The Rate of Convergence of AdaBoost

Indraneel Mukherjee
Computer Science Department
Princeton University
Princeton, NJ 08540 USA
imukherj@cs.princeton.edu

Cynthia Rudin
MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
rudin@mit.edu

Robert E. Schapire
Computer Science Department
Princeton University
Princeton, NJ 08540 USA
schapire@cs.princeton.edu

Abstract

The AdaBoost algorithm was designed to combine many "weak" hypotheses that perform slightly better than random guessing into a "strong" hypothesis that has very low error. We study the rate at which AdaBoost iteratively converges to the minimum of the "exponential loss." Unlike previous work, our proofs do not require a weak-learning assumption, nor do they require that minimizers of the exponential loss are finite. Our first result shows that at iteration t, the exponential loss of AdaBoost's computed parameter vector will be at most ε more than that of any parameter vector of ℓ_1-norm bounded by B in a number of rounds that is at most a polynomial in B and $1/\varepsilon$. We also provide lower bounds showing that a polynomial dependence on these parameters is necessary. Our second result is that within C/ε iterations, AdaBoost achieves a value of the exponential loss that is at most ε more than the best possible value, where C depends on the dataset. We show that this dependence of the rate on ε is optimal up to constant factors, that is, at least $\Omega(1/\varepsilon)$ rounds are necessary to achieve within ε of the optimal exponential loss.

1 Introduction

The AdaBoost algorithm of Freund and Schapire (1997) was designed to combine many "weak" hypotheses that perform slightly better than random guessing into a "strong" hypothesis that has very low error. Despite extensive theoretical and empirical study, basic properties of AdaBoost's convergence are not fully understood. In this work, we focus on one of those properties, namely, to find convergence rates that hold in the absence of any simplifying assumptions. Such assumptions, relied upon in much of the preceding work, make it easier to prove a fast convergence rate for AdaBoost, but often do not hold in the cases where AdaBoost is commonly applied.

AdaBoost can be viewed as a coordinate descent (or functional gradient descent) algorithm that iteratively minimizes an objective function $L : \mathbb{R}^n \to \mathbb{R}$ called the *exponential loss* (Breiman, 1999, Frean and Downs, 1998, Friedman et al., 2000, Friedman, 2001, Mason et al., 2000, Onoda et al., 1998, Rätsch et al., 2001, Schapire and Singer, 1999). 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 $y_i \in \{-1, +1\}$, and a finite (but typically very large) space of weak hypotheses $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$, where each $\hbar_j : \mathcal{X} \to \{-1, +1\}$, the exponential loss is defined as

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

where $\boldsymbol{\lambda} = \langle \lambda_1, \ldots, \lambda_N \rangle$ is a vector of weights or parameters. In each iteration, a coordinate descent algorithm moves some distance along some coordinate direction λ_j. For AdaBoost, the coordinate directions correspond to the individual weak hypotheses. Thus, on each round, AdaBoost chooses some weak hypothesis and step length, and adds these to the current weighted combination of weak hypotheses, which is equivalent to updating a single weight. The direction and step length are so chosen that the resulting vector $\boldsymbol{\lambda}^t$ in iteration t yields a lower value of the exponential loss than in the previous iteration, $L(\boldsymbol{\lambda}^t) < L(\boldsymbol{\lambda}^{t-1})$. This repeats until it reaches a minimizer if one exists. It was shown by Collins et al. (2002), and later by Zhang and Yu (2005), that AdaBoost asymptotically converges to the minimum possible exponential loss. That is,

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

However, that work did not address a convergence rate to the minimizer of the exponential loss.

Our work specifically addresses a recent conjecture of Schapire (2010) stating 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 $B > 0$,

$$L(\boldsymbol{\lambda}^t) \leq \min_{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|_1 \leq B} L(\boldsymbol{\lambda}) + \frac{\mathrm{poly}(\log N, m, B)}{t^c}. \tag{1}$$

In other words, 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 will typically be extremely large.) Along with an upper bound that is polynomial in these parameters, we also provide lower bound constructions showing some polynomial dependence on B and $1/\varepsilon$ is necessary. Without any additional assumptions on the exponential loss L, and without altering AdaBoost's minimization algorithm for L, the best known convergence rate of AdaBoost prior to this work that we are aware of is that of Bickel et al. (2006) who prove a bound on the rate of the form $O(1/\sqrt{\log t})$.

We provide also a convergence rate of AdaBoost to the minimum value of the exponential loss. Namely, within C/ϵ iterations, AdaBoost achieves a value of the exponential loss that is at most ϵ more than the best possible value, where C depends on the dataset. This convergence rate is different from the one discussed above in that it has better dependence on ϵ (in fact the dependence is optimal, as we show), and does not depend on the best solution within a ball of size B. However, this second convergence rate cannot be used to prove (1) since in certain worst case situations, we show the constant C may be larger than 2^m (although usually it will be much smaller).

Within the proof of the second convergence rate, we provide a lemma (called the *decomposition lemma*) that shows that the training set can be split into two sets of examples: the "finite margin set," and the "zero loss set." Examples in the finite margin set always make a positive contribution to the exponential loss, and they never lie too far from the decision boundary. Examples in the zero loss set do not have these properties. If we consider the exponential loss where the sum is only over the finite margin set (rather than over all training examples), it is minimized by a finite $\boldsymbol{\lambda}$. The fact that the training set can be decomposed into these two classes is the key step in proving the second convergence rate.

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). It may also be relevant to practitioners who wish to have a guarantee on the exponential loss value at iteration t (although, in general, minimization of the exponential loss need not be perfectly correlated with test accuracy).

There have been several works that make additional assumptions on the exponential loss in order to attain a better bound on the rate, but those assumptions are not true in general, and cases are known where each of these assumptions are violated. For instance, better bounds are proved by Rätsch et al. (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. There are many cases where L does not have a finite minimizer; in fact, one such case is provided by Schapire (2010). Shalev-Shwartz and Singer (2008) have proven bounds for a variant of AdaBoost. Zhang and Yu (2005) also have given rates of convergence, but their technique requires a bound on the change in the size of $\boldsymbol{\lambda}^t$ at each iteration that does not necessarily hold for AdaBoost. Many classic results are known on the convergence of iterative algorithms generally (see for instance Luenberger and Ye, 2008, 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, assumptions that do not generally hold in our setting. When the weak learning assumption holds, there is a parameter $\gamma > 0$ that governs the improvement of the exponential loss at each iteration. Freund and Schapire (1997) and Schapire and Singer (1999) showed that the exponential loss is at most $e^{-2t\gamma^2}$ after t rounds, so AdaBoost rapidly converges to the minimum possible loss under this assumption.

In Section 2 we summarize the coordinate descent view of AdaBoost. Section 3 contains the proof of the conjecture, with associated lower bounds proved in Section 4. Section 5 provides the C/ϵ convergence rate. The proof of the decomposition lemma is given in Section 6.

2 Coordinate Descent View of AdaBoost

From the examples $(x_1, y_1), \ldots, (x_m, y_m)$ and hypotheses $\mathcal{H} = \{h_1, \ldots, h_N\}$, AdaBoost iteratively computes the function $F : \mathcal{X} \to \mathbb{R}$, where $\mathrm{sign}(F(x))$ can be used as a classifier for a new instance x. The function F is a linear combination of the hypotheses. At each iteration t, AdaBoost chooses one of the weak hypotheses h_t from the set $\mathcal{H}$, and adjusts its coefficient by a specified value α_t. Then F is constructed after T iterations as: $F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$. Figure 1 shows the AdaBoost algorithm (Freund and Schapire, 1997).

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$
set $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$ 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 $h_t \in \mathcal{H}$ whose correlation $r_t \triangleq \mathbb{E}_{i \sim D_t}\left[y_i h_t(x_i)\right]$ has maximum magnitude $|r_t|$.
- Choose $\alpha_t = \frac{1}{2} \ln\left\{(1 + r_t) \,/\, (1 - r_t)\right\}$.
- 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: $F(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Figure 1: The boosting algorithm AdaBoost.

Since each h_t is equal to $\hbar_{j_t}$ for some j_t, F can also be written $F(x) = \sum_{j=1}^{N} \lambda_j \hbar_j(x)$ for a vector of values $\boldsymbol{\lambda} = \langle \lambda_1, \ldots \lambda_N \rangle$ (such vectors will sometimes also be referred to as *combinations*, since they represent combinations of weak hypotheses). In different notation, we can write AdaBoost as a coordinate descent algorithm on vector $\boldsymbol{\lambda}$. We define the *feature matrix* $\mathbf{M}$ elementwise by $M_{ij} = y_i \hbar_j(x_i)$, so that this matrix contains all of the inputs to AdaBoost (the training examples and hypotheses). Then the exponential loss can be written more compactly as:

$$L(\boldsymbol{\lambda}) = \frac{1}{m} \sum_{i=1}^{m} e^{-(\mathbf{M}\boldsymbol{\lambda})_i}$$

where $(\mathbf{M}\boldsymbol{\lambda})_i$, the i^{th} coordinate of the vector $\mathbf{M}\boldsymbol{\lambda}$, is the (unnormalized) *margin* achieved by vector $\boldsymbol{\lambda}$ on training example i.

Coordinate descent algorithms choose a coordinate at each iteration where the directional derivative is the steepest, and choose a step that maximally decreases the objective along that coordinate. To perform coordinate descent on the exponential loss, we determine the coordinate j_t at iteration t as follows, where $\mathbf{e}_j$ is a vector that is 1 in the j^{th} position and 0 elsewhere:

$$j_t \in \operatorname*{argmax}_j \left| \left(-\frac{dL(\boldsymbol{\lambda}^{t-1} + \alpha \mathbf{e}_j)}{d\alpha}\Big|_{\alpha=0} \right) \right| = \operatorname*{argmax}_j \frac{1}{m} \left| \sum_{i=1}^{m} e^{-(\mathbf{M}\boldsymbol{\lambda}^{t-1})_i} M_{ij} \right|. \tag{2}$$

It can be shown (see for instance Mason et al., 2000) that the distribution D_t chosen by AdaBoost at each round t puts weight $D_t(i)$ proportional to $e^{(-\mathbf{M}\boldsymbol{\lambda}^{t-1})_i}$. Eq. (2) can now be rewritten as

$$j_t \in \operatorname*{argmax}_j \left| \sum_i D_t(i) M_{ij} \right| = \operatorname*{argmax}_j \left| \mathbb{E}_{i \sim D_t}\left[M_{ij}\right] \right| = \operatorname*{argmax}_j \left| \mathbb{E}_{i \sim D_t}\left[y_i h_j(x_i)\right] \right|,$$

which is exactly the way AdaBoost chooses a weak hypothesis in each round (see Figure 1). The correlation $\sum_i D_t(i) M_{ij_t}$ will be denoted by r_t and its absolute value $|r_t|$ denoted by δ_t. The quantity δ_t is commonly called the *edge* for round t. The distance α_t to travel along direction j_t is chosen to minimize $L(\boldsymbol{\lambda}^{t-1} + \alpha_t \mathbf{e}_{j_t})$, and can be shown to be equal to $\alpha_t = \frac{1}{2} \ln\left(\frac{1+r_t}{1-r_t}\right)$ (see for instance Mason et al., 2000), just as in Figure 1. With this choice of step length, it can be shown (see for instance Freund and Schapire, 1997) that the exponential loss drops by an amount depending on the edge: $L(\boldsymbol{\lambda}^t) = L(\boldsymbol{\lambda}^{t-1})\sqrt{1 - \delta_t^2}$.

Our rate bounds also hold when the weak-hypotheses are confidence-rated, that is, giving real-valued predictions in $[-1, +1]$, so that $h : \mathcal{X} \to [-1, +1]$. In that case, the criterion for picking a weak hypothesis in each round remains the same, that is, at round t, an $\hbar_{j_t}$ maximizing the absolute correlation $j_t \in \text{argmax}_j \left| \sum_{i=1}^{m} e^{-(\mathbf{M}\boldsymbol{\lambda}^{t-1})_i} M_{ij} \right|$, is chosen, where M_{ij} may now be non-integral. An exact analytical line search is no longer possible, but if the step size is chosen in the same way,

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 + r_t}{1 - r_t}\right), \tag{3}$$

then Freund and Schapire (1997) and Schapire and Singer (1999) show that a similar drop in the loss is still guaranteed:

$$L(\boldsymbol{\lambda}^t) \leq L(\boldsymbol{\lambda}^{t-1})\sqrt{1 - \delta^2}. \tag{4}$$

With confidence rated hypotheses, other implementations may choose the step size in a different way. However, in this paper, by "AdaBoost" we will always mean the version in (Freund and Schapire, 1997, Schapire and Singer, 1999) which chooses step sizes as in (3), and enjoys the loss guarantee as in (4). That said, all our proofs work more generally, and are robust to numerical inaccuracies in the implementation. In other words, even if the previous conditions are violated by a small amount, similar bounds continue to hold, although we leave out explicit proofs of this fact to simplify the presentation.

3 Convergence to any target loss

In this section, we bound the number of rounds of AdaBoost required to get within ε of the loss attained by any parameter vector $\boldsymbol{\lambda}^*$ as a function of ε and the ℓ_1-norm $\|\boldsymbol{\lambda}^*\|_1$. The vector $\boldsymbol{\lambda}^*$ serves as a reference based on which we define the target loss $L(\boldsymbol{\lambda}^*)$, and its ℓ_1-norm is a measure of the difficulty of attaining the target loss. We prove a bound polynomial in $1/\varepsilon$, $\|\boldsymbol{\lambda}^*\|_1$ and the number of examples m, showing (1) holds, thereby resolving affirmatively the open problem posed in (Schapire, 2010). Later in the section we provide lower bounds showing how a polynomial dependence on both parameters is necessary.

Theorem 1 *For any $\boldsymbol{\lambda}^* \in \mathbb{R}^N$, AdaBoost achieves loss at most $L(\boldsymbol{\lambda}^*) + \varepsilon$ in at most $13\|\boldsymbol{\lambda}^*\|_1^6 \varepsilon^{-5}$ rounds.*

The high level idea behind the proof of the theorem is as follows. To show a fast rate, we require a large edge in each round, as indicated by (4). A large edge is guaranteed if the size of the current solution of AdaBoost is small. Therefore AdaBoost makes good progress if the size of its solution does not grow too fast. On the other hand, the increase in size of its solution is given by the step length, which in turn is proportional to the edge achieved in that round. Therefore, if the solution size grows fast, the loss also drops fast. Either way the algorithm makes good progress. In the rest of the section we make these ideas concrete through a sequence of lemmas.

We provide some more notation. Throughout, $\boldsymbol{\lambda}^*$ is fixed, and its ℓ_1-norm is denoted by B (matching the notation in Schapire, 2010). One key parameter is the suboptimality R_t of AdaBoost's solution measured via the logarithm of the exponential loss:

$$R_t \triangleq \ln L(\boldsymbol{\lambda}^t) - \ln L(\boldsymbol{\lambda}^*).$$

Another key parameter is the ℓ_1-distance S_t of AdaBoost's solution from the closest combination that achieves the target loss:

$$S_t \triangleq \inf_{\boldsymbol{\lambda}} \left\{ \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|_1 : L(\boldsymbol{\lambda}) \le L(\boldsymbol{\lambda}^*) \right\}.$$

We will also be interested in how they change as captured by

$$\Delta R_t \triangleq R_{t-1} - R_t, \qquad \Delta S_t \triangleq S_t - S_{t-1}.$$

Notice that ΔR_t is always non-negative since AdaBoost decreases the loss, and hence the suboptimality, in each round. Let T_0 be the bound on the number of rounds in Theorem 1. We assume without loss of generality that $R_0, \ldots, R_{T_0}$ and $S_0, \ldots, S_{T_0}$ are all strictly positive, since otherwise the theorem holds trivially. Also, in the rest of the section, we restrict our attention entirely to the first T_0 rounds of boosting. We first show that a $\mathrm{poly}(B, \varepsilon^{-1})$ rate of convergence follows if the edge is always polynomially large compared to the suboptimality.

Lemma 2 *If for some constants c_1, c_2, where $c_2 > 1/2$, the edge satisfies $\delta_t \ge B^{-c_1} R_{t-1}^{c_2}$ in each round t, then AdaBoost achieves at most $L(\boldsymbol{\lambda}^*) + \varepsilon$ loss after $2B^{2c_1}(\varepsilon \ln 2)^{1-2c_2}$ rounds.*

Proof: From the definition of R_t and (4) we have

$$\Delta R_t = \ln L(\boldsymbol{\lambda}^{t-1}) - \ln L(\boldsymbol{\lambda}^t) \ge -\frac{1}{2}\ln(1 - \delta_t^2). \tag{5}$$

Combining the above with the inequality $e^x \ge 1 + x$, and the assumption on the edge

$$\Delta R_t \ge -\frac{1}{2}\ln(1 - \delta_t^2) \ge \frac{1}{2}\delta_t^2 \ge \frac{1}{2} B^{-2c_1} R_{t-1}^{2c_2}.$$

Let $T = \lceil 2B^{2c_1}(\varepsilon \ln 2)^{1-2c_2} \rceil$ be the bound on the number of rounds in the lemma. If any of $R_0, \ldots, R_T$ is negative, then by monotonicity $R_T < 0$ and we are done. Otherwise, they are all non-negative. Then, applying Lemma 18 from the Appendix to the sequence $R_0, \ldots, R_T$, and using $c_2 > 1/2$ we get

$$R_T^{1-2c_2} \ge R_0^{1-2c_2} + c_2 B^{-2c_1} T > (1/2) B^{-2c_1} T \ge (\varepsilon \ln 2)^{1-2c_2} \implies R_T < \varepsilon \ln 2.$$

If either ε or $L(\boldsymbol{\lambda}^*)$ is greater than 1, then the lemma follows since $L(\boldsymbol{\lambda}^T) \leq L(\boldsymbol{\lambda}^0) = 1 < L(\boldsymbol{\lambda}^*) + \varepsilon$. Otherwise,

$$L(\boldsymbol{\lambda}^T) < L(\boldsymbol{\lambda}^*) e^{\varepsilon \ln 2} \leq L(\boldsymbol{\lambda}^*)(1+\varepsilon) \leq L(\boldsymbol{\lambda}^*) + \varepsilon,$$

where the second inequality uses $e^x \leq 1 + (1/\ln 2)x$ for $x \in [0, \ln 2]$. ■

We next show that large edges are achieved provided S_t is small compared to R_t.

Lemma 3 *In each round t, the edge satisfies* $\delta_t \geq R_{t-1}/S_{t-1}$.

Proof: For any combination $\boldsymbol{\lambda}$, define $p_{\boldsymbol{\lambda}}$ as the distribution on examples $\{1, \ldots, m\}$ that puts weight proportional to the loss $D_{\boldsymbol{\lambda}}(i) = e^{-(\mathbf{M}\boldsymbol{\lambda})_i}/(mL(\boldsymbol{\lambda}))$. Choose any $\boldsymbol{\lambda}$ suffering at most the target loss $L(\boldsymbol{\lambda}) \leq L(\boldsymbol{\lambda}^*)$. By non-negativity of relative entropy we get

$$\begin{aligned} 0 &\leq \mathtt{RE}(D_{\boldsymbol{\lambda}^{t-1}} \parallel D_{\boldsymbol{\lambda}}) = \sum_{i=1}^m D_{\boldsymbol{\lambda}^{t-1}} \ln \left(\frac{\frac{1}{m} e^{-(\mathbf{M}\boldsymbol{\lambda}^{t-1})_i}/L(\boldsymbol{\lambda}^{t-1})}{\frac{1}{m} e^{-(\mathbf{M}\boldsymbol{\lambda})_i}/L(\boldsymbol{\lambda})} \right) \\ &= -R_{t-1} + \sum_{i=1}^m D_{\boldsymbol{\lambda}^{t-1}}(i) \left(\mathbf{M}\boldsymbol{\lambda} - \mathbf{M}\boldsymbol{\lambda}^{t-1} \right)_i . \end{aligned} \tag{6}$$

Note that $D_{\boldsymbol{\lambda}^{t-1}}$ is the distribution D_t that AdaBoost creates in round t. The above summation can be rewritten as

$$\begin{aligned} \sum_{i=1}^m D_{\boldsymbol{\lambda}^{t-1}}(i) \sum_{j=1}^N \left(\lambda_j - \lambda_j^{t-1}\right) M_{ij} &= \sum_{j=1}^N \left(\lambda_j - \lambda_j^{t-1}\right) \sum_{i=1}^m D_t(i) M_{ij} \\ &\leq \left(\sum_{j=1}^N \left|\lambda_j - \lambda_j^{t-1}\right| \right) \max_j \left| \sum_{i=1}^m D_t(i) M_{ij} \right| = \delta_t \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t-1}\|_1 . \end{aligned} \tag{7}$$

Since the previous holds for any $\boldsymbol{\lambda}$ suffering less than the target loss, the last expression is at most $\delta_t S_{t-1}$. Combining this with (7) completes the proof. ■

To complete the proof of Theorem 1, we show S_t is small compared to R_t in rounds $t \leq T_0$ (during which we have assumed S_t, R_t are all positive). In fact we prove:

Lemma 4 *For any* $t \leq T_0$, $S_t \leq B^3 R_t^{-2}$.

This, along with Lemmas 2 and 3, immediately proves Theorem 1. The bound on S_t in Lemma 4 can be proven if we can first show S_t grows slowly compared to the rate at which the suboptimality R_t falls. Intuitively this holds since growth in S_t is caused by a large step, which in turn will drive down the suboptimality. In fact we can prove the following.

Lemma 5 *In any round* $t \leq T_0$, *we have* $\frac{2\Delta R_t}{R_{t-1}} \geq \frac{\Delta S_t}{S_{t-1}}$.

Proof: Firstly, it follows from the definition of S_t that $\Delta S_t \leq \|\boldsymbol{\lambda}^t - \boldsymbol{\lambda}^{t-1}\|_1 = |\alpha_t|$. Next, using (5) and (3) we may write $\Delta R_t \geq \Upsilon(\delta_t)\,|\alpha_t|$, where the function Υ has been defined in (Rätsch and Warmuth, 2005) as

$$\Upsilon(x) = \frac{-\ln(1-x^2)}{\ln\left(\frac{1+x}{1-x}\right)}.$$

It is known (Rätsch and Warmuth, 2005, Rudin et al., 2007) that $\Upsilon(x) \geq x/2$ for $x \in [0,1]$. Combining and using Lemma 3,

$$\Delta R_t \geq \delta_t \Delta S_t/2 \geq R_{t-1}\left(\Delta S_t/2S_{t-1}\right).$$

Rearranging completes the proof. ■

Using this we may prove Lemma 4.

Proof: We first show $S_0 \leq B^3 R_0^{-2}$. Note, $S_0 \leq \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\|_1 = B$, and by definition the quantity $R_0 = -\ln\left(\frac{1}{m}\sum_i e^{-(\mathbf{M}\boldsymbol{\lambda}^*)_i}\right)$. The quantity $(\mathbf{M}\boldsymbol{\lambda}^*)_i$ is the inner product of row i of matrix $\mathbf{M}$ with the vector $\boldsymbol{\lambda}^*$. Since the entries of $\mathbf{M}$ lie in $[-1,+1]$, this is at most $\|\boldsymbol{\lambda}^*\|_1 = B$. Therefore $R_0 \leq -\ln\left(\frac{1}{m}\sum_i e^{-B}\right) = B$, which is what we needed.

To complete the proof, we show that $R_t^2 S_t$ is non-increasing. It suffices to show for any t the inequality $R_t^2 S_t \leq R_{t-1}^2 S_{t-1}$. This holds by the following chain:

$$\begin{aligned} R_t^2 S_t &= (R_{t-1} - \Delta R_t)^2 (S_{t-1} + \Delta S_t) = R_{t-1}^2 S_{t-1} \left(1 - \frac{\Delta R_t}{R_{t-1}}\right)^2 \left(1 + \frac{\Delta S_t}{S_{t-1}}\right) \\ &\leq R_{t-1}^2 S_{t-1} \exp\left(-\frac{2\Delta R_t}{R_{t-1}} + \frac{\Delta S_t}{S_{t-1}}\right) \leq R_{t-1}^2 S_{t-1}, \end{aligned}$$

where the first inequality follows from $e^x \geq 1 + x$, and the second one from Lemma 5. ■

Although we achieve a rate polynomial in B and ε^{-1} as conjectured by Schapire (2010), the exponents are rather large, and (we believe) not tight. As evidence supporting this belief, we note that a minor modification to AdaBoost can converge much faster. This variant, which we call AdaBoost.S, is the same as AdaBoost except that at the end of each round, the current combination of weak hypotheses is scaled back, that is, multiplied by a scalar in $[0, 1]$ if doing so will reduce the exponential loss further (and in particular, choosing that scalar which causes the greatest decrease in the loss). With only this modification, the rate of AdaBoost.S can be bounded by B^2/ε using similar techniques to those given above. (Details omitted for lack of space.) As shown in the next section on rate lower bounds, this is nearly the best possible.

4 Lower bounds

Here we show that the dependence of the rate in Theorem 1 on the norm $\|\boldsymbol{\lambda}^*\|_1$ of a solution achieving target accuracy is necessary for a wide class of datasets. Although we prove our results for exponential loss, the arguments in this section hold more generally for any coordinate descent algorithm based on a loss function of the form $L(\boldsymbol{\lambda}) = (1/m) \sum_i \phi((\mathbf{M}\boldsymbol{\lambda})_i)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is any non-negative non-increasing function.

The first lemma connects the size of a competing solution to the required number of rounds of boosting, and shows that for a wide variety of datasets the convergence rate to a target loss can be lower bounded by the ℓ_1-norm of the smallest solution achieving that loss.

Lemma 6 *Suppose the feature matrix $\mathbf{M}$ corresponding to a dataset has two rows with $\{-1, +1\}$ entries that are complements of each other, that is, there are two examples on which any hypothesis gets one wrong and one correct prediction. Then the number of rounds required to achieve a target loss L^* is at least $\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq L^*\}/(2 \ln m)$.*

Proof: We first show that the two examples corresponding to the complementary rows in $\mathbf{M}$ both satisfy a certain margin boundedness property. Since each hypothesis predicts oppositely on these, in any round t their margins will be of equal magnitude and opposite sign. Unless both margins lie in $[-\ln m, \ln m]$, one of them will be smaller than $-\ln m$. But then the exponential loss $L(\boldsymbol{\lambda}^t) = (1/m) \sum_j e^{-(\mathbf{M}\boldsymbol{\lambda}^t)_j}$ in that round will exceed 1, a contradiction since the losses are non-increasing through rounds, and the loss at the start was 1. Thus, assigning one of these examples the index i, we have the absolute margin $|(\mathbf{M}\boldsymbol{\lambda}^t)_i|$ is bounded by $\ln m$ in any round t. Letting $\mathbf{M}(i)$ denote the ith row of $\mathbf{M}$, the step length α_t in round t therefore satisfies

$$|\alpha_t| = |M_{ij_t}\alpha_t| = |\langle \mathbf{M}(i), \alpha_t \mathbf{e}_{j_t} \rangle| = \left|(\mathbf{M}\boldsymbol{\lambda}^t)_i - (\mathbf{M}\boldsymbol{\lambda}^{t-1})_i\right| \leq \left|(\mathbf{M}\boldsymbol{\lambda}^t)_i\right| + \left|(\mathbf{M}\boldsymbol{\lambda}^{t-1})_i\right| \leq 2 \ln m,$$

and the statement of the lemma directly follows. ■

The next theorem constructs a feature matrix satisfying the properties of Lemma 6 and where additionally the smallest size of a solution achieving $L^* + \varepsilon$ loss is at least $\Omega(2^m) \ln(1/\varepsilon)$, for some fixed L^* and every $\varepsilon > 0$. This implies that when ε is a small constant (say $\varepsilon = 0.01$), and $\boldsymbol{\lambda}^*$ is some vector with loss $L^* + \varepsilon/2$, AdaBoost takes at least $\Omega(2^m / \ln m)$ steps to get within $\varepsilon/2$ of the loss achieved by $\boldsymbol{\lambda}^*$, that is, to within $L^* + \varepsilon$ loss. Since m and ε are independent quantities, this shows that a polynomial dependence of the convergence rate on the norm of the competing solution is unavoidable. Further this norm might be exponential in the number of training examples and weak hypotheses in the worst case, and hence the bound $\mathrm{poly}(\|\boldsymbol{\lambda}^*\|_1, 1/\varepsilon)$ in Theorem 1 cannot be replaced by $\mathrm{poly}(m, N, 1/\varepsilon)$.

Theorem 7 *Consider the following matrix $\mathbf{M}$ with m rows (or examples) labeled $0, \ldots, m-1$ and $m-1$ columns labeled $1, \ldots, m-1$ (assume $m \geq 3$). The square sub-matrix ignoring row zero is an upper triangular matrix, with 1's on the diagonal, -1's above the diagonal, and 0 below the diagonal. Therefore row 1 is $(+1, -1, -1, \ldots, -1)$. Row 0 is defined to be just the complement of row 1. Then, for any $\varepsilon > 0$, a loss of $2/m + \varepsilon$ is achievable on this dataset, but with large norms*

$$\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq 2/m + \varepsilon\} \geq (2^{m-2} - 1) \ln(1/(3\varepsilon)).$$

Therefore, by Lemma 6, the minimum number of rounds required for reaching loss at most $2/m + \varepsilon$ is at least $\left(\frac{2^{m-2}-1}{2 \ln m}\right) \ln(1/(3\varepsilon))$.

Proof: We first lower bound the norm of solutions achieving loss at most $2/m+\varepsilon$. Observe that since rows 0 and 1 are complementary, any solution's loss on just examples 0 and 1 will add up to at least $2/m$. Therefore, to get within $2/m+\varepsilon$, the margins on examples $2,\ldots,m-1$ should be at least $\ln\left((m-2)/(m\varepsilon)\right) \geq \ln(1/(3\varepsilon))$ (for $m \geq 3$). Now, the feature matrix is designed so that the margins due to a combination $\boldsymbol{\lambda}$ satisfy the following recursive relationships:

$$\begin{aligned}(M\boldsymbol{\lambda})_{m-1} &= \lambda_{m-1},\\ (M\boldsymbol{\lambda})_i &= \lambda_i - (\lambda_{i+1}+\ldots+\lambda_{m-1}),\ \text{for } 1 \leq i \leq m-2.\end{aligned}$$

Therefore, the margin on example $m-1$ is at least $\ln(1/(3\varepsilon))$ implies $\lambda_{m-1} \geq \ln(1/(3\varepsilon))$. Similarly, $\lambda_{m-2} \geq \ln(1/(3\varepsilon)) + \lambda_{m-1} \geq 2\ln(1/(3\varepsilon))$. Continuing this way,

$$\lambda_i \geq \ln\left(\frac{1}{3\varepsilon}\right) + \lambda_{i+1}+\ldots+\lambda_{m-1} \geq \ln\left(\frac{1}{3\varepsilon}\right)\left\{1+2^{(m-1)-(i+1)}+\ldots+2^0\right\} = \ln\left(\frac{1}{3\varepsilon}\right)2^{m-1-i},$$

for $i = m-1,\ldots,2$. Hence $\|\boldsymbol{\lambda}\|_1 \geq \ln(1/(3\varepsilon))(1+2+\ldots+2^{m-3}) = (2^{m-2}-1)\ln(1/(3\varepsilon))$.

We end by showing that a loss of at most $2/m+\varepsilon$ is achievable. The above argument implies that if $\lambda_i = 2^{m-1-i}$ for $i = 2,\ldots,m-1$, then examples $2,\ldots,m-1$ attain margin exactly 1. If we choose $\lambda_1 = \lambda_2+\ldots+\lambda_{m-1} = 2^{m-3}+\ldots+1 = 2^{m-2}-1$, then the recursive relationship implies a zero margin on example 1 (and hence example 0). Therefore the combination $\ln(1/\varepsilon)(2^{m-2}-1, 2^{m-3}, 2^{m-4}, \ldots, 1)$ achieves a loss $(2+(m-2)\varepsilon)/m \leq 2/m+\varepsilon$, for any $\varepsilon > 0$. ■

The above lower-bound examples use feature matrices with only integer entries, i.e., entries in $\{-1,0,+1\}$. The lower bounds can be much larger when the entries are real numbers in $[-1,+1]$. In fact, tiny perturbations to a feature matrix with integer entries on which AdaBoost converges very fast can lead to a new matrix with fractional entries that requires arbitrarily large convergence times. The proof of this fact is also based on the norm of the competing solution (details omitted). In the next section we investigate the dependence of the convergence rate on the other independent parameter ε, and show that $\Omega(1/\varepsilon)$ rounds are necessary.

5 Convergence to optimal loss

In the previous section, our rate bound depended on both the approximation parameter ε, as well as the size of the smallest solution achieving the target loss. For many datasets, the optimal target loss $\inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})$ cannot be realized by any finite solution. In such cases, if we want to bound the number of rounds needed to achieve within ε of the optimal loss, the only way to use Theorem 1 is to first decompose the accuracy parameter ε into two parts $\varepsilon = \varepsilon_1+\varepsilon_2$, find some finite solution $\boldsymbol{\lambda}^*$ achieving within ε_1 of the optimal loss, and then use the bound $\mathrm{poly}(1/\varepsilon_2, \|\boldsymbol{\lambda}^*\|_1)$ to achieve at most $L(\boldsymbol{\lambda}^*)+\varepsilon_2 = \inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda})+\varepsilon$ loss. However, this introduces implicit dependence on ε through $\|\boldsymbol{\lambda}^*\|_1$ which may not be immediately clear. In this section, we show bounds of the form C/ε, where the constant C depends only on the feature matrix $\mathbf{M}$, and not on ε. A similar approach to solving this problem was taken independently by Telgarsky (2011).

Theorem 8 *AdaBoost reaches within ε of the optimal loss in at most C/ε rounds, where C only depends on the feature matrix.*

Additionally, we show that this dependence on ε is optimal in Lemma 17 of the Appendix, where $\Omega(1/\varepsilon)$ rounds are shown to be necessary for converging to within ε of the optimal loss on a certain dataset. Finally, we note that the lower bounds in the previous section indicate that C can be $\Omega(2^m)$ in the worst case for integer matrices (although it will typically be much smaller), and hence this bound, though stronger than that of Theorem 1 with respect to ε, cannot be used to prove the conjecture in (Schapire, 2010), since the constant is not polynomial in the number of examples m.

Our techniques build upon earlier work on the rate of convergence of AdaBoost, which have mainly considered two particular cases. In the first case, the *weak learning assumption* holds, that is, the edge in each round is at least some fixed constant. In this situation, Freund and Schapire (1997) and Schapire and Singer (1999) show that the optimal loss is zero, that no solution with finite size can achieve this loss, but AdaBoost achieves at most ε loss within $O(\ln(1/\varepsilon))$ rounds. In the second case some finite combination of the weak classifiers achieves the optimal loss, and Rätsch et al. (2002), using results from Luo and Tseng (1992), show that AdaBoost achieves within ε of the optimal loss again within $O(\ln(1/\varepsilon))$ rounds.

Here we consider the most general situation, where the weak learning assumption may fail to hold, and yet no finite solution may achieve the optimal loss. The dataset used in Lemma 17 and shown in Figure 2 exemplifies this situation. Our main technical contribution shows that the examples in any dataset can be partitioned into a *zero-loss set* and *finite-margin set*, such that a certain form of the weak learning assumption holds within the zero-loss set, while the optimal loss considering only the finite-margin set can be obtained by some finite solution. The two partitions provide different ways of making progress in every round, and one of the two kinds of progress will always be sufficient for us to prove Theorem 8.

We next state our decomposition result, illustrate it with an example, and then state several lemmas quantifying the nature of the progress we can make in each round. Using these lemmas, we prove Theorem 8.

Lemma 9 *(Decomposition Lemma) For any dataset, there exists a partition of the set of training examples X into a (possibly empty)* zero-loss set *Z and a (possibly empty)* finite-margin set *$F = Z^c \triangleq X \setminus Z$ such that the following hold simultaneously :*

1. *For some positive constant $\gamma > 0$, there exists some vector $\boldsymbol{\eta}^\dagger$ with unit ℓ_1-norm $\|\boldsymbol{\eta}^\dagger\|_1 = 1$ that attains at least γ margin on each example in Z, and exactly zero margin on each example in F*

$$\forall i \in Z : (\mathbf{M}\boldsymbol{\eta}^\dagger)_i \geq \gamma, \qquad \forall i \in F : (\mathbf{M}\boldsymbol{\eta}^\dagger)_i = 0.$$

2. *The optimal loss considering only examples within F is achieved by some finite combination $\boldsymbol{\eta}^*$.*

3. *There is a constant $\mu_{\max} < \infty$, such that for any combination $\boldsymbol{\eta}$ with bounded loss on the finite-margin set, $\sum_{i \in F} e^{-(\mathbf{M}\boldsymbol{\eta})_i} \leq m$, the margin $(\mathbf{M}\boldsymbol{\eta})_i$ for any example i in F lies in the bounded interval $[-\ln m, \mu_{\max}]$.*

A proof is deferred to the next section. The decomposition lemma immediately implies that the vector $\boldsymbol{\eta}^* + \infty \cdot \boldsymbol{\eta}^\dagger$, which denotes $\left(\boldsymbol{\eta}^* + c\boldsymbol{\eta}^\dagger\right)$ in the limit $c \to \infty$, is an optimal solution, achieving zero loss on the zero-loss set, but only finite margins (and hence positive losses) on the finite-margin set (thereby justifying the names).

	$\hbar_1$	$\hbar_2$
a	$+$	$-$
b	$-$	$+$
c	$+$	$+$

Figure 2: A dataset requiring $\Omega(1/\varepsilon)$ rounds for convergence.

Before proceeding, we give an example dataset and indicate the zero-loss set, finite-margin set, $\boldsymbol{\eta}^*$ and $\boldsymbol{\eta}^\dagger$ to illustrate our definitions. Consider a dataset with three examples $\{a, b, c\}$ and two hypotheses $\{\hbar_1, \hbar_2\}$ and the feature matrix $\mathbf{M}$ in Figure 2. Here $+$ means correct ($M_{ij} = +1$) and $-$ means wrong ($M_{ij} = -1$). The optimal solution is $\infty \cdot (\hbar_1 + \hbar_2)$ with a loss of $2/3$. The finite-margin set is $\{a, b\}$, the zero-loss set is $\{c\}$, $\boldsymbol{\eta}^\dagger = (1/2, 1/2)$ and $\boldsymbol{\eta}^* = (0, 0)$; for this dataset these are unique. This dataset also serves as a lower-bound example in Lemma 17, where we show that $2/(9\varepsilon)$ rounds are necessary for AdaBoost to achieve loss at most $(2/3) + \varepsilon$.

Before providing proofs, we introduce some notation. By $\|\cdot\|$ we will mean ℓ_2-norm; every other norm will have an appropriate subscript, such as $\|\cdot\|_1, \|\cdot\|_\infty$, etc. The set of all training examples will be denoted by X. By $\ell^{\boldsymbol{\lambda}}(i)$ we mean the exp-loss $e^{-(\mathbf{M}\boldsymbol{\lambda})_i}$ on example i. For any subset $S \subseteq X$ of examples, $\ell^{\boldsymbol{\lambda}}(S) = \sum_{i \in S} \ell^{\boldsymbol{\lambda}}(i)$ denotes the total exp-loss on the set S. Notice $L(\boldsymbol{\lambda}) = (1/m)\ell^{\boldsymbol{\lambda}}(X)$, and that $D_{t+1}(i) = \ell^{\boldsymbol{\lambda}^t}(i)/\ell^{\boldsymbol{\lambda}^t}(X)$, where $\boldsymbol{\lambda}^t$ is the combination found by AdaBoost at the end of round t. By $\delta_S(\boldsymbol{\eta}; \boldsymbol{\lambda})$ we mean the edge obtained on the set S by the vector $\boldsymbol{\eta}$, when the weights over the examples are given by $\ell^{\boldsymbol{\lambda}}(\cdot)/\ell^{\boldsymbol{\lambda}}(S)$:

$$\delta_S(\boldsymbol{\eta}; \boldsymbol{\lambda}) = \left| \frac{1}{\ell^{\boldsymbol{\lambda}}(S)} \sum_{i \in S} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\boldsymbol{\eta})_i \right|.$$

In the rest of the section, by "loss" we mean the unnormalized loss $\ell^{\boldsymbol{\lambda}}(X) = mL(\boldsymbol{\lambda})$ and show that in C/ε rounds AdaBoost converges to within ε of the optimal unnormalized loss $\inf_{\boldsymbol{\lambda}} \ell^{\boldsymbol{\lambda}}(X)$, henceforth denoted by K. Note that this means AdaBoost takes C/ε rounds to converge to within ε/m of the optimal normalized loss, that is to loss at most $\inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) + \varepsilon/m$. Replacing ε by $m\varepsilon$, it takes $C/(m\varepsilon)$ steps to attain normalized loss at most $\inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) + \varepsilon$. Thus, whether we use normalized or unnormalized does not substantively affect the result in Theorem 8. The progress due to the zero-loss set is now immediate from Item 1 of the decomposition lemma:

Lemma 10 *In any round t, the maximum edge δ_t is at least $\gamma \ell^{\boldsymbol{\lambda}^{t-1}}(Z)/\ell^{\boldsymbol{\lambda}^{t-1}}(X)$, where γ is as in Item 1 of the decomposition lemma.*

Proof: Recall the distribution D_t created by AdaBoost in round t puts weight $D_t(i) = \ell^{\boldsymbol{\lambda}^{t-1}}(i)/\ell^{\boldsymbol{\lambda}^{t-1}}(X)$ on each example i. From Item 1 we get

$$\delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1}) = \left| \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in X} \ell^{\boldsymbol{\lambda}^{t-1}}(i)(\mathbf{M}\boldsymbol{\eta}^\dagger)_i \right| = \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in Z} \gamma \ell^{\boldsymbol{\lambda}^{t-1}}(i) = \gamma \left(\frac{\ell^{\boldsymbol{\lambda}^{t-1}}(Z)}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \right).$$

Since $(\mathbf{M}\boldsymbol{\eta}^\dagger)_i = \sum_j \eta_j^\dagger (\mathbf{M}\mathbf{e}_j)_i$, we may rewrite the edge $\delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1})$ as follows:

$$\begin{aligned}
\delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1}) &= \left| \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in X} \ell^{\boldsymbol{\lambda}^{t-1}}(i) \sum_j \eta_j^\dagger (\mathbf{M}\mathbf{e}_j)_i \right| \\
&= \left| \sum_j \eta_j^\dagger \frac{1}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)} \sum_{i \in X} \ell^{\boldsymbol{\lambda}^{t-1}}(i) (\mathbf{M}\mathbf{e}_j)_i \right| \\
&= \left| \sum_j \eta_j^\dagger \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1}) \right| \leq \sum_j \left| \eta_j^\dagger \right| \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1}).
\end{aligned}$$

Since the ℓ_1-norm of $\boldsymbol{\eta}^\dagger$ is 1, the weights $\left|\eta_j^\dagger\right|$ form some distribution p over the columns $1, \ldots, N$. We may therefore conclude

$$\gamma\left(\frac{\ell^{\boldsymbol{\lambda}^{t-1}}(Z)}{\ell^{\boldsymbol{\lambda}^{t-1}}(X)}\right) = \delta_X(\boldsymbol{\eta}^\dagger; \boldsymbol{\lambda}^{t-1}) \leq \mathbb{E}_{j \sim p}\left[\delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1})\right] \leq \max_j \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}^{t-1}) \leq \delta_t.$$

■

If the set F were empty, then Lemma 10 implies an edge of γ is available in each round. This in fact means that the weak learning assumption holds, and using (4), we can show an $O(\ln(1/\varepsilon)\gamma^{-2})$ bound matching the rate bounds of Freund and Schapire (1997) and Schapire and Singer (1999). So henceforth, we assume that F is non-empty. Note that this implies that the optimal loss K is at least 1 (since any solution will get non-positive margin on some example in F), a fact we will use later in the proofs.

Lemma 10 says that the edge is large if the loss on the zero-loss set is large. On the other hand, when it is small, Lemmas 11 and 12 together show how AdaBoost can make good progress using the finite margin set. Lemma 11 uses second order methods to show how progress is made in the case where there is a finite solution. Similar arguments, under additional assumptions, have earlier appeared in (Rätsch et al., 2002).

Lemma 11 *Suppose $\boldsymbol{\lambda}$ is a combination such that $m \geq \ell^{\boldsymbol{\lambda}}(F) \geq K$. Then in some coordinate direction the edge is at least $\sqrt{C_0\left(\ell^{\boldsymbol{\lambda}}(F) - K\right)/\ell^{\boldsymbol{\lambda}}(F)}$, where C_0 is a constant depending only on the feature matrix $\mathbf{M}$.*

Proof: Let $\mathbf{M}_F \in \mathbb{R}^{|F| \times N}$ be the matrix $\mathbf{M}$ restricted to only the rows corresponding to the examples in F. Choose $\boldsymbol{\eta}$ such that $\boldsymbol{\lambda} + \boldsymbol{\eta} = \boldsymbol{\eta}^*$ is an optimal solution over F. Without loss of generality assume that $\boldsymbol{\eta}$ lies in the orthogonal subspace of the null-space $\{\mathbf{u} : \mathbf{M}_F\mathbf{u} = \mathbf{0}\}$ of $\mathbf{M}_F$ (since we can translate $\boldsymbol{\eta}^*$ along the null space if necessary for this to hold). If $\boldsymbol{\eta} = \mathbf{0}$, then $\ell^{\boldsymbol{\lambda}}(F) = K$ and we are done. Otherwise $\|\mathbf{M}_F\boldsymbol{\eta}\| \geq \lambda_{\min}\|\boldsymbol{\eta}\|$, where $\lambda_{\min}^2$ is the smallest positive eigenvalue of the symmetric matrix $\mathbf{M}_F^T\mathbf{M}_F$ (exists since $\mathbf{M}_F\boldsymbol{\eta} \neq \mathbf{0}$). Now define $f : [0,1] \to \mathbb{R}$ as the loss along the (rescaled) segment $[\boldsymbol{\eta}^*, \boldsymbol{\lambda}]$

$$f(x) \triangleq \ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(F) = \sum_{i \in F} \ell^{\boldsymbol{\eta}^*}(i) e^{x(\mathbf{M}\boldsymbol{\eta})_i}.$$

This implies that $f(0) = K$ and $f(1) = \ell^{\boldsymbol{\lambda}}(F)$. Notice that the first and second derivatives of $f(x)$ are given by:

$$f'(x) = \sum_{i \in F} (\mathbf{M}_F\boldsymbol{\eta})_i \ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(i), \qquad f''(x) = \sum_{i \in F} (\mathbf{M}_F\boldsymbol{\eta})_i^2 \ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(i).$$

We next lower bound possible values of the second derivative as follows:

$$f''(x) = \sum_{i' \in F} (\mathbf{M}_F\boldsymbol{\eta})_{i'}^2 \ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(i') \geq \sum_{i' \in F} (\mathbf{M}_F\boldsymbol{\eta})_{i'}^2 \min_i \ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(i) \geq \|\mathbf{M}_F\boldsymbol{\eta}\|^2 \min_i \ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(i).$$

Since both $\boldsymbol{\lambda} = \boldsymbol{\eta}^* - \boldsymbol{\eta}$, and $\boldsymbol{\eta}^*$ suffer total loss at most m, by convexity, so does $\boldsymbol{\eta}^* - x\boldsymbol{\eta}$ for any $x \in [0,1]$. Hence we may apply Item 3 of the decomposition lemma to the vector $\boldsymbol{\eta}^* - x\boldsymbol{\eta}$, for any $x \in [0,1]$, to conclude that $\ell^{(\boldsymbol{\eta}^* - x\boldsymbol{\eta})}(i) = \exp\left\{-(\mathbf{M}_F(\boldsymbol{\eta}^* - x\boldsymbol{\eta}))_i\right\} \geq e^{-\mu_{\max}}$ on every example i. Therefore we have,

$$f''(x) \geq \|\mathbf{M}_F\boldsymbol{\eta}\|^2 e^{-\mu_{\max}} \geq \lambda_{\min}^2 e^{-\mu_{\max}} \|\boldsymbol{\eta}\|^2 \text{ (by choice of } \boldsymbol{\eta}) .$$

A standard second-order result is (see e.g. Boyd and Vandenberghe, 2004, eqn. (9.9))

$$|f'(1)|^2 \geq 2\left(\inf_{x \in [0,1]} f''(x)\right)(f(1) - f(0)) .$$

Collecting our results so far, we get

$$\sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\boldsymbol{\eta})_i = |f'(1)| \geq \|\boldsymbol{\eta}\| \sqrt{2\lambda_{\min}^2 e^{-\mu_{\max}} \left(\ell^{\boldsymbol{\lambda}}(F) - K\right)}.$$

Next let $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}/\|\boldsymbol{\eta}\|_1$ be $\boldsymbol{\eta}$ rescaled to have unit ℓ_1 norm. Then we have

$$\sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\tilde{\boldsymbol{\eta}})_i = \frac{1}{\|\boldsymbol{\eta}\|_1} \sum_i \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\boldsymbol{\eta})_i \geq \frac{\|\boldsymbol{\eta}\|}{\|\boldsymbol{\eta}\|_1} \sqrt{2\lambda_{\min}^2 e^{-\mu_{\max}} \left(\ell^{\boldsymbol{\lambda}}(F) - K\right)}.$$

Applying the Cauchy-Schwarz inequality, we may lower bound $\frac{\|\boldsymbol{\eta}\|}{\|\boldsymbol{\eta}\|_1}$ by $1/\sqrt{N}$ (since $\boldsymbol{\eta} \in \mathbb{R}^N$). Along with the fact $\ell^{\boldsymbol{\lambda}}(F) \leq m$, we may write

$$\frac{1}{\ell^{\boldsymbol{\lambda}}(F)} \sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\tilde{\boldsymbol{\eta}})_i \geq \sqrt{2\lambda_{\min}^2 N^{-1} m^{-1} e^{-\mu_{\max}}} \sqrt{\left(\ell^{\boldsymbol{\lambda}}(F) - K\right)/\ell^{\boldsymbol{\lambda}}(F)}.$$

If we define p to be a distribution on the columns $\{1, \ldots, N\}$ of $\mathbf{M}_F$ which puts probability $p(j)$ proportional to $|\tilde{\boldsymbol{\eta}}_j|$ on column j, then we have

$$\frac{1}{\ell^{\boldsymbol{\lambda}}(F)} \sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\tilde{\boldsymbol{\eta}})_i \leq \mathbb{E}_{j \sim p} \left| \frac{1}{\ell^{\boldsymbol{\lambda}}(F)} \sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\mathbf{e}_j)_i \right| \leq \max_j \left| \frac{1}{\ell^{\boldsymbol{\lambda}}(F)} \sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\mathbf{e}_j)_i \right|.$$

Notice the quantity inside the max is precisely the edge $\delta_F(\mathbf{e}_j; \boldsymbol{\lambda})$ in direction j. Combining everything, the maximum possible edge is

$$\max_j \delta_F(\mathbf{e}_j; \boldsymbol{\lambda}) \geq \sqrt{C_0 \left(\ell^{\boldsymbol{\lambda}}(F) - K\right)/\ell^{\boldsymbol{\lambda}}(F)},$$

where we define $C_0 = 2\lambda_{\min}^2 N^{-1} m^{-1} e^{-\mu_{\max}}$. ■

Lemma 12 *Suppose, at some stage of boosting, the combination found by AdaBoost is $\boldsymbol{\lambda}$, and the loss is $K + \theta$. Let $\Delta\theta$ denote the drop in the suboptimality θ after one more round; i.e., the loss after one more round is $K + \theta - \Delta\theta$. Then there are constants C_1, C_2 depending only on the feature matrix (and not on θ), such that if $\ell^{\boldsymbol{\lambda}}(Z) < C_1\theta$, then $\Delta\theta \geq C_2\theta$.*

Proof: Let $\boldsymbol{\lambda}$ be the current solution found by boosting. Using Lemma 11, pick a direction j in which the edge $\delta_F(\mathbf{e}_j; \boldsymbol{\lambda})$ restricted to the finite loss set is at least $\sqrt{2C_0(\ell^{\boldsymbol{\lambda}}(F) - K)/\ell^{\boldsymbol{\lambda}}(F)}$. We can bound the edge $\delta_X(\mathbf{e}_j; \boldsymbol{\lambda})$ on the entire set of examples as follows:

$$\begin{aligned}
\delta_X(\mathbf{e}_j; \boldsymbol{\lambda}) &= \frac{1}{\ell^{\boldsymbol{\lambda}}(X)} \left| \sum_{i \in F} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\mathbf{e}_j)_i + \sum_{i \in Z} \ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\mathbf{e}_j)_i \right| \\
&\geq \frac{1}{\ell^{\boldsymbol{\lambda}}(X)} \left(\left| \ell^{\boldsymbol{\lambda}}(F) \delta_F(\mathbf{e}_j; \boldsymbol{\lambda}) \right| - \sum_{i \in Z} \ell^{\boldsymbol{\lambda}}(i) \right) \text{ (using the triangle inequality)} \\
&\geq \frac{1}{\ell^{\boldsymbol{\lambda}}(X)} \left(\sqrt{2C_0(\ell^{\boldsymbol{\lambda}}(F) - K)\ell^{\boldsymbol{\lambda}}(F)} - \ell^{\boldsymbol{\lambda}}(Z) \right).
\end{aligned}$$

Now, $\ell^{\boldsymbol{\lambda}}(Z) < C_1\theta$, and $\ell^{\boldsymbol{\lambda}}(F) - K = \theta - \ell^{\boldsymbol{\lambda}}(Z) \geq (1 - C_1)\theta$. Further, we will choose $C_1 < 1$, so that $\ell^{\boldsymbol{\lambda}}(F) \geq K \geq 1$. Hence, the previous inequality implies

$$\delta_X(\mathbf{e}_j; \boldsymbol{\lambda}) \geq \frac{1}{K + \theta} \left(\sqrt{2C_0(1 - C_1)\theta} - C_1\theta \right).$$

Set $C_1 = \min \left\{ 1/2, (1/4)\sqrt{C_0/(2m)} \right\}$. Using $\theta \leq K + \theta = \ell^{\boldsymbol{\lambda}}(X) \leq m$, we can bound the square of the term in brackets on the previous line as

$$\begin{aligned}
\left(\sqrt{2C_0(1 - C_1)\theta} - C_1\theta \right)^2 &\geq 2C_0(1 - C_1)\theta - 2C_1\theta\sqrt{2C_0(1 - C_1)\theta} \\
&\geq 2C_0(1 - 1/2)\theta - 2\left((1/4)\sqrt{C_0/(2m)} \right)\theta\sqrt{2C_0(1 - 0)m} = C_0\theta/2.
\end{aligned}$$

So, if δ is the maximum edge in any direction, then

$$\delta \geq \delta_X(\mathbf{e}_j; \boldsymbol{\lambda}) \geq \sqrt{C_0\theta/(2(K+\theta)^2)} \geq \sqrt{C_0\theta/(2m(K+\theta))},$$

where, for the last inequality, we again used $K + \theta \leq m$. Therefore the loss after one more step is at most $(K+\theta)\sqrt{1-\delta^2} \leq (K+\theta)(1-\delta^2/2) \leq K + \theta - \frac{C_0}{4m}\theta$. Setting $C_2 = C_0/(4m)$ completes the proof. ■

Proof of Theorem 8. At any stage of boosting, let $\boldsymbol{\lambda}$ be the current combination, and $K + \theta$ be the current loss. We show that the new loss is at most $K + \theta - \Delta\theta$ for $\Delta\theta \geq C_3\theta^2$ for some constant C_3 depending only on the dataset (and not θ). To see this, either $\ell^{\boldsymbol{\lambda}}(Z) < C_1\theta$, in which case Lemma 12 applies, and $\Delta\theta \geq C_2\theta \geq (C_2/m)\theta^2$ (since $\theta = \ell^{\boldsymbol{\lambda}}(X) - K \leq m$). Or $\ell^{\boldsymbol{\lambda}}(Z) \geq C_1\theta$, in which case applying Lemma 10 yields $\delta \geq \gamma C_1\theta/\ell^{\boldsymbol{\lambda}}(X) \geq (\gamma C_1/m)\theta$. By (4), $\Delta\theta \geq \ell^{\boldsymbol{\lambda}}(X)(1-\sqrt{1-\delta^2}) \geq \ell^{\boldsymbol{\lambda}}(X)\delta^2/2 \geq (K/2)(\gamma C_1/m)^2\theta^2$. Using $K \geq 1$ and choosing C_3 appropriately gives the required condition.

If $K+\theta_t$ denotes the loss in round t, then the above claim implies $\theta_t - \theta_{t+1} \geq C_3\theta_t^2$. Applying Lemma 18 to the sequence $\{\theta_t\}$ we have $1/\theta_T - 1/\theta_0 \geq C_3 T$ for any T. Since $\theta_0 \geq 0$, we have $T \leq 1/(C_3\theta_T)$. Hence to achieve loss $K + \varepsilon$, C_3^{-1}/ε rounds suffice. ■

The hidden constant C in Theorem 8 depends on intrinsic properties of the feature matrix. When the matrix has real entries, the discussion in the previous section implies that C may be arbitrarily large. When the entries are restricted to $\{-1, 0, +1\}$ we can upper bound C by just a function of the dimensions of the matrix. These dimensions are the number of training examples and the number of weak hypotheses in the dataset (details omitted).

6 Proof of the decomposition lemma

Throughout this section we only consider (unless otherwise stated) *admissible* combinations $\boldsymbol{\lambda}$ of weak classifiers, which have loss $\ell^{\boldsymbol{\lambda}}(X)$ bounded by m (since such are the ones found by boosting). We prove Lemma 9 in three steps. We begin with a simple lemma that rigorously defines the zero-loss and finite-margin sets.

Lemma 13 *For any sequence $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots,$ of admissible combinations of weak classifiers, we can find a subsequence $\boldsymbol{\eta}_{(1)} = \boldsymbol{\eta}_{t_1}, \boldsymbol{\eta}_{(2)} = \boldsymbol{\eta}_{t_2}, \ldots,$ whose losses converge to zero on all examples in some fixed (possibly empty) subset Z (the zero-loss set), and losses bounded away from zero in its complement $X \setminus Z$(the finite-margin set)*

$$\forall x \in Z : \lim_{t\to\infty} \ell^{\boldsymbol{\eta}_{(t)}}(x) = 0, \qquad \forall x \in X \setminus Z : \inf_i \ell^{\boldsymbol{\eta}_{(t)}}(x) > 0. \tag{8}$$

Proof: We will build a zero-loss set and the final subsequence incrementally. Initially the set is empty. Pick the first example. If the infimal loss ever attained on the example in the sequence is bounded away from zero, then we do not add it to the set. Otherwise we add it, and consider only the subsequence whose t^{th} element attains loss less than $1/t$ on the example. Beginning with this subsequence, we now repeat with other examples. The final sequence is the required subsequence, and the examples we have added form the zero-loss set. ■

We apply Lemma 13 to some admissible sequence converging to the optimal loss (for instance, the one found by AdaBoost). Let us call the resulting subsequence $\boldsymbol{\eta}^*_{(t)}$, the obtained zero-loss set Z, and the finite-margin set $F = X \setminus Z$. The next lemma shows how to extract a single combination out of the sequence $\boldsymbol{\eta}^*_{(t)}$ that satisfies the properties in Item 1 of the decomposition lemma.

Lemma 14 *Suppose $\mathbf{M}$ is the feature matrix, Z is a subset of the examples, and $\boldsymbol{\eta}_{(1)}, \boldsymbol{\eta}_{(2)}, \ldots,$ is a sequence of combinations of weak classifiers such that Z is its zero loss set, and $X \setminus Z$ its finite loss set, that is,* (8) *holds. Then there is a combination $\boldsymbol{\eta}^\dagger$ of weak classifiers that achieves positive margin on every example in Z, and zero margin on every example in its complement $X \setminus Z$, that is:*

$$(\mathbf{M}\boldsymbol{\eta}^\dagger)_i \begin{cases} > 0 & \text{if } i \in Z, \\ = 0 & \text{if } i \in X \setminus Z. \end{cases}$$

Proof: Since the $\boldsymbol{\eta}_{(t)}$ achieve arbitrarily large positive margins on Z, $\|\boldsymbol{\eta}_{(t)}\|$ will be unbounded, and it will be hard to extract a useful single solution out of them. On the other hand, the rescaled combinations $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$ lie on a compact set, and therefore have a limit point, which might have useful properties. We formalize this next.

We prove the statement of the lemma by induction on the total number of training examples $|X|$. If X is empty, then the lemma holds vacuously for any $\boldsymbol{\eta}^\dagger$. Assume inductively for all X of size less than $m > 0$, and consider X of size m. Since translating a vector along the null space of $\mathbf{M}$, $\ker \mathbf{M} = \{\mathbf{x} : \mathbf{M}\mathbf{x} = \mathbf{0}\}$, has no effect on the margins produced by the vector, assume without loss of generality that the $\boldsymbol{\eta}_{(t)}$'s are

orthogonal to $\ker \mathbf{M}$. Also, since the margins produced on the zero loss set are unbounded, so are the norms of $\boldsymbol{\eta}_{(t)}$. Therefore assume (by picking a subsequence and relabeling if necessary) that $\|\boldsymbol{\eta}_{(t)}\| > t$. Let $\boldsymbol{\eta}'$ be a limit point of the sequence $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$, a unit vector that is also orthogonal to the null-space. Then firstly $\boldsymbol{\eta}'$ achieves non-negative margin on every example; otherwise by continuity for some extremely large t, the margin of $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$ on that example is also negative and bounded away from zero, and therefore $\boldsymbol{\eta}_{(t)}$'s loss is more than m, a contradiction to admissibility. Secondly, the margin of $\boldsymbol{\eta}'$ on each example in $X \setminus Z$ is zero; otherwise, by continuity, for arbitrarily large t the margin of $\boldsymbol{\eta}_{(t)}/\|\boldsymbol{\eta}_{(t)}\|$ on an example in $X \setminus Z$ is positive and bounded away from zero, and hence that example attains arbitrarily small loss in the sequence, a contradiction to (8). Finally, if $\boldsymbol{\eta}'$ achieves zero margin everywhere in Z, then $\boldsymbol{\eta}'$, being orthogonal to the null-space, must be $\mathbf{0}$, a contradiction since $\boldsymbol{\eta}'$ is a unit vector. Therefore $\boldsymbol{\eta}'$ must achieve positive margin on some non-empty subset S of Z, and zero margins on every other example.

Next we use induction on the reduced set of examples $X' = X \setminus S$. Since S is non-empty, $|X'| < m$. Further, using the same sequence $\boldsymbol{\eta}_{(t)}$, the zero-loss and finite-loss sets, restricted to X', are $Z' = Z \setminus S$ and $(X \setminus Z) \setminus S = X \setminus Z$ (since $S \subseteq Z$) $= X' \setminus Z'$. By the inductive hypothesis, there exists some $\boldsymbol{\eta}''$ which achieves positive margins on Z', and zero margins on $X' \setminus Z' = X \setminus Z$. Therefore, by setting $\boldsymbol{\eta}^\dagger = \boldsymbol{\eta}' + c\boldsymbol{\eta}''$ for a large enough c, we can achieve the desired properties. ■

Applying Lemma 14 to the sequence $\boldsymbol{\eta}^*_{(t)}$ yields some convex combination $\boldsymbol{\eta}^\dagger$ having margin at least $\gamma > 0$ (for some γ) on Z and zero margin on its complement, proving Item 1 of the decomposition lemma. The next lemma proves Item 2.

Lemma 15 *The optimal loss considering only examples within F is achieved by some finite combination $\boldsymbol{\eta}^*$.*

Proof: The existence of $\boldsymbol{\eta}^\dagger$ with properties as in Lemma 14 implies that the optimal loss is the same whether considering all the examples, or just examples in F. Therefore it suffices to show the existence of finite $\boldsymbol{\eta}^*$ that achieves loss K on F, that is, $\ell^{\boldsymbol{\eta}^*}(F) = K$.

Recall $\mathbf{M}_F$ denotes the matrix $\mathbf{M}$ restricted to the rows corresponding to examples in F. Let $\ker \mathbf{M}_F = \{\mathbf{x} : \mathbf{M}_F\mathbf{x} = 0\}$ be the null-space of $\mathbf{M}_F$. Let $\boldsymbol{\eta}^{(t)}$ be the projection of $\boldsymbol{\eta}^*_{(t)}$ onto the orthogonal subspace of $\ker \mathbf{M}_F$. Then the losses $\ell^{\boldsymbol{\eta}^{(t)}}(F) = \ell^{\boldsymbol{\eta}^*_{(t)}}(F)$ converge to the optimal loss K. If $\mathbf{M}_F$ is identically zero, then each $\boldsymbol{\eta}^{(t)} = \mathbf{0}$, and then $\boldsymbol{\eta}^* = \mathbf{0}$ has loss K on F. Otherwise, let λ^2 be the smallest positive eigenvalue of $\mathbf{M}_F^T\mathbf{M}_F$. Then $\|\mathbf{M}\boldsymbol{\eta}^{(t)}\| \geq \lambda\|\boldsymbol{\eta}^{(t)}\|$. By the definition of finite margin set, $\inf_{t\to\infty} \min_{i\in F} \ell^{\boldsymbol{\eta}^{(t)}}(i) = \inf_{t\to\infty} \min_{i\in F} \ell^{\boldsymbol{\eta}^*_{(t)}}(i) > 0$. Therefore, the norms of the margin vectors $\|\mathbf{M}\boldsymbol{\eta}^{(t)}\|$, and hence that of $\boldsymbol{\eta}^{(t)}$, are bounded. Therefore the $\boldsymbol{\eta}^{(t)}$'s have a (finite) limit point $\boldsymbol{\eta}^*$ that must have loss K over F. ■

As a corollary, we prove Item 3.

Lemma 16 *There is a constant $\mu_{\max} < \infty$, such that for any combination $\boldsymbol{\eta}$ that achieves bounded loss on the finite-margin set, $\ell^{\boldsymbol{\eta}}(F) \leq m$, the margin $(\mathbf{M}\boldsymbol{\eta})_i$ for any example i in F lies in the bounded interval $[-\ln m, \mu_{\max}]$.*

Proof: Since the loss $\ell^{\boldsymbol{\eta}}(F)$ is at most m, therefore no margin may be less than $-\ln m$. To prove a finite upper bound on the margins, we argue by contradiction. Suppose arbitrarily large margins are producible by bounded loss vectors, that is the set $\{(\mathbf{M}\boldsymbol{\eta})_i : \ell^{\boldsymbol{\eta}}(F) \leq m, 1 \leq i \leq m\}$ contains arbitrarily large elements. Then for some fixed example $x \in F$ there exists a sequence of combinations of weak classifiers, whose t^{th} element achieves more than margin t on x but has loss at most m on F. Applying Lemma 13 we can find a subsequence $\boldsymbol{\lambda}^{(t)}$ whose tail achieves vanishingly small loss on some non-empty subset S of F containing x, and bounded margins in $F \setminus S$. Applying Lemma 14 to $\boldsymbol{\lambda}^{(t)}$ we get some convex combination $\boldsymbol{\lambda}^\dagger$ which has positive margins on S and zero margin on $F \setminus S$. Let $\boldsymbol{\eta}^*$ be as in Lemma 15, a finite combination achieving the optimal loss on F. Then $\boldsymbol{\eta}^* + \infty \cdot \boldsymbol{\lambda}^\dagger$ achieves the same loss on every example in $F \setminus S$ as the optimal solution $\boldsymbol{\eta}^*$, but zero loss for examples in S. This solution is strictly better than $\boldsymbol{\eta}^*$ on F, a contradiction to the optimality of $\boldsymbol{\eta}^*$. Therefore our assumption is false, and some finite upper bound $\mu_{\max}$ on the margins $(\mathbf{M}\boldsymbol{\eta})_i$ of vectors satisfying $\ell^{\boldsymbol{\eta}}(F) \leq m$ exists. ■

7 Conclusion

In this paper we studied the convergence rate of AdaBoost with respect to the exponential loss. We showed upper and lower bounds for convergence rates to both an arbitrary target loss achieved by some finite combination of the weak hypotheses, as well as to the infimum loss which may not be realizable. For the first convergence rate, we showed a strong relationship exists between the size of the minimum vector achieving a target loss and the number of rounds of coordinate descent required to achieve that loss. In particular, we

showed that a polynomial dependence of the rate on the ℓ_1-norm B of the minimum size solution is necessary, and that a $\mathrm{poly}(B, 1/\varepsilon)$ upper bound holds, where ε is the accuracy parameter. For the second kind of convergence, using entirely separate techniques we derived an $O(1/\varepsilon)$ upper bound, and showed that this is tight up to constant factors. In the process, we showed a certain decomposition lemma that might be of independent interest. In the full version of the paper we study the hidden constants more carefully. We also combine the separate techniques for the two rate proofs to obtain further improved estimates.

Acknowledgments

This research was funded by the National Science Foundation under grants IIS-1016029 and IIS-1053407. We thank Nikhil Srivastava for informing us of the matrix used in Theorem 7, and Matus Telgarsky for many helpful discussions.

References

Peter L. Bartlett and Mikhail Traskin. AdaBoost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.

Peter J. Bickel, Ya'acov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Leo Breiman. Prediction games and arcing classifiers. *Neural Computation*, 11(7):1493–1517, 1999.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1/2/3), 2002.

Marcus Frean and Tom Downs. A simple cost function for boosting. Technical report, Department of Computer Science and Electrical Engineering, University of Queensland, 1998.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–374, April 2000.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), October 2001.

David G. Luenberger and Yinyu Ye. *Linear and nonlinear programming*. Springer, third edition, 2008.

Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, January 1992.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, 2000.

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

G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.

Gunnar Rätsch and Manfred K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, 2005.

Gunnar Rätsch, Sebastian Mika, and Manfred K. Warmuth. On the convergence of leveraging. In *Advances in Neural Information Processing Systems 14*, 2002.

Cynthia Rudin, Robert E. Schapire, and Ingrid Daubechies. Analysis of boosting algorithms using the smooth margin function. *Annals of Statistics*, 35(6):2723–2768, 2007.

Robert E. Schapire. The convergence rate of AdaBoost. In *The 23rd Conference on Learning Theory*, 2010. open problem.

Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, December 1999.

Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *21st Annual Conference on Learning Theory*, 2008.

Matus Telgarsky. The convergence rate of AdaBoost and friends. http://arxiv.org/abs/1101.4752, January 2011.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

Appendix

Lemma 17 *For any $\varepsilon < 1/3$, to get within ε of the optimum loss on the dataset in Table 2, AdaBoost takes at least $2/(9\varepsilon)$ steps.*

Proof: Note that the optimal loss is $2/3$, and we are bounding the number of rounds necessary to get within $(2/3) + \varepsilon$ loss for $\varepsilon < 1/3$. We will compute the edge in each round analytically. Let w_a^t, w_b^t, w_c^t denote the normalized-losses (adding up to 1) or weights on examples a, b, c at the beginning of round t, h_t the weak hypothesis chosen in round t, and δ_t the edge in round t. The values of these parameters are shown below for the first 5 rounds, where we have assumed (without loss of generality) that the hypothesis picked in round 1 is $\hbar_b$:

Round	w_a^t	w_b^t	w_c^t	h_t	δ_t
$t=1:$	$1/3$	$1/3$	$1/3$	$\hbar_b$	$1/3$
$t=2:$	$1/2$	$1/4$	$1/4$	$\hbar_a$	$1/2$
$t=3:$	$1/3$	$1/2$	$1/6$	$\hbar_b$	$1/3$
$t=4:$	$1/2$	$3/8$	$1/8$	$\hbar_a$	$1/4$
$t=5:$	$2/5$	$1/2$	$1/10$	$\hbar_b$	$1/5.$

Based on the patterns above, we first claim that for rounds $t \geq 2$, the edge achieved is $1/t$. In fact we prove the stronger claims, that for rounds $t \geq 2$, the following hold:

1. One of w_a^t and w_b^t is $1/2$.
2. $\delta_{t+1} = \delta_t/(1+\delta_t)$.

Since $\delta_2 = 1/2$, the recurrence on δ_t would immediately imply $\delta_t = 1/t$ for $t \geq 2$. We prove the stronger claims by induction on the round t. The base case for $t = 2$ is shown above and may be verified. Suppose the inductive assumption holds for t. Assume without loss of generality that $1/2 = w_a^t > w_b^t > w_c^t$; note this implies $w_b^t = 1 - (w_a^t + w_c^t) = 1/2 - w_c^t$. Further, in this round, $\hbar_a$ gets picked, and has edge $\delta_t = w_a^t + w_c^t - w_b^t = 2w_c^t$. Now for any dataset, the weights of the examples labeled correctly and incorrectly in a round of AdaBoost are rescaled during the weight update step in a way such that each add up to $1/2$. Therefore, $w_b^{t+1} = 1/2, w_c^{t+1} = w_c^t \left(\frac{1/2}{w_a^t + w_c^t}\right) = w_c^t/(1+2w_c^t)$. Hence, $\hbar_b$ gets picked in round $t+1$ and, as before, we get edge $\delta_{t+1} = 2w_c^{t+1} = 2w_c^t/(1+2w_c^t) = \delta_t/(1+\delta_t)$. The proof of our claim follows by induction.

Next we find the loss after each iteration. Using $\delta_1 = 1/3$ and $\delta_t = 1/t$ for $t \geq 2$, the loss after T rounds can be written as

$$\prod_{t=1}^{T} \sqrt{1-\delta_t^2} = \sqrt{1-(1/3)^2} \prod_{t=2}^{T} \sqrt{1-1/t^2} = \frac{2\sqrt{2}}{3} \sqrt{\prod_{t=2}^{T} \left(\frac{t-1}{t}\right)\left(\frac{t+1}{t}\right)}.$$

The product can be rewritten as follows:

$$\prod_{t=2}^{T} \left(\frac{t-1}{t}\right)\left(\frac{t+1}{t}\right) = \left(\prod_{t=2}^{T} \frac{t-1}{t}\right)\left(\prod_{t=2}^{T} \frac{t+1}{t}\right) = \left(\prod_{t=2}^{T} \frac{t-1}{t}\right)\left(\prod_{t=3}^{T+1} \frac{t}{t-1}\right).$$

Notice almost all the terms cancel, except for the first term of the first product, and the last term of the second product. Therefore, the loss after T rounds is

$$\frac{2\sqrt{2}}{3}\sqrt{\left(\frac{1}{2}\right)\left(\frac{T+1}{T}\right)} = \frac{2}{3}\sqrt{1+\frac{1}{T}} \geq \frac{2}{3}\left(1+\frac{1}{3T}\right) = \frac{2}{3} + \frac{2}{9T},$$

where the inequality holds for $T \geq 1$. Since the initial error is $1 = (2/3) + 1/3$, therefore, for any $\varepsilon < 1/3$, the number of rounds needed to achieve loss $(2/3) + \varepsilon$ is at least $2/(9\varepsilon)$. ■

Lemma 18 *Suppose $u_0, u_1, \ldots,$ are non-negative numbers satisfying*

$$u_t - u_{t+1} \geq c_0 u_t^{1+c_1},$$

for some non-negative constants c_0, c_1. Then, for any t,

$$\frac{1}{u_t^{c_1}} - \frac{1}{u_0^{c_1}} \geq c_1 c_0 t.$$

Proof: By induction on t. The base case is an identity. Assume the statement holds at iteration t. Then,

$$\frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_0^{c_1}} = \left(\frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_t^{c_1}}\right) + \left(\frac{1}{u_t^{c_1}} - \frac{1}{u_0^{c_1}}\right) \geq \frac{1}{u_{t+1}^{c_1}} - \frac{1}{u_t^{c_1}} + c_1 c_0 t \text{ (by inductive hypothesis).}$$

Thus it suffices to show $1/u_{t+1}^{c_1} - 1/u_t^{c_1} \geq c_1 c_0$. Multiplying both sides by $u_t^{c_1}$ and adding 1, this is equivalent to showing $(u_t/u_{t+1})^{c_1} \geq 1 + c_1 c_0 u_t^{c_1}$. We will in fact show the stronger inequality

$$(u_t/u_{t+1})^{c_1} \geq (1 + c_0 u_t^{c_1})^{c_1}. \tag{9}$$

Since $(1+a)^b \geq 1 + ba$ for a, b non-negative, (9) will imply $(u_t/u_{t+1})^{c_1} \geq (1 + c_0 u_t^{c_1})^{c_1} \geq 1 + c_1 c_0 u_t^{c_1}$, which will complete our proof. To show (9), we first rearrange the condition on u_t, u_{t+1} to obtain

$$u_{t+1} \leq u_t \left(1 - c_0 u_t^{c_1}\right) \implies \frac{u_t}{u_{t+1}} \geq \frac{1}{1 - c_0 u_t^{c_1}}.$$

Applying the fact $\left(1 + c_0 u_t^{c_1}\right)\left(1 - c_0 u_t^{c_1}\right) \leq 1$ to the previous equation we get,

$$\frac{u_t}{u_{t+1}} \geq 1 + c_0 u_t^{c_1}.$$

Since $c_1 \geq 0$, we may raise both sides of the above inequality to the power of c_1 to show (9), finishing our proof. ■