The Convergence Rate of AdaBoost

Robert E. Schapire
Princeton University
Department of Computer Science
schapire@cs.princeton.edu

Abstract. We pose the problem of determining the rate of convergence at which AdaBoost minimizes exponential loss.

Boosting is the problem of combining many "weak," high-error hypotheses to generate a single "strong" hypothesis with very low error. The AdaBoost algorithm of Freund and Schapire (1997) is shown in Figure 1. Here we are given m labeled training examples $(x_1, y_1), \ldots, (x_m, y_m)$ where the x_i's are in some domain $\mathcal{X}$, and the labels $y_i \in \{-1, +1\}$. On each round t, a distribution D_t is computed as in the figure over the m training examples, and a weak hypothesis $h_t : \mathcal{X} \to \{-1, +1\}$ is found, where our aim is to find a weak hypothesis with low weighted error ϵ_t relative to D_t. In particular, for simplicity, we assume that h_t minimizes the weighted error over all hypotheses belonging to some finite class of weak hypotheses $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$.

The final hypothesis H computes the sign of a weighted combination of weak hypotheses $F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$. Since each h_t is equal to $\hbar_{j_t}$ for some j_t, this can also be rewritten as $F(x) = \sum_{j=1}^{N} \lambda_j \hbar_j(x)$ for some set of values $\boldsymbol{\lambda} = \langle \lambda_1, \ldots \lambda_N \rangle$. It was observed by Breiman (1999) and others (Frean & Downs, 1998; Friedman et al., 2000; Mason et al., 1999; Onoda et al., 1998; Rätsch et al., 2001; Schapire & Singer, 1999) that AdaBoost behaves so as to minimize the exponential loss

$$L(\boldsymbol{\lambda}) = \frac{1}{m} \sum_{i=1}^{m} \exp\left(- \sum_{j=1}^{N} \lambda_j y_i \hbar_j(x_i) \right)$$

over the parameters $\boldsymbol{\lambda}$. In particular, AdaBoost performs coordinate descent, on each round choosing a single coordinate j_t (corresponding to some weak hypothesis $h_t = \hbar_{j_t}$) and adjusting it by adding α_t to it: $\lambda_{j_t} \leftarrow \lambda_{j_t} + \alpha_t$. Further, AdaBoost is greedy, choosing j_t and α_t so as to cause the greatest decrease in the exponential loss.

In general, the exponential loss need not attain its minimum at any finite $\boldsymbol{\lambda}$ (that is, at any $\boldsymbol{\lambda} \in \mathbb{R}^N$). For instance, for an appropriate choice of data (with $N = 2$ and $m = 3$), we might have

$$L(\lambda_1, \lambda_2) = \tfrac{1}{3} \left(e^{\lambda_1 - \lambda_2} + e^{\lambda_2 - \lambda_1} + e^{-\lambda_1 - \lambda_2} \right).$$

The first two terms together are minimized when $\lambda_1 = \lambda_2$, and the third term is minimized when $\lambda_1 + \lambda_2 \to +\infty$. Thus, the minimum of L in this case is attained when we fix $\lambda_1 = \lambda_2$, and the two weights together grow to infinity at the same pace.

Let $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2, \ldots$ be the sequence of parameter vectors computed by AdaBoost in the fashion described above. It is known that AdaBoost asymptotically converges to the minimum possible exponential loss (Collins et al., 2002). That is,

$$\lim_{t \to \infty} L(\boldsymbol{\lambda}^t) = \inf_{\boldsymbol{\lambda} \in \mathbb{R}^N} L(\boldsymbol{\lambda}).$$

However, it seems that only extremely weak bounds are known on the rate of convergence, for the most general case. In particular, Bickel, Ritov and Zakai (2006) prove a very weak bound of the form $O(1/\sqrt{\log t})$ on this rate. Much better bounds are proved by Rätsch, Mika and Warmuth (2002) using results from Luo and Tseng (1992), but these appear to require that the exponential loss be minimized by a finite $\boldsymbol{\lambda}$, and also depend on quantities that are not easily measured. Shalev-Shwartz and Singer (2008) prove bounds for a variant of AdaBoost. Zhang and Yu (2005) also give rates of convergence, but their technique requires a bound on the step sizes α_t. Many classic results are known on the convergence of iterative algorithms generally (see for instance, Luenberger and Ye (2008), or Boyd and Vandenberghe (2004)); however, these typically start by assuming that the minimum is attained at some finite point in the (usually compact) space of interest.

When the weak learning assumption holds, that is, when it is assumed that the weighted errors ϵ_t are all upper bounded by $1/2 - \gamma$ for some $\gamma > 0$, then it is known (Freund & Schapire, 1997; Schapire & Singer, 1999) that the exponential loss is at most $e^{-2t\gamma^2}$ after t rounds, so it clearly quickly converges to the

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$
space $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$ of weak hypotheses $\hbar_j : \mathcal{X} \to \{-1, +1\}$
Initialize: $D_1(i) = 1/m$ for $i = 1, \ldots, m$.
For $t = 1, \ldots, T$:

- Train weak learner using distribution D_t; that is, find weak hypothesis $h_t \in \mathcal{H}$ that minimizes the weighted error $\epsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$.
- Choose $\alpha_t = \frac{1}{2} \ln((1 - \epsilon_t)/\epsilon_t)$.
- Update, for $i = 1, \ldots, m$: $D_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i h_t(x_i))/Z_t$
where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis: $H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Figure 1: The boosting algorithm AdaBoost.

minimum possible loss in this case. However, here our interest is in the general case when the weak learning assumption might not hold.

This problem of determining the rate of convergence is relevant in the proof of the consistency of AdaBoost given by Bartlett and Traskin (2007), where it has a direct impact on the rate at which AdaBoost converges to the Bayes optimal classifier (under suitable assumptions).

We conjecture that there exists a positive constant c and a polynomial $\mathrm{poly}()$ such that for all training sets and all finite sets of weak hypotheses, and for all $B > 0$,

$$L(\boldsymbol{\lambda}^t) \leq \min_{\boldsymbol{\lambda}: \|\boldsymbol{\lambda}\|_1 \leq B} L(\boldsymbol{\lambda}) + \frac{\mathrm{poly}(\log N, m, B)}{t^c}.$$

Said differently, the conjecture states that the exponential loss of AdaBoost will be at most ε more than that of any other parameter vector $\boldsymbol{\lambda}$ of ℓ_1-norm bounded by B in a number of rounds that is bounded by a polynomial in $\log N$, m, B and $1/\varepsilon$. (We require $\log N$ rather than N since the number of weak hypotheses $N = |\mathcal{H}|$ will typically be extremely large.) The open problem is to determine if this conjecture is true or false, in general, for AdaBoost. The result should be general and apply in all cases, even when the weak learning assumption does not hold, and even if the minimum of the exponential loss is not realized at any finite vector $\boldsymbol{\lambda}$. The prize for a new result proving or disproving the conjecture is US$100.

References

Bartlett, P. L., & Traskin, M. (2007). AdaBoost is consistent. *Journal of Machine Learning Research*, *8*, 2347–2368.

Bickel, P. J., Ritov, Y., & Zakai, A. (2006). Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, *7*, 705–732.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Breiman, L. (1999). Prediction games and arcing classifiers. *Neural Computation*, *11*, 1493–1517.

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*.

Frean, M., & Downs, T. (1998). *A simple cost function for boosting* (Technical Report). Department of Computer Science and Electrical Engineering, University of Queensland.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, *38*, 337–374.

Luenberger, D. G., & Ye, Y. (2008). *Linear and nonlinear programming*. Springer. Third edition.

Luo, Z. Q., & Tseng, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, *72*, 7–35.

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Functional gradient techniques for combining hypotheses. In *Advances in large margin classifiers*. MIT Press.

Onoda, T., Rätsch, G., & Müller, K.-R. (1998). An asymptotic analysis of AdaBoost in the binary classification case. *Proceedings of the 8th International Conference on Artificial Neural Networks* (pp. 195–200).

Rätsch, G., Mika, S., & Warmuth, M. K. (2002). On the convergence of leveraging. *Advances in Neural Information Processing Systems 14*.

Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for AdaBoost. *Machine Learning*, *42*, 287–320.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, *37*, 297–336.

Shalev-Shwartz, S., & Singer, Y. (2008). On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *21st Annual Conference on Learning Theory*.

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