On the Equivalence of Weak Learnability and Linear Separability: New Relaxations and Efficient Boosting Algorithms

Shai Shalev-Shwartz
Toyota Technological Institute, Chicago, USA
SHAI@TTI-C.ORG

Yoram Singer
Google, Mountain View, USA
SINGER@GOOGLE.COM

Abstract

Boosting algorithms build highly accurate prediction mechanisms from a collection of low-accuracy predictors. To do so, they employ the notion of weak-learnability. The starting point of this paper is a proof which shows that weak learnability is equivalent to linear separability with ℓ_1 margin. While this equivalence is a direct consequence of von Neumann's minimax theorem, we derive the equivalence directly using Fenchel duality. We then use our derivation to describe a family of relaxations to the weak-learnability assumption that readily translates to a family of relaxations of linear separability with margin. This alternative perspective sheds new light on known soft-margin boosting algorithms and also enables us to derive several new relaxations of the notion of linear separability. Last, we describe and analyze an efficient boosting framework that can be used for minimizing the loss functions derived from our family of relaxations. In particular, we obtain efficient boosting algorithms for maximizing hard and soft versions of the ℓ_1 margin.

1 Introduction

Boosting is a popular and successful method for building highly accurate predictors from a set of low-accuracy base predictors. For an overview see for example [FS99, Sch03, MR03]. The first boosting algorithm was used for showing the equivalence between weak learnability and strong learnability [Sch90]. Weak learnability means that for any distribution over a set of examples there exists a single feature, also referred to as weak hypothesis, that performs slightly better than random guessing. Schapire [Sch90] was the first to show that if the weak learnability assumption holds then it is possible to construct a highly accurate classifier, to the point that it perfectly classifies all the examples in the training set. This highly accurate classifier is obtained by taking the sign of a weighted combination of weak hypotheses. Put another way, [Sch90] showed that if the weak learnability assumption holds then the set of examples is linearly separable.

Studying the generalization properties of the AdaBoost algorithm, Schapire et al. [SFBL97] showed that AdaBoost in fact finds a linear separator with a large margin. However, AdaBoost does not converge to the max margin solution [RW05, RSD07]. Interestingly, the equivalence between weak learnability and linear separability is not only qualitative but also quantitative: weak learnability with edge γ is equivalent to linear separability with an ℓ_1 margin of γ. We give a precise statement and a simple proof of the equivalence in Thm. 4. We note that the equivalence can be also derived from von Neumann's minimax theorem [vN28]. Nevertheless, our proof is instructive and serves as a building block for the derivation of our main results.

Since the weak learnability assumption is equivalent to linear separability, it implies that the weak-learnability assumption is non-realistic due to its high sensitivity to even small amounts of label noise. For example, assume that we have a dataset that is perfectly separable with a large margin with the exception of two examples. These two examples share the same instance but attain opposite labels. Since such a dataset is non-separable, the weak learnability assumption fails to hold as well. To cope with this problem, we must somehow relax the weak learnability, which is equivalent to relaxing the linear separability assumption. In this paper we propose a family of relaxations of the linear separability assumption, which stems from the equivalence of weak-learnability and linear-separability. The guiding tool is to first define a natural family of relaxations of the weak learnability assumption, and then analyze its implication on the separability assumption.

In addition to our analysis and relaxations outline above, we also propose and analyze an algorithmic framework for boosting that efficiently solve the problems derived from our family of relaxations. The algorithm finds an ϵ accurate solution after performing at most $O(\log(m)/\epsilon^2)$ iterations, where m is the number of training examples. The number of iterations upper bounds the number of different weak-hypotheses constituting the solution. Therefore, we cast a natural trade-off between the desired accuracy level, ϵ, of the (possibly relaxed) margin attained by the weight vector learned by the boosting algorithm, and the sparseness of the resulting predictor. In particular, we obtain new algorithms for maximizing the hard and soft ℓ_1 margin. We also provide an $O(m \log(m))$ procedure for entropic projections onto ℓ_∞ balls. Combined with this procedure, the total complexity of each iteration of our algorithm for minimizing the soft ℓ_1 margin is almost the same as the complexity of each iteration of AdaBoost, assuming that the complexity of each activation of the weak learning algorithm requires $\Omega(m)$ time.

Related Work As mentioned above, the equivalence between weak learnability and linear separability with ℓ_1 margin is a direct consequence of von Neumann's minimax theorem in game theory [vN28]. Freund and Schapire [FS96] were the first to use von Neumann's result to draw a connection between weak learnability and separability. They showed that if the weak learnability assumption holds then the data is linearly separable. The exact quantification of the weak learnability parameter and the ℓ_1 margin parameter was spelled out later in [RW05].

Schapire et al. [SFBL97] showed that the AdaBoost algorithm finds a large margin solution. However, as pointed out by [RW05, RSD07], AdaBoost does not converge to the max margin solution. Ratsch and Warmuth [RW05] suggested an algorithm called AdaBoost$_*$ which converges to the maximal margin solution in $O(\log(m)/\epsilon^2)$ iterations. The family of algorithms we propose in this paper entertains the same convergence properties. Rudin et al. [RSD07] provided a more accurate analysis of the margin attained by AdaBoost and also presented algorithms for achieving the max-margin solution. However, their algorithm may take $O(1/\epsilon^3)$ iterations to find an ϵ accurate predictor.

The above algorithms are effective when the data is linearly separable. Over the years, many boosting algorithms were suggested for non-separable datasets. We list here few examples. The LogLoss Boost algorithm [CSS02] tries to minimize the cumulative logistic loss, which is less sensitive to noise. MadaBoost [DW00] is another example of an algorithm that copes with non-separability. It does so by capping from the above the importance weights produced by the boosting algorithm. MadaBoost shares similarities with some of the relaxations presented in this paper. However, MadaBoost does not exploit the aforementioned equivalence and has a convergence rate that seems to be inferior to the rate obtained by the relaxations we consider in this paper. Another notable example for a boosting algorithm that works well in the non-separable case and is noise tolerant is the BrownBoost algorithm [Fre01]. BrownBoost uses the error-function (erf) as a margin-based loss function. The error-function reaches an asymptote when its input (margin in the context of BrownBoost) tends to $-\infty$. It thus constitutes a robust alternative to a convex loss function, including the LogLoss function. Since the error function is non-convex, all the results presented in this paper are not applicable to BrownBoost. In the support vector machine literature, the common relaxation of the separability assumption is obtained by using the hinge-loss (see for example [CST00]). Warmuth, Glocer and Ratsch [WGR07] recently proposed the SoftBoost algorithm that directly minimizes the hinge-loss function. The relaxation described in [WGR07] is a special case of the family of relaxations we present in this paper. The SoftBoost algorithm also builds on the idea of relaxing the weak learnability assumption by capping the maximal weight of a single example. A similar idea was also used by the SmoothBoost algorithm [Ser03]. Our presentation leads to an interesting perspective on this relaxation, showing that maximizing the margin while minimizing the hinge-loss is equivalent to maximizing the average margin of the k examples with the worst margin. This equivalence is also implied from the work presented in [WGR07]. More importantly, in this paper we present a much simple algorithm which does not employ a convex optimization procedure on each round of boosting. Our approach stands in contrast to the algorithm of [WGR07], which requires "totally corrective" updates (see also [WLR06]) and needs to solve a rather complex optimization problem on each iteration.

The family of boosting algorithms we derive is reminiscent of the boosting algorithm proposed by Zhang [Zha03]. However, our analysis is different and allows us to: (i) provide an analytic solution for the step size; (ii) tackle complicated loss functions, including cases when the loss function does not take an explicit form. Our analysis stems from the primal-dual view of online convex programming [SSS06a, SSS07, SS07] and also borrows ideas from the analysis given in [SVL07]. The main difference between our analysis and that of [SVL07, Zha03] is that we do not impose any assumption on the second order derivatives of the objective function. Instead, we rely on a duality argument and require a strongly convex assumption on the Fenchel conjugate of the loss function. As we show, in many interesting cases, it is simple to verify that our assumption holds, while it is very complex to analyze the second order derivatives of the loss function in hand.

Throughout this paper, we focus on the analysis of the empirical loss over the training set. There has been extensive work on obtaining generalization bounds for boosting algorithms and for margin-based hypotheses. We refer the reader for example to [SFBL97, MBB98, KPL01]. A complimentary question, left out of the scope of this paper, is whether the equivalence between weak learnability and linear separability with margin can be exploited for obtaining improved generalization bounds.

2 Notation and basic definitions

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a sequence of m examples, where for all i, $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{+1, -1\}$. Let $\mathcal{H}$ be a set of base hypotheses, namely, each $h \in \mathcal{H}$ is a function from $\mathcal{X}$ into $[+1, -1]$. For simplicity, we assume that $\mathcal{H}$ is finite and thus $\mathcal{H} = \{h_1, \ldots, h_n\}$. Let A be a matrix of size $m \times n$ over $[+1, -1]$ where the (i, j) entry of A is $A_{i,j} = y_i \, h_j(\mathbf{x}_i)$. We note that boosting algorithms solely use the matrix A and do not directly work with the set of examples. Therefore, throughout the rest of the paper we focus on the properties of the matrix A.

We denote column vectors with bold face letters, e.g. $\mathbf{d}$ and $\mathbf{w}$, and use the notation $\mathbf{d}^\dagger, \mathbf{w}^\dagger$ for denoting their corresponding row vectors. The inner product between vectors is denoted by $\langle \mathbf{d}, \mathbf{w} \rangle = \mathbf{d}^\dagger \mathbf{w}$. We denote by $A^\dagger$ the transpose of the matrix A. The vector obtained by multiplying a matrix A with a vector $\mathbf{d}$ is designated as $A\mathbf{d}$ and its ith element as $(A\mathbf{d})_i$.

The set of non-negative real numbers is denoted as $\mathbb{R}_+$ and the set of integers $\{1, \ldots, n\}$ as $[n]$. The m dimensional probability simplex is denoted by $\mathbb{S}^m = \{\mathbf{d} \in \mathbb{R}_+^m : \|\mathbf{d}\|_1 = 1\}$. We denote the m dimensional ℓ_1 ball of radius r by $\mathbb{B}_1^m(r) = \{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w}\|_1 \leq r\}$. For the unit ℓ_1 ball, we often omit r and use the shorthand $\mathbb{B}_1^m$. Similarly, we denote the m dimensional ℓ_p ball by $\mathbb{B}_p^m(r) = \{\mathbf{w} \in \mathbb{R}^m :$

$\|\mathbf{w}\|_p \leq r\}$ and again omit r whenever it is equals to 1.

Definition 1 (separability with ℓ_1 margin γ) *A matrix A is linearly separable with ℓ_1 margin γ if there exists $\mathbf{w} \in \mathbb{B}_1^n$ such that $\min_{i\in[m]}(A\mathbf{w})_i \geq \gamma$, and γ is the largest scalar that satisfies the above inequality, namely,*

$$\gamma \;=\; \max_{\mathbf{w}\in\mathbb{B}_1^n} \; \min_{i\in[m]} \, (A\mathbf{w})_i \; .$$

Definition 2 (γ-weak-learnability) *A matrix A is γ-weak-learnable if for all $\mathbf{d} \in \mathbb{S}^m$ there exists $j \in [n]$ such that $|(\mathbf{d}^\dagger A)_j| \geq \gamma$, and γ is the largest scalar that satisfies the above. Namely,*

$$\gamma \;=\; \min_{\mathbf{d}\in\mathbb{S}^m} \; \max_{j\in[n]} \, |(d^\dagger A)_j| \; .$$

We next give a few basic definitions from convex analysis. A set $S \subset \mathbb{R}^n$ is convex if for any two vectors $\mathbf{d}_1, \mathbf{d}_2$ in S, all the line between $\mathbf{d}_1$ and $\mathbf{d}_2$ is also in S, that is, $\{\alpha\mathbf{d}_1 + (1-\alpha)\mathbf{d}_2 : \alpha \in [0,1]\} \subseteq S$. A function $f : S \to \mathbb{R}$ is closed and convex if for any scalar r, the level set $\{\mathbf{d} : f(\mathbf{d}) \leq r\}$ is closed and convex. We allow functions to output $+\infty$ and denote by $\mathrm{dom}(f)$ the set $\{\mathbf{d} : f(\mathbf{d}) < +\infty\}$. The core of a set $C \in \mathbb{R}^n$, denoted $\mathrm{core}(C)$, is the set of all points in $\mathbf{x} \in C$ such that for all $\mathbf{d} \in \mathbb{R}^n$ there exists $\tau' > 0$ for which for all $\tau \in [0, \tau']$ we have $\mathbf{x} + \tau d \in C$. The Fenchel conjugate of a function $f : S \to \mathbb{R}$ is defined as

$$f^\star(\boldsymbol{\theta}) = \max_{\mathbf{d}\in S} \; \langle \mathbf{d}, \boldsymbol{\theta}\rangle - f(\mathbf{d}) \;\; . \qquad (1)$$

If f is closed and convex then $f^{\star\star} = f$.

Our derivation makes an extensive use of the following theorem.

Theorem 3 (Fenchel Duality: Theorem 3.3.5 in [BL06]) *Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^n :\to \mathbb{R} \cup \{\infty\}$ be two closed and convex functions and let A be a matrix of dimension $m \times n$. Then,*

$$\max_{\mathbf{w}} -f^\star(-A\mathbf{w}) - g^\star(\mathbf{w}) \;\leq\; \min_{\mathbf{d}} f(\mathbf{d}) + g(d^\dagger A) \;\; .$$

The above holds with equality if in addition we have

$$\mathbf{0} \in \mathrm{core}\left(\mathrm{dom}(g) - A^\dagger \mathrm{dom}(f)\right) \;\; .$$

We denote an arbitrary norm by $\|\cdot\|$ and its dual norm by $\|\cdot\|_\star$. That is,

$$\|\mathbf{w}\|_\star = \max_{\mathbf{d}:\|\mathbf{d}\|\leq 1} \langle \mathbf{w}, \mathbf{d}\rangle \;\; .$$

Two dual norms that we extensively use are $\|\mathbf{w}\|_1 = \sum_i |w_i|$ and $\|\mathbf{w}\|_\infty = \max_i |w_i|$.

For a set C, we denote by $I_C(\mathbf{d})$ the indicator function of C, that is, $I_C(\mathbf{d}) = 0$ if $\mathbf{d} \in C$ and otherwise $I_C(\mathbf{d}) = \infty$. The definition of $\|\mathbf{w}\|_\star$ implies that the Fenchel conjugate of $I_C(\mathbf{d})$ where $C = \{\mathbf{d} : \|\mathbf{d}\| \leq 1\}$, is the function $\|\cdot\|_\star$. To conclude this section, we would like to point the reader to Table 1 which summarizes our notations.

Table 1: Summary of notations.

$\mathbf{x}, \mathbf{x}^\dagger$	column vector and its transpose
$\langle \mathbf{x}, \mathbf{v}\rangle$	inner product ($= \mathbf{x}^\dagger \mathbf{v}$)
A	matrix of size $m \times n$
$\mathbb{S}^m$	m dimensional probability simplex
$\mathbb{B}_p^m(\nu)$	ℓ_p ball $\{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w}\|_p \leq \nu\}$
$I_C(\mathbf{d})$	indicator function ($= 0$ if $\mathbf{d} \in C$ and $= \infty$ else)
$[\mathbf{x}]_+$	vector whose ith element equals $\max\{0, x_i\}$
$\|\cdot\|$, $\|\cdot\|_\star$	norm and its dual norm
$f, f^\star$	function and its Fenchel conjugate
$\mathbf{e}^i$	all zeros vector except 1 in the ith position
$[m]$	the set $\{1, \ldots, m\}$

3 Weak-learnability and linear-separability

In this section we establish the equivalence between weak learnability and linear separability with ℓ_1 margin. As mentioned before, this result can be derived from von Neumann's minimax theorem. The purpose of the proof below is to underscore the duality between weak learnability and separability, which becomes useful in the next sections.

Theorem 4 *A matrix A is γ-weak-learnable if and only if it is linearly separable with ℓ_1 margin of γ.*

Proof: We prove the theorem using Fenchel duality (Thm. 3). For convenience, we refer to the optimization problem on the right (left) hand side of Thm. 3 as the primal (dual) optimization problem. Let f be the indicator function of the m-dimensional simplex, i.e. $f(\mathbf{d}) = 0$ if $\mathbf{d} \in \mathbb{S}^m$ and otherwise $f(\mathbf{d}) = \infty$, and let $g(\mathbf{w}) = \|\mathbf{w}\|_\infty$. Then, the primal problem is

$$P^\star = \min_{\mathbf{d}} f(\mathbf{d}) + g(\mathbf{d}^\dagger A) = \min_{\mathbf{d}\in\mathbb{S}^m} \|\mathbf{d}^\dagger A\|_\infty \;\; .$$

The definition of γ-weak-learnability conveys that A is $P^\star$-weak-learnable. Next, we turn to the dual problem. The Fenchel conjugate of g is the indicator function of the set $\mathbb{B}_1^n$ (see Sec. 2) and the Fenchel conjugate of f is

$$f^\star(\boldsymbol{\theta}) = \max_{\mathbf{d}\in\mathbb{R}^m} \langle \boldsymbol{\theta}, \mathbf{d}\rangle - f(\mathbf{d}) = \max_{\mathbf{d}\in\mathbb{S}^m} \langle \boldsymbol{\theta}, \mathbf{d}\rangle = \max_{i\in[m]} \theta_i \;\; .$$

Therefore,

$$D^\star = \max_{\mathbf{w}\in\mathbb{R}^n} -f^\star(-A\,\mathbf{w}) - g^\star(\mathbf{w}) = \max_{\mathbf{w}\in\mathbb{B}_1^n} \min_{i\in[m]} (A\mathbf{w})_i \; .$$

Definition 1 implies that A is separable with ℓ_1 margin of $D^\star$. To conclude our proof, it is left to show that $P^\star = D^\star$. First, we note that for $\mathbf{w} = \mathbf{0}$ the value of D is zero, and thus $D^\star \geq 0$. Therefore, if $P^\star = 0$ then $0 = P^\star \geq D^\star \geq 0$ so in this case we clearly have $P^\star = D^\star$. Assume now that $P^\star = \gamma > 0$. Based on Thm. 3 and the definition of the core operator, it suffices to show that for any vector $\mathbf{v}$ there exists $\tau' > 0$ such that for all $\tau \in [0, \tau']$ we have $\tau\,\mathbf{v} \notin \{A^\dagger\mathbf{d} : \mathbf{d} \in \mathbb{S}^m\}$. This property holds true since for any $\mathbf{d} \in \mathbb{S}^m$ we have $\|A^\dagger\mathbf{d}\|_\infty \geq P^\star$ while for sufficiently small τ' we must have $\|\tau\mathbf{v}\|_\infty < P^\star$ for all $\tau \in [0, \tau']$. ■

4 A family of relaxations

In the previous section we showed that weak learnability is equivalent to separability. The separability assumption is problematic since even a perturbation of a singe example can break it. In this section we propose a family of relaxations of the separability assumption. The motivation for these relaxations stems from the equivalence between weak-learnability and separability. The main idea is to first define a natural family of relaxations of the weak learnability assumption, and then analyze the implication to the separability assumption. To simplify the presentation, we start with a particular relaxation that was studied in [Ser03, WLR06]. We then generalize the example and describe the full family of relaxations.

4.1 A first relaxation: capped probabilities and soft margin

To motivate the first simple relaxation, consider a matrix A whose ith row equals to the negation of its jth row. That is, our training set contains an instance which appears twice, each time with a different label. Clearly, this training set is not separable even though the rest of the training set can be perfectly separable with a large margin. The equivalence between weak learnability and linear separability implies that A is also not weak learnable. To derive this property directly, construct the distribution $\mathbf{d}$ with $d_i = d_j = \frac{1}{2}$ (and $d_r = 0$ for $r \neq i$ and $r \neq j$) and note that $\mathbf{d}^\dagger A = \mathbf{0}$.

In the above example, the weak learnability assumption fails because we place excessive weight on the problematic examples i, j. Indeed, it was observed that AdaBoost overweighs examples, which partially explains its poor performance on noisy data. To overcome this problem, it was suggested (see for instance [Ser03, WLR06]) to restrict the set of admissible distributions by capping the maximum importance weight of each example. That is, the weak learner should return a weak hypothesis only when its input distribution satisfies $\|\mathbf{d}\|_\infty \leq \frac{1}{k}$, for a predefined integer $k \in [m]$.

Plugging the above restriction on $\mathbf{d}$ into Definition 2 we obtain the following relaxed weak learnability value,

$$\rho \;=\; \min_{\mathbf{d} \in \mathbb{S}^m : \|\mathbf{d}\|_\infty \leq \frac{1}{k}} \; \max_{j \in [n]} |(\mathbf{d}^\dagger A)_j| \;\; . \tag{2}$$

Assume that a matrix A satisfies the above with $\rho > 0$. The immediate question that surfaces is what is the implication on the separability properties of A? To answer this question, we need to refine the duality argument given in the proof of Thm. 4.

Let $f(\mathbf{d})$ be the indicator function of $\mathbb{S}^m \cap \mathbb{B}^m_\infty(\frac{1}{k})$ and let $g(\mathbf{w}) = \|\mathbf{w}\|_\infty$. The optimization problem given in Eq. (2) can be rewritten as $\min_{\mathbf{d}} f(\mathbf{d}) + g(\mathbf{d}^\dagger A)$. To derive the dual optimization problem, we find the Fenchel conjugate of f,

$$f^\star(\boldsymbol{\theta}) \;=\; \max_{\mathbf{d} \in \mathbb{S}^m : \|\mathbf{d}\|_\infty \leq \frac{1}{k}} \langle \mathbf{d}, \boldsymbol{\theta} \rangle \;\; .$$

To maximize the inner product $\langle \mathbf{d}, \boldsymbol{\theta} \rangle$ we should allocate the largest admissible weight to the largest element of $\boldsymbol{\theta}$, allocate the largest of the remaining weights to the second largest element of $\boldsymbol{\theta}$, and so on and so forth. For each $i \in [m]$, let $s_i(\boldsymbol{\theta})$ be the ith largest element of $\boldsymbol{\theta}$, that is, $s_1(\boldsymbol{\theta}) \geq s_2(\boldsymbol{\theta}) \geq \ldots$. Then, the above argument yields

$$f^\star(\boldsymbol{\theta}) \;=\; \frac{1}{k} \sum_{j=1}^{k} s_j(\boldsymbol{\theta}) \; .$$

Combining the form of $f^\star$ with Thm. 3 we obtain that the dual problem of Eq. (2) is

$$\max_{\mathbf{w} \in \mathbb{B}^n_1} \; \frac{1}{k} \sum_{j=0}^{k-1} s_{m-j}(A\mathbf{w}) \; . \tag{3}$$

Using the same technique as in the proof of Thm. 4 it is easy to verify that strong duality holds as well. We therefore obtain the following corollary.

Corollary 5 *Let A be a matrix and let $k \in [m]$. For a vector $\boldsymbol{\theta}$, let $\mathrm{AvgMin}_k(\boldsymbol{\theta})$ be the average of the k smallest elements of $\boldsymbol{\theta}$. Let ρ be as defined in Eq. (2). Then,*

$$\max_{\mathbf{w} \in \mathbb{B}^n_1} \; \mathrm{AvgMin}_k(A\mathbf{w}) \;=\; \rho \;\; .$$

Let us now discuss the role of the parameter k. First, if $k = 1$ then the function AvgMin_k reduces to the minimum over the vector provided as its argument, and therefore we revert back to the traditional definition of margin. When $k = m$, the only admissible distribution is the uniform distribution. In this case, it is easy to verify that the optimal weight vector associates $w_j = 1$ with the feature that maximizes $|(\mathbf{d}^\dagger A)_j|$ (while $\mathbf{d}$ being the uniform distribution) and $w_j = 0$ with the rest of the features. That is, the performance of the optimal strong hypothesis is equal to the performance of the best single weak hypothesis, and no boosting process takes place. The interesting regime is when k is proportional to m, for example $k = 0.1m$. In this case, if $\rho > 0$, then we are guaranteed that 90% of the examples can be separated by margin of at least ρ.

It is also possible to set k based on knowledge of the number of noisy examples in the training set and the separability level of the rest of the examples. For example, assume that all but ν of the examples are separable with margin γ. Then, the worst objective value that $\mathbf{w}$ can attain is, $\mathrm{AvgMin}_k(A\mathbf{w}) = \frac{-\nu + (k-\nu)\gamma}{k}$. Constraining the right hand side of this equality above to be at least $\frac{\gamma}{2}$ and solving for k yields that for $k \geq 2\nu(\gamma+1)/\gamma$ at least $m - k$ examples attain a margin value of at least $\gamma/2$.

4.2 A general relaxation scheme

We now generalize the above relaxation and present our general relaxation scheme. To do so, we first rewrite Eq. (2) as follows. Denote $C = \mathbb{B}^m_\infty(1/k)$ and recall that $I_C(\mathbf{d})$ is the indicator function of the set C. We can now rewrite Eq. (2) as

$$\rho \;=\; \min_{\mathbf{d} \in \mathbb{S}^m} \left(\max_{j \in [n]} |(\mathbf{d}^\dagger A)_j| + I_C(\mathbf{d}) \right) \;\; . \tag{4}$$

The general relaxation scheme is obtained by replacing I_C with a large family of functions. Before specifying the properties of allowed functions, let us first define the following generalized notion of weak learnability.

Definition 6 ((ρ, f)-weak-learnability) *Let f be an arbitrary function. A matrix A is (ρ, f)-weak-learnable if*

$$\rho \;=\; \min_{\mathbf{d}\in\mathbb{S}^m} \left(\max_{j\in[n]} |(\mathbf{d}^\dagger A)_j| + f(\mathbf{d}) \right) \;\; .$$

Intuitively, we can think on ρ as the minimum of the maximal edge plus a regularization term $f(\mathbf{d})$. In the case of capped importance weights, the regularization function is a barrier function that does not penalize distributions inside $\mathbb{B}^m_\infty(1/k)$ and places an infinite penalty for the rest of the distributions.

The following theorem shows how the fact that a matrix A is (ρ, f)-weak-learnable affects its separability properties. To remind the reader, we denote by $\mathbf{e}^i$ the vector whose ith element is 1 and the rest of its elements are zero. The notation $[\mathbf{x}]_+$ represents the vector whose ith element is $\max\{0, x_i\}$.

Theorem 7 *Let f be a convex function, ρ be a scalar, and A be a (ρ, f)-weak-learnable matrix. If the following assumptions hold,*
(i) $\min_{\mathbf{d}} f(\mathbf{d}) = 0$,
(ii) $\mathbf{0} \in \mathrm{core}(\mathrm{dom}(f))$,
(iii) $\forall \boldsymbol{\theta} \in \mathbb{R}^m,\ \forall i \in [m],\ \forall \alpha \in [0,1]$*, the Fenchel conjugate of f satisfies*

$$f^\star(\boldsymbol{\theta}) \;\geq\; f^\star(\boldsymbol{\theta} - \alpha\, \theta_i\, \mathbf{e}^i)$$

then,

$$\max_{\mathbf{w}\in\mathbb{B}^n_1, \gamma\in\mathbb{R}} \Big(\gamma - f^\star([\gamma - A\,\mathbf{w}]_+) \Big) \;=\; \rho \;\; .$$

The proof of the theorem is again based on the Fenchel duality theorem. The vector $[\gamma - A\,\mathbf{w}]_+$ appearing in the dual problem is the vector of hinge-losses. Before diving into the details of the proof, let us give two concrete family of functions that satisfy the requirement given in the theorem.

Example 1 *Let f be the indicator function of a ball of radius ν, $\{\mathbf{d} : \|\mathbf{d}\| \leq \nu\}$, where $\|\cdot\|$ is an arbitrary norm and ν is a scalar such that the intersection of this ball with the simplex is non-empty. Then, $f^\star(\mathbf{w}) = \nu\,\|\mathbf{w}\|_\star$ and the condition given in the theorem clearly holds. In this case, we obtain that*

$$\max_{\mathbf{w}\in\mathbb{B}^n_1, \gamma\in\mathbb{R}} \Big(\gamma - \nu\,\|[\gamma - A\,\mathbf{w}]_+\|_\star \Big) \;=\; \min_{d\in\mathbb{S}^m : \|\mathbf{d}\|\leq\nu} \|\mathbf{d}^\dagger A\|_\infty \; .$$

In particular, if $\|\cdot\|$ is the ℓ_∞ norm we obtain again the example of capped sample weights. Since the 1-norm and ∞-norm are dual we get that in the dual problem we are maximizing the margin parameter γ while minimizing the cumulative hinge-loss. Combining this fact with Corollary 5 we get that

$$\mathrm{AvgMin}_k(A\mathbf{w}) \;=\; \max_{\gamma\in\mathbb{R}} \Big(\gamma - \tfrac{1}{k} \sum_{i=1}^m [\gamma - (A\mathbf{w})_i]_+ \Big) \; .$$

The right hand side of the above is usually called the "soft-margin". The above equality tells us that the soft margin is equivalent to the average margin of the k worst examples (see also [WLR06, SSWB98]).

Example 2 *Let $f(\mathbf{d}) = \nu\,\|\mathbf{d}\|$ where $\|\cdot\|$ is an arbitrary norm and ν is a scalar. Then, $f^\star(\mathbf{w})$ is the indicator function of the ball of radius ν with respect to the dual norm $\{\mathbf{w} : \|\mathbf{w}\|_\star \leq \nu\}$. The condition given in the theorem clearly holds here as well and we obtain the dual problem*

$$\max_{\mathbf{w}\in\mathbb{B}^n_1, \gamma\in\mathbb{R}} \gamma \quad \text{s.t.} \quad \|[\gamma - A\,\mathbf{w}]_+\|_\star \leq \nu \; .$$

That is, we are now maximizing the margin subject to a constraint on the vector of hinge-losses.

We now turn to proving Thm. 7. First, we need the following lemma which characterizes the Fenchel conjugate of $f + I_{\mathbb{S}^m}$.

Lemma 8 *Assume that f satisfies the conditions given in Thm. 7 and denote $\tilde{f}(\mathbf{d}) = f(\mathbf{d}) + I_{\mathbb{S}^m}(\mathbf{d})$. Then,*

$$\tilde{f}^\star(\boldsymbol{\theta}) \;=\; -\max_{\gamma\in\mathbb{R}} \left(\gamma - f^\star([\gamma + \boldsymbol{\theta}]_+)\right) \;\; .$$

Proof: We first rewrite $\tilde{f}^\star$ as

$$\begin{aligned} \tilde{f}^\star(\boldsymbol{\theta}) &= \max_{\mathbf{d}} -f(\mathbf{d}) - (I_{\mathbb{S}^m}(\mathbf{d}) - \langle \boldsymbol{\theta}, \mathbf{d}\rangle) \\ &= -\left(\min_{\mathbf{d}} f(\mathbf{d}) + (I_{\mathbb{S}^m}(\mathbf{d}) - \langle \boldsymbol{\theta}, \mathbf{d}\rangle) \right) \end{aligned}$$

Denote $g(\mathbf{d}) = I_{\mathbb{S}^m}(\mathbf{d}) - \langle \boldsymbol{\theta}, \mathbf{d}\rangle$. It is easy to verify that $g^\star(\mathbf{x}) = \max_i(\theta_i + x_i)$. Next, note that $\mathbf{0} \in \mathrm{core}(\mathrm{dom}(f))$ by assumption and that $\mathrm{dom}(g) = \mathbb{S}^m$. Therefore, strong duality holds and we can use Thm. 3 which yields,

$$\begin{aligned} -\tilde{f}^\star(\boldsymbol{\theta}) &= \max_{\mathbf{x}} \left(-f^\star(\mathbf{x}) - g^\star(-\mathbf{x})\right) \\ &= \max_{\mathbf{x}} \Big(-f^\star(\mathbf{x}) - \max_i(\theta_i - x_i) \Big) \;\; . \end{aligned}$$

Let $C_\gamma = \{\mathbf{x} : \forall i, x_i \geq \theta_i + \gamma\}$. We show in the sequel that for any γ, the vector $[\boldsymbol{\theta} + \gamma]_+$ is a minimizer of $f^\star(\mathbf{x})$ over $\mathbf{x} \in C_\gamma$. Combining this with the above expression for $-\tilde{f}^\star(\boldsymbol{\theta})$ we get that

$$-\tilde{f}^\star(\boldsymbol{\theta}) \;=\; \max_{\gamma} \left(\gamma - f^\star([\boldsymbol{\theta} + \gamma]_+)\right) \;\; ,$$

as required. Therefore, it is left to show that the vector $[\boldsymbol{\theta} + \gamma]_+$ is indeed a minimizer of $f^\star(\mathbf{x})$ over C_γ. Clearly, $[\boldsymbol{\theta} + \gamma]_+ \in C$. In addition, for any $\mathbf{x} \in C_\gamma$ we can make a sequence of modifications to $\mathbf{x}$ until $\mathbf{x} = [\boldsymbol{\theta} + \gamma]_+$ as follows. Take some element i. If $x_i > [\theta_i + \gamma]_+$ then based on assumption (iii) of Thm. 7 we know that

$$f^\star\left(\mathbf{x} - \frac{x_i - [\theta_i + \gamma]_+}{x_i} x_i \mathbf{e}^i \right) \leq f^\star(\mathbf{x}) \; .$$

If $x_i < [\theta_i + \gamma]_+$ we must have that $[\theta_i + \gamma]_+ = 0$ since we assume that $\mathbf{x} \in C_\gamma$ and thus $x_i \geq \theta_i + \gamma$. Thus, $x_i < 0$ but now using assumption (iii) of Thm. 7 again we obtain that $f^\star(\mathbf{x} - x_i \mathbf{e}^i) \leq f^\star(\mathbf{x})$. Repeating this for every $i \in [m]$ makes $\mathbf{x}$ equals to $[\boldsymbol{\theta} + \gamma]_+$ while the value of $f^\star(\mathbf{x})$ is non-increasing along this process. We therefore conclude that $[\boldsymbol{\theta} + \gamma]_+$ is a minimizer of $f^\star(\mathbf{x})$ over $\mathbf{x} \in C_\gamma$ and our proof is concluded. ■

Based on the above lemma the proof of Thm. 7 is easily derived.

Proof:[of Thm. 7] The proof uses once more the Fenchel duality theorem. Define the function $\tilde{f}(\mathbf{d}) = f(\mathbf{d}) + I_{\mathbb{S}^m}(\mathbf{d})$. Therefore, Thm. 3 tells us that the dual of the problem $\min_{\mathbf{d}} \tilde{f}(\mathbf{d}) + \|\mathbf{d}^\dagger A\|_\infty$ is the problem $\max_{\mathbf{w}\in\mathbb{B}_1^n} \left(-\tilde{f}^\star(-A\mathbf{w})\right)$. Using Lemma 8 we obtain that the dual of the problem given in Definition 6 is the same maximization problem as stated in the theorem. To conclude the proof it is left to show that strong duality also holds here. First, using the assumption $\min_{\mathbf{d}} f(\mathbf{d}) = 0$ we get that $f^\star(\mathbf{0}) = 0$. By setting $\mathbf{w} = \mathbf{0}$ and $\gamma = 0$ we get that the dual problem is bounded below by zero. Thus, if $\rho = 0$ then strong duality holds. If $\rho > 0$ then we can use the fact that $\mathrm{dom}(\tilde{f}) \subseteq \mathrm{dom}(f)$ and therefore the same arguments as in the end of the proof of Thm. 4 holds here as well. ■

5 Boosting algorithms

In this section we derive a boosting algorithm for solving the max-relaxed-margin problem described in the previous section, namely,

$$\max_{\mathbf{w}\in\mathbb{B}_1^n} \max_{\gamma\in\mathbb{R}} \left(\gamma - f^\star([\gamma - A\,\mathbf{w}]_+)\right) \ . \tag{5}$$

The function $f^\star$ should satisfy the conditions stated in Thm. 7. In particular, if $f^\star(\mathbf{x}) = \nu\,\|\mathbf{x}\|_1$ we obtain the soft margin problem

$$\max_{\mathbf{w}\in\mathbb{B}_1^n} \max_{\gamma\in\mathbb{R}} \left(\gamma - \nu\sum_{i=1}^m [\gamma - (A\,\mathbf{w})_i]_+\right) \ , \tag{6}$$

while if $f^\star(\mathbf{x}) = \max_i x_i$ then we obtain the non-relaxed max margin problem

$$\max_{\mathbf{w}\in\mathbb{B}_1^n} \min_{i\in[m]} (A\,\mathbf{w})_i \ .$$

The boosting algorithm for solving Eq. (5) is described in Fig. 1. To simplify the presentation, let us first describe the algorithm for the non-relaxed max-margin problem, that is, $f^\star(\mathbf{x}) = \max_i x_i$. As we have shown in the proof of Thm. 4, the corresponding Fenchel conjugate $f(\mathbf{d})$ is the indicator function of $\mathbb{S}^m$. The algorithm initializes the weight vector to be the zero vector, $\mathbf{w}_1 = \mathbf{0}$. On round t, we define a distribution over the examples

$$\begin{aligned}\mathbf{d}_t &= \operatorname*{argmax}_{\mathbf{d}\in\mathbb{S}^m} \Big(\langle -A\,\mathbf{w}_t, \mathbf{d}\rangle - (f(\mathbf{d}) + \beta\,h(\mathbf{d}))\Big) \\ &= \operatorname*{argmin}_{\mathbf{d}\in\mathbb{S}^m} \Big(\langle A\,\mathbf{w}_t, \mathbf{d}\rangle + (f(\mathbf{d}) + \beta\,h(\mathbf{d}))\Big) \ ,\end{aligned}$$

where $h(\mathbf{d})$ is the relative entropy function. Since we are now dealing with the case $f(\mathbf{d}) = I_{\mathbb{S}^m}$, we can use Lemma 18 in the appendix and get that $\mathbf{d}_t$ is the gradient of the Fenchel conjugate of the function $\beta h(\mathbf{d})$. In the appendix we list several Fenchel conjugate pairs. In particular, the Fenchel conjugate of the relative entropy is the soft-max function

$$h^\star(\boldsymbol{\theta}) = \log\left(\frac{1}{m}\sum_{i=1}^m e^{\theta_i}\right) \ .$$

```
INPUT: matrix $A \in [+1,-1]^{m,n}$
       Relaxation function $f^\star$
       Desired accuracy $\epsilon$
DEFINE: $h(\mathbf{d}) = \sum_{i=1}^m d_i \log(d_i) + \log(m)$
        $f(d) =$ Fenchel conjugate of $f^\star$
INITIALIZE: $\mathbf{w}_1 = \mathbf{0},\ \beta = \frac{\epsilon}{2\log(m)}$
FOR $t = 1, 2, \ldots, T$
    $\mathbf{d}_t = \operatorname*{argmin}_{\mathbf{d}\in\mathbb{S}^m} \Big(\langle A\,\mathbf{w}_t, \mathbf{d}\rangle + (f(\mathbf{d}) + \beta\,h(\mathbf{d}))\Big)$
    $j_t \in \arg\max_j |(\mathbf{d}_t^\dagger A)_j|$
        (w.l.o.g. assume $\mathrm{sign}(\mathbf{d}_t^\dagger A)_{j_t} = 1$)
    $\eta_t = \max\left\{0, \min\left\{1, \frac{\beta\,\mathbf{d}_t^\dagger A(\mathbf{e}^{j_t} - \mathbf{w}_t)}{\|A(\mathbf{e}^{j_t} - \mathbf{w}_t)\|_\infty^2}\right\}\right\}$
    $\mathbf{w}_{t+1} = (1 - \eta_t)\mathbf{w}_t + \eta_t\,\mathbf{e}^{j_t}$
OUTPUT: $\mathbf{w}_{T+1}$
```

Figure 1: A Boosting Algorithm for maximizing the relaxed margin given in Eq. (5).

Using the property $(\beta h)^\star(\boldsymbol{\theta}) = \beta h^\star(\boldsymbol{\theta}/\beta)$ we obtain that

$$d_{t,i} \propto e^{-\frac{1}{\beta}(A\mathbf{w}_t)_i} \ .$$

That is, the log of the probability assigned to the ith example is negatively proportional to the margin of the example according to the current weight vector $\mathbf{w}_t$. Therefore, the algorithm allocates larger importance weights to the erroneous examples, in a similar fashion to the weighting scheme of examples of many other boosting algorithms, such as AdaBoost.

Next, we perform a step analogous to calling a weak-learner by finding a single column of A with the best edge. We would like to note that it is possible to extend the algorithm so that the weak learner may find a column whose edge is only approximately optimal. For simplicity we confine the description to weak learners that return the column with the largest edge. Finally, we set $\mathbf{w}_{t+1}$ to be the convex combination of $\mathbf{w}_t$ and the new hypothesis. The coefficient of the convex combination, denoted η_t, is calculated analytically based on our analysis. Note that the update form guarantees that $\|\mathbf{w}_t\|_1 \leq 1$ for all t.

The sole modification of the algorithm when running with other relaxation functions is concerned with the definition of $\mathbf{d}_t$. In Sec. 5.2 we further elaborate on how to solve the optimization problem which appears in the definition of $\mathbf{d}_t$. We provide a few general tools and also present an efficient procedure for the case where f is the indicator function of $\mathbb{B}_\infty^m(\nu)$.

The following theorem provides analysis of the rate of convergence of the algorithm.

Theorem 9 *Assume that the algorithm given in Fig. 1 is run for $T = \Omega(\log(m)/\epsilon^2)$ iterations. Then, the algorithm outputs an ϵ-accurate solution,*

$$\max_{\gamma} \left(\gamma - f^\star([\gamma - A\,\mathbf{w}_{T+1}]_+)\right) \ \geq\ \rho\ -\ \epsilon\ ,$$

where ρ is the optimal value of the solution as defined in Thm. 7.

Before turning into the proof of Thm. 9 let us first discuss its implications. First we note that the number of iterations of the algorithm upper bounds the number of non-zero elements of the solution. Therefore, we have a trade-off between the desired accuracy level, ϵ, and the level of sparsity of the solution, $\mathbf{w}_{T+1}$.

The algorithm can be used for maximizing the hard margin using $O(\log(m)/\epsilon^2)$ iterations. In this case, the algorithm shares the simplicity of the popular AdaBoost approach. The rate of convergence we obtain matches the rate of the AdaBoost$_\star$ described by Ratsch and Warmuth [RW05] and is better than the rate obtained in Rudin et al. [RSD07]. We note also that if A is γ-separable and we set $\epsilon = \gamma/2$ then we would find a solution with half the optimal margin in $O(\log(m)/\gamma^2)$ iterations. AdaBoost seemingly attains an exponentially fast decay of the empirical error of $e^{-\gamma^2 T}$. Thus, T should be at least $1/\gamma^2$. Further careful examination also reveals a factor of $\log(m)$ in the convergence rate of AdaBoost. Therefore, our algorithm attains the same rate of convergence of AdaBoost while both algorithms obtain a margin which is half of the optimal margin. (See also the margin analysis of AdaBoost described in Rudin et al. [RSD07].)

We can also use the algorithm for maximizing the soft margin given in Eq. (6). In Sec. 5.2 we show how to calculate $\mathbf{d}_t$ in $\tilde{O}(m)$ time. Therefore, the complexity of the resulting algorithm is roughly the same as the complexity of AdaBoost. The bound on the number of iterations that we obtain matches the bound of the SoftBoost algorithm, recently proposed by Warmuth et al. [WLR06]. However, our algorithm is simpler to implement and the time complexity of each iteration of our algorithm is substantially lower than the one described in [WLR06].

5.1 Proof of convergence rate

To motivate our proof technique, let us focus first on the max-margin case without any relaxation. As we showed before, the AdaBoost algorithm approximates the max operator, $\max_i \theta_i$, with a soft-max operator, $\log(\frac{1}{m}\sum_i e^{\theta_i})$, also known as the exp-loss. We can think of this approximation as another form of relaxation of the max margin. To distinguish this type of relaxation from the family of relaxations described in the previous section, we refer to it as an "algorithmic" relaxation, since this relaxation is driven by algorithmic factors and not directly by the concept of relaxing the margin. The algorithmic relaxation of AdaBoost encapsulates the following relaxation of weak learnability: replace the indicator function of the simplex with the relative entropy function over the simplex, which we denote by $h(\mathbf{d})$ (see also the definition in Fig. 1). The advantage of endowing the simplex with the relative entropy stems from the fact that the relative entropy is *strongly* convex with respect to the ℓ_1 norm, as we formally define now.

Definition 10 *A continuous function f is σ-strongly convex over a convex set S with respect to a norm $\|\cdot\|$ if S is contained in the domain of f and for all $\mathbf{v}, \mathbf{u} \in S$ and $\alpha \in [0,1]$ we have*

$$\begin{aligned} f(\alpha\,\mathbf{v} + (1-\alpha)\,\mathbf{u}) \quad &\le \quad \alpha\, f(\mathbf{v}) + (1-\alpha)\, f(\mathbf{u}) \\ &\quad - \frac{\sigma}{2}\,\alpha\,(1-\alpha)\,\|\mathbf{v}-\mathbf{u}\|^2 \ . \end{aligned}$$

In the above definition, if $\sigma = 0$ we revert back to the standard definition of convexity. Strong convexity quantifies the difference between the value of the function at the convex combination and the convex combination of the values of the function. The relative entropy is 1-strongly convex with respect to the ℓ_1 norm over the probabilistic simplex (see Lemma 16 in [SS07]). Few important properties of *strongly* convex functions are summarized in Lemma 18 (in the appendix). We use these properties in our proofs below.

Continuing with our motivating discussion, we view the algorithmic relaxation of AdaBoost as a replacement of the convex function $I_{\mathbb{S}^m}(\mathbf{d})$ by the strongly convex function $h(\mathbf{d})$. More generally, recall the definition $\tilde{f}(\mathbf{d}) = f(\mathbf{d}) + I_{\mathbb{S}^m}(\mathbf{d})$ from Sec. 4 and that solving Eq. (5) is equivalent to maximizing $-\tilde{f}^\star(-A\,\mathbf{w})$ over $\mathbf{w} \in \mathbb{B}_1^n$. As in the algorithmic relaxation of AdaBoost, we replace $\tilde{f}(\mathbf{d})$ by the function

$$\hat{f}(\mathbf{d}) \;=\; \tilde{f}(\mathbf{d}) + \beta\, h(\mathbf{d}) \ ,$$

where $\beta \in (0,1)$. Since for all $\mathbf{d} \in \mathbb{S}^m$ we have $0 \le h(\mathbf{d}) \le \log(m)$, by setting $\beta = \epsilon/(2\log(m))$ we obtain that

$$\forall \mathbf{d} \in \mathbb{S}^m,\ \ \hat{f}(\mathbf{d}) - \epsilon/2 \;\le\; \tilde{f}(\mathbf{d}) \;\le\; \hat{f}(\mathbf{d}) \ .$$

Using Lemma 19 in the appendix we obtain that

$$\forall \boldsymbol{\theta},\ \ \hat{f}^\star(\boldsymbol{\theta}) \;\le\; \tilde{f}^\star(\boldsymbol{\theta}) \;\le\; \hat{f}^\star(\boldsymbol{\theta}) + \epsilon/2 \ . \qquad (7)$$

The above implies that maximizing $-\hat{f}^\star(-A\mathbf{w})$ gives an $\epsilon/2$ accurate solution to the problem of maximizing $-\tilde{f}^\star(-A\mathbf{w})$. This argument holds for the entire family of functions discussed in Sec. 4. An appealing property of strong convexity that we exploit is that by adding a convex function to a strongly convex function we retain at least the same strong convexity level. Therefore, for all the functions $\tilde{f}(\mathbf{d})$ discussed in Sec. 4 the corresponding $\hat{f}(\mathbf{d})$ retains the strongly convex property of the relative entropy.

The algorithm in Fig. 1 is designed for maximizing $-\hat{f}^\star(-A\mathbf{w})$ over $\mathbb{B}_1^n$. Based on the above discussion, this maximization translates to an approximate maximization of $-\tilde{f}^\star(-A\,\mathbf{w})$. Using again Thm. 3 we obtain that

$$\max_{\mathbf{w}\in\mathbb{B}_1^n} -\hat{f}^\star(-A\,\mathbf{w}) \;\le\; \min_{\mathbf{d}} \hat{f}(\mathbf{d}) + \|\mathbf{d}^\dagger A\|_\infty \ .$$

Denote by $\mathcal{D}(\mathbf{w})$ and $\mathcal{P}(\mathbf{d})$ the dual and primal objective values of the above equation. We also denote by ϵ_t the sub-optimality value attained at iteration t of the algorithm, namely,

$$\epsilon_t = \max_{\mathbf{w}\in\mathbb{B}_1^n} \mathcal{D}(\mathbf{w}) - \mathcal{D}(\mathbf{w}_t) \ .$$

The following key lemma lower bounds the improvement of the algorithm in terms of its current sub-optimality.

Lemma 11 *Let ϵ_t be the sub-optimality value of the algorithm in Fig. 1 at iteration t and assume that $\epsilon_t \le 1$. Then, $\epsilon_t - \epsilon_{t+1} \;\ge\; \beta\,\epsilon_t^2/8$.*

Proof: Denote $\Delta_t = \epsilon_t - \epsilon_{t+1}$ and based on the definition of ϵ_t we clearly have that $\Delta_t = \mathcal{D}(\mathbf{w}_{t+1}) - \mathcal{D}(\mathbf{w}_t)$. To simplify our notation, we use the shorthand j for j_t and η for η_t. Since

$$\mathbf{w}_{t+1} = (1-\eta)\mathbf{w}_t + \eta \mathbf{e}^j$$

we get that

$$\Delta_t \;=\; \mathcal{D}(\mathbf{w}_t + \eta(\mathbf{e}^j - \mathbf{w}_t)) - \mathcal{D}(\mathbf{w}_t)\ \ .$$

Using the definition of $\mathcal{D}$ we further rewrite Δ_t as

$$\Delta_t \;=\; \hat{f}^\star(-A\mathbf{w}_t) - \hat{f}^\star(-A\mathbf{w}_t - \eta\, A\,(\mathbf{e}^j - \mathbf{w}_t))\ . \tag{8}$$

The key property that we use is that $\hat{f}^\star$ is the Fenchel conjugate of a β-strongly convex function over the simplex with respect to the ℓ_1 norm. Therefore, using Lemma 18 in the appendix, we know that for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$:

$$\hat{f}^\star(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - \hat{f}^\star(\boldsymbol{\theta}_1) \;\le\; \langle \nabla, \boldsymbol{\theta}_2\rangle + \frac{\|\boldsymbol{\theta}_2\|_\infty^2}{2\,\beta}\ ,$$

where $\nabla = \arg\max_{\mathbf{d}} \langle \boldsymbol{\theta}_1, \mathbf{d}\rangle - \hat{f}(\mathbf{d})$. Combining this property with Eq. (8) and using the definition of $\mathbf{d}_t$ we obtain that

$$\Delta_t \;\ge\; \eta\,\langle \mathbf{d}_t, A\,(\mathbf{e}^j - \mathbf{w}_t)\rangle - \frac{\eta^2\,\|A\,(\mathbf{e}^j - \mathbf{w}_t)\|_\infty^2}{2\,\beta}\ . \tag{9}$$

Using the assumption $A \in [+1,-1]^{m\times n}$, the fact that $\mathbf{w}_t \in \mathbb{B}_1^n$, and the triangle inequality we get that

$$\|A\,(\mathbf{e}^j - \mathbf{w}_t)\|_\infty \le 2$$

and thus

$$\Delta_t \;\ge\; \eta\,\langle \mathbf{d}_t, A\,(\mathbf{e}^j - \mathbf{w}_t)\rangle - 2\,\eta^2/\beta\ . \tag{10}$$

Next, we show that $\langle \mathbf{d}_t, A\,(\mathbf{e}^j - \mathbf{w}_t)\rangle = \mathcal{P}(\mathbf{d}_t) - \mathcal{D}(\mathbf{w}_t)$. To to so, we first use Lemma 17 to get that $\langle \mathbf{d}_t, -A\,\mathbf{w}_t\rangle = \hat{f}(\mathbf{d}_t) + \hat{f}^\star(-A\,\mathbf{w}_t)$ and second we use the definition of j to get that $\langle \mathbf{d}_t, A\,\mathbf{e}^j\rangle = \|\mathbf{d}_t^\dagger\,A\|_\infty$. Combining this with Eq. (10) yields

$$\Delta_t \;\ge\; \eta\,(\mathcal{P}(\mathbf{d}_t) - \mathcal{D}(\mathbf{w}_t)) - 2\,\eta^2/\beta\ . \tag{11}$$

The weak duality property tells us that $\mathcal{P}(\mathbf{d}_t) \ge \max_{\mathbf{w}\in\mathbb{B}_1^n} \mathcal{D}(\mathbf{w})$ and therefore $\Delta_t \ge \eta\,\epsilon_t - 2\,\eta^2/\beta$. Denote $\eta' = \epsilon_t\,\beta/4$ and note that $\eta' \in [0,1]$. Had we set $\eta_t = \eta'$ we could have obtained that $\Delta_t \ge \beta\,\epsilon_t^2/8$ as required. Since we set η_t to be the maximizer of the expression in Eq. (9) over $[0,1]$, we get an even larger value for Δ_t. This concludes our proof. ■

Based on Lemma 11 the proof of Thm. 9 easily follows.

Proof:[(of Thm. 9)] We first show that $\epsilon_1 \le 1$. To see this, we use the weak duality to get that $\epsilon_1 \le \mathcal{P}(\mathbf{d}_1) - \mathcal{D}(\mathbf{w}_1)$. Next, we recall that in the proof of Lemma 11 we have shown that for all t, $\mathcal{P}(\mathbf{d}_t) - \mathcal{D}(\mathbf{w}_t) = \langle \mathbf{d}_t, A(\mathbf{e}^{j_t} - \mathbf{w}_t)\rangle$. Since $\mathbf{w}_1 = \mathbf{0}$ we get that $\epsilon_1 \le \langle \mathbf{d}_1, A\mathbf{e}^{j_1}\rangle = \|\mathbf{d}_1^\dagger A\|_\infty \le 1$.

We can now apply Lemma 11 for $t = 1$ and get that $\epsilon_2 \le \epsilon_1$. By induction, we obtain that Lemma 11 holds for all t. Applying Lemma 20 (given in the appendix) we get that $\epsilon_t \le \frac{8}{\beta(t+1)}$.

Plugging the definition of $\beta = \epsilon/(2\log(m))$ into the upper bound on ϵ_{T+1} we get $\epsilon_{T+1} \le \frac{16\log(m)}{(T+2)\epsilon}$. Therefore, if $T+2 \ge 32\log(m)/\epsilon^2$ we get that $\epsilon_{T+1} \le \epsilon/2$. Finally, Let ϵ' be the error of $\mathbf{w}_{T+1}$ on the original $\tilde{f}$ then using Eq. (7) we obtain that $\epsilon' \le \epsilon_{T+1} + \epsilon/2 = \epsilon$. ■

5.2 Efficient implementation for soft margin

In this section we provide an efficient procedure for calculating the distribution $\mathbf{d}_t$ as described in Fig. 1 when $f(\mathbf{d})$ is the indicator function of $\{\mathbf{d} : \|\mathbf{d}\|_\infty \le \nu\}$. As we showed above, this case corresponds to the maximization of the soft margin.

We first present a lemma that provides us with an alternative method for finding $\mathbf{d}$, which is based on Bregman divergences. The Bregman divergence with respect to a convex function h between two vectors $\mathbf{d}$ and $\mathbf{d}_0$ is defined as,

$$B_h(\mathbf{d}\|\mathbf{d}_0) \;=\; h(\mathbf{d}) - h(\mathbf{d}_0) - \langle \nabla h(\mathbf{d}_0), \mathbf{d} - \mathbf{d}_0\rangle\ .$$

See [CZ97] for a rigorous definition of the Bregman divergence.

Lemma 12 *Let $h : S \to \mathbb{R}$ be a strongly convex and differentiable function, let f be a convex function, and denote $\hat{f} = h + f$. Let $\boldsymbol{\theta}$ be a vector and denote $\mathbf{d}_0 = \nabla h^\star(\boldsymbol{\theta})$, where $h^\star$ is the Fenchel conjugate of h. Then,*

$$\nabla \hat{f}^\star(\boldsymbol{\theta}) \;=\; \operatorname*{argmin}_{\mathbf{d}}\ (B_h(\mathbf{d}\|\mathbf{d}_0) + f(\mathbf{d}))\ \ .$$

Proof: Since h is strongly convex and differentiable we have that $\nabla h(\mathbf{d}_0) = \boldsymbol{\theta}$. Therefore,

$$\begin{aligned}
\nabla \hat{f}^\star(\boldsymbol{\theta}) \;&=\; \operatorname*{argmax}_{\mathbf{d}}\ \langle \mathbf{d}, \boldsymbol{\theta}\rangle - \hat{f}(\mathbf{d}) \\
&=\; \operatorname*{argmin}_{\mathbf{d}}\ h(\mathbf{d}) - \langle \mathbf{d}, \boldsymbol{\theta}\rangle + f(\mathbf{d}) \\
&=\; \operatorname*{argmin}_{\mathbf{d}}\ h(\mathbf{d}) - \langle \mathbf{d}, \nabla h(\mathbf{d}_0)\rangle + f(\mathbf{d}) \\
&=\; \operatorname*{argmin}_{\mathbf{d}}\ B_h(\mathbf{d}\|\mathbf{d}_0) + f(\mathbf{d})\ \ .
\end{aligned}$$

■

Applying the above lemma with $f = I_C$ for some convex set C we obtain the following corollary.

Corollary 13 *Assume that the conditions stated in Lemma 12 hold and that $f(\mathbf{d}) = I_C(\mathbf{d})$ for some convex set C. Then,*

$$\nabla(h+f)^\star(\boldsymbol{\theta}) \;=\; \operatorname*{argmin}_{\mathbf{d}\in C} B_h(\mathbf{d}\|\nabla h^\star(\boldsymbol{\theta}))\ \ .$$

We now get back to the problem of finding $\mathbf{d}_t$ when $f(\mathbf{d})$ is $I_C(\mathbf{d})$ for $C = \{\mathbf{d} : \|\mathbf{d}\|_\infty \le \nu\}$. Based on Corollary 13 we can first define the distribution vector $\mathbf{d}_0$ such that $\mathbf{d}_{0,i} \propto \exp(-\frac{1}{\beta}(A\mathbf{w}_t)_i)$ and then set

$$\mathbf{d}_t \;=\; \operatorname*{argmin}_{\mathbf{d}\in\mathbb{S}^m : \|\mathbf{d}\|_\infty \le \nu}\quad B_h(\mathbf{d}\|\mathbf{d}_0)\ \ . \tag{12}$$

We are therefore left with the problem of solving the entropic projection problem given in Eq. (12). A similar problem was tackled by Herbster and Warmuth [HW01], who provided $O(m\log(m))$ and $O(m)$ algorithms for performing entropic projections. For completeness, in the rest of this section we outline the simpler $O(m\log(m))$ algorithm. To do so, we first show that the entropic projection preserves the relative order of components of the projected vector.

Lemma 14 *Let $\mathbf{d}_t$ be the solution of Eq. (12) and let i, j be two indices such that $d_{0,i} > d_{0,j}$. Then, $d_{t,i} \ge d_{t,j}$.*

```
INPUT: A vector $\mathbf{d}_0 \in \mathbb{S}^m$ and a scalar $\nu \in (0,1)$
    Sort $\mathbf{d}_0$ in non-increasing order $\Rightarrow$ $\mathbf{u}$
INITIALIZE: $Z = \sum_{r=1}^{m} u_r$
FOR $i = 1, ..., m$
    $\theta = \frac{1 - \nu\,(i-1)}{Z}$
    IF $\theta u_i \leq \nu$
        BREAK
    ENDIF
    $Z \leftarrow Z - u_i$
ENDFOR
OUTPUT: $\mathbf{d}_t$ s.t. $d_{t,r} = \min\{\nu, \theta d_{0,r}\}$
```

Figure 2: An $O(m\log(m))$ Procedure for solving the Entropic Projection problem defined by Eq. (12).

Proof: Assume that the claim of the proof is not true. Let i and j be two indices which violate the claim, therefore $d_{t,i} < d_{t,j}$. We now construct a vector $\tilde{\mathbf{d}}$ which resides in $\mathbb{S}^m$ and whose components do not exceed ν. We set all the components of $\tilde{d}_t$, except for the ith and jth components, to be equal to the corresponding components of $\mathbf{d}_t$. Next, we set $\tilde{d}_{t,i} = d_{t,j}$ and $\tilde{d}_{t,j} = d_{t,i}$. Clearly, $\tilde{d}_t$ constitutes a feasible solution. Taking the difference between the Bregman divergence of the two vectors each to $\mathbf{d}_0$ we get,

$$B_h(\mathbf{d}_t \| d_0) - B_h(\tilde{\mathbf{d}}_t \| \mathbf{d}_0) = (d_j - d_i)\log(d_{0,i}/d_{0,j}) > 0 \ ,$$

which contradicts the fact that $\mathbf{d}_t$ is the vector attaining the smallest Bregman divergence to $\mathbf{d}_0$. ■

Without loss of generality, assume that $\mathbf{d}_0$ is sorted in a non-increasing order. Therefore, using Lemma 14 we know that $\mathbf{d}_t$ has the form $(\nu, \ldots, \nu, d_{t,i}, \ldots, d_{t,j}, 0, \ldots, 0)$ where for each $r \in \{i, \ldots, j\}$ we have $d_{t,r} \in (0, \nu)$. Moreover, the following lemma provides us with a simple way to find all the rest of the elements of $\mathbf{d}_t$.

Lemma 15 *Assume that $\mathbf{d}_0$ is sorted in a non-increasing order and that $\mathbf{d}_t = (\nu, \ldots, \nu, d_{t,i}, \ldots, d_{t,j}, 0, \ldots, 0)$. Then, for all $r \in \{i, \ldots, j\}$ we have*

$$d_{t,r} \ = \ \theta\, d_{0,r} \ \text{ where } \ \theta = \frac{1 - \nu\,(i-1)}{\sum_{r=i}^{j} d_{0,r}} \ .$$

Proof: Let $\mathbf{v}$ denotes the gradient of $B_h(\mathbf{d}\|\mathbf{d}_0)$ with respect to $\mathbf{d}$ at $\mathbf{d}_t$, namely,

$$v_i \ = \ \log(d_{t,i}) + 1 - \log(d_{0,i}) \ .$$

Let $I = \{i, \ldots, j\}$. Note that for the elements in I the optimization problem has a single linear equality constraint and the solution is in the interior of the set $(0,\nu)^{|I|}$. Therefore, using Corollary 2.1.3 in [BL06] we obtain that there exists a constant θ' such that for all $i \in I$, $v_i = \theta' - 1$ or equivalently

$$\forall i \in I, \ d_{t,i} = d_{t,0}\, e^{\theta' - 1} \ .$$

Let us denote $\theta = e^{\theta'-1}$. Using this form in the equation $\sum_i d_{t,i} = 1$ we get that,

$$1 = \sum_{r=1}^{m} d_{t,r} = \nu(i-1) + \theta \sum_{r=i}^{j} d_{0,r} \ ,$$

which immediately yields that θ attains the value stated in the lemma. ■

We are left with the problem of finding the indices i and j. The next lemma tells us that not a single element of the optimal vector attains a value of zero.

Lemma 16 *Assume that the vector $\mathbf{d}_0$ is provided in a non-increasing order of elements and that all of its elements are positive. Then, the optimal solution of Eq. (12) is of the form, $(\nu, \ldots, \nu, d_{t,i}, \ldots, d_{t,m})$ where $d_{t,m} > 0$.*

Proof: Plugging the value of θ from the previous lemma into the objective function and performing simple algebraic manipulations we obtain the following objective value,

$$B_h(\mathbf{d}_t \| \mathbf{d}_0) = \sum_{r=1}^{i-1} \nu \log(\tfrac{\nu}{d_{0,r}}) + (1 - \nu(i-1))\log(\theta) \ .$$

Therefore, the objective is monotonically increasing in θ. This in turn implies that we should set θ to be as small as possible in order to find the minimal Bregman divergence. Next, note that the value of θ as defined in Lemma 15 is decreasing as a function of j. The optimal solution is obtained for $j = m$. ■

Finally, we are left with the task of finding the index i. Once it is found we readily obtain θ, which immediately translates into a closed form solution for $\mathbf{d}_t$. Lemma 14 in conjunction with a property presented in the sequel, implies that the *first* index for which $\mathbf{d}_t$, as defined by Lemma 15 with $j = m$, constitutes the optimal index for i. The pseudocode describing the resulting efficient procedure for solving the problem in Eq. (12) is given in Fig. 2. The algorithm starts by sorting the vector $\mathbf{d}_0$. Then, it checks each possible index i of the sorted vector as the position to stop capping the weights. More formally, given an index i the algorithm checks whether $\mathbf{d}_t$ can take the form $(\nu, \ldots, \nu, d_{t,i}, \ldots, d_{t,m})$ where $d_{t,i} < \nu$. To check each index i the algorithm calculates θ as given by Lemma 15. The same lemma also implies that $d_{t,i} = \theta d_{0,i}$. Thus, if the assumption on the index i is correct, the following inequality must hold, $\nu > d_{t,i} = \theta d_{0,i}$. In case the index i under examination indeed satisfies the inequality the algorithm breaks out of the loop. Therefore, the algorithm outputs the feasible solution with the smallest number of weights at the bound ν. It thus remains to verify that the feasible solution with the smallest number of capped weights is indeed optimal. This property follows from a fairly straightforward yet tedious lemma which generalizes Lemma 3 from [SSS06b] and is thus omitted. Note also that the time complexity of the resulting algorithm is $O(m\,\log(m)))$ which renders it applicable to boosting-based applications with large datasets.

6 Discussion

The starting point of this paper was an alternative view of the equivalence of weak-learnability and linear-separability. This view lead us to derive new relaxations of the notion of margin, which are useful in the noisy non-separable case. In turn, the new relaxations of the margin motivated us to derive new boosting algorithms which maintain distributions over the examples that are restricted to a subset of the simplex. There are a few future direction research we plan to pursue. First, we would like to further explore additional constraints of the distribution $\mathbf{d}_t$, such as adding ℓ_2 constraints. We also would like to replace the relative entropy penalty for the distribution $\mathbf{d}_t$ with binary entropies of each of the components of $\mathbf{d}_t$ with respect to the two dimensional vector $(\frac{1}{2}, \frac{1}{2})$. The result is a boosting-based apparatus for the log-loss. Last, we would like to explore alternative formalisms for the primal problem that also modify the definition of the function $g(\mathbf{d}) = \|\mathbf{d}^\dagger A\|_\infty$, which may lead to a regularization term of the vector $\mathbf{w}$ rather than the domain constraint we currently have.

A Technical lemmas

The first lemma states a sufficient condition under which the Fenchel-Young inequality holds with equality. Its proof can be found in ([BL06], Proposition 3.3.4).

Lemma 17 *Let f be a closed and convex function and let $\partial f(\mathbf{w})$ be its differential set at $\mathbf{w}$. Then, for all $\boldsymbol{\theta} \in \partial f(\mathbf{w})$ we have, $f(\mathbf{w}) + f^\star(\boldsymbol{\theta}) \;=\; \langle \boldsymbol{\theta}, \mathbf{w} \rangle$.*

The next lemma underscores the importance of strongly convex functions. The proof of this lemma follows from Lemma 18 in [SS07].

Lemma 18 *Let f be a closed and σ-strongly convex function over S with respect to a norm $\|\cdot\|$. Let $f^\star$ be the Fenchel conjugate of f. Then, $f^\star$ is differentiable and its gradient satisfies $\nabla f^\star(\boldsymbol{\theta}) = \arg\max_{\mathbf{w}\in S}\ \langle \mathbf{w}, \boldsymbol{\theta}\rangle - f(\mathbf{w})$. Furthermore, for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^n$, we have*

$$f^\star(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - f^\star(\boldsymbol{\theta}_1) \;\le\; \langle \nabla f^\star(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 \rangle + \frac{1}{2\sigma}\|\boldsymbol{\theta}_2\|_\star^2$$

Lemma 19 *Let f, g be two functions and assume that for all $w \in S$ we have $g(\mathbf{w}) \ge f(\mathbf{w}) \ge g(\mathbf{w}) - c$ for some constant c. Then, $g^\star(\boldsymbol{\theta}) \le f^\star(\boldsymbol{\theta}) \le g^\star(\boldsymbol{\theta}) + c$.*

Proof: There exists some $\mathbf{w}'$ s.t.

$$\begin{aligned} g^\star(\boldsymbol{\theta}) &= \langle \mathbf{w}', \boldsymbol{\theta}\rangle - g(\mathbf{w}') \\ &\le \langle \mathbf{w}', \boldsymbol{\theta}\rangle - f(\mathbf{w}') \\ &\le \max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta}\rangle - f(\mathbf{w}) = f^\star(\boldsymbol{\theta}) \ . \end{aligned}$$

This proves the first inequality. The second inequality follows from the fact that the conjugate of $g(\mathbf{w}) - c$ is $g^\star(\boldsymbol{\theta}) + c$. ■

Lemma 20 *Let $1 \ge \epsilon_1 \ge \epsilon_2 \ge \ldots$ be a sequence such that for all $t \ge 1$ we have $\epsilon_t - \epsilon_{t+1} \ge r\,\epsilon_t^2$ for some constant $r \in (0, 1/2)$. Then, for all t we have $\epsilon_t \le \frac{1}{r(t+1)}$.*

Proof: We prove the lemma by induction. First, for $t = 1$ we have $\frac{1}{r(t+1)} = \frac{1}{2r} \ge 1$ and the claim clearly holds. Assume that the claim holds for some t. Then,

$$\epsilon_{t+1} \le \epsilon_t - r\epsilon_t^2 \le \tfrac{1}{r(t+1)} - \tfrac{1}{r(t+1)^2} \ , \tag{13}$$

where we used the fact that the function $x - rx^2$ is monotonically increasing in $[0, 1/(2r)]$ along with the inductive assumption. We can rewrite the right-hand side of Eq. (13) as

$$\tfrac{1}{r(t+2)} \left(\tfrac{(t+1)+1}{t+1} \cdot \tfrac{(t+1)-1}{t+1}\right) \;=\; \tfrac{1}{r(t+2)} \left(\tfrac{(t+1)^2-1}{(t+1)^2}\right) \ .$$

The term $\frac{(t+1)^2-1}{(t+1)^2}$ is smaller than 1 and thus $\epsilon_{t+1} \le \frac{1}{r(t+2)}$, which concludes our proof. ■

B Fenchel conjugate pairs

We now list a few useful Fenchel-conjugate pairs. Proofs can be found in ([BV04] Section 3.3, [BL06] Section 3.3., [SS07] Section A.3).

$f(\mathbf{d})$	$f^\star(\boldsymbol{\theta})$
$I_C(\mathbf{d})$ for $C = \{\mathbf{d} : \|\mathbf{d}\| \le \nu\}$	$\nu\,\|\boldsymbol{\theta}\|_\star$
$I_{\mathbb{S}^m}(\mathbf{d})$	$\max_i \theta_i$
$I_{\mathbb{S}^m}(\mathbf{d}) + \sum_{i=1}^m d_i \log(\frac{d_i}{1/m})$	$\log\left(\frac{1}{m}\sum_{i=1}^m e^{\theta_i}\right)$
$\frac{1}{2}\|\mathbf{d}\|^2$	$\frac{1}{2}\|\boldsymbol{\theta}\|_\star^2$
$c\,f(\mathbf{d})$ for $c > 0$	$c\,f^\star(\boldsymbol{\theta}/c)$
$f(\mathbf{d} + \mathbf{d}_0)$	$f^\star(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \mathbf{d}_0\rangle$
$f(c\,\mathbf{d})$ for $c \ne 0$	$f^\star(\boldsymbol{\theta}/c)$

References

[BL06] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CSS02] M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 47(2/3):253–285, 2002.

[CST00] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[CZ97] Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York, NY, USA, 1997.

[DW00] C. Domingo and O. Watanabe. Madaboost: A modification of adaboost. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[Fre01] Y. Freund. An adaptive version of the boost by majority algorithm. *Machine Learning*, 43(3):293–318, 2001.

[FS96] Y. Freund and R.E. Schapire. Game theory, online prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 325–332, 1996.

[FS99] Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[HW01] M. Herbster and M. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

[KPL01] V. Koltchinskii, D. Panchenko, and F. Lozano. Some new bounds on the generalization error of combined classifiers. In *Advances in Neural Information Processing Systems 14*, 2001.

[MBB98] Llew Mason, Peter Bartlett, and Jonathan Baxter. Direct optimization of margins improves generalization in combined classifiers. Technical report, Deparment of Systems Engineering, Australian National University, 1998.

[MR03] R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning*, pages 119–184. Springer, 2003.

[RSD07] C. Rudin, R.E. Schapire, and I. Daubechies. Analysis of boosting algorithms using the smooth margin function. *Annals of Statistics,*, 2007.

[RW05] G. Ratsch and M. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, pages 2153–2175, 2005.

[Sch90] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[Sch03] R.E. Schapire. The boosting approach to machine learning: An overview. In D.D. Denison, M.H. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*. Springer, 2003.

[Ser03] R.A. Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4:633–648, 2003.

[SFBL97] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 322–330, 1997. To appear, *The Annals of Statistics*.

[SS07] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

[SSS06a] S. Shalev-Shwartz and Y. Singer. Convex repeated games and fenchel duality. In *Advances in Neural Information Processing Systems 20*, 2006.

[SSS06b] S. Shalev-Shwartz and Y. Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7 (July):1567–1599, 2006.

[SSS07] S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning Journal*, 2007.

[SSWB98] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. Technical Report NC2-TR-1998-053, NeuroColt2, 1998.

[SVL07] A. Smola, S.V.N. Vishwanathan, and Q. Le. Bundle methods for machine learning. In *Advances in Neural Information Processing Systems 21*, 2007.

[vN28] J. von Neumann. Zur theorie der gesellschaftsspiele (on the theory of parlor games). *Math. Ann.*, 100:295–320, 1928.

[WGR07] M. Warmuth, K. Glocer, and G. Ratsch. Boosting algorithms for maximizing the soft margin. In *Advances in Neural Information Processing Systems 21*, 2007.

[WLR06] M. Warmuth, J. Liao, and G. Ratsch. Totally corrective boosting algorithms that maximize the margin. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

[Zha03] T. Zhang. Sequential greedy approximation for certain convex optimization problems. *IEEE Transaction on Information Theory*, 49:682–691, 2003.