[13] B. Kosko, *Neural Networks and Fuzzy Systems*, Prentice Hall, 1992.
[14] H. Ying, W. Siler and J. J. Buckley, "Fuzzy Control Theory: A Nonlinear Case," *Automatica*, vol. 26, no. 3, pp. 513–520, 1990.
[15] R. Langari, "A Nonlinear Formulation of a Class of Fuzzy Linguistic Control Algorithms," *Proc. American Control Conference*, Chicago Illinois, June 24–26 1992, pp. 2273–2278.
[16] G. F. Franklin, J. D. Powell and M. L. Workman, *Digital Control of Dynamic Systems*, Addison Wesley, 1990.

Cone Algorithm: An Extension of the Perceptron Algorithm

S. J. Wan

***Abstract*—The perceptron convergence theorem played an important role in the early development of machine learning. Mathematically, the perceptron learning algorithm is an iterative procedure for finding a separating hyperplane for a finite set of linearly separable vectors, or equivalently, for finding a separating hyperplane for a finite set of linearly contained vectors. In this paper, we show that the perceptron algorithm can be extended to a more general algorithm, called the cone algorithm, for finding a covering cone for a finite set of linearly contained vectors. A proof of the convergence of the cone algorithm is given. The relationship between the cone algorithm and other related algorithms is discussed. The equivalence of the problem of finding a covering cone for a set of linearly contained vectors and the problem of finding a solution cone for a system of homogeneous linear inequalities is established.**

***Index Terms*—Machine learning, perceptron, linearly separable sets, linearly contained sets, covering cones, solution cones, linear inequalities.**

I. INTRODUCTION

A set of vectors is *linearly contained* if all the vectors in the set are distributed on one side of a homogeneous hyperplane. A *covering cone* of a linearly contained set is a circular hypercone which encloses all the vectors in the set. The problem of finding a covering cone of a linearly contained set may arise in some applications such as machine learning [3], [9], [13], computational geometry [10], and stability analysis [2], [4], [7].

The perceptron learning algorithm was developed in the early 1960s for modeling the learning process of a neuron in the human brain [11]. Mathematically, it is an iterative procedure for finding a separating hyperplane for a finite set of linearly separable vectors [3], or equivalently, for finding a separating hyperplane for a finite set of linearly contained vectors [5], [9].

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ be a set of vectors in an n-dimensional Euclidean space R^n. Suppose each vector in X belongs to one of two classes X_1 or X_2. The set X is said to be *linearly separable* [3] if there exists a homogeneous hyperplane:

$$\mathbf{w}^T\mathbf{x} = \sum_{j=1}^{n} w_j x_j = 0 \tag{1}$$

Manuscript received February 28, 1992; revised March 14, 1993 and November 15, 1993.

The author was with the Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada S4S 0A2. He is now with the Imaging Research and Advanced Development, Eastman Kodak Company, Rochester, NY 14650-1907 USA.

IEEE Log Number 9403056.

so that for any $\mathbf{x}_i \in X$,

$$\mathbf{w}^T\mathbf{x}_i = \sum_{j=1}^{n} w_j x_{ij} \begin{cases} > 0 & \text{if } \mathbf{x}_i \in X_1 \\ < 0 & \text{if } \mathbf{x}_i \in X_2 \end{cases} \tag{2}$$

where T denotes the transpose of a vector, and $\mathbf{w}$ is the normal vector of the hyperplane.

The perceptron algorithm finds a separating hyperplane for a set of linearly separable vectors by iterations. It starts with an arbitrary normal vector $\mathbf{w}_0$. The normal vector is then modified according to the following correction rule:

$$\mathbf{w}_{k+1} = \begin{cases} \mathbf{w}_k + \mathbf{x}_i & \text{if } \mathbf{w}_k^T\mathbf{x}_i < 0 \quad \text{and } \mathbf{x}_i \in X_1 \\ \mathbf{w}_k - \mathbf{x}_i & \text{if } \mathbf{w}_k^T\mathbf{x}_i > 0 \quad \text{and } \mathbf{x}_i \in X_2 \end{cases} \tag{3}$$

The well-known perceptron convergence theorem is stated below [3], [5], [9].

Theorem 1: If X is linearly separable, the above procedure will converge to a vector $\mathbf{w}$ satisfying (2) in a finite number of iterations.

A set of vectors $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\}$ is said to be *linearly contained* [9] if all vectors in Y are distributed on one side of a homogeneous hyperplane. In other words, Y is linearly contained if there exists a separating hyperplane defined by (1) satisfying:

$$\mathbf{w}^T\mathbf{y}_i > 0, \qquad \text{for all } \mathbf{y}_i \in Y. \tag{4}$$

A linearly separable set X can be transferred to a linearly contained set Y by changing the sign of the vectors in one class, i.e.,

$$Y = \{\mathbf{x} \mid \mathbf{x} \in X_1\} \cup \{-\mathbf{x} \mid \mathbf{x} \in X_2\}.$$

Fig. 1 depicts a linearly contained set Y transferred from a linearly separable set X.

With a minor modification, the perceptron algorithm can be used for finding a separating hyperplane for a set Y of linearly contained vectors [5]. Starting with an arbitrary normal vector $\mathbf{w}_0$, the normal vector is then modified according to the following correction rule:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mathbf{y}_i, \qquad \text{if } \mathbf{w}_k^T\mathbf{y}_i \le 0 \tag{5}$$

or equivalently,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mathbf{y}_i, \qquad \text{if } \langle \mathbf{w}_k, \mathbf{y}_i \rangle \ge \theta, (\theta = 90^\circ) \tag{6}$$

where $\langle \mathbf{w}_k, \mathbf{y}_i \rangle$ represents the angle between $\mathbf{w}_k$ and $\mathbf{y}_i$. The perceptron convergence theorem in this case is stated as follows [5]:

Theorem 2: If Y is linearly contained, the above procedure will converge to a vector $\mathbf{w}$ satisfying (4) in a finite number of iterations.

In this paper, we show that the perceptron algorithm can be extended to a more general algorithm, called the cone algorithm, for finding a covering cone for a finite set of linearly contained vectors. A proof of the convergence of the cone algorithm is given. The relationship between the cone algorithm and other related algorithms is discussed. The equivalence of the problem of finding a covering cone for a set of linearly contained vectors and the problem of finding a solution cone for a system of homogeneous linear inequalities is established.

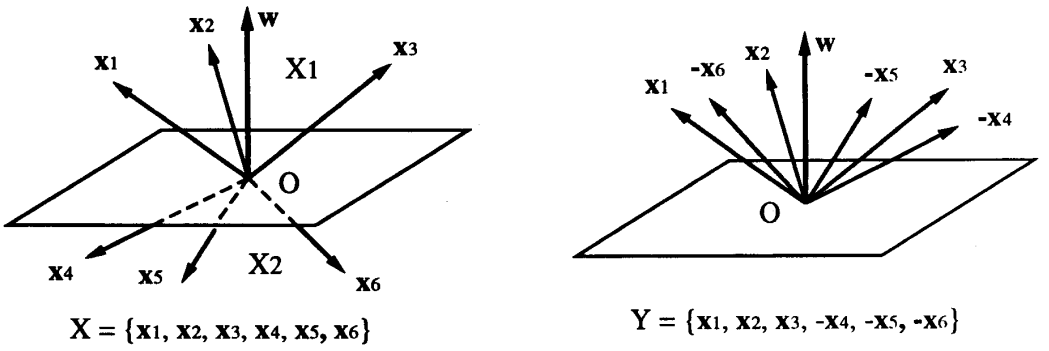

Fig. 1. A linearly contained set Y transferred from a linearly separable set X.

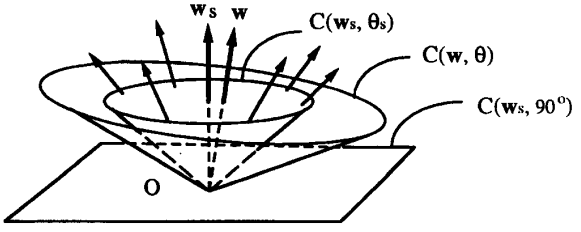

Fig. 2. Three covering cones of Y.

II. THE CONE ALGORITHM

In (6), the perceptron algorithm is expressed as a procedure for adjusting angles between the normal vector $\mathbf{w}_k$ of a hyperplane and the vectors in Y. The normal vector $\mathbf{w}_k$ is rotated towards $\mathbf{y}_i$ if the angle between $\mathbf{w}_k$ and $\mathbf{y}_i$ is larger than or equal to 90°. The perceptron convergence theorem guarantees that this procedure stop in a finite number of iterations. The problem that we are interested in is what would happen if we modify $\theta = 90°$ in (6) to $0° < \theta \leq 90°$. Does the perceptron algorithm still converge in this case? To answer this question, we first introduce the notion of covering cones.

A *hypercone* with axis $\mathbf{w}$ and angle θ in R^n is defined by:

$$C(\mathbf{w}, \theta) = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle \leq \theta, \mathbf{x} \in R^n\}$$

where $\mathbf{w} \neq \mathbf{0}$ and $0° \leq \theta \leq 90°$. A hypercone $C(\mathbf{w}, \theta)$ is said to be a *covering cone* of a set Y of linearly contained vectors if $\langle \mathbf{w}, \mathbf{y}_i \rangle \leq \theta$ for all $\mathbf{y}_i \in Y$. A covering cone of Y with the smallest angle is called the *smallest covering cone*, denoted by $C(\mathbf{w}_s, \theta_s)$. A covering cone of Y with the largest angle ($\theta = 90°$) is a halfspace bounded by the separating hyperplane $\mathbf{w}^T\mathbf{x} = 0$. Fig. 2 depicts three covering cones of Y.

By modifying $\theta = 90°$ in (6) to $0° < \theta \leq 90°$, the perceptron algorithm becomes a more general algorithm, called the *cone algorithm*, stated as follows:

The cone algorithm: Starting with an arbitrary axis $\mathbf{w}_0$, if a vector $\mathbf{y}_i$ in Y is not enclosed by the hypercone $C(\mathbf{w}_k, \theta)$, the axis $\mathbf{w}_k$ is modified by:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mathbf{y}_i, \qquad \text{if } \langle \mathbf{w}_k, \mathbf{y}_i \rangle \geq \theta \quad (0° < \theta \leq 90°). \tag{7}$$

The convergence of the cone algorithm is stated below.

Theorem 3 (the cone algorithm convergence theorem): Let Y be a set of linearly contained vectors and $C(\mathbf{w}_s, \theta_s)$ be the smallest covering cone of Y. If $\theta_s < \theta$, then each correction given by (7) will bring $\mathbf{w}_k$ closer to $\mathbf{w}_s$ when k is large enough, namely,

$$\langle \mathbf{w}_{k+1}, \mathbf{w}_s \rangle < \langle \mathbf{w}_k, \mathbf{w}_s \rangle, \qquad \text{if } k > K_0. \tag{8}$$

Proof: See the Appendix.

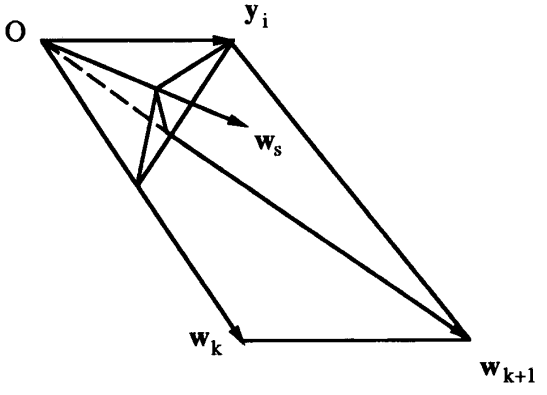

Fig. 3. An illustration of the convergence of the cone algorithm.

The convergence of the cone algorithm may be illustrated by Fig. 3. The length of the normal vector $\mathbf{w}_k$ can become arbitrarily large with the increase of k, but the length of each vector $\mathbf{y}_i$ in Y is fixed. When k is large enough, if $\langle \mathbf{w}_k, \mathbf{y}_i \rangle \geq \theta$, then $\mathbf{w}_k$ is rotated towards $\mathbf{y}_i$ for a small amount, which brings $\mathbf{w}_{k+1}$ closer to $\mathbf{w}_s$.

The convergence speed of the cone algorithm may be improved by introducing a proper coefficient ρ_k in the correction rule, namely,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \rho_k \mathbf{y}_i \qquad \text{if } \langle \mathbf{w}_k, \mathbf{y}_i \rangle, > \theta, (0° < \theta \leq 90°) \tag{9}$$

where ρ_k controls the rotation angle of $\mathbf{w}_k$ towards $\mathbf{y}_i$.

In what follows, we discuss the relationship between the cone algorithm and other related algorithms, and other related issues.

A. The Cone Algorithm Versus the Perceptron Algorithm

The only difference between the cone algorithm and the perceptron algorithm (the version for linearly contained sets) is the condition for modifying the vector $\mathbf{w}_k$. In the perceptron algorithm (refer to (6)), if a vector $\mathbf{y}_i \in Y$ is not located in the halfspace defined by the

hyperplane, then the normal vector $\mathbf{w}_k$ of the hyperplane is modified towards this vector. The perceptron algorithm stops when a separating hyperplane of Y is obtained. In the cone algorithm (refer to (7)), if a vector $\mathbf{y}_i \in Y$ is not located in the covering cone $C(\mathbf{w}_k, \theta)$, then the axis $\mathbf{w}_k$ of the covering cone is modified towards this vector. The cone algorithm stops when a covering cone of Y is obtained. From a geometric point of view, the perceptron algorithm can be viewed as a procedure of adjusting the normal vector of a hyperplane so that all the vectors in Y are distributed on one side of the hyperplane, while the cone algorithm adjusts the axis of a hypercone so that all the vectors in Y are enclosed by the hypercone. Because a hyperplane is a special case of a hypercone ($\theta = 90°$), the perceptron algorithm is a special case of the cone algorithm.

B. The Cone Algorithm Versus Other Related Algorithms

There are a number of gradient descent algorithms such as the relaxation procedures [1], [6] and variable increment procedures [3] designed for solving a system of *inhomogeneous* linear inequalities:

$$\mathbf{w}^T\mathbf{y}_i > b_i, \qquad i = 1, 2, \ldots, m, \tag{10}$$

where $b_1, b_2, \ldots, b_m$ are positive constants. These algorithms can be written in the following form:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \rho_k\mathbf{y}_i, \qquad \text{if } \mathbf{w}_k^T\mathbf{y}_i \le b_i, \tag{11}$$

where ρ_k is a parameter. Compared with these algorithms, the perceptron algorithm can be viewed as an approach for finding a solution vector $\mathbf{w}$ for a system of *homogeneous* linear inequalities:

$$\mathbf{w}^T\mathbf{y}_i > 0, \qquad i = 1, 2, \ldots, m. \tag{12}$$

It is clear that any solution of (10) is also a solution of (12), but the converse may not necessarily be true. Thus, the algorithms described in (11) are more general than the perceptron algorithm.

On the other hand, the cone algorithm is designed for finding a solution vector $\mathbf{w}$ for the following *nonlinear* inequalities:

$$\mathbf{w}^T\mathbf{y}_i \ge \|\mathbf{w}\|\|\mathbf{y}_i\| \cos\theta, \qquad i = 1, 2, \ldots, m. \tag{13}$$

If a solution $\mathbf{w}$ to (13) is found, a solution $\mathbf{w}'$ to (10) can be constructed based on $\mathbf{w}$ as follows:

$$\mathbf{w}' = \frac{b}{a}\mathbf{w}$$

where

$$b > \operatorname*{Max}_i b_i, \qquad a = \operatorname*{Min}_i \left(\mathbf{w}^T\mathbf{y}_i\right).$$

Conversely, given a solution $\mathbf{w}'$ to (10), one may not be able to construct a solution to (13) based on $\mathbf{w}'$. In other words, the cone algorithm can solve (10), but the algorithms described in (11) cannot solve (13). The cone algorithm is more general than these algorithms.

C. The Smallest Covering Cone

By decreasing the angle θ step by step, the cone algorithm can find a series of covering cones of Y approaching the smallest covering cone of Y. This may be done by setting the initial angle $\theta = 90°$. The angle is then decreased by a fixed quantity δ each time when the cone algorithm converges. The parameter δ can be set in advance according to the accuracy required. Because the cone algorithm will not converge when θ becomes smaller than θ_s, a terminating condition (for instance, a fixed number of passes over Y) should be added in the procedure to prevent an infinite loop.

To test the above procedure, we randomly created 1000 20-dimensional vectors in a hypercone with $\theta = 70°$. Half of these vectors are distributed on the boundary of the hypercone. This hypercone can be treated approximately as the smallest covering cone $C(\mathbf{w}_s, \theta_s)$ of the 1000 vectors. The cone algorithm is applied to this set of linearly contained vectors with the initial angle setting $\theta = 90°$. Decreasing the angle by $\delta = 5°$ at each step, five covering cones were obtained. Table I lists the angles between $\mathbf{w}_s$ and the axes of the five covering cones. It can be seen that when θ decreases, the axis $\mathbf{w}$ of the covering cone approaches $\mathbf{w}_s$.

TABLE I
THE CLOSENESS BETWEEN $C(\mathbf{w}_s, \theta_s)$ AND $C(\mathbf{w}, \theta)$.

Step	θ	θ_s	$\langle \mathbf{w}, \mathbf{w}_s \rangle$
1	90°	70°	30°15′
2	85°	70°	22°30′
3	80°	70°	13°54′
4	75°	70°	9°30′
5	70°	70°	0°45′

D. The Largest Solution Cone

The problem of finding a covering cone of $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\}$ is closely related to the one of finding a solution cone for a system of homogeneous linear inequalities:

$$\mathbf{w}^T\mathbf{y}_i \ge 0, \qquad i = 1, 2, \ldots, m. \tag{14}$$

If Y is linearly contained, there exist many solutions to (14). Geometrically, all the solutions can be obtained in the following way. Each vector $\mathbf{y}_i$ defines a halfspace bounded by a homogeneous hyperplane with $\mathbf{y}_i$ as its normal vector. The intersection of the m halfspaces forms the solution region of (14), denoted by S. It can be shown that the solution region S is a convex polyhedral cone in R^n [8, 12]. A *solution cone* of (14) is a hypercone enclosed by the solution region. A solution cone is said to be the *largest solution cone* if its angle is the largest among all the solution cones of (14).

Theorem 4 Let Y be a set of linearly contained vectors. $C(\mathbf{w}, \theta)$ is a covering cone of Y if and only if $C(\mathbf{w}, 90° - \theta)$ is a solution cone of (14).

Proof: See Appendix.

Theorem 4 indicates that there is a one-to-one correspondence between a covering cone of Y and a solution cone of (14). In a special case $\theta = \theta_s$, it states that $C(\mathbf{w}_s, \theta_s)$ is the smallest covering cone of Y if and only if $C(\mathbf{w}_s, 90° - \theta_s)$ is the largest solution cone of (14). Fig. 4 describes the relationship between the smallest covering cone of Y and the largest solution cone of (14). Note that the largest solution cone is enclosed by the solution region S.

From a stability point of view, the solution $\mathbf{w}_s$ is superior to any other solution in the solution region S because it can tolerate a maximum disturbance from all directions. In a noisy communication channel, the received signal $\mathbf{y}_i'$ may be different from the transmitted signal $\mathbf{y}_i$. However, if $\langle \mathbf{y}_i, \mathbf{y}_i' \rangle \le 90° - \theta_s$, $\mathbf{w}_s$ will remain to be a solution of the system. Note that $90° - \theta_s$ is the maximum disturbance angle that the system can tolerate.

III. CONCLUSION

It is shown that the perceptron algorithm can be extended to a more general algorithm, namely the cone algorithm, for finding a covering cone of a linearly contained set. It is also shown that there

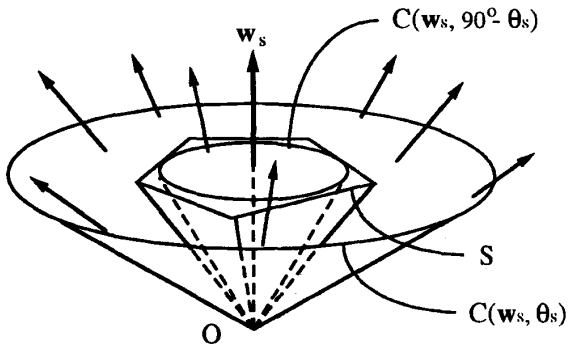

Fig. 4. The smallest covering cone and the largest solution cone.

is a one-to-one correspondence between a covering cone of a linearly contained set and a solution cone of a system of homogeneous linear inequalities. Compared with other gradient descent algorithms, the cone algorithm is a more general method for solving a system of inhomogeneous linear inequalities.

The main weakness of the cone algorithm is that it does not provide a bound on the number of iterations required for the algorithm to converge. This is an inherent weakness of the gradient descent type of algorithms, including the perceptron algorithm.

APPENDIX

Theorem 3: (the cone algorithm convergence theorem:) Let Y be a set of linearly contained vectors and $C(\mathbf{w}_s, \theta_s)$ be the smallest covering cone of Y. If $\theta_s < \theta$, then each correction given by (7) will bring $\mathbf{w}_k$ closer to $\mathbf{w}_s$ when k is large enough, namely,

$$\langle \mathbf{w}_{k+1}, \mathbf{w}_s \rangle < \langle \mathbf{w}_k, \mathbf{w}_s \rangle, \text{ if } k > K_0. \tag{8}$$

Proof: For convenience, we introduce the following unit vectors:

$$\hat{\mathbf{w}}_s = \frac{\mathbf{w}_s}{\|\mathbf{w}_s\|}, \qquad \hat{\mathbf{w}}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}, \qquad \hat{\mathbf{w}}_{k+1} = \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_{k+1}\|}$$

and notations:

$$\alpha = \cos\theta_s - \cos\theta, \qquad \beta = \min_i \|\mathbf{y}_i\|, \qquad \gamma = \max_i \|\mathbf{y}_i\|,$$
$$\varepsilon = \frac{\|\mathbf{w}_{k+1}\|}{\|\mathbf{w}_k\|} - 1.$$

From (7),

$$\hat{\mathbf{w}}_s - \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_k\|} = \hat{\mathbf{w}}_s - \frac{\mathbf{w}_k + \mathbf{y}_i}{\|\mathbf{w}_k\|} = (\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k) - \frac{\mathbf{y}_i}{\|\mathbf{w}_k\|}.$$

Hence,

$$\begin{aligned}\left\|\hat{\mathbf{w}}_s - \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_k\|}\right\|^2 &= \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2 - (\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k)^T \frac{2\mathbf{y}_i}{\|\mathbf{w}_k\|} + \frac{\|\mathbf{y}_i\|^2}{\|\mathbf{w}_k\|^2} \\ &= \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2 - \frac{2}{\|\mathbf{w}_k\|}\left(\hat{\mathbf{w}}_s^T\mathbf{y}_i - \hat{\mathbf{w}}_k^T\mathbf{y}_i\right) + \frac{\|\mathbf{y}_i\|^2}{\|\mathbf{w}_k\|^2}. \end{aligned} \tag{15}$$

Because $\mathbf{y}_i \in C(\mathbf{w}_s, \theta_s)$ and $\mathbf{y}_i \notin C(\mathbf{w}_k, \theta)$, we have

$$\hat{\mathbf{w}}_s^T\mathbf{y}_i \geq \|\mathbf{y}_i\|\cos\theta_s, \qquad \hat{\mathbf{w}}_k^T\mathbf{y}_i < \|\mathbf{y}_i\|\cos\theta. \tag{16}$$

Thus,

$$\hat{\mathbf{w}}_s^T\mathbf{y}_i - \hat{\mathbf{w}}_k^T\mathbf{y}_i > (\cos\theta_s - \cos\theta)\|\mathbf{y}_i\| = \alpha\|\mathbf{y}_i\|. \tag{17}$$

Combining (15) and (17) yields:

$$\begin{aligned}\left\|\hat{\mathbf{w}}_s - \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_k\|}\right\|^2 &< \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2 - \frac{2\alpha\|\mathbf{y}_i\|}{\|\mathbf{w}_k\|} + \frac{\|\mathbf{y}_i\|^2}{\|\mathbf{w}_k\|^2} \\ &= \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2 - \frac{\|\mathbf{y}_i\|}{\|\mathbf{w}_k\|^2}(2\alpha\|\mathbf{w}_k\| - \|\mathbf{y}_i\|) \\ &\leq \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2 - \frac{\beta}{\|\mathbf{w}_k\|^2}(2\alpha\|\mathbf{w}_k\| - \gamma). \end{aligned} \tag{18}$$

From (7) and (16),

$$\begin{aligned}\hat{\mathbf{w}}_s^T\mathbf{w}_{k+1} &= \hat{\mathbf{w}}_s^T\mathbf{w}_k + \hat{\mathbf{w}}_s^T\mathbf{y}_i \\ &\geq \hat{\mathbf{w}}_s^T\mathbf{w}_k + \|\mathbf{y}_i\|\cos\theta_s \\ &\geq \hat{\mathbf{w}}_s^T\mathbf{w}_k + \beta\cos\theta_s. \end{aligned} \tag{19}$$

Applying induction to (19), we obtain

$$\hat{\mathbf{w}}_s^T\mathbf{w}_k \geq k\beta\cos\theta_s + \lambda, \qquad \text{for } k > 0$$

where $\lambda = \hat{\mathbf{w}}_s^T\mathbf{w}_0$ is a finite real number. Since Y is linearly contained, β and $\cos\theta_s$ are greater than zero. By the assumption $\theta_s < \theta$, we have $\alpha > 0$. Thus, $\alpha\,\beta\cos\theta_s > 0$. If we choose

$$k > K_0 = \left\lceil \frac{\gamma - 2\alpha\lambda}{2\alpha\,\beta\cos\theta_s} \right\rceil$$

then,

$$\hat{\mathbf{w}}_s^T\mathbf{w}_k \geq \frac{\beta(\gamma - 2\alpha\lambda)\cos\theta_s}{2\alpha\,\beta\cos\theta_s} + \lambda = \frac{\gamma}{2\alpha}.$$

Because $\|\mathbf{w}_k\| \geq \hat{\mathbf{w}}_s^T\mathbf{w}_k$, it follows that

$$\|\mathbf{w}_k\| \geq \frac{\gamma}{2\alpha}, \qquad \text{or } 2\alpha\|\mathbf{w}_k\| - \gamma \geq 0. \tag{20}$$

Combining (18) and (20), we obtain the following inequality:

$$\left\|\hat{\mathbf{w}}_s - \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_k\|}\right\|^2 < \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2, \qquad \text{if } k > K_0. \tag{21}$$

Since for each correction either $\|\mathbf{w}_{k+1}\| > \|\mathbf{w}_k\|$ or $\|\mathbf{w}_{k+1}\| \leq \|\mathbf{w}_k\|$ holds, we consider each case separately.

Case 1: Assuming $\|\mathbf{w}_{k+1}\| > \|\mathbf{w}_k\|$.

In this case, we have

$$\varepsilon = \frac{\|\mathbf{w}_{k+1}\|}{\|\mathbf{w}_k\|} - 1 > 0.$$

Thus,

$$\begin{aligned}\left\|\hat{\mathbf{w}}_s - \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_k\|}\right\|^2 &= \left\|\hat{\mathbf{w}}_s - \frac{\|\mathbf{w}_{k+1}\|}{\|\mathbf{w}_k\|}\hat{\mathbf{w}}_{k+1}\right\|^2 \\ &= \|(\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}) - \varepsilon\hat{\mathbf{w}}_{k+1}\|^2 \\ &= \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}\|^2 \\ &\quad - 2\varepsilon\hat{\mathbf{w}}_{k+1}^T(\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}) + \varepsilon^2 \\ &> \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}\|^2 - 2\varepsilon\hat{\mathbf{w}}_{k+1}^T(\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}) \\ &= \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}\|^2 + 2\varepsilon(1 - \cos\langle\hat{\mathbf{w}}_s, \hat{\mathbf{w}}_{k+1}\rangle) \\ &\geq \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}\|^2. \end{aligned} \tag{22}$$

Combining (21) and (22) yields:

$$\|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_{k+1}\|^2 < \|\hat{\mathbf{w}}_s - \hat{\mathbf{w}}_k\|^2, \qquad \text{if } k > K_0. \tag{23}$$

Because $\hat{\mathbf{w}}_s$, $\hat{\mathbf{w}}_k$ and $\hat{\mathbf{w}}_{k+1}$ are unit vectors, it follows from (23) that

$$\langle\hat{\mathbf{w}}_s, \hat{\mathbf{w}}_{k+1}\rangle < \langle\hat{\mathbf{w}}_s, \hat{\mathbf{w}}_k\rangle, \qquad \text{if } k > K_0$$

which is the desired result.

Case 2: Assuming $\|\mathbf{w}_{k+1}\| \leq \|\mathbf{w}_k\|$.
From (19), $\widehat{\mathbf{w}}_s^T \mathbf{w}_{k+1} > \widehat{\mathbf{w}}_s^T \mathbf{w}_k$. Thus,

$$\widehat{\mathbf{w}}_s^T \widehat{\mathbf{w}}_{k+1} > \frac{\widehat{\mathbf{w}}_s^T \mathbf{w}_k}{\|\mathbf{w}_{k+1}\|} \geq \frac{\widehat{\mathbf{w}}_s^T \mathbf{w}_k}{\|\mathbf{w}_k\|} = \widehat{\mathbf{w}}_s^T \widehat{\mathbf{w}}_k,$$

or equivalently,

$$\langle \widehat{\mathbf{w}}_s, \widehat{\mathbf{w}}_{k+1} \rangle < \langle \widehat{\mathbf{w}}_s, \widehat{\mathbf{w}}_k \rangle.$$

This completes the proof. □

The following lemma is needed for the proof of Theorem 4.

Lemma 1: For any vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in R^n$, *the following inequality holds:*

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle. \tag{24}$$

Proof: Because the length of the vectors in (23) is immaterial, It is assumed, without loss of generality, that $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are unit vectors.

We first construct a vector $\mathbf{y}'$ based on $\mathbf{x}, \mathbf{y}, \mathbf{z}$ as follows:

$$\mathbf{y}' = a\mathbf{z} - b\mathbf{x} \tag{25}$$

where

$$a = \frac{\sin(\beta + \gamma)}{\sin \beta}, \quad b = \frac{\sin \gamma}{\sin \beta}, \quad \beta = \langle \mathbf{x}, \mathbf{z} \rangle, \quad \gamma = \langle \mathbf{z}, \mathbf{y} \rangle.$$

It then follows that

$$\begin{aligned} \|\mathbf{y}'\|^2 &= a^2 - 2ab\cos\beta + b^2 = 1 \\ \mathbf{z}^T\mathbf{y}' &= a - b\cos\beta = \cos\gamma \\ \mathbf{x}^T\mathbf{y}' &= a\cos\beta - b = \cos(\beta + \gamma) \end{aligned}$$

This means that $\mathbf{y}'$ is a unit vector,

$$\langle \mathbf{z}, \mathbf{y}' \rangle = \gamma = \langle \mathbf{z}, \mathbf{y} \rangle \tag{26}$$

and

$$\langle \mathbf{x}, \mathbf{y}' \rangle = \beta + \gamma = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle. \tag{27}$$

Next, we introduce another vector $\mathbf{z}'$ defined by:

$$\mathbf{z}' = d\mathbf{z}, \qquad d = \frac{\cos\left(\frac{\beta+\gamma}{2}\right)}{\cos\left(\frac{\beta-\gamma}{2}\right)}.$$

It can be shown that

$$\|\mathbf{x} - \mathbf{z}'\| + \|\mathbf{z}' - \mathbf{y}'\| = \|\mathbf{x} - \mathbf{y}'\|. \tag{28}$$

Note that $\mathbf{y}$ and $\mathbf{y}'$ are unit vectors. It follows from (26) that

$$\|\mathbf{z}' - \mathbf{y}\|^2 = \|\mathbf{z}'\|^2 - 2\|\mathbf{z}'\|\cos\gamma + 1 = \|\mathbf{z}' - \mathbf{y}'\|^2. \tag{29}$$

From the triangle inequality and (28), (29),

$$\|\mathbf{x}-\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{z}'\| + \|\mathbf{z}'-\mathbf{y}\| = \|\mathbf{x}-\mathbf{z}'\| + \|\mathbf{z}'-\mathbf{y}'\| = \|\mathbf{x}-\mathbf{y}'\|. \tag{30}$$

Because $\mathbf{x}, \mathbf{y}, \mathbf{y}'$ are unit vectors,

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \langle \mathbf{x}, \mathbf{y}' \rangle. \tag{31}$$

Substituting (27) into (31) yields (24). □

Theorem 4: Let Y *be a set of linearly contained vectors.* $C(\mathbf{w}, \theta)$ *is a covering cone of* Y *if and only if* $C(\mathbf{w}, 90^\circ - \theta)$ *is a solution cone of* (14).

Proof: We first show that $C(\mathbf{w}, 90^\circ - \theta)$ is a solution cone of (14) if $C(\mathbf{w}, \theta)$ is a covering cone of Y.

Because $C(\mathbf{w}, \theta)$ is a covering cone of Y, $\langle \mathbf{w}, \mathbf{y}_i \rangle \leq \theta$ for all $\mathbf{y}_i \in Y$. For any $\mathbf{x} \in C(\mathbf{w}, 90^\circ - \theta)$, we have $\langle \mathbf{x}, \mathbf{w} \rangle \leq 90^\circ - \theta$. It follows from Lemma 1 that for any $\mathbf{y}_i \in Y$,

$$\langle \mathbf{x}, \mathbf{y}_i \rangle \leq \langle \mathbf{x}, \mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{y}_i \rangle \leq (90^\circ - \theta) + \theta = 90^\circ,$$

or

$$\mathbf{x}^T\mathbf{y}_i \geq 0, \qquad i = 1, 2, \ldots, m. \tag{32}$$

This means that $\mathbf{x}$ is a solution of (14). Because (32) holds for any $\mathbf{x} \in C(\mathbf{w}, 90^\circ - \theta)$, we conclude that $C(\mathbf{w}, 90^\circ - \theta)$ is a solution cone of (14).

Conversely, suppose $C(\mathbf{w}, 90^\circ - \theta)$ is a solution cone of (14). We show that $C(\mathbf{w}, \theta)$ is a covering cone of Y.

For any $\mathbf{y}_i \in Y$, we can construct a vector $\mathbf{x}$ as follows:

$$\mathbf{x} = a\frac{\mathbf{w}}{\|\mathbf{w}\|} - b\frac{\mathbf{y}_i}{\|\mathbf{y}_i\|} \tag{33}$$

where

$$a = \frac{\sin(\beta + \gamma)}{\sin \beta}, \quad b = \frac{\sin \gamma}{\sin \beta}, \quad \beta = \langle \mathbf{w}, \mathbf{y}_i \rangle, \quad \gamma = 90^\circ - \theta.$$

We have,

$$\|\mathbf{x}\|^2 = a^2 - 2ab\cos\beta + b^2 = 1, \tag{34}$$

and,

$$\mathbf{x}^T\mathbf{w} = a - b\cos\beta = \cos\gamma$$

or

$$\langle \mathbf{x}, \mathbf{w} \rangle = \gamma = 90^\circ - \theta. \tag{35}$$

Eq. (35) indicates $\mathbf{x} \in C(\mathbf{w}, 90^\circ - \theta)$. Therefore,

$$\langle \mathbf{x}, \mathbf{y}_i \rangle \leq 90^\circ, \text{ for any } \mathbf{y}_i \in Y. \tag{36}$$

On the other hand, from (33),

$$\mathbf{x}^T\frac{\mathbf{y}_i}{\|\mathbf{y}_i\|} = a\cos\beta - b = \cos(\beta + \gamma)$$

which means $\langle \mathbf{x}, \mathbf{y}_i \rangle = \beta + \gamma$, or

$$\langle \mathbf{x}, \mathbf{y}_i \rangle = \langle \mathbf{w}, \mathbf{y}_i \rangle + \langle \mathbf{x}, \mathbf{w} \rangle. \tag{37}$$

Combining (35)–(37) yields

$$\langle \mathbf{w}, \mathbf{y}_i \rangle = \langle \mathbf{x}, \mathbf{y}_i \rangle - \langle \mathbf{x}, \mathbf{w} \rangle \leq 90^\circ - (90^\circ - \theta) = \theta, \quad \text{for all } \mathbf{y}_i \in Y. \tag{38}$$

Thus, we conclude that $C(\mathbf{w}, \theta)$ is a covering cone of Y. □

REFERENCES

[1] S. Agmon, "The relaxation method for linear inequalities," *Canadian Journal of Mathematics*, vol. 6, pp. 382–392, 1954.

[2] A. Deif, *Sensitivity Analysis in Linear Systems*. Berlin: Springer-Verlag, 1986.

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

[4] C. H. Mays, "Effects of adaptation parameters on convergence time and tolerance for adaptive threshold elements," *IEEE Trans. Electr. Comput., EC-13*, pp. 465–468, 1964.

[5] M. Minsky and S. Papert, *Perceptrons*. Cambridge, MA: MIT Press, 1988 (expanded edition).

[6] T. S. Motzkin and I. J. Schoenberg, "The relaxation method for linear inequalities," *Canadian Journal of Mathematics*, vol. 6, pp. 392–404, 1954.
[7] S. Muroga, *Threshold Logic and Its Applications*. John Wiley and Sons, 1971.
[8] K. Murty, *Linear and Combinatorial Programming*. New York: John Wiley and Sons, 1976.
[9] N. J. Nilsson, *Learning Machines*. McGraw-Hill, 1965.
[10] F. P. Preparate and M. I. Shamos, *Computational Geometry*. New York: Springer-Verlag, 1985.
[11] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, D.C: Spartan Books, 1962.
[12] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimension I*. Berlin: Springer-Verlag, 1970.