RUTCOR
RESEARCH
REPORT

Sparse Weighted Voting Classifier Selection and its LP Relaxations

Noam Goldberg[a] Jonathan Eckstein[b]

RRR 9-2010, May 2010

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/~rrr

[a]Faculty of Industrial Engineering and Management, Technion – IIT, Haifa 32000, Israel; noamgold@tx.technion.ac.il

[b]MSIS Department and RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway NJ 08854; jeckstei@rci.rutgers.edu

Rutcor Research Report
RRR 9-2010, May 2010

Sparse Weighted Voting Classifier Selection and its LP Relaxations

Noam Goldberg Jonathan Eckstein

Abstract. We consider a combinatorial optimization problem that generalizes the minimum disagreement halfspace problem; we seek to minimize the number of misclassifications of a weighted voting classifier, plus a penalty proportional to the density of the vector of weights. To justify the use of this optimization problem, we investigate its relation to the minimum description length principle and statistical learning theory generalization bounds. We prove that the optimum is at least as hard to approximately compute as minimum disagreement halfspace for a large class of penalty parameters. After formulating the problem as a mixed integer program (MIP), we show that common "soft margin" linear programming formulations for constructing weighted voting classsifiers are equivalent to an LP relaxation of our formulation. We illustrate that the LP relaxation can be very weak, with an exponential lower bound on the potential intergrality gap. We then prove that augmenting the optimization problem with certain simple valid inequalities tightens the relaxation considerably, yielding a linear upper bound on the gap for all values of the penalty parameter that exceed a sensible lower bound.

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. We thank Rob Schapire for helpful discussions. The first author would also like to thank Kristin Bennett for her comments during earlier stages of this work, Martin Milanic for his comments regarding a related feature selection problem, and participants of the Operations Research seminar at the Technion for their comments.

1 Introduction

Consider a binary classification problem with M training samples, each consisting of N real-valued attributes, represented as a matrix $A \in \mathbb{R}^{M\times N}$ 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 into a "positive" class $\Omega^+ = \{i \in \{1,\ldots,M\} \mid y_i = 1\}$ and a "negative" class $\Omega^- = \{1,\ldots,M\} \setminus \Omega^+$. We suppose we have potentially large set of base classifiers $h_u : \mathbb{R}^N \to \{-1,0,1\}$ indexed by the set $\mathcal{U} = \{1,\ldots,U\}$, and would like to train a weighted voting classifier

$$g(x) = \sum_{u\in\mathcal{U}} \lambda_u h_u(x),$$

for $\lambda \in \mathbb{R}^U_+$. We classify any new observation $x \in \mathbb{R}^N$ as either positive or negative based on $\operatorname{sgn}(g(x))$.

The literature of learning algorithms for classification has considered various loss functions as classification performance measures. Common loss functions include the empirical 0/1 loss

$$\ell(y_i, g, A_i) = \mathbf{I}(y_i \neq \operatorname{sgn}(g(A_i))), \tag{1}$$

where $\mathbf{I}(\cdot)$ is the 0/1 indicator function and A_i is the i$^{\text{th}}$ row of A, and the *soft margin* loss (with margin fixed to 1):

$$\ell(y_i, g, A_i) = (1 - y_i g(A_i))_+, \text{ where } (\cdot)_+ = \max\{\cdot, 0\}. \tag{2}$$

The *empirical risk minimization* strategy calls for minimization of the average of losses (1) over the training data (or equivalently the sum of such losses) in order to determine λ. This strategy leads to an $\mathcal{NP}$-hard problem; we elaborate on the complexity of minimizing the 0/1 loss in Section 3. From the learning perspective, even more important is that empirical risk minimization can result in overfitting and significantly larger losses with respect to the unseen test data. More robust approaches to classification mitigate this problem by regularizing, or adding a model complexity penalty. Everything else being the same, it seems desirable to follow Occam's principle and select models for which $\lambda_u \neq 0$ for the smallest possible subset of $\mathcal{U}$, that is, to minimize $\|\lambda\|_0$, where the "L_0-norm" $\|\cdot\|_0$ is defined to be the number of nonzero coefficients in a vector (and is thus not a true L_p norm). Optimally sparse solutions of linear systems may be obtained by directly minimizing or penalizing the L_0 norm of λ. Unfortunately, minimizing the L_0-norm of λ for linearly separable data is also known to be $\mathcal{NP}$-hard (Amaldi & Kann, 1998). Nevertheless, inspired by the *Minimum Description Length* (MDL) principle (Grünwald, 2007), we consider here the problem of minimizing the observation-wise sum of the 0/1 loss function (1) plus an L_0-norm penalty. The MDL approach attempts to balance the complexity of a model against the empirical loss by addressing the problem of model selection as a data compression optimization problem. The optimal model minimizes the sum of the size of an efficient encoding of the model, together with a list of the observations which the model misclassifies. Section 2 will further discuss MDL in attempt to motivate our optimization formulation.

In order to avoid an $\mathcal{NP}$-hard combinatorial optimization problem, the authors of various classification methods such as LP-Boost, Lasso, and Support Vector Machines (SVMs) (Demiriz et al., 2002; Friedman, 2008; Cortes & Vapnik, 1995; Bennett & Bredensteiner, 2000) suggest using the L_1 or L_2 norms of λ, instead of L_0. Provided one is using an appropriate loss function, such strategies have the computational advantage of involving only convex optimization problems. For example, minimizing the observation-wise sum of the soft margin loss function (2), plus an L_p penalty, for $p \geq 1$, yields the convex optimization problem

$$\min_{\lambda} \sum_{i=1}^{M} (1 - y_i g(A_i))_+ + \|\lambda\|_p . \tag{3}$$

Alternatively, greedy algorithms have also been proposed for sparse feature and model selection (Schölkopf & Smola, 2002; Zhang, 2008; Huang et al., Submitted 2008).

Soft margin formulations have been suggested by several authors such as Bennett and Mangasarian (1992) and Cortes and Vapnik (1995). Cortes and Vapnik first suggested maximizing a "soft margin" for Support Vector Machines (SVMs); their formulation essentially solves (3) with $p = 2$. In that work, the term "soft margin" meant that not all observations need to be separated and satisfy the same requirement of distance from the hyperplane. Graepel et al. (1999) and Rätsch et al. (2001) adapted the quadratic optimization formulation of SVMs with soft margins to a linear programming (LP) formulation. In particular, setting $p = 1$ in (3) gives rise to the following linear program:

$$\min \left\{ \sum_{u=1}^{U} \lambda_u + D \sum_{i=1}^{M} \xi_i \;\middle|\; \operatorname{diag}(y) H \lambda + \xi \geq \mathbf{1}, \text{ and } \lambda, \xi \geq 0 \right\}. \tag{4}$$

Here, D is an appropriately chosed constant; see for example Bennett and Bredensteiner (2000). In this formulation, we seek to minimize the sum of the L_1-norm of λ and a soft margin loss penalty equivalent to (2). The margin of observation i is equal to $y_i H_i \lambda$, where H_i is the i^{th} row of H, an $M \times U$ matrix whose elements are $H_{iu} = h_u(A_i)$. Each observation $i \in \{1, ..., M\}$ has a variable ξ_i which allows it to have a margin smaller than 1. The parameter D penalizes the magnitude of each margin deviation $\xi_i = 1 - y_i H_i \lambda$; these deviations are illustrated in Figure 1.

Next, we will further motivate the problem of minimizing the sum of (1) and a penalty proportional to $\|\lambda\|_0$. We present the optimization problem more formally and study its computational complexity in Section 3. We explore this formulation's continuous LP relaxation and prove its equivalence to common LP formulations such as (4) in Section 4, also showing that the ratio of the optimal integer solution objective value and the relaxation solution value can grow exponentially in M.

Finally, Section 5 shows how to augment the relaxation with certain valid inequalities, maintaining a polynomial time solution in terms of M and U, in order to tighten the integrality gap to a factor of at most M, for all sensible values of the penalty parameter.

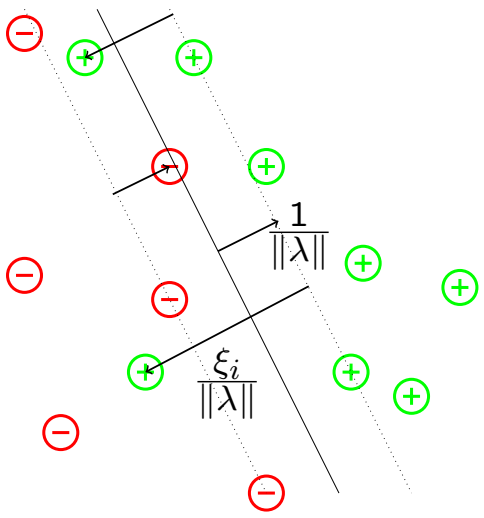

Figure 1: The margin deviations ξ_i illustrated in the soft margin LP formulation. The separating hyperplane is the solid line, and the supporting hyperplanes are the parallel dotted lines.

2 Statistical learning theory justification

We now motivate our approach by a brief appeal to statistical learning theory, which attempts to quantify prediction risk — the probability of error on the unseen test data for a given model. Classical bounds on prediction risk are expressed in terms of the Vapnik-Chervonenkis (VC) of the set of classifiers (Vapnik, 1998). Freund and Schapire (1997), drawing on work by Baum and Haussler (1989) for neural networks, have bounded the VC dimension of weighted voting classifiers in terms of the L_0-norm of λ and the VC dimension of the base classifiers:

Theorem 2.1 (Freund and Schapire). *The VC-dimension of the set of functions of the form*

$$\operatorname{sgn}(g(x)) = \operatorname{sgn}\left(\sum_{u\in\mathcal{U}} \lambda_u h_u(x)\right),$$

with $\|\lambda\|_0 \le T$ *is at most*

$$2(d+1)(T+1)\log((T+1)e),$$

where $d \ge 2$ *is the VC-dimension of the set of base classifiers* $\{\operatorname{sgn}(h_u(\cdot)) \mid u \in \mathcal{U}\}$.

Following later results by Bartlett (1998), Schapire et al. (1998), and Vapnik (1998, Chapters 9 and 10), algorithms for finding weighted voting classifiers that maximize the margin of separation, such as boosting and Support Vector Machines (SVMs), have been motivated by other risk bounds expressed in terms of the margin of separation, or equivalently the L_1 or L_2 norm of λ.

The connection between statistical learning theory, (L_2) margin maximization, and compression has been investigated by von Luxburg et al. (2004). Although SVMs are designed to

maximize the margin of separation in the space of features subject to a soft margin penalty, von Luxburg et al. found that not only were their compression-based risk bounds tighter than margin-based bounds, but they also resulted in better classification performance when used to fine-tune SVM parameters. Their analysis relates the idea of compression to the notion of separation margin by showing that the larger the L_2 margin of separation, the lower the precision needed to encode the voting weights λ, and the more the classifier description may be compressed. While von Luxburg et al. help to illuminate the relation between the L_2 margin of separation and compression, their work raises the question whether there can be more direct approaches to "compressing" weighted voting classifiers. For example, are there efficient methods that more directly optimize sparsity and attempt to minimize the length of the classifier description?

Here, we adopt a strategy inspired by an MDL two-part-code conception of learning: an imaginary sender and receiver are assumed to share the unlabeled data and to agree in advance on the set of candidate features. The sender has access to the labels, and wishes to minimize the length of a two-part code intended to transmit them to the receiver: the first part of the code describes the classifier g, and the second part encodes which observations are misclassified by g. An "optimal" g minimizes the total length of this code.

MDL can be related to statistical learning theory through the compression interpretation of Vapnik (1998), as well as that of Blum and Langford (2003). Let $\bar{L}(y, g)$ denote the length function of the code being used to encode the labels. For simplicity, assume that the number of test observations is also M, the same as the number of training observations; we denote the test data by $A' \in \mathbb{R}^{M \times N}$ and the associated labels by $y' \in \{-1, 1\}^M$. The resulting bound on the risk of g, by Blum and Langford (2003), is that with probability $1 - \delta$:

$$\tilde{R}[g] = \frac{1}{M} \sum_{i=1}^{M} \mathbf{I}(y'_i g(A'_i) \leq 0) \leq \frac{\bar{L}[y, g]}{M} - \frac{\ln \delta}{M}. \tag{5}$$

We write $\bar{L}[y, g] = \dot{L}[g] + L[y|g]$, where $\dot{L}[g]$ is the number of bits needed to encode the weighted voting classifier g, and $L[y|g]$ is the number of bits needed to encode the set of observations misclassified by g; note that $L[y|g] \leq \sum_{i=1}^{M} \mathbf{I}(y_i \neq \operatorname{sgn}(g(A_i)))\lceil \log M \rceil$. If $\mathcal{G}$ denotes the set of all weighted voting classifier models, then our theoretical objective is to solve the problem

$$\min_{g \in \mathcal{G}} \bar{L}[y, g] = \min_{g \in \mathcal{G}} \{\dot{L}[g] + L[y|g]\}.$$

To describe a classifier g, the sender must specify the elements of $\mathcal{U}(\lambda) = \{u \in \mathcal{U} \mid \lambda_u \neq 0\}$, requiring at most $\|\lambda\|_0 \lceil \log U \rceil$ bits. In addition to this set of features, one must also encode the numerical values of the hyperplane weights λ. Rather than directly encoding the elements of λ to potentially very high precision, we instead proceed as follows: let $S(\lambda) = \{i \in \{1, \ldots, M\} \mid y_i H_i \lambda \geq 1\}$ denote the set of observations correctly classified with the weights λ, with margin at least 1 (note that by scaling λ if necessary, we can make $S(\lambda)$ the set of all correctly classified observations). Next, consider some particular weight vector $\bar{\lambda}$ that we

might wish to encode, and define

$$\Lambda(\bar{\lambda}) = \left\{ \lambda \in \mathbb{R}^U \;\middle|\; \begin{array}{ll} \sum_{u \in \mathcal{U}} y_i H_i \lambda \geq 1 & \forall\, i \in S(\bar{\lambda}) \\ \lambda_u = 0 & \forall\, u : \bar{\lambda}_u = 0 \\ \lambda \geq 0 & \end{array} \right\}. \tag{6}$$

The weight vectors $\lambda \in \Lambda(\bar{\lambda})$ are those that perform at least as well as $\bar{\lambda}$, in the sense that they correctly classify all the observations correctly classified with $\bar{\lambda}$, without using any features except those appearing in $\bar{\lambda}$. Thus, if we transmit any element of $\Lambda(\bar{\lambda})$, that is at least as good in principle as encoding $\bar{\lambda}$. Noting that $\Lambda(\bar{\lambda})$ is a polyhedron, we will consider encoding any vertex (extreme point) of $\Lambda(\bar{\lambda})$. For example, given any $\bar{\lambda}$, one could identify a vertex of $\Lambda(\bar{\lambda})$ having minimum L_1 norm by solving the linear programming problem $\min_{\lambda \in \Lambda(\bar{\lambda})} \{\sum_{u \in \mathcal{U}} \lambda_u\}$; this choice will in fact maximize the L_∞ distance between the classifier hyperplane $\{x \mid \sum_{u \in \mathcal{U}} \lambda_u h_u(x) = 0\}$ and the margin hyperplanes $\{x \mid \sum_{u \in \mathcal{U}} \lambda_u h_u(x) = \pm 1\}$; see (Mangasarian, 1999; Bennett & Bredensteiner, 2000). We next observe that any vertex of $\Lambda(\bar{\lambda})$ may be encoded in a compact manner:

Theorem 2.2. *For $\bar{\lambda} \in \mathbb{R}^U_+$, any vertex λ of $\Lambda(\bar{\lambda})$ may be encoded in $\|\lambda\|_0 (\lceil \log M \rceil + 1) \leq \|\bar{\lambda}\|_0 (\lceil \log M \rceil + 1)$ bits, in addition to the encoding of $\widehat{\mathcal{U}}(\lambda) = \{u \in \mathcal{U} \mid \lambda_u \neq 0\}$.*

Proof. Any vertex λ of $\Lambda(\bar{\lambda})$ corresponds to the intersection of U linearly independent binding constraints in the definition (6), at least $\|\lambda\|_0$ of which cannot be the simple nonnegativity or equality constraints. By specifying the indices of any $\|\lambda\|_0$ of these binding constraints, along with corresponding labels y_i, the receiver may solve a linear system of $\|\lambda\|_0$ variables and the same number of equations for the weights λ. For any two-part code where the sender and receiver share the M data points, each can be identified using $\lceil \log M \rceil$ bits, while its label can be identified using one additional bit. □

Note that the observations selected to encode λ are *support vectors*, that is, observations with margin ± 1. If we encode the coefficients of λ as suggested in Theorem 2.2, the total length of the resulting encoded representation of g is:

$$\dot{L}[g] \leq \|\lambda\|_0 (\lceil \log U \rceil + \lceil \log M \rceil + 1). \tag{7}$$

Thus, by (7), minimizing $\|\lambda\|_0$ is equivalent to minimizing an upper bound on the code length $\dot{L}[g]$.

In Appendix A, we will explore a tighter code length bound that may apply when combining base classifiers of different *classes of risk*, meaning that some features are considered more complex, and hence require more bits to encode, than others. In this case, the portion $\|\lambda\|_0 \lceil \log U \rceil$ of the bound (7) that corresponds to encoding the set of features $\mathcal{U}(\lambda)$ is replaced by a complicated expression. In the optimization formulations we develop below, this means that each each feature would have a potentially different penalty parameter; in the body of this paper, we will restrict our analysis to the more basic setting inspired by the bound (7) in which every possible feature is assumed to have the same "cost", and is thus assigned an identical penalty parameter.

3 Sparse weighted voting classifier

3.1 Problem formulation

The problem of finding $\lambda \in \mathbb{R}^U$ so that g that minimizes the sum of (1) over $i = 1, \ldots, M$ is known as the *minimum disagreement halfspace* problem (MDH), and is $\mathcal{NP}$-hard (Höffgen et al., 1995; Arora et al., 1997). Mangasarian (1994) and Bennett and Bredensteiner (1997) have proposed heuristic approaches to this problem, based on nonlinear programming; both of these works refer to a hardness result in (Heath, 1992). The problem of minimizing the classification error is also known to be related to the *minimum infeasible subsystem* problem: finding a feasible subsystem by excluding the least number of inequalities from a given infeasible system $Ax \leq b$ (Amaldi & Kann, 1998; Pfetch, 2008).

We extend the minimum disagreement halfspace problem to penalize the L_0 norm of λ. Without losing generality, we require λ to be nonnegative; at the cost potentially doubling the size of U by including the negative "mirror image" $h_{u^-} = -h_u$ of each base classifier h_u, we may simply replace each unrestricted term $\lambda_u h_u(x)$ in the classifier with $\lambda_u h_u(x) + \lambda_{u^-} h_{u^-}(x) = (\lambda_u - \lambda_{u^-}) h_u(x)$, where $\lambda_u, \lambda_{u^-} \geq 0$. We call the more general problem including an L_0 penalty the *sparse weighted voting classifier* problem (SWVC). In the basic version of the problem, we penalize all non-zero components of λ uniformly through the same penalty parameter C. The problem can be stated as:

Sparse Weighted Voting Classifier (SWVC)

Input: A matrix $H \in \{-1, 0, 1\}^{M \times U}$ of base classifier labels, a corresponding vector $y \in \{-1, 1\}^M$ of sample labels, and a penalty parameter $0 \leq C < M$

Problem: To find a separating hyperplane, as specified by $\lambda \in \mathbb{R}^U_+$, such that the function $\sum_{i=1}^M \mathbf{I}(y_i H_i \lambda < 1) + C \left\|\lambda\right\|_0$ is minimized.

Note that we exclude the case that $C \geq M$ because it always has the trivial solution $\lambda = 0$.

3.2 Formulating SWVC as a mixed integer linear program

We now formulate SWVC as a Mixed Integer Program (MIP), using variables μ_u to indicate whether feature u is used, and variables ξ_i to indicate whether observation i is misclassified:

$$\min_{\xi,\mu,\lambda} \left\{ \sum_{i=1}^{M} \xi_i + C \sum_{u=1}^{U} \mu_u \;\middle|\; (\xi, \mu, \lambda) \in Q_{H,y} \cap \left(\{0,1\}^M \times \{0,1\}^U \times \mathbb{R}^U_+ \right) \right\}, \tag{8}$$

where $Q_{H,y}$ is a "soft margin" classification polyhedron defined as

$$Q_{H,y} = \left\{ (\xi, \mu, \lambda) \in \mathbb{R}^M_+ \times \mathbb{R}^U_+ \times \mathbb{R}^U_+ \;\middle|\; \begin{array}{l} \operatorname{diag}(y) H \lambda + (MK + 1)\xi \geq \mathbf{1} \\ \lambda \leq K\mu \end{array} \right\},$$

K being a suitably large constant, and $\operatorname{diag}(y)$ the diagonal matrix with nonzero entries $y_1, \ldots, y_M$. We show below that (8) is equivalent to SWVC for all large enough K. The

magnitude of the constant, however, will determine the (poor) quality of the MIP LP relaxation, which we will further explore in Section 4. Note that SWVC and thus formulation (8) always has a feasible solution. Therefore, since the objective value of any feasible solution to either SWVC or formulation (8) is bounded, each must always have an optimal solution. Thus, to prove the equivalence of (8) and SWVC, it is sufficient to show in the following theorem that the optimal solution values of SWVC and (8) are equal.

Theorem 3.1. *If $K \geq M^{M/2}$, then for any optimal solution $(\xi^*, \mu^*, \lambda^*)$ of (8), it must be that*

$$\sum_{i=1}^{M} \xi_i^* + C \sum_{u \in \mathcal{U}} \mu_u^* = \min_{\lambda \in \mathbb{R}_+^N} \sum_{i=1}^{M} \mathbf{I}(y_i H_i \lambda \leq 0) + C \left\|\lambda\right\|_0 = \sum_{i=1}^{M} \mathbf{I}(y_i H_i \lambda^* < 1) + C \left\|\lambda^*\right\|_0 .$$

To prove this result, we will require the following two Lemmas.

Lemma 3.2. *Suppose that for some subset of features $\Gamma \subseteq \mathcal{U}$ and some weight vector $\bar{\lambda} \geq 0$, the hyperplane $\bar{g}(x) = \sum_{u \in \Gamma} \bar{\lambda}_u h_u(x) = 0$ strictly separates $S^+ \subseteq \Omega^+$ and $S^- \subseteq \Omega^-$. Then there exists λ^* such that $0 \leq \lambda_u^* \leq M^{M/2}$ for all $u \in \Gamma$, and $g(x) = \sum_{u \in \Gamma} \lambda_u^* h_u(x) = 0$ also separates S^+ and S^-.*

Proof. By the hypothesis of linearly separability of the points S^+ and S^- with $\lambda \geq 0$, the linear system

$$\sum_{u \in \Gamma} y_i H_{iu} \lambda_u - s_i = 1 \qquad \forall\, i \in S^+ \cup S^- \tag{9}$$

$$\lambda, s \geq 0 \tag{10}$$

has a feasible solution. Thus, by the fundamental theorem of linear programming, this system of inequalities must have a basic feasible solution (BFS) $\lambda^*, s^* \geq 0$. Let B denote the corresponding basis, composed of linearly independent columns of $\begin{bmatrix}\mathrm{diag}(y)H & -I\end{bmatrix}$. Using Cramer's rule, it follows that every nonzero element λ_u^* of λ^* satisfies

$$\lambda_u^* = \frac{\det\left(B^{(u)}\right)}{\det(B)} \geq 0,$$

where $B^{(u)}$ is the matrix B with the column corresponding to feature u replaced by $\mathbf{1}$. Since the rank of B is bounded by $|S^+ \cup S^-| \leq M$, we have by Hadamard's bound (Brenner, 1972, for example) that $\left|\det\left(B^{(u)}\right)\right| \leq M^{M/2}$. Since B is a basis, $\det(B) \neq 0$. Thus, by the definition of the determinant and since $B_{ij} \in \{-1, 0, 1\}$,

$$|\det(B)| = \left| \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} B_{i\sigma(i)} \right| \geq 1,$$

and the claim follows. □

Lemma 3.3. *If there exists $\bar{\lambda} \geq 0$ such that the hyperplane $\sum_{u\in\mathcal{U}} \bar{\lambda}_u h_u(x) = 0$ strictly separates $S^+ \subseteq \Omega^+$ and $S^- \subseteq \Omega^-$, then there exists $\lambda^* \geq 0$ such that $\sum_{u\in\mathcal{U}} \lambda^*_u h_u(x) = 0$ strictly separates S^+ and S^-, and $\|\lambda^*\|_0 \leq |S^+| + |S^-|$.*

Proof. Let

$$Y = \left[\begin{array}{c|c} y_{i_1}H_{i_1} & \\ \vdots & -I \\ y_{i_k}H_{i_k} & \end{array}\right],$$

where $\{i_1, \ldots, i_k\} = S^+ \cup S^-$. By the hypothesis, the linear system (9)-(10) has a feasible solution and thus has a basic feasible solution (λ^*, s^*). The number of nonzero components in a basic feasible solution is at most the rank of the constraint matrix, in this case the matrix Y. So,

$$\|\lambda\|_0 \leq \|\lambda^*\|_0 + \|s^*\|_0 \leq \operatorname{rank}(Y) = |S^+| + |S^-|,$$

the equality following because Y has $|S^+| + |S^-|$ independent rows. □

Proof. [Proof of Theorem 3.1] By the constraints $\lambda_u \leq K\mu_u$ of formulation (8), we have $\|\lambda^*\|_0 \leq \sum_{u\in\mathcal{U}} \mu^*_u$. By the constraint $\operatorname{diag}(y)H\lambda^* + (MK+1)\xi^* \geq \mathbf{1}$, it also follows that $\sum_{i=1}^M \mathbf{I}(y_i H_i \lambda^* < 1) \leq \sum_{i=1}^M \xi^*_i$. Assume λ is optimal for SWVC; then

$$\sum_{i=1}^M \mathbf{I}(y_i H_i \lambda < 1) + C\|\lambda\|_0 \leq \sum_{i=1}^M \mathbf{I}(y_i H_i \lambda^* < 1) + C\|\lambda^*\|_0 \leq \sum_{i=1}^M \xi^*_i + C\sum_{u\in\mathcal{U}} \mu^*_u. \tag{11}$$

We now prove the reverse inequality between the first and last quantities. Let

$$S^+ = \left\{ i \in \Omega^+ \;\middle|\; y_i H_i \lambda \geq 0 \right\} \qquad S^- = \left\{ i \in \Omega^- \;\middle|\; y_i H_i \lambda \geq 0 \right\}$$

denote the sets which are correctly classified by the hyperplane corresponding to λ. Now, by Lemma 3.3 and the assumed optimality of λ for SWVC, we also have $\|\lambda\|_0 \leq M$. By Lemma 3.2 with $\Gamma = \mathcal{U}(\lambda) = \{u \in \mathcal{U} \mid \lambda_u \neq 0\}$, there exists λ'', with $\lambda''_u \leq M^{M/2} \leq K$ for all $u \in \mathcal{U}$, with the same support as λ, separating S^+ and S^-. Thus, $|y_i H_i \lambda''| \leq M^{M/2+1}$ for all $i = 1, \ldots, M$. Next, let

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

Then, for all $i \in \{1, \ldots, M\}$,

$$y_i H_i \lambda'' + \left(M^{M/2+1} + 1\right)\xi''_i \geq 1.$$

Thus, $(\xi'', \mu'', \lambda'')$ is feasible to (8), and $\sum_{i=1}^M \xi''_i = \sum_{i=1}^M \mathbf{I}(y_i H_i \lambda < 1)$. Therefore, by the optimality of $(\xi^*, \mu^*, \lambda^*)$ for (8),

$$\sum_{i=1}^M \xi^*_i + C\sum_{u\in\mathcal{U}} \mu^*_u \leq \sum_{i=1}^M \xi''_i + C\sum_{u\in\mathcal{U}} \mu''_u \leq \sum_{i=1}^M \mathbf{I}(y_i H_i \lambda < 1) + C\|\lambda\|_0.$$

Thus, all the relations in (11) hold with equality, and therefore λ^* is also an optimal solution to SWVC. □

3.3 Computational complexity

The SWVC problem generalizes the MDH problem, so it is at least as hard to solve computationally. Specifically, an MDH instance (H', y') can be reduced to (8) with $H = \begin{pmatrix} H' & -H' \end{pmatrix}$, $y = y'$ and $C = 0$. We will also refer to any solution (ξ, μ, λ) of (8), after applying this reduction to an MDH instance (H, y), as an *MDH solution*. Höffgen et al. (1995) showed by reduction of the set cover problem that MDH cannot be approximated within a factor better than $(1 - \epsilon) \log M$, for any $\epsilon > 0$, unless $\mathcal{NP} \subseteq \text{DTIME}(M^{\log \log M})$.[1] Arora et al. (1997) increased the inapproximability factor to $2^{\log^{1-\epsilon} M}$, by reduction of the label cover problem, while making the weaker assumption $\mathcal{NP} \nsubseteq \text{DTIME}(M^{\text{poly}(\log M)})$; see also (Amaldi & Kann, 1998). Finally, Dinur and Safra (2004) strengthened the inapproximability of the label cover problem used in the reduction to $2^{\log^{(1-o(1))} M}$, making the even weaker assumption $\mathcal{P} \neq \mathcal{NP}$. The same strengthened inapproximability result also applies to MDH, through the polynomial reduction of Arora et al. (1997).

Considering the special case of SWVC with $C = 0$ and appropriate negated duplicated columns, existing MDH inapproximability results directly apply to SWVC. Alternatively, by setting C to be any small positive constant such that $0 < C < 1/M$, we arrive at a special case of another $\mathcal{NP}$-hard problem, the problem of minimizing the number of relevant variables in a linear system (Amaldi & Kann, 1998). We state in the following theorem a more general inapproximability result for SWVC for any choice penalty $C \in O(M^\delta)$, where $0 \leq \delta < 1$, making use of the $2^{\log^{1-\epsilon} M}$-factor inapproximability for MDH (Arora et al., 1997; Amaldi & Kann, 1998) and the work of of Dinur and Safra (2004). Note that although the problem is $\mathcal{NP}$-hard in general and for the above-mentioned large class of penalty parameters, SWVC has the trivial solution $\lambda = 0$ and $\xi = \mathbf{1}$ whenever $C \geq M$. We will require the following lemmas to derive our inapproximability result:

Lemma 3.4. *Given an MDH instance with input $H \in \{-1, 0, 1\}^{M \times U}$, and $y \in \{-1, 1\}^M$, and some parameter value $C \geq 0$, then for any 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 $\sum_{i=1}^{Mk} \xi_i^* = k \left(\sum_{i=1}^{M} \hat{\xi}_i \right)$ and $\sum_{u=1}^{2U} \mu_u^* \leq M$.*

Proof. Given the input (H, y) for MDH, and a penalty C, construct an instance of SWVC (H', y'), by creating k corresponding duplicates of the row $\begin{bmatrix} H_i & -H_i \end{bmatrix}$ in the matrix H', and k duplicates of y_i in the vector y', for $i = 1, \ldots, M$. Without loss of generality, assume that the rows of H' and y' are indexed such that $H'_i = \begin{bmatrix} H_i & -H_i \end{bmatrix}$ and $y'_i = y_i$ for $i = 1, \ldots, M$. 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 solution of MDH, corresponding to formulation (8) with $C = 0$ and H replaced by $\begin{bmatrix} H & -H \end{bmatrix}$. The objective value of this solution is $z_{\text{MDH}} = \sum_{i=1}^{M} \hat{\xi}_i + 0 \sum_{u=1}^{2U} \hat{\mu}_u = \sum_{i=1}^{M} \hat{\xi}_i$. Now, since feasible solutions of MDH and SWVC must always exist, we only need to prove

[1] DTIME(n) is the class of problems that can be solved in deterministic time n.

$z_{\text{MDH}} = \sum_{i=1}^{M} \hat{\xi}_i = (1/k) \sum_{i=1}^{Mk} \xi_i^*$. Assume to the contrary that

$$z_{\text{MDH}} \neq \frac{1}{k} \sum_{i=1}^{Mk} \xi_i^*.$$

First, we note that we must have $z_{\text{MDH}} \leq (1/k) \sum_{i=1}^{Mk} \xi_i^*$, since otherwise we could construct an MDH solution with objective below z_{MDH}. On the other hand, if $z_{\text{MDH}} < (1/k) \sum_{i=1}^{Mk} \xi_i^*$, then by the integrality of the MDH objective,

$$z_{\text{MDH}} \leq \frac{1}{k} \sum_{i=1}^{Mk} \xi_i^* - 1 \tag{12}$$

By Lemma 3.3, since the sets $\{i \in \Omega^+ \mid \hat{\xi}_i = 0\}$ and $\{i \in \Omega^- \mid \hat{\xi}_i = 0\}$ are linearly separable, there exists λ corresponding to a hyperplane that separates $\{i \in \Omega^+ \mid \hat{\xi}_i = 0\}$ and $\{i \in \Omega^- \mid \hat{\xi}_i = 0\}$, with $\|\lambda\|_0 \leq M$. Then the optimal SWVC solution value z_{SWVC} must satisfy

$$\begin{aligned}
z_{\text{SWVC}} &= \sum_{i=1}^{Mk} \xi_i^* + C \sum_{u=1}^{2U} \mu_u^* \\
&\leq k \sum_{i=1}^{M} \hat{\xi}_i + C \sum_{u=1}^{2U} \mathbf{I}(\lambda_u \neq 0) && \text{[by optimality of } \xi^*, \mu^*] \\
&\leq k z_{\text{MDH}} + CM \\
&< k \left(z_{\text{MDH}} + 1\right) && \text{[by } k > CM] \\
&\leq \sum_{i=1}^{Mk} \xi_i^* && \text{[by (12)]} \\
&\leq z_{\text{SWVC}}.
\end{aligned}$$

From this contradiction, we conclude that $z_{\text{MDH}} = \sum_{i=1}^{Mk} \xi_i^* = (1/k) \sum_{i=1}^{M} \xi_i^*$. □

Lemma 3.5. *A polynomial-time $f(M)$-approximation factor for SWVC with penalty $C = C(M) \in O(M^\delta)$, for some $0 \leq \delta < 1$ 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. We will convert an MDH instance (H, y) to an SWVC instance (H', y', C), using the reduction of Lemma 3.4, for some $k > CM$. Let $M' = kM$ denote the number of observations in the resulting SWVC instance. Suppose, for some choice of $C \in O(M'^\delta)$ with $0 \leq \delta < 1$, that there is a polynomial-time $f(M')$-factor approximation algorithm for SWVC, and let (ξ, μ, λ) be the solution it returns for the instance (H', y', C). Furthermore, let $(\xi^*, \mu^*, \lambda^*)$ be an optimal SWVC solution for this instance. Let $(\hat{\xi}, \hat{\mu}, \hat{\lambda})$ denote the optimal MDH solution, and z_{MDH} denote its objective value.

We will assume that $z_{\text{MDH}} = \sum_{i=1}^{M} \hat{\xi}_i \geq \kappa$, for some constant $\kappa \in \mathbb{N}_+$; all MDH instances violating this assumption may be solved exactly in polynomial time by excluding each possible subset of observations up to size κ and finding the corresponding separating hyperplanes by linear programming. Now,

$$\begin{aligned}
z_{\text{MDH}} &= \sum_{i=1}^{M} \hat{\xi}_i = \frac{1}{k} \sum_{i=1}^{kM} \xi_i^* && \text{[by Lemma 3.4]} \\
&\leq \frac{1}{k} \left(\sum_{i=1}^{kM} \xi_i^* + C \sum_{u=1}^{2U} \mu_u^* \right) \leq \frac{1}{k} \left(\sum_{i=1}^{kM} \xi_i + C \sum_{u=1}^{2U} \mu_u \right) \\
&\leq \frac{f(kM)}{k} \left(\sum_{i=1}^{kM} \xi_i^* + C \sum_{u=1}^{2U} \mu_u^* \right) && \text{[by assumption]} \\
&\leq \frac{f(kM)}{k} \left(k \sum_{i=1}^{M} \hat{\xi}_i + CM \right) && \left[\begin{array}{l} \text{by optimality of } \xi^*, \mu^* \\ \text{and Lemma 3.3} \end{array} \right] \\
&\leq f(kM) \left(\sum_{i=1}^{M} \hat{\xi}_i + 1 \right) && \text{[by } k > CM\text{]} \\
&\leq f(kM)(1 + 1/\kappa) \sum_{i=1}^{M} \hat{\xi}_i,
\end{aligned}$$

Note that the bound $\sum_{u=1}^{2U} \mu_u^* \leq M$ follows from Lemma 3.3 with $S^+ \subseteq \Omega^+$ and $S^- \subseteq \Omega^-$, since any separating hyperplane for S^+, S^- must also separate the corresponding duplicates within the duplicate observations $\{M+1, \ldots, kM\}$. Now, by the hypothesis, $C \leq \gamma(kM)^\delta$, for some $\gamma \in O(1)$. For the analysis above to hold, we require $k > CM$; since $C \leq \gamma(kM)^\delta$, 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

$$k = \left\lceil \gamma^{1-\delta} M^{(1+\delta)/(1-\delta)} + \epsilon \right\rceil$$

where $\epsilon > 0$ is a small constant. Note that there exists some constant $\beta > 0$ such that $k \leq \beta M^{(1+\delta)/(1-\delta)}$ for all M. Constructing an SWVC instance as in Lemma 3.4 and then running the hypothesized approximation algorithm, we obtain an approximate solution with approximation ratio

$$f(kM)(1 + 1/\kappa) \leq f(\beta M^{1+(1+\delta)/(1-\delta)})(1 + 1/\kappa) = \alpha f(\beta M^{1+(1+\delta)/(1-\delta)}),$$

where we set $\alpha = 1 + 1/\kappa$. Since the choices of k and C are polynomially bounded in M, and the reduction of Lemma 3.4 is polynomial in k and U, this procedure is polynomial. □

Theorem 3.6. *For any penalty $C \in O(M^\delta)$ with $0 \leq \delta < 1$, 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 3.5, a polynomial-time $2^{\log^{1-\epsilon} M}$-factor approximation for SWVC, for some $\epsilon > 0$, yields a polynomial-time approximation for MDH with factor $\alpha 2^{\log^{1-\epsilon} \beta M^{1+(1+\delta)/(1-\delta)}}$, for some $\alpha, \beta \in O(1)$. Now,

$$\alpha 2^{\log^{1-\epsilon} \beta M^{1+(1+\delta)/(1-\delta)}} \leq \alpha 2^{[1+(1+\delta)/(1-\delta)]^{(1-\epsilon)} \log^{1-\epsilon} M} \leq 2^{\log^{1-\epsilon'} M}$$

for some $0 < \epsilon' \leq \epsilon$, $M_0 \in \mathsf{N}_+$ and all $M \geq M_0$. Unless $\mathcal{P} = \mathcal{NP}$, such an approximation factor contradicts the inapproximability of MDH following from the reduction of Arora et al. (1997); Amaldi and Kann (1998) and the strengthened inapproximability result of label cover (Dinur & Safra, 2004). □

4 SWVC relaxation and its integrality gap

For sufficiently large K, the continuous LP relaxation of the MIP formulation (8) may be stated as

$$\min \left\{ \sum_{i=1}^{M} \xi_i + C \sum_{u=1}^{U} \mu_u \;\middle|\; \begin{array}{l} \operatorname{diag}(y) H \lambda + (MK+1)\xi \geq \mathbf{1} \\ 0 \leq \lambda \leq K\mu \\ \xi \geq 0 \end{array} \right\}. \tag{13}$$

We next show that the LP relaxation is equivalent to the well known soft margin formulation (4) with appropriate choices of the penalties C and D. The theorem will also enable us to claim in Section 4 that our LP formulation provides a tightened relaxation of the discrete SWVC formulation (8), by introducing novel cutting planes which strengthen the relaxation of the soft margin LP.

Theorem 4.1. *For every instance (H, y), (ξ, λ) is an optimal solution of (4) if and only if $(\hat{\xi}, \hat{\mu}, \lambda)$ is an optimal solution of (13) where $\hat{\xi} = \xi/(MK+1)$, $\hat{\mu} = \lambda/K$ and $C = 1/(D(M + 1/K))$.*

Proof. Consider the map

$$\omega : (\lambda, \xi, D) \mapsto \left(\lambda, \tfrac{1}{MK+1}\xi, \tfrac{1}{K}\lambda, \tfrac{1}{D(M+1/K)}\right).$$

Take any $D > 0$ and (ξ, λ) that is a feasible solution of (4), and let $(\lambda, \hat{\xi}, \hat{\mu}, C) = \omega(\lambda, \xi, D)$, that is,

$$\hat{\xi} = \tfrac{1}{MK+1}\xi \qquad \hat{\mu} = \tfrac{1}{K}\lambda \qquad C = \tfrac{1}{D(M+1/K)}.$$

Now,

$$\operatorname{diag}(y) H \lambda + \xi = \operatorname{diag}(y) H \lambda + (MK+1)\hat{\xi} \geq \mathbf{1},$$

and $\hat{\mu} = \lambda/K$ imply that $(\lambda, \hat{\xi}, \hat{\mu})$ is feasible for (13). The objective value of the solution $(\lambda, \hat{\xi}, \hat{\mu})$ to (13) is

$$\begin{aligned}\sum_{i=1}^{M} \hat{\xi} + C \sum_{u=1}^{U} \hat{\mu}_u &= \tfrac{1}{MK+1} \textstyle\sum_{i=1}^{M} \xi_i + \tfrac{1}{D(M+1/K)} \sum_{u=1}^{U} \lambda_u/K \\ &= \tfrac{1}{D(MK+1)} \left(D \textstyle\sum_{i=1}^{M} \xi_i + \sum_{u=1}^{U} \lambda_u \right).\end{aligned}$$

Thus, ω maps feasible solutions of (4) to feasible solutions of (13), scaling the objective value by the constant $1/(D(MK+1))$. Conversely, if one takes any solution $(\lambda, \hat{\xi}, \hat{\mu})$ to (13) with $\hat{\mu} = \lambda/K$, then the inverse image of $(\lambda, \hat{\xi}, \hat{\mu}, C)$ under the map ω is a singleton $\{(\lambda, \xi, D)\}$ such that (λ, ξ) is feasible for (4), with objective value scaled by by $D(MK+1)$. The conclusion then follows by noting that in all optimal solutions of (13), one must have $\mu = \lambda/k$, since the nonnegative variables μ_u have positive objective coefficients, each appears only in the constraint $\mu_u \geq \lambda_u/K$, and the objective is being minimized. □

The weakness of the continuous relaxations of (8) is reflected in a large *integrality gap*. The integrality gap of our MIP formulation of SWVC may be defined as

$$\sup_{H,y} \left\{ \frac{z(H,y)}{z_{\mathrm{R}} R(H,y)} \right\},$$

where $z(H,y)$ and $z_{\mathrm{R}}(H,y)$ are the optimal solution values of the SWVC MIP and its continuous relaxation, respectively; see Vazirani (2003).

In Lemma 3.2, we proved a large upper bound for the required constant K, *i.e.*, that formulation (8) is correct for all $K \geq M^{M/2}$. Although such a K is sufficient it may not be necessary. The following lemma identifies a lower bound on possible values of K:

Lemma 4.2. *Let λ' be optimal for SWVC. In order for formulation* (8) *to have an optimal λ satisfying*

$$y_i H_i \lambda \geq 1 \Leftrightarrow y_i H_i \lambda' \geq 1 \qquad \text{for } i = 1, \ldots, M,$$

for all instances, one must have $K \geq 2^{M-1}$.

Proof. Consider an SWVC instance such that

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

By using all the features, it is possible to correctly classify every observation, and if $C \leq 1$, it will be optimal to do so. In (8), this means we must have an optimal solution (ξ, μ, λ) with $\xi = 0$ and $\mu = \mathbf{1}$. By construction, we must have $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 4, \ldots, \lambda_M = 2^{M-1}$. Thus, formulation (8) must have $K \geq 2^{M-1}$ □

Theorem 4.3. *The integrality gap of SWVC is at least* $M2^{M-1}$.

Proof. Consider the simple SWVC instance given by $C = 1$ and $\mathrm{diag}(y)H = I$ (the identity matrix), meaning that each base classifier covers only a single observation. Since each observation $i \in \{1, \ldots, M\}$ must be either classified correctly by the single classifier u with $y_i H_{iu} = 1$ and $\mu_u = 1$, or otherwise $\xi_i = 1$, this instance has an optimal integer solution of value M, where M of the μ_u and ξ_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 $i = 1, \ldots, M$, and $\mu = 0$, with value $M/(MK+1)$. Since Lemma (4.2) requires $K \geq 2^{M-1}$ in general, the integrality gap of SWVC satisfies

$$\sup_{H,y} \left\{ \frac{z(H,y)}{z_R(H,y)} \right\} \geq \frac{M}{M/(M(2^{M-1})+1)} \geq M2^{M-1}.$$

□

5 Tightening the integrality gap using valid inequalities

We now consider adding valid inequalities to (8) 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.*, $h_u(A_i) = y_i \neq h_u(A_{i'})$. Let $S_{i,i'} = \{u \in \mathcal{U} \mid h_u(A_i) = y_i \neq h_u(A_{i'})\}$ denote the set of base classifiers that correctly classify observation i and distinguish it from i'. We consider the following inequality for each pair of observations $(i, i') \in \Phi = (\Omega^+ \times \Omega^-) \cup (\Omega^- \times \Omega^+)$:

$$\xi_i + \xi_{i'} + \sum_{u \in S_{i,i'}} \mu_u \geq 1. \tag{14}$$

Intuitively, such a cutting plane implies 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 $S_{i,i'}$.

Theorem 5.1. *The inequalities* (14) *are valid, that is, they hold for all integer-feasible solutions of* (8).

Proof. Take any $(i, i') \in \Omega^+ \times \Omega^-$. If $\xi_i + \xi_{i'} \geq 1$ then (14) clearly holds. Otherwise, $i \in \Omega^+$ and $\xi_i = \xi_{i'} = 0$ imply that $\sum_{u \in \mathcal{U}} H_{iu}\lambda_u \geq 1$. Thus, $h_u(A_i)\lambda_u > 0$ for some $u \in \mathcal{U}$; $\lambda_u \geq 0$ and $h_u(A_i) = y_i = 1$ imply $0 < \lambda_u/K \leq \mu_u = 1$. The proof for $(i, i') \in \Omega^- \times \Omega^+$ is similar. □

In the following, we will denote by $\mathcal{A}$ some subset of pairs in Φ. Now, we let

$$\mathcal{R}(\mathcal{A}) = \left\{ (\xi, \mu, \lambda) \in \mathbb{R}^M_+ \times \mathbb{R}^U_+ \times \mathbb{R}^U_+ \;\middle|\; \xi_i + \xi_{i'} + \textstyle\sum_{u \in S_{i,i'}} \mu_u \geq 1,\ \forall (i, i') \in \mathcal{A} \right\}$$

denote the polyhedron implied by the cutting planes (14) corresponding to the pairs of observations in $\mathcal{A}$. As a direct consequence of Theorem 5.1,

$$Q_{H,y} \cap \{0,1\}^M \times \{0,1\}^U \subseteq Q_{H,y} \cap \mathcal{R}(\Phi) \subseteq Q_{H,y} \cap \mathcal{R}(\mathcal{A}) \subseteq Q_{H,y}.$$

We now consider the tightened relaxation which corresponds to (13), augmented by all possible cutting planes of the form (14):

$$\min \left\{ \sum_{i=1}^{M} \xi_i + C \sum_{u=1}^{U} \mu_u \mid (\xi, \mu, \lambda) \in Q_{H,y} \cap \mathcal{R}(\Phi) \right\}. \tag{15}$$

Let us denote the optimal objective value of the strengthened relaxation (15) by $z_{\mathrm{SR}}(H, y)$.

In typical learning applications of SWVC, one has $U \gg M$ and the complexity penalty satisfies $C \geq 1$. For $C \geq 1$, it is easy to show that the ratio of the optimal integral objective value and strengthened relaxation objective value is at most a multiplicative factor of M:

Theorem 5.2. *If $C \geq 1$ then $z(H, y)/z_{\mathrm{SR}}(H, y) \leq M$.*

Proof. Consider an optimal solution of the relaxation (15) (ξ, μ, λ). Now, if $C \geq 1$ the cuts (14) assure that

$$z_{SR}(H, y) = \sum_{i=1}^{M} \xi_i + C \sum_{u \in \mathcal{U}} \mu_u \geq \xi_i + \xi_i' + \sum_{u \in S_{ii'}} \mu_u \geq 1,$$

for some (i, i'). The optimal integral solution value, on the other hand, must satisfy $z(H, y) \leq M$. □

Clearly, the strengthened LP relaxation of (15) solves a tighter relaxation of the SWVC problem than the straightforward soft margin LP relaxations (13) and (4). Before we conclude, a few remarks ought to be made about the application of the valid inequalities (14) which we have proposed. First, since the number of cuts (14) is at most $|\Omega^+||\Omega^-|$, following the polynomial time solvability of LP, the solution of (15) can be found in time polynomial in the input. Although $|\Phi| \in O(M^2)$, significant computational improvements may be possible in practice by selecting a small subset of promising inequalities corresponding to $\mathcal{A} \subset \Phi$ and optimizing over all $(\xi, \mu, \lambda) \in Q_{H,y} \cap \mathcal{R}(\mathcal{A})$; see (Goldberg & Eckstein, 2009). On the other hand, we also note that in order to further tighten the relaxation, following work done on characterization of the set cover polytope (Balas & Ng, 1989; Cornuéjols & Sassano, 1989), one could further strengthen the inequalities (14) by introducing certain inequalities derived from triples of observation pairs. Finally, in (Goldberg & Eckstein, 2009), we suggested a more numerically stable variant of the formulation in (15), in which we fix $\|\lambda\|_1 = 1$ and the margin is fixed to equal a small parameter. We used this formulation to evaluate the practical effectiveness of inequalities (14) within a boosting algorithm. In empirical tests this formulation provided very competitive classification results while maintaining the sparsity of the weight vectors λ.

6 Conclusions and open problems

We have considered sparse weighted voting classifier selection from the point of view of combinatorial optimization and LP relaxations. When formulating the hard problem as a MIP we require a large constant for the correctness of the formulation. The magnitude of this constant, on the other hand, directly affects the quality of the continuous LP relaxation. In fact we observed that we could reinterpret existing soft margin LP formulations as an LP relaxation of the MIP formulation. The initial LP relaxation, however, has a large integrality gap of $\Omega(M2^{M-1})$ which also implies that existing LP formulations provide a poor approximation of the discrete solution. We suggested a tightening of the relaxation by augmenting the relaxation with inequalities that are valid for the discrete formulation. For a common case of complexity penalties, *i.e.*, where $C \geq 1$, we are able to show that the integrality gap of our tightened relaxation is at most M.

In our analysis we found that a sufficient condition for the correctness of the MIP was $K \geq M^{M/2}$. On the other hand we showed the necessary condition (or lower bound) $K \geq 2^{M-1}$. It would be interesting to attempt to "tighten" this gap in the general case. Also it would be interesting to investigate smaller values of K that are sufficient for the correctness of the MIP formulation in special cases of matrices H corresponding to particular types of base classifiers. Finally, it would be interesting to extend and apply our cutting planes to kernel methods and SVMs.

References

Amaldi, E., & Kann, V. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, *209*, 237–260.

Arora, S., Babai, L., Stern, J., & Sweedyk, Z. (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and Systems Sciences*, *54*, 317–331.

Balas, E., & Ng, S. M. (1989). On the set covering polytope: I. all the facets with coefficients in $\{0, 1, 2\}$. *Mathematical Programming*, *43*, 57–69.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, *44*, 526–536.

Baum, E., & Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, *1*, 151–160.

Bennett, K., & Bredensteiner, E. (1997). A parametric optimization method for machine learning. *INFORMS Journal of Computing*, *9*, 311–318.

Bennett, K., & Bredensteiner, E. (2000). Duality and geometry in SVM classifers. *Proceedings of the 17th International Conference on Machine Learning*, 57–64.

Bennett, K., & Mangasarian, O. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, *1*, 23–34.

Blum, A., & Langford, J. (2003). PAC-MDL bounds. *Proceedings of the 16th Annual Conference on Computational Learning Theory* (pp. 344–357).

Brenner, J. (1972). The Hadamard maximum determinant problem. *The American Mathematical Monthly*, *79*, 626–630.

Cornuéjols, G., & Sassano, A. (1989). On the 0, 1 facets of the set covering polytope. *Mathematical Programming: Series A and B*, *43*, 44–55.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Demiriz, A., Bennett, K., & Shawe-Taylor, J. (2002). Linear programming boosting via column generation. *Machine Learning*, *46*, 225–254.

Dinur, I., & Safra, S. (2004). On the hardness of approximating label-cover. *Information Processing Letters*, *89*, 247–254.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and Systems Sciences*, *55*, 119–139.

Friedman, J. (2008). *Fast sparse regression and classification* (Technical Report). Stanford University.

Goldberg, N., & Eckstein, J. (2009). *Tightened l_0-relaxation penalties for classification* (Technical Report RRR-23). Rutgers University.

Graepel, T., Herbrich, R., Schölkopf, B., Smola, A., Bartlett, P., Müller, K.-R., Obermayer, K., & Williamson, R. (1999). Classification on proximity data with LP-machines. *International Conference of Artificial Neural Networks*, 304–309.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press.

Heath, D. (1992). *A geometric framework for machine learning*. Doctoral dissertation, Johns Hopkins University.

Höffgen, K.-U., Simon, H., & Horn, K. V. (1995). Robust trainability of single neurons. *Journal of Computer and Systems Sciences*, *50*, 114–125.

Huang, C., Cheang, G. H., & Barron, A. R. (Submitted 2008). Risk of penalized least squares, greedy selection and L1 penalization for flexible function libraries. *Annals of Statistics*.

Mangasarian, O. (1994). Misclassification minimization. *Journal of Global Optimization*, *5*, 309–323.

Mangasarian, O. (1999). Arbitrary-norm separating plane. *Operations Research Letters*, *24*, 15–23.

Pfetch, M. E. (2008). Branch-and-cut for the maximum feasible subsystem problem. *SIAM Journal on Optimization*, *19*, 21–38.

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

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

Schölkopf, B., & Smola, A. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. The MIT Press.

Shawe-Taylor, J., Bartlett, P., Williamson, R., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, *44*, 1926–1940.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley and Sons.

Vazirani, V. V. (2003). *Approximation algorithms*. Springer Verlag.

von Luxburg, U., Bousquet, O., & Schölkopf, B. (2004). A compression approach to support vector model selection. *Journal of Machine Learning Research*, *5*, 293–323.

Zhang, T. (2008). Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Neural Information Processing Systems*.

A A derivation of generalized penalty parameters for base classifiers of different classes of risk

The objective function of (8) minimizes misclassification plus a complexity penalty proportional to the number of features used. Thus, the SWVC formulation (8) corresponds to minimizing an upper bound on the total code length when the features being combined have equal complexity penalties. However, we may also assign different penalties c_u to different features $u \in \mathcal{U}$, by generalizing our formulation to include penalties that vary with u.

The single penalty parameter C of formulation (8) was motivated in Section 2 where we derived (7) as a simple bound on a code length function in terms of $\|\lambda\|_0$. Here we

suggest a tighter bound assuming the sender and receiver share the knowledge of some partition of the base classifiers $\mathcal{U}$. Specifically, we further suppose that the base classifiers are partitioned into K subsets $\mathcal{U}_k$, $k = 1, \ldots, Q$, with each $\mathcal{U}_k$ representing the classifiers that have equal complexity or "risk"; this notion is related to the concept of *structural risk minimization* (SRM) (Vapnik, 1998; Shawe-Taylor et al., 1998). The corresponding weighted voting classifiers can then be decomposed into subsets $S_j = \mathrm{conv}(\{h_u \mid u \in \bigcup_{k=0}^{j} \mathcal{U}_j\})$, for $j = 1, \ldots, K$, where $\mathrm{conv}(\mathcal{U})$ denotes the convex hull of set $\mathcal{U}$. Each of the sets $\mathcal{U}_k$ corresponds to one of K tables in a *code book* (Vapnik, 1998), present at both the sender and receiver, and an element of the k^{th} table can be identified using $\lceil \log |\mathcal{U}_k| \rceil$ bits. In order to identify an element of $\mathcal{U}$ we need to specify $\|\lambda\|_0$ such table indices, and also specify which table each element corresponds to requiring at most $\lceil \log K \rceil$ bits. We can now derive a bound on the code length function as:

$$\dot{L}[g] \leq \|\lambda\|_0 \lceil \log K \rceil + \sum_{u \in \mathcal{U}: \lambda_u \neq 0} \lceil \log \left| \mathcal{U}_{k(u)} \right| \rceil + \|\lambda\|_0 \left(\lceil \log M \rceil + 1 \right)$$