To appear in *Optimization Methods and Software*

ROBUST LINEAR PROGRAMMING DISCRIMINATION OF TWO LINEARLY INSEPARABLE SETS

KRISTIN P. BENNETT and O. L. MANGASARIAN

Computer Sciences Department, University of Wisconsin,
1210 West Dayton Street, Madison, WI 53706

A single linear programming formulation is proposed which generates a plane that minimizes an average sum of misclassified points belonging to two disjoint points sets in n-dimensional real space. When the convex hulls of the two sets are also disjoint, the plane completely separates the two sets. When the convex hulls intersect, our linear program, unlike all previously proposed linear programs, is guaranteed to generate some error-minimizing plane, without the imposition of extraneous normalization constraints that inevitably fail to handle certain cases. The effectiveness of the proposed linear program has been demonstrated by successfully testing it on a number of databases. In addition, it has been used in conjunction with the multisurface method of piecewise-linear separation to train a feed-forward neural network with a single hidden layer.

KEY WORDS: Linear Programming, Pattern Recognition, Neural Networks

1 INTRODUCTION

We consider the two point-sets $\mathcal{A}$ and $\mathcal{B}$ in the n-dimensional real space R^n represented by the $m \times n$ matrix A and the $k \times n$ matrix B respectively. Our principal objective here is to formulate a single linear program with the following properties:

(i) If the convex hulls of $\mathcal{A}$ and $\mathcal{B}$ are disjoint, a strictly separating plane is obtained.

(ii) If the convex hulls of $\mathcal{A}$ and $\mathcal{B}$ intersect, a plane is obtained that minimizes some measure of misclassification points, for **all** possible cases.

(iii) No extraneous constraints are imposed on the linear program that rule out any specific case from consideration.

Most linear programming formulations[6,5,12,4] have property (i), however, none to our knowledge have properties (ii) and (iii). For example, the linear program of Mangasarian[6] fails to satisfy property (ii) for all linearly inseparable cases, while Smith's linear program[12] fails in satisfying (ii) when uniform weights are used in its objective function as originally proposed. The linear programs[5,4] fail in satisfying both (ii) and (iii). Our linear programming formulation on the other hand has

all three properties (i), (ii) and (iii). It is interesting to note that our proposed linear program (2.11) will always generate some error-minimizing plane even in the usually troublesome case when the means of the two sets are identical. For this case, among possible solutions to our linear program is the null solution. However, this null solution is never unique for our linear program and thus a useful alternative solution is always available. For example, such an alternative, the 45° line, is obtained computationally by our linear program for the classical counterexample of linear inseparability: the Exclusive-Or example[11]. (See Example 2.7 below.)

We outline our results now. In Section 2 we state our linear program (2.11) and establish that it possesses properties (i)-(iii) above in Theorems 2.5 and 2.6. On the other hand in Example 2.8 we show that ($\bar{w} = 0,\ \bar{\gamma} = 1$) uniquely solves Smith's linear program ((2.10b) with $\delta_1 = \delta_2$) and hence property (ii) is violated. Similarly in Remark 2.9 we give an example which violates property (ii) for Grinold's linear program[5] (2.20) and give conditions under which this is always true. In Section 3 we report on some computational results using our proposed linear program on the Wisconsin Breast Cancer Database and the Cleveland Heart Disease Database. See also Bennett-Mangasarian[1] for other computational results using linear programming on these databases.

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 $(x_+)_i := \max\{x_i,\ 0\},\ i = 1, \ldots, n$. The notation $A \in R^{m \times n}$ will signify a real $m \times n$ matrix. For such a matrix, A' will denote the transpose while A_i will denote the ith row. The 1-norm of x, $\sum_{i=1}^{n} |x_i|$, will be denoted by $\|x\|_1$, while the ∞-norm of x, $\max_{1 \le i \le n} |x_i|$, will be denoted by $\|x\|_\infty$. A vector of ones in a real space of arbitrary dimension will be denoted by e.

2 A ROBUST LINEAR PROGRAMMING SEPARATION

Our linear program is based on the following error-minimizing optimization problem

$$\min_{w,\gamma} \frac{1}{m}\left\|\left(-Aw + e\gamma + e\right)_+\right\|_1 + \frac{1}{k}\left\|\left(Bw - e\gamma + e\right)_+\right\|_1 \tag{2.1}$$

where $A \in R^{m \times n}$ represents the m points of the set $\mathcal{A}$, $B \in R^{k \times n}$ represents the k points of the set $\mathcal{B}$, w is the n-dimensional "weight" vector representing the normal to the optimal "separating" plane, and the real number γ is a threshold that gives the location of the separating plane: $wx = \gamma$. The choice of the weights $\frac{1}{m}$ and $\frac{1}{k}$ in (2.1) is critical (as we shall demonstrate below in Theorems 2.5 and 2.6) in that it sets it apart from Smith's linear program[12] where equal weights were proposed, and from other linear programming formulations[6,5,4]. Our choice we believe is a "natural" one in that the useless null solution $w = 0$ is not encountered computationally for linearly inseparable sets. This is theoretically justified (Theorem 2.5

below) because $w = 0$ cannot be a solution unless the following equality between the arithmetic means of $\mathcal{A}$ and $\mathcal{B}$ holds

$$\frac{eA}{m} = \frac{eB}{k} \tag{2.2}$$

However in this case, it is guaranteed that a nonzero optimal w exists in addition to $w = 0$ (Theorem 2.6 below).

We begin our analysis by justifying the use of the optimization problem (2.1) which minimizes the average of the misclassified points of $\mathcal{A}$ and $\mathcal{B}$ by the separating plane $xw = \gamma$. We define now linear separability for concreteness.

DEFINITION 2.1. **(Linear Separability)** *The point sets $\mathcal{A}$ and $\mathcal{B}$, represented by the matrices $A \in R^{m\times n}$ and $B \in R^{k\times n}$ respectively, are linearly separable if and only if*

$$\min_{1\le i\le m} A_i v > \max_{1\le i\le k} B_i v \quad \text{for some} \quad v \in R^n \tag{2.3}$$

or equivalently

$$Aw \ge e\gamma + e, \quad e\gamma - e \ge Bw \quad \text{for some} \quad w \in R^n, \ \gamma \in R \tag{2.4}$$

That (2.4) implies (2.3) is evident. To see the converse just note the relations

$$w := 2v/\nu, \ \nu := \min_{1\le i\le m} A_i v - \max_{1\le i\le k} B_i v > 0, \ \gamma := \min_{1\le i\le m} \frac{A_i v}{\nu} + \max_{1\le i\le k} \frac{B_i v}{\nu} \tag{2.5}$$

Note also that when the sets $\mathcal{A}$ and $\mathcal{B}$ are linearly separable, as defined by (2.4), the plane

$$\{x | wx = \gamma\} \tag{2.6}$$

is a strict separating plane with

$$Aw > e\gamma \quad \text{and} \quad e\gamma > Bw \tag{2.7}$$

With the above definitions the following lemma becomes evident.

LEMMA 2.2. *The sets $\mathcal{A}$ and $\mathcal{B}$ represented by $A \in R^{m\times n}$ and $B \in R^{k\times n}$ respectively are linearly separable if and only if the minimum value of (2.1) is zero, in which case $(w = 0, \ \gamma)$ cannot be optimal.*

Proof. Note that the minimum of (2.1) is zero if and only if

$$-Aw + e\gamma + e \le 0 \ \text{ and } \ Bw - e\gamma + e \le 0$$

which is equivalent to the linear separability definition (2.4). To see that $(w = 0, \ \gamma)$ cannot be optimal for (2.1), note that if we set $w = 0$ in (2.1) we get

$$\min_{\gamma} (1+\gamma)_+ + (1-\gamma)_+ = 2 > 0 \tag{2.8}$$

which contradicts the requirement that the minimum of (2.1) be zero for linearly separable $\mathcal{A}$ and $\mathcal{B}$. ∎

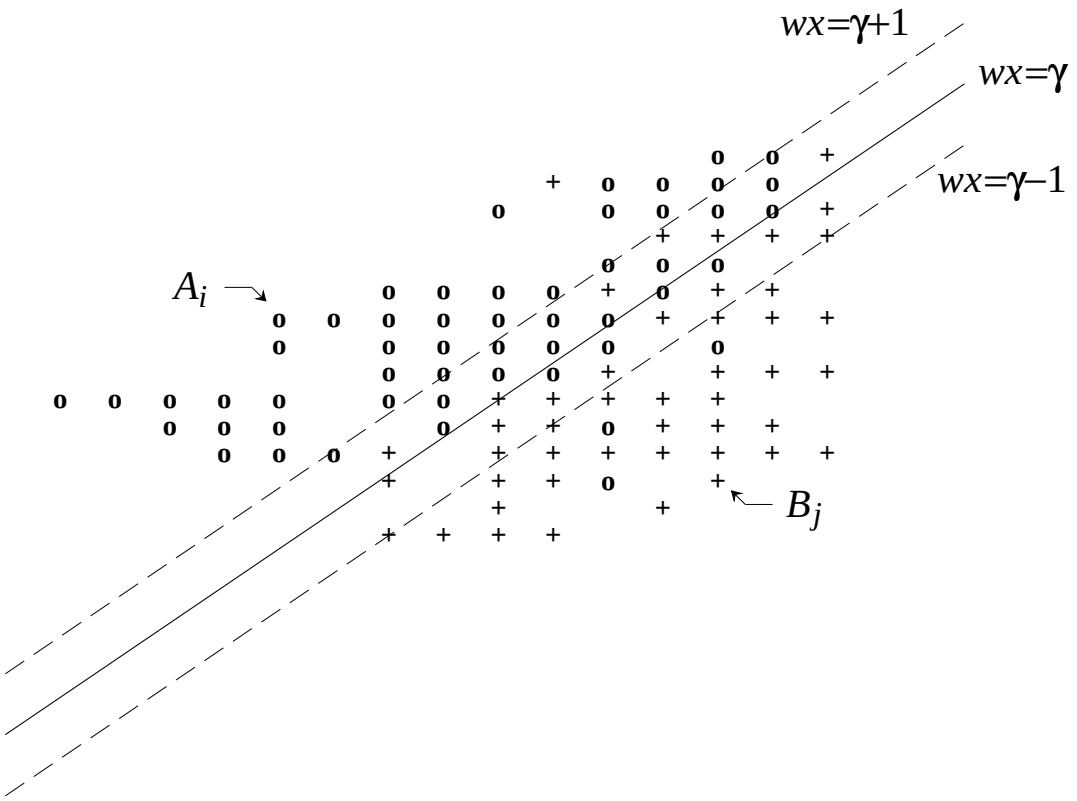

Figure 1: An Optimal Separator $wx = \gamma$ for Linearly Inseparable Sets: $\mathcal{A}$ (o) and $\mathcal{B}$ (+)

The import of Lemma 2.2 is that the optimization problem (2.1), which is equivalent to the linear program (2.11) below, will always generate a separating plane $wx = \gamma$ for linearly separable sets $\mathcal{A}$ and $\mathcal{B}$. For linearly inseparable set $\mathcal{A}$ and $\mathcal{B}$ the optimization problem (2.1) will generate an optimal separating plane $wx = \gamma$ which minimizes the average violations

$$\frac{1}{m}\sum_{i=1}^{m}(-A_i w + \gamma + 1)_+ + \frac{1}{k}\sum_{i=1}^{k}(B_i w - \gamma + 1)_+, \tag{2.1a}$$

of points of $\mathcal{A}$ which lie on the wrong side of the plane $wx = \gamma + 1$, that is in $\{x \,|\, wx < \gamma + 1\}$, and of points of $\mathcal{B}$ which lie on the wrong side of the plane $wx = \gamma - 1$, that is in $\{x \,|\, wx > \gamma - 1\}$. See Fig. 1.

Note also that the location of the plane $wx = \gamma$ obtained by minimizing the average violations (2.1a) can be further optimized by holding w fixed at the optimal

value and solving the one-dimensional optimization problem in γ

$$\min_{\min_i A_i w \le \gamma \le \max_j B_j w} \frac{1}{m}\sum_{i=1}^{m}(-A_i w+\gamma)_+ + \frac{1}{k}\sum_{k=1}^{k}(B_i w-\gamma)_+ \tag{2.1c}$$

This "secondary" optimization is not necessary in general, but for some problems it does improve the location of the optimal separator for a fixed orientation of the planes. The objective of the one-dimensional problem (2.1c) is a piecewise-linear convex function which can be easily minimized by evaluating the function at the breakpoints $\gamma = A_1 w, ..., A_m w, B_1 w, ..., B_k w$.

In order to set up the equivalent linear programming formulation to (2.1) we state first a simple lemma that relates a norm minimization problem such as (2.1) to a constrained optimization problem devoid of norms of plus-functions.

LEMMA 2.3. *Let $g: R^n \to R^m$, $h: R^n \to R^k$ and let S be a subset of R^n. The problems*

$$\min_{x \in S} \left\| g(x)_+ \right\|_1 + \left\| h(x)_+ \right\|_1 \tag{2.9a}$$

$$\min_{x \in S} \{ey + ez \,|\, y \ge g(x),\ y \ge 0,\ z \ge h(x),\ z \ge 0\} \tag{2.9b}$$

have identical solution sets.

Proof. The equivalence follows by noting that for the minimization problem (2.9b), the optimal y, z and x must be related through the equalities $y = g(x)_+$, $z = h(x)_+$. □

By using this lemma we can state an equivalent linear programming formulation to (2.1) as follows.

PROPOSITION 2.4. *For $\delta_1 > 0$, $\delta_2 > 0$, the error-minimizing problem*

$$\min_{w,\gamma} \delta_1 \left\| \left(-Aw + e\gamma + e\right)_+ \right\|_1 + \delta_2 \left\| \left(Bw - e\gamma + e\right)_+ \right\|_1 \tag{2.10a))}$$

is equivalent to the linear program

$$\min_{w,\gamma,y,z} \left\{ \delta_1 ey + \delta_2 ez \,\middle|\, Aw - e\gamma + y \ge e,\ -Bw + e\gamma + z \ge e,\ y \ge 0,\ z \ge 0 \right\} \tag{2.10b}$$

The linear program (2.10b) originally proposed by Smith[12] with equal weights $\delta_1 = \delta_2 = \dfrac{1}{m+k}$ does not possess all the properties found in our linear program with $\delta_1 = \dfrac{1}{m}$ and $\delta_2 = \dfrac{1}{k}$:

$$\min_{w,\gamma,y,z} \left\{ \frac{ey}{m} + \frac{ez}{k} \,\middle|\, Aw - e\gamma + y \ge e,\ -Bw + e\gamma + z \ge e,\ y \ge 0,\ z \ge 0 \right\} \tag{2.11}$$

The principal property that (2.11) has over other linear programs, including Smith's, is that for the linearly inseparable case it will *always* generate a nontrivial w *without* an extraneous constraint. To our knowledge no other linear programming

formulation has this property for linearly inseparable sets. We establish this property by first considering the linear program (2.10b) for arbitrary positive weights δ_1 and δ_2 and showing under what conditions $w = 0$ constitutes a solution of the problem.

THEOREM 2.5. **(Occurrence of the null solution w in linear program (2.10b))** *Let $\delta_2 k \geq \delta_1 m$. The linear program (2.10b) has a solution $(w = 0,\ \gamma, y, z)$ if and only if*

$$\frac{eA}{m} = vB,\ v \geq 0,\ ev = 1,\ v \leq \frac{\delta_2 e}{\delta_1 m}, \tag{2.12}$$

that is if the arithmetic mean of the points in $\mathcal{A}$ equals a convex combination of some points of $\mathcal{B}$. When $\delta_2 k = \delta_1 m$, (2.12) degenerates to

$$\frac{eA}{m} = \frac{eB}{k}, \tag{2.12a}$$

that is the arithmetic mean of the points in $\mathcal{A}$ equals the arithmetic mean of the points in $\mathcal{B}$.

Proof. We note first that $\delta_2 k \geq \delta_1 m$ does not result in any loss of generality because the roles of the sets $\mathcal{A}$ and $\mathcal{B}$ can be switched to obtain this inequality. Consider now the dual to the linear program (2.10b)

$$\max_{u,v} \left\{ eu + ev \middle| A'u - B'v = 0,\ -eu + ev = 0,\ 0 \leq u \leq \delta_1 e,\ 0 \leq v \leq \delta_2 e \right\} \tag{2.13}$$

The point $(w = 0, \delta, y, z)$ is optimal for the primal problem (2.10b) if and only if

$$\begin{aligned} 2\delta_1 m &= \min_{\gamma}\ \delta_1 m(1+\gamma)_+ + \delta_2 k(1-\gamma)_+ \\ &= \min_{\gamma,y,z} \left\{ \delta_1 ey + \delta_2 ez \middle| -e\gamma + y \geq e,\ e\gamma + z \geq e, (y,z) \geq 0 \right\} \end{aligned} \tag{2.14a}$$

$$= \max_{u,v} \left\{ eu + ev \middle| A'u - B'v = 0,\ -eu + ev = 0,\ 0 \leq u \leq \delta_1 e,\ 0 \leq v \leq \delta_2 e \right\} \tag{2.14b}$$

Since $eu = ev$ and $eu + ev = 2\delta_1 m$, it follows that $eu = ev = \delta_1 m$. Since $0 \leq u \leq \delta_1 e$, and so if any $u_i < \delta_1$ then $eu < \delta_1 m$ contradicting $eu = \delta_1 m$. Hence $u = \delta_1 e$ and $ev = eu = \delta_1 m$. By normalizing u and v by dividing by $\delta_1 m$ we obtain (2.12). When $\delta_2 k = \delta_1 m$, then from (2.12) we have that $0 \leq v \leq \frac{e}{k}$. Since $ev = 1$, it follows that $v = \frac{e}{k}$ and (2.12a) follows (2.12). □

This theorem gives a theoretical explanation to some observed computational experience, namely that Smith's linear program (2.10b) with $\delta_1 = \delta_2$, ended sometimes with the useless null w for real world linearly inseparable problems, whereas our linear program (2.11) never did. The reason for that is the rarity of the satisfaction of (2.12a) by real problems in contrast to the possibly frequent satisfaction of (2.12).

We now proceed to our next results which show that when the null vector $w = 0$ constitutes a solution to the linear program (2.10b), except for our proposed choice

of $\delta_1 = \dfrac{1}{m}$ and $\delta_2 = \dfrac{1}{k}$, such $w = 0$ can be unique and nothing can be done to alter it. (See Example 2.8 below.) However, for our linear program (2.11), even when the null w occurs in the rare case of (2.12a), e.g. in the contrived but classical Exclusive-Or example[11], there always exists an alternate non-null optimal w. (See Example 2.7 below.) These results are contained in the following theorem, examples and remarks.

THEOREM 2.6. **(Nonuniqueness of the null w solution to the linear program (2.11))** *The solution $(w = 0, \gamma, y, z)$ to (2.11) is not unique in w.*

Proof. Note from the first equality of (2.14a) with $\delta_1 m = \delta_2 k = 1$ that when $(\bar{w} = 0, \gamma, y, z)$ is a solution to (2.11), then $\bar{\gamma}$ can be any point in [-1, 1]. In particular, take $\bar{\gamma} = 0$. Then for this choice of $\bar{w} = 0$, $\bar{\gamma} = 0$, the corresponding optimal y, z for (2.11) are $\bar{y} = e$, $\bar{z} = e$ and the active constraints are the first two constraints of (2.11). Hence $(\bar{w}, \bar{\gamma}, \bar{y}, \bar{z})$ is unique in $\bar{w}$ if and only if the following system of linear inequalities has *no* solution (w, γ, y, z)

$$\begin{array}{ccc} \frac{e}{m}y + \frac{e}{k}z & \leq & \frac{e}{m}\bar{y} + \frac{\bar{z}}{k}\bar{z} \\ Aw - e\gamma + y & \geq & A\bar{w} - e\bar{\gamma} + \bar{y} = e \\ -Bw + e\gamma + z & \geq & -B\bar{w} + e\bar{\gamma} + \bar{z} = e \\ w & \neq & \bar{w} \end{array} \tag{2.15a}$$

This is equivalent to the following system of linear inequalities having no solution (w, γ, y, z) for *each* h in R^n :

$$\begin{array}{cc} -\frac{e}{m}(y - \bar{y}) - \frac{e}{k}(z - \bar{z}) & \geq 0 \\ A(w - \bar{w}) - e(\gamma - \bar{\gamma}) + (y - \bar{y}) & \geq 0 \\ -B(w - \bar{w}) + e(\gamma - \bar{\gamma}) + (z - \bar{z}) & \geq 0 \\ -h(w - \bar{w}) & > 0 \end{array} \tag{2.15b}$$

By Motzkin's theorem of the alternative[8], (2.15b) has no solution for a given h in R^n if and only if the following system of linear inequalities *does* have a solution (ζ, u, v) for that h in R^n :

$$\begin{array}{cc} A'u - B'v & = h \\ -eu + ev & = 0 \\ -\frac{1}{m}e\zeta + u & = 0 \\ -\frac{1}{k}e\zeta + v & = 0 \\ \zeta, u, v & \geq 0 \end{array} \tag{2.16}$$

Obviously it is possible to choose h in R^n such that (2.16) has no solution, since there are h in R^n that cannot be written as:

$$h = \frac{A'e\zeta}{m} - \frac{B'e\zeta}{k},\ \zeta \geq 0. \tag{2.17}$$

Hence (2.15b) has a solution for some h in R^n. Consequently (2.15a) has a solution and $\bar{w} = 0$ is not unique. ∎

We now apply this theorem to the classical Exclusive-Or example[11] for which condition (2.12a) is satisfied and hence $(\bar{w}=0, \bar{\gamma}, \bar{y}, \bar{z})$ is a solution to (2.11) which, however, is not unique in $\bar{w}=0$.

EXAMPLE 2.7. **(Exclusive-Or)**

$$A=\begin{bmatrix}0 & 0\\ 1 & 1\end{bmatrix}; \quad B=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}.$$

For this example $\frac{eA}{2}=\frac{eB}{2}$ and $(\bar{w}=0,\ \bar{\gamma}=0,\ \bar{y}=e,\ \bar{z}=e)$ is a solution to the linear program (2.11) which can also be written in the equivalent (2.1) formulation of

$$(2.18)\quad \min_{w,\gamma} \frac{1}{2}\left[(1+\gamma)_+ + (1+\gamma-w_1-w_2)_+ + (1-\gamma+w_1)_+ + (1-\gamma+w_2)_+\right] = 2$$

However, the point $\bar{w}=\begin{pmatrix}1\\1\end{pmatrix}$, $\bar{\gamma}=1$ is also optimal, because it gives the same minimum value of 2. The $-45°$ direction in the $w-$space associated with this solution is useful in the multisurface method of pattern separation[7,1] since it can be used to generate the first part of a piecewise-linear separator. In practice the linear program package returned the optimal point $\bar{w}=\begin{pmatrix}2\\-2\end{pmatrix}$, $\bar{\gamma}=-1$, which generates another $45°$ direction that can also be used for piecewise-linear separation.

We turn now to the case when $\delta_1 m \neq \delta_2 k$. In particular consider Smith's case of $\delta_1=\delta_2=\frac{1}{m+k}$. A similar analysis to that of Theorem 2.6 does not give guaranteed nonuniqueness of the solution $(\bar{w}=0, \bar{\gamma}, \bar{y}, \bar{z})$ in $\bar{w}=0$ for the linear program (2.10b) with $\delta_1=\delta_1=\frac{1}{m+k}$. In fact, to the contrary, the analysis shows that indeed $\bar{w}=0$ is unique under certain conditions which are satisfied by the following counterexample from Mangasarian *et al* [9] to Smith's claim[12] that his linear program (2.10b) with $\delta_1=\delta_2=\frac{1}{m+k}$ always generates a nonzero $\bar{w}$. In reality $\bar{w}=0$ is unique for this example.

EXAMPLE 2.8. **(Unique $\bar{w}=0$ for Smith's LP (2.10b),** $\delta_1=\delta_2=\frac{1}{m+k}$)

$$A=\begin{bmatrix}1\\2\end{bmatrix}, \quad B=\begin{bmatrix}-1\\0\\4\end{bmatrix}, \quad m=2, \quad k=3.$$

For this problem the equivalent norm-minimization problem (2.1) to the Smith's LP (2.10b) with $\delta_1=\delta_2=\frac{1}{m+k}$ is:

$$(2.19)\quad \begin{aligned} \min_{w,\gamma}\ f(w,\alpha) := \min_{w,\gamma}\quad & \tfrac{1}{5}\left[(-w+\gamma+1)_+ + (-2w+\gamma+1)_+\right.\\ & \left.+(-w-\gamma+1)_+ + (-\gamma+1)_+ + (4w-\gamma+1)_+\right] = \tfrac{4}{5}\end{aligned}$$

is achieved at $\bar{w} = 0$, $\bar{\gamma} = 1$. The uniqueness of this solution can be established by considering the subdifferential (see Section 5.1.4, p.127 of Polyak[10] and Equation 14.1.4, p.363 of Fletcher[3]) of the function $f(w, \gamma)$ at (0,1) which is given by

$$\partial f(0,1) = \frac{1}{5}\begin{bmatrix} -1-2-\ell_1+0\cdot\ell_2+4\ell_3 \\ 1+1-\ell_1-\ell_2-\ell_3 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} -3-\ell_1+4\ell_3 \\ 2-\ell_1-\ell_2-\ell_3 \end{bmatrix}$$

with $0 \le \ell \le e$. Since $0 \in \partial f(0,1)$ with

$$1 \ge \ell_3 \ge 0.8, \quad \ell_2 = 5(1-\ell_3), \quad \ell_1 = 4\ell_3 - 3,$$

it follows that $0 \in$ interior $(\partial f(0,1))$ and hence by Lemma 3, p.137 of Polyak[10], (0,1) is a unique solution of (2.19). (The uniqueness of (0,1) can also be shown by considering the linear program (2.10b)).

Grinold[5] proposed the following linear program

$$\min_{w,\gamma,\rho} \left\{ -\rho \,\middle|\, \begin{array}{c} Aw - e\gamma - e\rho \ge 0,\ -Bw + e\gamma - e\rho \ge 0, \\ (eA - eB)w + (k-m)\gamma = k + m \end{array} \right\} \tag{2.20}$$

In Mangasarian *et al* [9] the example $A = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, $B = \begin{bmatrix} \frac{1}{2} \\ 2 \\ 4 \end{bmatrix}$ was given to show that the solution $(\bar{w} = 0,\ \bar{\gamma} = 5,\ \bar{\rho} = -5)$ is unique in $\bar{w}$. In fact, it can be shown that $(\bar{w} = 0, \bar{\gamma}, \bar{\rho})$ is always a solution of (2.20) whenever there exist u such that

$$uA + \frac{eA - eB}{k-m} = 0, \quad eu = 1, \quad u \ge 0, \quad k > m \tag{2.21}$$

Furthermore, $\bar{w} = 0$ is unique if for each h in R^n

$$\left(A' + \frac{eA - eB}{k-m}e\right)u = h, \quad \text{has a solution} \quad u \ge 0 \tag{2.22}$$

3 COMPUTATIONAL COMPARISONS

In this section we give computational comparisons on two real-world databases: the Wisconsin Breast Cancer Database[14,13], and the Cleveland Heart Disease Database[2], using our proposed linear program (2.11), Smith's linear program (2.10b) with $\delta_1 = \delta_2 = \dfrac{1}{m+k}$ and the following linear programming formulation[9] for multisurface discrimination

$$\max_{1 \le i \le n} \max_{w,\alpha,\beta} \left\{ \alpha - \beta \,\middle|\, Aw \ge e\alpha,\ Bw \le e\beta,\ -e \le w \le e, w_i = \pm 1 \right\}. \tag{3.1}$$

Note that (3.1) can be solved by solving $2n$ linear programs. The corresponding linear separation obtained from (3.1) is

$$\bar{w}x = \frac{\bar{\alpha} + \bar{\beta}}{2} \tag{3.2}$$

where $(\bar{w}, \bar{\alpha}, \bar{\beta})$ is a solution of (3.1). Problem (3.1) is equivalent[9] to

$$\max_{w,\alpha,\beta} \left\{\alpha - \beta \middle| Aw \geq e\alpha,\ Bw \leq e\beta,\ \left\|w\right\|_\infty = 1\right\} \tag{3.3}$$

which in turn is easily seen to be the following problem

$$\max_{\|w\|_\infty = 1} \left(\min_{1 \leq i \leq m} A_i w - \max_{1 \leq i \leq k} B_i w \right) \tag{3.4}$$

Fig.2 summarizes the results obtained for two linearly inseparable databases using the three linear programs mentioned above. The Wisconsin Breast Cancer Database consists of 566 points of which 354 are benign and 212 are malignant, all in a 9-dimensional real space. The Cleveland Heart Disease Database consists of 197 points in a 13-dimensional real space, of which 137 are negative and 60 are positive. Our testing methodology consisted of dividing each set randomly into a training set consisting of 67% of the data and a testing set consisting of the remaining 33%. Each linear programming formulation was run on the training set and the resulting separator tested on the testing set. This was repeated ten times and the average results of the ten runs are depicted in Fig. 2. No "secondary" optimization using (2.1c) was performed for any method. For each database our linear program (2.11) (referred to as MSM1, multisurface method 1-norm) outperformed both Smith's linear program (2.10b) with $\delta_1 = \delta_2$, and the linear programming method of (3.1), referred to as MSM. The average run times for MSM1 and Smith are very close: 3.84 and 3.89 seconds respectively on a DEC station 5000/125 for the Wisconsin Breast Cancer Database, while the corresponding time for MSM was 53.54 seconds. For the Cleveland Heart Disease Database the corresponding times are: 2.82, 2.89 and 53.82 seconds respectively. Note that the percent error of MSM1 on the testing set was better than that of Smith and considerably better than that of MSM on both databases.

4 CONCLUSIONS

We have presented a robust linear program which always generates a linear surface as an "optimal" separator for two linearly inseparable sets. The "optimality" of the separator consists in minimizing a weighted average sum of the violations of the points lying on the wrong side of the separator. By using an appropriately weighted sum, we have overcome the problem of the null solution which has plagued previous linear programming approaches. These approaches either left this difficulty unresolved or imposed an extraneous linear constraint which never resolved the problem completely. The fact that computational results on real-world problems give an edge to our linear program over Smith's and a substantial edge over another linear programming approach makes it, in our opinion, a suitable linear program for the linearly inseparable case.

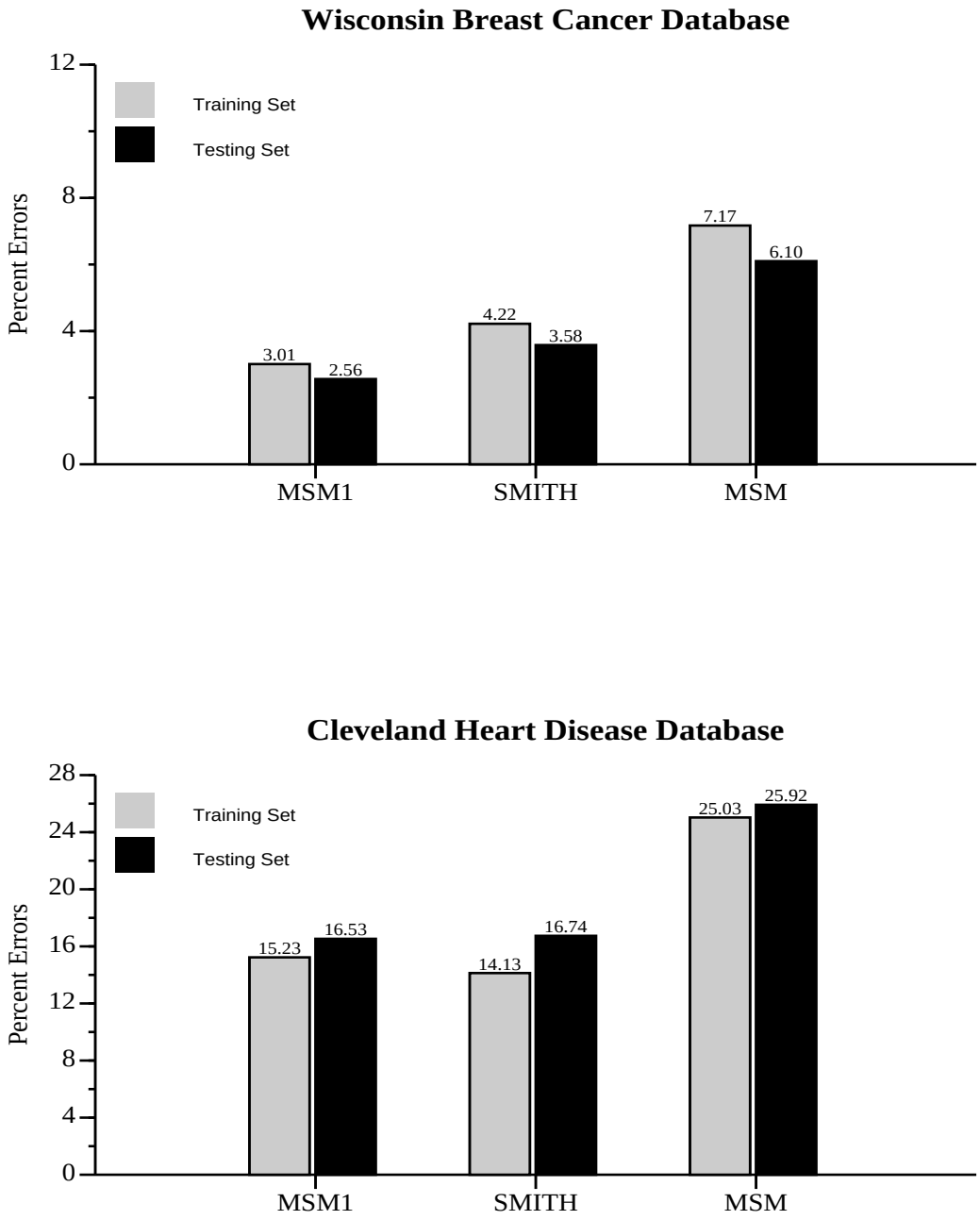

Figure 2: Comparison of Three Linear Programming Discriminators for Linearly Inseparable Sets

ACKNOWLEDGEMENTS

This material is based on research supported by Air Force Office of Scientific Research Grant AFOSR-89-0410, National Science Foundation Grant CCR-9101801, and Air Force Laboratory Graduate Fellowship Program Grant SSAN 531-56-2969.

REFERENCES

1. K. P. Bennett and O. L. Mangasarian, *Neural Network Training Via Linear Programming,* in P. M. Pardalos (Ed.), *Advances in Optimization and Parallel Computing*, North Holland, Amsterdam 1992, pp. 56-67.
2. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V. Froelicher, *International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease,* American Journal of Cardiology 64, 1989, pp. 304-310.
3. R. Fletcher, *Practical Methods in Optimization,* Second Edition, Wiley, New York 1987.
4. F. Glover, *Improved Linear Programming Models for Discriminant Analysis,* Decision Sciences, 21,4, 1990, pp. 771-785.
5. R. C. Grinold, *Mathematical Programming Methods of Pattern Classification,* Management Science 19, 1972, pp. 272-289.
6. O. L. Mangasarian, *Linear and Nonlinear Separation of Patterns by Linear Programming,* Operations Research 13, 1965, pp. 444-452.
7. O. L. Mangasarian, *Multisurface Method of Pattern Separation,* IEEE Transactions on Information Theory IT-14(6), 1968, pp. 801-807.
8. O. L. Mangasarian, *Nonlinear Programming,* McGraw-Hill, New York 1969, pp. 28-29.
9. O. L. Mangasarian, R. Setiono, and W. H. Wolberg, *Pattern Recognition via Linear Programming: Theory and Application to Medical Diagnosis,* in: *Large-Scale Numerical Optimization,* Thomas F. Coleman and Yuying Li, (Eds.), SIAM, Philadelphia 1990, pp. 22-30.
10. B. T. Polyak, *Introduction to Optimization,* Optimization Software, New York 1987.
11. D. E. Rumelhart, G. E. Hinton, and J. L. McClelland, *Learning Internal Representations,* in *Parallel Distributed Processing,* D. E. Rumelhart and J. L. McClelland (Eds.), Vol. I, M.I.T. Press, Cambridge, Massachusetts 1969, pp. 318-362.
12. F. W. Smith, *Pattern Classifier Design by Linear Programming,* IEEE Transactions on Computers C-17, 4, 1968, pp. 367-372.
13. W. H. Wolberg AND O. L. Mangasarian, *Multisurface Method of Pattern Separation Applied to Breast Cytology Diagnosis,* Proceeding of National Academy of Sciences U.S.A. 87, 1990, pp. 9193-9196.
14. W. H. Wolberg, M. S. Tanner and W. Y. Loh, *Diagnostic Schemes for Fine Needle Aspirates of Breast Masses,* Analytical and Quantitative Cytology and Histology 10, 1988, pp. 225-228.