Information Processing Letters 112 (2012) 481–486

Contents lists available at SciVerse ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

# Sparse weighted voting classifier selection and its linear programming relaxations

Noam Goldberg [a,*], Jonathan Eckstein [b]

[a] Mathematics and Computer Science Division, Argonne National Laboratory, United States
[b] MSIS Department and RUTCOR, Rutgers University, United States

ARTICLE INFO

*Article history:*
Received 14 January 2011
Received in revised form 8 March 2012
Accepted 8 March 2012
Available online 9 March 2012
Communicated by W.-L. Hsu

*Keywords:*
Machine learning
Computational complexity
Weighted voting classification
Sparsity
Integrality gap
Hardness of approximation

ABSTRACT

We consider the problem of minimizing the number of misclassifications of a weighted voting classifier, plus a penalty proportional to the number of nonzero weights. We first prove that its optimum is at least as hard to approximate as the minimum disagreement halfspace problem for a wide range of penalty parameter values. After formulating the problem as a mixed integer program (MIP), we show that common "soft margin" linear programming (LP) formulations for constructing weighted voting classsifiers are equivalent to an LP relaxation of our formulation. We show that this relaxation is very weak, with a potentially exponential integrality gap. However, we also show that augmenting the relaxation with certain valid inequalities tightens it considerably, yielding a linear upper bound on the gap for all values of the penalty parameter that exceed a reasonable threshold. Unlike earlier techniques proposed for similar problems (Bradley and Mangasarian (1998) [4], Weston et al. (2003) [14]), our approach provides bounds on the optimal solution value.



## 1. Introduction

This paper examines the relationship between continuous and discrete formulations of weighted voting classification problems. Consider a binary classification problem with $m$ training samples, each consisting of $\ell$ real-valued attributes, represented as a matrix $A \in \mathbb{R}^{m\times\ell}$ whose rows correspond to observations and whose columns correspond to attributes. We are also given a vector of labels $y \in \{-1,1\}^m$, defining a partition of the observations $M = \{1,\dots,m\}$ into a "positive" class $M^+ = \{i \in M \mid y_i = 1\}$ and a "negative" class $M^- = M \setminus M^+$. Using a potentially large set of base classifiers $h_j : \mathbb{R}^u \to \{-1,0,1\}$ indexed by the set $U = \{1,\dots,u\}$, we would like to train a weighted voting classifier $g(x) = \sum_{j\in U} \lambda_j h_j(x)$, for $\lambda \in \mathbb{R}^u_+$. A new test observation $x \in \mathbb{R}^N$ is classified as either positive or negative based on $\operatorname{sgn}(g(x))$. Note that $\lambda$ is constrained to be nonnegative; this restriction is common in weighted voting classification methods and also simplifies some of the problem formulations below. By including additional base classifiers of the form $-h_j$ in $U$ as necessary, there is no loss of generality from requiring $\lambda \geqslant 0$.

Optimization models for training such classifiers typically have two components in their objectives, one related to penalizing misclassified data points and another related to penalizing classifier complexity, or equivalently maximizing a margin of separation; see [6, Theorem 3.1]. To obtain tractable convex optimization problems, however, commonly used formulations only use continuous approximations, such as the $L_1$ norm of $\lambda$ as a surrogate for the number of nonzero elements in $\lambda$. In this paper, we consider the natural combinatorial formulation which more directly penalizes misclassification error and classifier complexity; its motivation may be traced back to error (generalization) bounds for boosting algorithms [8, Theorems 7–8]. Further, there has been significant renewed interest in solving this problem, and closely related variants, either heuristically or approximately [14,9].

* Corresponding author.
*E-mail addresses:* noamgold@mcs.anl.gov (N. Goldberg), jeckstei@rci.rutgers.edu (J. Eckstein).

0020-0190/$ – see front matter 
doi:10.1016/j.ipl.2012.03.004

Hence, we consider the *sparse weighted voting classifier* (SWVC) problem

$$\min_{\lambda\in\mathbb{R}^u_+} \sum_{i=1}^{m} \mathbf{I}(y_i H_i \lambda < 1) + C\|\lambda\|_0, \tag{1}$$

where $C \geqslant 0$ is a parameter, $\mathbf{I}(\cdot)$ is the binary indicator function, $\|\cdot\|_0$ denotes the "$L_0$ norm" which counts the number of nonzeroes in its argument, and $H_i$ is the $i$th row of $H$, an $m \times u$ matrix whose elements are $H_{ij}$, the label assigned to observation $i$ by classifier $j$.

Unfortunately, special cases of this problem are known to be $\mathcal{NP}$-hard; we discuss and extend these results in Section 3. For similar problems, Weston et al. [14], extending earlier work of Bradley and Mangasarian [4], propose minimizing a smooth, nonconvex approximation of the step function in order to heuristically approximate an $L_0$-norm penalty for $\lambda \in \mathbb{R}^u$. Unlike such techniques, our approach is based on a mixed integer programming (MIP) formulation, and provides bounds on the optimal value. We will relate it to continuous "soft margin" linear programming (LP) formulations of classification problems, such as [10]

$$\min\left\{ \sum_{j=1}^{u} \lambda_j + D\sum_{i=1}^{m} \xi_i \;\middle|\; \operatorname{diag}(y)H\lambda + \xi \geqslant \mathbb{1} \text{ and } \lambda, \xi \geqslant 0 \right\}. \tag{2}$$

Here, the *margin* of observation $i$ is $y_i H_i \lambda$, and margins smaller than 1 incur a penalty proportional to the positive parameter $D$.

## 2. MIP formulation

We now reformulate (1) as a MIP, using the binary variable $\mu_j$ to indicate that feature $j$ is used, and the binary variable $\xi_i$ to indicate that observation $i$ is misclassified. Letting $K$ be a suitably large constant, $\mathbb{1}$ denote a vector of ones, and $\operatorname{diag}(x)$ denote a diagonal matrix whose $i$th diagonal entry is $x_i$, the formulation is

$$\min_{\xi,\mu,\lambda} \sum_{i\in M} \xi_i + C\sum_{j\in U} \mu_j \tag{3a}$$

$$\text{s.t.} \quad \operatorname{diag}(y)H\lambda + (mK+1)\xi \geqslant \mathbb{1} \tag{3b}$$

$$\lambda \leqslant K\mu \tag{3c}$$

$$\xi \in \{0,1\}^m, \quad \mu \in \{0,1\}^u, \quad \lambda \geqslant 0. \tag{3d}$$

We now show that (3) is equivalent to (1) for large enough $K$. The magnitude of $K$, however, determines the (poor) quality of the MIP LP relaxation, which we examine in Section 4. Note that the objective values of (1) and (3) are both bounded below by 0, and that (3) always has a feasible solution (for example, $\lambda = 0$, $\xi = \mathbb{1}$). Therefore, the equivalence of SWVC and (3) for sufficiently large $K$ is established by the following:

**Proposition 2.1.** *If $K \geqslant m^{m/2}$, then every optimal solution $(\xi^*, \mu^*, \lambda^*)$ of (3) satisfies*

$$\sum_{i=1}^{m} \xi_i^* + C\sum_{j\in U} \mu_j^* = \min_{\lambda\in\mathbb{R}^u_+} \sum_{i=1}^{m} \mathbf{I}(y_i H_i \lambda < 1) + C\|\lambda\|_0 = \sum_{i=1}^{m} \mathbf{I}\big(y_i H_i \lambda^* < 1\big) + C\big\|\lambda^*\big\|_0.$$

**Proof.** By (3c), $\|\lambda^*\|_0 \leqslant \sum_{j\in U} \mu_j^*$. By (3b),

$$\sum_{i=1}^{m} \mathbf{I}\big(y_i H_i \lambda^* < 1\big) \leqslant \sum_{i=1}^{m} \xi_i^*.$$

If $\hat{\lambda}$ is optimal for SWVC, then

$$\sum_{i=1}^{m} \mathbf{I}(y_i H_i \hat{\lambda} < 1) + C\|\hat{\lambda}\|_0 \leqslant \sum_{i=1}^{m} \mathbf{I}\big(y_i H_i \lambda^* < 1\big) + C\big\|\lambda^*\big\|_0 \leqslant \sum_{i=1}^{m} \xi_i^* + C\sum_{j\in U} \mu_j^*. \tag{4}$$

We now prove the reverse inequality between the first and last quantities. Consider the linear system in $\lambda$ given by

$$\sum_{j\in U:\hat{\lambda}_j\neq 0} y_i H_{ij} \lambda_j \geqslant 1 \quad \forall i \in M\colon\ y_i H_i \hat{\lambda} \geqslant 1. \tag{5}$$

Let $B$ denote a basis matrix of this system, comprising a submatrix of $\operatorname{diag}(y)H$ with column indices $j \in U$ such that $\hat{\lambda}_j \neq 0$, and a subset of columns of $-I$ (corresponding to the slack variables when converting the inequalities to equalities). By a standard LP basis argument, there exists a $\bar{\lambda}$ solving (5) such that $\|\bar{\lambda}\|_0 \leqslant \|\hat{\lambda}\|_0 \leqslant m$ (with $\|\bar{\lambda}\|_0 = \|\hat{\lambda}\|_0$ when $C > 0$ by the optimality of $\hat{\lambda}$). Let $B^{(j)}$ denote the matrix $B$ with the column corresponding to feature $j$ replaced by $\mathbb{1}$; by Cramer's rule, $\bar{\lambda}_j = \det(B^{(j)})/\det(B)$ for all $j$ for which $\bar{\lambda}_j > 0$. Since the rank of $B$ is at most $m$, Hadamard's bound [5, for example] yields $|\det(B^{(j)})| \leqslant m^{m/2}$. Since $B$ is a basis, $\det(B) \neq 0$, and as $B$ is an integer matrix, we have $|\det(B)| \geqslant 1$. Hence, there exists $\lambda' \in \mathbb{R}^u$ with $\lambda'_j = \bar{\lambda}_j$ if $\hat{\lambda}_j \neq 0$ and $\lambda'_j = 0$ otherwise, such that $\|\lambda'\|_0 = \|\hat{\lambda}\|_0 = \|\bar{\lambda}\|$ and $\lambda'_j \leqslant m^{m/2} \leqslant K$ for all $j \in U$. Further, $|y_i H_i \lambda'| \leqslant m^{m/2+1}$ for all $i \in M$. Let

$$\xi'_i = \begin{cases} 1 & \text{if } y_i H_i \lambda' < 1, \\ 0 & \text{otherwise,} \end{cases} \qquad \mu'_j = \begin{cases} 1 & \text{if } \lambda'_j > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, for all $i \in M$, $y_i H_i \lambda' + (m^{m/2+1}+1)\xi'_i \geqslant 1$. Thus, $(\xi', \mu', \lambda')$ is feasible for (3), and

$$\sum_{i=1}^{m} \xi'_i = \sum_{i=1}^{m} \mathbf{I}(y_i H_i \hat{\lambda} < 1).$$

Therefore, by the optimality of $(\xi^*, \mu^*, \lambda^*)$ for (3),

$$\sum_{i=1}^{m} \xi_i^* + C\sum_{j\in U} \mu_j^* \leqslant \sum_{i=1}^{m} \xi'_i + C\sum_{j\in U} \mu'_j \leqslant \sum_{i=1}^{m} \mathbf{I}(y_i H_i \hat{\lambda} < 1) + C\|\hat{\lambda}\|_0.$$

Thus, all the relations in (4) hold with equality. □

The basis arguments of the proof resemble those of Muroga et al. [13], in particular Lemma 1 and Theorem 16; see also [1]. The results of [13], well known in the field of Boolean functions and logic gates, state that any weighted majority gate with $n$ inputs can be realized with integer weights bounded by $2^{-n}(n+1)^{(n+1)/2}$. This analysis, however, relies on the assumption that each column of $H$ has only $\{0,1\}$ or $\{0,-1\}$ entries, whereas we allow mixed $\{-1,0,1\}$ entries in each column. In the case $H$ conforms to the assumptions of [13], we note that its results can be used to tighten the $m^{m/2}$ bound above to $2^{-m+2}m^{m/2}$; however, this bound remains exponential in $m$.

## 3. Computational complexity and inapproximability

When $C=0$, minimizing the objective of (1) over $\lambda\in\mathbb{R}^u$ is known as the *minimum disagreement halfspace* problem (MDH), and is $\mathcal{NP}$-hard [12,3]. SWVC generalizes MDH, so it is at least as hard to solve computationally. Specifically, any MDH instance $(H',y')$ can be reduced to (3) with $H=(H' \quad -H')$, $y=y'$ and $C=0$. We will also refer to any solution $(\xi,\mu,\lambda)$ of (3), after applying this reduction to an MDH instance $(H,y)$, as an *MDH solution*. Arora et al. [3] showed that MDH is inapproximable to within any factor better than $2^{\log^{1-\epsilon}m}$, for $\epsilon>0$, assuming $\mathcal{NP}\not\subseteq \mathrm{DTIME}(m^{\mathrm{poly}(\log m)})$, by reduction of the label cover problem; see also [2] ($\mathrm{DTIME}(n)$ is the class of problems that can be solved in deterministic time $n$). Dinur and Safra [7] strengthened the inapproximability of label covering to $2^{\log^{(1-o(1))}m}$, assuming $\mathcal{P}\neq\mathcal{NP}$. This strengthened inapproximability result also applies to MDH, through the same reduction as [3], and thus to SWVC with $C=0$.

If $0<C<1/m$ and the input data are linearly separable, then SWVC is equivalent to a special case of another problem that is $\mathcal{NP}$-hard to approximate, minimizing the number of relevant variables in a linear system [2]. We omit this proof for brevity, and instead establish a more general inapproximability result for SWVC for $C=O(m^{\delta})$, where $0\leqslant\delta<1$, making use of the $2^{\log^{1-\epsilon}m}$-factor inapproximability for MDH [3,2] and the work of Dinur and Safra [7]. Note that SWVC has the trivial solution $\lambda=0$, $\xi=\mathbb{1}$ whenever $C\geqslant m$, so only smaller values of $C$ are of interest. For any binary vector of length at least $m$, define

$$S^+(\xi)=\left\{H_i \;\middle|\; i\in M^+,\xi_i=0\right\} \quad\text{and}$$
$$S^-(\xi)=\left\{H_i \;\middle|\; i\in M^-,\xi_i=0\right\}.$$

We will need the following lemmas:

**Lemma 3.1.** *Given an MDH instance $(H,y)$ with $H\in\{-1,0,1\}^{m\times u}$ and $y\in\{-1,1\}^m$, along with any $C\geqslant 0$ and some integer $k>Cm$, there exists a reduction, polynomial in $k$ and $u$, to an SWVC instance $H'\in\{-1,0,1\}^{mk\times 2u}$ and $y'\in\{-1,1\}^{mk}$, such that $(\hat\xi,\hat\mu,\hat\lambda)$ is an optimal MDH solution for $(H,y)$ if and only if the SWVC instance $(H',y',C)$ has an optimal solution $(\xi^*,\mu^*,\lambda^*)$, where $\sum_{i=1}^{mk}\xi_i^*=k(\sum_{i=1}^{m}\hat\xi_i)$.*

**Proof.** Construct an SWVC instance by having $y'$ consist of $k$ concatenated copies of $y$, and $H'$ consist of corresponding duplicate blocks of the form $[\,H \quad -H\,]$. Let $(\xi^*,\mu^*,\lambda^*)$ be an optimal SWVC solution for the input $(H',y',C)$. Let $(\hat\xi,\hat\mu,\hat\lambda)$ be an optimal MDH solution; its objective value is $z_{\mathrm{MDH}}=\sum_{i=1}^{m}\hat\xi_i+0\sum_{j=1}^{2u}\hat\mu_j=\sum_{i=1}^{m}\hat\xi_i$. Now, since feasible solutions of MDH and SWVC always exist and are bounded below by 0, we have only to prove that $z_{\mathrm{MDH}}=\sum_{i=1}^{m}\hat\xi_i=(1/k)\sum_{i=1}^{mk}\xi_i^*$. First, note that $z_{\mathrm{MDH}}\leqslant(1/k)\sum_{i=1}^{mk}\xi_i^*$, since otherwise there exists an MDH solution with objective below $z_{\mathrm{MDH}}$. To complete the proof, it remains only to rule out the possibility that $z_{\mathrm{MDH}}<(1/k)\sum_{i=1}^{mk}\xi_i^*$, which we now do by contradiction. Suppose that $z_{\mathrm{MDH}}<(1/k)\sum_{i=1}^{mk}\xi_i^*$. Then, the integrality of $\xi^*$ and $z_{\mathrm{MDH}}$ yields $z_{\mathrm{MDH}}\leqslant\frac{1}{k}\sum_{i=1}^{mk}\xi_i^*-1$. Since $(\hat\xi,\hat\mu,\hat\lambda)$ is a solution of (3), $S^+(\hat\xi)$ and $S^-(\hat\xi)$ are linearly separable. Therefore, since $|S^+(\hat\xi)|+|S^-(\hat\xi)|\leqslant|M^+|+|M^-|=m$, a standard LP basis argument (see also [13, Lemma 1]) implies that $S^+(\hat\xi)$ and $S^-(\hat\xi)$ are separable by some hyperplane whose weight vector $\lambda\geqslant 0$ satisfies $\|\lambda\|_0\leqslant m$. Then, denoting the SWVC optimal value by $z_{\mathrm{SWVC}}$,

$$\begin{aligned}
z_{\mathrm{SWVC}}&=\sum_{i=1}^{mk}\xi_i^*+C\sum_{j=1}^{2u}\mu_j^*\\
&\leqslant k\sum_{i=1}^{m}\hat\xi_i+C\sum_{j=1}^{2u}\mathbf{I}\,(\lambda_j\neq 0)\\
&\leqslant kz_{\mathrm{MDH}}+Cm<k\,(z_{\mathrm{MDH}}+1)\\
&\leqslant\sum_{i=1}^{mk}\xi_i^*\leqslant z_{\mathrm{SWVC}}.
\end{aligned}$$

The first two inequalities on the last line follow respectively from $k>Cm$ and $z_{\mathrm{MDH}}\leqslant(1/k)\sum_{i=1}^{mk}\xi_i^*-1$. Since the result is a contradiction, we must have $z_{\mathrm{MDH}}=(1/k)\sum_{i=1}^{mk}\xi_i^*=\sum_{i=1}^{m}\hat\xi_i$. □

**Lemma 3.2.** *A polynomial-time $f(m)$-approximation factor for SWVC with penalty $C=C(m)\in O(m^{\delta})$, for some $0\leqslant\delta<1$ and $f:\mathbb{N}_+\to\mathbb{R}_+$, implies a polynomial-time $\alpha f(\beta m^{1+(1+\delta)/(1-\delta)})$-approximation factor for MDH for some $\alpha,\beta\in O(1)$.*

**Proof.** Given any MDH instance $(H,y)$, take $(H',y',C)$ to be the corresponding instance of SWVC, using the reduction of Lemma 3.1, for some integer $k>Cm$. Let $m'=km$ denote the number of observations in this SWVC instance. Let $(\xi^*,\mu^*,\lambda^*)$ be an optimal SWVC solution for $(H',y',C)$. By a standard LP basis argument (see also [13, Lemma 1]) it must be possible to separate linearly separable sets of observations $S^+(\xi^*)$ and $S^-(\xi^*)$ using at most $m\geqslant|S^+(\xi^*)|+|S^-(\xi^*)|$ features, and this same separator also separates the duplicate observations with indices $\{m+1,\ldots,km\}$. Therefore, we may take $(\xi^*,\mu^*,\lambda^*)$ to be an optimal SWVC solution such that $\sum_{j=1}^{2u}\mu_j^*\leqslant m$ (which holds for all optimal solutions whenever $C>0$).

Let $(\hat\xi,\hat\mu,\hat\lambda)$ be an optimal MDH solution, with $z_{\mathrm{MDH}}$ denoting its objective value, and let $(\xi,\mu,\lambda)$ denote

the hypothesized approximate SWVC solution. Now, by Lemma 3.1, $z_{\mathrm{MDH}} = \sum_{i=1}^{m} \hat{\xi}_i = \frac{1}{k} \sum_{i=1}^{km} \xi_i^*$; thus,

$$\begin{aligned} z_{\mathrm{MDH}} &\leqslant \frac{1}{k}\left(\sum_{i=1}^{km} \xi_i^* + C\sum_{j=1}^{2u} \mu_j^*\right) \leqslant \frac{1}{k}\left(\sum_{i=1}^{km} \xi_i + C\sum_{j=1}^{2u} \mu_j\right) \\ &\leqslant \frac{f(km)}{k}\left(\sum_{i=1}^{km} \xi_i^* + C\sum_{j=1}^{2u} \mu_j^*\right) \\ &\leqslant \frac{f(km)}{k}\left(k\sum_{i=1}^{m} \xi_i^* + Cm\right) \leqslant f(km)\left(\sum_{i=1}^{m} \xi_i^* + 1\right). \end{aligned}$$

The last two inequalities use respectively that $\sum_{j=1}^{2u} \mu_j^* \leqslant m$, and $k > Cm$.

Now, by the hypothesis, $C \leqslant \gamma m^{\delta}$, for some constant $\gamma$, and so $C \leqslant \gamma (km)^{\delta}$, since $k \geqslant 1$ and $\delta \geqslant 0$. For the analysis above to hold, we require $k > Cm$; since $C \leqslant \gamma (km)^{\delta}$, this condition will hold if we have $k > (\gamma (km)^{\delta})m$. Solving for $k$, this condition is equivalent to $k > \gamma^{1-\delta} m^{(1+\delta)/(1-\delta)}$. Specifically, let us choose $k = \lceil \gamma^{1-\delta} m^{(1+\delta)/(1-\delta)} + \epsilon \rceil$ where $\epsilon > 0$ is a small constant. Note that there exists some constant $\beta > 0$ such that $k \leqslant \beta m^{(1+\delta)/(1-\delta)}$ for all $m$. Now select some (constant) integer $\tau > 0$, and let $\alpha = 1 + 1/\tau$. If $z_{\mathrm{MDH}} \geqslant \tau$, then $z_{\mathrm{MDH}} \leqslant f(km)(1 + 1/\tau) \sum_{i=1}^{m} \xi_i^*$, and by running the hypothesized approximation algorithm we obtain an approximate

$$\begin{aligned} f(km)(1 + 1/\tau) &\leqslant f\left(\beta m^{1+(1+\delta)/(1-\delta)}\right)(1 + 1/\tau) \\ &= \alpha f\left(\beta m^{1+(1+\delta)/(1-\delta)}\right) \end{aligned}$$

factor solution. On the other hand, if $z_{\mathrm{MDH}} < \tau$, then we may obtain an exact solution in polynomial time by excluding each possible subset of observations of size less than $\tau$, and applying a polynomial time LP algorithm to attempt to construct a separating hyperplane for the remaining observations. Since the choices of $C$ and $k$ are polynomially bounded in $m$, and the reduction of Lemma 3.1 is polynomial in $k$ and $u$, the entire procedure is polynomial-time. □

**Proposition 3.3.** *For any penalty $C \in O(m^{\delta})$ with $0 \leqslant \delta < 1$, and $\epsilon > 0$, the SWVC problem cannot be approximated in polynomial time within a factor of $2^{\log^{1-\epsilon} m}$ unless $\mathcal{P} = \mathcal{NP}$.*

**Proof.** By Lemma 3.2, a polynomial-time $2^{\log^{1-\epsilon} m}$-factor approximation for SWVC, for some $\epsilon > 0$, yields a polynomial-time approximation for the MDH problem with factor $\alpha 2^{\log^{1-\epsilon}(\beta m^{1+(1+\delta)/(1-\delta)})}$, for some $\alpha, \beta \in O(1)$. Now,

$$\begin{aligned} \alpha 2^{\log^{1-\epsilon}(\beta m^{1+(1+\delta)/(1-\delta)})} &\leqslant \alpha 2^{[1+(1+\delta)/(1-\delta)]^{(1-\epsilon)} \log^{1-\epsilon} m} \\ &\leqslant 2^{\log^{1-\epsilon'} m} \end{aligned}$$

for some $0 < \epsilon' \leqslant \epsilon$, and all $m \geqslant m_0$, for some $m_0 \geqslant 1$. Unless $\mathcal{P} = \mathcal{NP}$, such an approximation factor contradicts the inapproximability of MDH following from the reduction of [3,2] and the strengthened inapproximability result for label covering in [7]. □

## 4. The soft margin relaxation and its integrality gap

Consider the following relaxation of (3):

$$\min_{\xi,\mu,\lambda \geqslant 0} \left\{ \sum_{i=1}^{m} \xi_i + C\sum_{j \in U} \mu_j \;\middle|\; \begin{array}{l} \mathrm{diag}(y)H\lambda + (mK+1)\xi \geqslant \mathbb{1} \\ \lambda \leqslant K\mu \end{array} \right\}. \tag{6}$$

Because it omits the constraints $\xi \leqslant \mathbb{1}$, $\mu \leqslant \mathbb{1}$, this linear program is a weaker relaxation of (3) than the customary linear programming relaxation; we call it the *soft margin relaxation*. We now justify this terminology by showing that (6) is equivalent to the standard soft margin classifier LP (2) when the penalties $C$ and $D$ are of the appropriate ratio.

**Proposition 4.1.** *For every instance $(H, y)$, $(\xi, \lambda)$ is an optimal solution of (2) if and only if $(\hat{\xi}, \hat{\mu}, \lambda)$ is an optimal solution of (6) with $C = 1/D(m + 1/K))$, where $\hat{\xi} = \xi/(mK + 1)$ and $\hat{\mu} = \lambda/K$.*

**Proof.** Consider the map

$$\omega : (\lambda, \xi, D) \mapsto \left(\lambda, \frac{1}{mK+1}\xi, \frac{1}{K}\lambda, \frac{1}{D(m+1/K)}\right).$$

Take any $D > 0$ and $(\xi, \lambda)$ that is a feasible solution of (2), and let $(\lambda, \hat{\xi}, \hat{\mu}, C) = \omega(\lambda, \xi, D)$. Now, $\mathrm{diag}(y)H\lambda + \xi = \mathrm{diag}(y)H\lambda + (mK+1)\hat{\xi} \geqslant \mathbb{1}$ and $\hat{\mu} = \lambda/K$ imply that $(\lambda, \hat{\xi}, \hat{\mu})$ is feasible for (6), with objective value

$$\begin{aligned} &\sum_{i=1}^{m} \hat{\xi}_i + C\sum_{j=1}^{u} \hat{\mu}_j \\ &\quad = \frac{1}{mK+1}\sum_{i=1}^{m} \xi_i + \frac{1}{D(m+1/K)}\sum_{j=1}^{u} \lambda_j/K \\ &\quad = \frac{1}{D(mK+1)}\left(D\sum_{i=1}^{m} \xi_i + \sum_{j=1}^{u} \lambda_j\right). \end{aligned}$$

Thus, $\omega$ maps feasible solutions of (2) to feasible solutions of (6), scaling the objective by $C = 1/(D(mK+1))$. Conversely, if one takes any solution $(\lambda, \hat{\xi}, \hat{\mu})$ to (6) which has $\hat{\mu} = \lambda/K$, then the inverse image of $(\lambda, \hat{\xi}, \hat{\mu}, C)$ under $\omega$ is a singleton $\{(\lambda, \xi, D)\}$ such that $(\lambda, \xi)$ is feasible for (2), with objective value scaled by $D(mK+1)$. The conclusion then follows by noting that all optimal solutions of (6) must have $\mu = \lambda/k$, since the nonnegative variables $\mu_j$ have positive objective coefficients, each appears only in the constraint $\mu_j \geqslant \lambda_j/K$, and the objective is being minimized. □

The strength of a relaxation is typically characterized by its *gap*, the ratio between its optimal objective value and that of the original problem. In the case of the LP relaxation of a MIP, this ratio is called the *integrality gap*. For a given input $(H, y)$, define $z(H, y)$ to be the optimal value of (3), and $z_{\mathrm{R}}(H, y)$ to be the optimal value of (6). We now show that (6) is an extremely weak relaxation of (3):

**Proposition 4.2.** $\sup_{H,y}\{z(H,y)/z_R R(H,y)\} \geqslant mK+1$.

**Proof.** Consider the simple SWVC instance given by $C=1$ and $\operatorname{diag}(y)H = I$ (the identity matrix), meaning that each base classifier covers only a single observation. Since each observation $i \in M$ must be either classified correctly by the single classifier $u$ with $y_i H_{ij} = 1$ and $\mu_j = 1$, or otherwise $\xi_i = 1$, this instance has an optimal integer solution of value $m$, where $m$ of the $\mu_j$ and $\xi_i$ variables assume a value of one and all of the remaining variables are zero. The relaxation, however, has the feasible solution $\xi_i = 1/(mK+1)$ for $i \in M$, and $\mu = 0$, with objective value $m/(mK+1)$. Thus, for this instance, we have $z(H,y)/z_R(H,y) \geqslant m/(m/(mK+1)) = mK+1$. □

To relate this result to the SWVC problem (1), one must consider the magnitude of $K$. The lower bound on $K$ from Proposition 2.1 is sufficient for (3) to be equivalent to SWVC, but may be much larger than necessary. We now establish some necessary lower bounds on $K$.

**Proposition 4.3.** *If order for* (3) *to be equivalent to the SWVC problem* (1) *for all choices of* $(H,y,C)$ *with* $C < q$, *for some integer* $q \in [0,m]$, *it is necessary to have* $K \geqslant 2^{\lceil m/q\rceil - 1}$.

**Proof.** Construct an SWVC instance $(H,y,C)$ with $m$ observations and $u = \lceil m/q\rceil$ features so that column $j$ of $\operatorname{diag}(y)H$ contains

- 0 in rows $1,\ldots,(j-1)q$,
- $+1$ in rows $(j-1)q+1,\ldots,\min\{jq,m\}$,
- $-1$ in all subsequent rows (if any).

By using all the features, it is possible to correctly classify every observation, obtaining an objective value of $C\cdot u$. This is the only optimal choice, since any solution that uses $r < u$ features must misclassify at least $(u-r)q$ observations, and hence its objective value is at least

$$(u-r)q + C\cdot r > C\cdot u$$

(because $q > C$). Therefore, if $(\xi^*,\mu^*,\lambda^*)$ is optimal for (3), we must have $\xi^* = 0$ and $\mu^* = \mathbb{1}$. It then follows from (3b) that $\lambda_1^* = 1, \lambda_2^* = 2, \lambda_3^* = 4, \ldots, \lambda_{\lceil m/q\rceil}^* = 2^{\lceil m/q\rceil - 1}$. Thus, formulation (3) must have $K \geqslant 2^{\lceil m/q\rceil-1}$, or it prohibits the optimal SWVC solution. □

As a simple example of the construction in the proof of Proposition 4.3, setting $q=1$ yields an instance $(H,y,C)$ such that

$$\operatorname{diag}(y)H = \begin{pmatrix} +1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & +1 & 0 & \ldots & 0 & 0 \\ -1 & -1 & +1 & \ldots & 0 & 0 \\ & & & \ddots & & \\ -1 & -1 & -1 & \ldots & +1 & 0 \\ -1 & -1 & -1 & \ldots & -1 & +1 \end{pmatrix},$$

and $C < 1$, implying that we need $K \geqslant 2^{m-1}$ for the formulation to be correct. As a simple corollary of Propositions 4.2 and 4.3, we conclude that the soft margin relaxation gap is in general at least exponential in $m/\lceil C\rceil$. Finally, for instances with very small $C$, specifically $C < 1/m$, one can use the results of [11,1] to demonstrate an even larger gap. We omit these results for brevity.

## 5. Tightening the relaxation

We now consider adding valid inequalities to (3) in order to strengthen its relaxation. We say that a base classifier $h$ *distinguishes* between a pair $(i,i')$ if it classifies them differently but classifies at least one of them correctly, *e.g.*, $h_j(A_i) = y_i \neq h_j(A_{i'})$. Let $S_{i,i'} = \{j \in U \mid h_j(A_i) = y_i \neq h_j(A_{i'})\}$ denote the set of base classifiers that correctly classify observation $i$ and distinguish it from $i'$. Consider the following inequality for each pair of observations $(i,i') \in \Phi = (M^+ \times M^-) \cup (M^- \times M^+)$:

$$\xi_i + \xi_{i'} + \sum_{j \in S_{i,i'}} \mu_j \geqslant 1. \tag{7}$$

The interpretation of this inequality is that either we misclassify at least one of the of the observations $i$ or $i'$, or we need to distinguish between the two using at least one of the distinguishing features in $S_{i,i'}$.

**Proposition 5.1.** *The inequalities* (7) *are valid, that is, they hold for all integer-feasible solutions of* (3).

**Proof.** Take any $(i,i') \in M^+ \times M^-$. If $\xi_i + \xi_{i'} \geqslant 1$, then (7) clearly holds. Otherwise, $i \in M^+$ and $\xi_i = \xi_{i'} = 0$ imply that $\sum_{j\in U} H_{ij}\lambda_j \geqslant 1$. Thus, $h_j(A_i)\lambda_j > 0$ for some $j \in U$; $\lambda_j \geqslant 0$ and $h_j(A_i) = y_i = 1$ imply $0 < \lambda_j/K \leqslant \mu_j = 1$. The proof for $(i,i') \in M^- \times M^+$ is similar. □

We now consider the *tightened relaxation* consisting of the linear program (6), augmented by all possible cutting planes of the form (7). We denote the optimal objective value of the tightened relaxation by $z_{TR}(H,y)$.

In typical learning applications of SWVC, each feature added to the model should explain at least a single additional observation, implying that $C \geqslant 1$. In this case, it is straightforward to prove a bound on the gap of the tightened relaxation:

**Proposition 5.2.** *If* $C \geqslant 1$, *then* $z(H,y)/z_{TR}(H,y) \leqslant m$.

**Proof.** Consider an optimal solution $(\xi,\mu,\lambda)$ of the tightened relaxation. Now, if $C \geqslant 1$, the cuts (7) assure that $z_{TR}(H,y) = \sum_{i=1}^m \xi_i + C\sum_{j\in U}\mu_j \geqslant \xi_i + \xi_i' + \sum_{j\in S_{ii'}}\mu_j \geqslant 1$, for some $(i,i')$. On the other hand, $\xi = \mathbb{1}$, $\mu = 0$, $\lambda = 0$ is feasible for (3) and attains the objective value $m$, so $z(H,y) \leqslant m$. □

Thus, the cuts (7) provably tighten the soft margin relaxation (6), equivalent to the standard soft-margin linear program (2), reducing its gap from exponential to a small polynomial in the input.

We now make some remarks about the practical application of the tightened relaxation: first, note that since the

number of cuts (7) is at most $2|M^+||M^-|$, the polynomial-time solvability of LP implies that the solution of the tightened relaxation may be found in time polynomial in the input. Although the total number of cutting planes is in $O(m^2)$, significant practical improvements are possible by using only a carefully selected subset, for example by iteratively identifying violated inequalities (7), adding them to the formulation, and reoptimizing. Finally, in [9], we suggested a variant of the tightened relaxation formulation, preferable for numerical reasons, which dispenses with the large constant $K$ and constrains $\|\lambda\|_1 = 1$, while fixing the margin equal to a small parameter. We used this formulation to evaluate the practical effectiveness of inequalities of the form (7) within a cut and column generation boosting algorithm.

## 6. Conclusion

Generalization error bounds of weighted voting classifier techniques can be expressed in terms of the training data errors and number of nonzero entries of the weight vector; see [8] and references therein. Our results here extend previous computational complexity results that only addressed the problems of minimizing each quantity independently of the other. Our extension shows that minimizing the number of misclassifications plus a penalty proportional to the number of nonzeroes is equally hard for a large range of penalty parameter values; this result is significant because the problem possesses a trivial solution for a sufficiently large penalty. We related an LP relaxation of our problem to previous algorithmic work, and proved an exponential lower bound on its integrality gap. On the other hand, we were able to prove a linear upper bound on the gap when the formulation is augmented by a polynomial number of novel inequalities. We believe that these results have practical significance for the design of sparse weighted voting classifier algorithms: in [9], we implemented a similar approach to a closely related formulation, with empirical results showing competitive classification performance while maintaining weight vector sparsity. This technique was less sensitive to parameter settings than prior methods omitting inequalities of the form (7).

## Acknowledgements

This material is based upon work funded in part by the U.S. Department of Homeland Security under Grant Award Number 2008-DN-077-ARI001-02, the Daniel Rose Technion-Yale Initiative for Research on Homeland Security and Counter-Terrorism, and the Council of Higher Education, State of Israel. We thank Rob Schapire for helpful discussions, and also thank the anonymous referees for comments that helped improve the presentation of these results. The first author would also like to thank Martin Milanic, Ilan Newman, and Asaf Levin for their comments.

## References

[1] N. Alon, V. Vu, Anti-hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs, J. Combin. Theory Ser. A 79 (1997) 133–160.
[2] E. Amaldi, V. Kann, On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems, Theoret. Comput. Sci. 209 (1998) 237–260.
[3] S. Arora, L. Babai, J. Stern, Z. Sweedyk, The hardness of approximate optima in lattices, codes, and systems of linear equations, J. Comput. System Sci. 54 (1997) 317–331.
[4] P. Bradley, O. Mangasarian, Feature selection via mathematical programming, INFORMS J. Comput. 10 (1998) 209–217.
[5] J. Brenner, The Hadamard maximum determinant problem, Amer. Math. Monthly 79 (1972) 626–630.
[6] A. Demiriz, K. Bennett, J. Shawe-Taylor, Linear programming boosting via column generation, Mach. Learn. 46 (2002) 225–254.
[7] I. Dinur, S. Safra, On the hardness of approximating label-cover, Inform. Process. Lett. 89 (2004) 247–254.
[8] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1997) 119–139.
[9] N. Goldberg, J. Eckstein, Boosting classifiers with tightened $L_0$-relaxation penalties, in: Proceedings of the Twenty-Seventh International Conference on Machine Learning, 2010, pp. 383–390.
[10] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.R. Müller, K. Obermayer, R. Williamson, Classification on proximity data with LP-machines, in: International Conference of Artificial Neural Networks, 1999, pp. 304–309.
[11] J. Hastad, On the size of weights for threshold gates, SIAM J. Discrete Math. 7 (1994) 484–492.
[12] K.U. Höffgen, H. Simon, K.V. Horn, Robust trainability of single neurons, J. Comput. System Sci. 50 (1995) 114–125.
[13] S. Muroga, I. Toda, S. Takasu, Theory of majority decision elements, J. Franklin Inst. 271 (1961) 376–418.
[14] J. Weston, A. Elisseeff, B. Schölkopf, M. Tipping, Use of the zero norm with linear models and kernel methods, J. Mach. Learn. Res. 3 (2003) 1439–1461.