Potential-Based Agnostic Boosting

Adam Tauman Kalai
Microsoft Research
adum@microsoft.com

Varun Kanade
Harvard University
vkanade@fas.harvard.edu

Abstract

We prove strong noise-tolerance properties of a potential-based boosting algorithm, similar to MadaBoost (Domingo and Watanabe, 2000) and SmoothBoost (Servedio, 2003). Our analysis is in the agnostic framework of Kearns, Schapire and Sellie (1994), giving polynomial-time guarantees in presence of arbitrary noise. A remarkable feature of our algorithm is that it can be implemented without reweighting examples, by randomly relabeling them instead. Our boosting theorem gives, as easy corollaries, alternative derivations of two recent nontrivial results in computational learning theory: agnostically learning decision trees (Gopalan et al, 2008) and agnostically learning halfspaces (Kalai et al, 2005). Experiments suggest that the algorithm performs similarly to MadaBoost.

1 Introduction

Boosting procedures attempt to improve the accuracy of general machine learning algorithms, through repeated executions on reweighted data. Aggressive reweighting of data may lead to poor performance in the presence of certain types of noise [1]. This has been addressed by a number of “robust” boosting algorithms, such as SmoothBoost [2, 3] and MadaBoost [4] as well as boosting by branching programs [5, 6]. Some of these algorithms are potential-based boosters, i.e., natural variants on AdaBoost [7], while others are perhaps less practical but have stronger theoretical guarantees in the presence of noise.

The present work gives a simple potential-based boosting algorithm with guarantees in the (arbitrary noise) agnostic learning setting [8, 9]. A unique feature of our algorithm, illustrated in Figure 1, is that it does not alter the distribution on unlabeled examples but rather it alters the labels. This enables us to prove a strong boosting theorem in which the weak learner need only succeed for one distribution on unlabeled examples. To the best of our knowledge, earlier weak-to-strong boosting theorems have always relied on the ability of the weak learner to succeed under arbitrary distributions. The utility of our boosting theorem is demonstrated by re-deriving two non-trivial results in computational learning theory, namely agnostically learning decision trees [10] and agnostically learning halfspaces [11], which were previously solved using very different techniques.

The main contributions of this paper are, first, giving the first provably noise-tolerant analysis of a potential-based boosting algorithm, and, second, giving a distribution-specific boosting theorem that does not require the weak learner to learn over all distributions on \( x \in X \). This is in contrast to recent work by Long and Servedio, showing that convex potential boosters cannot work in the presence of random classification noise [12]. The present algorithm circumvents that impossibility result in two ways. First, the algorithm has the possibility of negating the current hypothesis and hence is not technically a standard potential-based boosting algorithm. Second, weak agnostic learning is more challenging than weak learning with random classification noise, in the sense that an algorithm which is a weak-learner in the random classification noise setting need not be a weak-learner in the agnostic setting.

Related work. There is a substantial literature on robust boosting algorithms, including algorithms already mentioned, MadaBoost, SmoothBoost, as well as LogitBoost [13], BrownBoost [14], Nad-
1.1 Preliminaries

In the agnostic setting, we consider learning with respect to a distribution over \( X \times Y \). For simplicity, we will take \( X \) to be finite or countable and \( Y = \{ -1, 1 \} \). Formally, learning is with respect to some class of functions, \( C \), where each \( c \in C \) is a binary classifier \( c : X \to \{ -1, 1 \} \). There is an arbitrary distribution \( \mu \) over \( X \) and an arbitrary target function \( f : X \to [-1, 1] \). Together these determine an arbitrary joint distribution \( D = \langle \mu, f \rangle \) over \( X \times \{ -1, 1 \} \) where \( D(x, y) = \mu(x) \frac{1 + yf(x)}{2} \), i.e., \( f(x) = \mathbb{E}_D[y|x] \). The error and correlation\(^1\) of a classifier \( h : X \to \{ -1, 1 \} \) with respect to \( D \), are

\[^{1}\text{This quantity is typically referred to as edge in the boosting literature. However, } \text{corr}(h, D) = 2 \text{ edge}(h, D) \text{ according to the standard notation, hence we use the notation corr.}\]
for any $f$.

Since our boosting procedure does not change the distribution over unlabeled examples, it offers setting where $x_1$, may make $q(x,y) \sim D$ sample from $h$.

Definition 1. A learning algorithm is a $f$ since it has to succeed in the agnostic model (where no assumption is made on realistic as it does not assume noiseless data. Finding such a weak learner may be quite challenging examples (the algorithm may only have guarantees for certain distributions), and (b) it is more.

We now define the distribution $D$ to a $\gamma$-optimal hypotheses to one which outputs a nearly $1$-optimal hypothesis, even for small

$\gamma$-weakly accurate classifier [23] for PAC (noiseless) learning is simply one whose correlation is $\gamma$ (for some $\gamma \in (0, 1)$). A different definition of weakly accurate classifier is appropriate in the agnostic setting. Namely, for some $\gamma \in (0, 1)$, $h : X \rightarrow \{-1, 1\}$ is said to be $\gamma$-optimal for $C$ (and $D$) if,

$\text{cor}(h, D) \geq \gamma \text{cor}(C, D)$

Hence, if the labels are totally random then a weak hypothesis need not have any correlation over random guessing. On the other hand, in a noiseless setting, where $\text{cor}(C) = 1$, this is equivalent to a $\gamma$-weakly accurate hypothesis. The goal is to boost from an algorithm capable of outputting $\gamma$-optimal hypotheses to one which outputs a nearly $1$-optimal hypothesis, even for small $\gamma$.

Let $D$ be a distribution over $X \times \{-1, 1\}$. Let $w : X \times \{-1, 1\} \rightarrow [0, 1]$ be a weighting function.

We now define the distribution $D$ relabeled by $w$, $R_{D,w}$. Procedurally, one can think of generating a sample from $R_{D,w}$ by drawing an example $(x, y)$ from $D$, then with probability $w(x, y)$, outputting $(x, y)$ directly, and with probability $1 - w(x, y)$, outputting $(x, y')$ where $y'$ is uniformly random in $\{-1, 1\}$. Formally,

$R_{D,w}(x, y) = D(x, y) \left( w(x, y) + \frac{1 - w(x, y)}{2} \right) + D(x, -y) \left( \frac{1 - w(x, -y)}{2} \right)$

Note that $D$ and $R_{D,w}$ have the same marginal distributions over unlabeled examples $x \in X$. Also, observe that, for any $D$, $w$, and $h : X \rightarrow \mathbb{R}$,

$\mathbb{E}_{(x, y) \sim R_{D,w}} [h(x)y] = \mathbb{E}_{(x, y) \sim D} [h(x)yw(x, y)]$ (1)

This can be seen by the procedural interpretation above. When $(x, y)$ is returned directly, which happens with probability $w(x, y)$, we get a contribution of $h(x)y$, but $E[h(x)y'] = 0$ for uniform $y' \in \{-1, 1\}$.

It is possible to describe traditional supervised learning and active (query) learning in the same framework. A general $(m, q)$-learning algorithm is given $m$ unlabeled examples $(x_1, \ldots, x_m)$, and may make $q$ label queries to a query oracle $L : X \rightarrow \{-1, 1\}$, and it outputs a classifier $h : X \rightarrow \{-1, 1\}$. The queries may be active, meaning that queries may only be made to training examples $x$, or membership queries meaning that arbitrary examples $x \in X$ may be queried. The active query setting where $q = m$ is the standard supervised learning setting where all $m$ labels may be queried. One can similarly model semi-supervised learning.

Since our boosting procedure does not change the distribution over unlabeled examples, it offers two advantages: (1) Agnostic weak learning may be defined with respect to a single distribution $\mu$ over unlabeled examples, and (2) The weak learning algorithms may be active (or use membership queries). In particular, the agnostic weak learning hypothesis for $C$ and $\mu$ is that for any $f : X \rightarrow [-1, 1]$, given examples from $D = \langle \mu, f \rangle$, the learner will output a $\gamma$-optimal classifier for $C$. The advantages of this new definition are: (a) it is not with respect to every distribution on unlabeled examples (the algorithm may only have guarantees for certain distributions), and (b) it is more realistic as it does not assume noiseless data. Finding such a weak learner may be quite challenging since it has to succeed in the agnostic model (where no assumption is made on $f$), however it may be a bit easier in the sense that the learning algorithm need only handle one particular $\mu$.

Definition 1. A learning algorithm is a $(\gamma, \epsilon_0, \delta)$ agnostic weak learner for $C$ and $\mu$ over $X$ if, for any $f : X \rightarrow [-1, 1]$, with probability $\geq 1 - \delta$ over its random input, the algorithm outputs $h : X \rightarrow [-1, 1]$ such that, if $D = \langle \mu, f \rangle$,

$\text{cor}(h, D) = \mathbb{E}_{x \sim \mu} [h(x)f(x)] \geq \gamma \left( \sup_{c \in C} \mathbb{E}_{x \sim \mu} [c(x)f(x)] \right) - \epsilon_0$
Lastly, we define $\text{sign}(z)$ to be 1 if $z \geq 0$ and $-1$ if $z < 0$.

## 2 Formal boosting procedure and main results

The formal boosting procedure we analyze is given in Figure 2.

**AGNOSTIC BOOSTER**

Inputs: $\langle x_1, \ldots, x_{Tm+q} \rangle$, $T, s \geq 1$, label oracle $L : X \rightarrow \{-1, 1\}$, $(m, q)$-learner $W$.

Output: classifier $h : X \rightarrow \{-1, 1\}$.

1. Let $H^0 = 0$
2. Query the labels of the first $s$ examples to get $y_1 = L(x_1), \ldots, y_s = L(x_s)$.
3. For $t = 1, \ldots, T$:
   a) Define $w^t(x, y) = -\phi'(H^{t-1}(x)y) = \min\{1, \exp(-H^{t-1}(x)y)\}$
      
      Define $L^t : X \rightarrow \{-1, 1\}$ by: 
      i) On input $x \in X$, let $y = L(x)$. 
      ii) With probability $w^t(x, y)$, return $y$. 
      iii) Otherwise return $-1$ or $1$ with equal probability.
   b) Let $g^t = W(\langle x_{s+(t-1)m+1}, \ldots, x_{s+tm} \rangle, L^t)$
   c) Let 
      i) $\alpha^t = \frac{1}{s} \sum_{i=1}^{s} g^t(x_i) w^t(x_i, y_i)$ 
      ii) $\beta^t = \frac{1}{s} \sum_{i=1}^{s} -\text{sign}(H^{t-1}(x_i)) w^t(x_i, y_i)$ 
   d) If $\alpha^t \geq \beta^t$,
      
      $h^t = g^t$; 
      $\gamma^t = \alpha^t$;
      
      Else,
      
      $h^t = -\text{sign}(H^{t-1})$; 
      $\gamma^t = \beta^t$;
   e) $H^t(x) = H^{t-1}(x) + \gamma^t h^t(x)$
4. Output $h = \text{sign}(H^T)$ where $\tau$ is chosen so as to minimize empirical error on $\langle (x_1, y_1), \ldots, (x_s, y_s) \rangle$

Figure 2: Formal Boosting by Relabeling Procedure.

**Theorem 1.** If $W$ is a $(\gamma, \epsilon_0, \delta)$ weak learner with respect to $C$ and $\mu$, $s = \frac{200}{\sqrt{\gamma \epsilon_0}} \log \left( \frac{1}{\delta} \right)$, $T = \frac{29}{\sqrt{\gamma \epsilon_0}}$, Algorithm AGNOSTIC BOOSTER (Figure 2) with probability at least $1 - 4\delta T$ outputs a hypothesis $h$ satisfying:

$$\text{cor}(h, D) \geq \text{cor}(C, D) - \frac{\epsilon_0}{\gamma} - \epsilon$$

Recall that $\epsilon_0$ is intended to be very small, e.g., $O(m^{-1/2})$. Also note that the number of calls to the query oracle $L$ is $s$ plus $T$ times the number of calls made by the weak learner (if the weak learner is active, then so is the boosting algorithm). We show that two recent non-trivial results, viz. agnostically learning decision trees and agnostically learning halfspaces follow as corollaries to Theorem 1. The two results are stated below:

**Theorem 2** ([10]). Let $C$ be the class of binary decision trees on $\{-1, 1\}^n$ with at most $t$ leaves, and let $U$ be the uniform distribution on $\{-1, 1\}^n$. There exists an algorithm that when given $t, n, \epsilon, \delta > 0$, and a label oracle for an arbitrary $f : \{-1, 1\}^n \rightarrow [-1, 1]$, makes $q = \text{poly}(nt/(\epsilon \delta))$ membership queries and, with probability $\geq 1 - \delta$, outputs $h : \{-1, 1\}^n \rightarrow \{-1, 1\}$ such that for $U_f = \langle U, f \rangle$, $\text{err}(h, U_f) \leq \text{err}(C, U_f) + \epsilon$. 

4
Theorem 3 ([11]). For any fixed $\epsilon > 0$, there exists a univariate polynomial $p$ such that the following holds: Let $n \geq 1$, $C$ be the class of halfspaces in $n$ dimensions, let $\mathcal{U}$ be the uniform distribution on $\{-1,1\}^n$, and $f : \{-1,1\}^n \rightarrow [-1,1]$ be an arbitrary function. There exists a polynomial-time algorithm that, when given $m = p(n \log(1/\delta))$ labeled examples from $\mathcal{U}_f = (\mathcal{U}, f)$, outputs a classifier $h : \{-1,1\}^n \rightarrow \{-1,1\}$ such that $\text{err}(h, \mathcal{U}_f) \leq \text{err}(C, \mathcal{U}_f) + \epsilon$. (The algorithm makes no queries.)

Note that a related theorem was shown for halfspaces over log-concave distributions over $X = \mathbb{R}^n$. The boosting approach here similarly generalizes to that case in a straightforward manner. This illustrates how, from the point of view of designing provably efficient agnostic learning algorithms, the current boosting procedure may be useful.

3 Analysis of Boosting Algorithm

This section is devoted to the analysis of algorithm AGNOSTIC BOOSTER (see Fig 2). As is standard, the boosting algorithm can be viewed as minimizing a convex potential function. However, the proof is significantly different than the analysis of AdaBoost [7], where they simply use the fact that the boosting algorithm can be viewed as minimizing a convex potential function. However, the proof (see Fig 2). As is standard, the boosting algorithm can be viewed as minimizing a convex potential function. However, the proof is significantly different than the analysis of AdaBoost [7], where they simply use the fact that the potential is an upper-bound on the error rate.

Our analysis has two parts. First, we define a conservative relabeling, such as the one we use, to be one which never relabels/downweights examples that the booster currently misclassifies. We show that for a conservative reweighting, either the weak learner will make progress, returning a hypothesis correlated with the relabeled distribution or $-\text{sign}(H'^{-1})$ will be correlated with the relabeled distribution.

Second, if we find a hypothesis correlated with the relabeled distribution, then the potential on round $t$ will be noticeably lower than that of round $t - 1$. This is essentially a simple gradient descent analysis, using a bound on the second derivative of the potential. Since the potential is between 0 and 1, it can only drop so many rounds. This implies that $\text{sign}(H'^t)$ must be a near-optimal classifier for some $t$ (though the only sure way we have of knowing which one to pick is by testing accuracy on held-out data).

The potential function we consider, as in MadaBoost, is defined by $\phi : \mathbb{R} \rightarrow \mathbb{R}$,

$$\phi(z) = \begin{cases} 1 - z & \text{if } z \leq 0 \\ e^{-z} & \text{if } z > 0 \end{cases}$$

Define the potential of a (real-valued) hypothesis $H$ with respect to a distribution $\mathcal{D}$ over $X \times \{-1,1\}$ as:

$$\Phi(H, \mathcal{D}) = \mathbb{E}_{(x, y) \sim \mathcal{D}} [\phi(yH(x))]$$

(2)

Note that $\Phi(H^0, \mathcal{D}) = \Phi(0, \mathcal{D}) = 1$. We will show that the potential decreases every round of the algorithm. Notice that the weights in the boosting algorithm correspond to the derivative of the potential, because $-\phi'(z) = \min\{1, \exp(-z)\} \in [0, 1]$. In other words, the weak learning step is essentially a gradient descent step.

We next state a key fact about agnostic learning in Lemma 1.

Definition 2. Let $h : X \rightarrow \{-1,1\}$ be a hypothesis. Then weighting function $w : X \times \{-1,1\} \rightarrow [0,1]$ is called conservative for $h$ if $w(x, -h(x)) = 1$ for all $x \in X$.

Note that, if the hypothesis is $\text{sign}(H'(x))$, then a weighting function defined by $-\phi'(H'(x)y)$ is conservative if and only if $\phi'(z) = -1$ for all $z < 0$. We first show that relabeling according to a conservative weighting function is good in the sense that, if $h$ is far from optimal according to the original distribution, then after relabeling by $w$ it is even further from optimal.

Lemma 1. For any distribution $\mathcal{D}$ over $X \times \{-1,1\}$, classifiers $c, h : X \rightarrow \{-1,1\}$, and any weighting function $w : X \times \{-1,1\} \rightarrow [0,1]$ conservative for $h$,

$$\text{cor}(c, R_{\mathcal{D}, w}) - \text{cor}(h, R_{\mathcal{D}, w}) \geq \text{cor}(c, \mathcal{D}) - \text{cor}(h, \mathcal{D})$$
Thus the above equation implies the lemma.

Finally, consider two cases. In the first case, when \(1 - w(x, y) > 0\), we have \(h(x)y = 1\) while \(c(x)y \leq 1\). The second case is \(1 - w(x, y) = 0\). In either case, \((c(x) - h(x))y(1 - w(x, y)) \leq 0\). Thus the above equation implies the lemma.

We will use Lemma 1 to show that the weak learner will return a useful hypothesis. The case in which the weak learner may not return a useful hypothesis is when \(\text{cor}(C, R_D, w) = 0\), when the optimal classifier on the reweighted distribution has no correlation. This can happen, but in this case it means that either our current hypothesis is close to optimal, or \(h = \text{sign}(H^{-1})\) is even worse than random guessing, and hence we can use its negation as a weak agnostic learner.

We next explain how a \(\gamma\)-optimal classifier on the reweighted distribution decreases the potential. We will use the following property linear approximation of \(\phi\).

**Lemma 2.** For any \(x, \delta \in \mathbb{R}\), \(\phi(x + \delta) - \phi(x) - \phi'(x)\delta \leq \delta^2/2\).

**Proof.** This follows from Taylor’s theorem and the fact the function \(\phi\) is differentiable everywhere, and that the left and right second derivatives exist everywhere and are bounded by 1.

Let \(h^t : X \rightarrow \{-1, 1\}\) be the weak hypothesis that the algorithm finds on round \(t\). This may either be the hypothesis returned by the weak learner \(W\) or \(-\text{sign}(H^{-1})\). The following lemma lower bounds the decrease in potential caused by adding \(\gamma h^t\) to \(H^{-1}\). We will apply the following Lemma on each round of the algorithm to show that the potential decreases on each round, as long as the weak hypothesis \(h^t\) has non-negligible correlation and \(\gamma^t\) is suitably chosen.

**Lemma 3.** Consider any function \(H : X \rightarrow \mathbb{R}\), hypothesis \(h : X \rightarrow [-1, 1], \gamma \in \mathbb{R}\), and distribution \(D\) over \(X \times \{-1, 1\}\). Let \(D' = R_D, w\) be the distribution \(D\) relabeled by \(w(x, y) = -\phi'(yH(x))\). Then,

\[
\Phi(H, D) - \Phi(H + \gamma h, D) \geq \gamma \text{cor}(h, D') - \frac{\gamma^2}{2}
\]

**Proof.** For any \((x, y) \in X \times \{-1, 1\}\), using Lemma 2 we know that:

\[
\phi(H(x)y) - \phi((H(x) + \gamma h(x))y) \geq \gamma h(x)y(-\phi'(H(x)y)) - \frac{\gamma^2}{2}
\]

In the step above we use the fact that \(h(x)^2y^2 \leq 1\). Taking expectation over \((x, y)\) from \(D\),

\[
\Phi(H, D) - \phi(H + \gamma h, D) \geq \mathbb{E}_{(x, y) \sim D} [h(x)y(-\phi'(H(x)y))] - \frac{\gamma^2}{2}
\]

\[
= \mathbb{E}_{(x, y) \sim D'} [h(x)y] - \frac{\gamma^2}{2}
\]

In the above we have used Eq. (1). We are done, by definition of \(\text{cor}(h, D')\).

Using all the above lemmas, we will show that the algorithm AGNOSTIC BOOSTER returns a hypothesis with correlation (or error) close to that of the best classifier from \(C\). We are now ready to prove the main theorem.

**Proof of Theorem 1.** Suppose \(\exists c \in C\) such that \(\text{cor}(c, D) > \text{cor}(\text{sign}(H^{-1}), D) + \frac{\delta_0}{\gamma} + \epsilon\), then applying Lemma 1 to \(H^{-1}\) and setting \(w^t(x, y) = -\phi'(H^{-1}(x)y)\), we get that

\[
\text{cor}(c, R_{D, w^t}) > \text{cor}(\text{sign}(H^{-1}), R_{D, w^t}) + \frac{\epsilon_0}{\gamma} + \epsilon
\]

In this case we want to show that the algorithm successfully finds \(h^t\) with \(\text{cor}(h^t, R_{D, w^t}) \geq \frac{\gamma \epsilon}{3}\).
Let \( g^t \) be the hypothesis returned by the weak learner \( W \). From Step 3c) in the algorithm:

\[
\alpha^t = \frac{1}{s} \sum_{i=1}^{s} g(x_i) w^t(x_i, y_i); \quad \beta^t = \frac{1}{s} \sum_{i=1}^{s} -\text{sign}(H^{t-1})(x_i) w^t(x_i, y_i)
\]

When \( s = \frac{200}{\gamma^2} \log \left( \frac{1}{\delta} \right) \), by Chernoff-Hoeffding bounds we know that \( \alpha^t \) and \( \beta^t \) are within an additive \( \frac{2\epsilon}{\gamma} \) of \( \text{cor}(g^t, R_{D,w^t}) \) and \( \text{cor}(-\text{sign}(H^{t-1}), R_{D,w^t}) \) respectively with probability at least \( 1 - 2\delta \). As defined in Step 3d) in the algorithm, let \( \gamma^t = \max(\alpha^t, \beta^t) \). We allow the algorithm to fail with probability \( 3\delta \) at this stage, possibly caused by the weak-learner and the estimation of \( \alpha^t, \beta^t \).

Consider two cases: First that \( \text{cor}(c, R_{D,w^t}) \geq \frac{3\epsilon}{\gamma} + \frac{\epsilon}{3} \), in this case by the weak learning assumption, \( \text{cor}(g^t, R_{D,w^t}) \geq \frac{2\epsilon}{\gamma} \). In the second case, if this does not hold, then \( \text{cor}(-\text{sign}(H^{t-1}), R_{D,w^t}) \geq \frac{\epsilon}{3} \) using (3). Thus, even after taking into account the fact that the empirical estimates may be off from the true correlations by \( \frac{2\epsilon}{\gamma^2} \), we get that \( \text{cor}(h^t, R_{D,w^t}) \geq \frac{2\epsilon}{\gamma} \) and that \( |\gamma^t - \text{cor}(h^t, R_{D,w^t})| \leq \frac{2\epsilon}{\gamma^2} \).

Using this and Lemma 3, we get that by setting \( H^t = H^{t-1} + \gamma^t h^t \) the potential decreases by at least \( \frac{2\epsilon^2}{\gamma^2} \).

When \( t = 0 \) and \( H^0 = 0 \), \( \Phi(H^0, D) = 1 \). Since for any \( H : X \rightarrow \mathbb{R} \), \( \Phi(H, D) > 0 \); we can have at most \( T = \frac{29}{\gamma^2} \) rounds. This guarantees that when the algorithm is run for \( T \) rounds, on some round \( t \) the hypothesis \( \text{sign}(H^t) \) will have correlation at least \( \sup_{c \in C} \text{cor}(c, D) - \frac{\epsilon_0}{\gamma} - \frac{2\epsilon}{3} \). For \( s = \frac{200}{\gamma^2} \log \left( \frac{1}{c} \right) \) the empirical estimate of the correlation of the constructed hypothesis on each round is within an additive \( \frac{\epsilon}{\gamma} \) of its true correlation, allowing a further failure probability of \( \delta \) each round. Thus the final hypothesis \( H^T \) which has the highest empirical correlation satisfies,

\[
\text{cor}(H^T, D) \geq \sup_{c \in C} \text{cor}(c, D) - \frac{\epsilon_0}{\gamma} - \epsilon
\]

Since there is a failure probability of at most \( 4\delta \) on each round, the algorithm succeeds with probability at least \( 1 - 4T\delta \).

\[\square\]

4 Applications

We show that recent agnostic learning analyses can be dramatically simplified using our boosting algorithm. Both of the agnostic algorithms are distribution-specific, meaning that they only work on one (or a family of) distributions \( \mu \) over unlabeled examples.

4.1 Agnostically Learning Decision Trees

Recent work has shown how to agnostically learn polynomial-sized decision trees using membership queries, by an \( L_1 \) gradient-projection algorithm [10]. Here, we show that learning decision trees is quite simple using our distribution-specific boosting theorem and the Kushilevitz-Mansour membership query parity learning algorithm as a weak learner [24].

**Lemma 4.** Running the KM algorithm, using \( q = \text{poly}(n, t, 1/\epsilon_0) \) queries, and outputting the parity with largest magnitude of estimated Fourier coefficient, is a \( (\gamma = 1/t, \epsilon_0) \) agnostic weak learner for size-\( t \) decision trees over the uniform distribution.

The proof of this Lemma is simple using results in [24] and is given in Appendix A. Theorem 2 now follows easily from Lemma 4 and Theorem 1.

4.2 Agnostically Learning Halfspaces

In the case of learning halfspaces, the weak learner simply finds the degree-\( d \) term, \( \chi_S(x) \) with \( |S| \leq d \), with greatest empirical correlation \( \frac{1}{m} \sum_{i=1}^{m} \chi_S(x_i) y_i \) on a data set \( (x_1, y_1), \ldots, (x_m, y_m) \).

The following lemma is useful in analyzing it.

**Lemma 5.** For any \( \epsilon > 0 \), there exists \( d \geq 1 \) such that the following holds. Let \( n \geq 1 \), \( C \) be the class of halfspaces in \( n \) dimensions, let \( \mathcal{U} \) be the uniform distribution on \( \{-1, 1\}^n \), and \( f : \{-1, 1\}^n \rightarrow [-1, 1] \) be an arbitrary function. Then there exists a set \( S \subseteq [n] \) of size \( |S| \leq d = \frac{2\epsilon}{\epsilon_0} \) such that \( |\text{cor}(\chi_S, \mathcal{U}f)| \geq (\text{cor}(C, \mathcal{U}f) - \epsilon_0)/n^d \).
Using results from [25] the proofs of Lemma 5 and Theorem 3 are straightforward and are given in Appendix B.

5 Experiments

We performed preliminary experiments with the new boosting algorithm presented here on 8 datasets from UCI repository [26]. We converted multi-class problems into binary classification problems by arbitrarily grouping classes, and ran Adaboost, Madaboost and Agnostic Boost on these datasets, using stumps as weak learners. Since stumps can accept weighted examples, we passed the exact weighted distribution to the weak learner.

Our experiments were performed with *fractional relabeling*, which means the following. Rather than keeping the label with probability $w^t(x,y)$ and making it completely random with the remaining probability, we added both $(x,y)$ and $(x,-y)$ with weights $(1 + w^t(x,y))/2$ and $(1 - w^t(x,y))/2$ respectively. Experiments with random relabeling showed that random relabeling performs much worse than fractional relabeling.

Table 1 summarizes the final test error on the datasets. In the case of pima and german datasets, we observed overfitting and the reported test errors are the minimum test error observed for all the algorithms. In all other cases the test error rate at the end of round 500 is reported. Only pendigits had a test dataset, for the rest of the datasets we performed 10-fold cross validation. We also added random classification noise of 5%, 10% and 20% to the datasets and ran the boosting algorithms on the modified dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No Added Noise</th>
<th>5% noise</th>
<th>10% Noise</th>
<th>20% Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>sonar</td>
<td>12.4</td>
<td>14.8</td>
<td>15.3</td>
<td>23.9</td>
</tr>
<tr>
<td>ionosphere</td>
<td>8.6</td>
<td>9.1</td>
<td>8.1</td>
<td>15.8</td>
</tr>
<tr>
<td>pima</td>
<td>23.7</td>
<td>23.0</td>
<td>23.6</td>
<td>26.1</td>
</tr>
<tr>
<td>german</td>
<td>23.1</td>
<td>23.6</td>
<td>23.1</td>
<td>28.5</td>
</tr>
<tr>
<td>waveform</td>
<td>10.4</td>
<td>10.2</td>
<td>10.3</td>
<td>14.9</td>
</tr>
<tr>
<td>magic</td>
<td>14.7</td>
<td>14.9</td>
<td>14.5</td>
<td>18.2</td>
</tr>
<tr>
<td>letter</td>
<td>17.4</td>
<td>18.2</td>
<td>18.3</td>
<td>20.9</td>
</tr>
<tr>
<td>pendigits</td>
<td>7.4</td>
<td>7.3</td>
<td>8.2</td>
<td>12.1</td>
</tr>
</tbody>
</table>

Table 1: Final test error rates of Adaboost, Madaboost and Agnostic Boosting on 8 datasets. The first column reports error rates on the original datasets, and the next three report errors on datasets with 5%, 10% and 20% classification noise added.

6 Conclusion

We show that potential-based agnostic boosting is possible in theory, and also that this may be done without changing the distribution over unlabeled examples. We show that non-trivial agnostic learning results, for learning decision trees and halfspaces, can be viewed as simple applications of our boosting theorem combined with well-known weak learners. Our analysis can be viewed as a theoretical justification of noise tolerance properties of algorithms like Madaboost and Smoothboost. Preliminary experiments show that the performance of our boosting algorithm is comparable to that of Madaboost and Adaboost. A more thorough empirical evaluation of our boosting procedure using different weak learners is part of future research.

References


