Rejoinder

By Gábor Lugosi and Nicolas Vayatis

Pompeu Fabra University, Barcelona and Université Paris 6–Pierre et Marie Curie

We thank the discussants for the interesting comments which shed light on many different aspects of boosting and related methods for classification and regression. In this rejoinder we summarize what we have learnt about boosting since the writing of the paper, in great part thanks to these discussion pieces.

The new and elegant proof of the consistency theorem of Koltchinskii is not only amusing but also shows the way how many seemingly different classifiers, including regularized boosting and support vector machines, can be analyzed in a single framework. The main message of Bartlett, Jordan, and McAuliffe is similar in that they consider so-called large-margin classification methods which minimize a certain empirical loss function of the margin different from the empirical probability of error and characterize the loss functions which lead to consistent classification. The generality of these conditions is surprising and again, develops a unified treatment that encompasses not only various versions of boosting methods but also support vector machines and related kernel-based methods.

We agree with Freund and Schapire that consistency is just a minimal requirement and does not explain the good practical behavior of boosting. Once consis-

AMS 2000 subject classifications.
Key words and phrases.

tency is established, attention should be turned to a finer analysis. Koltchinskii points out the importance of establishing rates of convergence. However, it is not completely obvious what the reasonable assumptions are for the distribution in high-dimensional classification problems. We share the view of Friedman, Hastie, Rosset, Tibshirani, and Zhu that sparsity should play a key role. We believe that the analysis of consistency provides valuable insight into the behavior of boosting. Indeed, building partly on the techniques of the discussed papers by Zhang and us, and on the recent paper of Bartlett, Jordan, and McAuliffe (cited in their discussion), in a recent joint work with Gilles Blanchard [1] we have been able to derive rate-of-convergence results for regularized boosting methods similar to the ones studied in our paper. As it turns out, some regularized boosting methods produce classifiers whose probability of error converges to the Bayes error at a rate independent of the dimension (faster than $O(n^{-1/4})$ and sometimes as fast as $O(n^{-1/2})$) for large classes of distributions. This is an interesting feature not shared by classical nonparametric methods such as the k-nearest neighbor classifier, as it is also pointed out by Freund and Schapire. The distributions under which such a rate of convergence holds are those for which the function f^* minimizing the cost function $A(f) = \mathbb{E}\phi(-f(X)Y)$ can be approximated arbitrarily (say, in the L_∞ sense) by linear combinations of base functions with coefficients bounded in L_1. The characterization of these distributions is far from being trivial in general, but in some cases it is well understood. As an example, we cite the following special case from [1]:

COROLLARY 1. *Let $X \in \mathbb{R}^d$ with $d \geq 2$. There exists a regularized boosting classifier $\widehat{f}_n$ based on the logit cost function and decision stumps such that if there exist functions $f_1, \ldots, f_d : \mathbb{R} \to \mathbb{R}$ and a positive constant B such that the sum of the*

total variations of the f_i is bounded by Bd and such that $\log \frac{\eta(x)}{1-\eta(x)} = \sum_{i=1}^{d} f_i(x^{(i)})$ then for every n, with probability at least $1 - 1/n^2$, the probability of error $L(\widehat{f}_n)$ satisfies

$$L(\widehat{f}_n) - L^* \leq C\sqrt{d \log d} n^{-\frac{1}{2(2-\alpha)}\left(\frac{V_d+2}{V_d+1}\right)}$$

where C is a universal constant, $V_d \leq 2\log_2(2d)$ and the value of $\alpha \in [0,1]$ depends on the distribution.

This result quantifies the observation of Friedman, Hastie, and Tibshirani [2] who pointed out a close relationship between boosting and additive logistic regression. The example described by Bühlmann and Yu fits exactly in the framework of this corollary and explains the good behavior of LogitBoost in their simulations. Interestingly, the same result is not true when the exponential cost function is used. In that case, even though the rate of convergence in terms of the sample size remains the same, the dimension-dependent constant in front grows exponentially rapidly with d. It is a remarkable fact that the dimensionality only appears in the multiplicative constant of the rate of convergence. We believe that, even though now we are closer to the understanding of boosting and related methods, there is still a lot to discover and interesting unexplored questions abound.

Freund and Schapire point out that in very high dimensional problems boosting may not be computationally feasible if the base class is one of the usual classes (e.g., decision trees with $d+1$ extremal nodes) which guarantee universal consistency. In such cases one may have to resort to smaller base classes such as decision stumps. The corollary above shows that boosting based on stumps has an excellent behavior if the distribution happens to follow an additive logistic model. However, one should proceed with care when using such "incomplete" base classes. It is shown in [1] that boosting (and other large-margin methods which minimize an empir-

ical cost functional) may have a catastrophic behavior if the function f^* cannot be approximated by linear combinations of base functions in the sense that the resulting classifier may have a probability of error which is much larger not only than the Bayes error but also than the error of the best classifier realizable by linear combinations of base classifiers. Thus, an interesting open problem is to find "simple" base classes which are dense in the sense that all possible classifiers can be approximated by convex combinations of base classifiers. In a recent manuscript [4] we show the existence of such a class of VC dimension 2, independently of the dimension of the space. While the construction given in that paper is probably of little practical value, a better understanding of the tradeoff between computational complexity and approximation ability is an important challenge.

Another important issue that Freund and Schapire raise is that by minimizing an empirical cost function such as the exponential or the logit functions one implicitly estimates the whole conditional probability function $\eta(x)$ (more precisely, a monotone function of it). By doing that, one does more than necessary since in binary classification the only thing that matters is whether $\eta(x)$ is greater or smaller than $1/2$. The results of Bartlett, Jordan, and McAuliffe refine this point of view by showing that under conditions on the behavior of $\eta(x)$ around $1/2$ (introduced by Tsybakov) the rate of convergence of boosting methods speeds up considerably. (The constant α in the corollary above is determined by the behavior of $\eta(x)$ in the vicinity of $1/2$.) There is one convex cost function, the "hinge loss" used by support vector machines, which has the distinguishing property that its minimizer is the Bayes classifier g^* itself, see Lin [3]. Thus, as opposed to boosting, support vector classifiers do just what they are supposed to do, and do not "waste energy" in estimating the function $\eta(x)$ in irrelevant ranges. However, this does not necessarily

mean that support vector machines perform better as for the hinge loss it seems to become more difficult to approximate the minimizer f^* by linear combinations of base classifiers. Once again, the relationship of minimizers of different empirical cost functions is complex, very far from being well understood.

The discussion of Bühlmann and Yu tackles algorithmic issues of regularized boosting procedures. In our experiments, we used MarginBoost.L1 as a convergent algorithm giving a nearly optimal output in the λ-blowup of the convex hull of the base class (for a *fixed* value λ of the smoothing parameter). Running this algorithm for various values of λ revealed that this smoothing parameter was effectively acting as a relevant complexity measure even for small sample sizes. The discussion of Friedman, Hastie, Rosset, Tibshirani, and Zhu, pointing out the connection of regularized boosting methods with L_1-penalty to Tibshirani's Lasso, provides a strong intuition on how the practical problem of finding efficient greedy algorithms can be dealt with.

Bühlmann and Yu also comment on the importance of distinguishing between regularizing by an explicit constraint on the sum (or other norm) of the weights and by early stopping. This is an important and difficult question. The very interesting results of Bickel and Ritov show in a general framework that stopping by cross validation works in a strong sense. While early stopping is alluring from a practical point of view (it reduces to AdaBoost, plus a stopping rule), its theoretical analysis is more problematic. Indeed, in most cases, it turns out that there is an optimal value for the smoothing parameter $\lambda = \lambda^*$ (corresponding to the L_1-norm of the weights of the optimal combination). The successive iterations in AdaBoost can be conceived as drawing a path in the space of the weights crossing the iso-surfaces defined by constant values of the L_1-norm of the weights, and

early stopping returns an output on this path which may be close to the optimal vector of weights. Since there is no known guarantee that during the iterations the weight vector passes through a near-optimal value for the best choice λ^*, it seems to be difficult to derive rate-of-convergence results such as the corollary above for AdaBoost or LogitBoost with early stopping. To better understand the relationship between explicitly regularized boosting and early stopped AdaBoost is a challenging problem that requires a careful study of the approximation properties of the iterative construction of the boosting estimator based on highly redundant dictionaries of base classifiers. We entirely agree with Friedman, Hastie, Rosset, Tibshirani, and Zhu, who point out the importance of sparsity. We believe that these perspectives motivate interesting research at the interface of statistics, optimization and approximation theories.

REFERENCES

[1] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. manuscript, 2003.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337-374, 2000.

[3] Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259-275, 2002.

[4] G. Lugosi and S. Mendelson. A note on the richness of convex hulls of VC classes. manuscript, 2003.