Technical Reports on Mathematical and Computing Sciences:TR-C194
title: Learning r-of-k Functions by Boosting
author: Kohei Hatano and Osamu Watanabe
affiliation:
Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology
W8-25, Tokyo 152-8552
email: watanabe@is.titech.ac.jp
acknowledgments to financial supports:
Supported in part by a Grant-in-Aid for Scientific Research on Priority Areas
"Statical-Mechanical Approach to Probabilistic Information Processing" 2002-2005.

Abstract. We investigate further improvement of boosting in the case that the target concept belongs to the class of r-of-k threshold Boolean functions, which answers "+1" if at least r of k relevant variables are positive, and answers "−1" otherwise. Given m examples of a r-of-k function and literals as base hypotheses, popular boosting algorithms (e.g., AdaBoost [FS97]) construct a consistent final hypothesis in $O(k^2 \log m)$ iterations. While this convergence speed is tight in general, we show that a modification of AdaBoost (confidence-rated AdaBoost [SS99] or InfoBoost [Asl00]) can make use of the property of r-of-k functions that make less error on one-side to find a consistent final hypothesis in $O(kr \log m)$ iterations. Our result extends the previous investigation by Hatano and Warmuth [HW04] and gives more general examples where confidence-rated AdaBoost or InfoBoost has an advantage over AdaBoost.

1 Introduction

Boosting is a method of constructing strongly accurate classifiers by combining weakly accurate base classifiers, and it is now one of the most fundamental techniques in machine learning. A typical boosting algorithm proceeds in rounds as follows: At each round, it assigns a distribution on the given data, and finds a moderately accurate base classifier with respect to the distribution. Then it updates the distribution so that the last chosen classifier is insignificant for it. It has been shown that this algorithm converges relatively quickly under the assumption that a base hypothesis with error less than $1/2 - \gamma$ can be found at each iteration. For example, under this assumption, Freund showed [Fre95] a boosting algorithm and proved that it yields a ε-approximation of the correct hypothesis in $O((1/\gamma^2)\log(1/\varepsilon))$ iterations. Freund also proved that the result is in a sense tight; that is, if the final hypothesis is of a majority of weak hypotheses (like his boosting algorithm does), then any boosting algorithm needs, in the worst case, $\Omega((1/\gamma^2)\log(1/\varepsilon))$ iterations to obtain an ε-approximation. While Freund's result is tight in general, it is still possible to improve boosting for some particular cases. For example, Dasgupta and Long [DL03] showed that the above boosting algorithm by Freund works more efficiently when the set of base classifiers contains the subset of good "diverse" classifiers. This paper studies one of such cases; here we consider the situation where error occurs less often in one side.

Let us see the following illustrative example. Consider learning k-disjunctions. That is, given m labeled examples for a target k-disjunction f over n Boolean variables, our task is to find a hypothesis consistent with f on these examples. Let $\mathcal{H}$ be the set of very simple Boolean functions that are either a function defined by one literal or a constant $+1$ or -1, and let $\mathcal{H}_+$ be the subset of $\mathcal{H}$ consisting of functions defined by one Boolean variable, i.e., a positive literal. By using functions in $\mathcal{H}$ as base classifiers, a popular boosting algorithm (e.g., AdaBoost [FS97]) can find a consistent hypotheses in $O(k^2 \log m)$ iterations. In contrast, a use of the greedy set-covering algorithm can accomplish the same task in $O(k \log m)$ iterations, as learning of k-disjunctions can be reduced to the set-covering problem [Nat91]. Then it is natural to ask whether boosting algorithms can beat the greedy set-covering algorithm in this learning problem. Hatano and Warmuth [HW04] showed that a modification of AdaBoost, called confidence-rated AdaBoost [SS99] or InfoBoost [Asl00] (in the following, we simply call it InfoBoost), can achieve $O(k \log m)$ bound. Moreover, they showed that AdaBoost does need $\Omega(k^2 \log m)$ iterations under some technical conditions.

In the above case, we may assume that all base hypotheses in $\mathcal{H}_+$ have one-sided error property; that is, their positive predictions are always correct, in other words, all these base hypotheses have 0 false positive error. InfoBoost is designed so that it can make use of this situation, and this is why it performs better than AdaBoost. In general, when this one-sided error property can be assumed for good base hypotheses, InfoBoost can reduce the number of boosting iterations from $O((1/\gamma^2) \log m)$ to $O((1/\gamma) \log m)$. Besides this extreme case, we may naturally expect that InfoBoost can also make use of "partial" one-sided error property. This is the question we discuss in this paper.

As a concrete example of "partial" one-sided error situations, we consider here the problem of learning r-of-k threshold functions. For any r and k, a r-of-k function (over n Boolean variables) consists of k relevant variables, and it answers $+1$ if at least r of the k variables are positive, and otherwise answers -1. We learn these functions by using again $\mathcal{H}$ as the set of base classifiers. Note that k-disjunctions are nothing but 1-of-k functions; hence, our learning task is a generalization of learning k-disjunctions. Also for small r, we may expect that hypotheses in $\mathcal{H}_+$ have *on average* small (but may not be 0) false positive error. We show that InfoBoost can still make use of this one-sided error property and find a consistent hypothesis in $O(kr \log m)$ iterations. Thus, the running time of InfoBoost is from $O(k \log m)$ (when $r = O(1)$, e.g., k-disjunctions) to $O(k^2 \log m)$ (when $r = \Omega(k)$, e.g., k-variable majority functions).

Technically our contribution is to develop a way to show a base hypothesis with reasonable advantage *and* small error probability on one side from the fact that $\mathcal{H}$ has *on average* reasonable advantage and small false positive error.

2 Preliminaries

In this section, we give basic definitions. Throughout this paper, the domain of interest is $\mathcal{X} = \{-1, +1\}^n$. By a "Boolean function" we will mean a function from $\mathcal{X}$ to $\{-1, +1\}$. We use X_i, $i = 1 \ldots$, to denote Boolean variables, which naturally define functions from $\mathcal{X}$ to $\{-1, +1\}$. Let F_k^r be the set of r-of-k functions, i.e., Boolean functions on k Boolean variables that return $+1$ if at least r of these k variables are assigned $+1$. For any $f \in F_k^r$, $\mathrm{Rel}(f)$ is the set of the relevant Boolean variables of f.

For any $f \in F_k^r$, a pair $(x, f(x))$ is called an *example* of f. A set of such pairs is called a *sample*. We will use $\mathcal{H} = \{X_1, \ldots, X_n, +1, -1\}$ as the set of base hypotheses (or weak hypotheses).

For any distribution D, any Boolean function f, and any hypothesis $h \in \mathcal{H}$, we define

$$\begin{array}{lll} \mathrm{err}|_D^+(f,h) &=& \Pr_D[\, f(x) = +1 \,|\, h(x) = -1\,], \quad \mathrm{err}_D^+(f,h) = \Pr_D[\, f(x) = +1, h(x) = -1\,], \\ \mathrm{err}|_D^-(f,h) &=& \Pr_D[\, f(x) = -1 \,|\, h(x) = +1\,], \quad \mathrm{err}_D^-(f,h) = \Pr_D[\, f(x) = -1, h(x) = +1\,], \\ \multicolumn{3}{l}{\mathrm{error}_D(f,h) = \mathrm{err}_D^+(f,h) + \mathrm{err}_D^-(f,h).} \end{array}$$

We will omit D and f in the notation when it is clear from the context.

Given: A set S of m examples $\{\,(x,y) \,|\, x \in \mathcal{X}, y = f(x)\,\}$.
We denote elements of S as (x_1, y_1), ..., (x_m, y_m).

begin

1. Let $D_1(i) = 1/m$, for $i = 1, ..., m$. (I.e., $D_1(i) = D_0(x_i)$.)
2. For $t = 1$ to T, repeat the following procedures:
   a) Select $h_t \in \mathcal{H}$ that minimizes Z_t defined below (see (1)).
   b) Let $\alpha_t(+1) = \frac{1}{2} \ln \frac{1-\mathrm{err}|^-(h_t)}{\mathrm{err}|^-(h_t)}$ and $\alpha_t(-1) = \frac{1}{2} \ln \frac{1-\mathrm{err}|^+(h_t)}{\mathrm{err}|^+(h_t)}$.
   c) Define the next distribution D_{t+1} as follows for each $i = 1, ..., m$,
   $$D_{t+1}(i) = \frac{D_t(i) \exp\{-y_i \cdot \alpha_t(h_t(x_i)) \cdot h_t(x_i)\}}{Z_t},$$
   (Z_t defined by (1) is the normalization factor.)
3. Output the final hypothesis f_{out} defined by
   $$f_{\mathrm{out}}(x) = sign\left(\sum_{t=1}^{T} \alpha_t(h_t(x_i)) \cdot h_t(x_i) \right).$$

Figure 1: A Boosting Algorithm

For our learning algorithm, we will consider InfoBoost [Asl00] (or confidence-rated AdaBoost [SS99]) with $\mathcal{H}$ as the set of weak hypothesis; Figure 2 shows the description of this boosting algorithm on a given sample $S = \{\,(x,y) \,|\, x \in \mathcal{X}, y = f(x)\,\}$ of m examples for some target concept f. Recall that the goal of this algorithm is to obtain a (combined) hypothesis f_{out} that is consistent with f on S, which we will simply call a *consistent hypothesis*. More precisely, for any $x \in \mathcal{X}$, define $D_0(x) = 1/m$ if x appears in S and $D_0(x) = 0$ otherwise. Then the consistent hypothesis, which we want to get, is f_{out} such that $\mathrm{error}_{D_0}(f, f_{\mathrm{out}}) = 0$. For this goal, the following boosting property is known [SS99, Asl00].

Theorem 1. For any $T \geq 1$, let f_{out} be the hypothesis that the boosting algorithm yields after T iterations. Then the following bound holds for the training error of this hypothesis.

$$\mathrm{error}_{D_0}(f, f_{\mathrm{out}}) \;\leq\; \prod_{t=1}^{T} Z_t,$$

Here Z_t is the normalization factor of the tth iteration, which is defined as follows.

$$Z_t \;=\; \sum_{1 \leq i \leq m} D_t(i) \exp\{-y_i \alpha(h_t(x_i)) h_t(x_i)\}. \tag{1}$$

3 Main Result

Theorem 1 gives the convergence speed of the boosting algorithm of Figure 2. In our setting, it is easy to bound Z_t by $\sqrt{1-1/O(k^2)}$ for any r-of-k target function, from which it is shown that the algorithm yields a consistent hypothesis in $O(k^2 \log m)$ steps. Our technical contribution is to improve this bound to $\sqrt{1-1/O(rk)}$, thereby improving the upper bound for the boosting steps to $O(kr \log m)$.

In the following, for any $r \geq 1$ and $k \geq 1$, consider any r-of-k function f for our target concept, let it be fixed, and we will estimate the normalization factor Z_t of each step $t \geq 1$. Our goal is to prove the following bound. (For Z_1, by definition we have $Z_1 \leq 1$.)

Theorem 2. For any $t \geq 2$, we have

$$Z_t \ \leq \ \sqrt{1 - \frac{1}{128rk}}.$$

We first show some key properties used to derive this bound. In the following, let any $t \geq 2$ be fixed, and let us simply denote Z_t as Z. Also let D be the distribution D_t defined in our algorithm at the tth iteration. Precisely speaking, D_t is defined on $i = 1, ..., m$; but we may regard this as a distribution D_t on $\mathcal{X}$ such that $D_t(x) = D_t(i)$ if $x = x_i$ for some $i = 1, ..., m$, and $D_t(x) = 0$ otherwise. Let D denote this distribution. First we recall that D is *balanced* (with respect to f). That is, the following fact holds.

Fact 1. For our distribution D, we have $\Pr_D[f(x) = +1] = \Pr_D[f(x) = -1]$.

Using this property and the assumption that f is a r-of-k function, the following weak hypothesis can be guaranteed. In the following, let $\mathcal{H}'$ be the nonconstant hypotheses of $\mathcal{H}$; that is, $\mathcal{H}' = \{X_1, \ldots, X_n\}$. Also when $h = X_i$, we simply write $\mathrm{err}|_i^+$ and $\mathrm{err}|_i^-$ for $\mathrm{err}|_D^+(f,h)$ and $\mathrm{err}|_D^-(f,h)$ respectively.

Lemma 1. There exists a hypothesis $h \in \mathcal{H}'$ (in fact, some $h \in \mathrm{Rel}(f)$) with $\mathrm{error}(h) \leq \dfrac{1}{2} - \dfrac{1}{2k}$.

Proof. Consider the total number of mistakes made by hypotheses X_i in $\mathrm{Rel}(f)$. For any positive instance, the total number of mistakes is at most $k-r$. Hence, we have $\sum_{X_i \in \mathrm{Rel}(f)} \mathrm{err}_i^+ = \sum_{X_i \in \mathrm{Rel}(f)} \Pr_D\{X_i(x) = -1, f(x) = 1\} \leq (k-r) \cdot \Pr_D\{f(x) = 1\} = (k-r)/2$, because D is balanced. By a similar argument for negative instances, we have $\sum_{X_i \in \mathrm{Rel}(f)} \mathrm{err}_i^- \leq (r-1)/2$. In total, we have

$$\sum_{X_i \in \mathrm{Rel}(f)} \mathrm{error}(X_i) \ \leq \ \frac{k-r}{2} + \frac{r-1}{2} \ = \ \frac{k-1}{2}.$$

On the other hand, $\mathrm{Rel}(f)$ has k variables; therefore, by the averaging argument, there should be at least one X_i satisfying the lemma. □

This bound is somewhat standard. In fact, using this bound, one can bound Z by $\sqrt{1-1/(2k)^2}$. While this bound cannot be improved in general, we may be able to have better bounds if the error is estimated separately for positive and negative instances, which is the advantage of using InfoBoost. For example, if the target function f is the disjunction of some k variables (i.e.,

$r = 1$), then any $h \in \mathrm{Rel}(f)$ does not make a mistake on negative instances; thus, it follows above that there exists $h \in \mathcal{H}'$ whose general error probability is bounded by $1/2 - 1/2k$ *and* whose error probability on negative instances (i.e., $\mathrm{err}^-(h)$) is 0. The bound $O(k \log m)$ [HW04] is derived from this observation. Here we would like to generalize this argument. In the case of k-disjunctions (or 1-of-k functions), it is guaranteed that base classifiers in $\mathcal{H}'$ have good advantage on average and *all* of them have 0 false positive error. But now the assumption we can use is that they have good advantage and small false positive error both *on average*. Our technical contribution is to derive a similar consequence from this assumption.

Lemma 2. There exists a hypothesis $h \in \mathcal{H}'$ (in fact, $h \in \mathrm{Rel}(f)$) such that the following holds for some integer a, $1 \leq a \leq 4k$ with constant $c_0 = 2$.

$$\mathrm{error}(h) \leq \frac{1}{2} - \frac{a}{8k}, \quad \text{and} \tag{2}$$

$$\mathrm{err}^-(h) \leq \frac{c_0(r-1)a^2}{k}. \tag{3}$$

Proof. Consider k hypotheses $h_1, ..., h_k$ in $\mathrm{Rel}(f)$; that is, each h_i is a Boolean variables relevant to the target function f. Consider first the case that there are more than $k/2$ "relatively good" h_i in $\mathrm{Rel}(f)$ that satisfies $\mathrm{error}(h_i) \leq 1/2 - 1/4k$. As discussed in the proof of Lemma 1, we have $\sum_{h_i \in \mathrm{Rel}(f)} \mathrm{err}_i^- \leq (r-1)/2$. Thus, among such relatively good h_i, there must be some h_i such that $\mathrm{err}_i^- \leq ((r-1)/2) \div (k/2) = (r-1)/k$. This clearly satisfies the lemma with $a = 1$.

Consider next the case that we do not have so many relatively good hypotheses. For each h_i, let $w_i = (1/2 - 1/(2k)) - \mathrm{error}(h_i) < 1/2$. Here note that some h_i may have negative w_i; for example, a hypothesis h_i that is not "relatively good" has weight $w_i < -1/(4k)$. Also note that the total, i.e., $\sum_{i=1}^k w_i$ must be at least 0, because $\mathrm{error}(h_i)$ is at most $1/2 - 1/(2k)$ on average (see the proof of Lemma 1).

Consider the set H of hypotheses h_i that have nonnegative weights, and let s be the number of such hypotheses. Since more than $k/2$ hypotheses are not "relatively good", the total weights of all negative w_i's is less than $-(1/(4k))(k/2) = -1/8$. Thus, letting $\sigma = \sum_{h_i \in H} w_i$, we have $\sigma \geq 1/8$. Now we use Lemma A in Appendix with these w_i's, s, σ, and $t = 8k$. Then from the lemma, we have either (i) $s \geq k/2$, or (ii) there exists some $a \leq 4k$ for which the following holds.

$$\left| \left\{ i \,:\, w_i \geq \frac{a}{8k} \right\} \right| \geq \frac{k}{4a^2}.$$

Note, however, that the case (i) cannot happen because we assume that there are not such many good hypotheses. Hence consider the case (ii). For the parameter a for which the above bound holds, consider hypotheses h_i's such that $w_i \geq a/8k$. Let H' be the set of such hypotheses. Clearly, any h_i in H' satisfies (2). Again since H' has $s \geq k/(4a^2)$ hypotheses and the sum of their err_i^-'s, i.e., $\sum_{h_i \in H'} \mathrm{err}_i^-$, is still at most $(r-1)/2$, there must be some hypothesis in H' such that $\mathrm{err}_i^- \leq 2a^2(r-1)/k$, satisfying the lemma. □

Now we are ready to prove our theorem.

Proof of Theorem 2. Let err^-, $\mathrm{err}|^-$, err^+, $\mathrm{err}|^+$, and err denote respectively $\mathrm{err}_D^-(h)$, $\mathrm{err}|_D^-(h)$, $\mathrm{err}_D^+(h)$, $\mathrm{err}|_D^+(h)$ and $\mathrm{error}_D(h)$ for the hypothesis h selected at the tth iteration under the distribution D $(= D_t)$. First we restate (1) as follows (see, e.g., [Asl00]).

$$Z \;=\; 2\left\{ \sum_{i:h(x_i)=+1} D(i)\sqrt{(1-\mathrm{err}|^-)\mathrm{err}|^-} + \sum_{i:h(x_i)=-1} D(i)\sqrt{(1-\mathrm{err}|^+)\mathrm{err}|^+} \right\}.$$

By using the fact that D is balanced, it is easy to restate it further as

$$Z \;=\; 2\sqrt{\left(\frac{1}{2}-\mathrm{err}^+\right)\mathrm{err}^-} + 2\sqrt{\left(\frac{1}{2}-\mathrm{err}^-\right)\mathrm{err}^+}.$$

We consider two cases.

(Case 1: $\mathcal{H}$ has a quite good hypothesis)
Suppose that $\mathcal{H}$ has a good hypothesis so that we have $\mathrm{err} < 1/2 - 1/(16\sqrt{rk})$ for the selected h. Recall that $\mathrm{err} = \mathrm{err}^+ + \mathrm{err}^+$; then it is easy to see that, for any fixed value of err, Z takes the maximum when $\mathrm{err}^- = \mathrm{err}^+ = \mathrm{err}/2$. Thus, considering this situation, we have

$$Z \;\le\; 2\sqrt{\mathrm{err}\,(1-\mathrm{err})} \;\le\; \sqrt{1-\frac{1}{64rk}}.$$

(Case 2: $\mathcal{H}$ does not have such good hypothesis)
That is, suppose that $\mathcal{H}$ has no hypothesis whose error is less than $1/2 - 1/(16\sqrt{rk})$. Together with this assumption, it follows from Lemma 2 we have some hypothesis $h \in \mathcal{H}$ satisfying the conditions (2) and (3) for some integer a, $1 \le a \le (1/2)\sqrt{(k/r)}$. Then by Lemma B given in Appendix, an upper bound of Z is obtained by assuming the equalities in the conditions (2) and (3). Note that the constraints of Lemma B are satisfied for our choice of a. Thus, for any $r > 1$, we have[1]

$$\begin{aligned}
Z \;&=\; 2\sqrt{\left(\frac{1}{2}-\mathrm{err}^-\right)\mathrm{err}^+} + 2\sqrt{\mathrm{err}^-\left(\frac{1}{2}-\mathrm{err}^+\right)} \\
&=\; 2\sqrt{\left(\sqrt{\left(\frac{1}{2}-\mathrm{err}^-\right)\mathrm{err}^+} + \sqrt{\mathrm{err}^-\left(\frac{1}{2}-\mathrm{err}^+\right)}\right)^2} \\
&=\; 2\sqrt{\frac{\mathrm{err}^- + \mathrm{err}^+}{2} - 2\mathrm{err}^-\mathrm{err}^+ + 2\sqrt{\mathrm{err}^-\mathrm{err}^+\left(\frac{1}{2}-\mathrm{err}^-\right)\left(\frac{1}{2}-\mathrm{err}^+\right)}} \\
&=\; 2\sqrt{\frac{\mathrm{error}}{2} - 2\mathrm{err}^-\mathrm{err}^+ + 2\sqrt{\mathrm{err}^-\mathrm{err}^+\left(\frac{1}{4}-\frac{\mathrm{err}^- + \mathrm{err}^+}{2} + \mathrm{err}^-\mathrm{err}^+\right)}} \\
&\le\; \sqrt{1-\frac{a}{4k} - 8\mathrm{err}^-\mathrm{err}^+ + 8\sqrt{\mathrm{err}^-\mathrm{err}^+\left(\frac{a}{16k} + \mathrm{err}^-\mathrm{err}^+\right)}} \\
&=\; \sqrt{1-\frac{a}{4k} - 8\mathrm{err}^-\mathrm{err}^+ + 8\mathrm{err}^-\mathrm{err}^+\sqrt{\left(\frac{a}{16k\mathrm{err}^-\mathrm{err}^+} + 1\right)}}
\end{aligned}$$

[1]Here we explain the case $r > 1$. The case $r = 1$ can be treated by much simpler analysis, which can be found in [HW04].

Since it holds by using the Tayler series that $\sqrt{1+x} \leq 1 + \frac{x}{2} - \frac{x^2}{16}$ for x $(0 \leq x < 1)$, we have

$$\begin{aligned} Z &\leq \sqrt{1 - \frac{a}{4k} - 8\mathrm{err}^-\mathrm{err}^+ + 8\mathrm{err}^-\mathrm{err}^+\left(1 + \frac{1}{2}\cdot\frac{a}{16k\mathrm{err}^-\mathrm{err}^+} - \frac{1}{16}\frac{a^2}{(16k\mathrm{err}^-\mathrm{err}^+)^2}\right)} \\ &= \sqrt{1 - \frac{a^2}{512k^2\mathrm{err}^-\mathrm{err}^+}}. \end{aligned}$$

Then by substituting $\mathrm{err}^- = \frac{1}{2} - \frac{a}{8k} - \frac{2a^2(r-1)}{k} = \frac{4k - a - 16a^2(r-1)}{8k}$ and $\mathrm{err}^+ = \frac{2a^2(r-1)}{k}$, we obtain

$$\begin{aligned} Z &\leq \sqrt{1 - \frac{4}{512(r-1)(4k-a-16(r-1)a^2)}} \\ &< \sqrt{1 - \frac{1}{512kr}}. \end{aligned}$$

□

Finally we state the following corollary that is derived from Theorem 1 and Theorem 2 by a standard analysis.

Corollary 3. For any r and k, $1 \leq r \leq k/2$ and $2 \leq k \leq n$, and for any target r-of-k function f, the boosting algorithm given in Figure 2 yields a hypothesis that is consistent with f if $T \geq (1024kr\log m + 1)$.

4 Concluding Remarks

We proved that for r-of-k functions, InfoBoost can find consistent hypotheses in $O(rk\log m)$ iterations (where m is the number of examples). On the other hand, AdaBoost seems to need $O(k^2\log m)$ iterations (which has been indeed proved conditionally in [HW04]); thus, our result indicates that InfoBoost performs better than AdaBoost when we may assume less error on one classification side. A problem of theoretical interest is to show the tightness of our analysis. (We conjecture that our upper bound is tight except for the constant factor.)

Another interesting open problem on InfoBoost, which is also important in practice, is its generalization error. We are still unaware of the way to bound the generalization error of InfoBoost. Recall that Schapire et al. [SFBL98] showed that "margin" can be used to estimate the generalization error of a hypothesis defined as a linear combination of hypotheses. While this analysis is applicable for hypotheses produced by InfoBoost, it may be the case that InfoBoost produces a hypothesis with "small margin", which is not enough to guarantee a small generalization error (see [SS99] for the detail). One way to avoid arguments on generalization error is to introduce the boosting by filtering framework [Sch91, Fre95]. In this framework, instead of being given a set of examples in advance, the learner can directly draw examples randomly according to a fixed unknown distribution. Several researchers pointed out that AdaBoost does not fit this scheme because it constructs too "skew" distributions from which it is hard to sample examples [DW00, BG01, Gav03]. This also accounts for InfoBoost. They proposed variants of AdaBoost that create "smooth" distributions enabling us to sample efficiently. One of the interesting problems is to extend some of these ideas to InfoBoost.

Acknowledgements

We thank Manfred Warmuth and Phil Long for helpful discussions and leading us to study our problem.

References

[Asl00] J.A. Aslam, Improving algorithms for boosting, in *Proc. 13th Annu. Conference on Comput. Learning Theory*, ACM, 200–207, 2000.

[BG01] N.H. Bshouty and D. Gavinsky, On boosting with optimal poly-bounded distributions, in *Proc. 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, Lecture Notes in Artificial Intelligence 2111, 490–506, 2001.

[DL03] S. Dasgupta and P.M. Long, Boosting with diverse base classifiers, in *Learning Theory and Kernel Machines, 16th Annual Conference on Learning Theory and 7th Kernel Workshop*, Lecture Notes in Artificial Intelligence 2777, Springer, 273–287, 2003.

[DW00] C. Domingo and O. Watanabe, MadaBoost: a modification of AdaBoost, in *Proc. 13th Annu. Conference on Computational Learning Theory*, ACM, 180–189, 2000.

[Fre95] Y. Freund, Boosting a weak learning algorithm by majority, *Inform. Comput.*, 121(2), 256–285, 1995.

[FS97] Y. Freund and R. E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *J. Comput. Syst. Sci.*, 55(1), 119–139, 1997.

[Gav03] D. Gavinsky, Optimally-smooth adaptive boosting and application to agnostic learning, *Journal of Machine Learning Research* 4, 101-117, 2003.

[HW04] K. Hatano and M.K. Warmuth. Boosting versus covering, in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. Saul, and B. Schölkopf eds.), MIT Press, 2004.

[Nat91] B.K. Natarajan. *Machine Learning: A Theoretical Approach*, Morgan Kaufmann, 1991.

[Sch91] R.E. Schapire, The strength of weak learnability, *Machine Learning* 5(2), 197–227, 1990.

[SS99] R.E. Schapire and Y. Singer, Improved boosting algorithms using confidence-rated predictions, *Machine Learning* 37(3), 297–336, 1999.

[SFBL98] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, *The Annals of Statistics* 26(5), 1651-1686, 1998.

Appendix

Lemma A. Let $\{w_1, ..., w_s\}$ be a set of weights such that $0 \le w_i \le 1/2$ for all i, $1 \le i \le s$, and $\sum_{i=1}^{s} w_i = \sigma$. Then for any $t > 0$, either $s \ge \sigma t/2$, or there exists some a $(1 \le a \le t/2)$, such that the following holds.

$$\left|\left\{ i : w_i \ge \frac{a}{t} \right\}\right| \;\ge\; \frac{d_0 \sigma t}{a^2}.$$

Where we may set $d_0 = 1/4$.

Remark. Note that the average weight $(\sum w_i)/s = \sigma/s$. Then the Markov inequality claims that $|\{w_i \ge a/t\}|$ $(= |\{w_i \ge \alpha(\sigma/s)\}| \le s(1/\alpha) =)$ $\sigma t/a$ for all $a > 0$, where $\alpha = sa/(\sigma t)$. Thus, the lemma is regarded as a lower bound version of the Markov inequality. We can improve this bound by replacing the denominator a^2 with smaller one, e.g., $a(\log a)^2$; but on the other hand, it is also easy to see that we cannot use the denominator a like the Markov inequality.

Proof. Assume that the lemma does not hold (i.e., neither $s \ge \sigma t/2$ nor some a exists satisfying the lemma), and we will lead a contradiction. For any integer $j \ge 0$, define $n(j) = d_0\sigma t/(2^{2j})$. We may assume that $w_1 \le w_2 \le \cdots \le w_s$. Then we have $w_i < 1/k$ for the first $s_0 \ge s - n(0)$ weights, because otherwise, we would have more than $n(0)$ weights that are larger than $1/k$, contradicting our assumption. Similarly, we have $w_i < 2/t$ for the next $s_1 \ge s - s_0 - n(1)$ weights, $w_i < 2^2/t$ for the next $s_2 \ge s - s_0 - s_1 - n(2)$ weights, and so on. It is easy to see that $\sum w_i$ becomes largest if $s_0 = n(0)$, $s_1 = s - s_0 - n(1)$, ... Thus, estimate $\sum w_i$ with $s_0 = n(0)$, $s_1 = s - s_0 - n(1) = s - (s - n(0)) - n(1) = n(0) - n(1)$, $s_2 = s - s_0 - s_1 - n(2) = n(1) - n(2)$, ... Then we have

$$\begin{aligned}
\sum_{i=1}^{s} w_i &< \frac{1}{t}\cdot(s - n(0)) + \frac{2}{t}\cdot(n(0) - n(1)) + \frac{2^2}{t}\cdot(n(1) - n(2)) + \frac{2^3}{t}\cdot(n(2) - n(3)) + \cdots \\
&= \frac{1}{t}\left(s + n(0) + 2n(1) + 2^2 n(2) + 2^3 n(3) \cdots\right) \\
&= \frac{s}{t} + (d_0\sigma)\left(\frac{1}{1} + \frac{2}{2^2} + \frac{2^2}{2^4} + \frac{2^3}{2^6} + \cdots\right) \\
&= \frac{s}{t} + (d_0\sigma)\left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots\right),
\end{aligned}$$

which is, from our assumption on s and the choice of d_0, less than σ. A contradiction. □

Lemma B. Consider the following optimization problem.

$$\text{Maximize} \quad V(b,c) \;=\; \sqrt{b\left(\frac{1}{2} - c\right)} + \sqrt{\left(\frac{1}{2} - b\right)c}$$

subject to

$$\begin{cases} b + c \le \dfrac{1}{2} - \gamma, \\ c \le \dfrac{1}{4} - \dfrac{\gamma}{2} - \delta, \\ b, c \ge 0, \end{cases}$$

for given γ and δ such that $\gamma > 0$ and $0 < \delta \le (1/2 - \gamma)/2 = 1/4 - \gamma/2$. Then the maximum is obtained if $b + c = 1/2 - \gamma$ and $c = 1/4 - \gamma/2 - \delta$.

Proof. For any β, $0 < \beta \le 1/2 - \gamma$, define $Z(\beta)$ by

$$Z(\beta) \;=\; \max_{b,c} V(b,c) \;=\; \max_{b,c} \sqrt{b\left(\frac{1}{2} - c\right)} + \sqrt{\left(\frac{1}{2} - b\right) c},$$

where $b, c \ge 0$ are chosen so that $b + c = \beta$, and $c \le 1/4 - \gamma/2 - \delta$. We consider the following two cases.

(Case 1: $\beta \le 1/2 - \gamma - 2\delta$)
In this case, since $\beta/2 \le 1/4 - \gamma/2 - \delta$, it is possible to choose $b = c = \beta/2$, which maximize $V(b,c) = \sqrt{(1-\beta)\beta}$. Hence, $Z(\beta) = \sqrt{(1-\beta)\beta}$. Since this $Z(\beta)$ is monotone increasing, its maximum Z_1 is obtained with $\beta = 1/2 - \gamma$, and we have

$$Z_1 \;=\; \sqrt{\frac{1}{4} - (r + 2\delta)^2}.$$

(Case 2: $\beta > 1/2 - \gamma - 2\delta$)
Since $V(\beta - c, c)$ is monotone increasing for $c \le 1/4 - \gamma/2 - \delta$, it is maximized when $c = 1/4 - \gamma/2 - \delta$; hence, we have

$$Z(\beta) \;=\; \sqrt{\left(\beta - \frac{1}{4} + \frac{\gamma}{2} + \delta\right)\left(\frac{1}{4} + \frac{\gamma}{2} + \delta\right)} + \sqrt{\left(\frac{3}{4} - \beta - \frac{\gamma}{2} - \delta\right)\left(\frac{1}{4} - \frac{\gamma}{2} - \delta\right)}.$$

Again this $Z(\beta)$ is monotone increasing, we have the maximum

$$Z_2 \;=\; \sqrt{\left(\frac{1}{4} + \delta\right)^2 - \frac{\gamma^2}{4}} + \sqrt{\left(\frac{1}{4} - \delta\right)^2 - \frac{\gamma^2}{4}},$$

when $\beta = 1/2 - \gamma$ (i.e., $b + c = 1/2 - \gamma$ and $c = 1/4 - \gamma/2 - \delta$).

Now we show that $Z_1 \le Z_2$ for any δ such that $0 < \delta < 1/4 - \gamma/2$. Let $g(x) = \sqrt{x}$ and $\Delta = 1/16 - (\delta + \gamma/2)^2$. Then we have

$$Z_2 - Z_1 = g(\Delta + p) + g(\Delta - q) - 2g(\Delta), \tag{4}$$

where we let $p = 2\delta^2 + \delta\gamma + \delta/2$, and $q = -2\delta^2 - \delta\gamma + \delta/2$. Note that $p, q > 0$ by the definition of δ and γ. Since the function $g(x)$ is concave and monotone increasing, it follows that $g(\Delta + p) \ge g(\Delta) + g'(\Delta + p)p$ and $g(\Delta - q) \ge g(\Delta) - g'(\Delta - q)q$. Therefore the quantity (4) is bounded from below by

$$\begin{aligned}
& g'(\Delta + p)p - g'(\Delta - q)q \\
= & \frac{2\delta^2 + \delta\gamma + \frac{\delta}{2}}{2\sqrt{\left(\frac{1}{4} + \delta\right)^2 - \frac{\gamma^2}{4}}} - \frac{-2\delta^2 - \delta\gamma + \frac{\delta}{2}}{2\sqrt{\left(\frac{1}{4} - \delta\right)^2 - \frac{\gamma^2}{4}}} \\
= & \frac{2\delta\left(\delta + \gamma + \frac{1}{2}\right)\sqrt{\left(\frac{1}{4} - \delta\right)^2 - \frac{\gamma^2}{4}}}{2\sqrt{\left(\frac{1}{4} + \delta\right)^2 - \frac{\gamma^2}{4}}\sqrt{\left(\frac{1}{4} - \delta\right)^2 - \frac{\gamma^2}{4}}} - \frac{\delta(-2\delta - \gamma + \frac{1}{2})\sqrt{\left(\frac{1}{4} + \delta\right)^2 - \frac{\gamma^2}{4}}}{2\sqrt{\left(\frac{1}{4} + \delta\right)^2 - \frac{\gamma^2}{4}}\sqrt{\left(\frac{1}{4} - \delta\right)^2 - \frac{\gamma^2}{4}}}.
\end{aligned}$$

In order to show that $Z_2 - Z_1 \geq 0$, it is sufficent to prove that

$$\left(\delta+\gamma+\frac{1}{2}\right)^2\left\{\left(\frac{1}{4}-\delta\right)^2-\frac{\gamma^2}{4}\right\}-\left(-\delta-\gamma+\frac{1}{2}\right)^2\left\{\left(\frac{1}{4}+\delta\right)^2-\frac{\gamma^2}{4}\right\}$$

is positive. Let us denote the quantity above as $\widetilde{\Delta}$. Then we get

$$\begin{aligned}
\widetilde{\Delta} &= \left\{(2\delta+\gamma)^2+(2\delta+\gamma)+\frac{1}{4}\right\}\left\{\left(\frac{1}{16}+\delta^2-\frac{\gamma^2}{4}\right)-\frac{\delta}{2}\right\} \\
&- \left\{(2\delta+\gamma)^2-(2\delta+\gamma)+\frac{1}{4}\right\}\left\{\left(\frac{1}{16}+\delta^2-\frac{\gamma^2}{4}\right)+\frac{\delta}{2}\right\} \\
&= 2(2\delta+\gamma)\left(\frac{1}{16}+\delta^2-\frac{\gamma^2}{4}\right)-2\cdot\frac{\delta}{2}\left\{(2\delta+\gamma)^2+\frac{1}{4}\right\} \\
&= \frac{\gamma^2}{8}-2\delta^2\gamma-2\delta\gamma^2-\frac{\gamma^3}{2} \\
&= \frac{\gamma}{2}\left\{\left(\frac{1}{2}\right)^2-(2\delta+\gamma)^2\right\} > 0,
\end{aligned}$$

as desired. $\square$