

REJOINDER

WENXIN JIANG

Northwestern University

May 13, 2003

1. Comments and discussions. I thank the discussants for their insightful comments and new contributions, and thank Jon A. Wellner for arranging this discussion. I also congratulate the authors of the other papers under discussion and thank them for their significant works that are both independent and also collaborative in a general sense. My paper is a small step built on important previous works of other people: Freund and Schapire (1997) invent the popular AdaBoost algorithm. Brieman (1997) and Schapire and Singer (1998) identify the exponential criterion in relation to AdaBoost. Friedman, Hastie and Tibshirani (2000) find the minimizer of the exponential criterion in the population case. Breiman (2000) solves the difficult convergence problem for AdaBoost in the population case, showing that the iterations indeed approach the right minimum.

When I start to work on this topic a few years ago, AdaBoost seems to me a mysterious and interesting puzzle. Although it performs well and is often resistant against overfitting, I soon find that in all analytic examples that I can work out, AdaBoost can overfit when run for a sufficiently long time, e.g., of order $t \sim n^2 \log n$, where n is the sample size, see Jiang (2001, 2002). [Of course, overfitting has been noticed in experiments or anticipated conceptually by other people as well, e.g., Grove and Schuurmans (1998), Mason et al. (1999), Friedman et al. (2000), Bickel (2001, private communication).] So when Breiman brought to my attention his paper proving that the population version ($n = \infty$, roughly speaking) of AdaBoost does *not* overfit and is consistent at $t \rightarrow \infty$, there seemed to be an apparent contradiction that I had to resolve.

To find a compromise, I regard Breiman's situation as $t \prec n$ (since n is already infinite), while the situations considered in Jiang (2002) are $t \succ n$. The situations are different. It then occurs to me that Breiman's result suggests that a consistent AdaBoost solution may be obtained in the finite sample situation as well, if t is chosen to increase with n at a rate that is *not too fast*, so as to prevent the overfitting situation that I considered before.

Starting with this conjecture, I wrote up this note. It was originally intended to be a short communication, since I was not satisfied with the restrictive framework, conditions and results. However, to my relief, I noticed follow-up works that has made significant improvements in several directions. So now I think this short communication is at least very successful in this regard, i.e., to induce other works that are better and more comprehensive.

Several interesting new results are described in the discussions. Bartlett, Jordan and McAuliffe obtain general comparison theorem with sufficient and necessary conditions relating the consistency in prediction and the consistency in minimizing the 'working' cost function. Bickel and Ritov, in a very efficient way, outline a consistency proof for boosting with truncation with general cost function, and justify the use of cross-validation for implementing the truncation. Bühlman and Yu point out the computational advantage of the truncation method and introduce some of their promising works in this direction that are independent of Bickel and Ritov: Bühlman's work on consistency of L_2 Boost, and Zhang and Yu's work on convergence and consistency of boosting with general cost function.

Friedman, Hastie, Rosset, Tibshirani and Zhu investigate the close relationship between boosting and L_1 penalty and show how this might benefit in the case of 'sparsity'. This connection was also

discussed by Bühlman and Yu and both discussions refer to the interesting work by Efron et al. on boosting with infinitesimal steps. Koltchinskii provides an alternative proof for the consistency result in Lugosi and Vayatis' paper under discussion. Freund and Schapire raise several interesting points that are common to all papers under discussion. I will only focus on a few of their remarks. I am sure the other authors can provide better replies and I will rely on them to respond to the other points.

Freund and Schapire rightly pointed out that the consistency results do not seem to explain the good finite sample performances of the boosting algorithms when handling high-dimensional data. They recommend explanations from margin theory. On the other hand, most other discussants implicitly or explicitly (e.g., Bartlett, Jordan and McAuliffe; Koltchinskii) suggest future works in studying convergence rates. I also think the study of convergence rates is more promising. The currently available margin bounds do not compare to the Bayes error, and typically cannot be tight in the case of noisy data.

Freund and Schapire also raised the question on regularization of AdaBoost: is it unnecessary or potentially benefiting. I tend to agree on the second choice. Especially in noisy cases, regularized variants have been reported to lead to improvements. See, e.g., Mason et al. (1999); Bühlman and Yu (2000); Friedman (2001); Krieger, Long and Wyner (2001).

2. Future directions. Future efforts are most effective if both experimental people and theorists (they can be the same persons of course) collaborate very closely. Analytic studies alone can reveal some insights but are often limited to idealized cases. Experimental studies sometimes fail to generate information that might be important in reaching a good understanding.

For (an old) example, as far as I know, there is still no complete understanding of the most mysterious behavior of AdaBoost: *In some situations the training error becomes zero, but the prediction error still continues to decrease.* Partial explanation was made in Schapire et al. (1998) based on semi-empirical upperbounds. Theoretical studies obtained exact solutions only in the one dimensional case, where it is proved that AdaBoost with trees generates zero training error in finite time and converges to a nearest neighbor rule [see Jiang (2001)]. On the other hand, in higher dimensions, a rule that fits training sample perfectly can be very different from the nearest neighbor rule. *After a perfect fit on training sample, does the prediction error approach something that is about the same as the nearest neighbor error, or not as good, or magically better?* Apparently, only in the last possibility will this mystery be worth studying, for otherwise one could use a nearest neighbor rule to do as good a job or better. However, in the experimental results that reported such mystery, in no case was the noise level or nearest neighbor error reported. The reporting of the nearest neighbor error in such cases with zero training error could help us understand when will such mystery occur (in noisy or noiseless cases), and whether this mystery is worth studying (whether it will beat the natural benchmark, the nearest neighbor error). A coordinated effort in both experiments and theory seems needed.

As many discussants point out, another promising direction is the study of convergence rates for variants of boosting algorithms, possibly regularized, in various combinations of base learners and situations of data (e.g., noisy or not, sparseness or denseness, as Friedman, Hastie, Rosset, Tibshirani and Zhu commented). There are already some preliminary steps made in this direction, e.g., Bühlmann and Yu (2001); Jiang (2001); Mannor, Meir, and Zhang (2002). *Studying the convergence rates in various interesting situations for various methods could further our knowledge on when boosting will work well, and when it can be improved and how.* In the next section I will try to explain

these points by a simple example.

3. A simple example. The following example involves a regularization method that averages over AdaBoost predictions from several subsamples. It is motivated from slightly different bag-boosting schemes described by Bühlman and Yu (2000) and Krieger, Long and Wyner (2001), and is closer to the latter reference. I became interested in this regularization scheme due to the good performances reported in the experiments of Krieger et al., and due to the modularity of its implementation. Again, I can only obtain analytic results in some idealized one-dimensional case described below.

Consider a setup similar to Section 5.2 of Jiang (2001), $X \sim Unif[0, 1]$. The base hypothesis space is the space of ‘stumps’, or a more general space which contains piecewise constant hypotheses, having splits chosen from mid-data points. The regularization scheme involves averaging subsample predictions as follows.

ALGORITHM. (*Averaged AdaBoost from subsamples*).

- (i) Divide the training sample $(X_i, Z_i)_1^n$ randomly into K subsamples of size m . (For convenience we assume $n = Km$.)
- (ii) For subsample $k = 1, \dots, K$, run AdaBoost t -steps and define the resulting prediction rule (at any $x \in [0, 1]$) as $\hat{z}_k^{(t)}(x)$.
- (iii) Compute the average of these predictions $\bar{z}_K^{(t)} = K^{-1} \sum_{k=1}^K \hat{z}_k^{(t)}$ and use $\hat{Z}_K^t(x) \equiv \text{sgn}(\bar{z}_K^{(t)}(x))$ to predict the value of the unknown label Z for a future observation with $X = x$.

We will study cases with large t 's so that AdaBoost already overfits the individual subsamples, and investigate how averaging over K subsample results remedies the overfit. Here K provides additional freedom for regularizing AdaBoost and measures the level of regularization (non-regularized AdaBoost has $K = 1$). We will consider what convergence rate of the resulting prediction can be achieved, in the following three situations. Denote the (conditional) *probability function* $\pi(x) = P(Z = 1 | X = x)$.

- (A). (*Noiseless with finite number of jumps*).
 $\pi(x) \in \{0, 1\}$ for all x and is piecewise constant with at most J jumps. (J is a positive integer.)
- (B). (*Lipschitz*).
 $|\pi(x) - \pi(x')| < D\epsilon$ whenever $|x - x'| < \epsilon$, for all small ϵ .
- (C). (*Finite number of finite ‘sign-changes’*).
 $|\pi(x) - 0.5| \geq \delta$ for all x , for some $\delta \in (0, 0.5]$. Also, $\text{sgn}\{\pi(x^-) - 0.5\} \neq \text{sgn}\{\pi(x^+) - 0.5\}$ for at most J locations of x . (J is a positive integer.)

PROPOSITION. (*Averaged AdaBoost & convergence rates*).

Denote $L = P[\hat{Z}_K^t(X) \neq Z]$ as the prediction error, $L^* = E \min\{\pi(X), 1 - \pi(X)\}$ as the Bayes error. The following results hold for all $t \geq 2m^2 \log(m + 1)$, where $m = n/K$ is the size of the subsamples, t is the number of boosting steps.

- (a1) [Remark 3(c), Jiang (2000)]. For the noiseless class (A), taking $K = 1$ (no subsampling) leads to

$$L - L^* \leq 2Jn^{-1} \log n \{1 + o_n(1)\}.$$

(a2) [Theorem 3, Jiang (2001)]. In the general noisy situations, however, taking $K = 1$ for our current case of large t can lead to inconsistency:

$$L - L^* = E[2|\pi(X) - 0.5|\min\{\pi(X), 1 - \pi(X)\}] + o_n(1).$$

(b1) For the Lipschitz case (B), denote the ‘noise level’ $E\text{var}(Y|X) = \sigma^2$ and assume $\sigma > 0$ [here $Y = (Z + 1)/2$ is the binary response]. Let the ‘signal to noise ratio’ $\text{SNR} = D/\sigma$ where D is the Lipschitz constant. Then, taking $K \approx (n/\text{SNR})^{2/3}$ leads to

$$L - L^* \leq 2\sigma(n/\text{SNR})^{-1/3} \log(n/\text{SNR})^{1/3} \{1 + o_n(1)\}.$$

(b2) For smooth cases in (B) with continuous derivative $\pi'(\cdot)$, result (b1) can be strengthened by taking $\text{SNR} = E|\pi'(X)|/\sigma$ (assume $E|\pi'(X)| > 0$) in the formulas of K and $L - L^*$.

(c) For the possibly noisy case (C) with finite number of finite ‘sign-changes’, if we take $K \approx (2\delta^2)^{-1} \log n$, we have

$$L - L^* \leq 0.5J\delta^{-4}n^{-1}(\log n)^2\{1 + o_n(1)\}.$$

REMARKS.

1. These results show that *different data situations entail different regularization strategies and allow different convergence rates*. An important indicator for characterizing various situations is the noise level, which can be defined in several ways: as the average conditional variance (which is essentially the nearest neighbor error), or average conditional standard deviation, or the Bayes error. When $\pi \in \{0, 1\}$, all such measures should indicate zero noise.
2. In the noiseless case (A), boosting without regularization ($K = 1$) is already near optimal, i.e., achieves a rate $n^{-1} \log n$ that is within $\log n$ from the minimax rate n^{-1} for noiseless learning [see, e.g., Devroye et al. (1996) Theorem 14.1, or Jiang (2000) Proposition 4].
3. In the noisy situations, non-regularized boosting is generally inconsistent. However, for noisy cases in (B), regularization with averaged subsample predictions can achieve a near optimal rate $n^{-1/3} \log n$, within $\log n$ from the minimax rate $n^{-1/3}$ for a Lipschitz family [see, e.g., Yang (1999)].
4. In the noisy case described in (C), a better rate $n^{-1}(\log n)^2$ can be obtained, which is within $(\log n)^2$ to the best possible [the minimax rate $1/n$ for the noiseless case (A) holds here too since $(A) \subset (C)$]. This has no actual contradiction to the minimax result $n^{-1/3}$ above for case (B), since (B) can allow ‘difficult’ functions such as $\pi = 0.5 + 0.5x^2 \sin(x^{-1})$, which can cross 0.5 infinitely often (lack of ‘sparsity’), and can become arbitrarily close to 0.5, while such probability functions are excluded from (C).
5. Note that the results in this proposition suggests different levels of regularization for different types of problems. In noisy situations, overfitting is prevented by averaging over multiple subsamples. However, case (C) suggests the use of much less subsamples than case (B). But if the data are noiseless after all, it is possible that regularization might actually hurt the performance.

6. Even when restricted to consider smooth probability functions in case (B), results (b1,2) still suggest that one should use more subsamples when there is higher noise σ , and less when σ is low. Prior knowledge on the noise level, or knowledge on σ from a 2-stage procedure (using the maximum possible σ -value 0.5 in the first stage) might help.
7. In case (B), the regularization level used in results (b1,2) actually produce reliable estimation of the probability function $\pi(x)$. The average of AdaBoost predictions $\bar{z}_K^{(t)}(x)$, before the sign transformation, estimates the mean function $E(Z|X = x) = 2\pi(x) - 1$ at a near optimal rate.
8. The implementation of this regularization scheme is simple and modular, since it only involves manipulation of outcomes from the standard AdaBoost algorithm. Subsamples can also be processed in parallel. Although all the results are derived for one dimensional X , we suspect that the performance of this algorithm, with suitable choice of the number of subsamples and the number of boosting steps, will also be good in higher dimensions. A similar ‘bag-boosting’ algorithm, where the subsamples can overlap, has shown good numerical performance in higher dimensions in both noisy and noiseless situations [Krieger et al. (2001)].
9. The key to the proof of these results is to notice that for sufficiently large t , the Adaboost prediction from each subsample becomes the nearest neighbor rule. Then averaging these nearest neighbor rules, based on independent subsamples, can be easily shown to prevent overfit, even when the subsamples have already been overfitted individually. In practice, using smaller t (and K) may also generate good performance, see Krieger et al. (2001).

For a particular dataset, using a ‘dataset-based’ procedure (such as cross-validation) to determine the level of regularization can generate better results than just relying on what is suggested by convergence rate analyses. However, the procedure can then become either more computationally involved or less efficiently utilizing the whole data. A purely ‘dataset-based’ approach also provides less understanding to the general behaviors of boosting algorithms, compared to a ‘situation-based’ convergence rate analysis. It may also be possible to combine the two approaches in some sense, by using some kind of a 2-stage analysis (see Remark 6 above), or by incorporating the knowledge from a convergence rate analysis to reduce the range of searching for a best regularization parameter in cross-validation.

Clearly, further studies on regularization schemes are needed, with attention paid to all three sides: the *theoretical side* (on consistency and convergence rates in various interesting situations), the *numerical side* (on finite sample performance), as well as the *practical side* (on the ease of implementation and computation).

ADDITIONAL REFERENCES

- BREIMAN, L. (1997). Prediction games and arcing algorithms. *Technical Report 504, Statistics Department, University of California at Berkeley.*
- BÜHLMANN, P. AND YU, B. (2000). Discussion. Additive logistic regression: a statistical view of boosting, by Friedman, J., Hastie, T. and Tibshirani, R., *Annals of Statistics*, **28** 377-386.
- BÜHLMANN, P. AND YU, B. (2001). Boosting with the L2 Loss: Regression and Classification. *Technical Report 605, Statistics Department, University of California at Berkeley.* (To appear in *J. Amer. Statist. Assoc.*).
- FRIEDMAN, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29** 1189-1232.

- JIANG, W. (2000). Does boosting overfit: views from an exact solution. *Technical Report 00-03, Department of Statistics, Northwestern University*.
(Downloadable at <http://neyman.stats.nwu.edu/jiang/boost/boost.onedim.ps>.)
- JIANG, W. (2001). Some theoretical aspects of boosting in the presence of noisy data. *Proceedings of the Eighteenth International Conference on Machine Learning, 2001*. Morgan Kaufmann 234-241. (Also listed as Technical Report 01-01, Department of Statistics, Northwestern University.)
(Downloadable at <http://neyman.stats.nwu.edu/jiang/boost/boost.icml.ps>.)
- KRIEGER, A., LONG, C. AND WYNER, A. (2001). Boosting noisy data. *Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann 274-281.
- MANNOR, S., MEIR, R. AND ZHANG, T. (2002). The consistency of greedy algorithms for classification. J. Kivinen and R. H. Sloan (Eds.): COLT 2002, *Lecture Notes in Artificial Intelligence* **2375** 319-333.
- SCHAPIRE, R. E. AND SINGER, Y. (1998). Improved boosting algorithms using confidence-rated predictions. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 1998* 80-91.
- YANG, Y. (1999). Minimax nonparametric classification—Part I: rates of convergence. *IEEE Trans. Info. Theory*, **45** 2271-2284.

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