Journal of Machine Learning Research 14 (2013) 2315-2347 Submitted 8/11; Revised 5/13; Published 8/13

The Rate of Convergence of AdaBoost

Indraneel Mukherjee IMUKHERJ@CS.PRINCETON.EDU
Princeton University
Department of Computer Science
Princeton, NJ 08540 USA

Cynthia Rudin RUDIN@MIT.EDU
Massachusetts Institute of Technology
MIT Sloan School of Management
Cambridge, MA 02139 USA

Robert E. Schapire SCHAPIRE@CS.PRINCETON.EDU
Princeton University
Department of Computer Science
Princeton, NJ 08540 USA

Editor: Tong Zhang

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 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 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 data set. 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.

Keywords: AdaBoost, optimization, coordinate descent, convergence rate

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. This definition can also be extended to the setting where the weak hypotheses are confidence rated, that is, they output real numbers, with the sign denoting prediction and magnitude indicating the confidence in that prediction. 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, the work by Collins et al. (2002) did not address a rate of convergence to the minimum exponential loss.

Our work specifically addresses a recent conjecture of Schapire (2010) stating that there exists a positive constant c and a polynomial poly() such that for all training sets and all finite sets of weak hypotheses, and for all $B > 0$,

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

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 data set. 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).

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 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 proved 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 of Schapire (2010), with associated lower bounds proved in Section 3.3. Section 4 provides the C/ε convergence rate. The proof of the decomposition lemma is given in Section 4.2.

2. Coordinate Descent View of AdaBoost

From the examples $(x_1, y_1), \ldots, (x_m, y_m)$ and hypotheses $\mathcal{H} = \{\hbar_1, \ldots, \hbar_N\}$, AdaBoost iteratively computes the function $F : \mathcal{X} \to \mathbb{R}$, where $\text{sign}(F(x))$ 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).

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

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 $r_t \triangleq \mathbb{E}_{i \sim D_t}[y_i h_t(x_i)]$ has maximum magnitude $|r_t|$.
- Choose $\alpha_t = \frac{1}{2} \ln\{(1 + r_t)/(1 - r_t)\}$.
- Update, for $i = 1, \ldots, m$: $D_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i h_t(x_i))/Z_t$
  where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

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

Figure 1: The boosting algorithm AdaBoost.

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:

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

We can show that this is equivalent to the weak learning step of AdaBoost. Unraveling the recursion in Figure 1 for AdaBoost's weight vector D_t, we can see that $D_t(i)$ is proportional to

$$\exp\left(-\sum_{t'<t} \alpha_{t'} y_i h_{t'}(x_i) \right).$$

The term in the exponent can also be rewritten in terms of the vector $\boldsymbol{\lambda}^t$, where λ_j^t is the sum of α_t's where hypothesis $\hbar_j$ was chosen: $\sum_{t'<t} \alpha_{t'} \mathbf{1}_{[\hbar_j = h_{t'}]} = \lambda_{t-1,j}$. The term in the exponent is:

$$\sum_{t'<t} \alpha_{t'} y_i h_{t'}(x_i) = \sum_j \sum_{t'<t} \alpha_{t'} \mathbf{1}_{[\hbar_j = h_{t'}]} y_i \hbar_j(x_i) = \sum_j \lambda_j^{t-1} M_{ij} = (\mathbf{M}\boldsymbol{\lambda}^{t-1})_i,$$

where $(\cdot)_i$ denotes the ith component of a vector. This means $D_t(i)$ is proportional to $e^{-(\mathbf{M}\boldsymbol{\lambda}^{t-1})_i}$. Equation (2) can now be rewritten as

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

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 found for coordinate descent via a linesearch (see for instance Mason et al., 2000):

$$0 = -\frac{dL(\boldsymbol{\lambda}^t + \alpha_t \mathbf{e}_{j_t})}{d\alpha_t} = \sum_i e^{-\left(M(\boldsymbol{\lambda}^t+\alpha_t\mathbf{e}_{j_t})\right)_i} M_{ij_t}$$

and dividing both sides by the normalization factor,

$$0 = \sum_{i:M_{ij}=1} D_t(i)e^{-\alpha_t} - \sum_{i:M_{ij}=-1} D_t(i)e^{\alpha_t} = \frac{(1+r_t)}{2}e^{-\alpha_t} - \frac{(1-r_t)}{2}e^{\alpha_t},$$

that is

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

just as in Figure 1. Thus, AdaBoost is equivalent to coordinate descent on $L(\boldsymbol{\lambda})$. With this choice of step length, it can be shown (Freund and Schapire, 1997) that the exponential loss drops by an amount depending on the edge. First notice that for any value of α_t we have:

$$\begin{aligned} L(\boldsymbol{\lambda}^t) &= L\left(\boldsymbol{\lambda}^{t-1} + \alpha_t \mathbf{e}_{\mathbf{j_t}}\right) \\ &= \left(\sum_{i:M_{ij_t}=1} D_t(i)e^{-\alpha_t} + \sum_{i:M_{ij_t}=-1} D_t(i)e^{\alpha_t}\right) L(\boldsymbol{\lambda}^{t-1}) \\ &= \left(\frac{(1+r_t)}{2}e^{-\alpha_t} + \frac{(1-r_t)}{2}e^{\alpha_t}\right) L(\boldsymbol{\lambda}^{t-1}). \end{aligned} \tag{3}$$

Plugging in the choice of α_t calculated before, we have

$$L(\boldsymbol{\lambda}^t) = \left(\sqrt{(1+r_t)(1-r_t)}\right) L(\boldsymbol{\lambda}^{t-1}) = \left(\sqrt{1-r_t^2}\right) L(\boldsymbol{\lambda}^{t-1}) = \left(\sqrt{1-\delta_t^2}\right) L(\boldsymbol{\lambda}^{t-1}).$$

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,

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right), \tag{4}$$

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

$$L(\boldsymbol{\lambda}^t) \le L(\boldsymbol{\lambda}^{t-1})\sqrt{1-\delta_t^2}. \tag{5}$$

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) and Schapire and Singer (1999) which chooses step sizes as in (4), and enjoys the loss guarantee as in (5). 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.

Before proceeding to the statements and proofs of convergence we make a few technical observations that will simplify all the proofs considerably. All the convergence statements in this paper are of the following form. Within a specific number of rounds T, AdaBoost will achieve loss at most L_0 for some non-negative L_0: that is, $L(\boldsymbol{\lambda}^T) \leq L_0$. The non-negativity is necessary since the exponential-loss L only takes non-negative values, and hence the minimum attainable value is 0. Since the loss is non-decreasing through various rounds of AdaBoost, we may assume, for the sake of proving the kind of bound mentioned above, that the losses $L(\boldsymbol{\lambda}^1), \ldots, L(\boldsymbol{\lambda}^T)$ are all strictly greater than zero. Otherwise, within T rounds the minimum possible loss of zero has already been attained and there is nothing to prove. By virtue of (5), the positivity assumption on the losses in turn implies that we may assume that the edges $\delta_1, \ldots, \delta_T$ are all strictly less than 1. Finally, note that $\delta_t = 0$ implies that the optimal solution has been attained. To have a nontrivial convergence problem, we may assume that all the edges are positive. Thus, we assume throughout that $0 < \delta_t < 1$ to ensure that the statements and proofs are non-trivial.

3. First Convergence Rate: 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 a 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 we will show that its ℓ_1-norm measures the difficulty of attaining the target loss in a specific sense. 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.

3.1 Upper Bound

The main result of this section is the following rate upper bound.

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 (5). A large edge is guaranteed if the size of the current solution $\|\boldsymbol{\lambda}^t\|_1$ 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:

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

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's loss is always decreasing, thus the suboptimality also decreases in each round. We assume without loss of generality that $R_0, \ldots, R_t$ and $S_0, \ldots, S_t$ are all strictly positive up to at least $t = 13\|\boldsymbol{\lambda}^*\|_1^6 \varepsilon^{-5}$, since otherwise the theorem holds trivially. In the rest of the section, we restrict our attention entirely to rounds of boosting when these positivity conditions hold. We first show that a poly(B, ε^{-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 $c_2 \ge 1$, 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.*

We will need the following expression within the proofs. From the definition of R_t and (5) we have

$$\Delta R_t = \ln L(\boldsymbol{\lambda}^{t-1}) - \ln L(\boldsymbol{\lambda}^t) \ge -\frac{1}{2} \ln(1 - \delta_t^2). \tag{6}$$

Proof Combining (6) 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 $R_0, \ldots, R_T$ are negative, then by monotonicity $R_T < 0$ and we are done. Otherwise, they are all non-negative. Then, applying Lemma 32 from the Appendix to the sequence $R_0, \ldots, R_T$, and using $c_2 \ge 1$ we get

$$R_T^{1-2c_2} \ge R_0^{1-2c_2} + \left(c_2 - \frac{1}{2}\right) B^{-2c_1} T \ge (1/2) B^{-2c_1} T \ge (\varepsilon \ln 2)^{1-2c_2} \implies R_T \le \varepsilon \ln 2.$$

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

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

where the second inequality uses $e^x \le 1 + (1/\ln 2)x$ 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 $\delta_t S_{t-1} \ge R_{t-1}$.*

Proof For any combination $\boldsymbol{\lambda}$, define $D_{\boldsymbol{\lambda}}$ 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 $L(\boldsymbol{\lambda}) \le L(\boldsymbol{\lambda}^*)$. By non-negativity of relative entropy we get

$$\begin{aligned} 0 \le \mathtt{RE}(D_{\boldsymbol{\lambda}^{t-1}} \parallel D_{\boldsymbol{\lambda}}) &= \sum_{i=1}^m D_{\boldsymbol{\lambda}^{t-1}}(i) \ln \left(\frac{e^{-(\mathbf{M}\boldsymbol{\lambda}^{t-1})_i} / mL(\boldsymbol{\lambda}^{t-1})}{e^{-(\mathbf{M}\boldsymbol{\lambda})_i} / mL(\boldsymbol{\lambda})} \right) \\ &\le -R_{t-1} + \sum_{i=1}^m D_{\boldsymbol{\lambda}^{t-1}}(i) \left(\mathbf{M}\boldsymbol{\lambda} - \mathbf{M}\boldsymbol{\lambda}^{t-1} \right)_i. \end{aligned} \tag{7}$$

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

$$\begin{aligned}\sum_{i=1}^{m} D_{\boldsymbol{\lambda}^{t-1}}(i) \sum_{j=1}^{N}\left(\lambda_j - \lambda_j^{t-1}\right) M_{ij} &= \sum_{j=1}^{N}\left(\lambda_j - \lambda_j^{t-1}\right)\sum_{i=1}^{m} D_t(i) M_{ij} \\ &\leq \left(\sum_{j=1}^{N}\left|\lambda_j - \lambda_j^{t-1}\right|\right)\max_j \left|\sum_{i=1}^{m} D_t(i) M_{ij}\right| \\ &= \delta_t \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t-1}\|_1. \end{aligned} \tag{8}$$

Since the previous holds for any $\boldsymbol{\lambda}$ suffering less than the target loss, we can choose in particular a $\boldsymbol{\lambda}$ so that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{t-1}\|_1$ is arbitrarily close to S_{t-1} showing that 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 will need that S_t is small compared to R_t. In fact we prove:

Lemma 4 *Whenever $R_0, \ldots, R_t$ and $S_0, \ldots, S_t$ are all strictly positive, $S_t \leq B^3 R_t^{-2}$.*

Before we prove Lemma 4, we show how to prove Theorem 1.

Proof [Of Theorem 1] Again if the positivity conditions on $R_0, \ldots, R_t$ and $S_0, \ldots, S_t$ do not hold, then the result is trivial. Thus, assume these quantities are positive at least for $t \leq 13\|\boldsymbol{\lambda}^*\|_1^6 \varepsilon^{-5}$. Combining Lemma 3 and Lemma 4 yields

$$\delta_t \geq B^{-3} R_{t-1}^3.$$

Notice this matches the condition of Lemma 2 for $c_1 = c_2 = 3$. Lemma 2 provides the desired bound on the number of rounds:

$$2(\varepsilon \ln 2)^{1-2\cdot 3} B^{2\cdot 3} < 13 B^6 \varepsilon^{-5}.$$

■

We still need to prove Lemma 4. The bound on S_t in Lemma 4 can be proved if we can first show S_t grows slowly compared to the rate at which the suboptimality R_t falls. Intuitively, when the edge δ_t is large, it leads to a large step size, causing growth in S_t, and a proportionately larger improvement in the suboptimality. To be precise, we prove the following about the relationship between S_t and R_t.

Lemma 5 *Whenever $R_0, \ldots, R_t$ and $S_0, \ldots, S_t$ are all strictly positive, 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 (6) and (4) 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,

$$\Delta R_t \geq |\alpha_t| \delta_t/2 \geq \delta_t \Delta S_t/2 \geq R_{t-1}\Delta S_t/2S_{t-1}.$$

Rearranging completes the proof. ■

Using this we prove Lemma 4.

Proof [Of Lemma 4] We first show $S_0 \leq B^3 R_0^{-2}$. Note, $S_0 \leq \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^0\|_1 = B$, so that it suffices to show $R_0 \leq B$. 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. ■

This completes the proof of Theorem 1. Although our bound provides a rate polynomial in B, ε^{-1} as desired by the conjecture in Schapire (2010), the exponents are rather large, and (we believe) not tight. One possible source of slack is the bound on S_t in Lemma 4. Qualitatively, the distance S_t to some solution having target loss should decrease with rounds, whereas Lemma 4 only says it does not increase too fast. Improving this will directly lead to a faster convergence rate. In particular, showing that S_t never increases would imply a B^2/ε rate of convergence. Whether or not the monotonicity of S_t holds, we believe that the obtained rate bound is probably true, and state it as a conjecture.

Conjecture 6 *For any $\boldsymbol{\lambda}^*$ and $\varepsilon > 0$, AdaBoost converges to within $L(\boldsymbol{\lambda}^*) + \varepsilon$ loss in $O(B^2/\varepsilon)$ rounds, where the order notation hides only absolute constants.*

As evidence supporting the conjecture, we show in the next section how a minor modification to AdaBoost can achieve the above rate.

3.2 Faster Rates For A Variant

In this section we introduce a new algorithm, AdaBoost.S , which will enjoy the much faster rate of convergence mentioned in Conjecture 6. 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. The code is largely the same as in Section 2, maintaining a combination $\boldsymbol{\lambda}^{t-1}$ of weak hypotheses, and greedily choosing α_t and $\hbar_{j_t}$ on each round to form a new combination $\tilde{\boldsymbol{\lambda}}^t = \boldsymbol{\lambda}^{t-1} + \alpha_t \hbar_{j_t}$. However, after creating the new combination $\tilde{\boldsymbol{\lambda}}^t$, the result is multiplied by the value s_t in $[0,1]$ that causes the greatest decrease in the exponential loss: $s_t = \operatorname{argmin}_s L(s\tilde{\boldsymbol{\lambda}}^t)$, and then we assign $\boldsymbol{\lambda}^t = s_t \tilde{\boldsymbol{\lambda}}^t$. Since $L(s\tilde{\boldsymbol{\lambda}}^t)$, as a function of s, is convex, its minimum on $[0,1]$ can be found easily, for instance, using a simple binary search. The new distribution D_{t+1} on the examples is constructed using $\boldsymbol{\lambda}^t$ as before; the weight $D_{t+1}(i)$ on example i is proportional to its exponential loss $D_{t+1}(i) \propto e^{-(\mathbf{M}\boldsymbol{\lambda}^t)_i}$. With this modification we prove the following:

Theorem 7 *For any* $\boldsymbol{\lambda}^*, \varepsilon > 0$*, AdaBoost.*S *achieves at most* $L(\boldsymbol{\lambda}^*) + \varepsilon$ *loss within* $3\|\boldsymbol{\lambda}^*\|_1^2/\varepsilon$ *rounds.*

The proof is similar to that in the previous section. Reusing the same notation, note that the proof of Lemma 2 continues to hold (with very minor modifications to that are straightforward). Next we can exploit the changes in AdaBoost.*S* to show an improved version of Lemma 3. Intuitively, scaling back has the effect of preventing the weights on the weak hypotheses from becoming "too large", and we may show

Lemma 8 *In each round* t*, the edge satisfies* $\delta_t \geq R_{t-1}/B$*.*

Proof We will reuse parts of the proof of Lemma 3. Setting $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ in (7) we may write

$$R_{t-1} \leq \sum_{i=1}^{m} D_{\boldsymbol{\lambda}^{t-1}}(i) \left(\mathbf{M}\boldsymbol{\lambda}^*\right)_i + \sum_{i=1}^{m} -D_{\boldsymbol{\lambda}^{t-1}}(i) \left(\mathbf{M}\boldsymbol{\lambda}^{t-1}\right)_i .$$

The first summation can be upper bounded as in (8) by $\delta_t \|\boldsymbol{\lambda}^*\| = \delta_t B$. We will next show that the second summation is non-positive, which will complete the proof. The scaling step was added just so that this last fact would be true.

If we define $G : [0,1] \to \mathbb{R}$ to be $G(s) = L\left(s\tilde{\boldsymbol{\lambda}}^t\right) = (1/m)\sum_i e^{-s(\mathbf{M}\tilde{\boldsymbol{\lambda}}^t)_i}$, then observe that the scaled derivative $G'(s)/G(s)$ is exactly equal to the second summation. Since $G(s) \geq 0$, it suffices to show the derivative $G'(s) \leq 0$ at the optimum value of s, denoted by s^*. Since G is a strictly convex function ($\forall s : G''(s) > 0$), it is either strictly increasing or strictly decreasing throughout $[0,1]$, or it has a local minimum. In the case when it is strictly decreasing throughout, then $G'(s) \leq 0$ everywhere. Otherwise, if G has a local minimum, then $G'(s) = 0$ at s^*. We finish the proof by showing that G cannot be strictly increasing throughout $[0,1]$. If it were, we would have $L(\tilde{\boldsymbol{\lambda}}^t) = G(1) > G(0) = 1$, an impossibility since the loss decreases through rounds. ■

The above lemma implies the conditions in Lemma 2 hold if we set $c_1 = c_2 = 1$. The result of Lemma 2 now implies Theorem 7, where we used that $2\ln 2 < 3$.

In experiments we ran, the scaling back never occurs. For such data sets, AdaBoost and AdaBoost.*S* are identical. We believe that even for contrived examples, the rescaling could happen only a few times, implying that both AdaBoost and AdaBoost.*S* would enjoy the convergence rates of Theorem 7. In the next section, we construct rate lower bound examples to show that this is nearly the best rate one can hope to show.

3.3 Lower Bounds

Here we show that the dependence of the rate in Theorem 1 on the norm $\|\boldsymbol{\lambda}^*\|_1$ of the solution achieving target accuracy is necessary for a wide class of data sets. The arguments in this section are not tailored to AdaBoost, but hold more generally for any coordinate descent algorithm, and can be readily generalized to any loss function $\tilde{L}$ of the form $\tilde{L}(\boldsymbol{\lambda}) = (1/m)\sum_i \phi(-\mathbf{M}\boldsymbol{\lambda})$, where $\phi : \mathbb{R} \to \mathbb{R}$ is any non-decreasing function. For instance, with the exponential loss, ϕ is the exponential function, and the lower-bound results have a logarithmic term in them. For general non-decreasing functions ϕ, similar arguments yield bounds which are identical to the ones in this section, except the logarithmic terms are replaced by the inverse function ϕ^{-1}.

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

Lemma 9 *Suppose the feature matrix* $\mathbf{M}$ *corresponding to a data set has two rows with* $\{-1,+1\}$ *entries which 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

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

and the statement of the lemma directly follows. ■

When the weak hypotheses are *abstaining* (Schapire and Singer, 1999), a hypothesis can make a definitive prediction that the label is -1 or $+1$, or it can "abstain" by predicting zero. No other levels of confidence are allowed, and the resulting feature matrix has entries in $\{-1,0,+1\}$. The next theorem constructs a feature matrix satisfying the properties of Lemma 9 and where additionally the smallest size of a solution achieving $L^* + \varepsilon/m$ loss is at least $\Omega(2^m)\ln(1/\varepsilon)$, for some fixed L^* and every $\varepsilon > 0$.

Theorem 10 *Consider the following matrix* $\mathbf{M}$ *with* m *rows (or examples) (for* $m \geq 2$*) labeled* $0,\ldots,m-1$ *and* $m-1$ *columns labeled* $1,\ldots,m-1$*. 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* $2/m + \varepsilon/m$ *is achievable on this data set, but with large norms*

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

Therefore, by Lemma 9, the minimum number of rounds required for reaching loss at most $2/m+\varepsilon$ *is at least* $\left(\frac{2^{m-2}-1}{2\ln m}\right)\ln(1/\varepsilon)$.

A picture of the matrix constructed in the above lemma for $m = 6$ is shown in Figure 2. Theorem 10 shows that when ε is a small constant (say $\varepsilon = 0.01$), and $\boldsymbol{\lambda}^*$ is some vector with loss $L^* + \varepsilon/(2m)$, AdaBoost takes at least $\Omega(2^m/\ln m)$ steps to get within $\varepsilon/(2m)$ of the loss achieved by $\boldsymbol{\lambda}^*$. Since m and ε are independent quantities, this shows that worst case bounds can be exponential in m. Thus, a polynomial dependence on the norm of the reference solution is unavoidable, and this norm might be exponential in the number of training examples in the worst case.

$$\begin{array}{r} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ m-1=5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5=m-1 \end{array} \\ \begin{pmatrix} - & + & + & + & + \\ + & - & - & - & - \\ 0 & + & - & - & - \\ 0 & 0 & + & - & - \\ 0 & 0 & 0 & + & - \\ 0 & 0 & 0 & 0 & + \end{pmatrix} \end{array}$$

Figure 2: The matrix used in Theorem 10 when $m = 6$. Note that $+$ is the same as $+1$, referring to a correct prediction, and $-$ the same as -1, referring to an incorrect prediction.

Corollary 11 *Consider feature matrices containing only* $\{-1,0,+1\}$ *entries. If, for some constants* c *and* β*, the bound in Theorem 1 can be replaced by* $O\left(\|\boldsymbol{\lambda}^*\|_1^c \varepsilon^{-\beta}\right)$ *for all such matrices, then* $c \geq 1$*. Further, for such matrices, the bound* $\text{poly}(1/\varepsilon, \|\boldsymbol{\lambda}^*\|_1)$ *in Theorem 1 cannot be replaced by a uniform* $\text{poly}(1/\varepsilon, m, N)$ *bound that holds for all values of* m, ε *and* N.

We now prove Theorem 10.

Proof of Theorem 10. We first lower bound the norm of solutions achieving loss at most $2/m + \varepsilon$. Observe that since rows 0 and 1 are complementary, then $-(\mathbf{M}\boldsymbol{\lambda})_0 = (\mathbf{M}\boldsymbol{\lambda})_1$, and any solution's loss on just examples 0 and 1 will add up to $1/m[\exp(-(\mathbf{M}\boldsymbol{\lambda})_1) + \exp(-(\mathbf{M}\boldsymbol{\lambda})_2)]$ which is at least $2/m$. Therefore, to get within $2/m + \varepsilon/m$, the margins on examples $2, \ldots, m-1$ should be at least $\ln(1/\varepsilon)$. Now, the feature matrix is designed so that the margins due to a combination $\boldsymbol{\lambda}$ satisfy the following recursive relationships:

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

Therefore, the margin of example $m-1$ is at least $\ln(1/\varepsilon)$, then by the first equation above, $\lambda_{m-1} \geq \ln(1/\varepsilon)$. Similarly, $\lambda_{m-2} \geq \ln(1/\varepsilon) + \lambda_{m-1} \geq 2\ln(1/\varepsilon)$. Continuing this way, if the margin of example i is at least $\ln(1/\varepsilon)$, then

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

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

We end by showing that a loss of at most $2/m + \varepsilon/m$ 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 $(2^{m-2} - 1, 2^{m-3}, 2^{m-4}, \ldots, 1)$ achieves margins $(0, 0, 1, 1, 1, \ldots, 1)$. Hence when scaled by a factor of $\ln((m-2)/\varepsilon)$, this combination achieves a loss $\left(2 + (m-2)\frac{\varepsilon}{m-2}\right)/m = 2/m + \varepsilon/m$, for any $\varepsilon > 0$. ■

We finally show that if the weak hypotheses are confidence-rated with arbitrary levels of confidence, so that the feature matrix is allowed to have non-integral entries in $[-1, +1]$, then the

$$\begin{pmatrix} -1 & +1 \\ +1 & -1 \\ -1+\nu & +1 \\ +1 & -1+\nu \end{pmatrix}$$

Figure 3: A picture of the matrix used in Theorem 12.

minimum norm of a solution achieving a fixed accuracy can be arbitrarily large. Our constructions will satisfy the requirements of Lemma 9, so that the norm lower bound translates into a rate lower bound.

Theorem 12 *Let* $\nu > 0$ *be an arbitrary number, and let* $\mathbf{M}$ *be the (possibly) non-integral matrix with 4 examples and 2 weak hypotheses shown in Figure 3. Then for any* $\varepsilon > 0$*, a loss of* $1/2+\varepsilon$ *is achievable on this data set, but with large norms*

$$\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq 1/2+\varepsilon\} \geq 2\ln(1/(4\varepsilon))\nu^{-1}.$$

Therefore, by Lemma 9, the number of rounds required to achieve loss at most $1/2+\varepsilon$ *is at least* $\ln(1/(4\varepsilon))\nu^{-1}/\ln(m)$.

Proof We first show a loss of $1/2+\varepsilon$ is achievable. Observe that the vector $\boldsymbol{\lambda} = (c,c)$, with $c = \nu^{-1}\ln(1/(2\varepsilon))$, achieves margins $0,0,\ln(1/(2\varepsilon)),\ln(1/(2\varepsilon))$ on examples $1,2,3,4$, respectively. Therefore $\boldsymbol{\lambda}$ achieves loss $1/2+\varepsilon$. We next show a lower bound on the norm of a solution achieving this loss. Observe that since the first two rows are complementary, the loss due to just the first two examples is at least $1/2$. Therefore, any solution $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ achieving at most $1/2+\varepsilon$ loss overall must achieve a margin of at least $\ln(1/(4\varepsilon))$ on both the third and fourth examples. By inspecting these two rows, this implies

$$\lambda_2 - \lambda_1 + \lambda_1\nu \geq \ln(1/(4\varepsilon)),$$
$$\lambda_1 - \lambda_2 + \lambda_2\nu \geq \ln(1/(4\varepsilon)).$$

Adding the two equations we find

$$\nu(\lambda_1+\lambda_2) \geq 2\ln(1/(4\varepsilon)) \implies \lambda_1+\lambda_2 \geq 2\nu^{-1}\ln(1/(4\varepsilon)).$$

By the triangle inequality, $\|\boldsymbol{\lambda}\|_1 \geq \lambda_1+\lambda_2$, and the lemma follows. ∎

Note that if $\nu = 0$, then an optimal solution is found in zero rounds of boosting and has optimal loss 1. However, even the tiniest perturbation $\nu > 0$ causes the optimal loss to fall to $1/2$, and causes the rate of convergence to increase drastically. In fact, by Theorem 12, the number of rounds required to achieve any fixed loss below 1 grows as $\Omega(1/\nu)$, which is arbitrarily large when ν is infinitesimal. We may conclude that with non-integral feature matrices, the dependence of the rate on the norm of a reference solution is absolutely necessary.

Corollary 13 *When using confidence rated weak-hypotheses with arbitrary confidence levels, the bound* $\mathrm{poly}(1/\varepsilon, \|\boldsymbol{\lambda}^*\|_1)$ *in Theorem 1 cannot be replaced by any function of purely m, N and* ε *alone.*

The construction in Figure 3 can be generalized to produce data sets with any number of examples that suffer the same poor rate of convergence as the one in Theorem 12. We discussed the smallest such construction, since we feel that it best highlights the drastic effect non-integrality can have on the rate.

In this section we saw how the norm of the reference solution is an important parameter for bounding the convergence rate. In the next section we investigate the optimal dependence of the rate on the parameter ε and show that $\Omega(1/\varepsilon)$ rounds are necessary in the worst case.

4. Second Convergence Rate: 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 data sets, 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 $\text{poly}(1/\varepsilon_2, \|\boldsymbol{\lambda}^*\|_1)$ 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. Additionally, we show that this dependence on ε is optimal in Lemma 31 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 data set. 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.

4.1 Upper Bound

The main result of this section is the following rate upper bound. A similar approach to solving this problem was taken independently by Telgarsky (2011).

Theorem 14 *AdaBoost reaches within ε of the optimal loss in at most C/ε rounds, where C only depends on the feature matrix.*

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 data set used in Lemma 31 and shown in Figure 4 exemplifies this situation. Our main technical contribution shows that the examples in any data set 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 14.

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 14.

Lemma 15 *(Decomposition Lemma) For any data set, 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 *$F \triangleq Z^c = X \setminus Z$ such that the following hold simultaneously :*

1. *There exists some positive constant $\gamma > 0$, and 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 $\boldsymbol{\eta}^*$. (Note that $\boldsymbol{\eta}^*$ may not be unique. There may be a whole subspace of vectors like $\boldsymbol{\eta}^*$ that achieve the optimal loss on F.)*

3. *There is a constant $\mu_{\max} < \infty$, such that for any combination $\boldsymbol{\eta}$ with bounded loss on the finite-margin set, specifically obeying $\sum_{i \in F} e^{-(\mathbf{M}\boldsymbol{\eta})_i} \leq m$, 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 4: Matrix $\mathbf{M}$ for a data set requiring $\Omega(1/\varepsilon)$ rounds for convergence.

Before proceeding, we give an example data set and indicate the zero-loss set, finite-margin set, $\boldsymbol{\eta}^*$ and $\boldsymbol{\eta}^\dagger$ to illustrate our definitions. Consider a data set with three examples $\{a,b,c\}$ and two hypotheses $\{\hbar_1, \hbar_2\}$ and the feature matrix $\mathbf{M}$ in Figure 4. 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 data set these are unique. This data set also serves as a lower-bound example in Lemma 31, where we show that $2/(9\varepsilon)$ rounds are necessary for AdaBoost to achieve loss at most $(2/3) + \varepsilon$, where the infimum of the loss is 2/3. Observe that this data set is similar to the data set in Theorem 10 in that the zero-loss set (corresponding to the complementary rows) and finite-loss sets (remaining examples) are both non-empty.

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 say at most $\inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) + \varepsilon/m$ loss. 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 14. The progress due to the zero-loss set is now immediate from Item 1 of the decomposition lemma:

Lemma 16 *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

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

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

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

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

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

■

If the set F were empty, then Lemma 16 implies an edge of γ is available in each round. This in fact means that the weak learning assumption holds, and using (5), 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). This is in fact the "separable" case where positive and negative examples can be separated by AdaBoost. Henceforth, we assume that F is non-empty. This implies that the optimal loss K is at least 1 (since we are in the nonseparable case, and any solution will get non-positive margin on some example in F), a fact that we will use later in the proofs. In the separable case, the minimum normalized margin becomes an important performance instead of the exponential loss. Convergence with respect to the normalized margin has been studied by Schapire et al. (1998), Grove and Schuurmans (1998), Rudin et al. (2004), Rätsch and Warmuth (2005), and others.

Lemma 16 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 17 and 18 together show how AdaBoost can make good progress using the finite margin set. Lemma 17 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 17 *Suppose $\boldsymbol{\lambda}$ is a combination such that $m \geq \ell^{\boldsymbol{\lambda}}(F) \geq K$. Then in some coordinate direction $\vec{e}_j$ the edge $\delta_F(\vec{e}_j;\boldsymbol{\lambda})$ 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 $\{\vec{u} : \mathbf{M}_F\vec{u}=\mathbf{0}\}$ 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 $\mathbf{M}_F^T\mathbf{M}_F$ (which 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 $[\boldsymbol{\eta}^*,\boldsymbol{\lambda}]$

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

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:

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

Since $f''(x)\geq 0$, f is convex; further it attains the minimum value at 0. Therefore f is increasing, and $f'(x)\geq 0$ at all points. We next lower bound possible values of the second derivative as follows:

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

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 $\ell^{(\boldsymbol{\eta}^*-x\boldsymbol{\eta})}(i)=\exp\{-(\mathbf{M}_F(\boldsymbol{\eta}^*-x\boldsymbol{\eta}))_i\}\geq e^{-\mu_{\max}}$ on every example i. Therefore we have

$$f''(x)\geq\|\mathbf{M}_F\boldsymbol{\eta}\|^2e^{-\mu_{\max}}\geq\lambda_{\min}^2e^{-\mu_{\max}}\|\boldsymbol{\eta}\|^2 \text{ (by choice of } \boldsymbol{\eta}\text{)}. \tag{9}$$

A standard second-order result for a twice-differentiable function f is (see, e.g., Boyd and Vandenberghe, 2004, Equation (9.9))

$$\left|f'(1)\right|^2 \geq 2\left(\inf_{x\in[0,1]} f''(x)\right)(f(1)-f(0)). \tag{10}$$

Collecting our results so far, including the definition of f' and the fact that it is non-negative, and equations (9) and (10), we get

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

Next let $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}/\|\boldsymbol{\eta}\|_1$ be $\boldsymbol{\eta}$ rescaled to have unit ℓ_1 norm. Then dividing the previous equation by $\|\boldsymbol{\eta}\|_1$

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

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

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

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 $\left|\tilde{\boldsymbol{\eta}}_j\right|$ on column j, then we have

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

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

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

where we define

$$C_0 = 2\lambda_{\min}^2 N^{-1}m^{-1}e^{-\mu_{\max}}. \tag{11}$$

■

Lemma 18 *There are constants C_1, C_2 depending only on the feature matrix M with the following property. Suppose, at any 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; that is, the loss after one more round is $K+\theta-\Delta\theta$. In this situation, if $\ell^{\boldsymbol{\lambda}}(Z) \leq C_1\theta$, for $C_1 \leq 1$, then $\Delta\theta \geq C_2\theta$.*

Proof Let $\boldsymbol{\lambda}$ be the current solution found by boosting. Using Lemma 17, pick a direction j in which the edge $\delta_F(\vec{e}_j;\boldsymbol{\lambda})$ restricted to the finite margin set is at least

$$\delta_F(\vec{e}_j;\boldsymbol{\lambda}) \geq \sqrt{C_0(\ell^{\boldsymbol{\lambda}}(F)-K)/\ell^{\boldsymbol{\lambda}}(F)}. \tag{12}$$

We can bound the edge $\delta_X(\vec{e}_j;\boldsymbol{\lambda})$ on the entire set of examples as follows:

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

We can bound the first summation as

$$\left|\sum_{i\in F}\ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\vec{e}_j)_i\right| = \ell^{\boldsymbol{\lambda}}(F)\delta_F(\vec{e}_j;\boldsymbol{\lambda}) \geq \sqrt{C_0(\ell^{\boldsymbol{\lambda}}(F)-K)\ell^{\boldsymbol{\lambda}}(F)} \text{ (by (12))}.$$

We can bound the second summation as

$$\left|\sum_{i\in Z}\ell^{\boldsymbol{\lambda}}(i)(\mathbf{M}\vec{e}_j)_i\right| \leq \sum_{i\in Z}\ell^{\boldsymbol{\lambda}}(i)\|\mathbf{M}\|_\infty \leq \ell^{\boldsymbol{\lambda}}(Z).$$

Substituting the two bounds above,

$$\delta_X(\vec{e}_j;\boldsymbol{\lambda}) \geq \frac{1}{\ell^{\boldsymbol{\lambda}}(X)}\left(\sqrt{C_0(\ell^{\boldsymbol{\lambda}}(F)-K)\ell^{\boldsymbol{\lambda}}(F)} - \ell^{\boldsymbol{\lambda}}(Z)\right). \tag{13}$$

Now, assume as in the statement of the lemma that $\ell^{\boldsymbol{\lambda}}(Z) \leq C_1\theta$. Thus, $\ell^{\boldsymbol{\lambda}}(F) - K = \theta - \ell^{\boldsymbol{\lambda}}(Z) \geq (1-C_1)\theta$. Set $C_1 = \min\left\{1/4, (1/4)\sqrt{C_0/m}\right\}$. Then we have

$$\begin{aligned}
\sqrt{C_0(\ell^{\boldsymbol{\lambda}}(F)-K)\ell^{\boldsymbol{\lambda}}(F)} &\geq \sqrt{C_0(1-C_1)\theta} \\
\ell^{\boldsymbol{\lambda}}(Z) &\leq C_1\theta.
\end{aligned}$$

The last inequality holds because $\theta \leq m$, which follows since the total loss $K+\theta$ at any stage of boosting is less than the initial loss m. Using the above, we simplify (13) as follows:

$$\delta_X(\vec{e}_j;\boldsymbol{\lambda}) \geq \frac{1}{\ell^{\boldsymbol{\lambda}}(X)}\left(\sqrt{C_0(\ell^{\boldsymbol{\lambda}}(F)-K)\ell^{\boldsymbol{\lambda}}(F)} - \ell^{\boldsymbol{\lambda}}(Z)\right) \geq \frac{1}{K+\theta}\left(\sqrt{C_0(1-C_1)\theta} - C_1\theta\right)$$

Using $\theta \leq m$, we can bound the square of the term in brackets on the previous line as

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

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

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

where, for the last inequality, we again used $K+\theta \leq m$. Therefore, from (5), 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 14. 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 data set (and not θ). To see this, either $\ell^{\boldsymbol{\lambda}}(Z) \leq C_1\theta$, in which case Lemma 18 applies, and $\Delta\theta \geq C_2\theta \geq (C_2/m)\theta^2$ (since $\theta = \ell^{\boldsymbol{\lambda}}(X) - K \leq m$). Or $\ell^{\boldsymbol{\lambda}}(Z) > C_1\theta$, in which case applying Lemma 16 yields $\delta \geq \gamma C_1\theta/\ell^{\boldsymbol{\lambda}}(X) \geq (\gamma C_1/m)\theta$. By (5), $\Delta\theta \geq \ell^{\boldsymbol{\lambda}}(X)(1-\sqrt{1-\delta^2}) \geq \ell^{\boldsymbol{\lambda}}(X)\delta^2/2 \geq (K/2)(\gamma C_1/m)^2\theta^2$. Using $K \geq 1$ and choosing C_3 appropriately as $\min\left\{C_2/m, (1/2)(\gamma C_1/m)^2\right\}$ gives the required condition for $\Delta\theta \geq C_3\theta^2$. Note that plugging in the estimates for C_1 and C_2 from the proof of Lemma 18 yields

$$C_3 \geq \max\left\{\frac{C_0}{4m^2}, \frac{\gamma^2 C_0}{32m^3}\right\} = \frac{C_0}{4m^2}\max\left\{1, \frac{\gamma^2}{8m}\right\}. \tag{14}$$

If $K+\theta_t$ denotes the loss in round t, then the above claim implies $\theta_t - \theta_{t+1} \geq C_3\theta_t^2$. We will show that $T = C_3^{-1}/\varepsilon$ rounds suffice for the loss to be at most ε, that is, $\theta_T \leq \varepsilon$. Since θ_t is non-increasing and non-negative, if it is non-positive at any point before T iterations, then the bound follows trivially. So assume θ_t's are positive for $t \leq T$. Applying Lemma 32 to the sequence $\{\theta_t\}$ we have $1/\theta_T - 1/\theta_0 \geq C_3 T = 1/\varepsilon$. Since $\theta_0 \geq 0$, we have $\theta_T \leq \varepsilon$, completing the proof. ■

4.2 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 these are the ones found by boosting). We prove Lemma 15 in three steps. We begin with a simple lemma that rigorously defines the zero-loss and finite-margin sets.

Lemma 19 *For any infinite 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 whose losses are bounded away from zero in its complement $X \setminus Z$(the finite-margin set)*

$$\forall x \in Z : \lim_{t\to\infty} \ell^{\boldsymbol{\eta}_{(t)}}(x) = 0, \qquad\qquad \forall x \in X \setminus Z : \inf_t \ell^{\boldsymbol{\eta}_{(t)}}(x) > 0. \tag{15}$$

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. ■

Lemma 19 can be applied to any sequence to yield a new sequence with respect to which the examples can be decomposed into zero-loss and finite-margin sets satisfying (15). This way we can get *nicer* sequences out of ones with possibly complicated tail behavior. The next lemma shows, that given such a nice sequence, one may extract a single vector that satisfies properties similar to those required in Item 1 of the decomposition lemma.

Lemma 20 *Let* $\mathbf{M}$ *be a feature matrix, with rows indexed by examples in* X*. Suppose* 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 margin set, that is,* (15) *holds with respect to the entire sequence itself. 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:*

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

Proof Firstly assume Z is non-empty, since otherwise setting $\boldsymbol{\eta}^\dagger$ to the zero-vector proves the lemma. Since the $\boldsymbol{\eta}_{(t)}$ achieve arbitrarily large positive margins on Z, the sequence $\|\boldsymbol{\eta}_{(t)}\|$ 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 $|X|$. To be more precise, the lemma makes an assertion about feature matrices M, whose rows are indexed by the set X. We will prove this assertion for all feature matrices M by induction on the number $|X|$ of rows it contains. If X is empty, then the lemma holds vacuously for any $\boldsymbol{\eta}^\dagger$. Assume the statement of the lemma holds inductively for all subsets of X of size less than $m > 0$, meaning that a vector analogous to $\boldsymbol{\eta}^\dagger$ exists for the subset, and consider X of size m.

First, we find a unit vector $\boldsymbol{\eta}'$ that we will show has a positive margin on a non-empty subset S of Z and zero margins on X/Z. Since translating a vector along the null space of $\mathbf{M}$, $\ker\mathbf{M} = \{\vec{x} : \mathbf{M}\vec{x} = \mathbf{0}\}$, has no effect on the margins produced by the vector, assume without loss of generality that the $\boldsymbol{\eta}_{(t)}$'s are orthogonal to $\ker\mathbf{M}$. Since the margins produced on Z, which we have assumed is non-empty, 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, which is 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 (15). 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 feature matrix restricted to 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 $S \subseteq Z$) $= X' \setminus Z'$. The set X' is smaller than the set X, and thus the inductive hypothesis holds for X', meaning that there exists some $\boldsymbol{\eta}''$ that achieves positive margins on every element in Z', and zero margins on every element in $X' \setminus Z' = X \setminus Z$. Therefore, by setting $\boldsymbol{\eta}^\dagger = \boldsymbol{\eta}' + c\boldsymbol{\eta}''$ for a suitable c, we can achieve a positive margin on every element in $S \cup Z' = Z$, and zero margins on every element in $X \setminus Z$, completing the proof. ■

We may now use the previous two results to prove Item 1 of the decomposition lemma. First, we apply Lemma 19 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 $F = X \setminus Z$. Now, applying Lemma 20 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 21 *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 20 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 $\ker \mathbf{M}_F = \{\vec{x} : \mathbf{M}_F \vec{x} = 0\}$ 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 $K = |F|$ on F. Otherwise, let λ^2 be the smallest positive eigenvalue of the symmetric matrix $\mathbf{M}_F^T \mathbf{M}_F$. Then $\|\mathbf{M}\boldsymbol{\eta}^{(t)}\| \geq \lambda \|\boldsymbol{\eta}^{(t)}\|$. By the definition of the finite margin set, $\inf_t \min_{i \in F} \ell^{\boldsymbol{\eta}^{(t)}}(i) = \inf_t \min_{i \in F} \ell^{\boldsymbol{\eta}^*_{(t)}}(i) > 0$. 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 22 *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 arbitrarily large elements are present in the set $\{(\mathbf{M}\boldsymbol{\eta})_i : \ell^{\boldsymbol{\eta}}(F) \leq m, 1 \leq i \leq m\}$. 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 19 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 20 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 21, 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. ■

The proof of the decomposition theorem is complete.

4.3 Investigating The Constants

In this section, we try to estimate the constant C in Theorem 14. We show that it can be arbitrarily large for adversarial feature matrices with real entries (corresponding to confidence rated weak hypotheses), but has an upper-bound doubly exponential in the number of examples when the feature matrix has $\{-1,0,+1\}$ entries only. We also show that this doubly exponential bound cannot be improved without significantly changing the proof in the previous section.

By inspecting the proofs, in particular equations (11) and (14), and seeing that $1/C_3$ is C from Theorem 14, we can bound the constant in Theorem 14 as follows.

Corollary 23 *The constant C in Theorem 14 that emerges from the proofs is*

$$C \le \frac{2m^3 N e^{\mu_{\max}}}{\lambda_{\min}^2} \max\left\{1, \frac{8m}{\gamma^2}\right\},$$

where m is the number of examples, N is the number of hypotheses, γ *and* $\mu_{\max}$ *are as given by Items 1 and 3 of the decomposition lemma, and* $\lambda_{\min}^2$ *is the smallest positive eigenvalue of* $\mathbf{M}_F^T\mathbf{M}_F$ *(*$\mathbf{M}_F$ *is the feature matrix restricted to the rows belonging to the finite margin set F).*

Our bound on C will be obtained by in turn bounding the quantities $\lambda_{\min}^{-1}, \gamma^{-1}, \mu_{\max}$. These are strongly related to the singular values of the feature matrix $\mathbf{M}$, and in general cannot be easily measured. In fact, when $\mathbf{M}$ has real entries, we have already seen in Section 3.3 that the rate can be arbitrarily large, implying these parameters can have very large values. Even when the matrix $\mathbf{M}$ has integer entries (that is, $-1,0,+1$), the next lemma shows that these quantities can be exponential in the number of examples.

Lemma 24 *There are examples of feature matrices with* $-1,0,+1$ *entries and at most m rows or columns (where* $m > 10$*) for which the quantities* $\gamma^{-1}, \lambda^{-1}$ *and* $\mu_{\max}$ *are at least* $\Omega(2^m/m)$*.*

Proof We first show the bounds for γ and λ. Let $\mathbf{M}$ be an $m \times m$ upper triangular matrix with $+1$ on the diagonal, and -1 above the diagonal. Let $\vec{y} = (2^{m-1}, 2^{m-2}, \ldots, 1)^T$, and $\mathbf{b} = (1,1,\ldots,1)^T$. Then $\mathbf{M}\vec{y} = \mathbf{b}$, although the $\vec{y}$ has much bigger norm than $\mathbf{b}$: $\|\vec{y}\| \ge 2^{m-1}$, while $\|\mathbf{b}\| = m$. Since $\mathbf{M}$ is invertible, by the definition of $\lambda_{\min}$, we have $\|\mathbf{M}\vec{y}\| \ge \lambda_{\min}\|\vec{y}\|$, so that $\lambda_{\min}^{-1} \ge \|\vec{y}\|/\|\mathbf{M}\vec{y}\| \ge 2^m/m$. Next, note that $\vec{y}$ produces all positive margins $\mathbf{b}$, and hence the zero-loss set consists of all the examples. In particular, if $\boldsymbol{\eta}^\dagger$ is as in Item 1 of the decomposition lemma, then the vector $\gamma^{-1}\boldsymbol{\eta}^\dagger$ achieves margin greater than 1 on each example: $\mathbf{M}(\gamma^{-1}\boldsymbol{\eta}^\dagger) \ge \mathbf{b}$. On the other hand, our matrix is very similar to the one in Theorem 10, and the same arguments in the proof of that theorem can be used to show that if for some $\vec{x}$ we have $(\mathbf{M}\vec{x}) \ge \mathbf{b}$ entry-wise, then $\vec{x} \ge \vec{y}$. This implies that $\gamma^{-1}\|\boldsymbol{\eta}^\dagger\|_1 \ge \|\vec{y}\|_1 = (2^m - 1)$. Since $\boldsymbol{\eta}^\dagger$ has unit ℓ_1-norm, the bound on γ^{-1} follows too.

Next we provide an example showing $\mu_{\max}$ can be $\Omega(2^m/m)$. Consider an $m \times (m-1)$ matrix $\mathbf{M}$. The bottom row of $\mathbf{M}$ is all $+1$. The upper $(m-1)\times(m-1)$ submatrix of $\mathbf{M}$ is a lower triangular matrix with -1 on the diagonal and $+1$ below the diagonal. Observe that if $\vec{y}^T = (2^{m-2}, 2^{m-3}, \ldots, 1, 1)$, then $\vec{y}^T\mathbf{M} = \mathbf{0}$. Therefore, for any vector $\vec{x}$, the inner product of the margins $\mathbf{M}\vec{x}$ with $\vec{y}$ is zero: $\vec{y}^T\mathbf{M}\vec{x} = 0$. This implies that achieving positive margin on any example forces some other example to receive negative margin. By Item 1 of the decomposition lemma, the zero loss set in this data set is empty since there cannot be an $\boldsymbol{\eta}^\dagger$ with both positive and zero margins and no negative margins. Thus, all the examples belong to the finite margin set. Next, we choose a combination with at most m loss that nevertheless achieves $\Omega(2^m/m)$ positive margin on some example. Let $\vec{x}^T = (1,2,4,\ldots,2^{m-2})$. Then $(\mathbf{M}\vec{x})^T = (-1,-1,\ldots,-1,2^{m-1}-1)$. Then the margins using $\varepsilon\vec{x}$ are $(-\varepsilon,\ldots,-\varepsilon,\varepsilon(2^{m-1}-1))$ with total loss $(m-1)e^{\varepsilon} + e^{\varepsilon(1-2^{m-1})}$. Choose $\varepsilon = 1/(2m) \le 1$, so that the loss on examples corresponding to the first $m-1$ rows is at most $e^{\varepsilon} \le 1 + 2\varepsilon = 1 + 1/m$, where the first inequality holds since $\varepsilon \in [0,1]$. For $m > 10$, the choice of ε guarantees $1/(2m) = \varepsilon \geq (\ln m)/(2^{m-1}-1)$, so that the loss on the example corresponding to the bottom most row is $e^{-\varepsilon(2^{m-1}-1)} \leq e^{-\ln m} = 1/m$. Therefore the net loss of $\varepsilon\vec{x}$ is at most $(m-1)(1+1/m)+1/m = m$. On the other hand the margin of the example corresponding to the last row is $\varepsilon(2^{m-1}-1) = (2^{m-1}-1)/(2m) = \Omega(2^m/m)$. ■

The above result implies any bound on C derived from Corollary 23 will be at least $2^{\Omega(2^m/m)}$ in the worst case. This does not imply that the best bound one can hope to prove is doubly exponential, only that our techniques in the previous section do not admit anything better. We next show that the bounds in Lemma 24 are nearly the worst possible.

Lemma 25 *Suppose each entry of* $\mathbf{M}$ *is* $-1, 0$ *or* $+1$. *Then each of the quantities* $\lambda_{\min}^{-1}, \gamma^{-1}$ *and* $\mu_{\max}$ *are at most* $2^{O(m\ln m)}$.

The proof of Lemma 25 is rather technical, and we defer it to the Appendix. Lemma 25 and Corollary 23 together imply a convergence rate of $2^{2^{O(m\ln m)}}/\varepsilon$ to the optimal loss for integer matrices. This bound on C is exponentially worse than the $\Omega(2^m)$ lower bound on C we saw in Section 3.3, a price we pay for obtaining optimal dependence on ε. In the next section we will see how to obtain $\text{poly}(2^{m\ln m}, \varepsilon^{-1})$ bounds, although with a worse dependence on ε. We end this section by showing, just for completeness, how a bound on the norm of $\boldsymbol{\eta}^*$ as defined in Item 2 of the decomposition lemma follows as a quick corollary to Lemma 25.

Corollary 26 *Suppose* $\boldsymbol{\eta}^*$ *is as given by Item 2 of the decomposition lemma. When the feature matrix has only* $-1, 0, +1$ *entries, we may bound* $\|\boldsymbol{\eta}^*\|_1 \leq 2^{O(m\ln m)}$.

Proof Note that every entry of $\mathbf{M}_F\boldsymbol{\eta}^*$ lies in the range $[-\ln m, \mu_{\max} = 2^{O(m\ln m)}]$, and hence $\|\mathbf{M}_F\boldsymbol{\eta}^*\| \leq 2^{O(m\ln m)}$. Next, we may choose $\boldsymbol{\eta}^*$ orthogonal to the null space of $\mathbf{M}_F$; then $\|\boldsymbol{\eta}^*\| \leq \lambda_{\min}^{-1}\|\mathbf{M}_F\boldsymbol{\eta}^*\| \leq 2^{O(m\ln m)}$. Since $\|\boldsymbol{\eta}^*\|_1 \leq \sqrt{N}\|\boldsymbol{\eta}^*\|$, and the number of possible columns N with $\{-1, 0, +1\}$ entries is at most 3^m, the proof follows. ■

5. Improved Estimates

The goal of this section is to show how the ideas introduced in the paper can be applied in ways other than presented so far to produce new and stronger results. By combining techniques from Sections 3 and 4, we obtain both new upper bounds for convergence to the optimal loss, as well as more general lower bounds for convergence to an arbitrary target loss. We also indicate what we believe might be the optimal bounds for either situation.

We first show how the finite rate bound of Theorem 1 along with the decomposition lemma yields a new rate of convergence to the optimal loss. The proof includes choosing a useful target $\boldsymbol{\lambda}^*$ for Theorem 1. Although the dependence on ε is worse than in Theorem 14, the dependence on m is nearly optimal. We will need the following key application of the decomposition lemma.

Lemma 27 *When the feature matrix has* $-1, 0, +1$ *entries, for any* $\varepsilon > 0$, *there is some solution with* ℓ_1*-norm at most* $2^{O(m\ln m)}\ln(1/\varepsilon)$ *that achieves within* ε *of the optimal loss.*

Proof Let $\boldsymbol{\eta}^*, \boldsymbol{\eta}^\dagger, \gamma$ be as given by the decomposition lemma. Let $c = \min_{i\in Z}(\mathbf{M}\boldsymbol{\eta}^*)_i$ be the minimum margin produced by $\boldsymbol{\eta}^*$ on any example in the zero-loss set Z. Then $\boldsymbol{\eta}^* - c\boldsymbol{\eta}^\dagger$ produces non-negative margins on Z, since $\mathbf{M}\boldsymbol{\eta}^* - c\mathbf{M}\boldsymbol{\eta}^\dagger \geq \mathbf{0}$, and it attains the optimal margins on the finite margin set F, since $\mathbf{M}\boldsymbol{\eta}^\dagger = \mathbf{0}$ on F. Therefore, the vector $\boldsymbol{\lambda}^* = \boldsymbol{\eta}^* + \left(\ln(1/\varepsilon)\gamma^{-1} - c\right)\boldsymbol{\eta}^\dagger$ achieves at least $\ln(1/\varepsilon)$ margin on every example in Z, and optimal margins on the finite loss set F. Hence $L(\boldsymbol{\lambda}^*) \leq \inf_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) + \varepsilon$. Using $|c| \leq \|\mathbf{M}\boldsymbol{\eta}^*\| \leq m\|\boldsymbol{\eta}^*\|$,

$$\begin{aligned}\|\boldsymbol{\lambda}^*\|_1 &\leq \|\boldsymbol{\eta}^*\| + \ln(1/\varepsilon)\gamma^{-1}\|\boldsymbol{\eta}^\dagger\| + |c|\|\boldsymbol{\eta}^\dagger\| \\ &\leq \|\boldsymbol{\eta}^*\| + \ln(1/\varepsilon)\gamma^{-1}\cdot 1 + m\|\boldsymbol{\eta}^*\|\cdot 1.\end{aligned}$$

Combining this with the results in Corollary 26 and Lemma 25, we may conclude the vector $\boldsymbol{\lambda}^*$ has ℓ_1-norm at most $2^{O(m\ln m)}\ln(1/\varepsilon)$. ■

We may now invoke Theorem 1 to obtain a $2^{O(m\ln m)}\ln^6(1/\varepsilon)\varepsilon^{-5}$ rate of convergence to the optimal solution. Rate bounds with similar dependence on m and slightly better dependence on ε can be obtained by modifying the proof in Section 4 to use first order instead of second order techniques. In that way we may obtain a $\text{poly}(\lambda_{\min}^{-1}, \gamma^{-1}, \mu_{\max})\varepsilon^{-3} = 2^{O(m\ln m)}\varepsilon^{-3}$ rate bound. We omit the rather long but straightforward proof of this fact. Finally, note that if Conjecture 6 is true, then Lemma 27 provides a bound for B in Conjecture 6, implying a $2^{O(m\ln m)}\ln(1/\varepsilon)\varepsilon^{-1}$ rate bound for converging to the optimal loss, which is nearly optimal in both m and ε. We state this as an independent conjecture.

Conjecture 28 *For feature matrices with $-1, 0, +1$ entries, AdaBoost converges to within ε of the optimal loss within $2^{O(m\ln m)}\varepsilon^{-(1+o(1))}$ rounds.*

We next focus on lower bounds on the convergence rate to arbitrary target losses discussed in Section 3. We begin by showing the rate dependence on the norm of the solution as given in Lemma 9 holds for much more general data sets.

Lemma 29 *Suppose a feature matrix has only ± 1 entries, and the finite loss set is non-empty. Consider any coordinate descent procedure, that iteratively chooses a sequence of vectors $\boldsymbol{\lambda}^1, \ldots, \boldsymbol{\lambda}^t, \ldots$, such that successive elements $\boldsymbol{\lambda}^t$ and $\boldsymbol{\lambda}^{t+1}$ of this sequence differ in at most one coordinate, and the loss on this sequence is non-increasing. Then, the number of rounds required by such a procedure to achieve a target loss ϕ^* is at least*

$$\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq \phi^*\}/(2 + 2\ln m).$$

Proof It suffices to upper-bound the step size $|\alpha_t|$ in any round t by at most $2 + 2\ln m$; as long as the step size is smaller than or equal to $2 + 2\ln m$, it will take at least $\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq \phi^*\}/(2 + 2\ln m)$ steps for $\|\boldsymbol{\lambda}\|_1$ to be at least $\inf\{\|\boldsymbol{\lambda}\|_1 : L(\boldsymbol{\lambda}) \leq \phi^*\}$. To see this, recall that (3) shows that a step α_t causes the loss to change by a factor of $f(\alpha_t)$ given by:

$$f(\alpha_t) = \frac{(1+r_t)}{2}e^{-\alpha_t} + \frac{(1-r_t)}{2}e^{\alpha_t},$$

where r_t denotes the correlation in direction j_t in which the step is taken. We find the maximum magnitude of α_t that will still allow $f(\alpha_t)$ to be at most 1. Notice that $f''(\alpha_t) = f(\alpha_t)$ is always positive, since $|r_t| = \delta_t \leq 1$. Therefore f is strictly convex, and $f(\alpha_t) \leq 1$ for α_t lying in some interval. Since $f(\alpha_t) = 1$ at $\alpha_t = 0$ and $\alpha_t = \ln((1+r_t)/(1-r_t))$, the desired maximum magnitude is the latter value. Therefore,

$$|\alpha_t| \leq \left|\ln\left(\frac{1+r_t}{1-r_t}\right)\right| = \ln\left(\frac{1+|r_t|}{1-|r_t|}\right) = \ln\left(\frac{1+\delta_t}{1-\delta_t}\right).$$

Further, by (5), we have that the optimal step in direction j_t would cause the loss to change as follows: $L(\boldsymbol{\lambda}^t) \leq L(\boldsymbol{\lambda}^{t-1})\sqrt{1-\delta_t^2}$. On the other hand, before the step, the loss is at most 1, $L(\boldsymbol{\lambda}^{t-1}) \leq 1$, and after the step the loss is at least $1/m$, that is, $L(\boldsymbol{\lambda}^t) \geq 1/m$. This last fact comes from having at least one example with at most margin 0. Combining these inequalities we get

$$1/m \leq L(\boldsymbol{\lambda}^t) \leq L(\boldsymbol{\lambda}^{t-1})\sqrt{1-\delta_t^2} \leq \sqrt{1-\delta_t^2},$$

that is, $\sqrt{1-\delta_t^2} \geq 1/m$. Now the step length can be bounded as

$$|\alpha_t| \leq \ln\left(\frac{1+\delta_t}{1-\delta_t}\right) = 2\ln(1+\delta_t) - \ln(1-\delta_t^2) \leq 2\delta_t + 2\ln m \leq 2 + 2\ln m.$$

■

We end by showing a new lower bound for the convergence rate to an arbitrary target loss studied in Section 3. Corollary 11 implies that the rate bound in Theorem 1 has to be at least polynomially large in the norm of the solution. We now show that a polynomial dependence on ε^{-1} in the rate is unavoidable too. This shows that rates for competing with a finite solution are different from rates on a data set where the optimum loss is achieved by a finite solution, since in the latter we may achieve a $O(\ln(1/\varepsilon))$ rate.

Corollary 30 *Consider any data set (e.g., the one in Figure 4) for which $\Omega(1/\varepsilon)$ rounds are necessary to get within ε of the optimal loss. If there are constants c and β such that for any $\boldsymbol{\lambda}^*$ and ε, a loss of $L(\boldsymbol{\lambda}^*)+\varepsilon$ can be achieved in at most $O(\|\boldsymbol{\lambda}^*\|_1^c \varepsilon^{-\beta})$ rounds, then $\beta \geq 1$.*

Proof The decomposition lemma implies that $\boldsymbol{\lambda}^* = \boldsymbol{\eta}^* + \ln(2/\varepsilon)\boldsymbol{\eta}^\dagger$ with ℓ_1-norm $O(\ln(1/\varepsilon))$ achieves loss at most $K+\varepsilon/2$ (recall K is the optimal loss). Suppose the corollary fails to hold for constants c and $\beta \leq 1$. Then $L(\boldsymbol{\lambda}^*)+\varepsilon/2 = K+\varepsilon$ loss can be achieved in $O(\varepsilon^{-\beta}/\ln^c(1/\varepsilon)) = o(1/\varepsilon)$ rounds, contradicting the $\Omega(1/\varepsilon)$ lower bound in Lemma 31 in the appendix. ■

6. 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 absolutely necessary, and that a poly$(B, 1/\varepsilon)$ upper bound holds, where ε is the accuracy parameter. The actual rate we derived has rather large exponents, and we discussed a minor variant of AdaBoost that achieves a much tighter and near optimal rate.

For the second kind of convergence, using entirely separate techniques, we derived a C/ε 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. We also studied the constants and showed how they depend on certain intrinsic parameters related to the singular values of the feature matrix. We estimated the worst case values of these parameters, and when considering feature

Convergence rate with respect to:	Reference solution (Section 3)	Optimal solution (Section 4)
Upper bounds:	$13B^6/\varepsilon^5$ [Theorem 1]	$\mathrm{poly}(e^{\mu_{\max}}, \lambda_{\min}^{-1}, \gamma^{-1})/\varepsilon \leq 2^{2^{O(m \ln m)}}/\varepsilon$ [Corollary 23, Lemma 25]
		$\mathrm{poly}(\mu_{\max}, \lambda_{\min}^{-1}, \gamma^{-1})/\varepsilon^3 \leq 2^{O(m \ln m)}/\varepsilon^3$ [Section 5]
Lower bounds with:	$(B/\varepsilon)^{1-\nu}$ for any $\nu > 0$ [Corollaries 11 and 30]	$\max\left\{\frac{2^m \ln(1/\varepsilon)}{\ln m}, \frac{2}{9\varepsilon}\right\}$ [Theorem 1, Lemma 31]
a) $\{0, \pm 1\}$ entries	$O(2^m/\ln m)\ln(1/\varepsilon)$ [Theorem 10]	
b) real entries	Can be arbitrarily large even when m, N, ε are held fixed [Corollary 13]	
Conjectured upper bounds:	$O(B^2/\varepsilon)$ [Conjecture 6]	$2^{O(m \ln m)}/\varepsilon^{1+o(1)}$, if entries in $\{0, \pm 1\}$ [Conjecture 28]

Figure 5: Summary of our most important results and conjectures regarding the convergence rate of AdaBoost. Here m refers to the number of training examples, and ε is the accuracy parameter. The quantity B is the ℓ_1-norm of the reference solution used in Section 3. The parameters $\lambda_{\min}$, γ and $\mu_{\max}$ depend on the data set and are defined and studied in Section 4.

matrices with only $\{-1, 0, +1\}$ entries, this led to a bound on the rate constant C that is doubly exponential in the number of training examples. Since this is rather large, we also included bounds polynomial in both the number of training examples and the accuracy parameter ε, although the dependence on ε in these bounds is non-optimal.

Finally, for each kind of convergence, we conjectured tighter bounds that are not known to hold presently. A table containing a summary of the results in this paper is included in Figure 5.

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 10. We also thank Aditya Bhaskara and Matus Telgarsky for many helpful discussions.

Appendix A. Proofs

We provide proofs for results that have appeared before.

A.1 Lower Bound For Convergence To Optimal Loss

Lemma 31 *For any $\varepsilon < 1/3$, to get within ε of the optimum loss on the data set in Table 4, 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 data set, 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 $1/2$ after the rescaling. 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 $2/(9\varepsilon)$. ■

A.2 A Useful Technical Result

Here we prove a technical result that was used for proving the various rate upper bounds.

Lemma 32 *Suppose $u_0, u_1, \ldots,$ are positive numbers satisfying*

$$u_t - u_{t+1} \geq c_0 u_t^{1+p}, \tag{16}$$

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

$$\frac{1}{u_t^p} - \frac{1}{u_0^p} \geq pc_0 t.$$

Proof By induction on t. The base case $t = 0$ is an identity. Assume the statement holds at iteration t. Then, by the inductive hypothesis,

$$\frac{1}{u_{t+1}^p} - \frac{1}{u_0^p} = \left(\frac{1}{u_{t+1}^p} - \frac{1}{u_t^p}\right) + \left(\frac{1}{u_t^p} - \frac{1}{u_0^p}\right) \geq \frac{1}{u_{t+1}^p} - \frac{1}{u_t^p} + pc_0 t.$$

Thus it suffices to show $1/u_{t+1}^p - 1/u_t^p \geq pc_0$. Multiplying both sides by u_t^p and adding 1, this is equivalent to showing $(u_t/u_{t+1})^p \geq 1 + pc_0 u_t^p$. We will in fact show the stronger inequality

$$(u_t/u_{t+1})^p \geq \exp\left(c_0 u_t^p\right)^p = \exp\left(pc_0 u_t^p\right). \tag{17}$$

Because of the exponential inequality, $e^x \geq 1 + x$, (17) will imply $(u_t/u_{t+1})^p \geq \exp\left(pc_0 u_t^p\right) \geq 1 + pc_0 u_t^p$, which will complete our proof. To show (17), we first rearrange the condition (16) on u_t, u_{t+1} to obtain

$$u_{t+1} \leq u_t\left(1 - c_0 u_t^p\right) \implies \frac{u_t}{u_{t+1}} \geq \frac{1}{1 - c_0 u_t^p} \geq \frac{1}{\exp\left(-c_0 u_t^p\right)} = \exp\left(c_0 u_t^p\right),$$

where the last inequality again uses the exponential inequality. Notice in dividing by u_{t+1} and $(1 - c_0 u_t^p)$, the inequality does not flip since both terms are positive: u_t, u_{t+1} are positive according to the conditions of the lemma, and $(1 - c_0 u_t^p)$ is positive because of the inequality on the left side of the implication in the above. Since $p \geq 0$, we may raise both sides of the above inequality to the power of p to show (17), finishing our proof. ■

A.3 Proof Of Lemma 25

In this section we prove Lemma 25, by separately bounding the quantities $\lambda_{\min}^{-1}$, γ^{-1} and $\mu_{\max}$, through a sequence of lemmas. We will use the next result repeatedly.

Lemma 33 *If $\mathbf{A}$ is an $n \times n$ invertible matrix with $-1, 0, +1$ entries, then $\min_{\vec{x}:\|\vec{x}\|=1} \|\mathbf{A}\vec{x}\|$ is at least $1/n! = 2^{-O(n \ln n)}$.*

Proof It suffices to show that $\|\mathbf{A}^{-1}\vec{x}\| \leq n!$ for any $\vec{x}$ with unit norm. Now $\mathbf{A}^{-1} = \mathrm{adj}(\mathbf{A})/\det(\mathbf{A})$ where $\mathrm{adj}(\mathbf{A})$ is the adjoint of $\mathbf{A}$, whose i, j-th entry is the i, jth cofactor of $\mathbf{A}$ (given by $(-1)^{i+j}$ times the determinant of the $n-1 \times n-1$ matrix obtained by removing the ith row and jth column of $\mathbf{A}$), and $\det(\mathbf{A})$ is the determinant of $\mathbf{A}$. The determinant of any $k \times k$ matrix G can be written as $\sum_\sigma \mathrm{sgn}(\sigma)\prod_{i=1}^k G_{i,\sigma(i)}$, where σ ranges over all the permutations of $1,\ldots,k$. Therefore each entry of $\mathrm{adj}(\mathbf{A})$ is at most $(n-1)!$, and $\det(\mathbf{A})$ is a non-zero integer. Therefore $\|\mathbf{A}^{-1}\vec{x}\| = \|\mathrm{adj}(\mathbf{A})\vec{x}\|/\det(\mathbf{A}) \leq n!\|\vec{x}\|$, and the proof is complete. ■

We first show our bound holds for $\lambda_{\min}$.

Lemma 34 *Suppose* $\mathbf{M}$ *has* $-1,0,+1$ *entries, and let* $\mathbf{M}_F,\lambda_{\min}$ *be as in Corollary 23. Then* $\lambda_{\min} \geq 1/m!$.

Proof Let $\mathbf{A}$ denote the matrix $\mathbf{M}_F$. It suffices to show that $\mathbf{A}$ does not squeeze too much the norm of any vector orthogonal to the null-space $\ker\mathbf{A} \triangleq \{\boldsymbol{\eta} : \mathbf{A}\boldsymbol{\eta} = \mathbf{0}\}$ of $\mathbf{A}$, that is, $\|\mathbf{A}\boldsymbol{\lambda}\| \geq (1/m!)\|\boldsymbol{\lambda}\|$ for any $\boldsymbol{\lambda} \in \ker\mathbf{A}^\perp$. We first characterize $\ker\mathbf{A}^\perp$ and then study how $\mathbf{A}$ acts on this subspace.

Let the rank of $\mathbf{A}$ be $k \leq m$ (notice $\mathbf{A} = \mathbf{M}_F$ has N columns and fewer than m rows). Without loss of generality, assume the first k columns of $\mathbf{A}$ are independent. Then every column of $\mathbf{A}$ can be written as a linear combination of the first k columns of $\mathbf{A}$, and we have $\mathbf{A} = \mathbf{A}'[\mathbf{I}|\mathbf{B}]$ (that is, the matrix $\mathbf{A}$ is the product of matrices $\mathbf{A}'$ and $[\mathbf{I}|\mathbf{B}]$), where $\mathbf{A}'$ is the submatrix consisting of the first k columns of $\mathbf{A}$, $\mathbf{I}$ is the $k\times k$ identity matrix, and $\mathbf{B}$ is some $k\times(N-k)$ matrix of linear combinations (here $|$ denotes concatenation). The null-space of $\mathbf{A}$ consists of $\vec{x}$ such that $\mathbf{0} = \mathbf{A}\vec{x} = \mathbf{A}'[\mathbf{I}|\mathbf{B}]\vec{x} = \mathbf{A}'(\vec{x}_k + \mathbf{B}\vec{x}_{-k})$, where $\vec{x}_k$ is the first k coordinates of $\vec{x}$, and $\vec{x}_{-k}$ the remaining $N-k$ coordinates. Since the columns of $\mathbf{A}'$ are independent, this happens if and only if $\vec{x}_k = -\mathbf{B}\vec{x}_{-k}$. Therefore $\ker\mathbf{A} = \left\{(-\mathbf{B}\vec{z},\vec{z}) : \vec{z}\in\mathbb{R}^{N-k}\right\}$. Since a vector $\vec{x}$ lies in the orthogonal subspace of $\ker\mathbf{A}$ if it is orthogonal to every vector in the latter, we have

$$\ker\mathbf{A}^\perp = \left\{(\vec{x}_k,\vec{x}_{-k}) : \langle\vec{x}_k,\mathbf{B}\vec{z}\rangle = \langle\vec{x}_{-k},\vec{z}\rangle, \forall\vec{z}\in\mathbb{R}^{N-K}\right\}.$$

We next see how $\mathbf{A}$ acts on this subspace. Recall $\mathbf{A} = \mathbf{A}'[\mathbf{I}|\mathbf{B}]$ where $\mathbf{A}'$ has k independent columns. By basic linear algebra, the row rank of $\mathbf{A}'$ is also k, and assume without loss of generality that the first k rows of $\mathbf{A}'$ are independent. Denote by $\mathbf{A}_k$ the $k\times k$ submatrix of $\mathbf{A}'$ formed by these k rows. Then for any vector $\vec{x}$,

$$\|\mathbf{A}\vec{x}\| = \|\mathbf{A}'[\mathbf{I}|\mathbf{B}]\vec{x}\| = \|\mathbf{A}'(\vec{x}_k+\mathbf{B}\vec{x}_{-k})\| \geq \|\mathbf{A}_k(\vec{x}_k+\mathbf{B}\vec{x}_{-k})\| \geq \frac{1}{k!}\|\vec{x}_k+\mathbf{B}\vec{x}_{-k}\|,$$

where the last inequality follows from Lemma 33. To finish the proof, it suffices to show that $\|\vec{x}_k+\mathbf{B}\vec{x}_{-k}\| \geq \|\vec{x}\|$ for $\vec{x}\in\ker\mathbf{A}^\perp$. Indeed, by expanding out $\|\vec{x}_k+\mathbf{B}\vec{x}_{-k}\|^2$ as inner product with itself, we have

$$\|\vec{x}_k+\mathbf{B}\vec{x}_{-k}\|^2 = \|\vec{x}_k\|^2 + \|\mathbf{B}\vec{x}_{-k}\|^2 + 2\langle\vec{x}_k,\mathbf{B}\vec{x}_{-k}\rangle \geq \|\vec{x}_k\|^2 + 2\|\vec{x}_{-k}\|^2 \geq \|\vec{x}\|^2,$$

where the first inequality follows since $\vec{x}\in\ker\mathbf{A}^\perp$ implies $\langle\vec{x}_k,\mathbf{B}\vec{x}_{-k}\rangle = \langle\vec{x}_{-k},\vec{x}_{-k}\rangle$. ■

To show the bounds on γ^{-1} and $\mu_{\max}$, we will need an intermediate result.

Lemma 35 *Suppose* $\mathbf{A}$ *is a matrix, and* $\mathbf{b}$ *a vector, both with* $-1,0,1$ *entries. If* $\mathbf{A}\vec{x} = \mathbf{b}, \vec{x}\geq\mathbf{0}$ *is solvable, then there is a solution satisfying* $\|\vec{x}\| \leq k\cdot k!$, *where* $k = \mathrm{rank}(\mathbf{A})$.

Proof Pick a solution $\vec{x}$ with maximum number of zeroes. Let J be the set of coordinates for which x_i is zero. We first claim that there is no other solution $\vec{x}'$ which is also zero on the set J. Suppose there were such an $\vec{x}'$. Note any point $\vec{p}$ on the infinite line joining $\vec{x},\vec{x}'$ satisfies $\mathbf{A}\vec{p} = \mathbf{b}$, and $\vec{p}_J = \mathbf{0}$ (that is, $p_{i'} = 0$ for $i'\in J$). If i is any coordinate not in J such that $x_i \neq x'_i$, then for some point $\vec{p}^i$

along the line, we have $\vec{p}^i_{J\cup\{i\}} = \mathbf{0}$. Choose i so that $\vec{p}^i$ is as close to $\vec{x}$ as possible. Since $\vec{x} \geq \mathbf{0}$, by continuity this would also imply that $\vec{p}^i \geq \mathbf{0}$. But then $\vec{p}^i$ is a solution with more zeroes than $\vec{x}$, a contradiction.

The claim implies that the reduced problem $\mathbf{A}'\tilde{\vec{x}} = \mathbf{b}, \tilde{\vec{x}} \geq \mathbf{0}$, obtained by substituting $\vec{x}_J = \mathbf{0}$, has a unique solution. Let $k = \mathrm{rank}(\mathbf{A}')$, $\mathbf{A}_k$ be a $k \times k$ submatrix of $\mathbf{A}'$ with full rank, and $\mathbf{b}_k$ be the restriction of $\mathbf{b}$ to the rows corresponding to those of $\mathbf{A}_k$ (note that $\mathbf{A}'$, and hence $\mathbf{A}_k$, contain only $-1,0,+1$ entries). Then, $\mathbf{A}_k\tilde{\vec{x}} = \mathbf{b}_k, \tilde{\vec{x}} \geq \mathbf{0}$ is equivalent to the reduced problem. In particular, by uniqueness, solving $\mathbf{A}_k\tilde{\vec{x}} = \mathbf{b}_k$ automatically ensures the obtained $\vec{x} = (\tilde{\vec{x}}, \mathbf{0}_J)$ is a non-negative solution to the original problem, and satisfies $\|\vec{x}\| = \|\tilde{\vec{x}}\|$. But, by Lemma 33,

$$\|\tilde{\vec{x}}\| \leq k!\|\mathbf{A}_k\tilde{\vec{x}}\| = k!\|\mathbf{b}_k\| \leq k \cdot k!.$$

■

The bound on γ^{-1} follows easily.

Lemma 36 *Let $\gamma, \boldsymbol{\eta}^\dagger$ be as in Item 1 of Lemma 15. Then $\boldsymbol{\eta}^\dagger$ can be chosen such that $\gamma \geq 1/\left(\sqrt{N}m \cdot m!\right) \geq 2^{-O(m\ln m)}$.*

Proof We know that $\mathbf{M}(\boldsymbol{\eta}^\dagger/\gamma) = \mathbf{b}$, where $\mathbf{b}$ is zero on the set F and at least 1 for every example in the zero loss set Z (as given by Item 1 of Lemma 15). Since $\mathbf{M}$ is closed under complementing columns, we may assume in addition that $\boldsymbol{\eta}^\dagger \geq \mathbf{0}$. Introduce slack variables z_i for $i \in Z$, and let $\tilde{\mathbf{M}}$ be $\mathbf{M}$ augmented with the columns $-\vec{e}_i$ for $i \in Z$, where $\vec{e}_i$ is the standard basis vector with 1 on the ith coordinate and zero everywhere else. Then, by setting $\vec{z} = \mathbf{M}(\boldsymbol{\eta}^\dagger/\gamma) - \mathbf{b}$, we have a solution $(\boldsymbol{\eta}^\dagger/\gamma, \vec{z})$ to the system $\tilde{\mathbf{M}}\vec{x} = \mathbf{b}, \vec{x} \geq \mathbf{0}$. Applying Lemma 35, we know there exists some solution $(\vec{y}, \vec{z}')$ with norm at most $m \cdot m!$ (here $\vec{z}'$ corresponds to the slack variables). Observe that $\vec{y}/\|\vec{y}\|_1$ is a valid choice for $\boldsymbol{\eta}^\dagger$ yielding a γ of $1/\|\vec{y}\|_1 \geq 1/(\sqrt{N}m \cdot m!)$. ■

To show the bound for $\mu_{\max}$ we will need a version of Lemma 35 with strict inequality.

Corollary 37 *Suppose $\mathbf{A}$ is a matrix, and $\mathbf{b}$ a vector, both with $-1,0,1$ entries. If $\mathbf{A}\vec{x} = \mathbf{b}, \vec{x} > \mathbf{0}$ is solvable, then there is a solution satisfying $\|\vec{x}\| \leq 1 + k \cdot k!$, where $k = \mathrm{rank}(\mathbf{A})$.*

Proof Using Lemma 35, pick a solution to $\mathbf{A}\vec{x} = \mathbf{b}, \vec{x} \geq \mathbf{0}$ with norm at most $k \cdot k!$. If $\vec{x} > \mathbf{0}$, then we are done. Otherwise let $\vec{y} > \mathbf{0}$ satisfy $\mathbf{A}\vec{x} = \mathbf{b}$, and consider the segment joining $\vec{x}$ and $\vec{y}$. Every point $\vec{p}$ on the segment satisfies $\mathbf{A}\vec{p} = b$. Further any coordinate becomes zero at most once on the segment. Therefore, there are points arbitrarily close to $\vec{x}$ on the segment with positive coordinates that satisfy the equation, and these have norms approaching that of $\vec{x}$. ■

We next characterize the feature matrix $\mathbf{M}_F$ restricted to the finite-loss examples, which might be of independent interest.

Lemma 38 *If $\mathbf{M}_F$ is the feature matrix restricted to the finite-loss examples F (as given by Item 2 of Lemma 15), then there exists a positive linear combination $\vec{y} > \mathbf{0}$ such that $\mathbf{M}_F^T\vec{y} = \mathbf{0}$.*

Proof Item 3 of the decomposition lemma states that whenever the loss $\ell^{\vec{x}}(F)$ of a vector is bounded by m, then the largest margin $\max_{i\in F}(\mathbf{M}_F\vec{x})_i$ is at most $\mu_{\max}$. This implies that there is no vector $\vec{x}$ such that $\mathbf{M}_F\vec{x} \geq \mathbf{0}$ and at least one of the margins $(\mathbf{M}_F\vec{x})_i$ is positive; otherwise, an arbitrarily large multiple of $\vec{x}$ would still have loss at most m, but margin exceeding the constant $\mu_{\max}$. In other words,

$\mathbf{M}_F\vec{x} \geq \mathbf{0}$ implies $\mathbf{M}_F\vec{x} = \mathbf{0}$. In particular, the subspace of possible margin vectors $\left\{\mathbf{M}_F\vec{x} : \vec{x} \in \mathbb{R}^N\right\}$ is disjoint from the convex set Δ_F of distributions over examples in F, which consists of points in $\mathbb{R}^{|F|}$ with all non-negative and at least one positive coordinates. By the Hahn-Banach Separation theorem, there exists a hyperplane separating these two bodies, that is, there is a $\vec{y} \in \mathbb{R}^{|F|}$, such that for any $\vec{x} \in \mathbb{R}^N$ and $\vec{p} \in \Delta_F$, we have $\langle \vec{y}, \mathbf{M}_F\vec{x}\rangle \leq 0 < \langle \vec{y}, \vec{p}\rangle$. By choosing $\vec{p} = \vec{e}_i$ for various $i \in F$, the second inequality yields $\vec{y} > \mathbf{0}$. Since $\mathbf{M}_F\vec{x} = -\mathbf{M}_F(-\vec{x})$, the first inequality implies that equality holds for all $\vec{x}$, that is, $\vec{y}^T\mathbf{M}_F = \mathbf{0}^T$. ■

We can finally upper-bound $\mu_{\max}$.

Lemma 39 *Let $F, \mu_{\max}$ be as in Items 2,3 of the decomposition lemma. Then $\mu_{\max} \leq \ln m \cdot |F|^{1.5} \cdot |F|! \leq 2^{O(m \ln m)}$.*

Proof Pick any example $i \in F$ and any combination $\boldsymbol{\lambda}$ whose loss on F, $\sum_{i\in F} e^{-(\mathbf{M}\boldsymbol{\lambda})_i}$, is at most m. Let $\mathbf{b}$ be the i^{th} row of $\mathbf{M}$, and let $\mathbf{A}^T$ be the matrix $\mathbf{M}_F$ without the ith row. Then Lemma 38 says that $\mathbf{A}\vec{y} = -\mathbf{b}$ for some positive vector $\vec{y} > \mathbf{0}$. This implies the margin of $\boldsymbol{\lambda}$ on example i is $(\mathbf{M}\boldsymbol{\lambda})_i = -\vec{y}^T\mathbf{A}^T\boldsymbol{\lambda}$. Since the loss of $\boldsymbol{\lambda}$ on F is at most m, each margin on F is at least $-\ln m$, and therefore $\max_{i\in F}\left(-\mathbf{A}^T\boldsymbol{\lambda}\right)_i \leq \ln m$. Hence, the margin of example i can be bounded as $(\mathbf{M}\boldsymbol{\lambda})_i = \left\langle \vec{y}^T, -\mathbf{A}^T\boldsymbol{\lambda}\right\rangle \leq \ln m \|\vec{y}\|_1$. Using Corollary 37, we can find $\vec{y}$ with bounded norm, $\|\vec{y}\|_1 \leq \sqrt{|F|}\|\vec{y}\| \leq \sqrt{|F|}(1 + k \cdot k!)$, where $k = \text{rank}(\mathbf{A}) \leq \text{rank}(\mathbf{M}_F) \leq |F|$. The proof follows. ■

References

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

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

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1/2/3), 2002.

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

Y. Freund and R. 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, Aug. 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–374, Apr. 2000.

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

A. J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998.

D. G. Luenberger and Y. 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, Jan. 1992.

L. Mason, J. Baxter, P. Bartlett, and M. 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 and M. K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, 2005.

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

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

C. Rudin, I. Daubechies, and R. E. Schapire. The dynamics of AdaBoost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5:1557–1595, 2004.

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

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

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

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

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

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

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