$$W_i^T Y > W_j^T Y \quad (j = 1, \ldots, L:\ j \neq i) \quad \rightarrow \quad Y \in \omega_i \,, \tag{8.19}$$

where Y and W are defined, in terms of X, V, and v_0 of (8.1), by

$$y_0 = 1 \quad \text{and} \quad y_i = x_i \quad (i = 1, \ldots, n) \,,$$

$$w_i = v_i \quad (i = 0, 1, \ldots, n) \,. \tag{8.20}$$

When all $Y \in \omega_i$ $(i = 1, \ldots, L)$ satisfy (8.19), we call *these L classes linearly separable*. An algorithm to adjust these W's is given as follows:

(1) *If* $W_i^T Y > W_j^T Y$ $(j = 1, \ldots, L: j \neq i)$ for $Y \in \omega_i$, then

$$W_k' = W_k \quad (k = 1, \ldots, L) \,. \tag{8.21}$$

(2) *If* $W_\ell^T Y > W_i^T Y$ and $W_i^T Y > W_j^T Y$ for $Y \in \omega_i$, then

$$W_\ell' = W_\ell - cY, \quad W_i' = W_i + cY, \quad W_j' = W_j \quad (j \neq i, \ell) \,. \tag{8.22}$$

This multiclass problem can be reformulated as a two-class problem, if we extend our dimension to $n \times L$ as

$$W = [W_1^T \ldots W_{i-1}^T W_i^T W_{i+1}^T \ldots W_{\ell-1}^T W_\ell^T W_{\ell+1}^T \ldots W_L^T]^T \,, \tag{8.23}$$

$$Z = [0^T \ldots 0^T \quad Y^T \; 0^T \ldots 0^T \; -Y^T 0^T \ldots 0^T]^T \,, \tag{8.24}$$

for the Y of (8.22). Then, for a reduced sequence of samples, $Z_1^*, Z_2^*, \ldots$, we can obtain a corresponding sequence of W, $W_1^*, W_2^*, \ldots$. These W_k^*'s are related by

$$W_{k+1}^* = W_k^* + cZ_k^* \,. \tag{8.25}$$

Equation (8.25) is equivalent to (8.22). Since (8.25) is the same as (8.11), the convergence of (8.25) and, consequently, the convergence of (8.22) is guaranteed by the proof presented in the previous subsection.

As discussed in Chapter 4, a piecewise linear classifier is often used for separating many classes. Unfortunately, the convergence proof for a piecewise linear classifier is not known. However, similar algorithms to adjust the W's can be found in some references [3],[4].

8.2 Stochastic Approximation

The successive estimation algorithm of the last section does not always converge when the observation vectors are not linearly separable, as seen in Example 2. This fact leads us to look for an estimation algorithm for which convergence is guaranteed. *Stochastic approximation* is a technique that has been developed to find the root or the optimum point of a regression function in random environments [5],[6]. Stochastic approximation can be used for parameter estimation in pattern recognition, and convergence is guaranteed under very general circumstances. It is usually difficult, however, to discuss the rate of convergence.

Before we begin a detailed discussion, let us examine a simple example. Suppose we want to estimate the expected vector from a finite number of observation vectors. Suppose, further, that we want to use a successive estimate. Now the nonsuccessive estimate $\hat{\mathbf{M}}_N$ of the expected vector, based on N observation vectors, $\mathbf{X}_1, \ldots, \mathbf{X}_N$, is given by

$$\hat{\mathbf{M}}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \,. \tag{8.26}$$

The equation can be modified to

$$\hat{\mathbf{M}}_N = \frac{N-1}{N} \left\{ \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbf{X}_i \right\} + \frac{1}{N} \mathbf{X}_N$$

$$= \frac{N-1}{N} \hat{\mathbf{M}}_{N-1} + \frac{1}{N} \mathbf{X}_N \,. \tag{8.27}$$

That is, $\hat{\mathbf{M}}_N$ can be calculated with a new sample $\mathbf{X}_N$ if we store only $\hat{\mathbf{M}}_{N-1}$ and N. Also, the effect of the new sample on the sample mean vector should decrease, with an increase in N, as follows:

$$\mathbf{X}_1, \ \frac{1}{2}\mathbf{X}_2, \ \frac{1}{3}\mathbf{X}_3, \ \ldots, \ \frac{1}{N}\mathbf{X}_N \,. \tag{8.28}$$

The sequence of coefficients 1, 1/2, 1/3, ..., 1/N, ... is known as a *harmonic sequence*.

The above simple example suggests the following basic approach to successive estimation.

(1) When the mathematical expression for an estimate is available, we may obtain the successive expression of the estimate by separating the estimate calculated from $(N-1)$ samples and the contribution of the Nth sample.

(2) Even when we have to use a search process, in order to minimize or maximize a certain criterion, we may diminish the effect of the Nth sample by using a coefficient which is a decreasing function of N.

Root-Finding Problem

The simplest form of stochastic approximation is seen in finding a root of a *regression function*. This process is also called the *Robbins-Monro*

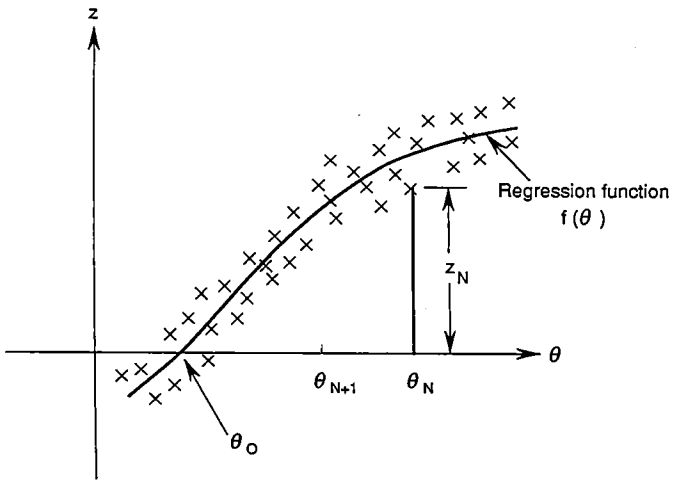

Fig. 8-3 Root-finding problem.

method [7]. Let $\boldsymbol{\theta}$ and $\mathbf{z}$ be two random variables with some correlation, as shown in Fig. 8-3. Our problem is to find the root of the regression function $f(\theta)$, which is defined by

$$f(\theta) = E\{\mathbf{z}|\theta\} . \tag{8.29}$$

If we can collect all samples for a fixed θ and estimate $E\{\mathbf{z}|\theta\}$, then finding the root of $f(\theta)$ can be carried out by a root-finding technique for a deterministic function such as the Newton method. However, when it is

predetermined that only one sample is observed for a given θ and we try to change θ accordingly, the observation of $f(\theta)$ is very noisy and may introduce an erroneous adjustment of θ, particularly around the root.

In the Robbins-Monro method, the new successive estimate $\boldsymbol{\theta}_{N+1}$ based on the present estimate $\boldsymbol{\theta}_N$ and a new observation $\mathbf{z}_N$ is given by

$$\boldsymbol{\theta}_{N+1} = \boldsymbol{\theta}_N - a_N \mathbf{z}_N , \tag{8.30}$$

where we assume, without losing any generality, that θ approaches θ_0, the root of (8.29), from the high side; that is, $f(\theta) > 0$ for $\theta > \theta_0$ and $f(\theta) < 0$ for $\theta < \theta_0$, as shown in Fig. 8-3. Also, a_N is assumed to be a sequence of positive numbers which satisfy the following conditions:

$$(1) \quad \lim_{N\to\infty} a_N = 0 , \tag{8.31}$$

$$(2) \quad \sum_{N=1}^{\infty} a_N = \infty , \tag{8.32}$$

$$(3) \quad \sum_{N=1}^{\infty} a_N^2 < \infty . \tag{8.33}$$

Although we will see later how these conditions for a_N are used for the convergence proof, the physical meaning of these equations can be described as follows. Equation (8.31) is similar to the $1/N$ term discussed earlier and allows the process to settle down in the limit. On the other hand, (8.32) insures that there is enough corrective action to avoid stopping short of the root. Equation (8.33) guarantees the variance of the accumulated noise to be finite so that we can correct for the effect of noise.

With a sequence of a_N satisfying (8.31) through (8.33), $\boldsymbol{\theta}_N$ of (8.30) converges toward θ_0 in the mean-square sense and with probability 1, that is,

$$\lim_{N\to\infty} E\{(\boldsymbol{\theta}_N - \theta_0)^2\} = 0 , \tag{8.34}$$

$$\lim_{N\to\infty} Pr\{\boldsymbol{\theta}_N = \theta_0\} = 1 . \tag{8.35}$$

The harmonic sequence of (8.28) is a suitable candidate for $\{a_N\}$. More generally, a sequence of the form

$$a_N = \frac{1}{N^k} \quad 1 \geq k > \frac{1}{2} \tag{8.36}$$

satisfies (8.31) through (8.33), although it is not the only possible sequence.

Before discussing the convergence of the Robbins-Monro method, let us consider a feedback system analogous to this process.

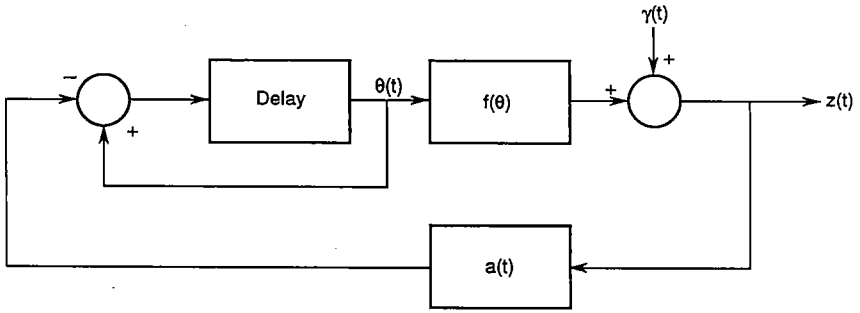

Fig. 8-4 Equivalent feedback circuit.

Figure 8-4 shows an equivalent feedback circuit, where $\gamma(t)$ is a noise process. Instead of a fixed feedback gain, we have a time-decreasing feedback gain $a(t)$. From the conventional design concept of a feedback circuit, one can notice that the decreasing $a(t)$ could guarantee the stability of the circuit without hunting but could also result in a slower response.

Convergence Proof of the Robbins-Monro Method

The convergence of the Robbins-Monro method is proved as follows. First let us divide $\mathbf{z}_N$ into two parts: the regression function $f(\boldsymbol{\theta}_N)$ and noise $\boldsymbol{\gamma}_N$. Then, (8.30) is rewritten as

$$\boldsymbol{\theta}_{N+1} = \boldsymbol{\theta}_N - a_N f(\boldsymbol{\theta}_N) - a_N \boldsymbol{\gamma}_N \,, \tag{8.37}$$

where

$$\boldsymbol{\gamma}_N = \mathbf{z}_N - f(\boldsymbol{\theta}_N) \,. \tag{8.38}$$

Then, from the definition of the regression function $f(\theta)$ in (8.29), $\boldsymbol{\gamma}_N$ is a random variable with zero mean as

$$E\{\boldsymbol{\gamma}_N \mid \boldsymbol{\theta}_N\} = E\{\mathbf{z}_N \mid \boldsymbol{\theta}_N\} - f(\boldsymbol{\theta}_N) = 0 \,. \tag{8.39}$$

Also, it is reasonable to assume that the variance of $\boldsymbol{\gamma}_N$ is bounded, that is,

$$E\{\boldsymbol{\gamma}_N^2\} \leq \sigma^2 \tag{8.40}$$

and that $\boldsymbol{\gamma}_N$ and $\boldsymbol{\theta}_N$ are statistically independent.

Next, let us study the difference between θ_0 and $\boldsymbol{\theta}_N$. From (8.37), we have

$$(\boldsymbol{\theta}_{N+1} - \theta_0) = (\boldsymbol{\theta}_N - \theta_0) - a_N f(\boldsymbol{\theta}_N) - a_N \boldsymbol{\gamma}_N \,. \tag{8.41}$$

Taking the expectation of the square of (8.41)

$$E\{(\boldsymbol{\theta}_{N+1} - \theta_0)^2\} - E\{(\boldsymbol{\theta}_N - \theta_0)^2\}$$

$$= a_N^2 E\{f^2(\boldsymbol{\theta}_N)\} + a_N^2 E\{\boldsymbol{\gamma}_N^2\} - 2a_N E\{(\boldsymbol{\theta}_N - \theta_0) f(\boldsymbol{\theta}_N)\} \,. \tag{8.42}$$

Therefore, repeating (8.42), we obtain

$$E\{(\boldsymbol{\theta}_N - \theta_0)^2\} - E\{(\boldsymbol{\theta}_1 - \theta_0)^2\}$$

$$= \sum_{i=1}^{N-1} a_i^2 [E\{f^2(\boldsymbol{\theta}_i)\} + E\{\boldsymbol{\gamma}_i^2\}] - 2\sum_{i=1}^{N-1} a_i E\{(\boldsymbol{\theta}_i - \theta_0) f(\boldsymbol{\theta}_i)\} \,. \tag{8.43}$$

We assume that the regression function is also bounded in the region of interest as

$$E\{f^2(\boldsymbol{\theta}_N)\} \leq b \,. \tag{8.44}$$

Then, (8.43) is bounded by

$$E\{(\boldsymbol{\theta}_N - \theta_0)^2\} - E\{(\boldsymbol{\theta}_1 - \theta_0)^2\}$$

$$\leq (b + \sigma^2) \sum_{i=1}^{N-1} a_i^2 - 2\sum_{i=1}^{N-1} a_i E\{(\boldsymbol{\theta}_i - \theta_0) f(\boldsymbol{\theta}_i)\} \,. \tag{8.45}$$

Let us examine (8.45) term by term. First, since $E\{(\boldsymbol{\theta}_N - \theta_0)^2\}$ is positive and assuming $\boldsymbol{\theta}_1$ is selected so that $E\{(\boldsymbol{\theta}_1 - \theta_0)^2\}$ is finite, the left-hand side of (8.45) is bounded from below. The first term on the right-hand side of (8.45) is finite because of (8.33).

Recall from Fig. 8-3 that the regression function satisfies:

$$
\begin{aligned}
f(\theta) &> 0 \text{ if } (\theta - \theta_0) > 0 , \\
f(\theta) &= 0 \text{ if } (\theta - \theta_0) = 0 , \\
f(\theta) &< 0 \text{ if } (\theta - \theta_0) < 0 .
\end{aligned} \tag{8.46}
$$

Therefore,

$$(\theta - \theta_0) f(\theta) \geq 0 , \tag{8.47}$$

and

$$E\{(\boldsymbol{\theta} - \theta_0) f(\boldsymbol{\theta})\} \geq 0 . \tag{8.48}$$

Now consider the following proposition:

$$\lim_{i \to \infty} E\{(\boldsymbol{\theta}_i - \theta_0) f(\boldsymbol{\theta}_i)\} = 0 . \tag{8.49}$$

If (8.49) does not hold, then, because of (8.32), the last term of (8.45) tends toward $-\infty$. But this contradicts the fact that the left-hand side of (8.45) is bounded from below. Hence, (8.49) must hold. Since (8.47) holds for all θ's, (8.49) is equivalent to

$$\lim_{i \to \infty} Pr\{\boldsymbol{\theta}_i = \theta_0)\} = 1 . \tag{8.50}$$

Thus, the convergence with probability 1 is proved. The convergence in mean-square sense has also been proved but this proof is omitted here.

Minimum-Point-Finding Problem

The Robbins-Monro method can be easily modified to seek the minimum point of a regression function instead of the root. As is well known, the minimum point or the optimum point of a function $f(\theta)$ is a root of $df(\theta)/d\theta$. Therefore, if we can measure $df(\theta)/d\theta$, we can apply the Robbins-Monro method directly. Unfortunately, in most applications, the measurement of $df(\theta)/d\theta$ is not available. Therefore, we measure the derivative experimentally and modify $\boldsymbol{\theta}_N$ as

$$\boldsymbol{\theta}_{N+1} = \boldsymbol{\theta}_N - a_N \frac{\mathbf{z}(\boldsymbol{\theta}_N + c_N) - \mathbf{z}(\boldsymbol{\theta}_N - c_N)}{2c_N} . \tag{8.51}$$

This successive equation is called the *Kiefer-Wolfowitz method* [8]. Figure 8-5 illustrates the Kiefer-Wolfowitz method.

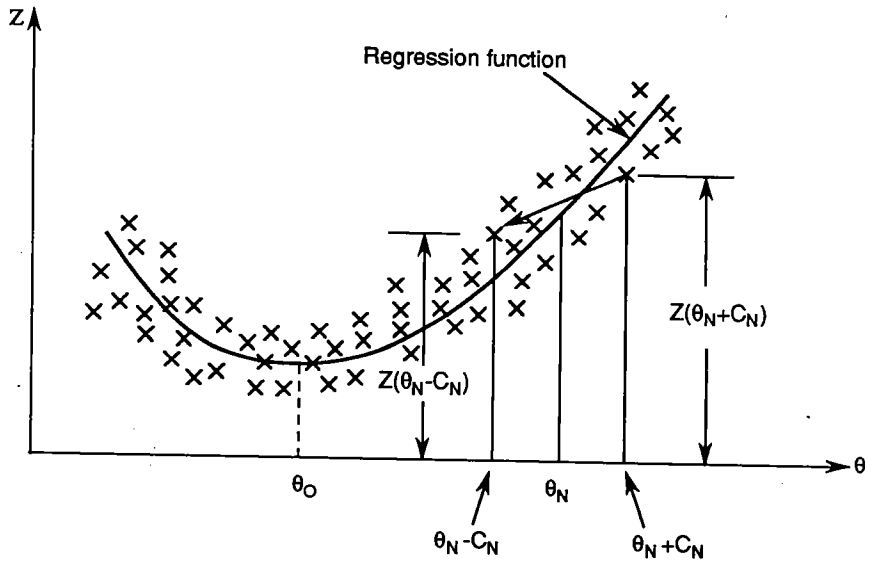

Fig. 8-5 Minimum-point-finding problem.

Both a_N and c_N are sequences of positive numbers. They must vanish in the limit, that is,

$$\lim_{N \to \infty} a_N = 0 , \tag{8.52}$$

$$\lim_{N \to \infty} c_N = 0 , \tag{8.53}$$

so that the process eventually converges. In order to make sure that we have enough corrective action to avoid stopping short of the minimum point, a_N should satisfy

$$\sum_{N=1}^{\infty} a_N = \infty . \tag{8.54}$$

Also, to cancel the accumulated noise effect we must have

$$\sum_{N=1}^{\infty} \left(\frac{a_N}{c_N} \right)^2 < \infty . \tag{8.55}$$

With a_N and c_N satisfying these conditions, it has been proven that $\boldsymbol{\theta}_N$ of (8.51) converges to θ_0 both in the mean-square sense and with probability 1, provided that we have a bounded variance for the noise and a bounded slope for the regression function. The proof is similar to the one for root-finding but is omitted here.