Journal of Machine Learning Research 9 (2008) 165-170 Published 2/08

Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, *JMLR 9*:131–156, 2008

Andreas Buja BUJA.AT.WHARTON@GMAIL.COM
Statistics Department
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104-6340

Werner Stuetzle WXS@STAT.WASHINGTON.EDU
Statistics Department
University of Washington
Seattle, WA 98195-4322

Editor: Yoav Freund

We thank the authors for writing a thought-provoking piece that may ruffle the feathers of recent orthodoxies in boosting. We also thank JMLR for publishing this article! Since the late 1990s, boosting has undergone the equivalent of a simultaneous X-ray, fMRI and PET exam, and the common view these days is that boosting is a kind of model fitting. As such, it is subjected to assumptions that are common in non-parametric statistics, such as: limiting the complexity of the base learner, building up complexity gradually by optimization, and preventing overfitting by early stopping or by regularizing the criterion with a complexity penalty. The theories backing this up use VC dimensions and other measures to show that, if the complexity of fits grows sufficiently slowly, asymptotic guarantees can be given. Into this orthodox scene Mease and Wyner throw one of the most original mind bogglers we have seen in a long time: “if stumps are causing overfitting, be willing to try larger trees.” In other words, if boosting a low-complexity base learner leads to overfit, try a higher-complexity base learner; boosting it might just not overfit. Empirical evidence backs up the claim.

Is this counterintuitive wisdom so surprising? Yes, if seen from the point of view of orthodoxy, but less so when reviving some older memories. We may remind ourselves how boosting’s fame arose in statistics when the late Leo Breiman stated in a discussed 1998 Annals of Statistics article (based on a 1996 report) that boosting algorithms are “the most accurate ... off-the-shelf classifiers on a wide variety of data sets.” We should further remind ourselves what this praise was based on: boosting of the full CART algorithm by Breiman himself, and boosting of the full C4.5 algorithm by others. In other words, the base learners were anything but ‘weak’ in the sense of today’s orthodoxy, where ’weak’ means ‘low complexity, low variance, and generally high bias.’ (Few people today use PAC theory’s untenable notion of weak learner, which was gently demolished by Breiman in the appendix of this same article.) Breiman’s (1998b, p. 802) 02) major conclusion at the time was: “The main effect of both bagging and [boosting] is to reduce variance.” It appears, therefore, that his notion of ‘weak learner’ was one of ‘high complexity, high variance, and low bias’! This was before the low-variance orthodoxy set in and erased the memories of the early boosting experiences.

Unfortunately, soon thereafter Breiman saw his own assumptions thrown into question when he learned from Schapire et al.’s (1998) work that excellent results could also be achieved by boosting stumps. This experience was later reinforced when Friedman et al. (2000) introduced the interpretation of boosting as model fitting: the base learner now had to be weak in the sense of low variance. Ever since, theoretical attempts at 'explaining boosting' have relied on low complexity of the base learner and controlling complexity of the final classifier to assure good generalization properties. These 'explanations,' however, have never been able to explain why boosting is relatively immune to overfitting, even when not stopped and not regularized and used with a high complexity base learner.

Mease and Wyner's achievement is to pull the messy truth out from under the rug of the low-variance orthodoxy. They do so with the equivalent of boy scout tools, some simple but telling simulations, which reinforce the idea that our reasonings about early stopping, regularization, low variance of the base learner, and the specifics of the surrogate loss function, are not or not the only essence of boosting. To explain why this is so, Mease and Wyner do not give us hard theory, but they point in a direction, essentially by recovering memories that predate the low-variance orthodoxy: "self-averaging" for variance reduction, which is the principle behind bagging and random forests.

While variance reduction is an aspect that has been ignored by the low-variance orthodoxy, the orthodoxy's implicit dogma, that boosting can reduce bias, is also true. As asserted and documented empirically a decade ago by Schapire et al. (1998, Section 5.3), boosting can do both. Depending on the data and the base learner, the effect that dominates may be bias reduction *or* variance reduction. In this regard Schapire et al.'s (1998) simulation results as summarized in their Table 1 (p. 1673) are illuminating, and had we taken them seriously sooner, we would be less surprised by Mease and Wyner's messages. Arguing against Breiman (1998b), Schapire et al. used the table to make the now orthodox point that boosting can reduce bias. An unprejudiced look shows, however, that the winner in all four scenarios is boosting C4.5, not boosting stumps, and when C4.5 is the base learner the overwhelming story is indeed variance reduction. With this information, the Mease-Wyner mind boggler is a touch less mind boggling indeed: From the combined evidence of Breiman (1998b) and Schapire et al. (1998), we should expect that boosting high-variance base learners generally outperforms boosting low-variance base learners. For the practitioner the recommendation should be to boost CART or C4.5. In theoretical terms, one should let most of the bias removal be done by the base learner and take advantage of boosting's variance removal; at the same time, boosting may further reduce the base learner's bias by another notch if that is possible.

Where does this leave us in terms of theory? The implications of Mease and Wyner's unorthodoxies stand: Complexity controlling theories of bias removal are off the mark; they are not incorrect but misleading, and they ignore a whole other dimension that matters hugely for the practice of boosting. What we need is a theory that explains bias and variance reduction in a single framework. We do not even know of a unified general theory of variance reduction, although some interesting work has been done in the area of bagging (Bühlmann and Yu, 2002) and random forests (Amit et al., 2001). The real jackpot, however, would be a theory that explains how and when boosting reduces bias *and* variance.

Meanwhile we are left with some tantalizing clues, above all Breiman's hunch (1999, p. 3): "AdaBoost has no random elements .... But just as a deterministic random number generator can give a good imitation of randomness, my belief is that in its later stages AdaBoost is emulating a random forest." If born out, this conjecture would have theoretical and practical implications. For one, it would mean that the initial stages of boosting may remove bias, whereas the later stages remove variance. According to Breiman (1998b, p. 803), boosting a high-variance base learner does not yield convergence but exhibits "back and forth rocking" of the weights, and "This variability may be an essential ingredient of successful [boosting] algorithms." Breiman implies that at some point boosting iterations turn into a pseudo-random process whose behavior may resemble more the purely random iterations of bagging than those of a minimization process. This random process may be able to achieve the self-averaging effect of variance reduction that is so prominent when boosting high-variance base learners. If this view is correct, one may have to rethink the role of the surrogate loss function that is minimized by boosting. Its main role is to produce structured weights in the iterations, but with noisy errors, these weights may for practical purposes be as much random as they are systematic. This insight jibes with Wyner's (2002) malicious experiments in which he doubled the step size of discrete AdaBoost with C4.5, thereby assuring that the exponential loss never decreased and in fact provably remained at a constant level; his empirical results indicated that on average this SOR (successively over-relaxed) form of boosting performs as well as regular AdaBoost. These results may be taken as evidence that the minimization aspect is of little importance for a high-variance base learner; of greater importance may be a pseudo-random aspect of the reweighting scheme that achieves variance reduction similar to bagging, just more successfully due to a sort of adaptivity in the reweighting that improves over the purely random resampling of bagging.

If the pseudo-random aspect of boosting is critical for high-variance base learners, one may draw consequences and implement boosting with proper pseudo-random processes. So did Breiman. He didn't attempt a theory of boosting for high-variance base learners, and instead he put his intuitions to use in further proposals such as in his work on "half & half bagging" (Breiman, 1998a), apparently with success. Another example that benefited from Breiman's inspiration was Friedman's (2002) "stochastic gradient boosting" which inhibits convergence of boosting by computing gradient steps from random subsamples drawn without replacement. Friedman (ibid., p. 9) observes improvements over deterministic boosting in a majority of situations, above all for small samples and "high capacity" (high variance) base learners. He admits that "the reason why this randomization produces improvement is not clear," but suggests "that variance reduction is an important ingredient." Friedman goes on to suggest that stochastic AdaBoost with sampling from the weights rather than reweighting may have similar variance-reducing effects. In early boosting approaches such sampling (with replacement) was performed to match the given sample size, but Friedman suggests that further variance reduction could be gained by choosing smaller resamples.

An implication of Breiman's hunch is that the real difference between LogitBoost and AdaBoost is not so much due to the differences in loss functions as to the minimization method, at least when the base learner has relatively high variance, or generally in the late stages of boosting. AdaBoost can be interpreted as constrained gradient descent on the exponential loss, whereas LogitBoost is Newton descent on the logistic loss (Friedman et al., 2000). The two minimization schemes produce very different reweighting schemes, and they work off different working responses during the iterations. We are currently ignorant about whether LogitBoost develops pseudo-random behavior late in the iterations, similar to AdaBoost. If it does, the cause may be traced to the base learner, and the phenomenon may be robust to the specifics not only of loss functions but of algorithms as well.

Another implication of Breiman's hunch is that boosting does both, reduce bias and variance, in the same problem, but each primarily at different stages of the boosting iterations. If it is true that variance reduction occurs during later iterations, then this should go a long way to explain boosting's relative immunity to overfitting. By comparison, conventional fitting mechanisms only know how to do one thing: follow the data ever more closely, thereby continually reduce bias and continually accumulate variance. According to orthodoxy, therefore, the art is to find the proper balance, and to this end auxiliary devices such as early stopping, regularization penalties and cross-validation come into play. Boosting seems to be different, but we do not have the theory yet to prove it.

All that we said so far is based on out-of-sample classification error. A peculiarity of classification error is that it is not the criterion being minimized in-sample because of its discontinuous nature. The role of minimizing a smooth surrogate loss function is to trace a path that leads to low classification error, but the surrogate loss is not of interest in itself. Yet, for the variance reducing properties of the resulting classifier, the surrogate loss is of interest. First of all, the surrogate loss should keep decreasing because for example discrete AdaBoost is constrained gradient descent with line search (Friedman et al., 2000). This explains why in terms of the surrogate loss, the fitted class probability estimates end up vastly overfitting the data, confirming the orthodox view in terms of the surrogate loss. Yet, two phenomena are also observed: in terms of out-of-sample classification error, no overfitting is taking place, and, according to Breiman, no convergence of the weights is taking place. The bouncing of the weights would indicate that, in spite of a well-behaved convex loss function, the descent directions chosen by the base learner become erratic. Such behavior would be plausible if the base learner is of the high-variance type, but the specifics of why the variance component of out-of-sample classification error is improved is not explained. It is quite clear, though, that explaining boosting's variance reduction would be a greater achievement than explaining its bias reduction. Bias can be largely taken care of by the base learner, variance can't.

If the peculiarities we observe in boosting are due to the use of two loss functions, one may ask whether any lessons learned carry over to other parts of statistics. The "statistical view" has indeed produced generalizations of boosting to other areas, such as regression: Bühlmann and Yu's L_2-boosting (2003), Friedman's gradient boosting in their deterministic and stochastic forms (2001; 2002), and boosting of exponential and survival models by Ridgeway (1999). In these single-loss function contexts, the paradoxical phenomena should no longer be visible, as they aren't for boosting if judged in terms of the surrogate loss function. Yet, Friedman's stochastic gradient boosting shows that adding an element of variance reduction with randomization may just be what the doctor ordered in most statistical model fitting contexts even with a single loss function. We should therefore aim for a variance reduction theory for all of statistics, reaching beyond classification.

Another question that may be raised for binary, or categorical response data in general, is whether classification error is as desirable a loss function as suggested by the attention it has received. Classification error is a bottom line number that may be appropriate in industrial contexts where real large scale engineering problems are solved, for example, in document retrieval. One might characterize these contexts as "the machine learner's black box problems." There do exist other contexts, though, and one might characterize them as "the problems of the interpreting statistician." When interpretation is the problem, attaining the last percent of classification accuracy is not the goal. Instead, one hopes to develop a functional form that reasonably fits the data but also "speaks," that is, lends itself to statements about what variables are associated with the categorical response. Fitting good conditional class probabilities takes on greater importance because associations and effects can then be measured in terms of differences in the logits of class 1 (for example) for a unit or percentage difference in the predictor variables. Interpretability is a problem for non-parametric model fits such as boosted trees. The decomposition of complex fits into interpretable components, for example with an ANOVA decomposition as suggested by Friedman et al. (2000), takes on considerable importance. In the end, one may want to produce a few telling plots explaining functional form and a few numbers summarizing the strengths of various associations. When fitting models for conditional class probabilities, the surrogate loss becomes the primary loss function because it can be interpreted as a loss function for fitted class probabilities. It is one of the achievements of Friedman et al. (2000) to have shown that this is true for exponential loss as much as for logistic loss, even though there is a misperception, as pointed out by Mease and Wyner, that LogitBoost was specifically designed to recover class probabilities that AdaBoost couldn't. Exponential loss does similar things as logistic loss in Friedman et al.'s analysis, and they provide the appropriate link functions for both. All this is relevant only if minimization of the "surrogate/now-primary loss" is prevented from overfitting, with cross-validated early stopping, penalization, or variance-reducing randomization, and it comes at the cost of diminished classification performance, one of Mease and Wyner's points.

Diminished classification performance when estimating class probabilities is easily explained; it is due to a compromise that class probability estimation has to strike. It effectively attempts good classification simultaneously at *all* misclassification cost ratios. (Note that this ratio is assumed to be one in most of the boosting literature.) This statement can be made precise in a technical sense: unbiased loss functions for class probabilities, so-called "proper scoring rules," are weighted mixtures of cost-weighted misclassification losses Buja et al. (2005). After mapping exponential and logistic losses to probability scales with their associated inverse link functions, they turn into proper scoring rules and therefore exhibit the mixture structure just described. It follows that both loss functions attempt compromises across classification problems with non-equal misclassification costs. Both loss functions give inordinate attention to extreme cost ratios, but exponential loss even more so than logistic loss. At any rate, the nature of the compromise is such that no cost ratio, in particular not equal costs, is served optimally if the exponential or logistic losses are tuned to high out-of-sample performance. By comparison, overfitting these losses in-sample seems to provide benefits in terms of classification error. However, once we change our priorities from black-box performance to interpretation, and hence from classification to class probability estimation, we may prefer tuning surrogate loss and accept the increased classification error.

Anticipating an objection by Mease and Wyner, we should disclose that one of us (Buja) collaborated with them on an article that is relevant here (Mease et al., 2007). In this work we traveled the opposite of the usual direction by composing class probability estimates from layered classification regions, estimated at a grid of misclassification cost ratios. Presumably such class probability estimation inherits superior performance from boosting in classification. When interpretation is the goal, however, a simple functional form that "speaks" might be more desirable than the increased performance of the layered estimates we provide in our joint proposal. The problem is that our proposal inherits the interpretative disadvantages of boosted classification regions, which tend to be jagged around the edges and pockmarked with holes—not a credible feature when it comes to interpretation.

We started this discussion joining Mease and Wyner in their argument against today's boosting orthodoxy. We ended by questioning the single-minded reliance on classification error as the only yard stick of performance. Still, Mease and Wyner's call should be heard because the orthodoxy misattributes the causes of boosting's success and makes invalid recommendations.

References

Y. Amit, G. Blanchard, and K. Wilder. Multiple randomized classifiers: Mrcl. Technical report, University of Chicago, 2001.

L. Breiman. Half & half bagging and hard boundary points. Technical Report 534, Statistics Dept., Univ. of California, Berkeley, 1998a.

L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998b.

L. Breiman. Random forests—random features. Technical Report 567, Statistics Dept., Univ. of California, Berkeley, 1999.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Peter Bühlmann and Bin Yu. Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, The Wharton School, University of Pennsylvania, 2005.

J. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

J. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38:367–378, 2002.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–407, 2000.

D. Mease, A. Wyner, and A. Buja. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8:409–439, 2007.

G. Ridgeway. The state of boosting. *Computing Science and Statistics*, 31:172–181, 1999.

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

A. Wyner. Boosting and the exponential loss. In *Proceedings of the Ninth Annual Conference on AI and Statistics*, 2002.