4.3 The Perceptron Convergence Theorem

For the development of the error-correction learning algorithm for a single-layer perceptron, we find it more convenient to work with the modified signal-flow graph model of Fig. 4.4. In this second model, which is equivalent to that of Fig. 4.2, the threshold $\theta(n)$ is treated as a synaptic weight connected to a fixed input equal to -1. We may thus define the $(p + 1)$-by-1 input vector

$$\mathbf{x}(n) = [-1, x_1(n), x_2(n), \ldots, x_p(n)]^T \tag{4.3}$$

Correspondingly, we define the $(p + 1)$-by-1 weight vector

$$\mathbf{w}(n) = [\theta(n), w_1(n), w_2(n), \ldots, w_p(n)]^T \tag{4.4}$$

Accordingly, the linear combiner output is written in the compact form

$$v(n) = \mathbf{w}^T(n)\mathbf{x}(n) \tag{4.5}$$

For fixed n, the equation $\mathbf{w}^T\mathbf{x} = 0$, plotted in a p-dimensional space with coordinates x_1, x_2, $\ldots$, x_p, defines a hyperplane as the decision surface between two different classes of inputs.

Suppose then the input variables of the single-layer perceptron originate from two *linearly separable classes* that fall on the opposite sides of some hyperplane. Let X_1 be the subset of training vectors $\mathbf{x}_1(1)$, $\mathbf{x}_1(2)$, $\ldots$ that belong to class $\mathscr{C}_1$, and let X_2 be the

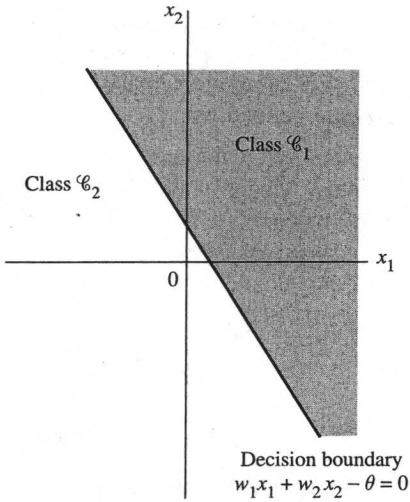

FIGURE 4.3 Illustrating linear reparability for a two-dimensional, two-class pattern-classification problem.

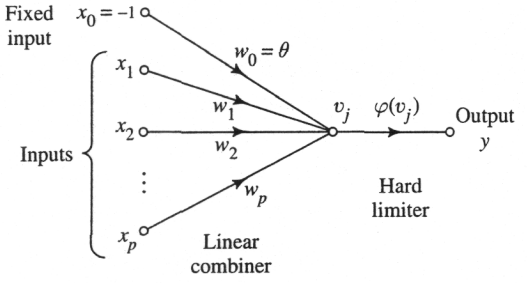

FIGURE 4.4 Equivalent signal-flow graph of the perceptron.

subset of training vectors $\mathbf{x}_2(1)$, $\mathbf{x}_2(2)$, . . . that belong to class $\mathscr{C}_2$. The union of X_1 and X_2 is the complete training set X. Given the sets of vectors X_1 and X_2 to train the classifier, the training process involves the adjustment of the weight vector $\mathbf{w}$ in such a way that the two classes $\mathscr{C}_1$ and $\mathscr{C}_2$ are separable. These two classes are said to be *linearly separable* if a realizable setting of the weight vector $\mathbf{w}$ exists. Conversely, if the two classes $\mathscr{C}_1$ and $\mathscr{C}_2$ are known to be linearly separable, then there exists a weight vector $\mathbf{w}$ such that we may state

$$\mathbf{w}^T\mathbf{x} \geq 0 \text{ for every input vector } \mathbf{x} \text{ belonging to class } \mathscr{C}_1$$

and (4.6)

$$\mathbf{w}^T\mathbf{x} < 0 \text{ for every input vector } \mathbf{x} \text{ belonging to class } \mathscr{C}_2$$

Given the subsets of training vectors X_1 and X_2, the training problem for the elementary perceptron is then to find a weight vector $\mathbf{w}$ such that the two inequalities of Eq. (4.6) are satisfied.

The algorithm for adapting the weight vector of the elementary perceptron may now be formulated as follows:

1. If the nth member of the training vector, $\mathbf{x}(n)$, is correctly classified by the weight vector $\mathbf{w}(n)$ computed at the nth iteration of the algorithm, no correction is made to the weight vector of the perceptron, as shown by

   $$\mathbf{w}(n+1) = \mathbf{w}(n) \quad \text{if } \mathbf{w}^T(n)\mathbf{x}(n) \geq 0 \text{ and } \mathbf{x}(n) \text{ belongs to class } \mathscr{C}_1$$

   and (4.7)

   $$\mathbf{w}(n+1) = \mathbf{w}(n) \quad \text{if } \mathbf{w}^T(n)\mathbf{x}(n) < 0 \text{ and } \mathbf{x}(n) \text{ belongs to class } \mathscr{C}_2$$

2. Otherwise, the weight vector of the perceptron is updated in accordance with the rule

   $$\mathbf{w}(n+1) = \mathbf{w}(n) - \eta(n)\mathbf{x}(n) \quad \text{if } \mathbf{w}^T(n)\mathbf{x}(n) \geq 0 \text{ and } \mathbf{x}(n) \text{ belongs to class } \mathscr{C}_2$$

   and (4.8)

   $$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta(n)\mathbf{x}(n) \quad \text{if } \mathbf{w}^T(n)\mathbf{x}(n) < 0 \text{ and } \mathbf{x}(n) \text{ belongs to class } \mathscr{C}_1$$

   where the *learning-rate parameter* $\eta(n)$ controls the adjustment applied to the weight vector at iteration n.

If $\eta(n) = \eta > 0$, where η is a constant independent of the iteration number n, we have a *fixed increment adaptation rule* for the perceptron.

In the sequel, we first prove the convergence of a fixed increment adaptation rule for which $\eta = 1$. Clearly, the value of η is unimportant, so long as it is positive. A value

of $\eta \neq 1$ merely scales the pattern vectors without affecting their separability. The case of a variable $\eta(n)$ is considered later in the section.

The proof is presented for the initial condition $\mathbf{w}(0) = \mathbf{0}$. Suppose that $\mathbf{w}^T(n)\mathbf{x}(n) < 0$ for $n = 1, 2, \ldots$ and an input vector $\mathbf{x}(n)$ that belongs to the subset X_1. That is, the perceptron incorrectly classifies the vectors $\mathbf{x}(1), \mathbf{x}(2), \ldots$, since the second condition of Eq. (4.6) is violated. Then, with the constant $\eta(n) = 1$, we may use the second line of Eq. (4.8) to write

$$\mathbf{w}(n + 1) = \mathbf{w}(n) + \mathbf{x}(n) \qquad \text{for } \mathbf{x}(n) \text{ belonging to class } \mathscr{C}_1 \tag{4.9}$$

Given the initial condition $\mathbf{w}(0) = \mathbf{0}$, we may iteratively solve this equation for $\mathbf{w}(n + 1)$, obtaining the result

$$\mathbf{w}(n + 1) = \mathbf{x}(1) + \mathbf{x}(2) + \cdots + \mathbf{x}(n) \tag{4.10}$$

Since the classes $\mathscr{C}_1$ and $\mathscr{C}_2$ are assumed to be linearly separable, there exists a solution $\mathbf{w}_0$ for which $\mathbf{w}^T\mathbf{x}(n) > 0$ for the vectors $\mathbf{x}(1), \ldots, \mathbf{x}(n)$ belonging to the subset X_1. For a fixed solution $\mathbf{w}_0$, we may then define a positive number α by the relation

$$\alpha = \min_{\mathbf{x}(n) \in X_1} \mathbf{w}_0^T\mathbf{x}(n) \tag{4.11}$$

where $\mathbf{x}(n) \in X_1$ stands for "$\mathbf{x}(n)$ belongs to subset X_1." Hence, multiplying both sides of Eq. (4.10) by the row vector $\mathbf{w}_0^T$, we get

$$\mathbf{w}_0^T\mathbf{w}(n + 1) = \mathbf{w}_0^T\mathbf{x}(1) + \mathbf{w}_0^T\mathbf{x}(2) + \cdots + \mathbf{w}_0^T\mathbf{x}(n) \tag{4.12}$$

Accordingly, in light of the definition given in Eq. (4.11), we have

$$\mathbf{w}_0^T\mathbf{w}(n + 1) \geq n\alpha \tag{4.13}$$

Next, we make use of an inequality known as the Cauchy–Schwarz inequality. Given two vectors $\mathbf{w}_0$ and $\mathbf{w}(n + 1)$, the *Cauchy–Schwarz inequality* states that

$$\|\mathbf{w}_0\|^2 \|\mathbf{w}(n + 1)\|^2 \geq [\mathbf{w}_0^T\mathbf{w}(n + 1)]^2 \tag{4.14}$$

where $\|\cdot\|$ denotes the Euclidean norm of the enclosed argument vector, and the inner product $\mathbf{w}_0^T\mathbf{w}(n + 1)$ is a scalar quantity. We now note from Eq. (4.13) that $[\mathbf{w}_0^T\mathbf{w}(n + 1)]^2$ is equal to or greater than $n^2\alpha^2$. Moreover, from Eq. (4.14) we note that $\|\mathbf{w}_0\|^2\|\mathbf{w}(n + 1)\|^2$ is equal to or greater than $[\mathbf{w}_0^T\mathbf{w}(n + 1)]^2$. It follows therefore that

$$\|\mathbf{w}_0\|^2 \|\mathbf{w}(n + 1)\|^2 \geq n^2\alpha^2$$

or, equivalently,

$$\|\mathbf{w}(n + 1)\|^2 \geq \frac{n^2\alpha^2}{\|\mathbf{w}_0\|^2} \tag{4.15}$$

Next, we follow another development route. In particular, we rewrite Eq. (4.9) in the form

$$\mathbf{w}(k + 1) = \mathbf{w}(k) + \mathbf{x}(k) \qquad \text{for } k = 1, \ldots, n \quad \text{and} \quad \mathbf{x}(k) \in X_1 \tag{4.16}$$

Hence, taking the squared Euclidean norm of both sides of Eq. (4.16), we get

$$\|\mathbf{w}(k + 1)\|^2 = \|\mathbf{w}(k)\|^2 + \|\mathbf{x}(k)\|^2 + 2\mathbf{w}^T(k)\mathbf{x}(k) \tag{4.17}$$

But, under the assumption that the perceptron incorrectly classifies an input vector $\mathbf{x}(k)$ belonging to the subset X_1, we have $\mathbf{w}^T(k)\mathbf{x}(k) < 0$. We therefore deduce from Eq. (4.17) that

$$\|\mathbf{w}(k + 1)\|^2 \leq \|\mathbf{w}(k)\|^2 + \|\mathbf{x}(k)\|^2$$

or, equivalently,

$$\|\mathbf{w}(k+1)\|^2 - \|\mathbf{w}(k)\|^2 \leq \|\mathbf{x}(k)\|^2, \qquad k = 1, \ldots, n \tag{4.18}$$

Adding these inequalities for $k = 1, \ldots, n$, and assuming that the initial condition $\mathbf{w}(0) = \mathbf{0}$, we get the following condition:

$$\begin{aligned} \|\mathbf{w}(n+1)\|^2 &\leq \sum_{k=1}^{n} \|\mathbf{x}(k)\|^2 \\ &\leq n\beta \end{aligned} \tag{4.19}$$

where β is a positive number defined by

$$\beta = \max_{\mathbf{x}(k) \in X_1} \|\mathbf{x}(k)\|^2 \tag{4.20}$$

Equation (4.19) states that the squared Euclidean norm of the weight vector $\mathbf{w}(n+1)$ grows at most linearly with the number of iterations n.

Clearly, the second result of Eq. (4.19) is in conflict with the earlier result of Eq. (4.15) for sufficiently large values of n. Indeed, we can state that n cannot be larger than some value $n_{\max}$ for which Eqs. (4.15) and (4.19) are both satisfied with the equality sign. That is, $n_{\max}$ is the solution of the equation

$$\frac{n_{\max}^2 \alpha^2}{\|\mathbf{w}_0\|^2} = n_{\max}\beta$$

Solving for $n_{\max}$ for a solution vector $\mathbf{w}_0$, we find that

$$n_{\max} = \frac{\beta \|\mathbf{w}_0\|^2}{\alpha^2} \tag{4.21}$$

We have thus proved that for $\eta(n) = 1$ for all n, and $\mathbf{w}(0) = \mathbf{0}$, and given that a solution vector $\mathbf{w}_0$ exists, the rule for adapting the synaptic weights connecting the associator units to the response unit of the perceptron must terminate after at most $n_{\max}$ iterations. Note also from Eqs. (4.11), (4.20), and (4.21) that there is *no* unique solution for $\mathbf{w}_0$ or $n_{\max}$.