Journal of Machine Learning Research 9 (2008) 59-64 Published 1/08

Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, *JMLR 9*:1–26, 2008

Jerome Friedman JHF@STANFORD.EDU
Trevor Hastie HASTIE@STANFORD.EDU
Robert Tibshirani TIBS@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305

Editor: Yoav Freund

1. Introduction

This is an interesting and thought-provoking paper. We especially appreciate the fact that the authors have supplied R code for their examples, as this allows the reader to understand and assess their ideas. The paper inspired us to re-visit many of these issues underlying boosting methods. However in the end we do not believe that the examples provided in the paper contradict our statistical view, although other views may well prove informative.

2. Our Statistical View of Boosting

Friedman et al. (2000) and our book (Hastie et al., 2001) argue that boosting methods have three important properties that contribute to their success:

1. they fit an additive model in a flexible set of basis functions
2. they use a suitable loss function for the fitting process
3. they regularize by forward stagewise fitting; with shrinkage this mimics an L_1 (lasso) penalty on the weights.

In many cases the paper ascribes consequences of this statistical view that are not the case. For example, it does not follow that smaller trees are necessarily better than larger ones for noisier problems (Sections 3.2 and 4.2), that the basis should necessarily be restricted as described in Sections 3.6 and 4.6, or that regularization should be based on the loss function used for fitting (Sections 3.5 and 4.5). To the extent possible model selection should be based on the ultimate loss associated with the application. Also, there is no requirement that test error have a unique minimum as a function of the number of included terms (Sections 3.4 and 4.4). However, to the extent that these are commonly held beliefs, the paper provides a valuable service by pointing out that they need not hold in all applications.

There is no direct relation between the application of shrinkage and overfitting (Sections 3.7 and 4.7). Heavy shrinkage emulates L_1 regularization, whereas its absence corresponds to stagewise

fitting approximating L_0 regularization. There is nothing in the statistical view that requires L_1 to be superior to L_0 in every application, although this is often the case. The best regularizer depends on the problem: namely the nature of the true target function, the particular basis used, signal-to-noise ratio, and sample size.

Finally, there is nothing in our statistical interpretation suggesting that boosting is similar to one nearest neighbor classification (Sections 3.9 and 4.9).

None-the-less, the paper does provide some interesting examples that appear to contradict the statistical interpretation. However these examples may have been carefully chosen, and the effects seems to vanish under various perturbations of the problem.

3. Can the "Wrong" Basis Work Better than the Right One?

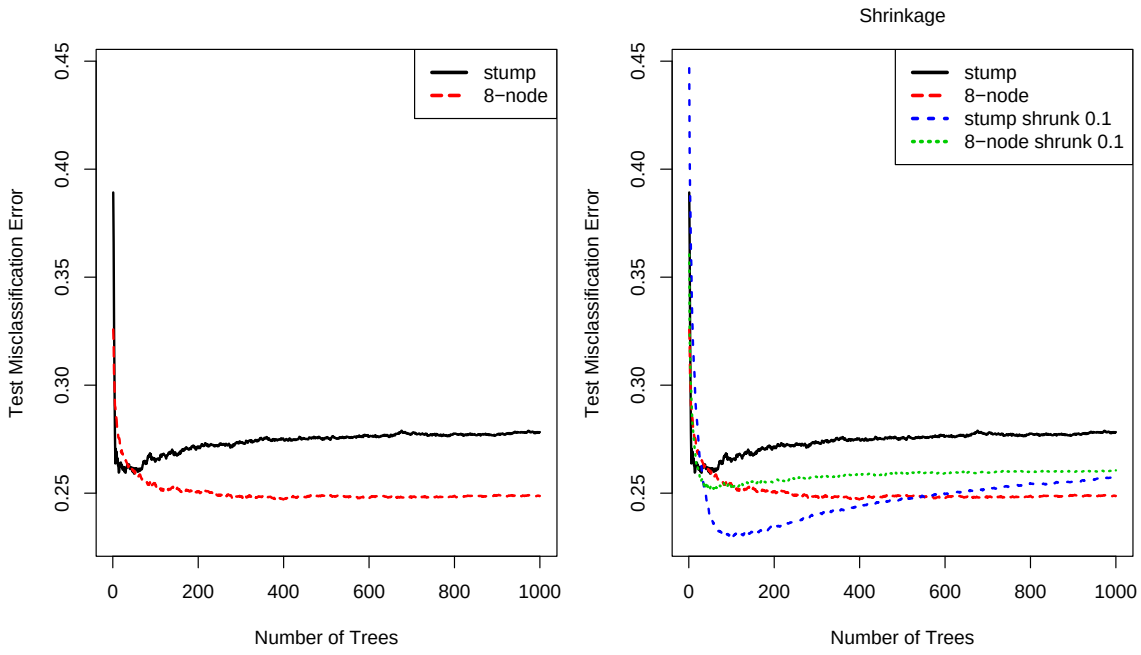

Figure 1: Average test misclassification error for 20 replications of Mease and Wyner's example used in their Figure 1. We used the package GBM in R, with the "adaboost" option. The left panel shows that 8-node trees outperform stumps. The right panel shows that stumps with shrinkage win handily.

The left panel of Figure 1 shows a version of the paper's Figure 1. We see that boosting with 8 node trees seems to outperform stumps, despite the fact that the generative model is additive in the predictor variables. However the right panel shows what happens to both stumps and 8 nodes trees when shrinkage is applied. Here shrinkage helps in both cases, and we see that stumps with shrinkage work the best of all.

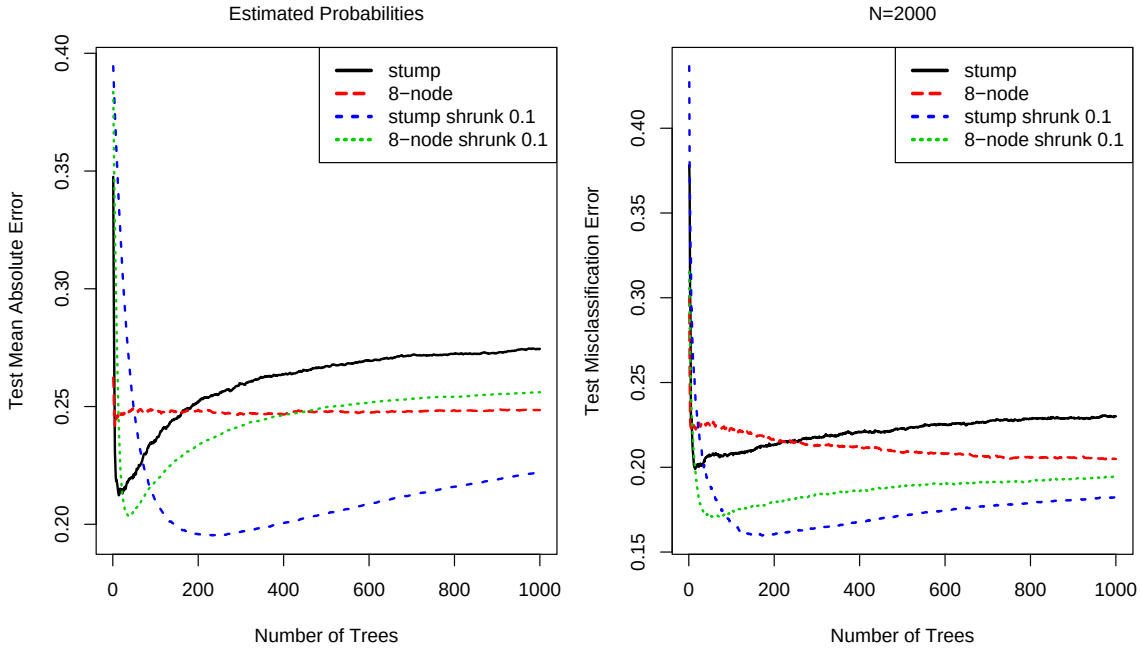

Figure 2: Left panel: average absolute deviations of the fitted probabilities from the true probabilities for the same simulations as in Figure 1. Right panel: average test misclassification error for the same simulations as in Figure 1, except using 2000 rather than 200 training examples.

We are not sure why unshrunken 8 node trees outperform unshrunken stumps in this example. As in the paper, we speculate that the extra splits in the 8 node tree might act as a type of regularizer, and hence they help avoid the overfitting displayed by unshrunken stumps in this example. All but the first split will tend to be noisy attempts at the other variables, which when averaged will have a "bagging" effect.

However this explanation becomes less convincing and indeed the effect itself seems to fade when we look more deeply. Figure 2 [left panel] shows the average absolute error in the estimated probabilities, while Figure 2[right panel] shows what happens when we increase the sample size to 2000. In Figure 3[left panel] we use the Bernoulli loss rather than exponential of Adaboost, and Figure 3[right panel] shows results for the regression version of this problem. In every case, the effect noted by the authors goes away and both the correct bases and shrinkage help performance. We repeated these runs on the second simulation example of Section 4, and the results were similar. Thus the effect illustrated by the authors is hard to explain, and seems to hold only for misclassification error. It depends on a very carefully chosen set of circumstances. Most importantly, we have to remember the big picture. Looking at the right panel of Figure 1, which method would anyone choose? Clearly, shrunken stumps work best here, just as might be expected from the statistical view.

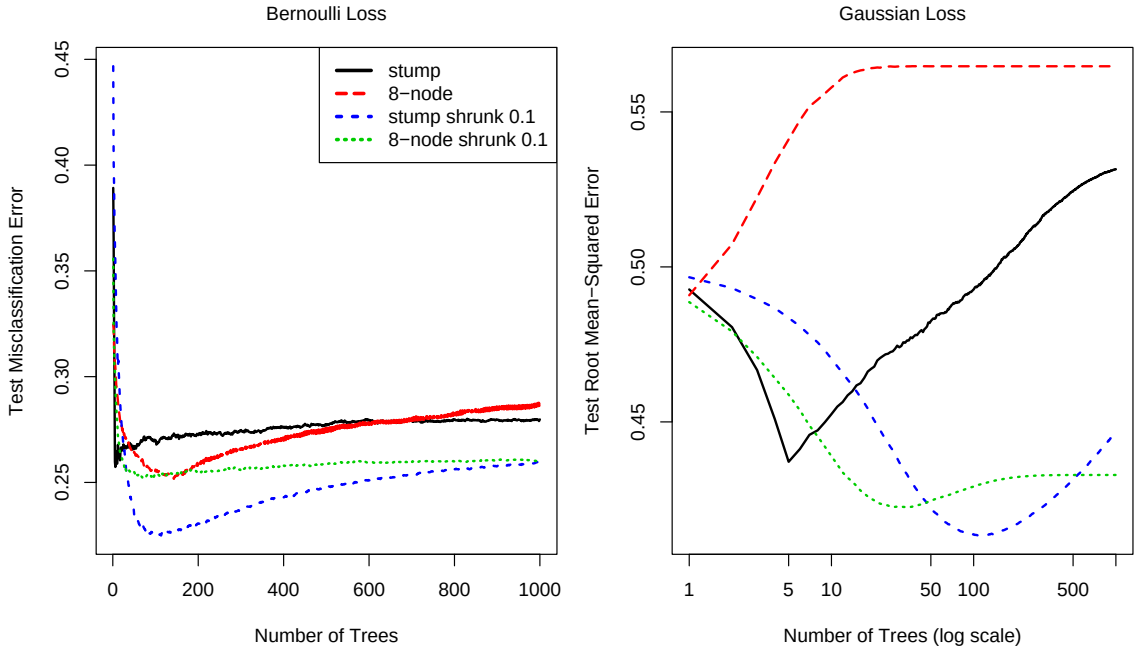

Figure 3: Left panel: test misclassification error when boosting with Bernoulli loss for the same simulations as in Figure 1. Right panel: root mean-squared test error when boosting with squared-error loss for the same simulations as in Figure 1 (legend as in left panel).

Figure 4 shows the fitted probabilities over the 20 runs, separately for each class, when using 250 shrunken stumps. Here 250 was chosen since it corresponds to the minimum in Figure 2[left panel]. This is an appropriate tradeoff curve if we are interested in probabilities; test deviance would also be fine. We see that the estimates are biased toward 0.5, which is expected when regularization is used. Hence they are underfit, rather than overfit.

A similar argument can be made concerning the paper's Figure 3. Yes, AdaBoost works better than Logitboost in this example. But using the statistical view of boosting, we have moved on and developed better methods like gradient boosting (Friedman, 2001) that typically outperform both of these methods.

Hastie et al. (2007) add further support to (3) of the statistical interpretation of boosting: they show that the incremental forward stagewise procedure used in boosting (with shrinkage) optimizes a criterion similar to but smoother than the L_1 penalized loss.

4. Conclusion

No theory, at least initially, can fully explain every observed phenomenon. Everything about regularized regression is not yet fully understood. There is still considerable ongoing research in the literature concerning the interplay between the target function, basis used, and regularization method. Hopefully, some of the apparent anomalies illustrated in this paper will eventually be explained with

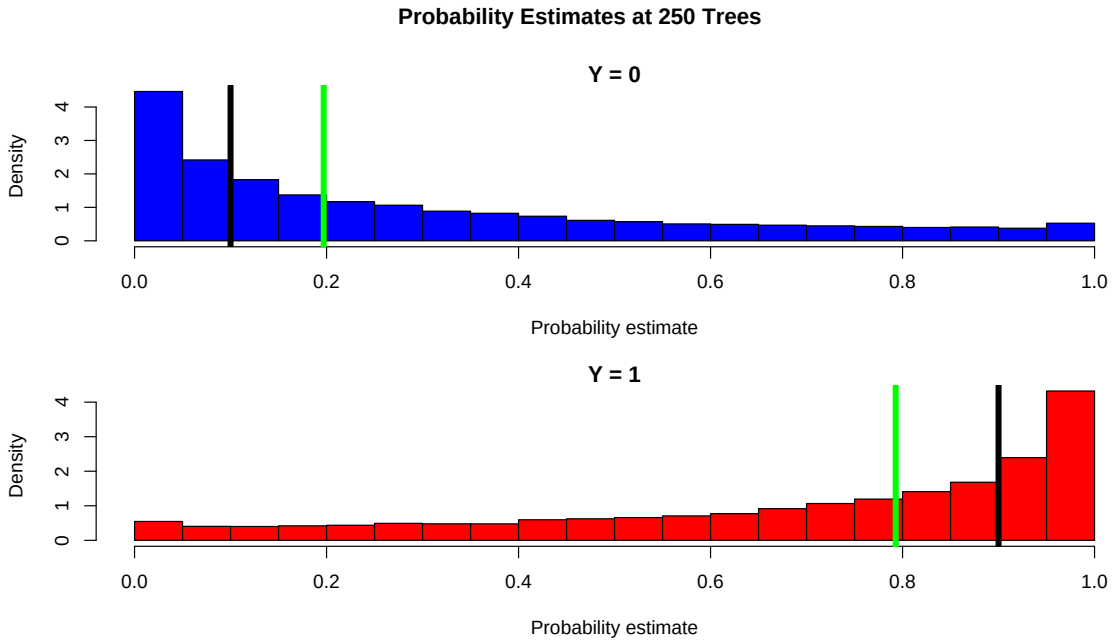

Figure 4: Fitted probabilities shown for the two classes, at 250 shrunken stumps. The vertical black bars are the target probabilities for this problem, and the green bars are the median of the estimates in each class.

a more thorough understanding of these issues. The paper provides a service in reminding us that there is still work remaining.

Although we would not begin to suggest that our statistical view of boosting has anywhere near the substance or importance of the Darwin's theory of evolution, the latter provides a useful analogy. The proponents of Intelligent Design point out that the theory of evolution does not seem to explain certain observed biological phenomena. And therefore they argue that evolution must be wrong despite the fact that it *does* explain an overwhelming majority of observed phenomena, and without offering an alternative *testable* theory.

We are sure that the authors will mount counter-arguments to our remarks, and due to the (standard format) of this discussion, they will have the last word. We look forward to *constructive* counter-arguments and alternative explanations for the success of boosting methods that can be used to extend their application and produce methods that perform better in practice (as in the right panel of Figure 1).

References

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

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

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer Verlag, New York, 2001.

T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29 (electronic), 2007. DOI: 10.1214/07-EJS004.