Arbitrary-Norm Separating Plane *

O. L. Mangasarian†

Abstract

A plane separating two point sets in n-dimensional real space is constructed such that it minimizes the sum of arbitrary-norm distances of misclassified points to the plane. In contrast to previous approaches that used surrogates for distance-minimization, the present work is based on a precise norm-dependent explicit closed form for the projection of a point on a plane. This projection is used to formulate the separating-plane problem as a minimization of a convex function on a unit sphere in a norm dual to that of the arbitrary norm used. For the 1-norm, the problem can be solved in polynomial time by solving $2n$ linear programs or by solving a bilinear program. For a general p-norm, the minimization problem can be transformed via an exact penalty formulation to minimizing the sum of a convex function and a bilinear function on a convex set. For the one and infinity norms, a finite successive linearization algorithm can be used for solving the exact penalty formulation.

1 Introduction

One of the fundamental problems of machine learning is that of discriminating between two finite point sets in n-dimensional real space R^n. When the convex hulls of the two sets do not intersect, a single linear program can construct a strict separating plane such that each of the two open halfspaces generated by the plane contains one of the two sets [6, 12, 2]. Such a plane corresponds to a perceptron and can also be obtained by the iterative perceptron learning algorithm [22, 10] which can be interpreted as the Motzkin-Schoenberg iterative scheme for solving consistent linear inequalities [21]. When the convex hulls of the two sets intersect the iterative scheme fails because the underlying linear inequalities are inconsistent, while the linear programming approach must be provided with an error criterion to be minimized. We propose as a criterion here the sum of arbitrary-norm distances to the separating plane of points lying on the wrong side of the plane. Unfortunately, if precise distances are used, linearity of the objective function is lost, as will be shown in this work. What many of the previous approaches have tended to use, are distance surrogates that maintained linearity of the problem [9, 8, 20, 2] but did not measure distances of violating points to the separating plane. Invariably such approaches have to contend with the null solution that does not generate any separating plane. In contrast, our approach here excludes the null solution and depends on a closed form formula given in Theorem 2.1 below for the projection of a point in R^n onto a given plane using an arbitrary norm. In Section 3 we formulate the problem of obtaining a separating plane that minimizes the sum of distances of points on the wrong side of the plane, as that of minimizing a convex function on a unit sphere in a norm dual to that

*Mathematical Programming Technical Report 97-07, May 1997. Revised April 1998. This material is based on research supported by NSF Grant CCR-9322479 and AFOSR Grant F49620-97-1-0326.

†Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706, *olvi@cs.wisc.edu*. The author gratefully acknowledges the gracious hospitality of the Mathematics Department of the University of California at San Diego during his sabbatical leave January-May 1997.

used for measuring distance to a projection on the separating plane (problem (17)). In Theorem 3.1 the nonconvex sphere constraint is replaced by two unit-ball convex constraints and a bilinear constraint. This allows us to reformulate the problem as an exact penalty problem for the 1-norm and ∞-norm that leads to a problem of minimizing the sum of a convex function and a bilinear function on a convex set (Theorem 3.2). This in turn leads to a finitely terminating successive linear programming algorithm for the 1-norm problem (Algorithm 3.3). The 1-norm problem can also be solved exactly by $2n$ linear programs (problem (22)).

A word about our notation and background material. All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. For $x \in R^n$, $|x|$ will denote the vector in R^n with components that are the absolute values of the components x_i of x. The scalar product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$. For a linear program $\min\limits_{x \in X} c'x$ the notation arg vertex $\min\limits_{x \in X} c'x$, will denote the set of vertex solutions of the linear program. For $x \in R^n$ and $p \in [1, \infty)$, the norm $\|x\|_p$ will denote the p-norm: $(\sum\limits_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ and $\|x\|_\infty$ will denote $\max\limits_{1 \le i \le n} |x_i|$. For $x \in R^n$, $(x_+)_i = \max\{0, x_i\}$, $i = 1, \ldots, n$. For an $m \times n$ matrix A, A_i will denote the ith row of A and A_{ij} will denote the element in row i and column j. The identity matrix in a real space of arbitrary dimension will be denoted by I, while a column vector of ones of arbitrary dimension will be denoted by e. The symbol $:=$ will denote a definition of the term appearing to the left of the symbol by the term appearing to the right of the symbol. For a general norm $\|\cdot\|$ on R^n, the dual norm $\|\cdot\|'$ on R^n is defined as

$$\|x\|' := \max_{\|y\|=1} x'y, \tag{1}$$

from which follows the generalized Cauchy-Schwarz inequality,

$$\pm x'y \le |x'y| \le \|x\|'\|y\|. \tag{2}$$

For $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, the p-norm and q-norm are dual norms by the classical Hölder inequality [1]. A norm $\|\cdot\|$ on R^n is said to be monotonic (or absolute) if either of the following equivalent conditions hold:

$$\begin{array}{rcl} x, y \in R^n,\ |x| \le |y| & \Longrightarrow & \|x\| \le \|y\| \\ \|\,|x|\,\| & = & \|x\|\ \forall x \in R^n \end{array} \tag{3}$$

For $p \in [1, \infty]$, the p-norm is monotonic. It is well known [23] that for $x \in R^n$ the general norm $\|x\|$ is convex on R^n, has bounded level sets in R^n, and is continuous on R^n.

2 Arbitrary-Norm Projection on a Plane

In this section we derive an explicit expression for the projection of a point on a given plane using an arbitrary norm. We will use this expression to derive a mathematical programming formulation of the separating plane problem. We first give an explicit form for the projection of an arbitrary point on a given plane using a general norm for measuring the distance between the point and its projection.

Theorem 2.1 Arbitrary-Norm Projection on a Plane. *Let $q \in R^n$ be any point in R^n not on the plane:*

$$P := \{x \mid w'x = \gamma\},\ 0 \ne w \in R^n,\ \gamma \in R. \tag{4}$$

A projection $p(q) \in P$ using a general norm $\|\cdot\|$ on R^n is given by:

$$p(q) = q - \frac{w'q-\gamma}{\|w\|'}y(w), \tag{5}$$

where $\|\cdot\|'$ is the dual norm to $\|\cdot\|$ and:

$$y(w) \in arg \max_{\|y\|=1} w'y. \tag{6}$$

Consequently, the distance between q and its projection $p(q)$ is given by:

$$\|q-p(q)\| = \frac{|w'q-\gamma|}{\|w\|'}. \tag{7}$$

Proof The proof consists of showing that $p(q) \in P$ and that it satisfies the right inequality of the Karush-Kuhn-Tucker saddlepoint sufficient optimality criterion [14, Equation 5.1.4, Theorem 5.3.1] for some Lagrange multiplier $\lambda \in R$:

$$\|p(q)-q\| \leq \|x-q\| - \lambda(w'x-\gamma),\ \forall x \in R^n. \tag{8}$$

To show that $p(q) \in P$ note that:

$$w'p(q)-\gamma = w'q + \frac{\gamma-w'q}{\|w\|'}w'y(w) - \gamma = w'q + \frac{\gamma-w'q}{\|w\|'}\|w\|' - \gamma = 0. \tag{9}$$

Hence $p(q) \in P$. To show that (8) holds, define:

$$\lambda := \frac{1}{\|w\|'} \cdot \frac{|\gamma-w'q|}{(\gamma-w'q)}. \tag{10}$$

So we need to show that:

$$\|\frac{\gamma-w'q}{\|w\|'}y(w)\| \leq \|x-q\| - \frac{1}{\|w\|'}\frac{|\gamma-w'q|}{\gamma-w'q} \cdot (w'x-\gamma),\ \forall x \in R^n, \tag{11}$$

or equivalently that:

$$|\gamma-w'q| \cdot \|y(w)\| + (w'x-\gamma)\frac{|\gamma-w'q|}{(\gamma-w'q)} \leq \|w\|' \cdot \|x-q\|,\ \forall x \in R^n. \tag{12}$$

Since $\|y(w)\| = 1$, inequality (12) is equivalent to:

$$\pm\, w'(x-q) \leq \|w\|' \cdot \|x-q\|,\ \forall x \in R^n, \tag{13}$$

which follows immediately from the generalized Cauchy-Schwarz inequality or equivalently from the definition of the dual norm. Hence (8) holds, and by the Karush-Kuhn-Tucker saddlepoint sufficiency theorem [14, Theorem 5.3.1] $p(q)$ as given by (5) is a projection of q on P. Formula (7) for the distance between q and its projection $p(q)$ follows from (5) by noting that $\|y(w)\| = 1$. $\diamondsuit$

It can be easily shown from the definition (1) of the dual norm or from [23, Corollary 23.5.3] that $y(w)$ of (6) is a subgradient $\partial\|w\|'$ of the dual norm $\|w\|'$ and hence (5) can be rewritten as:

$$q-p(q) = (w'q-\gamma)\frac{\partial\|w\|'}{\|w\|'} \tag{5a}$$

As immediate consequences of the above theorem we give projections of q on P using the common one, two and infinity norms.

Corollary 2.2 1-Norm Projection. *For the 1-norm* $\|\cdot\|_1$, $w'y(w) = \|w\|_\infty$ *and hence*

$$\left.\begin{array}{l} y_i(w) = 0 \;\; for \; |w_i| \neq \|w\|_\infty \\ y_i(w) = \nu_i \; for \; w_i = \|w\|_\infty \\ y_i(w) = -\nu_i \; for \; w_i = -\|w\|_\infty \end{array}\right\}, \; \nu_i \geq 0, \; \sum_{i=1}^{n} \nu_i = 1.$$

Furthermore, since $\|y(w)\|_1 = 1$, *it follows from (5) and (7) that:*

$$p(q) = q - \frac{w'q - \gamma}{\|w\|_\infty} y(w), \;\; \& \;\; \|q - p(q)\|_1 = \frac{|w'q - \gamma|}{\|w\|_\infty}.$$

Corollary 2.3 2-Norm Projection. *For the 2-norm* $\|\cdot\|_2$, $w'y(w) = \|w\|_2$ *and hence*

$$y(w) = \frac{w}{\|w\|_2}.$$

Furthermore, since $\|y(w)\|_2 = 1$, *it follows from (5) and (7) that:*

$$p(q) = q - \frac{w'q - \gamma}{\|w\|_2} y(w), \;\; \& \;\; \|q - p(q)\|_2 = \frac{|w'q - \gamma|}{\|w\|_2}.$$

Corollary 2.4 ∞-Norm Projection. *For the infinity-norm* $\|\cdot\|_\infty$, $w'y(w) = \|w\|_1$ *and hence*

$$\left.\begin{array}{l} y_i(w) = 1 \;\; for \; w_i > 0 \\ y_i(w) = -1 \; for \; w_i \leq 0 \end{array}\right\}.$$

Furthermore, since $\|y(w)\|_\infty = 1$, *it follows from (5) and (7) that:*

$$p(q) = q - \frac{w'q - \gamma}{\|w\|_1} y(w), \;\; \& \;\; \|q - p(q)\|_\infty = \frac{|w'q - \gamma|}{\|w\|_1}.$$

If we make use of standard inequalities between various norms such as [24, p 170]:

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq n^{\frac{1}{2}} \|x\|_2 \leq n\|x\|_\infty, \tag{14}$$

and the fact that for a fixed $x \in R^n$, the p-norm is a nonincreasing function of $p \in (0, \infty]$ [1, p 18], we obtain the following corollary to Theorem 2.1 relating norm-dependent distances between the point q and its projection $p(q)$ on the plane P.

Corollary 2.5 Inequalities Between Norm-Dependent Distances to Projections. *Let* $d(q,P)_\ell$ *denote the distance between the point* q *and its projection* $p(q)$ *on the plane* P *using the norm* $\|\cdot\|_\ell$, $\ell \in [1, \infty]$. *Then*

$$d(q,P)_1 \geq d(q,P)_2 \geq d(q,P)_\infty \geq n^{-\frac{1}{2}} d(q,P)_2 \geq n^{-1} d(q,P)_1 \tag{15}$$

and

$$d(q,P)_{\ell_1} \geq d(q,P)_{\ell_2}, \; for \; \ell_1, \; \ell_2 \in [1, \infty], \; \ell_2 > \ell_1. \tag{16}$$

We demonstrate now the above projections with a very simple example.

Example 2.6

$$P = \{x \in R^2 \mid x_1 + x_2 = 4\},\ q = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

$\boxed{\textbf{1-norm } \|\cdot\|_1}$

$$y(w) = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix},\ \nu_1 \geq 0,\ \nu_2 \geq 0,\ \nu_1 + \nu_2 = 1$$

$$p(q) = \begin{bmatrix} 5 \\ 3 \end{bmatrix} - \frac{8-4}{1} \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} = \begin{bmatrix} 5 - 4\nu_1 \\ 3 - 4\nu_2 \end{bmatrix},\ \nu_1 \geq 0,\ \nu_2 \geq 0,\ \nu_1 + \nu_2 = 1,\ \|q - p(q)\|_1 = 4.$$

$\boxed{\textbf{2-norm } \|\cdot\|_2}$

$$y(w) = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$p(q) = \begin{bmatrix} 5 \\ 3 \end{bmatrix} - \frac{8-4}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix},\ \|q - p(q)\|_2 = 2\sqrt{2}.$$

$\boxed{\boldsymbol{\infty}\textbf{-norm } \|\cdot\|_\infty}$

$$y(w) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$p(q) = \begin{bmatrix} 5 \\ 3 \end{bmatrix} - \frac{8-4}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix},\ \|q - p(q)\|_\infty = 2.$$

We note that the inequalities (15) and (16) of Corollary 2.5 are satisfied by the norm-dependent distances computed in Example 2.6.

With the above projection results we can now turn our attention to separating planes based on arbitrary norms.

3 Arbitrary-Norm Separating Plane as a Mathematical Program

We consider two finite point sets $\mathcal{A}$ and $\mathcal{B}$ in R^n represented respectively by the matrices $A \in R^{m\times n}$ and $B \in R^{k\times n}$. We wish to construct a plane P defined by (4) so that most of the points of $\mathcal{A}$ are on one side of P and most of the points of $\mathcal{B}$ are on the other side of P. We will attempt to achieve this by minimizing the sum of arbitrary-norm distances to P of points of $\mathcal{A}$ on the wrong side of P and distances of points to P of $\mathcal{B}$ on the wrong side of the plane as well. If we assume, without loss of generality, that the points of $\mathcal{A}$ should lie on side of P in the direction of the normal w to P and the points of $\mathcal{B}$ on the other side, we then have the following optimization problem, where the variables (w, γ) have been normalized by dividing by $\|w\|'$, the dual of the general norm $\|\cdot\|$ used:

$$\min_{(w,\gamma)\in R^{n+1}} \left\{ \left\| \begin{bmatrix} -Aw + e\gamma \\ Bw - e\gamma \end{bmatrix}_+ \right\|_1 \;\middle|\; \|w\|' = 1 \right\}. \tag{17}$$

We note that the *fixed* 1-norm in the objective function, which is used to obtain the sum of norm-dependent distances to the plane P of misclassified points of $\mathcal{A}$ and $\mathcal{B}$, is not related to the norm appearing in the constraint. However, the norm in the constraint is an *arbitrary* norm on R^n that is dual to the norm used in measuring distances of misclassified points to the plane P. We also note that the objective function of (17) is convex on R^n for any monotonic norm including the 1-norm. This follows from the easily established fact that the function $f(x) := \|h(x)_+\|$ is convex on R^n for any monotonic norm $\|\cdot\|$ on R^m and any convex function $h : R^n \longrightarrow R^m$ on R^n. The feasible region of (17), which is the unit sphere in the dual norm $\|\cdot\|'$, is however not convex. It is precisely this essential nonconvex condition that has been either ignored in most previous work [12, 9, 8, 2] or used heuristically [13, 20] to enforce nonzeroness of w but not as a distance-normalization constraint. Thus in these papers, the sum of the distances of misclassified points to the separating plane has not been the real objective function that has been minimized. In [5] a 2-norm error term is used, instead of the 1-norm error term of [2], which again is not a measure of the distance of misclassified points to the separating plane in either [2] or [5]. However, if the 1-norm is used to measure actual distance to the separating plane, then the resulting dual norm appearing in (17) is the ∞-norm and (17) can be solved by solving $2n$ linear programs as shown in (22) below.

In order to alleviate the nonconvexity of the constraint $\|w\|' = 1$ of (17) to some extent, we introduce an equivalent formulation of (17) which is convex except for a single bilinear constraint. This will allow us to construct a penalty problem that is convex except for a bilinear penalty term. Such bilinear reformulations have been successfully used to solve other difficult problems [3, 16, 4]. We state and establish this equivalent reformulation of (17).

Theorem 3.1 *The mathematical program (17) is equivalent to the following:*

$$\min_{(w,\gamma,t)\in R^{n+1+n}} \left\{ \left\| \begin{bmatrix} -Aw + e\gamma \\ Bw - e\gamma \end{bmatrix}_+ \right\|_1 \middle| \; \|w\|' \leq 1, \; \|t\| \leq 1, \; w't \geq 1 \right\}. \tag{18}$$

Proof We establish the equivalence of the mathematical programs (17) and (18) by showing that the constraints of each program imply the constraints of the other program.

($\Longrightarrow$) Let $\|w\|' = 1$. Then $(w, t := y(w))$ satisfies the constraints of (18), where

$$y(w) = \arg\max_{\|y\|=1} \; w'y = \|w\|' = 1.$$

This is so because

$$\|t\| = \|y(w)\| = 1, \; \|w\|' = 1, \; w't = w'y(w) = \|w\|' = 1,$$

and hence the constraints of (18) are satisfied by $(w, t := y(w))$. Consequently:

$$\text{min of (18)} \; \leq \; \text{min of (17)}.$$

($\Longleftarrow$) Let (w, t) satisfy the constraints of (18). Then by the generalized Cauchy-Schwarz inequality:

$$1 \leq w't \leq \|w\|' \cdot \|t\| \leq 1.$$

Since $\|w\|' \leq 1$, $\|t\| \leq 1$, it follows that $\|w\|' = 1$ and $\|t\| = 1$. Hence w satisfies the constraint of (17) and

$$\text{min of (17)} \; \leq \; \text{min of (18)}.$$

$\diamondsuit$

We can reformulate problem (18) as a penalty problem for which the only nonconvexity is a bilinear term in the objective function. In addition, for the cases of the one and infinity norms, the penalty function formulation can be shown to be exact. We state this result as the following theorem.

Theorem 3.2 *The mathematical program (18) can be reformulated as the following penalty problem:*

$$\min_{(w,\gamma)\in R^{n+1}} \left\{ \left\| \begin{bmatrix} -Aw+e\gamma \\ Bw-e\gamma \end{bmatrix}_+ \right\|_1 - \alpha w't \;\middle|\; \|w\|' \leq 1,\ \|t\| \leq 1 \right\},\ \alpha > 0 \tag{19}$$

For a general norm, the sequence $\{(w(\alpha_i), \gamma(\alpha_i), t(\alpha_i))\}$ of penalty problem solutions of (19), for a sequence $\{\alpha_i\} \uparrow \infty$ of penalty parameters, has an accumulation $(\bar{w}, \bar{\gamma}, \bar{t})$, and each such accumulation point solves (18). For the one and infinity norms, some fixed vertex solution solves (19) as well as (18) for each α_{i_j} of some subsequence $\{\alpha_{i_j}\}$ of $\{\alpha_i\} \uparrow \infty$.

Proof From the first two constraints of (18) we have that

$$w't \leq \|w\|'\|t\| \leq 1,$$

and hence $1-w't \geq 0$. Thus the exterior penalty term $\alpha(1-w't)_+$ for the constraint $w't \geq 1$ of (18) is equivalent to $\alpha(1-w't)$. Dropping the constant term α, gives the penalty term $-\alpha w't$ of (19). The first part of the theorem then follows from standard exterior penalty function results such as [15, Theorem 2.8]. We establish the second part for the 1-norm only, the proof for the ∞-norm is similar. For the 1-norm, problems (18) and (19) can be written as the following programs with a single bilinear term in the constraints and the objective function respectively:

$$\min_{(w,t,s,\gamma,y,z)\in R^{3n+1+m+k}} \left\{ e'y+e'z \;\middle|\; \begin{array}{l} -Aw+e\gamma \leq y,\ y \geq 0,\ Bw-e\gamma \leq z,\ z \geq 0, \\ -e \leq w \leq e,\ \ -s \leq t \leq s,\ e's \leq 1,\ w't \geq 1 \end{array} \right\} \tag{20}$$

$$\min_{(w,t,s,\gamma,y,z)\in R^{3n+1+m+k}} \left\{ e'y+e'z-\alpha w't \;\middle|\; \begin{array}{l} -Aw+e\gamma \leq y,\ y \geq 0,\ Bw-e\gamma \leq z,\ z \geq 0, \\ -e \leq w \leq e,\ \ -s \leq t \leq s,\ e's \leq 1 \end{array} \right\} \tag{21}$$

Since the quadratic objective function of (21) is bounded below by $-\alpha\|w\|_\infty \cdot \|t\|_1 = -\alpha$, it follows that the quadratic program (21) has a solution [7]. Since the constraints of (21) are separable in (w,γ,y,z) and (t,s), it follows that (21) has a vertex solution because from any solution point a vertex of the (w,γ,y,z) constraints can be obtained by a single linear program in (w,γ,y,z), and second linear program in (t,s) from that vertex gives the desired vertex solution of (21) [3, Proposition 2.1]. By [19, Theorem 2.1], or more simply by noting that some fixed vertex of the feasible region of (21) solves (21) for a subsequence $\{\alpha_{i_j}\} \uparrow \infty$ and hence must also solve (20).$\diamondsuit$

We turn our attention now to algorithmic aspects for solving 1-norm separating plane. First we note that it can be solved in polynomial time because the solution of (17) for the 1-norm can be obtained by the best solution of the following $2n$ polynomially solvable [11, 25] linear programs

$$\min_{(w,\gamma,y,z)\in R^{n+1+m+k}} \left\{ e'y+e'z \;\middle|\; \begin{array}{l} -Aw+e\gamma \leq y,\ y \geq 0,\ Bw-e\gamma \leq z,\ z \geq 0, \\ -e \leq w \leq e,\ w_i = \pm 1 \end{array} \right\},\ i=1,\ldots,n. \tag{22}$$

If n is large, one may wish to resort to the following finitely terminating successive linearization algorithm for solving the bilinear formulation (21) of the 1-norm separating plane problem.

Algorithm 3.3 *Let $\alpha > 0$ and start with arbitrary (w^0,γ^0), such that $-e \leq w^0 \leq e$, $y^0 = (-Aw^0+e\gamma^0)_+$, $z^0 = (Bw^0-e\gamma^0)_+$ and arbitrary t^0 such that $\|t^0\|_1 \leq 1$. Having $(w^i,\gamma^i,t^i,y^i,z^i,s^i)$ compute $(w^{i+1},\gamma^{i+1},t^{i+1},y^{i+1},z^{i+1},s^{i+1})$ as a vertex solution of the following linear program:*

$$\min_{(w,t,s,\gamma,y,z)\in R^{3n+1+m+k}} \left\{ \begin{array}{l} e'(y-y^i)+e'(z-z^i) \\ -\alpha((w^i)'(t-t^i)+(w-w^i)'t^i) \end{array} \;\middle|\; \begin{array}{l} -Aw+e\gamma \leq y,\ y \geq 0,\ Bw-e\gamma \leq z,\ z \geq 0, \\ -e \leq w \leq e,\ ,\ -s \leq t \leq s,\ e's \leq 1, \end{array} \right\} \tag{23}$$

Stop when the minimum of (23) is zero.

It follows by an argument similar to that of [3, Algorithm 2.2] that the above algorithm terminates at a vertex that satisfies the minimum principle necessary optimality condition for (21).

4 Summary & Conclusion

By using an exact expression for an arbitrary-norm projection of a point on a plane we are able to formulate precisely the separating plane problem as a minimization of a piecewise-linear convex function on a unit sphere. For $p = 1$ the problem can be solved in polynomial time by solving $2n$ linear programs. For $p = 1$ and $p = \infty$ the problem can be reduced via an exact penalty reformulation to minimizing a bilinear function on separable linear constraints. A successive linearization algorithm for the bilinear problem terminates in a finite number of steps at a stationary point of the bilinear exact penalty formulation. Successful experience in solving difficult NP-hard problems by such a bilinear formulation [3, 16, 17, 4] indicates that such an approach might be viable in the present context as well. An interesting open question that is worth investigating is the following. In [18] it was shown that if we append the additional consistent constraint $\|w\|_\infty \leq \frac{1}{n}$ to (17) for $p \in (1, \infty]$, an associated decision problem with (17) is NP-complete. Whether (17) can also be associated with an NP-complete decision problem without the additional constraint is an open question.

References

[1] E. F. Beckenbach and R. Bellman. *Inequalities.* Springer–Verlag, Berlin, 1961.

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

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

[4] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Clustering via concave minimization. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems -9-*, pages 368–374, Cambridge, MA, 1997. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-03.ps.Z.

[5] R. Bramley and B. Winnicka. Solving linear inequalities in a least square sense. *SIAM Jornal on Scientific Computing*, 17:275–286, 1996.

[6] A. Charnes. Some fundamental theorems of perceptron theory and their geometry. In J. T. Lou and R. H. Wilcox, editors, *Computer and Information Sciences*, pages 67–74, Washington, D.C., 1964. Spartan Books.

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

[8] F. Glover. Improved linear programming models for discriminant analysis. *Decision Sciences*, 21:771–785, 1990.

[9] R. C. Grinold. Mathematical methods for pattern classification. *Management Science*, 19:272–289, 1972.

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

[11] N. Karmarkar. A new polynomial time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

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

[13] O. L. Mangasarian. Multi-surface method of pattern separation. *IEEE Transactions on Information Theory*, IT-14:801–807, 1968.

[14] O. L. Mangasarian. *Nonlinear Programming.* McGraw–Hill, New York, 1969. Reprint: SIAM Classic in Applied Mathematics 10, 1994, Philadelphia.

[15] 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.

[16] O. L. Mangasarian. Misclassification minimization. *Journal of Global Optimization*, 5:309–323, 1994.

[17] O. L. Mangasarian. The linear complementarity problem as a separable bilinear program. *Journal of Global Optimization*, 6:153–161, 1995.

[18] O. L. Mangasarian. Arbitrary-norm separating plane. Technical Report 97-07, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, May 1997. *Operations Research Letters*, submitted. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07.ps.Z.

[19] O. L. Mangasarian and J.-S. Pang. Exact penalty functions for mathematical programs with linear complementarity constraints. *Optimization*, 42:1–8, 1997. //ftp.cs.wisc.edu/math-prog/tech-reports/96-06.ps.Z.

[20] 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.

[21] T. S. Motzkin and I. J. Schoenberg. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:393–404, 1954.

[22] N. J. Nilsson. *Learning Machines.* MIT Press, Cambridge, Massachusetts, 1966.

[23] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

[24] G. W. Stewart. *Introduction to Matrix Computations.* Academic Press, New York, 1973.

[25] Robert J. Vanderbei. *Linear Programming: Foundations and Extensions.* Kluwer Academic Publishers, Hingham, MA, 1997.