1}{2}\ln\left(\frac{1+\Phi_S}{1-\Phi_S}\right).$$

Details are referred to [8]. This theorem also shows that the Newton-like iteration converges to the optimal Bayes procedure. Straightforward calculation also shows that, under the same setting, the generalized population version AdaBoost as in [10] converges to the optimal Bayes procedure as well.

4. Conclusion and Discussion

[10] renders population version AdaBoost as a Newton update minimizing an approximate exponential criterion. Under a normal-normal setting, the classification problem is recasted as a Bayesian minimization problem. We derive an iterative algorithm and contrast with the population AdaBoost derived in [10]. It is shown

that the optimal Bayes procedure can be obtained via an iterative Newton update minimizing exponential criterion. In addition, the step sizes of AdaBoost are shown to be highly effective and lead to a one-step convergence. Contrast to many population theoretical results in the literature, we do not impose assumption on the base learners nor the functional relation between Y and X–but the distributional assumption on the response and the explanatory variables. With suitable choices of the hyperparameter, the distribution can represent rather noisy data. Even under these circumstances, the consistent results remain and neither early stopping nor regularization in step sizes are required. This is very different from other population theoretical results for boosting-like algorithms. Again, we warn the readers that our results are of population version thus may not directly imply similar results in finite-sample implementations.

The distributional assumption mainly serves as an alternative theoretical assumption for analysis of the convergence. This is different from many existing literature where the assumptions are imposed on the base learners or the regularization. On this regard, our results suggest a possible "statistical view" that can reconcile with [16]. The readers are referred to [7] for more discussion. At this stage, we do not claim this assumption is immediately applicable in practice. Nonetheless, this model can be more appealing with suitable transformations on the explanatory variables. Furthermore, when the sampling distribution and prior distribution can be suitably modeled, plug-in Bayes procedure with good estimated parameter/hyperparameter might be an alternative acceptable classifier.

REFERENCES

1. P. L. Bartlett, M. I. Jordan and J. D. McAuliffe, Convexity, classification and risk bounds, *Journal of American Statistical Association*, **101** (2006), 138-156.
2. P. L. Bartlett and M. Traskin, AdaBoost is consistent, *Journal of Machine Learning and Research*, **8** (2007), 2347-2368.
3. P. J. Bickel, Y. Ritov and A. Zakai, Some theory for generalized boosting algorithms, *Journal of Machine Learning Research*, **7** (2006), 705-732.
4. L. Breiman, Prediction games and arcing algorithms, *Neural Computation*, **11(7)** (1999), 1493-1517.
5. L. Breiman, Population theory for boosting ensemble, *Annals of Statistics*, **32** (2004), 1-11.
6. P. Bühlmann and B. Yu, Boosting with the l_2-loss: Regression and classification, *Journal of American Statistical Association*, **98** (2003), 324-339.
7. P. Bühlmann and B. Yu, Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, JMLR 9: 131-156, 2008, *Journal of Machine Learning and Research*, **9** (2008), 187-194.

8. W. D. Chen and C. A. Tsao, *Consistency of boosting under normal distributional assumptions, Technical Report*, Department of Applied Math., National Dong Hwa University, 2008.

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

10. J. H. Friedman, T. Hastie and R. Tibishirani, Additive logistic regression: a statistical view of boosting, *Annals of Statistics*, **28** (2000), 337-407.

11. W. Jiang, On weak base hypotheses and their implications for boosting regression and classification, *Annals of Statistics*, **30** (2002), 51-73.

12. W. Jiang, Process consistency for AdaBoost, *Annals of Statistics*, **32** (2004), 30-55.

13. G. Lugosi and N. Vayatis, On the Bayes-risk consistency of regularized boosting methods, *Annals of Statistics*, **32** (2004), 30-55.

14. S. Mannor, R. Meir and T. Zhang, Greedy algorithms for classification-consistency, converge rates, and adaptivity, *Journal of Machine Learning Research*, **4** (2003), 713-742.

15. L. Mason, J. Baxter, P. L. Bartlett and M. Frean, *Boosting algorithms as gradient descent*, Advances in Neural Information Processing Systems, 12, MIT, 1999, pp. 512-518.

16. D. Mease and A. Wyner, Evidence contrary to the statistical view of boosting, *Journal of Machine Learning Research*, **8** (2007), 409-439.

17. R. Meir and G. Rätsch, *An introduction to boosting and leveraging*, Advanced Lectures on Machine Learning, LNCS, Springer-Verlag, 2003, pp. 119-184.

18. R. E. Schapire, Y. Freund, P. Bartlett and W. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, *Annals of Statistics*, **26(5)** (1998), 1651-1686.

19. T. Zhang, Statistical behaviour and consistency of classification methods based on convex risk minimization, *Annals of Statistics*, **32** (2004), 56-134.

20. T. Zhang and B. Yu, Boosting with early stopping: Convergence and consistency, *Annals of Statistics*, **33** (2005), 1538-1579.

C. Andy Tsao
Department of Applied Mathematics,
National Dong Hwa University,
Hualien 97401,
Taiwan
E-mail: chtsao@mail.ndhu.edu.tw

W. Drago Chen
Center of General Education,
Lan Yang Institute of Technology,
Yilan 26141,
Taiwan
E-mail: drangochen@gmail.com