Margins, Shrinkage, and Boosting

Matus Telgarsky MTELGARS@CS.UCSD.EDU
Department of Computer Science and Engineering, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0404

Abstract

This manuscript shows that AdaBoost and its immediate variants can produce approximate maximum margin classifiers simply by scaling step size choices with a fixed small constant. In this way, when the unscaled step size is an optimal choice, these results provide guarantees for Friedman's empirically successful "shrinkage" procedure for gradient boosting (Friedman, 2000). Guarantees are also provided for a variety of other step sizes, affirming the intuition that increasingly regularized line searches provide improved margin guarantees. The results hold for the exponential loss and similar losses, most notably the logistic loss.

1. Introduction

AdaBoost and related boosting algorithms greedily aggregate many simple predictors into a single accurate predictor (Freund & Schapire, 1997). One explanation for the efficacy of boosting is that it not only seeks aggregates with low empirical risk, but moreover that it prefers good margins, which leads to improved generalization (Schapire et al., 1997). Since AdaBoost does not attain maximum margins on general instances, a push was made to develop methods which carry such a guarantee (Rätsch & Warmuth, 2005; Shalev-Shwartz & Singer, 2008; Rudin et al., 2007).

This work shows that margin maximization may be achieved by scaling back the step size. The intuition for this result is simple (cf. Figure 1): when (equivalently) considered as steps in a coordinate descent procedure, the iterates, depicted as a path, approximate the path of constrained optima (for all possible choices of constraint). By scaling back the step size, the optimal path is more finely approximated.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

As there have been many proposed step sizes for these methods, this manuscript will study four separate choices, deriving improved bounds for the more regularized choices. While it has been shown before that regularized step sizes have good generalization and asymptotically good margins (Zhang & Yu, 2005), this manuscript shows that straightforward step choices achieve these margins at rates matching explicitly margin-maximizing boosting methods.

This practice of scaling back weights was proposed by Friedman (2000, Section 5), who referred to it as a shrinkage scheme (Copas, 1983). This scheme is effective, and adopted in practice (see for instance Bradski (2000, Class `CvGBTrees`) and Pedregosa et al. (2011, Class `GradientBoostingClassifier`)); the purpose of this manuscript is to provide theoretical guarantees.

1.1. Outline

After summarizing the main content, this introduction closes with connections to related work; thereafter, Section 2 recalls the core algorithm, defines the class of loss functions, and provides the four step sizes.

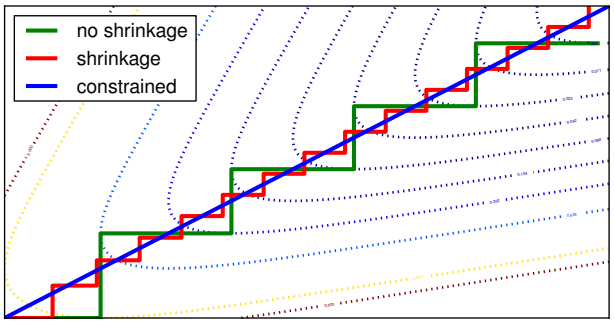

Figure 1. The blue diagonal line is the empirical risk minimizer subject to varying l^1 constraints, and is also a maximum margin choice. The green line takes optimal steps, and grossly overshoots the optimal path. By applying mild shrinkage, the red line approximates the maximum margin choice much more finely.

As boosting is generally studied under the weak learning assumption (a separability condition), the dominant study in this manuscript is also under the condition of separability, and appears in Section 3. The first step is to show that shrinkage does not drastically change the rate of convergence of the empirical risk under these methods. The more involved study is on the topic of margins, and the final subsection compares these bounds to those of other methods.

General (potentially nonseparable) instances are discussed in Section 4. Once again, the first step is a convergence rate guarantee, which again matches those without shrinkage. This section also demonstrates that, under a certain decomposition of boosting problems, the algorithm is still achieving margins on a separable sub-component of the problem.

The manuscript closes with some discussion in Section 5. All proofs are relegated to appendices (in the supplementary material).

1.2. Related Work

Three close works proposed regularized line searches for boosting. First, Friedman (2000) gave the same scheme as is considered here (albeit with only the optimal line search); follow-up work has been mainly empirical, and the questions of convergence rates and margin guarantees do not appear in the literature. Second, Zhang & Yu (2005) also considered regularized line searches, but with a goal of proving consistency; margin maximization is proved as a byproduct, and the analogous results here hold under fewer conditions, and come with rates for the more stringent step sizes. A third work, due to Rätsch et al. (2001), also proves margin maximizing properties of regularized line searches, but again without rates.

As mentioned in the introduction, margin maximization properties of AdaBoost have received extensive study; an excellent survey of results with pointers to other literature is provided by Schapire & Freund (2012, Chapter 5). Amongst these, a crucial result, due to Rudin et al. (2004), provides a concrete input to AdaBoost which yields suboptimal margins (which is used in Section 3.3); that work also studies the evolution of these margins as a dynamical system, a topic which will reappear in Section 5.

The primary contribution of this manuscript is to exhibit margin maximization, thus a natural comparison is to other algorithms with this same guarantee, for instance the works of Rätsch & Warmuth (2005), Shalev-Shwartz & Singer (2008), and Rudin et al. (2007) (or again refer to Schapire & Freund (2012, Chapter 5, Bibliographic Notes) for a more extensive summary). This manuscript will briefly compare with the methods of Shalev-Shwartz & Singer (2008), which subsume some earlier results and match the best guarantees, along with giving a simple, general, greedy scheme. The key distinction between previous work and the present work is firstly that the algorithmic modifications here are minor (in particular, the form of unregularized empirical risk minimization is unchanged), and that properties of an existing, widely used method are discerned (namely, the shrinkage procedure presented by Friedman (2000)).

As is standard in the above works, this manuscript is only concerned with convergence of empirical quantities.

In order to prove convergence rates, this work relies heavily on techniques due to Telgarsky (2012). In particular, the scheme to prove convergence rates of empirical risk, detailed properties of splitting out a *hard core* from a boosting instance (cf. Section 4), and the notion of relative curvature (cf. Section 2.1) are all due to Telgarsky (2012). The intent of the present manuscript is to establish margin properties, and in this regard it departs from Telgarsky (2012); by contrast, the convergence rates of empirical risk presented here are thus trivial, but included since they did not appear explicitly in the literature. It is worth mentioning that these methods produce bad constants when applied to the logistic loss; unfortunately, previous work also suffers in this case (for instance, the work of Collins et al. (2002) provided only convergence of empirical risk, and not rates).

2. Algorithms and Notation

First some basic notation. Let $\{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{X} \times \{-1, +1\}$ denote an m-point sample. Take $\mathcal{H}_0$ to denote the collection of weak learners; it is assumed that $h \in \mathcal{H}_0$ satisfies $h(\mathcal{X}) \subseteq [-1, +1]$, and that $\mathcal{H}_0$ has some form of bounded complexity, meaning specifically that the set of vectors $\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}_0\}$ is finite; this for instance holds if there is a fixed finite set of outputs from $\mathcal{H}_0$, e.g., each h is binary. Consequently, let $\mathcal{H} = \{h_j\}_{j=1}^n$ denote the effective finite set of hypothesis, and collect the responses on the sample into a matrix $A \in [-1, +1]^{m \times n}$ with $A_{ij} = -y_i h_j(x_i)$.

Boosting finds a weighting $\lambda \in \mathbb{R}^n$ of $\mathcal{H}$, which corresponds to a regressor $x \mapsto \sum_{j=1}^n \lambda_j h_j(x)$, and thus a binary classification rule after thresholding. The corresponding (l_1 minimum) margin $\mathcal{M}(A\lambda)$ over the sam-

ple with respect to λ is

$$\mathcal{M}(A\lambda) := \min_{i\in[m]} \frac{-\mathbf{e}_i^\top A\lambda}{\|\lambda\|_1} = \min_{i\in[m]} \frac{y_i \sum_{j=1}^n \lambda_j h_j(x_i)}{\|\lambda\|_1}.$$

Let γ denote the best (largest) achievable margin; equivalently (Shalev-Shwartz & Singer, 2008), γ is the weak learning rate (which justifies the choice of l_1 margins):

$$\gamma := \max_{\substack{\lambda\in\mathbb{R}^n \\ \|\lambda\|_1=1}} \mathcal{M}(A\lambda) = \max_{\substack{\lambda\in\mathbb{R}^n \\ \|\lambda\|_1=1}} \min_{i\in[m]} -\mathbf{e}_i^\top A\lambda$$
$$= \min_{w\in\Delta_m} \max_{j\in[n]} \left| \sum_{i=1}^m w_i y_i h_j(x_i) \right| = \min_{w\in\Delta_m} \|A^\top w\|_\infty.$$

When $\gamma > 0$, the instance is considered separable; classically, this condition is termed the *weak learning assumption* (Kearns & Valiant, 1989; Freund & Schapire, 1997).

2.1. The Family of Loss Functions

The class $\mathbb{L}$ will effectively be "functions similar to the exponential loss". Some of this is for analytic convenience, but some of this appears to be essential, and thus a bit of motivation is appropriate.

Optimization problems typically take advantage of curvature (e.g., strong convexity) to establish a convergence rate. The analysis here instead uses a relative form of curvature: it suffices for, say, the Hessian to not be too small relative to the gap between the current primal objective value and the primal optimum. In this sense, the exponential loss is ideal, as it is a fixed point of the differentiation operator.

Definition 2.1. Given a loss $\ell : \mathbb{R} \to \mathbb{R}_{++}$ (where $\mathbb{R}_{++}$ denotes positive reals), let $C_\ell(z) \geq 1$ (with potentially $C_\ell(z) = \infty$) be the tightest positive constant so that, for every $x \leq z$: $C_\ell(z)^{-1} \leq \exp(x)/\ell^{(i)}(x) \leq C_\ell(z)$ for $i \in \{0, 1, 2\}$ (the zeroth, first, and second derivatives). ◊

Since $C_\ell(z)$ is defined to be the tightest constant, it follows that $y \leq z$ implies $C_\ell(y) \leq C_\ell(z)$.

From here, the class of loss functions may be defined.

Definition 2.2. Let $\mathbb{L}$ contain all functions $\ell : \mathbb{R} \to \mathbb{R}_+$ which are twice continuously differentiable, strictly convex, and have $C_\ell(z) < \infty$ for all $z \in \mathbb{R}$. Additionally, if $\lim_{z\to-\infty} C_\ell(z) = 1$, then $\ell \in \mathbb{L}_\infty$. ◊

Crucially, the two classes $\mathbb{L}$ and $\mathbb{L}_\infty$ both contain the exponential and logistic losses.

Proposition 2.3. $\{x \mapsto \exp(x), x \mapsto \ln(1 + \exp(x))\} \subseteq \mathbb{L}_\infty$.

Algorithm 1 BOOST.
Input: loss ℓ, matrix $A \in [-1, +1]^{m\times n}$.
Output: Weighting sequence $\{\lambda_t\}_{t=0}^\infty$.

Initialize $\lambda_0 := 0$.
for $t = 1, 2, \ldots$: **do**
Choose column (weak learner)

$$j_t := \arg\max_j |\nabla\mathcal{L}(A\lambda_{t-1})^\top A\mathbf{e}_j|.$$

Set descent direction $v_t \in \{\pm\mathbf{e}_{j_t}\}$, whereby

$$\nabla\mathcal{L}(A\lambda_{t-1})^\top A v_t = -\|\nabla\mathcal{L}(A\lambda_{t-1})^\top A\|_\infty.$$

Find α_t via line search.
Update $\lambda_t := \lambda_{t-1} + \alpha_t v_t$.
end for

One way to interpret this is to say "in the limit, logistic loss is the same as exponential loss". Unfortunately, this treatment of the logistic loss ends up being quite unfair, in the sense that the bounds are not accurately representative of the behavior of the algorithm (see Section 3.3). It is, however, unclear how to better deal with the logistic loss.

Lastly, the relevant primal objective function may be defined.

Definition 2.4. Given $\ell \in \mathbb{L}$ and vector $z \in \mathbb{R}^m$, define $\mathcal{L}(z) := m^{-1} \sum_{i=1}^m \ell(z_i)$, whereby the primal optimization problem for boosting is to minimize $\mathcal{L}(A\lambda)$ over the domain $\mathbb{R}^n$. For convenience, define $\bar{\mathcal{L}}_A := \inf_{\lambda\in\mathbb{R}^n} \mathcal{L}(A\lambda)$. ◊

2.2. Algorithm

The algorithm appears in Algorithm 1. Before defining the various step sizes, two more definitions are in order.

Definition 2.5. For every t, define $\gamma_t := \|A^\top\nabla\mathcal{L}(A\lambda_{t-1})\|_\infty / \|\nabla\mathcal{L}(A\lambda_{t-1})\|_1$. (Note that $1 \geq \gamma_t \geq \gamma$.) ◊

Additionally, rather than depending on parameter $C_\ell(z)$ for a carefully chosen z, the following definition suffices.

Definition 2.6. For $t \geq 1$, define $C_t := C_\ell(\ell^{-1}(m\mathcal{L}(A\lambda_{t-1})))$. ◊

The significance of C_t is as follows. Since the algorithm itself is coordinate descent, and moreover since every line search will be shown to guarantee descent, every candidate λ considered in round t will satisfy $\mathcal{L}(A\lambda) \leq \mathcal{L}(A\lambda_{t-1})$; thus, for every $i \in [m]$,

$\ell(\mathbf{e}_i^\top A\lambda) \le m\mathcal{L}(A\lambda) \le m\mathcal{L}(A\lambda_{t-1})$, and so $\mathbf{e}_i^\top A\lambda \le \ell^{-1}(m\mathcal{L}(A\lambda_{t-1}))$, where the inverse is well-defined since ℓ is a bijection between $\mathbb{R}$ and $\mathbb{R}_{++}$ by definition of $\mathbb{L}$ (otherwise $C_\ell(z) = \infty$).

The collection of step sizes considered here are as follows, in order of least to most aggressive. Throughout these step sizes, $\nu \in (0, 1]$ will denote a shrinkage parameter.

Quadratic upper bound. Rather than performing an optimal line search, i.e., rather than minimizing $\alpha \mapsto \mathcal{L}(A(\lambda_{t-1} + \alpha v_t))$, a quadratic upper bound of this univariate function may be minimized, which has a closed form solution (cf. the proof of Lemma 3.2). In particular, define the step size $\alpha_t^{\mathrm{Q}}(\nu) := \nu\gamma_t/C_t^4$. This choice is pleasant algorithmically only when C_t is easy to compute (for instance, $C_t = 1$ for the exponential loss). In general, however, it is useful as an analytic aid, since most step sizes here can be lower bounded by it. This step size was introduced by Telgarsky (2012, Appendix D.3).

Wolfe. The Wolfe line search is a standard tool from nonlinear optimization (Nocedal & Wright, 2006, chapter 3), and for convex problems it may be implemented with binary search (Telgarsky, 2012, Appendix D.1). More precisely, this choice is a set of step sizes $\alpha_t^{\mathrm{W}}(\nu)$ satisfying two conditions. First, the step is explicitly disallowed from being too large:

$$\begin{aligned}&\mathcal{L}(A(\lambda_{t-1} + \alpha v_t))\\ &\quad\le \mathcal{L}(A\lambda_{t-1}) - \alpha(1-\nu/2)\|A^\top \nabla\mathcal{L}(A\lambda_{t-1})\|_\infty.\end{aligned} \tag{2.7}$$

Second, the step should be approximately optimal (in terms of the line search problem):

$$\begin{aligned}&\nabla\mathcal{L}(A(\lambda_{t-1} + \alpha v_t))^\top A v_t\\ &\quad\ge -(1-\nu/4)\|\nabla\mathcal{L}(A\lambda_{t-1})^\top A\|_\infty.\end{aligned} \tag{2.8}$$

(Requiring the reverse inequality (with the right hand side negated) yields the Strong Wolfe Conditions, which are not necessary here.) In contrast to $\alpha_t^{\mathrm{Q}}(\nu)$, the Wolfe step does not require knowledge of C_t, but will yield nearly identical bounds; in fact, computation of the Wolfe step requires only function evaluations, gradient evaluations, and knowledge of ν, A,v_t, λ_t.

AdaBoost. Following the scheme of AdaBoost, define $\alpha_t^{\mathrm{A}}(\nu) := \frac{\nu}{2}\ln(\frac{1+\gamma_t}{1-\gamma_t})$, where convention is followed and $\gamma_t = 1$ is ignored. Unfortunately, even though γ_t is loss-dependent, this step will only yield rates with the exponential loss. However, it will be instrumental in analyzing the fully optimizing step size, presented next. This step size was introduced with the original presentation of AdaBoost (Freund & Schapire, 1997), though the analysis here will rather follow a slightly later treatment (Schapire & Singer, 1999).

Optimal. Let $\alpha_t^{\mathrm{O}}(1)$ be a minimizer to $\alpha \mapsto \mathcal{L}(A(\lambda_{t-1} + \alpha v_t))$, which, as in the case of $\alpha_t^{\mathrm{A}}(\nu)$, is assumed to exist. For $\nu \in (0, 1)$, set $\alpha_t^{\mathrm{O}}(\nu) = \nu\alpha_t^{\mathrm{O}}(1)$. When A is binary and $\ell = \exp$, $\alpha_t^{\mathrm{O}}(\nu) = \alpha_t^{\mathrm{A}}(\nu)$, though in general this is not true. This step size (with shrinkage!) was suggested by Friedman (2000) for use with the logistic loss.

To close, note that $\alpha_t^{\mathrm{Q}}(\nu)$ and $\alpha_t^{\mathrm{O}}(\nu)$ have a simple relationship.

Proposition 2.9. *If $A \in [-1, +1]^{m\times n}$ and $\ell \in \mathbb{L}$, then $\alpha_t^{\mathrm{Q}}(\nu) \le \alpha_t^{\mathrm{O}}(\nu)$.*

3. The Separable Case

This section considers the setting of separability, meaning the weak learning assumption is satisfied ($\gamma > 0$). The three subsections respectively provide convergence rates in empirical risk, basic margin guarantees, and close with some discussion.

3.1. Convergence of Empirical Risk

The basic guarantee is that all of these line search methods, for any loss in $\mathbb{L}$ and with arbitrary shrinkage, exhibit the same basic convergence rate as AdaBoost.

Theorem 3.1. *Let boosting matrix A with corresponding $\gamma > 0$ and shrinkage parameter $\nu \in (0, 1]$ be given. Given any $\ell \in \mathbb{L}$, any $\epsilon > 0$, and iterates $\{\lambda_t\}_{t\ge 0}$ consistent with $\alpha_t^{\mathrm{Q}}(\nu)$, $\alpha_t^{\mathrm{W}}(\nu)$, $\alpha_t^{\mathrm{O}}(\nu)$, or $\alpha_t^{\mathrm{A}}(\nu)$ with $\ell = \exp$, then $\mathcal{O}(\frac{1}{\gamma^2}\ln(\frac{1}{\epsilon}))$ iterations suffice to ensure $\mathcal{L}(A\lambda_t) \le \epsilon$, where the $\mathcal{O}(\cdot)$ suppresses terms depending on C_1 and ν.*

The proof is in the appendix, but a basic discussion will appear here for each step size. The proofs are straightforward, as they should be: convergence analyses typically prove a bound for one step, and then iterate the bound. As such, taking $1/\nu$ steps which are ν-factor as long as the original should do at least as well as the original (which is indeed the exhibited trade-off).

First is the quadratic upper bound, which implicitly gives an upper bound for the optimal step as well. The

proof follows a standard scheme from convex optimization of lower and upper bounding a potential function based on the gradient; the specifics use the relative curvature properties of $\mathbb{L}$, and follow the analysis of Telgarsky (2012, Section 6.1, Appendix D).

Lemma 3.2. *Consider the setting of Theorem 3.1, but with each step size α_t satisfying $\alpha_t^{\mathrm{Q}}(\nu) \leq \alpha_t \leq \alpha_t^{\mathrm{O}}(\nu)$. Then for any t, t_0 with $t \geq t_0$,*

$$\mathcal{L}(A\lambda_{t+1}) \leq \mathcal{L}(A\lambda_{t_0}) \exp\left(-\frac{\nu(2-\nu)}{2C_{t_0+1}^6} \sum_{i=t_0+1}^{t+1} \gamma_i^2\right).$$

The reason for the parameter t_0 is to mitigate the horrendous dependence on C_{t_0}, which is potentially very large. In particular, consider $\ell \in \mathbb{L}_\infty$, meaning $\lim_{z\to-\infty} C_\ell(z) = 1$. C_1 may be quite bad, but convergence still happens. It follows that $C_t \to 1$, and thus, by choosing some large t_0, the bound provides that perhaps there is an initially slow convergence phase, but eventually it is very fast. That is to stay, Lemma 3.2 may be applied multiple times to give a more refined picture of the convergence, particularly in the case that $\ell \in \mathbb{L}_\infty$, which guarantees the constants are eventually near 1.

Next, the Wolfe step size has a similar guarantee (and the analysis once again heavily relies on techniques due to Telgarsky (2012, 6.1, Appendix D)).

Lemma 3.3. *Consider the setting of Theorem 3.1, but with $\alpha_t \in \alpha_t^{\mathrm{W}}(\nu)$. Then for any t, t_0 with $t \geq t_0$,*

$$\mathcal{L}(A\lambda_{t+1}) \leq \mathcal{L}(A\lambda_{t_0}) \exp\left(-\frac{\nu(2-\nu)}{8C_{t_0+1}^6} \sum_{i=t_0+1}^{t+1} \gamma_i^2\right).$$

(The denominator blows up by a factor 4 due to extra halves introduced into the Wolfe conditions, specifically to adjust around the natural Wolfe parameters being within $(0,1)$ and not $(0,1]$.)

Lastly, consider $\alpha_t^{\mathrm{A}}(\nu)$. As in the statement of Theorem 3.1, this step size is only shown to work with the exponential loss. This may be an artifact of the analysis, however, which perhaps follows too closely the treatment of Schapire & Singer (1999), which only considers the exponential loss; for instance, a slightly modified step size can be used to show convergence with the logistic loss (Collins et al., 2002).

Lemma 3.4. *Consider the setting of Theorem 3.1, but with $\alpha_t \in \alpha_t^{\mathrm{A}}(\nu)$ Then for any t, t_0 with $t \geq t_0$,*

$$\mathcal{L}(A\lambda_{t+1}) \leq \mathcal{L}(A\lambda_{t_0}) \prod_{i=t_0+1}^{t+1} C_i^3 \left(1 - \frac{\nu}{2}\gamma_i^2\right).$$

3.2. Margin Maximization

The margin rates here follow a simple pattern: the more regularized the step size, the faster the convergence to a good margin. While no lower bounds are presented, this is an interesting and intuitive correspondence (in particular, consistent with Figure 1). Unfortunately, the unconstrained step sizes only have asymptotic convergence (no rates), so the umbrella theorem for this subsection is also asymptotic.

Theorem 3.5. *Let boosting matrix A with corresponding $\gamma > 0$ and shrinkage parameter $\nu \in (0,1]$ be given. Given any $\ell \in \mathbb{L}_\infty$, any $\epsilon > 0$, and iterates $\{\lambda_t\}_{t\geq 0}$ consistent with $\alpha_t^{\mathrm{Q}}(\nu)$, $\alpha_t^{\mathrm{W}}(\nu)$, $\alpha_t^{\mathrm{A}}(\nu)$ with $\ell = \exp$, or $\alpha_t^{\mathrm{O}}(\nu)$ with binary $A \in \{-1,+1\}^{m\times n}$, then there exists T so that for all $\mathcal{M}(A\lambda_t) \geq \gamma - \epsilon$ for all $t \geq T$.*

In contrast with the convergence rates of empirical risk (e.g., Theorem 3.1), the condition $\ell \in \mathbb{L}_\infty$ is made, rather than simply $\ell \in \mathbb{L}$ (with improved constants when $\ell \in \mathbb{L}_\infty$). This can be interpreted to say: the analysis depends heavily upon the structure of the exponential loss. While this condition is likely unnecessary, on the other extreme it is important for the loss to be strictly convex; if for instance the hinge loss is used, then minimization can stop at any point achieving zero error, in particular at one with poor margin properties.

Returning to task, the quadratic upper bound comes first.

Lemma 3.6. *Suppose the setting of Theorem 3.5, but with $\alpha_t = \alpha_t^{\mathrm{Q}}(\nu)$. Additionally let $t \geq t_0$ be given with $t \geq \frac{2C_1^6 \ln(m)}{\gamma^2\nu(2-\nu)}$ (whereby all margins are nonnegative by Lemma 3.2). Then*

$$\mathcal{M}(A\lambda_{t+1}) \geq \gamma\left(\frac{2-\nu}{2C_{t_0+1}^6}\right) - \frac{\ln(c_0)}{(t+1)\nu\gamma},$$

where

$$c_0 := \max\left\{1, mC_{t_0+1}\mathcal{L}(A\lambda_{t_0}) \exp\left(\frac{\nu(2-\nu)}{2C_{t_0+1}^6} \sum_{i=1}^{t_0} \gamma_i^2\right)\right\}.$$

To interpret this bound, first consider the simplifying case that $\ell = \exp$, whereby $C_t = 1$ for all t. Additionally taking $t_0 = 0$, it follows that $c_0 = m$, and the bound is simply

$$\mathcal{M}(A\lambda_{t+1}) \geq \gamma\left(1 - \frac{\nu}{2}\right) - \frac{\ln(m)}{(t+1)\nu\gamma};$$

in particular, $\mathcal{M}(A\lambda_t) \to \gamma$ as $\nu \to 0$ and $t\nu \to \infty$. For some other $\ell \in \mathbb{L}_\infty$, the denominator term $C_{t_0+1}^6$ also

presents an obstacle to establishing margin maximization; but note that $t_0 \to \infty$ suffices, since it combines with $\ell \in \mathbb{L}_\infty$ via Theorem 3.1 to grant $C_{t_0} \to 1$.

The proof of Lemma 3.6 does not have to work too hard, as the step size appears prominently in the convergence rate bound (cf. Lemma 3.2). As will be discussed in Section 3.3, the rate is nearly ideal.

The Wolfe search exhibits a similar rate.

Lemma 3.7. *Suppose the setting of Theorem 3.5, but with $\alpha_t = \alpha_t^{\mathrm{W}}(\nu)$. Additionally let $t \geq t_0$ be given with $t \geq \frac{8C_1^6 \ln(m)}{\gamma^2 \nu(2-\nu)}$ (whereby all margins are nonnegative by Lemma 3.3). Then*

$$\mathcal{M}(A\lambda_{t+1}) \geq \gamma \left(\frac{2-\nu}{2C_{t_0+1}^2} \right) - \frac{4C_1 \ln(c_0)}{(t+1)\nu\gamma},$$

where

$$c_0 := \max \left\{ 1, mC_{t_0+1}\mathcal{L}(A\lambda_{t_0}) \exp\left(\frac{(2-\nu)\gamma}{2C_{t_0+1}^2} \sum_{i=1}^{t_0+1} \alpha_i \right) \right\}.$$

The preceding two step choices, $\alpha_t^{\mathrm{Q}}(\nu)$ and $\alpha_t^{\mathrm{W}}(\nu)$, had explicit regularization: the first stops as soon as the steepest matching quadratic turns upward, and the second refuses to go beyond a boundary (cf. eq. (2.7)).

On the other hand, the choices $\alpha_t^{\mathrm{A}}(\nu)$ and $\alpha_t^{\mathrm{O}}(\nu)$ are only constrained by the data. Recall that one way to derive $\alpha_t^{\mathrm{A}}(\nu)$ is in the case of binary $A \in \{-1,+1\}^{m\times n}$ and $\ell = \exp$, where it is crucial that each weak learner is wrong on at least one example: this prevents steps from being too large. The techniques in the following proof follow those used in the margin bounds for regular AdaBoost (and are asymptotic there as well). It is worth noting that not only is this bound the worst, but the analysis is the trickiest.

Lemma 3.8. *Consider the setting of Theorem 3.5, but now $\ell = \exp$ and $\alpha_t = \alpha_t^{\mathrm{A}}(\nu)$. Then for any $\epsilon \in (0,\gamma]$, there exists T so that $\mathcal{M}(A\lambda_t) \geq \gamma - \epsilon$ for all $t \geq T$.*

Similarly, $\alpha_t^{\mathrm{O}}(\nu)$ is only implicitly regularized. The condition that $A \in \{-1,+1\}^{m\times n}$ prevents the negative, constraining examples from having too little influence.

Lemma 3.9. *Consider the setting of Theorem 3.5, but now $\ell = \exp$, the matrix A is binary, and $\alpha_t = \alpha_t^{\mathrm{O}}(\nu)$. Then for any $\epsilon > 0$, there exists T so that $\mathcal{M}(A\lambda_t) \geq \gamma - \epsilon$ for all $t \geq T$.*

The above lemmas together provide the proof of Theorem 3.5. But before closing, note that while the results for the unconstrained step sizes were only asymptotic, it is possible to derive a rate for the more modest goal of margins closer to $\gamma/3$.

Proposition 3.10. *Consider the setting of Theorem 3.5, but specialized with $\ell = \exp$ and $\alpha_t = \alpha_t^{\mathrm{A}}(\nu)$. Let a target margin value $\theta < \gamma$ be given. If $\theta < \gamma/(1+\gamma)$ (e.g., it suffices that $\theta < \gamma/2$), then*

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[\frac{-\boldsymbol{e}_i A\lambda_t}{\|\lambda_t\|_1} < \theta \right] \leq \exp\left(\frac{-t\nu(\gamma^2 - \theta\gamma(2+\gamma))}{2} \right).$$

In particular, if $\theta < \gamma/(2+\gamma)$ (e.g., it suffices that $\theta < \gamma/3$) and $t > 2\ln(m)/(\nu(\gamma^2 - \theta\gamma(2+\gamma)))$, then $\mathcal{M}(A\lambda_t) \geq \theta$.

Note, of course, that this bound has the severe analytic artifact of demonstrating no benefit of shrinkage!

3.3. Discussion

To get a sense of these margin bounds, first recall Freund's lower bound on boosting methods in the separable case, which states that $\Omega(\frac{1}{\gamma^2}\ln(\frac{1}{\tau}))$ iterations are necessary to achieve classification error $\tau > 0$ (Freund, 1995, Section 2). Setting $\tau = 1/m$, it follows that $\Omega(\ln(m)/\gamma^2)$ iterations are necessary to achieve any nonnegative margin. By comparison, with $\alpha_t^{\mathrm{W}}(\nu)$ and $\ell = \exp$, just $12\ln(m)/\gamma^2$ iterations with choice $\nu = 1/2$ suffice to reach margin $\gamma/2$ (by Lemma 3.7). More generally, $\alpha_t^{\mathrm{W}}(\nu)$ reaches margin $\gamma(1-\nu)$ with $8\ln(m)/(\nu\gamma)^2$ iterations (if step size $\alpha_t^{\mathrm{Q}}(\nu)$ is used, then $2\ln(m)/(\nu\gamma)^2$ iterations suffice by Lemma 3.6).

The explicit margin-maximizing method of Shalev-Shwartz & Singer (2008) requires $t \geq 32\ln(m)/\epsilon^2$ iterations to achieve margin $\gamma - \epsilon$, where $\epsilon \in (0,\gamma)$. By comparison, converting the above multiplicative bound into an additive bound, step size $\alpha_t^{\mathrm{W}}(\epsilon/\gamma)$ requires $8\ln(m)/\epsilon^2$ iterations. While this bound is slightly better, the comparison is not fair, since $\alpha_t^{\mathrm{W}}(\epsilon/\gamma)$ requires knowledge of γ in the choice of shrinkage parameter ν. (Pessimistically taking $\nu = \epsilon$ gives an additive guarantee, but with a poor rate.) Consequently, it can be reasoned that shrinkage methods achieve excellent margins, but are best suited for multiplicative guarantees.

Another question is how accurately the bounds presented here depict the methods provided. As a brief sanity check, the methods may be run on a problem instance where AdaBoost demonstrably does not achieve maximum margins. The particular instance tested here is a binary matrix $A \in \{-1,+1\}^{8\times 8}$ due to Rudin et al. (2004, Theorem 7); recall that AdaBoost, in the present notation (with A binary), corresponds to $\ell = \exp$ and step size $\alpha_t^{\mathrm{A}}(1) = \alpha_t^{\mathrm{O}}(1)$ (no shrinkage). Two plots are provided.

1. Figure 2 is a sanity check, showing that $\ell = \exp$

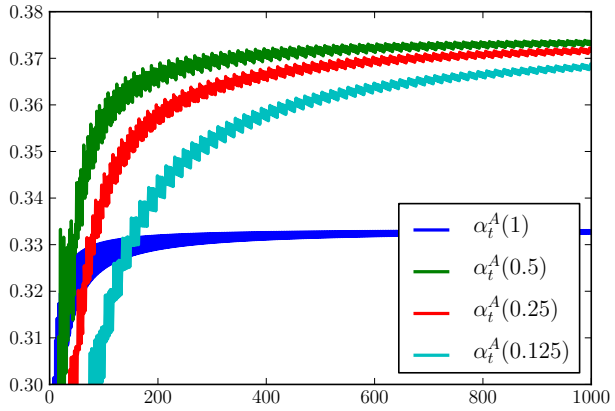

Figure 2. Sanity check: shrinkage leads to margin maximization.

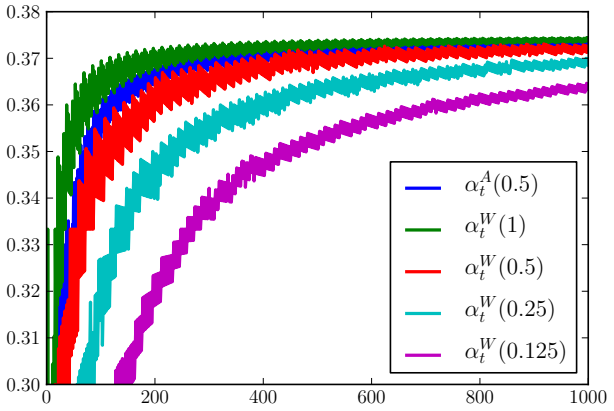

Figure 3. Sanity check: the Wolfe search effectively maximizes margins.

and $\alpha_t^{\mathrm{A}}(1) = \alpha_t^{\mathrm{O}}(1)$ may not achieve maximum margins, but shrinkage overcomes this.

2. Figure 3 demonstrates that the Wolfe search (with $\ell = \exp$) is indeed effective, but demanding higher accuracy comes at a price.

These plots will be discussed further in Section 5. Additional tests with this matrix demonstrated that the method of Shalev-Shwartz & Singer (2008) indeed performs a tiny bit worse than the Wolfe search, but of course one example is not terribly indicative. Perhaps most importantly, a test with the logistic loss showed that the bound is loose: the logistic loss performs well, and does not suffer a startup cost as indicated by the bounds.

4. The General Case

The last technical contribution of this manuscript is to briefly consider the general case (which is potentially nonseparable). Similarly to the separable case, this section will establish convergence rates for empirical risk, margin guarantees, and briefly discuss the connection to existing margin maximizing methods. But first, it is necessary to discuss the structure of the general case, and in particular to develop what margins mean without separability.

This section hinges upon the following decomposition of a boosting instance. This decomposition partitions a boosting instance, specifically its examples $\{(x_i, y_i)\}_{i=1}^m$, into a hard subset $H(A)$, and an easy subset $H(A)^c$. The easy subset alone is separable, and thus margins will be measured there. Although the analysis will rely heavily on properties of this decomposition due to Telgarsky (2012), the decomposition itself has appeared, with various guarantees, in numerous places (Goldreich & Levin, 1989; Impagliazzo, 1995; Mukherjee et al., 2011). The notation $H(A)$ reflects the fact that this structure has no relation to the choice of $\ell \in \mathbb{L}$.

Definition 4.1. (Cf. Telgarsky (2012, Definition 5.1, 5.7).) Given a boosting problem encoded in a matrix $A \in \mathbb{R}^{m\times n}$, a set of examples (rows) $H(A) \subseteq [m]$ is a *hard core* for A (and the corresponding boosting problem) if it satisfies the following properties.

- There exists a weighting $\hat{\lambda} \in \mathbb{R}^n$ with $\mathbf{e}_i^\top A\hat{\lambda} < 0$ for $i \in H(A)^c$ and $\mathbf{e}_i^\top A\hat{\lambda} = 0$ for $i \in H(A)$.
- Every weighting $\lambda \in \mathbb{R}^n$ with $\mathbf{e}_i^\top A\lambda < 0$ for some $i \in H(A)$ also has $\mathbf{e}_k^\top A\lambda > 0$ for some $k \in H(A)$.

Additionally, define a row-wise partition of A into matrices A_0, A_+, where A_+ has the examples in $H(A)$, and A_0 has the examples in $H(A)^c$. ♢

The second property provides that $H(A)$ is difficult: positive margins on some examples force negative margins on others. On the other hand, the complement $H(A)^c$ is easy, and moreover can be solved without affecting $H(A)$.

Proposition 4.2. *(Cf. Telgarsky (2012, Proposition 5.8, Theorem 5.9).) For any $A \in \mathbb{R}^{m\times n}$, a hard core $H(A)$ always exists, and is unique.*

With the decomposition in place, the aforementioned guarantees may be stated. The first, as in the separable case, is convergence of empirical risk. There is hardly anything to do here; the groundwork from Sec-

tion 3 can be plugged directly into existing techniques to generate this theorem (Telgarsky, 2012, Section 6).

Theorem 4.3. *Let general boosting matrix A be given (i.e., potentially $\gamma = 0$), along with shrinkage parameter $\nu \in (0, 1]$, any $\ell \in \mathbb{L}$, and target suboptimality $\epsilon > 0$. Suppose step sizes $\{\alpha_t\}_{t\geq 0}$ are consistent with $\alpha_t^{\mathrm{Q}}(\nu)$, $\alpha_t^{\mathrm{W}}(\nu)$, $\alpha_t^{\mathrm{O}}(\nu)$, or $\alpha_t^{\mathrm{A}}(\nu)$ with $\ell = \exp$ and A binary. Then $\mathcal{O}(\frac{1}{\epsilon})$ iterations suffice to reach suboptimality $\epsilon > 0$.*

If the instance is either separable (i.e., $\gamma > 0$ as in Section 3) or attains its minimizer (i.e., $|H(A)| = m$ (Telgarsky, 2012, Theorem 5.5)), then the rate improves to $\mathcal{O}(\ln(\frac{1}{\epsilon}))$.

Lastly come the margin guarantees. As stated above, $H(A)^c$, considered alone, is separable; note furthermore that the definition of hard core provides the existence of a weighting $\hat{\lambda}$. which has positive margins over $H(A)^c$, but abstains entirely over $H(A)$. Consequently, an approximate minimizer to $\mathcal{L}(A\cdot)$ can always add in a scaling of $\hat{\lambda}$ and improve its empirical risk while simultaneously improving margins over $H(A)^c$. Consequently, it is natural to expect the methods here to achieve positive margins over $H(A)^c$. Note that the following result only shows that some positive margins are attained, and neither assert some sense under which they are maximal, nor does it provide rates.

Theorem 4.4. *Let general boosting matrix A be given with $1 \leq |H(A)| \leq m - 1$ (i.e., the problem is neither separable, nor is the minimizer attainable). Let shrinkage parameter $\nu \in (0, 1]$ and any $\ell \in \mathbb{L}_\infty$ be given. Suppose step sizes $\{\alpha_t\}_{t\geq 0}$ are consistent with $\alpha_t^{\mathrm{Q}}(\nu)$, $\alpha_t^{\mathrm{W}}(\nu)$, $\alpha_t^{\mathrm{O}}(\nu)$ with $\ell = \exp$ and binary A, or $\alpha_t^{\mathrm{A}}(\nu)$ with $\ell = \exp$ and binary A., Then there exists $\hat{\gamma} > 0$ so that every example off the hard core (i.e., $i \in H(A)^c$) has margin at least $\hat{\gamma}$ for all large t.*

To close, consider once again the comparison to explicit margin maximizing boosting methods as presented by Shalev-Shwartz & Singer (2008). There is no point in discussing the specific method discussed in Section 3.3, whose optimal objective value is exactly γ, which in this case is zero, and the method may happily quit without iterating. Indeed, a primary contribution of Shalev-Shwartz & Singer (2008) is not only to address this issue, but show how the same general boosting scheme can be instantiated for the aforementioned method, as well as methods with tolerance to nonseparability.

Indeed, consider the "soft-margin" boosting method (Shalev-Shwartz & Singer, 2008), originally due to Warmuth et al. (2006), which, roughly speaking, has a parameter controlling how many examples to give up on. This is in contrast to the methods here, which not only have a fixed data-dependant structure they try less hard on (the hard core $H(A)$), but moreover the particular margins achieved over the hard core are determined by the loss function $\ell \in \mathbb{L}$. It is of course worth mentioning that the margin analysis in the nonseparable case here is by comparison very incomplete, providing no rates and not even identifying exactly what positive margins are attained.

5. Discussion

This manuscript immediately raises a number of questions. Perhaps foremost is the general question of the impact of margins on the efficacy of boosting. Although margins certainly provide an intuitive theory, it is still unclear how much they directly correlate with good algorithms (Reyzin & Schapire, 2006).

Next, the bounds for the logistic loss are not tight. As there do not appear to be any more forgiving analyses of the logistic loss, the natural question is whether there are new techniques which provide a better characterization.

Lastly, Figure 2 shows a threshold effect: shrinkage 1 does not lead to the right margin, but 1/2 and smaller suffices to reach the maximum margin. (Indeed, experimentation reveals the threshold to be roughly 0.92.) It should be possible to clarify this behavior from the perspective of dynamical systems: smaller steps dodge bad attractors (Rudin et al., 2004; 2007).

Acknowledgements

The author thanks Daniel Hsu and the ICML reviewers for helpful comments and discussions. The author is also deeply indebted to Robert Schapire for numerous discussions, insight, and for suggesting study of the unconstrained step size (at the time, guarantees were only in place for the other choices!). This work was graciously supported by the NSF under grant IIS-0713540.

References

Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Collins, Michael, Schapire, Robert E., and Singer, Yoram. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Copas, J. B. Regression, prediction and shrinkage.

Journal of the Royal Statistical Society, Series B (Methodological), 45(3):311–354, 1983.

Freund, Yoav. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2): 256–285, 1995.

Freund, Yoav and Schapire, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55 (1):119–139, 1997.

Friedman, Jerome H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

Goldreich, Oded and Levin, Leonid. A hard-core predicate for all one-way functions. STOC, pp. 25–32, 1989.

Impagliazzo, Russell. Hard-core distributions for somewhat hard problems. In *FOCS*, pp. 538–545, 1995.

Kearns, Michael and Valiant, Leslie. Cryptographic limitations on learning finite automata and boolean formulae. STOC, pp. 433–444, 1989.

Mukherjee, Indraneel, Rudin, Cynthia, and Schapire, Robert. The convergence rate of AdaBoost. In *COLT*, 2011.

Nocedal, Jorge and Wright, Stephen J. *Numerical optimization*. Springer, 2 edition, 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

Rätsch, Gunnar and Warmuth, Manfred. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Reyzin, Lev and Schapire, Robert E. How boosting the margin can also boost classifier complexity. In *In Proceedings of the 23rd International Conference on Machine Learning*, pp. 753–760, 2006.

Rudin, Cynthia, Daubechies, Ingrid, and Schapire, Robert E. The dynamics of AdaBoost: cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5:1557–1595, 2004.

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

Schapire, Robert E. and Freund, Yoav. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

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

Schapire, Robert E., Freund, Yoav, Barlett, Peter, and Lee, Wee Sun. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pp. 322–330, 1997.

Shalev-Shwartz, Shai and Singer, Yoram. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *COLT*, pp. 311–322, 2008.

Steele, J. Michael. *The Cauchy-Schwarz Master Class*. Cambridge University Press, 2004.

Telgarsky, Matus. A primal-dual convergence analysis of boosting. 2012. `arXiv:1101.4752v3 [cs.LG]`.

Warmuth, Manfred K., Liao, Jun, and Rätsch, Gunnar. Totally corrective boosting algorithms that maximize the margin. In *ICML*, pp. 1001–1008, 2006.

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