Misclassification Minimization

O. L. Mangasarian*

Revised April 6, 1994

Abstract

The problem of minimizing the number of misclassified points by a plane, attempting to separate two point sets with intersecting convex hulls in n-dimensional real space, is formulated as a linear program with equilibrium constraints (LPEC). This general LPEC can be converted to an exact penalty problem with a quadratic objective and linear constraints. A Frank-Wolfe-type algorithm is proposed for the penalty problem that terminates at a stationary point or a global solution. Novel aspects of the approach include: (i) A linear complementarity formulation of the step function that "counts" misclassifications, (ii) Exact penalty formulation without boundedness, nondegeneracy or constraint qualification assumptions, (iii) An exact solution extraction from the sequence of minimizers of the penalty function for a finite value of the penalty parameter for the general LPEC and an explicitly exact solution for the LPEC with uncoupled constraints, and (iv) A parametric quadratic programming formulation of the LPEC associated with the misclassification minimization problem.

1 Introduction

We consider the fundamental problem of machine learning of discriminating between the elements of two point sets $\mathcal{A}$ and $\mathcal{B}$ in the n-dimensional real space R^n. These sets are represented by the $m \times n$ and $k \times n$ matrices A and B respectively. When the convex hulls of $\mathcal{A}$ and $\mathcal{B}$ are disjoint, a single linear program [14, 4] or the classical iterative perceptron algorithm [20, 9] will obtain a separating plane in a finite number of steps. In the general and usual case of intersecting convex hulls, the perceptron algorithm merely obtains a bounded sequence of iterates [7]. However, the linear programming formulation [4] obtains an approximate separating plane that minimizes some norm of the **distances** of misclassified points to the approximate separating plane. Although this approach has been quite successful in important real world applications [18, 21] and in the training of neural networks [3], the approximate separating plane does not minimize the number of misclassified points, as do some machine learning approaches [19]. In neural networks [12], misclassification minimization is achieved by using the sigmoid error function $\dfrac{1}{1+e^{-\alpha x}}$, with a positive α, to approximate a step function. Hence, minimization of the sum of distances of misclassified points by a linear program is merely a surrogate for misclassification minimization. In the present work, we propose a precise mathematical programming formulation of the nonconvex problem of **minimizing the number of misclassified points.** This is done in Section 2 where we first propose a simple linear complementarity formulation of the step function (Lemma 2.1) and then use

*Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706, email: *olvi@cs.wisc.edu.* This material is based on research supported by Air Force Office of Scientific Research Grant F49620-94-1-0036 and National Science Foundation Grants CCR-9101801 and CDA-9024618.

this result to formulate the misclassification minimization as a linear program with an equilibrium (linear complementarity) constraint (Proposition 2.2). This LPEC is an important special case of mathematical programs with equilibrium constraints (MPEC) [11, 1] studied comprehensively recently in [13]. Section 3 deals with methods for solving a general LPEC. We convert our LPEC to an exact penalty problem and show (Theorem 3.2) that for a finite value of the penalty parameter an exact solution of the LPEC is contained in a minimizer of the penalty problem. Corollary 3.3 shows how to extract an exact solution from two minimizers of the penalty function. In Algorithm 3.4 we show how to solve the bilinear program, that constitutes the penalty problem, by a Frank-Wolfe type algorithm and establish its convergence in Theorem 3.5. In Section 4 we specialize to an LPEC with uncoupled equilibrium constraints (LPUEC) (32) which covers the misclassification minimization problem (12). Theorem 4.1 gives an explicitly exact penalty solution of this problem without any boundedness assumption that was required for the less general Stackelberg problem [2]. We propose Algorithm 4.2 as a finite stepless partial Frank-Wolfe algorithm for solving this problem. Its finite termination to an exact solution or a stationary point is established in Theorem 4.3. This algorithm is proposed for the solution of the parametric quadratic programming reformulation (40) of the misclassification minimization problem (12). This quadratic reformulation is proposed in order to overcome the stationarity of the penalty function for (12) at certain feasible points. Section 5 of the paper concludes with some brief remarks.

A word about our notation now. For a vector x in the n-dimensional real space R^n, x_+ will denote the vector in R^n with components $(x_+)_i := \max\{x_i, 0\}$, $i = 1, \ldots, n$. Similarly x_* will denote the vector in R^n with components $(x_*)_i := (x_i)_*$, $i = 1, \ldots, n$, where $()_*$ is the step function defined in (4) below. The notation $A \in R^{m\times n}$ will signify a real $m \times n$ matrix. For such a matrix, A^T will denote the transpose while A_i will denote row i. For two vectors x and y in R^n, xy will denote the scalar product, while $x \perp y$ will denote $xy = 0$. A vector of ones in a real space of arbitrary dimension will be denoted by e. The notation $arg \min_{x\in S} f(x)$ will denote the set of minimizers of $f(x)$ on the set S, while the notation $arg\ vertex\ partial \min_{x\in S} f(x)$ will denote the set of vertices of S that approximately minimize f on S, and in particular any vertex of S may be taken to be in this set. The empty set, as well as an empty vector will be denoted by $\emptyset$.

2 Misclassification Minimization as a Linear Program with Equilibrium Constraint (LPEC)

When the point sets $\mathcal{A}$ and $\mathcal{B}$, represented by the $m \times n$ and $k \times n$ matrices A and B respectively have disjoint convex hulls, they can be strictly separated [14] by a plane

$$wx = \gamma \tag{1}$$

in R^n where w is some weight vector constituting the normal to the separating plane and γ locates the plane relative to the origin. The plane (1) separates the sets $\mathcal{A}$ and $\mathcal{B}$ by the strict inequalities

$$Aw > e\gamma,\ e\gamma > Bw \tag{2}$$

where e is a vector of ones of appropriate dimension. The system (2), upon normalization, is equivalent to

$$Aw - e\gamma - e \geq 0,\ -Bw + e\gamma - e \geq 0 \tag{3}$$

Inequalities (3) state that the points A_i, $i = 1, \ldots, m$, lie in the open halfspace $\{x|xw > \gamma\}$ in R^n, while the points B_i, $i = 1, \ldots, k$, lie on the open halfspace $\{x|xw < \gamma\}$ in R^n. If we define the step function $()_* : R \to R$ as

$$(\xi)_* = \left\langle \begin{array}{ll} 1 & if\ \ \xi > 0 \\ 0 & if\ \ \xi \leq 0, \end{array} \right. \tag{4}$$

then the system (3) is equivalent to

$$e(-Aw + e\gamma + e)_* + e(Bw - e\gamma + e)_* = 0 \tag{5}$$

In fact, the left-hand side of (5) **counts** the number of misclassified points. For the linearly separable case of nonintersecting convex hulls, no points are misclassified, and hence the equality in (5) is obtained. In the more general case where the sets of $\mathcal{A}$ and $\mathcal{B}$ have intersecting convex hulls, and thus are linearly inseparable, equation (5) can be replaced by the minimization problem

$$\min_{w,\gamma} e(-Aw + e\gamma + e)_* + e(Bw - e\gamma + e)_* \tag{6}$$

This problem has a zero minimum if and only if the plane $xw = \gamma$ strictly separates the sets $\mathcal{A}$ and $\mathcal{B}$. Otherwise this plane minimizes the number of misclassified points, that is it minimizes

$$c(w,\gamma) := \text{cardinality} \left\{ (i,j) \,\middle|\, \begin{array}{c} A_i w - \gamma - 1 < 0, -B_j w + \gamma - 1 < 0 \\ 1 \leq i \leq m,\ 1 \leq j \leq k \end{array} \right\} \tag{7}$$

We immediately note that (6) is always solvable since there exists only a finite number of twofold partitions of $\mathcal{A} \bigcup \mathcal{B}$ that are linearly separable. Any such partition that minimizes the number of misclassified points solves (6). Our objective is to reduce (6) to a mathematical programming problem. To do that we begin by a representation of the step function (4) as a complementarity condition via the plus function $()_+$ as follows.

Lemma 2.1 ***Characterization of the step function*** $()_*$ *For* $r \in R^m,\ u \in R^m,\ a \in R^m$ *and* e*, a vector of ones in* R^m *:*

$$r = (a)_*,\ u = a_+ \Longleftrightarrow \begin{pmatrix} r \\ u \end{pmatrix} = \begin{pmatrix} r - u + a \\ r + u - e \end{pmatrix}_+ \tag{8}$$

Proof The points $(a)_*$ and $(a)_+$ solve respectively the dual linear programs

$$\max_r \{ar | 0 \leq r \leq e\} \text{ and } \min_u \{eu | u \geq a,\ u \geq 0\} \tag{9}$$

The right hand side of the equivalence (8) is merely the Karush-Kuhn-Tucker necessary and sufficient optimality conditions for r and u to solve (9), where use has been made of the elementary equivalence

$$c = d_+ \Longleftrightarrow c - d \geq 0,\ c \geq 0,\ c(c-d) = 0 \tag{10}$$

for $c \in R^m$ and $d \in R^m$. □

We now combine Lemma 2.1 and the minimization problem (6) and make use of (10), to obtain the following misclassification minimization characterization.

Proposition 2.2 ***Misclassification Minimization as a Linear Program with Equilibrium Constraints (LPEC)*** *A plane* $xw = \gamma$ *minimizes the number of misclassifications* $c(w,\gamma)$ *as* *defined by (7) if and only if* (w, γ, r, u, s, v) *solves the following linear program with equilibrium constraints:*

$$
\begin{array}{ll}
\underset{w,\gamma,r,u,s,v}{\mathit{minimize}} & er+es \\
\mathit{subject\ to} & \begin{pmatrix} r \\ u \end{pmatrix} = \begin{pmatrix} r-u-Aw+e\gamma+e \\ r+u-e \end{pmatrix}_+, \begin{pmatrix} s \\ v \end{pmatrix} = \begin{pmatrix} s-v+Bw-e\gamma+e \\ s+v-e \end{pmatrix}_+
\end{array}
\tag{11}
$$

or equivalently

$$
\begin{array}{lrr}
\underset{w,\gamma,r,u,s,v}{\mathit{minimize}} & er+es & \\
 & u+Aw-e\gamma-e \geq 0 & v-Bw+e\gamma-e \geq 0 \\
 & r \geq 0 & s \geq 0 \\
\mathit{subject\ to} & r(u+Aw-e\gamma-e)=0 & s(v-Bw+e\gamma-e)=0 \\
 & -r+e \geq 0 & -s+e \geq 0 \\
 & u \geq 0 & v \geq 0 \\
 & u(-r+e)=0 & v(-s+e)=0
\end{array}
\tag{12}
$$

We note that problem (12) is a linear program with equilibrium (linear complementarity) constraints (LPEC) and is a special case of the more general mathematical program with equilibrium constraints (MPEC) studied in detail by Luo, Pang, Ralph and Wu [13]. Being linear, our problem is endowed with some features not possessed by the more general MPECs, principally exactness of a penalty formulation without boundedness of the feasible region and without assuming nondegeneracy. We discuss these properties and an algorithm for solving (12) in the next section.

3 Linear Programs with Equilibrium Constraints (LPEC)

We consider the general LPEC

$$
\begin{array}{lr}
\underset{x,y}{\mathit{minimize}} & cx+dy \\
 & Mx+Ny+q \geq 0 \\
\text{subject to} & x(Mx+Ny+q)=0 \\
 & x,\ y \geq 0
\end{array}
\tag{13}
$$

where $M \in R^{\ell_1 \times \ell_1}$ and $N \in R^{\ell_1 \times \ell_2}$. Our misclassification minimization problem (12) is exactly of this type if we make the identification

$$
\begin{array}{c}
x = \begin{pmatrix} r \\ u \\ s \\ v \end{pmatrix},\ y = \begin{pmatrix} t \\ \delta \\ \xi \end{pmatrix},\ \begin{pmatrix} w \\ \gamma \end{pmatrix} = \begin{pmatrix} t-e\xi \\ \delta-\xi \end{pmatrix},\ M = \begin{pmatrix} 0 & I & 0 & 0 \\ -I & 0 & 0 & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & -I & 0 \end{pmatrix}, \\
N = \begin{pmatrix} A & -e & -Ae+e \\ 0 & 0 & 0 \\ -B & e & Be-e \\ 0 & 0 & 0 \end{pmatrix},\ c = \begin{pmatrix} e \\ 0 \\ e \\ 0 \end{pmatrix},\ d = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},\ q = \begin{pmatrix} -e \\ e \\ -e \\ e \end{pmatrix}
\end{array}
\tag{14}
$$

We note that in (14), $M + M^T = 0$. Hence M is a skew symmetric positive semidefinite matrix, reflecting its linear programming origin (9). We also note that the linear objective function of (13) is bounded below by zero on the nonempty feasible region when the identifications (14) are made. This motivates the following simple existence results for (13).

Theorem 3.1 ***Existence of Solution for LPEC*** *A solution to (13) exists if it is feasible and its objective function is bounded below on its nonempty feasible region.*

Proof The nonempty feasible region of (13) consists of the union of a finite number of polyhedral sets in $R^{\ell_1+\ell_2}$. Since $cx+dy$ is bounded below on each of these polyhedral sets it attains its minimum on each of them and its global minimum on their union. $\square$

We now convert the LPEC (13) into a penalty problem of minimizing a quadratic function on a polyhedral set and show that for sufficiently large but finite value of the penalty parameter an exact solution can be identified. For that purpose it is convenient to define the following constraint sets:

$$\begin{aligned} S &:= \{(x,y) \mid Mx+Ny+q \geq 0,\ (x,y) \geq 0,\ x(Mx+Ny+q)=0\} \\ S_0 &:= \{(x,y) \mid Mx+Ny+q \geq 0,\ (x,y) \geq 0\} \end{aligned} \tag{15}$$

We also define the following penalty function

$$P((x,y),\alpha) := cx+dy+\alpha x(Mx+Ny+q),\ \alpha \geq 0 \tag{16}$$

We will now show that for $\alpha \geq \bar{\alpha}$, for some $\bar{\alpha} > 0$, minimizing $P((x,y),\alpha)$ over S_0 will identify an exact solution of the LPEC (13). We state this result as follows.

Theorem 3.2 ***Exactness of LPEC Penalization*** *Let the feasible region S be nonempty and let $cx+dy$ be bounded below on S_0 of (15). There exists $\bar{\alpha} > 0$ such that for any fixed $\alpha \geq \bar{\alpha}$*

$$\begin{bmatrix} x(\alpha) \\ \\ y(\alpha) \end{bmatrix} = \begin{bmatrix} \frac{a^i}{\alpha} + x_0^i \\ \\ \frac{b^i}{\alpha} + y_0^i \end{bmatrix},\ \textit{for some } i \in \{1,\ldots,\ell\}, \tag{17}$$

where

$$(x(\alpha),\ y(\alpha)) \in \arg \min_{(x,y)\in S_0} P((x,y),\alpha) \tag{18}$$

Here ℓ and the vectors a^i, b^i, x_0^i, y_0^i, $i = 1,\ldots,\ell$, depend on the LPEC problem data only: M, N, C, d and q. Furthermore $(x_0^i,\ y_0^i)$, $i = 1,2,\ldots,\ell$, solve the LPEC (13) and such that for any fixed $\alpha \geq \bar{\alpha}$

$$x(\alpha)(Mx(\alpha)+Ny(\alpha)+q) = \frac{a^i(Ma^i+Nb^i)}{\alpha^2},\ \textit{for some } i \in \{1,\ldots,\ell\} \tag{19}$$

Proof First note that $cx+dy$ is bounded below on S_0 and $x(Mx+Ny+q) \geq 0$ on S_0. It follows that the penalty problem of (18) is solvable by some $(x(\alpha),\ y(\alpha))$ for each $\alpha \geq 0$. Hence for $\alpha > 0$, $x(\alpha)$, $y(\alpha)$ and some $u(\alpha) \in R^{\ell_1}$ satisfy the following Karush-Kuhn-Tucker conditions for (18)

$$0 \leq \begin{pmatrix} x \\ y \\ u \end{pmatrix} \perp \begin{pmatrix} (M+M^T) & N & -M^T \\ N^T & 0 & -N^T \\ M & N & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ u \end{pmatrix} + \begin{pmatrix} \frac{c}{\alpha}+q \\ \frac{d}{\alpha} \\ q \end{pmatrix} \geq 0 \tag{20}$$

The linear complementarity problem (20) has a vertex solution [16, Lemma 2], and hence a basic solution, for each $\alpha > 0$. Since (20) has a finite number of bases, it follows that on the set $\{\alpha \mid \alpha \geq \bar{\alpha}\}$, for some $\bar{\alpha} > 0$, only a finite number of basic solutions of (20) appear infinitely often, and no other basic solution appears a finite or infinite number of times. Let the $(2\ell_1 + \ell_2) \times (2\ell_1 + \ell_2)$ matrices

$$B^1,\ B^2, \ldots,\ B^\ell, \tag{21}$$

correspond to these basic solutions and such that

$$\begin{bmatrix} x(\alpha) \\ y(\alpha) \\ u(\alpha) \end{bmatrix} = B^i \begin{bmatrix} \frac{c}{\alpha} + q \\ \frac{d}{\alpha} \\ q \end{bmatrix},\ \text{for some } i \in \{1, \ldots, \ell\},\ \alpha \geq \bar{\alpha} \tag{22}$$

It follows from Theorem 2.8 of [17] by letting $\alpha \to \infty$, that

$$\begin{bmatrix} x_0^i \\ y_0^i \end{bmatrix} := \begin{bmatrix} x(\infty) \\ y(\infty) \end{bmatrix} := \begin{bmatrix} B_1^i \\ B_2^i \end{bmatrix} \begin{bmatrix} q \\ 0 \\ q \end{bmatrix},\ \text{for some } i \in \{1, \ldots, \ell\}, \tag{23}$$

solve LPEC (13). This establishes (17) with $\begin{bmatrix} a^i \\ b^i \end{bmatrix} = \begin{bmatrix} B_1^i \\ B_2^i \end{bmatrix} \begin{bmatrix} c \\ d \\ 0 \end{bmatrix}$ and $\begin{bmatrix} x_0^i \\ y_0^i \end{bmatrix}$ as given by (23). To establish (19) we employ Theorem 2.8 [17] again which states that

$$\lim_{\alpha \to \infty} \alpha x(\alpha)(Mx(\alpha) + Ny(\alpha) + q) = 0 \tag{24}$$

Since

$$x(\alpha)(Mx(\alpha) + Ny(\alpha) + q) = \frac{a^i g^i}{\alpha^2} + \frac{a^i h^i + x_0^i g^i}{\alpha} + x_0^i h^i,\quad i \in \{1, \ldots, \ell\} \tag{25}$$

where

$$g^i := Ma^i + Nb^i,\ h^i := Mx_0^i + Ny_0^i + q,\quad i \in \{1, \ldots, \ell\}$$

It follows from (24) and (25) that

$$x_0^i h^i = 0 \text{ and } a^i h^i + x_0^i g^i = 0,\quad i \in \{1, \ldots, \ell\} \tag{26}$$

Combining (25) and (26) gives (19). $\square$

Equation (19) shows that the complementarity residual decreases quadratically with increasing penalty parameter values. With the help of Theorem 3.2, an exact solution of LPEC (13) can be recovered from (23) as follows.

Corollary 3.3 *Let $\alpha_2 > \alpha_1 \geq \bar{\alpha} > 0$ be such that the corresponding solutions $(x(\alpha_2),\ y(\alpha_2),\ u(\alpha_2))$ and $(x(\alpha_1),\ y(\alpha_1),\ u(\alpha_1))$ of (22) have the same basis. Then $(x_0^i,\ y_0^i)$ solves LPEC (13) where*

$$x_0^i = \frac{\alpha_2 x(\alpha_2) - \alpha_1 x(\alpha_1)}{\alpha_2 - \alpha_1} \tag{27}$$

$$y_0^i = \frac{\alpha_2 y(\alpha_2) - \alpha_1 y(\alpha_1)}{\alpha_2 - \alpha_1} \tag{28}$$

Proof From (17), since α_2 and α_1 generate solutions with the same basis it follows that

$$x(\alpha_1) = \frac{a^i}{\alpha_1} + x_0^i \tag{29}$$

$$x(\alpha_2) = \frac{a^i}{\alpha_2} + x_0^i \tag{30}$$

Subtracting (29) from (30) and solving a^i we get

$$a^i = \alpha_1\alpha_2 \frac{x(\alpha_1) - x(\alpha_2)}{\alpha_2 - \alpha_1}$$

Substituting for a^i in (29) gives (27). The expression (28) for y_0 is similarly obtained. □

Corollary 3.3 is useful in obtaining an exact solution of LPEC (13) by monitoring repeated bases for $\alpha \geq \bar{\alpha}$, and using (27)-(28) to get the desired solution. Therefore, it remains to prescribe an algorithm for solving the penalty problem (18) for α sufficiently large. We propose a Frank-Wolfe algorithm [10, 6] for solving (18) similar to that employed in [5]. This approach was quite successful in solving the nonconvex bilinear separability problem on many test cases [5]. For completeness we give the algorithm here. For convenience, however, we first define $P(z, \alpha)$ as the penalty function of (16), that is

$$P(z, \alpha) := cx + dy + \alpha x(Mx + Ny + q), \; z = \begin{pmatrix} x \\ y \end{pmatrix} \tag{31}$$

Algorithm 3.4 ***Frank-Wolfe Algorithm for Solving (18)*** *Fix* $\alpha > 0$. *Start with any* $z^0 \in S^0$. *Determine* z^{j+1} *from* z^j *as follows.*

- $\bar{z}^j \in \arg\ \mathit{vertex} \min_{z \in S^0} \nabla_z P(z^j, \alpha) z$
- *Stop if* $\nabla_z P(z^j, \alpha)(\bar{z}^j - z^j) = 0$
- $z^{j+1} = (1 - \lambda^j) z^j + \lambda^j \bar{z}^i$, *where* $\lambda^j \in \arg \min_{0 \leq \lambda \leq 1} P(((1-\lambda) z^j + \lambda \bar{z}^j), \alpha)$

Convergence of this algorithm is established in [5] without any convexity assumption on $P(z, \alpha)$. We state this convergence result without proof here.

Theorem 3.5 ***Convergence of Frank-Wolfe Algorithm*** *Algorithm 3.4 terminates at some* z^j *that satisfies the minimum principle necessary optimality condition:* $\nabla_z P(z^j, \alpha)(z - z^j) \geq 0$ *for all* $z \in S^0$, *or each accumulation point* $\bar{z}$ *of the sequence* $\{z^j\}$ *satisfies the minimum principle.*

4 Linear Programs with Uncoupled Equilibrium Constraints (LPUEC)

In this section we specialize the general LPEC (13) to the case of uncoupled equilibrium constraints, that is

$$\begin{array}{ll} \underset{x,y}{\mathit{minimize}} & c_1x_1 + c_2x_2 + d_1y_1 + d_2y_2 \\ & M_{12}x_2 + N_{11}y_1 + q_1 \geq 0 \\ & M_{21}x_1 + N_{22}y_2 + q_2 \geq 0 \\ \text{subject to} & x_1(M_{12}x_2 + N_{11}y_1 + q_1) = 0 \\ & x_2(M_{21}x_1 + N_{22}y_2 + q_2) = 0 \\ & x_1, \; x_2, \; y_1, \; y_2 \geq 0 \end{array} \tag{32}$$

We note immediately that (32) includes our misclassification minimization problem (12) if we rewrite the latter as follows

$$
\begin{array}{ll}
\underset{r,s,u,v,t,\delta,\xi}{minimize} & er + es \\
\text{subject to} & 0 \leq \begin{pmatrix} r \\ s \\ u \\ v \end{pmatrix} \perp \left(\begin{pmatrix} 0 & 0 & I & 0 & A & -e & -Ae+e \\ 0 & 0 & 0 & I & -B & e & Be-e \\ -I & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -I & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} r \\ s \\ u \\ v \\ t \\ \delta \\ \xi \end{pmatrix} + \begin{pmatrix} -e \\ -e \\ e \\ e \end{pmatrix} \right) \geq 0 \\
& (t, \delta, \xi) \geq 0
\end{array}
\tag{33}
$$

Here we have used the substitution

$$
\begin{pmatrix} w \\ \gamma \end{pmatrix} = \begin{pmatrix} t - e\xi \\ \delta - \xi \end{pmatrix}, \ (t, \delta, \xi) \geq 0
$$

as was done in (14). We take in (32)

$$
x_1 = \begin{pmatrix} r \\ s \end{pmatrix}, \ x_2 = \begin{pmatrix} u \\ v \end{pmatrix}, \ y_1 = \begin{pmatrix} t \\ \delta \\ \xi \end{pmatrix}, \ y_2 = (\emptyset)
$$

We note that consequences of the uncoupled LPUEC (32) include the following:

(i) An exact penalty formulation which is explicit. That is, an exact solution is obtained for any sufficiently large value of the penalty parameter α without regard to a repeated basis, as was the case in Corollary 3.3.

(ii) A stepless partial Frank-Wolfe Algorithm 4.2 will terminate in a finite number of steps at a stationary point of the penalty problem or at a global solution of (32). This is an improvement over the stepless full Frank-Wolfe algorithm proposed for the linear Stackelberg problem, a special case of (32), with an exact penalty also, but where each linear program was solved completely [2]. Our experience with partial solution of the linear programming subproblems for the closely related bilinear programming problem [5] leads us to believe that partial solution of the linear programming subproblems is preferable especially when one is far from a solution point or stationary point, and the linear subproblems are merely crude surrogates for the nonlinear problem.

(iii) No boundedness of the feasible region is needed for exactness of the penalty function, as was required in [2], for the linear Stackelberg problem and in [13] for the MPEC, nor a constraint qualification as needed in [13].

We now formulate the penalty problem (18) associated with (32) and note its key distinguishing feature that its feasible region S_0 consists of uncoupled constraint regions S_{01} and S_{02}. In particular, we have

$$
(x(\alpha), y(\alpha)) \in \arg \min_{(x,y)\in S_0} P((x,y),\alpha) \tag{34}
$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},\ y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},\ c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},\ d = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix},\ q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix},$$
$$M = \begin{pmatrix} 0 & M_{12} \\ M_{21} & 0 \end{pmatrix},\ N = \begin{pmatrix} N_{11} & 0 \\ 0 & N_{22} \end{pmatrix} \tag{35}$$

$$P((x,y),\alpha) := cx + dy + \alpha x(Mx + Ny + q) \tag{36}$$

$$S := \{(x,y) | Mx + Ny + q \geq 0,\ (x,y) \geq 0,\ x(Mx + Ny + q) = 0\} \tag{37}$$

$$\begin{aligned} S_0 &:= S_{01} \times S_{02} \\ S_{01} &:= \{(x_2,\, y_1) | M_{12}x_2 + N_{11}y_1 + q_1 \geq 0,\ (x_2,\, y_1) \geq 0\} \\ S_{02} &:= \{(x_1,\, y_2) | M_{21}x_1 + N_{22}y_2 + q_2 \geq 0,\ (x_1,\, y_2) \geq 0\} \end{aligned} \tag{38}$$

We show first that as a consequence of the fact that $S_0 = S_{01} \times S_{02}$, the penalty problem (34) always has a vertex solution and that for sufficiently large α, an exact solution of LPUEC (32) is obtained.

Theorem 4.1 ***Explicit Exactness of LPUEC Penalization*** *Let LPUEC (32) have a nonempty feasible region S, and let its objective function be bounded below on S_0. Then*

(i) LPUEC (32) has a solution which is a vertex of S_0.

(ii) For $\alpha \geq \bar{\alpha}$ for some $\bar{\alpha} > 0$, the penalty function $P(\cdot,\alpha)$ attains a minimum on S_0 at $(\bar{x},\bar{y})$ which is a vertex of S_0, and every such vertex is a solution of LPUEC (32).

Proof Note that (ii) implies (i) and hence we need only establish the latter.

We establish first that $\min\limits_{(x,y)\in S_0} P((x,y),\alpha)$ has a vertex solution for each $\alpha \geq 0$. Since $cx + dy$ is bounded below on S_0 by assumption, and $x(Mx + Ny + q) \geq 0$ on S_0, it follows that the quadratic penalty function $P((x,y),\alpha)$ has a minimum solution $(\bar{x}(\alpha),\, \bar{y}(\alpha))$ in the polyhedral set S_0 for $\alpha \geq 0$ [10]. Hence the linear program

$$\min_{(x_2,y_1)\in S_{01}} P(((\bar{x}_1(\alpha),x_2),\ (y_1,\bar{y}_2(\alpha))),\alpha)$$

has a vertex $(x_2(\alpha), y_1(\alpha))$ of S_{01} as solution and such that

$$P(((\bar{x}_1(\alpha),x_2(\alpha)),\ (y_1(\alpha),\bar{y}_2(\alpha))),\alpha) = P((\bar{x}(\alpha),\bar{y}(\alpha)),\alpha)$$

Similarly, the linear program

$$\min_{(x_1,y_2)\in S_{02}} P(((x_1,x_2(\alpha)),\ (y_1(\alpha),y_2)),\alpha)$$

has a vertex $(x_1(\alpha), y_2(\alpha))$ of S_{02} as solution and such that

$$P((x(\alpha),y(\alpha)),\alpha) = P((\bar{x}_1(\alpha),x_2(\alpha),y_1(\alpha),\bar{y}_2(\alpha)),\alpha) = P((\bar{x}(\alpha),\bar{y}(\alpha)),\alpha)$$

Hence $((x_2(\alpha), y_1(\alpha)),\ (x_1(\alpha), y_2(\alpha)))$ is a vertex of $S_{01} \times S_{02}$ and consequently a vertex solution of $\min\limits_{(x,y)\in S_0} P((x,y),\alpha)$.

We now establish exactness of the penalty problem $\min\limits_{(x,y)\in S_0} P((x,y),\alpha)$. Since S_0 has a finite number of vertices, one vertex, say $(\bar{x}, \bar{y})$ will satisfy

$$(\bar{x},\bar{y}) \in \arg\text{vertex} \min_{(x,y)\in S_0} P((x,y),\alpha) \qquad \alpha \geq \bar{\alpha} > 0$$

for some $\bar{\alpha} > 0$. By Theorem 2.5 [17] we have that

$$\bar{x}(M\bar{x} + N\bar{y} + q) = 0$$

Consequently, for all $(x,y) \in S$ and $\alpha \geq \bar{\alpha}$

$$c\bar{x} + d\bar{y} = P((\bar{x},\bar{y}),\alpha) \leq P((x,y),\alpha) = cx + dy$$

Hence $(\bar{x}, \bar{y})$ solves LPUEC (32). $\square$

We now give a finite algorithm for solving LPUEC (32) that terminates either at a global solution or a point satisfying the minimum principle necessary optimality condition [15] for $\min\limits_{(x,y)\in S_0} P((x,y),\alpha)$. The algorithm consists of partially solving a succession of linear programs, and as such can be considered as a stepless Frank-Wolfe algorithm [10]. The ideas are similar to those of Algorithm 2.1 [5] for separable bilinear programs.

Algorithm 4.2 ***Stepless Partial Frank-Wolfe for LPUEC (32)*** *For $\alpha > 0$ start with $(x^0, y^0) \in S$. Determine (x^{i+1}, y^{i+1}) from (x^i, y^i) as follows:*

$$(x_2^{i+1}, y_1^{i+1}) \in \arg\textit{vertex partial} \min_{(x_2,y_1)\in S_{01}} P(((x_1^i, x_2), (y_1, y_2^i)), \alpha)$$

$$(x_1^{i+1}, y_2^{i+1}) \in \arg\textit{vertex partial} \min_{(x_1,y_2)\in S_{02}} P(((x_1, x_2^{i+1}), (y_1^{i+1}, y_2)), \alpha)$$

and *such that $P((x^{i+1}, y^{i+1}),\alpha) < P((x^i, y^i),\alpha)$. Stop when impossible.*

In the above, "*arg vertex partial* min" denotes the set of vertices of the respective feasible regions that give a value of the objective that is no greater than $P(((x_1^i, x_2^i), (y_1^i, y_2^i)), \alpha)$ and $P(((x_1^i, x_2^{i+1}), (y_1^{i+1}, y_2^i)), \alpha)$, respectively. We establish now finite termination of Algorithm 4.2

Theorem 4.3 ***Finite Termination of Algorithm 4.2*** *Let LPUEC (32) have a nonempty feasible region S, and let its objective function be bounded below on S_0. For $\alpha \geq \bar{\alpha}$ for some $\bar{\alpha} > 0$, Algorithm 4.2 terminates in a finite number of steps at a solution of LPUEC (32) or at a point $(x_1^i, x_2^{i+1}, y_1^{i+1}, y_2^i)$ that satisfies the minimum principle necessary optimality condition [15] for $\min\limits_{(x,y)\in S_0} P((x,y),\alpha)$:*

$$P(((x_1^i, x_2),\ (y_1, y_2^i)), \alpha) \geq P(((x_1^i, x_2^{i+1}),\ (y_1^{i+1}, y_2^i)), \alpha) \leq P(((x_1, x_2^{i+1}),\ (y_1^{i+1}, y_2,)), \alpha), \forall (x,y) \in S_0 \tag{39}$$

Proof If for some i, $P((x^{i+1}, y^{i+1}),\alpha) \not< P((x^i, y^i),\alpha)$, then each of the two linear programs of Algorithm 4.2 must have been solved to optimality, and $P(((x_1^i, x_2),\ (y_1, y_2^i)), \alpha) \geq P((x^i, y^i), \alpha) = P(((x_1^i, x_2^{i+1}),\ (y_1^{i+1}, y_2^i)), \alpha) = P((x^{i+1}, y^{i+1}), \alpha) \leq P(((x_1, x_2^{i+1}), (y_1^{i+1}, y_2)), \alpha), \forall (x,y) \in S_0$. From this the minimum principle (39) follows. Since there are a finite number of vertices of $S_0 = S_{01} \times S_{02}$, and since for each vertex visited by Algorithm 4.2 the penalty function $P((x,y),\alpha)$, strictly decreases, no vertex of S_0 is repeated. Thus, Algorithm 4.2 terminates at $(x_1^i, x_2^{i+1}, y_1^{i+1}, y_2^i)$ which is either a global minimum of $P((x,y),\alpha)$ on S_0 or $(x_1^i, x_2^{i+1}, y_1^{i+1}, y_2^i)$ satisfies the minimum principle (39). In the former case, it follows by Theorem 4.1 that $(x_1^i, x_2^{i+1}, y_1^{i+1}, y_2^i)$ solves LPUEC (32). $\square$

We return now to the misclassification minimization problem (12). Curiously, it turns out that the penalty function (36) for (12), has a stationary point on S_0 for each arbitrary (w,γ) defining the plane (1) and appropriately chosen (r,u,s,v). We skip the algebra that shows this, but give an intuitive justification of this curious fact. This may also explain why optimization algorithms, including the classical backpropagation algorithm of neural networks [12], may be slow when applied directly to the misclassification error function of (6) or the equivalent formulation (12), instead of the parametric quadratic minimization (40) proposed below. Suppose that the plane $wx = \gamma$ does not pass through any of the points of either set $\mathcal{A}$ or $\mathcal{B}$. A small perturbation in either w or γ or both will not change the number of misclassified points. Hence for such (w,γ) and its perturbation, the constraints of (12), which merely generate r and s that count misclassifications, remain satisfied by the perturbed (w,γ) and some corresponding (r,u,s,v), but the objective function $er + es$ remains constant. Thus (w,γ) is stationary. To avoid this difficulty we consider the following parametric reformulation of problem (12)

$$\begin{array}{lll}
\underset{w,\gamma,r,u,s,v}{minimize} & \multicolumn{2}{l}{\frac{1}{m}[r(Aw - e\gamma - e) + eu] + \frac{1}{k}[s(-Bw + e\gamma - e) + ev]} \\
 & u + Aw - e\gamma - e \geq 0 & v - Bw + e\gamma - e \geq 0 \\
 & r \geq 0 & s \geq 0 \\
\text{subject to} & -r + e \geq 0 & -s + e \geq 0 \\
 & u \geq 0 & v \geq 0 \\
 & & er + es \leq \delta \\
 & & \delta \in [0, \infty)
\end{array} \qquad (40)$$

Here the weights $\frac{1}{m}$ and $\frac{1}{k}$ are used to average the complementarity condition over the cardinalities m and k of the sets $\mathcal{A}$ and $\mathcal{B}$ respectively. This is motivated by the fact that when $\delta = 0$, problem (40) reduces precisely to problem (2.11) of [4] which is guaranteed to generate a non-null w that will minimize the average of the distances of misclassified points (assuming $\| A_i \|_2 = 1,\ \| B_j \|_2 = 1,\ i = 1,\ldots,m,\ j = 1,\ldots,k$). Note that problem (40) is solvable for each $\delta \in [0,\infty)$, because its quadratic objective is bounded below on its nonempty feasible region [10]. Furthermore $f(\delta)$, the nonnegative minimum value of (40), is a nonincreasing function of δ that decreases to zero at $\delta = \bar{\delta}$, for some $\bar{\delta} \geq 0$, which is the global minimum of problem (12). This is so, because the objective function of (40), which consists of the complementarity constraints of (12), becomes zero first at some $\bar{\delta}$, the smallest value of δ, for which the constraints of (12) are satisfied and $er + es$ is minimized, thus giving the smallest number of misclassified points. Hence problem (12) is reduced to solving problem (40) for $\delta = \bar{\delta}$, the smallest nonnegative root of $f(\delta) = 0$, that is

$$\bar{\delta} = \min_{\delta \geq 0} \ \{\delta \mid f(\delta) = 0\} \qquad (41)$$

Obviously if $\bar{\delta} = 0$, then the sets $\mathcal{A}$ and $\mathcal{B}$ are linearly separable and no points are misclassified. Our proposed procedure is to solve (40) by Algorithm 4.2 for increasing values of δ until $f(\delta) = 0$. A one dimensional secant method [8] for finding a nonnegative root of $f(\delta) = 0$ may be helpful. We note that $f(\delta) = 0$ for $\delta \geq \bar{\delta}$.

5 Conclusion

We have formulated the problem of minimizing the number of misclassified points by a plane attempting to separate two point sets in R^n as a problem of minimizing of a nonconvex quadratic function subject to linear inequalities. The quadratic program is either a penalty problem with a finite value of the penalty parameter obtained from a linear program with equilibrium constraints (LPEC), or a parametric nonconvex quadratic minimization problem with linear constraints. Our exact penalty formulation of the general LPEC, both with coupled and uncoupled constraints, require no assumptions on the problem other than feasibility and boundedness from below on the objective function. These assumptions are automatically satisfied by the LPEC associated with the misclassification minimization.

When the constraints of the LPEC are coupled, our exact penalty formulation (16) requires some calculation to obtain an exact solution to the original LPEC. However, when the constraints are uncoupled as in (32), which is the case for the misclassification minimization problem (12), our exact penalty (16) directly yields an explicit solution to the LPEC for a finite value of the penalty parameter. The method that we propose for solving the misclassification minimization problem is the linear-programming-based Algorithm 4.2, applied to the parametric quadratic program (40) as a bilinear program. We think that this effective approach for solving the misclassification minimization problem is worthy of further study and numerical testing.

Acknowledgement

I am grateful to my Ph.D. student Chunhui Chen for reading this paper and making valuable suggestions.

References

[1] G. Anandalingam and T.L. Friesz. Hierarchical optimization: An introduction. *Annals of Operations Research*, 34:1–11, 1992.

[2] G. Anandalingam and D.J. White. A solution method for the linear static stackelberg problem using penalty functions. *IEEE Transactions on Automatic Control*, 35(10):1170–1173, 1990.

[3] K.P. Bennett and O.L. Mangasarian. Neural network training via linear programming. In P. M. Pardalos, editor, *Advances in Optimization and Parallel Computing*, pages 56–67, Amsterdam, 1992. North Holland.

[4] K.P. Bennett and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[5] K.P. Bennett and O.L. Mangasarian. Bilinear separation of two sets in n-space. *Computational Optimization & Applications*, 2:207–227, 1993.

[6] C. Berge and A. Ghouila-Houri. *Programming, Games and Transportation Networks*. Wiley, New York, 1965.

[7] H.D. Block and S.A. Levin. On the boundedness of an iterative procedure for solving a system of linear inequalities. *Proceedings of the American Mathematical Society*, 26:229–235, 1970.

[8] J.E. Dennis and R.B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, N.J., 1983.

[9] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[10] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[11] P.T. Harker and J.-S. Pang. Existence of optimal solutions to mathematical programs with equilibrium constraints. *Operations Research Letters*, 7:61–64, 1988.

[12] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, California, 1991.

[13] Z.-Q. Luo, J.-S. Pang, D. Ralph, and S.-Q. Wu. Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints. Technical Report 275, Communications Research Laboratory, McMaster University, Hamilton, Ontario, Hamilton, Ontario L8S 4K1, Canada, 1993.

[14] O.L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

[15] O.L. Mangasarian. *Nonlinear Programming*. McGraw–Hill, New York, 1969.

[16] O.L. Mangasarian. Characterization of linear complementarity problems as linear programs. *Mathematical Programming Study*, 7:74–87, 1978.

[17] O.L. Mangasarian. Some applications of penalty functions in mathematical programming. In R. Conti, E. De Giorgi, and F. Giannessi, editors, *Optimization and Related Fields*, pages 307–329. Springer–Verlag, Heidelberg, 1986. Lecture Notes in Mathematics 1190.

[18] O.L. Mangasarian, R. Setiono, and W.H. Wolberg. Pattern recognition via linear programming: Theory and application to medical diagnosis. In T. F. Coleman and Y. Li, editors, *Large-Scale Numerical Optimization*, pages 22–31, Philadelphia, Pennsylvania, 1990. SIAM. Proceedings of the Workshop on Large-Scale Numerical Optimization, Cornell University, Ithaca, New York, October 19-20, 1989.

[19] S. Murthy, S. Kasif, S. Salzberg, and R. Beigel. OC1: Randomized induction of oblique decision trees. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 322–327, Cambridge, MA 02142, 1993. The AAAI Press/The MIT Press.

[20] F. Rosenblatt. The perceptron–a perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York, January 1957.

[21] W. H. Wolberg and O.L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences,U.S.A.*, 87:9193–9196, 1990.