Mach Learn (2010) 80: 141–163
DOI 10.1007/s10994-010-5173-z

On the equivalence of weak learnability and linear separability: new relaxations and efficient boosting algorithms

Shai Shalev-Shwartz · Yoram Singer

Received: 15 March 2009 / Accepted: 1 November 2009 / Published online: 4 May 2010

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. The equivalence is a direct consequence of von Neumann's minimax theorem. Nonetheless, 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.

Keywords Boosting · Margin · Linear separability · Minimax theorem

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 Freund and Schapire (1999), Schapire (2003), Meir and Rätsch (2003). The first boosting algorithm was used for showing the equivalence between weak learnability and strong learnability (Schapire

Editors: Sham Kakade and Ping Li.

A short version of this paper appeared in COLT 2008.

S. Shalev-Shwartz (✉)
School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel
e-mail: shais@cs.huji.ac.il

Y. Singer
Google Inc., Mountain View, USA
e-mail: singer@google.com

1990). 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 (1990) 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 building a majority tree of weak hypotheses. Shortly after Schapire's seminal paper, Freund (2001) devised the boost-by-majority algorithm which directly implied 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. (1997) showed that AdaBoost in fact finds a linear separator with a large margin. However, AdaBoost does not converge to the max margin solution (Ratsch and Warmuth 2005; Rudin et al. 2007). 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 Theorem 4. We note that the equivalence can be also derived from von Neumann's minimax theorem (von Neumann 1928). 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 except for merely 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 outlined 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 of weak learnability and linear separability with ℓ_1 margin is a direct consequence of von Neumann's minimax theorem in game theory (von Neumann 1928). Freund and Schapire (1996) 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 later spelled out in Ratsch and Warmuth (2005).

Schapire et al. (1997) showed that the AdaBoost algorithm finds a large margin solution. However, as pointed out by Ratsch and Warmuth (2005), Rudin et al. (2007), AdaBoost does not converge to the max margin solution. Ratsch and Warmuth (2005) 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. (2007) 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, numerous boosting algorithms were suggested for non-separable datasets. We list here few examples. The LogLoss Boost algorithm (Collins et al. 2002) tries to minimize the cumulative logistic loss, which is less sensitive to noise. MadaBoost (Domingo and Watanabe 2000) is another example of an algorithm that copes with non-separability. It does so by capping from 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 considered to be noise tolerant is the BrownBoost algorithm (Freund 2001). 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 Cristianini and Shawe-Taylor 2000). Warmuth et al. (2007) recently proposed the SoftBoost algorithm that directly minimizes the hinge-loss function. The relaxation described in Warmuth et al. (2007) 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 (Servedio 2003). 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 Warmuth et al. (2007). More importantly, in this paper we present a simpler algorithm which does not employ a convex optimization procedure on each round of boosting. Our approach stands in contrast to the algorithm of Warmuth et al. (2006, 2007, 2008), which requires "totally corrective" updates and entails solving a rather complex optimization problem on each iteration. See also the discussion in Sect. 5.2.

The family of boosting algorithms we derive is reminiscent of the boosting algorithm proposed by Zhang (2003). 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 (Shalev-Shwartz and Singer 2006b, 2007; Shalev-Shwartz 2007) and also borrows ideas from the analysis given in Smola et al. (2007). The main difference between our analysis and that of Smola et al. (2007), Zhang (2003) 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 rather 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 Schapire et al. (1997), Mason et al. (1998), Koltchinskii et al. (2001). 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 $\mathbf{w}$. We use the notation $\mathbf{d}^T, \mathbf{w}^T$ for denoting their corresponding row vectors and by A^T the transpose of the matrix A. The inner product between vectors is denoted by $\langle \mathbf{d}, \mathbf{w} \rangle = \mathbf{d}^T \mathbf{w}$. 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 equal 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 greatest scalar that satisfies the above inequality, namely,

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

Rewriting the above using the more familiar notation of examples $(\mathbf{x}_i, y_i)$ and feature vectors $\boldsymbol{\phi}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_n(\mathbf{x}))$ we obtain

$$\gamma = \max_{\mathbf{w} \in \mathbb{B}_1^n} \min_i y_i \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i) \rangle .$$

This is similar to the (hard-margin) Support Vector Machine (Vapnik 1998), with the important difference that here we constraint $\mathbf{w}$ to be in the ℓ_1 unit ball while in Support Vector Machine we constraint $\mathbf{w}$ to be in the ℓ_2 unit ball. Next, we formally define weak learnability.

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 $|(\mathbf{d}^T A)_j| \geq \gamma$, and γ is the greatest scalar that satisfies the above inequality, namely,

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

The quantity $(\mathbf{d}^T A)_j = \sum_{i=1}^m d_i y_i h_j(\mathbf{x}_i)$ is often called the *edge* of the jth hypothesis. The analysis of many boosting algorithms (including AdaBoost) assumes that for any distribution there exists an hypothesis with an edge of at least γ.

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 $\mathbf{x} \in C$ 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

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

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: Borwein and Lewis (2006, Theorem 3.3.5)) *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,*

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

The above holds with equality if in addition we have

$$\mathbf{0} \in \mathrm{core}\left(\mathrm{dom}(g) - A^T \mathrm{dom}(f)\right),$$

where $-$ denotes the set exclusion operator.

Note that the first part of the theorem is analogous to the notion of weak-duality while additional condition yields the equivalent of strong duality. 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.

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. The theorem is by no means new (see for instance Freund and Schapire 1996) and its role is to pave the road for the analysis presented in the sequel and underscore the usage of Fenchel duality which is used extensively throughout the paper.

Table 1 Summary of notations

$\mathbf{x}, \mathbf{x}^T$	column vector and its transpose
$\langle \mathbf{x}, \mathbf{v} \rangle$	inner product ($= \mathbf{x}^T \mathbf{v}$)
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 $\{1, \ldots, m\}$

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 (Theorem 3). For convenience, we refer to the optimization problem on the right (left) hand side of Theorem 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

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

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 Sect. 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 Theorem 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 $\tau \mathbf{v} \notin \{A^T \mathbf{d} : \mathbf{d} \in \mathbb{S}^m\}$. This property holds true since for any $\mathbf{d} \in \mathbb{S}^m$ we have $\|A^T \mathbf{d}\|_\infty \geq P^\star$ 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 single 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 Servedio (2003), Warmuth et al. (2006). 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 with opposing labels. 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 $\mathbf{d}^T A = \mathbf{0}$.

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 over-weighs examples, which partially explains its poor performance on noisy data. To overcome this problem, it was suggested (see for instance Servedio 2003, Warmuth et al. 2006) 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,

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

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 Theorem 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 (2) can be rewritten as $\min_{\mathbf{d}} f(\mathbf{d}) + g(\mathbf{d}^T A)$. 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 greatest admissible weight to the greatest element of $\boldsymbol{\theta}$, allocate the greatest of the remaining weights to the second greatest element of $\boldsymbol{\theta}$, and so on and so forth. For each $i \in [m]$, let $s_i(\boldsymbol{\theta})$ be the ith greatest 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 Theorem 3 we obtain that the dual problem of (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 Theorem 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 $\text{AvgMin}_k(\boldsymbol{\theta})$ be the average of the k smallest elements of $\boldsymbol{\theta}$. Let ρ be as defined in (2). Then,*

$$\max_{\mathbf{w} \in \mathbb{B}_1^n} \text{AvgMin}_k(A\mathbf{w}) = \rho .$$

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 $|(\mathbf{d}^T A)_j|$ (while $\mathbf{d}$ being the uniform distribution) and $w_j = 0$ for 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 with 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, $\text{AvgMin}_k(A\mathbf{w}) = \frac{-\nu+(k-\nu)\gamma}{k}$. 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 (2) as follows. Denote $C = \mathbb{B}_\infty^m(1/k)$ and recall that $I_C(\mathbf{d})$ is the indicator function of the set C. We can now rewrite (2) as

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

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

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

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 $\mathbb{B}_\infty^m(1/k)$ 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. Assume that the following conditions 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,

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

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 families 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. To apply the theorem, we need to verify that the three conditions hold. Clearly, $\min_d f(\mathbf{d}) = 0$ and the vector $\mathbf{0}$ is in the core of the domain of f (which is simply the ball of radius ν). Furthermore, $f^\star(\mathbf{w}) = \nu\|\mathbf{w}\|_\star$ and therefore the third condition given in the theorem holds as well. Applying the theorem we obtain that:

$$\max_{\mathbf{w} \in \mathbb{B}_1^n, \gamma \in \mathbb{R}} \left(\gamma - \nu\|[\gamma - A\mathbf{w}]_+\|_\star\right) = \min_{d \in \mathbb{S}^m : \|\mathbf{d}\| \leq \nu} \|\mathbf{d}^T A\|_\infty.$$

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 norms 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

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

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 Warmuth et al. 2006, Schölkopf et al. 1998).

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

$$\max_{\mathbf{w} \in \mathbb{B}_1^n, \gamma \in \mathbb{R}} \gamma \quad \text{s.t. } \|[\gamma - A\mathbf{w}]_+\|_\star \leq \nu.$$

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

We now turn to proving Theorem 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 Theorem 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 $\mathbf{0} \in \text{core}(\text{dom}(f))$ by assumption and that $\text{dom}(g) = \mathbb{S}^m$. Therefore, strong duality holds and we can use Theorem 3 which yields,

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

We note that the above equality also follows from properties of the Fenchel conjugate of the infimal convolution operator (see for example Borwein and Lewis 2006). 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 Theorem 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 Theorem 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 Theorem 7 is readily derived.

Proof of Theorem 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, Theorem 3 tells us that the dual of the problem $\min_{\mathbf{d}} \tilde{f}(\mathbf{d}) + \|\mathbf{d}^T A\|_\infty$ is the problem $\max_{\mathbf{w}\in\mathbb{B}_1^n}(-\tilde{f}^\star(-A\mathbf{w}))$. 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 Theorem 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 Theorem 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 (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 Theorem 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

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

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

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

INPUT: matrix $A \in [+1,-1]^{m,n}$

Relaxation function $f^\star$

Desired accuracy ϵ

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 $t = 1, 2, \ldots$

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

$$j_t \in \arg\max_j |(\mathbf{d}_t^T A)_j|$$

(w.l.o.g. assume $\operatorname{sign}(\mathbf{d}_t^T A)_{j_t} = 1$)

IF $\langle \mathbf{d}_t, A\,(\mathbf{e}^{j_t} - \mathbf{w}_t) \rangle \leq \epsilon$

RETURN $\mathbf{w}_t$

ELSE

$$\eta_t = \max\left\{0, \min\left\{1, \frac{\beta\, \mathbf{d}_t^T A(\mathbf{e}^{j_t} - \mathbf{w}_t)}{\|A(\mathbf{e}^{j_t} - \mathbf{w}_t)\|_\infty^2}\right\}\right\}$$

$$\mathbf{w}_{t+1} = (1 - \eta_t)\mathbf{w}_t + \eta_t\, \mathbf{e}^{j_t}$$

Fig. 1 A Boosting Algorithm for maximizing the relaxed margin given in (5)

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 (see also Warmuth et al. 2008). For simplicity we confine the description to weak learners that return the column with the greatest edge. Next, we check whether a stopping condition is met. (We later on prove that by the time the stopping condition is met we have obtained an ϵ-accurate solution.) 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 Sect. 5.3 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 *The algorithm given in Fig. 1 terminates after at most* $32\log(m)/\epsilon^2$ *iterations and returns an* ϵ*-accurate solution, namely,*

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

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

Before turning to the proof of Theorem 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.

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 (2005) and is faster than the rate obtained in Rudin et al. (2007). 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. (2007).)

We can also use the algorithm for maximizing the soft margin given in (6). In Sect. 5.3 we show how to calculate $\mathbf{d}_t$ in $\tilde{O}(m)$ time (where $\tilde{O}(\cdot)$ designates the asymptotic complexity up to logarithmic terms). 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. (2006). 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 Warmuth et al. (2006). See also the discussion in Sect. 5.2.

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 view 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

$$f(\alpha\mathbf{v} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{v}) + (1-\alpha) f(\mathbf{u}) - \frac{\sigma}{2}\alpha(1-\alpha)\|\mathbf{v}-\mathbf{u}\|^2.$$

In the above definition, if $\sigma = 0$ we revert back to the standard definition of convexity. Intuitively, when $S = \mathbb{R}$, σ is a lower bound on the second order derivative of f. 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. (The strong convexity of the relative entropy is a well established result. For a concrete proof see for example Lemma 16 in Shalev-Shwartz 2007.) A 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 Sect. 4 and that solving (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 $0 \leq h(\mathbf{d}) \leq \log(m)$, by setting $\beta = \epsilon/(2\log(m))$ we obtain that

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

Using Lemma 19 in the appendix, the above also implies that

$$\forall \boldsymbol{\theta}, \quad \hat{f}^\star(\boldsymbol{\theta}) \leq \tilde{f}^\star(\boldsymbol{\theta}) \leq \hat{f}^\star(\boldsymbol{\theta}) + \epsilon/2. \tag{7}$$

Thus, 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 Sect. 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 Sect. 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 $-\tilde{f}^\star(-A\mathbf{w})$. So, our next step is to show how Fig. 1 solves the problem: $\max_{\mathbf{w}\in\mathbb{B}_1^n} -\hat{f}^\star(-A\mathbf{w})$.

Let us denote

$$\mathcal{D}(\mathbf{w}) = -\hat{f}^\star(-A\mathbf{w}) \quad \text{and} \quad \mathcal{P}(\mathbf{d}) = \hat{f}(\mathbf{d}) + \|\mathbf{d}^T A\|_\infty. \tag{8}$$

Using again Theorem 3 we obtain that $\mathcal{P}$ and $\mathcal{D}$ are primal and dual objective values, that is,

$$\max_{\mathbf{w}\in\mathbb{B}_1^n} \mathcal{D}(\mathbf{w}) \leq \min_{\mathbf{d}} \mathcal{P}(\mathbf{d}).$$

We also denote by ϵ_t the dual's 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 lemma states that by the time the stopping condition of the algorithm in Fig. 1 is met, the algorithm has obtained an ϵ-accurate solution.

Lemma 11 *For all t we have $\epsilon_t = \max_{\mathbf{w}\in\mathbb{B}_1^n} \mathcal{D}(\mathbf{w}) - \mathcal{D}(\mathbf{w}_t) \leq \langle \mathbf{d}_t, A(\mathbf{e}^{j_t} - \mathbf{w}_t)\rangle$.*

Proof The weak duality property tells us that $\mathcal{P}(\mathbf{d}_t) \geq \max_{\mathbf{w} \in \mathbb{B}_1^n} \mathcal{D}(\mathbf{w})$ and therefore $\epsilon_t \leq \mathcal{P}(\mathbf{d}_t) - \mathcal{D}(\mathbf{w}_t)$. We prove the lemma by showing that $\mathcal{P}(\mathbf{d}_t) - \mathcal{D}(\mathbf{w}_t) = \langle \mathbf{d}_t, A(\mathbf{e}^j - \mathbf{w}_t) \rangle$. First, using Lemma 17, stated in the appendix, we get that for any pair, $\hat{f}$, $\hat{f}^\star$, the following equality holds,

$$\langle \mathbf{d}_t, -A\mathbf{w}_t \rangle = \hat{f}(\mathbf{d}_t) + \hat{f}^\star(-A\mathbf{w}_t).$$

Next, recall that from (8) we have that, $\hat{f}(\mathbf{d}_t) + \hat{f}^\star(-A\mathbf{w}_t) = \mathcal{P}(\mathbf{d}_t) - \|\mathbf{d}_t^T A\|_\infty - \mathcal{D}(\mathbf{w}_t)$. Next, we use the definition of j_t to rewrite $\|\mathbf{d}_t^T A\|_\infty$ as $\langle \mathbf{d}_t, A\mathbf{e}^{j_t} \rangle$. We thus obtain that $\epsilon_t \leq \mathcal{P}(\mathbf{d}_t) - \mathcal{D}(\mathbf{w}_t) = \langle \mathbf{d}_t, -A\mathbf{w}_t \rangle + \|\mathbf{d}_t^T A\|_\infty = \langle \mathbf{d}_t, A(\mathbf{e}^{j_t} - \mathbf{w}_t) \rangle$ which concludes the proof. □

Next, the following lemma lower bounds the improvement made by the boosting algorithm on each iteration in terms of its current sub-optimality. The lemma essentially conveys that if the stopping condition is not met then the update makes substantial progress towards the correct solution.

Lemma 12 *Let ϵ_t be the sub-optimality value of the algorithm in Fig. 1 at iteration t and assume that $\epsilon_t \leq 1$. Then, $\epsilon_t - \epsilon_{t+1} \geq \beta \epsilon_t^2 / 8$.*

Proof Let us denote by Δ_t the difference $\epsilon_t - \epsilon_{t+1}$. From the definition of ϵ_t, it clearly holds 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 $\mathcal{D}(\cdot)$ we further rewrite Δ_t as

$$\Delta_t = \hat{f}^\star(-A\mathbf{w}_t) - \hat{f}^\star(-A\mathbf{w}_t - \eta A(\mathbf{e}^j - \mathbf{w}_t)). \tag{9}$$

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 from the appendix, we know that for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$,

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

where $\nabla = \arg\max_{\mathbf{d}} \langle \boldsymbol{\theta}_1, \mathbf{d} \rangle - \hat{f}(\mathbf{d})$. We now set $\boldsymbol{\theta}_1 = -A\mathbf{w}_t$ and $\boldsymbol{\theta}_2 = -\eta A(\mathbf{e}^j - \mathbf{w}_t)$ and apply the above strong-convexity inequality to (9) while using the definition of $\mathbf{d}_t$ to obtain that,

$$\Delta_t \geq \eta \langle \mathbf{d}_t, A(\mathbf{e}^j - \mathbf{w}_t) \rangle - \frac{\eta^2 \|A(\mathbf{e}^j - \mathbf{w}_t)\|_\infty^2}{2\beta}. \tag{10}$$

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

$$\|A(\mathbf{e}^j - \mathbf{w}_t)\|_\infty \leq \|A\mathbf{e}^j\|_\infty + \|A\mathbf{w}_t\|_\infty \leq \max_{i,j} |A_{i,j}|(1 + \|\mathbf{w}_t\|_1) \leq 2,$$

and thus

$$\Delta_t \geq \eta \langle \mathbf{d}_t, A(\mathbf{e}^j - \mathbf{w}_t) \rangle - 2\eta^2/\beta. \tag{11}$$

Combining Lemma 11 with the above inequality yields,

$$\Delta_t \geq \eta \epsilon_t - 2\eta^2/\beta. \tag{12}$$

Denote $\eta' = \epsilon_t \beta/4$ and note that $\eta' \in [0, 1]$. Had we set $\eta_t = \eta'$ we would have obtained that $\Delta_t \geq \beta \epsilon_t^2/8$ as required. Since we set η_t to be the maximizer of the expression in (10) over $[0, 1]$, we obtain a lower bound on Δ_t which is at least as large as the bound given by (12). This concludes our proof. □

Based on Lemma 12 the proof of Theorem 9 easily follows.

Proof of Theorem 9 First, using Lemma 11, the initialization of $\mathbf{w}_1 = \mathbf{0}$, and the assumption $A \in [-1, +1]^{m \times n}$, we get that

$$\epsilon_1 \leq \langle \mathbf{d}_1, A(\mathbf{e}^{j_1} - \mathbf{w}_1) \rangle = \langle \mathbf{d}_1, A\mathbf{e}^{j_1} \rangle \leq 1.$$

We can now apply Lemma 12 for $t = 1$ and get that $\epsilon_2 \leq \epsilon_1$. By induction, we obtain that Lemma 12 holds for all t. Applying Lemma 20 (provided in the appendix) we get that $\epsilon_t \leq \frac{8}{\beta(t+1)}$.

Plugging the definition of $\beta = \epsilon/(2\log(m))$ into the upper bound on ϵ_t we get $\epsilon_t \leq \frac{16\log(m)}{(t+1)\epsilon}$. Therefore, if $t + 1 \geq 32\log(m)/\epsilon^2$ we get that $\epsilon_t \leq \epsilon/2$. Finally, let ϵ' be the error of $\mathbf{w}_t$ with respect to the original function $\tilde{f}$. Then, using (7) we obtain that $\epsilon' \leq \epsilon_t + \epsilon/2 \leq \epsilon$. □

5.2 Corrective vs. totally corrective updates

The boosting algorithm outlined in Fig. 1 has the advantage that the runtime of each iteration is small. Recently, Warmuth et al. (2006, 2008) proposed an alternative family of algorithms for maximizing the soft margin given in (6). Their algorithms share a common construction which distills to replacing the indicator function of the simplex with the relative entropy function, $h(\mathbf{d})$. The main difference between the algorithmic skeleton presented in this paper and the algorithms in Warmuth et al. (2006, 2008) is that the latter are "totally corrective". Totally corrective algorithms readjust the weights of the features induced thus far so as to minimize the boosting loss function. Once the weights have been optimized a new feature is introduced through a call to the weak-learner.

We first show that the proof of the iteration bound given in Theorem 9 can be seamlessly adapted to the setting of totally corrective updates. More precisely, consider the algorithm from Fig. 1, where the two last lines are replaced with the following totally corrective update rule,

$$\mathbf{w}_{t+1} = \operatorname*{argmax}_{\mathbf{w} \in \mathbb{B}_1^n} \mathcal{D}(\mathbf{w}) \quad \text{s.t. } \forall i \notin \{j_r : r \leq t\},\ w_i = 0. \tag{13}$$

Clearly, the increase of $\mathcal{D}(\mathbf{w})$ at each iteration due to this update is at least as large as the increase attained for $\mathcal{D}(\mathbf{w})$ had we set $\mathbf{w}_{t+1}$ as given in Fig. 1. Therefore, Lemma 12 holds for the totally corrective update as well. The rest of the proof of Theorem 9 remains intact.

Since both the standard (called hence forth corrective) and totally corrective updates share the same iteration bound, and since the runtime of each iteration is much smaller

for the corrective update, it seems that the corrective update is preferable over the totally corrective update. However, experimental evidence presented by Warmuth et al. (2008) indicates that the totally corrective updates may yield improved convergence. We leave further research on the theoretical understanding of totally corrective updates to future work.

5.3 Efficient implementation for soft margins

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 $\{\mathbf{d} : \|\mathbf{d}\|_\infty \leq \nu\}$. 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 Censor and Zenios (1997) for a rigorous definition of the Bregman divergence.

Lemma 13 *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,*

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

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 14 *Assume that the conditions stated in Lemma 13 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 $C = \{\mathbf{d} : \|\mathbf{d}\|_\infty \leq \nu\}$. Based on Corollary 14 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

$$\mathbf{d}_t = \operatorname*{argmin}_{\mathbf{d} \in \mathbb{S}^m : \|\mathbf{d}\|_\infty \leq \nu} B_h(\mathbf{d}\|\mathbf{d}_0). \tag{14}$$

We are therefore left with the problem of solving the entropic projection problem given in (14). A similar problem was tackled by Herbster and Warmuth (2001), who provided $O(m\log(m))$ and $O(m)$ algorithms for performing entropic projections. We would like though to distill the connection of previous work to our setting. Thus, for completeness, in the rest of this section we describe the simpler, sorting-based, $O(m\log(m))$ algorithm. We deviate though from the description given in Herbster and Warmuth (2001) and adapt the analysis given in Duchi et al. (2008) to our setting. Since we mostly recast known results, we give the core derivation of the algorithm and provide proofs in the appendix for central lemmas.

To make the connection to previous work, we first need to show that the entropic projection preserves the relative order of components of the projected vector. Since this lemma was not provided in Herbster and Warmuth (2001) we give its formal statement below while its proof is given in the appendix.

Lemma 15 *Let* $\mathbf{d}_t$ *be the solution of* (14) *and* i, j *be two indices such that* $d_{0,i} > d_{0,j}$, *then*, $d_{t,i} \geq d_{t,j}$.

Assume, without loss of generality, that $\mathbf{d}_0$ is sorted in a non-increasing order. Therefore, using Lemma 15 we know that $\mathbf{d}_t$ takes 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 $\mathbf{d}_t$. This lemma as well is a variation on a similar lemma for Euclidean projections from Duchi et al. (2008) and its proof is also deferred to the appendix.

Lemma 16 *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*

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

We now face the problem of finding the indices i and j. As we now show, all of the elements of optimal vector are non-zero. Formally, the optimal solution of (14) is of the form, $(\nu, \ldots, \nu, d_{t,i}, \ldots, d_{t,m})$ where $d_{t,m} > 0$. To show this property we plug the value of ξ from the previous lemma into the objective function and after simple algebraic manipulations we obtain the following objective value,

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

Therefore, the objective is monotonically increasing in ξ. This in turn implies that we should set ξ to be as small as possible in order to find the minimal Bregman divergence. Last, note that the value of ξ as defined in Lemma 16 is decreasing as a function of j. Therefore, the optimal solution is obtained for $j = m$, meaning $d_{t,m} > 0$.

Finally, we are left with the task of finding the index i. Once it is found we readily obtain ξ, which immediately translates into a closed form solution for $\mathbf{d}_t$. Lemma 15 in conjunction with a property presented in the sequel, implies that the *first* index for which $d_{t,i} < \nu$, where $\mathbf{d}_t$ is as defined by Lemma 16 with $j = m$, constitutes the optimal index for i. The pseudo-code describing the resulting efficient procedure for solving the problem in (14) is given in Fig. 2. The algorithm starts by sorting the vector $\mathbf{d}_0$. Then, it checks each

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 $i = 1, ..., m$
$\xi = \dfrac{1 - \nu\,(i-1)}{Z}$
IF $\xi u_i \leq \nu$
BREAK
ENDIF
$Z \leftarrow Z - u_i$
ENDFOR
OUTPUT: $\mathbf{d}_t$ s.t. $d_{t,r} = \min\{\nu, \xi d_{0,r}\}$

Fig. 2 A sorting-based procedure for the entropic projection problem defined by (14)

possible index i of the sorted vector as the position to stop capping the weights. Concretely, 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 ξ as given by Lemma 16. The same lemma also implies that $d_{t,i} = \xi d_{0,i}$. Thus, if the assumption on the index i is correct, the following inequality must hold, $\nu > d_{t,i} = \xi d_{0,i}$. 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 Lemma 3 in Shalev-Shwartz and Singer (2006a). 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. Moreover, since we simply need to search for the index i which satisfies the conditions of the lemmas above, the time complexity can be reduced to $O(m)$ by replacing sorting with a generalized median search. This improvement was described in Herbster and Warmuth (2001) for entropic projections and in Duchi et al. (2008) for Euclidean projections.

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 $g(\mathbf{d}) = \|\mathbf{d}^T A\|_\infty$, which may lead to a regularization term of the vector $\mathbf{w}$ rather than the domain constraint we currently have.

Appendix 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 Borwein and Lewis (2006, 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 Shalev-Shwartz (2007, Lemma 18).

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*

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

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

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

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

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

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

$$\epsilon_{t+1} \leq \epsilon_t - r\epsilon_t^2 \leq \frac{1}{r(t+1)} - \frac{1}{r(t+1)^2}, \tag{15}$$

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 (15) as

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

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

Proof of Lemma 15 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 $\mathbf{d}_0$. □

Proof of Lemma 16 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 Borwein and Lewis (2006, Corollary 2.1.3) we obtain that there exists a constant ξ' such that for all $i \in I$, $v_i = \xi' - 1$ or equivalently

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

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

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

which immediately yields that ξ attains the value stated in the lemma. □

Appendix B: Fenchel conjugate pairs

We now list a few useful Fenchel-conjugate pairs. Proofs can be found in Boyd and Vandenberghe (2004, Sect. 3.3), Borwein and Lewis (2006, Sect. 3.3), Shalev-Shwartz (2007, Sect. A.3).

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

References

Borwein, J., & Lewis, A. (2006). *Convex analysis and nonlinear optimization*. Berlin: Springer.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Censor, Y., & Zenios, S. A. (1997). *Parallel optimization: theory, algorithms, and applications*. New York: Oxford University Press.

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

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.

Domingo, C., & Watanabe, O. (2000). Madaboost: a modification of adaboost. In *Proceedings of the thirteenth annual conference on computational learning theory*.

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on machine learning*.

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

Freund, Y., & Schapire, R. E. (1996). Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on computational learning theory* (pp. 325–332).

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

Herbster, M., & Warmuth, M. (2001). Tracking the best linear predictor. *Journal of Machine Learning Research*, *1*, 281–309.

Koltchinskii, V., Panchenko, D., & Lozano, F. (2001). Some new bounds on the generalization error of combined classifiers. In *Advances in neural information processing systems, 14*.

Mason, L., Bartlett, P., & Baxter, J. (1998). *Direct optimization of margins improves generalization in combined classifiers* (Technical report). Department of Systems Engineering, Australian National University.

Meir, R., & Rätsch, G. (2003). An introduction to boosting and leveraging. In Mendelson, S., & Smola, A. (Eds.), *Advanced lectures on machine learning* (pp. 119–184). Berlin: Springer.

Ratsch, G., & Warmuth, M. (2005). Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, *6*, 2153–2175.

Rudin, C., Schapire, R. E., & Daubechies, I. (2007). Analysis of boosting algorithms using the smooth margin function. *Annals of Statistics*.

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

Schapire, R. E. (2003). The boosting approach to machine learning: an overview. In Denison, D. D., Hansen, M. H., Holmes, C., Mallick, B., & Yu, B. (Eds.), *Nonlinear estimation and classification*. Berlin: Springer.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1997). Boosting the margin: a new explanation for the effectiveness of voting methods. In *Machine learning: proceedings of the fourteenth international conference* (pp. 322–330).

Schölkopf, B., Smola, A., Williamson, R., & Bartlett, P. (1998). *New support vector algorithms* (Technical Report NC2-TR-1998-053). NeuroColt2.

Servedio, R. A. (2003). Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, *4*, 633–648.

Shalev-Shwartz, S. (2007). *Online learning: theory, algorithms, and applications*. PhD thesis, The Hebrew University.

Shalev-Shwartz, S., & Singer, Y. (2006a). Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, *7*(7), 1567–1599.

Shalev-Shwartz, S., & Singer, Y. (2006b). Convex repeated games and Fenchel duality. In *Advances in neural information processing systems, 20*.

Shalev-Shwartz, S., & Singer, Y. (2007). A primal-dual perspective of online learning algorithms. *Machine Learning Journal*.

Smola, A., Vishwanathan, S. V. N., & Le, Q. (2007). Bundle methods for machine learning. In *Advances in neural information processing systems, 21*.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

von Neumann, J. (1928). Zur Theorie der Gesellschaftsspiele (On the theory of parlor games). *Math. Ann.*, *100*, 295–320.

Warmuth, M., Liao, J., & Ratsch, G. (2006). Totally corrective boosting algorithms that maximize the margin. In *Proceedings of the 23rd international conference on machine learning* (pp. 1001–1008).

Warmuth, M., Glocer, K., & Ratsch, G. (2007). Boosting algorithms for maximizing the soft margin. In *Advances in neural information processing systems, 21*.

Warmuth, M., Glocer, K., & Vishwanathan, S. V. N. (2008). Entropy regularized lpboost. In *Algorithmic learning theory (ALT)*.

Zhang, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Transaction on Information Theory*, *49*, 682–691.