9

R. L. Kashyap
C. C. Blaydon
K. S. Fu

STOCHASTIC APPROXIMATION

I. Introduction

The goal of many adaptive or learning systems is to find or "learn" the value of certain parameters which minimize a certain criterion function. If the criterion function is known precisely, the point of minimum can be found by any one of the standard numerical techniques. In many problems, the criterion function is not known explicitly. Even in such cases, we can determine the minimum of the function recursively by the steepest descent method if we are in a position to observe the *exact* value of the function for any given setting of the parameters by means of a suitable experiment. But in almost all experiments, the results are observed only after they have been corrupted by measurement noise. In such circumstances, the stochastic analogs of the steepest descent method may have to be used to obtain the minimum of the function. These analogs are the stochastic approximation methods.

In our discussion, we would like to minimize a deterministic function $J(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter under our control. The function $J(\boldsymbol{\theta})$ is completely unknown; however, for every $\boldsymbol{\theta}$, we can observe a random variable $z(\boldsymbol{\theta})$ by performing a suitable experiment so that $E[z(\boldsymbol{\theta})|\,\boldsymbol{\theta}] = J(\boldsymbol{\theta})$. Sometimes, we may even be able to observe $\mathbf{y}(\boldsymbol{\theta})$, a noisy value of the gradient function where $\mathbf{y}(\boldsymbol{\theta})$ obeys the relation $E(\mathbf{y} \mid \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. We

have to devise a computing scheme to determine θ^o, the value of θ which minimizes $J(\theta)$ based on the observed values of $z(\theta)$ or $\mathbf{y}(\theta)$. Since every experiment has a cost associated with it, it is important that our method should give a satisfactory result with as few observations as possible, by processing them in an optimal fashion. Moreover, if we make the scheme recursive so that it alternately involves the experimental evaluation of $z(\theta_k)$ or $\mathbf{y}(\theta_k)$ for a given value of θ_k (the current estimate of θ^0 at the kth stage), and the updating of the estimate θ_k on the basis of the additional information; then we can think of the scheme as possessing "learning" properties, since at each stage it gives an estimate of θ^o which is superior to the one at the previous stage, and yields the true value θ^o in the event of carrying on the computation for an infinite length of time.

To clarify the roles of $J(\theta)$, $z(\theta)$ and $\mathbf{y}(\theta)$, let us consider the following example in pattern classification. A random pattern $\mathbf{x}$ can be from one of two possible classes, ω_1 and ω_2. For every $\mathbf{x}$, define the class indicator variable $d(\mathbf{x})$ as follows:

$$\begin{aligned} d(\mathbf{x}) &= 1 \quad \text{if} \quad \mathbf{x} \sim \omega_1 \\ &= 0 \quad \text{if} \quad \mathbf{x} \sim \omega_2 \end{aligned}$$

For classifying any given pattern with unknown class, we need a decision rule. The decision rule which minimizes the probability of misclassification is given below in terms of the probability function $p(d(\mathbf{x}) = 1 \mid \mathbf{x})$:

$$\text{classify } \mathbf{x} \text{ in class } \omega_1 \text{ if } p(d(\mathbf{x}) = 1 \mid \mathbf{x}) > 0.5$$

$$\text{classify } \mathbf{x} \text{ in class } \omega_2 \text{ if } p(d(\mathbf{x}) = 0 \mid \mathbf{x}) < 0.5$$

Since we do not know the probability function $p(d(\mathbf{x}) = 1 \mid \mathbf{x})$, we might be willing to settle for an approximation $\theta'\mathbf{x}$ for this function. The corresponding decision scheme is given below:

$$\omega_1 \text{ if } \theta'\mathbf{x} > 0.5$$

classify $\mathbf{x}$ in class

$$\omega_2 \text{ if } \theta'\mathbf{x} < 0.5$$

where we want to choose the value of θ to minimize the criterion $J_1(\theta)$

$$\begin{aligned} J_1(\theta) &= E\{\theta'\mathbf{x} - p(d(\mathbf{x}) = 1 \mid \mathbf{x})\}^2 \\ &= E\{\theta'\mathbf{x} - E\{d(\mathbf{x}) \mid \mathbf{x}\}\}^2 \\ &= E\{\theta'\mathbf{x} - d(\mathbf{x})\}^2 + E\{d(\mathbf{x}) - E\{d(\mathbf{x}) \mid \mathbf{x}\}\}^2 \end{aligned}$$

Let

$$J_2(\theta) = E\{\theta'\mathbf{x} - d(\mathbf{x})\}^2$$

Clearly, we may minimize $J_2(\theta)$ instead of $J_1(\theta)$. But $J_2(\theta)$ is not known to us explicitly. However, for any given value θ, we can evaluate the random variable $z(\theta)$ by observing a random pattern $\mathbf{x}$ with known classification $d(\mathbf{x})$

$$z(\theta) = (\theta'\mathbf{x} - d(\mathbf{x}))^2$$

$$E\{z(\theta)|\ \theta\} = J_2(\theta)$$

In this particular instance, we can observe, in addition, the noisy gradient value $\mathbf{y}(\theta)$,

$$\mathbf{y}(\theta) = 2(\theta'\mathbf{x} - d(\mathbf{x}))\mathbf{x} \qquad E\{\mathbf{y}(\theta)|\ \theta\} = \nabla_\theta J_2(\theta)$$

If a sequence of pattern vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ and their respective classifications $(d(\mathbf{x}_1), d(\mathbf{x}_2), \ldots)$ are available, then samples of $z(\theta)$ and $\mathbf{y}(\theta)$ are

$$z(\theta) : \{(\theta'\mathbf{x}_1 - d(\mathbf{x}_1))^2, (\theta'\mathbf{x}_2 - d(\mathbf{x}_2))^2, \ldots\}$$

$$\mathbf{y}(\theta) : \{2(\theta'\mathbf{x}_1 - d(\mathbf{x}_1))\mathbf{x}_1, 2(\theta'\mathbf{x}_2 - d(\mathbf{x}_2))\,\mathbf{x}_2, \ldots\}$$

Using the noisy gradient values, $\mathbf{y}(\theta)$, the gradient scheme for locating the minimum of $J_2(\theta)$ can be written as follows,

$$\theta_{k+1} = \theta_k - \rho_k \mathbf{y}(\theta_k)$$

$$= \theta_k - \rho_k(\theta_k'\mathbf{x}_k - d(\mathbf{x}_k))\mathbf{x}_k$$

where $\{\rho_k\}$ is a suitable scalar gain sequence. Based on the nature of the observations $\mathbf{y}(\theta)$ and $z(\theta)$ we divide the problems into three categories:

(i) For every θ, the samples of random variable $z(\theta)$ and $\mathbf{y}(\theta) = \nabla_\theta z(\theta)$ are available.

(ii) For every θ, the samples of $\mathbf{y}(\theta)$ are not directly observable. However, we can observe a related random variable $\bar{\mathbf{y}}(\theta)$ so that $\| E\{\bar{\mathbf{y}}(\theta)\} - E\{\mathbf{y}(\theta)\}\|$ goes to zero asymptotically.

(iii) For every θ, we can observe only samples of $z(\theta)$.

In this chapter, we will discuss stochastic approximation methods for minimizing $J(\theta) = E\{z(\theta)|\ \theta\}$ for the three cases described above. Each case will be illustrated with a particular learning system. There are several surveys of stochastic approximation methods that give convergence conditions in all of their generality (Schmetterer, 1961; Loginov,

1966). In our discussion, we will only emphasize those sets of conditions which are easily verified in real situations. In addition to the review of stochastic approximation for minimizing functions, we will also discuss its application to recovering functions from noisy measurements of their values, and, describe the relationship of stochastic approximation to other techniques such as the potential function method.

II. Algorithms for Finding Zeroes of Functions

In some cases, the learning scheme can find the minimum of $J(\theta) = E\{z(\theta)|\ \theta\}$ by finding the zero of $\nabla_\theta J(\theta)$. The problem treated in this section is the solution of Eq. (9.1) where $\mathbf{y}(\theta)$ is a random vector and $\mathbf{r}(\theta)$ is an *unknown function*

$$\nabla_\theta J(\theta) \triangleq \mathbf{r}(\theta) \triangleq E\{\mathbf{y}(\theta)|\ \theta\} = \mathbf{0} \tag{9.1}$$

Our first assumption is (A1).

(A1): For every θ, a random vector $\mathbf{y}(\theta)$ is observable where $\mathbf{r}(\theta) \triangleq E\{\mathbf{y}(\theta)|\ \theta\}$ has a unique zero at the value $\theta = \theta^o$ and θ^o lies in the region given below

$$\theta_i^{(1)} \leqslant \theta_i{}^o \leqslant \theta_i^{(2)} \qquad i = 1,\ldots n$$

If $\mathbf{r}(\theta)$ were known, the gradient scheme for finding the zero of $\mathbf{r}(\theta)$ would be:

$$\theta_{k+1} = \theta_k - \rho \mathbf{r}(\theta_k) \tag{9.2}$$

Since $\mathbf{r}(\theta)$ is unknown, we replace $\mathbf{r}(\theta_k)$ in Eq. (9.2) by the random variable $\mathbf{y}(\theta)$, using Eq. (9.1)

$$\theta_{k+1} = \theta_k - \rho_k \mathbf{y}(\theta_k) \tag{9.3}$$

Instead of the fixed gain ρ in Eq. (9.2), we use the variable gain sequence ρ_k in Eq. (9.3) to compensate for the "error" involved in replacing the deterministic function $\mathbf{r}(\theta)$ by the random function $\mathbf{y}(\theta)$. The gain ρ_k in the algorithm must be chosen as in Eq. (9.4), where $\Sigma {\rho_k}^2 < \infty$ ensures that the sum of the squares of the correction terms in Eq. (9.3) is finite, but $\Sigma \rho_k = \infty$ assures that the sum of corrections terms may be infinite

$$\rho_k \geqslant 0, \qquad \sum_k \rho_k = \infty \qquad \sum_k {\rho_k}^2 < \infty \tag{9.4}$$

To make sure that the corrections applied in Eq. (9.3) drive θ_k in the right direction on the average, culminating in the zero of $\mathbf{r}(\theta)$, we need

the following assumption (A2), which says that $\mathbf{r}(\theta)$ behaves like a linear function for values of θ near θ^o.

$$\text{(A2):} \qquad \underset{\epsilon<\|\theta-\theta^o\|<\epsilon^{-1}}{\text{Inf.}} [(\theta - \theta^o)' \mathbf{r}(\theta)] > 0 \qquad \forall \epsilon > 0$$

A further assumption (A3) on $\mathbf{y}(\theta)$ is necessary.

$$\text{(A3):} \qquad E\{\| \mathbf{y}(\theta)\|^2\} \leqslant h(1 + \| \theta - \theta^o \|^2) \qquad h > 0$$

The assumption (A3) assures that the variance of $\mathbf{y}(\theta)$ is bounded and that $\| \mathbf{y}(\theta)\|^2$ is bounded above by a quadratic function for all θ.

If all three assumptions are satisfied, then the algorithm in Eq. (9.3) with the gain as in Eq. (9.4) converges to the unique value θ^o which makes $\mathbf{r}(\theta)$ zero in mean square and with probability one. In other words

$$\lim_{k\to\infty} E\{\| \theta_k - \theta^o \|^2\} = 0$$

and

$$\text{Prob.}\{\lim_{k\to\infty} \theta_k = \theta^o\} = 1$$

The proof of such convergence is given in Appendix 1; it is based on the proof of Theorem 1 of Gladyshev (1965).

It is important to note that we do not need the successive samples $\mathbf{y}(\theta_1)$, $\mathbf{y}(\theta_2)$,... to be independent of one another. The only restrictive assumption that has been made is (A3) and that can be relaxed. If this assumption is relaxed, an additional assumption about the independence of the samples must be made. In this case, replace (A3) with (A3)′ and (A4).

$$\text{(A3)′:} \qquad E\{\| \eta_k \|^2\} < k_1 \qquad \forall k > k_2 > 0$$

where $\eta_k \triangleq \mathbf{y}(\theta_k) - \mathbf{r}(\theta_k)$

(A4): The random variables η_1 , η_2 ,... are independent of one another.

Again, with the assumptions (A1), (A2), (A3)′ and (A4) the algorithm in Eq. (9.3) converges in mean square and with probability one to the unique zero of $\mathbf{r}(\theta)$. The convergence proof follows that of Theorem 3 of Venter (1966).

Finally, there is occasionally the case where the random variable $\mathbf{y}(\theta)$ cannot be directly observed. Even then, we can develop a scheme to locate the zero provided a random variable $\bar{\mathbf{y}}(\theta)$ can be observed, where

$$\bar{\mathbf{y}}(\theta_k) = \mathbf{y}(\theta_k) + \eta_k$$

Here the random sequence $\{\eta_k\}$ is a "noise" masking the observation of $\mathbf{y}(\theta)$. The algorithm in Eq. (9.3) now has to use the observable $\bar{\mathbf{y}}(\theta)$ and becomes

$$\theta_{k+1} = \theta_k - \rho_k \bar{\mathbf{y}}(\theta_k) \tag{9.3}'$$

Since the biases in the "noise" η_k could cause this algorithm to converge to values other than θ^o, some restrictions must be placed on η_k. Either of the following restrictions (A1)′ or (A1)″ will serve. They are

$$\text{(A1)}': \qquad \lim_{k\to\infty} \| E\{\eta_k \mid \eta_{k-1}, \eta_{k-2}, \ldots, \eta_1\}\| = 0$$

or

$$\text{(A1)}'': \qquad \lim_{k\to\infty} E\{\| E\{\eta_k \mid \eta_{k-1}, \eta_{k-2}, \ldots, \eta_1\}\|^2\} = 0$$

Under the new assumptions (A1)′, (A2) and (A3)′, the algorithm in Eq. (9.3)′ with gains in Eq. (9.4), converges to the unique zero θ^o of $\mathbf{r}(\theta)$ with probability one. If assumption (A1)′ is replaced by the stronger assumption (A1)″, then the algorithm/in Eq. (9.3)′ with gains in Eq. (9.4) converges both in mean square and with probability one. For a proof, see Venter (1966).

It is interesting to note that the condition $E\{\eta_k\} = \mathbf{0}$ is not sufficient to satisfy either assumption (A1)′ or (A1)″. However, the condition that $E\{\eta_k\} = \mathbf{0}$ and the additional condition that the sequence $\{\eta_k\}$ be independent is sufficient (but not necessary) to satisfy (A1)′ and (A1)″. This means that conditions (A1) and (A3)′ are special cases of (A1)′ and (A1)″.

III. Kiefer–Wolfowitz Schemes

In the previous section, we were interested in finding the extremum of the function $J(\theta) = E(z(\theta)|\ \theta)$ based only on the observed samples of $\mathbf{y}(\theta) = \nabla_\theta z(\theta)$. In this section, we will determine methods for finding the extremum of $J(\theta)$ using only samples of the random variable $z(\theta)$. Such methods are called Kiefer–Wolfowitz (KW) algorithms.

The KW schemes use the estimates of the stochastic gradient $(\nabla_\theta z(\theta))$ computed from neighboring values of the function. To compute the stochastic gradient at a value θ_k, we need to observe $2n$ random samples of $z(\theta)$ which are $z(\theta_k \pm c_k \mathbf{e}_j), j = 1, \ldots, n$: where n is the dimension of θ; $c_1, c_2, \ldots$ is a suitably chosen positive sequence; and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are the n-dimensional unit vectors,

$$\mathbf{e}_1' = (1, 0, \ldots, 0) \qquad \mathbf{e}_2' = (0, 1, 0, \ldots, 0) \cdots \mathbf{e}_n' = (0, 0, \ldots, 0, 1)$$

Thus,

$$\hat{\nabla}_{\theta} z(\theta_k) = \sum_{j=1}^{n} \mathbf{e}_j \frac{1}{2c_k} [z(\theta_k + c_k \mathbf{e}_j) - z(\theta_k - c_k \mathbf{e}_j)] \tag{9.5}$$

Using the estimate of the gradient, the computing scheme can be written as follows:

$$\theta_{k+1} = \theta_k - \rho_k \hat{\nabla}_{\theta} z(\theta_k) \tag{9.6}$$

The gains ρ_k and c_k decay to zero in the following manner:

$$\left.\begin{aligned} \Sigma \rho_k = \infty \qquad \Sigma \rho_k c_k < \infty \qquad \Sigma \left(\frac{\rho_k}{c_k}\right)^2 < \infty \\ \rho_k < d_1 c_k^2 \qquad 0 < d_1 < \infty \end{aligned}\right\} \tag{9.7}$$

The condition $\Sigma(\rho_k/c_k)^2 < \infty$, as before, is imposed to ensure that the sum of the squares of the correction terms is finite. Moreover, if we expand the gradient $\nabla_{\theta} z(\theta)$ in Eq. (9.5) in a Taylor series about θ_k and retain only the first two terms, then the condition $\Sigma \rho_k c_k < \infty$ insures that the contributions of the second-order derivatives are finite.

The assumptions for the convergence given in the original pioneering paper by Kiefer–Wolfowitz (1952) are highly restrictive, especially the following condition of linear unimodality:

$$(\theta - \theta^o)' \frac{dJ}{d\theta} > 0 \qquad \forall \theta \tag{9.8}$$

We will give the assumptions used by Venter (1967b) in his proof of KW schemes:

(B1): $J(\theta)$ and its second-order derivatives are bounded over R^n.

(B2): θ^o is a local minimum of $J(\theta)$; i.e., for some $\epsilon > 0$, we have $J(\theta^o) \leqslant J(\theta)$ for all $\theta \neq \theta^o$ in the set $\{\theta : \| \theta - \theta^o \| < \epsilon\}$.

(B3): For every $\theta \in R^n$, $\nabla_{\theta} J(\theta) \neq 0$ if $\theta \neq \theta^o$; that is, θ^o is the only stationary point of $J(\theta)$.

(B4): $E[(\eta(\theta))^2] < \infty \; \forall \theta \in R^n$ where $J(\theta)$ has been redefined as

$$z(\theta) = J(\theta) + \eta(\theta) \qquad E\{\eta(\theta)\} = 0$$

If these conditions are met, then the algorithm in Eq. (9.6) with gains as in Eq. (9.7) converges with probability one to either θ^o or to infinity by Venter's theorem. There is no possibility of oscillation of infinite duration for θ_k.

The assumptions (B1) through (B3) are required for convergence even of the deterministic gradient algorithm in which $J(\theta)$ is explicitly known. The only auxiliary condition required for convergence of the stochastic algorithm is the assumption (B4) about the finiteness of the variance of the noise. The four assumptions assure the convergence of algorithm in Eq. (9.6) to the minimum of $J(\theta)$ or to $\pm\infty$.

Further, when θ is scalar, we can show that $\| \theta_k \| < \infty$ and thus prove convergence to the local minimum of $J(\theta)$ with probability one. If we want to rule out the possibility of $\| \theta_k \| \to \infty$ when the dimension of θ is greater than one, then we have to impose **one** of the following conditions (B5), (B5)′, or (B5)″.

$$\text{(B5):} \qquad \lim_{t\to\infty} \inf_{\|\theta-\theta^o\|>t} \{\| \nabla_\theta J(\theta)\|\} > 0$$

(B5) is not satisfied by such a simple function as $J(\theta) = \{-\| \theta - \theta^o \|^2\}$. In such cases, we use condition (B5)′.

(B5)′: For some $t > 0$, $(\theta - \theta^o)' \nabla_\theta J(\theta) \leqslant 0$ for all θ such that $\| \theta - \theta^o \| > t$

In many problems, the condition (B5) can be replaced by a weaker condition (B5)″.

$$\text{(B5)″:} \qquad \lim_{t\to\infty} \inf_{\|\theta-\theta^o\|>t} \{\| \theta - \theta^o \| \, \| \nabla_\theta J(\theta)\|\} > 0$$

Note that the condition (B5)′ is stronger than (B5)″.

Finally, we will comment briefly on the choice of the gains ρ_k and c_k. The usual choice is given below

$$\rho_k = \frac{\rho_1}{k} \qquad c_k = \frac{c_1}{k^\gamma} \qquad 0 < \gamma < 1/2 \qquad \rho_1, c_1 > 0$$

IV. Recovery of Functions from Noisy Measurements

In the previous two sections, we discussed the problem of finding the extremum of an unknown function when we had noisy observations of that function. In this section, we will discuss two related problems.

Problem (A): Let

$$z_k = f_k(\theta^o) + \eta_k \qquad k = 1, 2, \ldots$$

where $f_1(\theta)$, $f_2(\theta)$,... are *known* functions of a parameter θ whose true value θ^o is unknown. The noise η_k is zero mean and independent. The problem is to recover θ from the scalar measurements $z_1, z_2, \ldots$.

Problem (B): Let

$$z_k = f(\mathbf{x}_k) + \eta_k$$

where $f(\mathbf{x})$ is an *unknown* scalar function of the argument $\mathbf{x}$, which is a random m-vector. For every sample value of the argument (that is, $\mathbf{x}_k$), we can make a noisy measurement z_k of the function $f(\mathbf{x}_k)$. η_k is the corrupting additive noise. The problem is to estimate an approximation of $f(\cdot)$ from the sample pairs $\{z_k , \mathbf{x}_k; k = 1, 2,...\}$.

Problems (A) and (B) do not directly fit in the category of problems considered in Sections II and III. For example, in problems of type (A), the regression function $f_k(\theta)$ changes with the time index k (but θ remains the same), unlike Section II wherein the regression function did not change. Similarly, in problems of type (B) we deal with pairs of random variables $\{z_k , \mathbf{x}_k\}$ and not with only one random variable as in Section II.

A. *Estimation of the Parameters of a Known Time-Varying Function*

In this problem, the measurement at the kth stage is z_k and is given by

$$z_k = f_k(\theta^o) + \eta_k$$

where $E(\eta_k) = 0$ and $f_k(\theta)$, $k = 1, 2,...$ are known functions of θ.

The unknown n-vector θ^o has to be estimated. θ_k is the estimate of θ^o based on the measurements $z_1 , z_2 ,..., z_{k-1}$. We want to find a recursive relationship for finding θ_{k+1} from the measurement z_k and θ_k so that θ_k tends to θ^o as k goes to infinity.

Define the vector $\mathbf{g}_k(\theta)$ as

$$\mathbf{g}_k(\theta) = \nabla_\theta f_k(\theta) \tag{9.9}$$

The expansion of $f_k(\theta)$ about θ_k can then be written as

$$f_k(\theta) = f_k(\theta_k) + \mathbf{g}_k'(\xi_k)(\theta - \theta_k) \tag{9.10}$$

where ξ_k is an n-vector whose tip lies on the line joining θ and θ_k. We can define a "transformed" observation $\bar{z}_k$ as

$$\bar{z}_k \triangleq \mathbf{g}_k'(\xi_k)\theta^o + \eta_k \tag{9.11}$$

where

$$\bar{z}_k = z_k - f_k(\theta_k) + \mathbf{g}_k'(\xi_k)\theta_k \tag{9.12}$$

Equations (9.11) and (9.12) form the basis for the recursion scheme.

There are two methods for the estimation of θ: a first-order method and a second-order method.

1. *First-order method.* One way of estimating θ^o from the transformed observations in Eq. (9.11) would be to use a gradient algorithm to minimize the stochastic criterion function:

$$J_k(\theta) = (\eta_k)^2 = (\bar{z}_k - \mathbf{g}_k'(\xi_k)\theta)^2$$

The gradient algorithm is

$$\begin{aligned}\theta_{k+1} &= \theta_k - \rho_k \nabla_\theta J(\theta)|_{\theta=\theta_k} \\ &= \theta_k + \rho_k \mathbf{g}_k(\xi_k)[\bar{z}_k - \mathbf{g}_k'(\xi_k)\theta_k] \\ &= \theta_k + \rho_k \mathbf{g}_k(\xi_k)[z_k - f_k(\theta_k)]\end{aligned} \tag{9.13}$$

The two principal questions are how to choose the scalar sequence ρ_k and how to choose the points ξ_k at which to evaluate the derivatives $\mathbf{g}_k(\xi_k)$. The choice of ρ_k can be made to depend on ξ_k by choosing the gain sequence to be

$$\rho_k = 1 \Big/ \sum_{j=1}^{k} \| \mathbf{g}_j(\xi_j)\|^2 \tag{9.14}$$

This reduces the problem to the selection of $\{\xi_k, k = 1, 2,...\}$ which can be done in three ways.

(a) Deterministic gain sequence: Let $\xi_k = \theta_0$, a constant vector. In this case θ_0 is the a priori or nominal estimate of θ^o.

(b) Adaptive gain sequence: $\xi_k = \theta_k$. This is what is obtained by minimizing the criterion $J = (\xi_k - f_k(\theta))^2$ directly without linearizing about ξ_k.

(c) Quasi-adaptive gain sequence: $\xi_k = \theta(j_k)$ where j_k is a nondecreasing sequence of integers with $j_k \leqslant k$. The corresponding gain ρ_k is between the deterministic and adaptive gains. One usually starts with $\xi_k = \theta_0$, but as θ_k converges closer to θ^o, it is better to change (say at $k = n$) to an adaptive gain sequence. This will help speed up the convergence.

In many cases, we know that the θ^o lies in the cube

$$c_i \leqslant \theta_i{}^o \leqslant d_i \qquad i = 1,..., n \tag{9.15}$$

where $\theta_i{}^o$ is the ith component of θ^o.

In these cases, we can achieve faster convergence by modifying the algorithm in Eq. (9.13) so that θ_k is never outside the cube for any k:

$$\theta_{k+1} = \{\theta_k + \rho_k \mathbf{g}_k(\xi_k)(z_k - f_k(\theta_k))\}_P$$

where $\{\mathbf{x}\}_P$ means the orthogonal projection of vector $\mathbf{x}$ onto the cube defined in Eq. (9.15); that is to say,

$$\{\mathbf{x}\}_P = \begin{bmatrix} \{x_1\}_P \\ \{x_2\}_P \\ \vdots \\ \{x_n\}_P \end{bmatrix}$$

where

$$\begin{aligned} \{x_i\}_P &= c_i \quad \text{if} \quad x_i \leqslant c_i \\ &= x_i \quad \text{if} \quad c_i \leqslant x_i \leqslant d_i \\ &= d_i \quad \text{if} \quad x_i \geqslant d_i \end{aligned}$$

2. *Second-order method.* Instead of just following the gradient of

$$J_k(\theta) = (\bar{z}_k - \mathbf{g}_k'(\xi_k)\theta)^2$$

it would be better to use an algorithm that minimizes

$$\sum_{j=1}^{k} J_j + \| \theta - \theta_0 \|_R^2$$

where θ_0 is the a priori value for θ^o and R is a positive definite weighting matrix. The minimization algorithm is

$$\theta_{k+1} = \left[\sum_{j=1}^{k} \mathbf{g}_j(\xi_j)\, \mathbf{g}_j'(\xi_j) + R\right]^{-1}\left[\sum_{j=1}^{k} \mathbf{g}_j(\xi_j)\bar{z}_j + R\theta_0\right]$$

which can be written recursively as

$$\begin{aligned} \theta_{k+1} &= \theta_k + P_k \mathbf{g}_k(\xi_k)[\bar{z}_k - \mathbf{g}_k'(\xi_k)\theta_k] \\ &= \theta_k + P_k \mathbf{g}_k(\xi_k)[z_k - f_k(\theta_k)] \end{aligned} \tag{9.16}$$

where the $n \times n$ matrix P_k obeys the following scheme

$$\begin{aligned} P_k &= P_{k-1} - \frac{P_{k-1}\mathbf{g}_k(\xi_k)\, \mathbf{g}_k'(\xi_k) P_{k-1}}{1 + \mathbf{g}_k'(\xi_k)\, P_{k-1}\mathbf{g}_k(\xi_k)} \\ P_0 &= R \end{aligned}$$

The sequence ξ_k can also be chosen to be deterministic, adaptive, or quasi-adaptive as in the first order algorithm.

3. *Convergence conditions.* Albert and Gardner (1967) have described the sufficient conditions required for the convergence of the algorithms in Eqs. (9.13) and (9.16). Qualitatively, the most important condition concerns the direction and not the magnitude of the vectors $\mathbf{g}_k(\xi_k)$. This condition requires that the vectors $\mathbf{g}_k(\xi_k)$ repeatedly span the n-space (θ has dimension n). If this condition is not met, the algorithm may converge but perhaps not to the true value θ^o. See Albert and Gardner (1967: 125–126) for the complete conditions and convergence proof.

B. *Estimation of Functions from Values Observed at Randomly Selected Points*

In the previous problem, the functions $f_k(\theta)$ were known and we wanted to estimate θ^o based on noisy observations of $f_k(\theta^o)$. In the present problem, the noisy observations are z_k, given by

$$z_k = f(\mathbf{x}_k) + \eta_k \tag{9.17}$$

where the function $f(\mathbf{x})$ is unknown and the argument $\mathbf{x}$ is a random vector. The problem is to obtain a suitable approximation for the function $f(\cdot)$ from the observations z_k. At each stage k we can observe z_k and the argument $\mathbf{x}_k$. We know nothing of the statistics of the random argument vector $\mathbf{x}$ or the noise η except that

$$E\{\eta_k \mid \mathbf{x}_k\} = 0$$

We will be content with seeking a "best" approximation to $f(\cdot)$. The approximation $\hat{f}(\cdot)$ will be chosen to have the form

$$\sum_{i=1}^{n} \theta_i \varphi_i(\mathbf{x}) = \boldsymbol{\theta}' \boldsymbol{\varphi}(\mathbf{x})$$

where $\varphi_1(\mathbf{x})$, $\varphi_2(\mathbf{x})$,... and $\varphi_n(\mathbf{x})$ are a set of n linearly independent functions (and θ is an undetermined parameter). The problem is to find the value for θ that makes $\hat{f}(\cdot)$ "best" in the sense that the mean square approximation error $E\{(f(\mathbf{x}) - \boldsymbol{\theta}'\boldsymbol{\varphi}(\mathbf{x})^2\}$ is a minimum. The value of θ which minimizes this error is

$$\begin{aligned} \boldsymbol{\theta}^o &= \text{Arg.}[\underset{\boldsymbol{\theta}}{\text{Min.}}\, E\{f(\mathbf{x}) - \boldsymbol{\theta}'\boldsymbol{\varphi}(\mathbf{x})\}^2] \\ &= [E\{\boldsymbol{\varphi}(\mathbf{x})\, \boldsymbol{\varphi}'(\mathbf{x})\}]^{-1}\, E\{\boldsymbol{\varphi}(\mathbf{x}) f(\mathbf{x})\} \end{aligned}$$

where we have assumed the indicated inverse to exist. The corresponding optimal approximation $\hat{f}(\mathbf{x})$ is

$$\hat{f}(\mathbf{x}) = \boldsymbol{\theta}^{o\prime}\boldsymbol{\varphi}(\mathbf{x})$$

We cannot calculate this expression explicitly because we do not know the function $f(\mathbf{x})$ or the statistics of $\mathbf{x}$. Instead note that $\boldsymbol{\theta}^o$ can be written as

$$\boldsymbol{\theta}^o = \text{Arg.}[\underset{\boldsymbol{\theta}}{\text{Min.}}\ E\{z - \boldsymbol{\theta}'\boldsymbol{\varphi}(\mathbf{x})\}^2]$$

In other words, $\boldsymbol{\theta}^o$ is the solution of the following equation:

$$E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\boldsymbol{\theta}^o - \boldsymbol{\varphi}(\mathbf{x})z\} = 0 \tag{9.18}$$

The algorithm of Section II, Eq. (9.3) for the solution of Eq. (9.18) for $\boldsymbol{\theta}^o$ can be written as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \rho_k\boldsymbol{\varphi}(\mathbf{x}_k)[z_k - \boldsymbol{\theta}_k'\boldsymbol{\varphi}(\mathbf{x}_k)] \tag{9.19}$$

with the gain sequence ρ_k obeying Eq. (9.4).

In addition to the conditions that $E\{\eta \mid \mathbf{x}\} = 0$ and that the samples $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are independent, we also must assume that $E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\}$ and $E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\ \boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\}$ exist, are positive definite, and that $E\{\boldsymbol{\varphi}(\mathbf{x})f(\mathbf{x})\}$ and $E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\ \boldsymbol{\varphi}(\mathbf{x})f(\mathbf{x})\}$ exist. If these conditions are met, then $\boldsymbol{\theta}_k$ converges to $\boldsymbol{\theta}^o$ with probability one and in mean square (for a proof see Appendix 2).

Alternately, we can use a second order method to solve Eq. (9.18). Note that the derivative with respect to $\boldsymbol{\theta}$ of the left hand side of Eq. (9.18) is $E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\}$; hence, Newton's method for solving Eq. (9.18) can be written as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \rho_k[E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\}]^{-1}\ \boldsymbol{\varphi}(\mathbf{x}_k)\{z(k) - \boldsymbol{\theta}_k'\boldsymbol{\varphi}(\mathbf{x}_k)\} \tag{9.20}$$

At the kth stage, a good approximation for $E\{\boldsymbol{\varphi}(\mathbf{x})\ \boldsymbol{\varphi}'(\mathbf{x})\}$ is

$$\left\{\frac{1}{k}\sum_{j=1}^{k}\boldsymbol{\varphi}_j\boldsymbol{\varphi}_j' + B_0^{-1}\right\},$$

where B_0^{-1} is the a priori estimate of the covariance matrix of $\boldsymbol{\varphi}(\mathbf{x})$; hence, the algorithm in Eq. (9.20) can be written as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + B_k\boldsymbol{\varphi}(\mathbf{x}_k)\{z_k - \boldsymbol{\theta}_k'\boldsymbol{\varphi}(\mathbf{x}_k)\} \tag{9.21}$$

where*

$$B_k = \left\{\frac{1}{k}\sum_{j=1}^{k} \varphi_j \varphi_j' + B_0^{-1}\right\}^{-1} \tag{9.22}$$

Equation (9.22) can be written in a form which does not involve the explicit inversion of a matrix at every instant

$$B_k = B_{k-1} - \frac{B_{k-1}\varphi_k \varphi_k' B_{k-1}}{1 + \varphi_k' B_{k-1} \varphi_k}$$

The same assumptions that assured convergence of the first-order algorithm, Eq. (9.18), also assure the convergence of algorithm (9.20) with probability one.

It has been observed experimentally that algorithm (9.21) has a faster rate of convergence than that of (9.20). We will elaborate upon this fact later.

If the function $f(\mathbf{x})$ did possess an expansion of the following type

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \varphi_i(\mathbf{x}) = \boldsymbol{\alpha}' \boldsymbol{\varphi}(\mathbf{x})$$

then $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}^o$ would be identical; i.e., the best approximation to the function $f(\mathbf{x})$ is identical with the function $f(\mathbf{x})$ itself.

C. *The Method of Potential Functions*†

In the previous section, we developed algorithms for obtaining the best linear approximation to the unknown function $f(\mathbf{x})$. These algorithms do not lead to the unknown function itself, except in special cases. In this section we will treat the same problem as that treated in Section IVB and develop an algorithm to recover the function $f(\mathbf{x})$ itself from the noisy observations. These algorithms are obtained by the method of the potential functions, introduced by Aizerman *et al.* (1964). The method has been used for the recovery of unknown, uniformly bounded continuous functions which may be either deterministic or stochastic. The unknown function is recovered as an infinite sum of certain known functions.

With very little loss of generality, the unknown function $f(\mathbf{x})$ can be assumed to have the following infinite expansion

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} \theta_j \varphi_j(\mathbf{x}) \tag{9.23}$$

---

* $\varphi_j = \varphi(\mathbf{x}_j)$.

† Additional discussions on the method of potential functions are given in Chapters 3 and 5.

where the functions $\varphi_1(\mathbf{x})$, $\varphi_2(\mathbf{x})$,... are orthonormal functions with a measure $\mu(\cdot)$ satisfying Eq. (9.24), and θ_1 , θ_2 ,... are unknown parameters.

$$\int_{\Omega_x} \varphi_i(\mathbf{x})\, \varphi_j(\mathbf{x})\, d\mu(\mathbf{x}) = \delta_{ij} \tag{9.24}$$

In Eq. (9.23) Ω_x denotes the region where the function $f(\cdot)$ is defined. The unknown coefficients θ_1 , θ_2 ,... must satisfy the condition in Eq. (9.25) so that $f(\mathbf{x})$ may exist:

$$\sum_{i=1}^{\infty} \theta_i^2 < \infty \tag{9.25}$$

We will introduce a known function $K(\mathbf{x}, \mathbf{y})$ of 2, n-vectors $\mathbf{x}$ and $\mathbf{y}$, the so-called potential function, which obeys the following conditions:

$$\begin{aligned} &K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x}) \\ &\operatorname*{Max}_{\mathbf{x}}.\ K(\mathbf{x}, \mathbf{x}) \leqslant c_2 \end{aligned} \tag{9.26}$$

and

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i^2 \varphi_i(\mathbf{x})\, \varphi_i(\mathbf{y}) \tag{9.27}$$

where

$$\lambda_i < \infty$$

A typical potential function obeying Eqs. (9.26) and (9.27) is $\exp[-c^2 \| \mathbf{x} - \mathbf{y} \|^2]$. We will have to return to the choice of the potential function later.

Let us define $f_k(\mathbf{x})$ as the estimate of the unknown function $f(\mathbf{x})$, at the kth stage using observations $\{\mathbf{x}_i , z_i , i = 1, 2,..., k - 1\}$ where z_i obeys Eq. (9.17). The estimate $f_k(\mathbf{x})$ is assumed to obey the following recursion equation:

$$f_{k+1}(\mathbf{x}) = f_k(\mathbf{x}) + \gamma_k K(\mathbf{x}, \mathbf{x}_k) \tag{9.28}$$

Thus, $K(\mathbf{x}, \mathbf{x}_k)$ represents the weighting function; it is based on $\mathbf{x}_k$, the point at which the latest measurement is made. The term γ_k is the (numerical) weight attached to the correction term $K(\mathbf{x}, \mathbf{x}_k)$; it incorporates the information about the latest observation pair $(\mathbf{x}_k , z_k)$. γ_k should be so chosen that $f_k(\mathbf{x})$ tends to the unknown $f(\mathbf{x})$ in some sense. Such a choice for γ_k for our problem is,

$$\gamma_k = \rho_k(z_k - f(\mathbf{x}_k)) \tag{9.29}$$

where ρ_k is the usual scalar gain sequence obeying Eq. (9.4). Thus, the estimate $f_{k+1}(\mathbf{x})$ is a weighted sum of $(k+1)$ known functions, $K(\mathbf{x}, \mathbf{x}_1),\ldots, K(\mathbf{x}, \mathbf{x}_k)$, and $f_0(\mathbf{x})$; that is to say,

$$f_{k+1}(\mathbf{x}) = f_0(\mathbf{x}) + \gamma_1 K(\mathbf{x}, \mathbf{x}_1) + \gamma_2 K(\mathbf{x}, \mathbf{x}_2) + \cdots + \gamma_k K(\mathbf{x}, \mathbf{x}_k)$$

The weights γ_k are recursively computed from Eq. (9.29). Braverman (1965) has shown that the algorithm in Eqs. (9.28) and (9.29) converges to the unknown function $f(\mathbf{x})$ in probability.

There are two methods for constructing the potential function. In the first method, we start with a set of functions $\psi_i(\mathbf{x})$, $i = 1, 2,\ldots$ which are complete in some sense and construct the series

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i^2 \psi_i(\mathbf{x})\, \psi_i(\mathbf{y}) \tag{9.30}$$

where the numbers λ_i, $i = 1, 2,\ldots$ are so chosen that the infinite series in Eq. (9.30) can be summed analytically.

The second method is to select a symmetrical function of two variables $\mathbf{x}$ and $\mathbf{y}$ and use it as a potential function as long as it can be expanded in an infinite series as in Eq. (9.27). It is convenient to regard $K(\mathbf{x}, \mathbf{y})$ as a distance between the points $\mathbf{x}$ and $\mathbf{y}$. According to Braverman (1965), the potential function can have the following form

$$K(\mathbf{x}, \mathbf{y}) = g(\| \mathbf{z} \|), \qquad \mathbf{z} = \mathbf{x} - \mathbf{y}, \qquad \| \mathbf{z} \| = \sqrt{z_1^2 + \cdots + z_n^2}$$

where $g(\| \mathbf{z} \|)$ can be any function which has a multidimensional fourier transform that is positive everywhere. Two typical examples of $K(\mathbf{x}, \mathbf{y})$ constructed in this manner are given below.

(i) $$K(\mathbf{x}, \mathbf{y}) = \exp[-c \| \mathbf{x} - \mathbf{y} \|^2] \qquad c > 0$$

(ii) $$K(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \| \mathbf{x} - \mathbf{y} \|^2} \qquad \text{if} \quad \| \mathbf{x} - \mathbf{y} \|^2 < 1$$

The potential function method has the advantage that it imposes very few restrictions on the unknown function. The disadvantage is that it involves iteration in function space which gives rise to serious storage problems since the storage of the estimate $f_k(\mathbf{x})$ at the kth stage involves the storage of all the previous k potential functions $K(\mathbf{x}, \mathbf{x}_1),\ldots, K(\mathbf{x}, \mathbf{x}_{k-1})$. An infinite number of iterations (or an infinite sum of functions) finally converges to the correct form for the unknown function. The methods of stochastic approximation, instead, estimate parameters and they involve only a storage of an n-dimensional vector.

The potential function method reduces to the stochastic approximation when the function to be estimated can be exactly expressed as a finite sum

$$f(\mathbf{x}) = \boldsymbol{\theta}'\boldsymbol{\varphi}(\mathbf{x}) = \sum_{i=1}^{n} \theta_i \varphi_i(\mathbf{x}) \tag{9.31}$$

This will be demonstrated by chosing the potential function as follows

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \varphi_i(\mathbf{x})\, \varphi_i(\mathbf{y}) = \boldsymbol{\varphi}'(\mathbf{x})\, \boldsymbol{\varphi}(\mathbf{y}) \tag{9.32}$$

Let

$$f_k(\mathbf{x}) = \boldsymbol{\theta}_k'\boldsymbol{\varphi}(\mathbf{x}) \tag{9.33}$$

Substituting Eqs. (9.31)–(9.33) in the algorithm of Eqs. (9.28) and (9.29) we get

$$\boldsymbol{\varphi}'(\mathbf{x})[\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k - \rho_k \boldsymbol{\varphi}(\mathbf{x}_k)(z_k - \boldsymbol{\theta}_k'\boldsymbol{\varphi}(\mathbf{x}_k)) = 0 \tag{9.34}$$

Removing $\boldsymbol{\varphi}'(\mathbf{x})$ from Eq. (9.34) gives us the stochastic approximation scheme in Eq. (9.19) for recovering the unknown parameter $\boldsymbol{\theta}^o$.

V. Convergence Rates

One measure of the rate of convergence of an algorithm is the mean square error $E\{\| \boldsymbol{\theta}_k - \boldsymbol{\theta}^o \|^2\}$ where $\boldsymbol{\theta}^o$ is the true unknown value and $\boldsymbol{\theta}_k$ is the estimate of $\boldsymbol{\theta}^o$ at the kth stage. This measure is convenient for evaluation in most of the algorithms discussed earlier. We will consider the algorithms of the Robbins–Munro type and those of Kiefer–Wolfowitz type separately.

A. *Algorithms of the Type of Robbins–Munro*

In this class we include not only algorithms mentioned in Section II, but also the algorithm in Eq. (9.19) in Section IV. We will consider only the case in which the gain ρ_k is of the form

$$\rho_k = \rho_1/k, \qquad \rho_1 > 0$$

Let $b_k = E\{\| \boldsymbol{\theta}_k - \boldsymbol{\theta}^o \|^2\}$

and $$E\{[\mathbf{y}_k - E\{y(\boldsymbol{\theta}_k)\}]^2\} = \frac{1}{k^{(d-1)}} \qquad d \geqslant 1$$

It is possible to show that b_k obeys the following difference equation

$$b_{k+1} \leqslant \left(1 - \frac{d_1}{k}\right) b_k + d_2 k^{-(1+d)}$$

where d_1 , $d_2 > 0$.
The above equation has the following asymptotic solution (Chung, 1954; Venter, 1966)*

$$\begin{aligned} b_k &= O(k^{-d}) \quad \text{if} \quad d_1 > d > 0 \\ &= O(k^{-d_1}) \quad \text{if} \quad d_1 = d > 0 \\ &= O(k^{-d_1}) \quad \text{if} \quad d > d_1 > 0 \end{aligned}$$

The usual value of d is unity. This is so in the case of the algorithm in Eq. (9.19) in Section IV when the variance of the measurement noise is constant.

B. *Algorithms of the Type of Kiefer–Wolfowitz*

In addition to the assumptions made in the propositions in Section III, we need the following conditions:

(i) Let $r(\theta) = \nabla_\theta J(\theta)$ where $J(\theta)$ is the criterion function being minimized. There exist two positive numbers k_0 and k_1 so that

$$k_0 \| \theta - \theta^o \|^2 \leqslant (\theta - \theta^o)' \, r(\theta) \qquad \text{and} \qquad \| r(\theta) \| \leqslant k_1 \| \theta - \theta^o \|$$

(ii) Let $J(\theta)$ possess derivatives of order not less than $(s + 1)$, $s \geqslant 2$; and

$$\rho_k = \frac{\rho_1}{k^\alpha} \qquad c_k = \frac{c_1}{k^\gamma}$$

$$\rho_1 > 0 \qquad c_1 > 0 \qquad 0 < \alpha \leqslant 1 \qquad 0 < \gamma < \alpha/2$$

$$2k_0\rho_1 > \beta \qquad \text{if} \quad \alpha = 1 \qquad \text{where} \quad \beta = \min\{2s\gamma, \alpha - 2\gamma\}$$

Then $E\{\| \theta_k - \theta^o \|^2\} = O(k^{-\beta})$.

Hence, the speed of Kiefer–Wolfowitz (KW) scheme is considerably less than that of Robbins–Munro (RM); therefore, KW should be used only if we cannot formulate the scheme as an RM procedure.

C. *Least Squares Algorithms*

We will consider only the algorithm in Eqs. (9.21) and (9.22). Experimentally this algorithm is observed to converge faster than the first order

* $g(n) \triangleq O(g(n))$ means Lim. Sup $| g(n)/h(n) | < + \infty$.

algorithm in Eq. (9.19). But the upper bound for the mean square error $E\{\| \theta_k - \theta^o \|^2\}$ is still of the form $1/k$, as before, and we cannot do better than this on account of the Cramer–Rao inequality. However, we can demonstrate the superiority of the scheme in Eqs. (9.21) and (9.22) in the following manner (Wagner, 1968):

Let us assume that there exists an integer $t > 1$ such that

$$E\{| \varphi_i(\mathbf{x})|^{2t}\} < \infty \qquad \forall i = 1,..., n$$

$$E\{| f(\mathbf{x})|^{2t}\} < \infty$$

$$E\{| \eta_i |^t\} < \infty \qquad \forall i$$

Then the probability that θ_k lies outside a sphere of radius ϵ centered around θ^o obeys the following upper bound

$$\text{Prob.}[\| \theta_k - \theta^o \| \geqslant \epsilon] \leqslant \frac{\alpha_1(\epsilon)}{k^{t-1}}$$

$$\text{Prob.} \left\{ \bigcup_{j=k}^{\infty} [\| \theta_j - \theta^o \| \geqslant \epsilon] \right\} \leqslant \frac{\alpha_2(\epsilon)}{k^{t-2}}$$

where $\alpha_1(\epsilon)$ and $\alpha_2(\epsilon)$ are positive constants depending on ϵ. In other words, the more moments which can be assumed to exist for the quantities $| \varphi_i(\mathbf{x})|$, $| f(\mathbf{x})|$ and $| \eta_k |$ the higher will be the convergence rate. When the above mentioned functions have moments of all orders (that is, $t = \infty$), then θ_k tends to θ^o at an exponential rate, i.e.,

$$\text{Prob.}[\| \theta_k - \theta^o \| \geqslant \epsilon] \leqslant b_1(\epsilon)(\rho(\epsilon))^k$$

$$\text{Prob.} \left[\bigcup_{j=k}^{\infty} \| \theta_j - \theta^o \| \geqslant \epsilon \right] \leqslant b_2(\epsilon)(\rho(\epsilon))^k$$

where $b_1(\epsilon)$, $b_2(\epsilon)$ are positive functions and $0 < \rho(\epsilon) < 1$.

VI. Methods of Accelerating Convergence

Two approaches have been suggested for accelerating the convergence of stochastic approximation procedures. The first approach is to accelerate convergence by selecting a proper weighting sequence $\{\rho_k\}$. The choice of the weighting sequence based on information concerning the behavior of the regression function, intuitively speaking, should improve the rate of convergence. We will limit our discussion to problems in which θ is scalar. Historically, the first method of accelerating the convergence of a stochastic approximation procedure was proposed by Kesten (1958). The basic idea is that when the estimate is far from the

sought quantity θ^0 there will be few changes of sign of $(\theta_k - \theta_{k-1})$. Near the goal, θ^0, one would expect overshooting to cause oscillation from one side of θ^0 to the other. Kesten proposed, using the number of sign changes of $(\theta_k - \theta_{k-1})$, to indicate whether the estimate is near or far from θ^0. Specifically, the gain γ_k is not decreased if $(\theta_k - \theta_{k-1})$ retains its sign. Mathematically, the Kesten algorithm is,

$$\theta_{k+1} = \theta_k - \gamma_k y(\theta_k) \tag{9.35}$$

where $\gamma_1 = \rho_1$, $\gamma_2 = \rho_2$, ..., $\gamma_k = \rho_{s(k)}$

$$s(k) = 2 + \sum_{j=1}^{k} \Phi[(\theta_k - \theta_{k-1})(\theta_{k-1} - \theta_{k-2})]$$

and

$$\Phi[x] = \begin{cases} 1 & \text{if } \begin{cases} x < 0 \\ x > 0 \end{cases} \\ 0 \end{cases}$$

The algorithm in Eq. (9.35) converges with probability one.

Fabian (1965) has proposed the following accelerated algorithms:

$$\theta_{k+1} = \theta_k + \rho_k \operatorname{sgn}[y(\theta_k)] \tag{9.36}$$

for RM schemes, and

$$\theta_{k+1} = \theta_k + \frac{\rho_k}{c_k} \operatorname{sgn}[z(\theta_k + c_k) - z(\theta_k - c_k)] \tag{9.37}$$

for KW schemes. Algorithms (9.36) and (9.37) converge to their sought quantities, respectively, only in a comparatively narrow class of problems in which the distribution function of the random variable y is symmetric with respect to θ. Another scheme of accelerating convergence proposed by Fabian is analogus to the steepest descent method.

The second approach for accelerating convergence is by taking more observations at each stage of iteration. Intuitively speaking, the additional observations made at each stage will help us in exploring the regression function in greater detail than the original stochastic approximation procedure, and, consequently, the extra information can be utilized to improve the rate of convergence. Venter (1967a) and Fabian (1965) have proposed accelerated algorithms for RM and KW procedures, respectively. For example, Venter's procedure estimates the slope of the regression function by taking two observations at each stage and using this information to improve the rate of convergence and

the asymptotic variance of the Robbins–Munro procedure. The recursive algorithm is of the form

$$\theta_{k+1} = \theta_k - \rho_k \alpha_k^{-1} \tfrac{1}{2}(y_k' + y_k'') \tag{9.38}$$

where y_k' and y_k'' are random variables whose conditional distributions given y_j', y_j'', $j = 1,\ldots,(k-1)$ are identical to those of $y(\theta_k + c_k)$ and $y(\theta_k - c_k)$ respectively; $\rho_k = 1/k$, $c_k = ck^{-\gamma}$, $c > 0$, and $0 < \gamma < 1/2$; and, α_k is an estimate of the slope α of the regression function, $r(\theta)$, that is defined as follows:

Assume that $0 < a < \alpha < b < \infty$ with a and b known.
Let

$$b_k = k^{-1} \sum_{j=1}^{k} (y_k' - y_k'')/2c_k \tag{9.39a}$$

and

$$\alpha_k = \begin{cases} a & \text{if } b_k \leqslant a \\ b_k & \text{if } a \leqslant b_k \leqslant b \\ b & \text{if } b_k \geqslant b \end{cases} \tag{9.39b}$$

The algorithm in Eq. (9.38) converges with probability one.

The same idea can be carried over to the KW procedures. In this case, three observations are taken at each stage of an iteration, and the appropriate second-order differences of the observations are used to estimate the second-order derivative of the regression function at the maximum (or minimum). This information would be utilized to determine the next estimate θ_{k+1} of the maximum (or minimum). In a similar manner, Fabian, showed that the KW procedure can be modified in such a way as to be almost as speedy as the RM procedure. The modification consists of taking more observations at every stage of an iteration and utilizing this information to eliminate (smooth out) the effect of all higher-order derivates of the regression function.

When the unknown parameter $\boldsymbol{\theta}$ is a vector, it is very hard to develop schemes which will yield a faster rate of convergence than the standard algorithms in Sections II and III. One suggestion is to use the stochastic approximation algorithms in conjunction with Monte Carlo schemes for optimizing the multivariate functions (Newbold, 1967).

One general comment regarding all accelerating schemes is in order. In all of them, the upper bound for the mean square error $E\{\| \boldsymbol{\theta}_k - \boldsymbol{\theta}^o \|^2\}$ cannot decay faster than $(c_1/(k + c_2))$ where $c_1, c_2 > 0$. This statement follows from the Cramer–Rao inequality. In essence, the different schemes lead to different values of c_1 and c_2 in the above bound.

VII. Conclusion

We have concentrated our attention on developing methods of finding the minimum of stochastic functions. Our algorithms need very little information about the measurement noise. The unknown quantity was always a finite dimensional vector which did not vary with time.

We can easily modify the algorithms (Dupac, 1965) so as to estimate parameters which are slowly time-varying and are varying in either a deterministic manner, or in a random fashion with their variances going to zero as time tends to infinity. However, if the parameter is a stochastic process, with variance not decreasing with time, then the problem is one in nonlinear estimation of stochastic processes and is, therefore, outside the scope of this chapter.

APPENDIX 1

Proof of the Convergence of the Basic Stochastic Approximation Scheme

In this appendix, we will demonstrate that the θ_k given by the algorithm in Eq. (3) [rewritten below as Eq. (40)] will converge to the true value θ^o under the assumptions (A1)–(A3) of Section II. The proof is based on a technique developed by Gladyshev (1965).

$$\theta_{k+1} = \theta_k - \rho_k \mathbf{y}(\theta_k) \tag{9.40}$$

Let

$$\tilde{\theta}_k = \theta_k - \theta^o \tag{9.41}$$

Subtract θ^o from both sides of Eq. (9.40), square both sides of Eq. (9.40) and take expectations using Eq. (9.1) and assumption (A3).

$$\begin{aligned} &E\{\| \tilde{\theta}_{k+1} \|^2 \mid \theta_1, \ldots, \theta_k\} \\ &\quad = \| \tilde{\theta}_k \|^2 - 2\rho_k \tilde{\theta}_k' \mathbf{r}(\theta_k) + \rho_k^2 E\{\| \mathbf{y}(\theta_k) \|^2 \mid \theta_1, \ldots, \theta_k\} \\ &\quad \leqslant \| \tilde{\theta}_k \|^2 - 2\rho_k \tilde{\theta}_k' \mathbf{r}(\theta_k) + \rho_k^2 h(1 + \| \tilde{\theta}_k \|^2) \end{aligned} \tag{9.42}$$

By assumption (A2), Eq. (9.42) becomes

$$E\{\| \tilde{\theta}_{k+1} \|^2 \mid \theta_1, \ldots, \theta_k\} \leqslant (1 + h\rho_k^2) \| \tilde{\theta}_k \|^2 + h\rho_k^2 \tag{9.43}$$

Let us define the scalar α_k to be

$$\alpha_k = \| \tilde{\theta}_k \|^2 \prod_{j=k}^{\infty} (1 + h\rho_j^2) + \sum_{j=k}^{\infty} h\rho_j^2 \prod_{m=j+1}^{\infty} (1 + h\rho_m^2) \tag{9.44}$$

Using Eqs. (9.43) and (9.44), we can write a recursive inequality for α_k

$$E\{\alpha_{k+1} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\} \leqslant \alpha_k \tag{9.45}$$

Now let us take the conditional expectation on both sides of Eq. (9.45), given $\alpha_1, \ldots, \alpha_k$:

$$E\{\alpha_{k+1} \mid \alpha_1, \ldots, \alpha_k\} \leqslant \alpha_k \tag{9.46}$$

Inequality (9.46) shows that α_k is a semi-martingale where

$$E\{\alpha_{k+1}\} \leqslant E\{\alpha_k\} \leqslant \cdots \leqslant E\{\alpha_1\} < \infty \tag{9.47}$$

According to the theory of semi-martingales [Doob, 1953], the sequence α_k converges with probability one and, hence, by Eqs. (9.44) and (9.4), $\| \tilde{\boldsymbol{\theta}}_k \|^2$ tends to some random variable, say ξ, with probability one. It remains to show that $\xi = 0$ with probability one.

For this, we note that the boundedness of the sequence $E\{\| \tilde{\boldsymbol{\theta}}_k \|^2\}$ follows from Eqs. (9.47), (9.44) and (9.4). Further, let us take expectation over both sides of Eq. (9.42):

$$E \| \tilde{\boldsymbol{\theta}}_{k+1} \|^2 - E \| \tilde{\boldsymbol{\theta}}_k \|^2 \leqslant \rho_k{}^2 h(1 + E \| \tilde{\boldsymbol{\theta}}_k \|^2) - 2\rho_k E\{\tilde{\boldsymbol{\theta}}_k{}' \mathbf{r}(\boldsymbol{\theta}_k)\} \tag{9.48}$$

Repeated use of Eq. (9.48) gives us

$$E \| \tilde{\boldsymbol{\theta}}_{k+1} \|^2 - E \| \tilde{\boldsymbol{\theta}}_1 \|^2 \leqslant \sum_{j=1}^{k} h\rho_j{}^2(1 + E \| \tilde{\boldsymbol{\theta}}_j \|^2) - 2 \sum_{j=1}^{k} \rho_j E\{\tilde{\boldsymbol{\theta}}_j{}' \mathbf{r}(\boldsymbol{\theta}_j)\} \tag{9.49}$$

Equation (9.48), the boundedness of $E \| \tilde{\boldsymbol{\theta}}_k \|^2$, and Eq. (9.4) imply that

$$\sum_{j=1}^{k} \rho_j E\{\tilde{\boldsymbol{\theta}}_j' \mathbf{r}(\boldsymbol{\theta}_j)\} < \infty \tag{9.50}$$

Since $\tilde{\boldsymbol{\theta}}_j{}' \mathbf{r}(\boldsymbol{\theta}_j)$ is nonnegative by assumption (A2), Eq. (9.50) and the property of the $\{\rho_k\}$ sequence implies the existence of a subsequence $\{k_j\}$ such that

$$\tilde{\boldsymbol{\theta}}_{k_j}' \mathbf{r}(\boldsymbol{\theta}_{k_j}) \to 0 \text{ with probability one} \tag{9.51}$$

Equation (9.51), Assumption (A1) and the fact that $\tilde{\boldsymbol{\theta}}_k$ tends to some random variable imply $\tilde{\boldsymbol{\theta}}_k$ tends to $\mathbf{0}$ with probability one.

To prove mean square convergence, we note that by Assumption (A2) there exists a positive constant $d_2(\epsilon)$ so that

$$(\tilde{\theta}_k)'\mathbf{r}(\theta_k) \geqslant d_2(\epsilon)\| \tilde{\theta}_k \|^2 \quad \forall \| \tilde{\theta}_k \| < \epsilon, \qquad \epsilon > 0 \tag{9.52}$$

On account of the definition of the $\{\rho_k\}$ sequence, there exists an integer k, so that

$$2d_2(\epsilon) - \rho_k h \triangleq d_3(\epsilon) > 0 \qquad \forall k > k_1 \tag{9.53}$$

Using Eqs. (9.52), (9.53) and (9.47), it is easy to show

$$E\| \tilde{\theta}_{k+1} \|^2 \leqslant E\| \tilde{\theta}_k \|^2\{1 - \rho_k d_3(\epsilon)\} + \rho_k{}^2 h, \qquad \forall k > k_1 \tag{9.54}$$

Let us define the scalar sequence $\{\beta_k\}$,

$$\beta_k = (E\| \tilde{\theta}_k \|^2 \prod_{j=k}^{\infty} (1 - d_3\rho_j)) + \sum_{j=k}^{\infty} h\rho_j{}^2 \prod_{m=j+1}^{\infty} (1 - d_3\rho_j) \tag{9.55}$$

From Eqs. (9.55) and (9.54), we get

$$\beta_{k+1} \leqslant \beta_k \qquad \forall k > k_1 ;$$

thus, β_k is bounded from above. We see in Eq. (9.55) that the second term is bounded as k tends to infinity, and that the term $\prod_{j=k}^{\infty} (1 - d_3\rho_j)$ tends to infinity with k; hence, if β_k is to be bounded, the term $E\{\| \tilde{\theta}_k \|^2\}$ should tend to zero as k goes to infinity; thus completing the proof of mean square convergence.

APPENDIX 2

Convergence Proof of the Function Recovery Algorithm in Eq. (9.19)

The algorithm in Eq. (9.19) is rewritten here, as Eq. (9.56):

$$\theta_{k+1} = \theta_k + \rho_k \varphi(\mathbf{x}_k)[z_k - \theta_k{}'\varphi(\mathbf{x}_k)] \tag{9.56}$$

where $z_k = f(\mathbf{x}_k) + \eta_k$.

Let

$$\theta^o = [E\{\varphi(\mathbf{x})\, \varphi'(\mathbf{x})\}]^{-1} E[\varphi(\mathbf{x}) f(\mathbf{x})] \tag{9.57}$$

We will show that the sequence $\{\theta_k\}$ given by the algorithm in Eq. (9.56)

tends to θ^o defined above in the mean square sense and with probability one if the following conditions (i)–(v) are satisfied:

(i) The samples $\mathbf{x}_k$, $k = 1, 2,...,$ are statistically independent.
(ii) $E\{\varphi(\mathbf{x})\,\varphi'(\mathbf{x})\}$ and $E\{\varphi(\mathbf{x})\,\varphi'(\mathbf{x})\,\varphi(\mathbf{x})\,\varphi'(\mathbf{x})\}$ must be finite and positive definite.
(iii) $\| E\{\varphi(\mathbf{x})f(\mathbf{x})\}\| < \infty$.
$\| E\{\varphi(\mathbf{x})\,\varphi'(\mathbf{x})\,\varphi(\mathbf{x})f(\mathbf{x})\}\| < \infty$.
(iv) $E\{\eta_k^2\} < \infty$ and $E\{\eta_k\} = 0$.
(v) $E\{\| \theta_0 \|^2\} < \infty$.

Let

$$\begin{aligned} \tilde{\theta}_k &= \theta^o - \theta_k\,, \qquad \varphi_i = \varphi(\mathbf{x}_i) \\ \xi_i &= z_i - \theta^{o\prime}\varphi_i = f(\mathbf{x}_i) - \theta^{o\prime}\varphi_i + \eta_i \end{aligned} \tag{9.58}$$

The algorithm in Eq. (9.56) can be written as

$$\tilde{\theta}_{k+1} = (I - \rho_k\varphi_k\varphi_k')\tilde{\theta}_k - \rho_k\varphi_k\xi_k \tag{9.59}$$

Taking expectations on both sides of Eq. (9.59), noting $E(\varphi_k\xi_k) = \mathbf{0}$ and using Condition (i), we get

$$E\{\tilde{\theta}_{k+1}\} = [I - \rho_k E\{\varphi_k\varphi_k'\}]\,E(\tilde{\theta}_k\} \tag{9.60}$$

Taking the Euclidean norm on both sides of Eq. (9.60), we get

$$\| E\tilde{\theta}_{k+1} \| \leqslant (1 - d_4\rho_k)\| E\tilde{\theta}_k \| \tag{9.61}$$

where d_4 is the minimum eigenvalue of the matrix $E\{\varphi\varphi'\}$. Equation (9.61) in conjunction with the property of $\{\rho_k\}$ in Eq. (9.4) implies that

$$\lim_{k\to\infty} \| E\tilde{\theta}_k \| = 0 \tag{9.62}$$

Let us scalar multiply Eq. (9.59) with itself and take the conditional expectation:

$$\begin{aligned} E\{\| \tilde{\theta}_{k+1} \|^2 \mid \theta_k\} &= E\{\| \theta_k \|^2_{(I-\rho_1\varphi_k\varphi_k')^2} \mid \theta_k\} \\ &\quad + \rho_k^2[E\{\xi_k^2 \| \varphi_k \|^2 \mid \theta_k\} + 2E\{\tilde{\theta}_k'\varphi_k\varphi_k'\xi_k \mid \theta_k\} \\ &\quad - 2\rho_k E\{\tilde{\theta}_k'\varphi_k\xi_k \mid \theta_k\} \end{aligned} \tag{9.63}$$

Let us consider the various terms in Eq. (9.63) separately.*

* $\lambda_{\max}[A] \triangleq$ maximum eigenvalue of matrix A.

$$E\{\| \tilde{\theta}_k \|^2_{(I-\rho_k\varphi_k\varphi_k')^2} \mid \theta_k\} = \| \tilde{\theta}_k \|^2_{E\{(I-\rho_k\varphi_k\varphi_k')^2\}}$$

$$\leqslant \| \tilde{\theta}_k \|^2_{\lambda_{\max}[E\{(I-\rho_k\varphi_k\varphi_k')^2\}]}$$

$$\leqslant \| \tilde{\theta}_k \|^2_{\lambda_{\max}[I-(2-\epsilon_1)\rho_k E\{\varphi_k\varphi_k'\}]} \quad \forall k > k_1 \quad \text{and} \quad 0 < \epsilon_1 < 2$$

$$\leqslant \| \tilde{\theta}_k \|^2(1 - d_5\rho_k) \tag{9.64}$$

where $d_5 = (2 - \epsilon_1)\, d_4$

By Conditions (i), (iii), and (iv),

$$E\{\xi_k{}^2 \| \varphi_k \|^2 \mid \theta_k\} \leqslant E\{(f(\mathbf{x}_k) - \theta^{o\prime}\varphi_k)^2 \| \varphi_k \|^2 + \eta_k{}^2 \| \varphi_k \|^2\} \leqslant d_6 < \infty \tag{9.65}$$

where d_6 is a constant.

Further,

$$E(\tilde{\theta}_k{}'\varphi_k\varphi_k{}'\varphi_k\xi_k \mid \theta_k) = (E\{\tilde{\theta}_k\})'\, E\{\varphi_k\varphi_k{}'\varphi_k\xi_k\} = 0 \tag{9.66}$$

Similarly,

$$E\{\tilde{\theta}_k{}'\varphi_k\xi_k \mid \theta_k\} = 0 \tag{9.67}$$

Substituting Eqs. (9.64)–(9.67) in Eq. (9.63), we get

$$E\{\| \tilde{\theta}_{k+1} \|^2 \mid \theta_k\} \leqslant \| \tilde{\theta}_k \|^2(1 - d_5\rho_k) + d_6\rho_k{}^2 \tag{9.68}$$

Taking expectations on either side of Eq. (9.68), we get

$$E \| \tilde{\theta}_{k+1} \|^2 \leqslant (1 - d_5\rho_k) \| \tilde{\theta}_k \|^2 + d_6\rho_k{}^2 \tag{9.69}$$

We have already shown in Appendix 1 that the $E \| \tilde{\theta}_k \|^2$, given by Eq. (9.69), tends to zero with k, thus completing the proof.

The proof of convergence with probability one is omitted since it is given in Kashyap and Blaydon (1966).

References

Albert, A. and Gardner, L., "Stochastic Approximation and Nonlinear Regression." M.I.T. Press, Cambridge, Mass., 1967.

Aizerman, M. A., Braverman, F. M., and Rozonoer, L. I., The probability problem in pattern recognition learning and the method of potential functions. *Automation and Remote Control 25*, No. 9 (1964a).

Aizerman, M. A., Braverman, F. M., and Rozonoer, L. I., The method of potential functions in the problem of determining the characteristics of a function generator from randomly observed points. *Automation and Remote Control 25*, No. 12 (1964b).

Braverman, E. M., On the method of potential functions. *Automation and Remote Control 26*, No. 12 (1965).

Chung, K. L., On a stochastic approximation method. *Ann. Math. Stat. 25*, No. 4 (1954).

Dupac, V., A dynamic stochastic approximation method. *Ann. Math. Stat. 38*, Feb. (1967).

Fabian, V., Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Stat. 36*, Dec. (1965).

Gladyshev, E. A., On stochastic approximation. *Theory of Prob. and Appl. 10*, No. 2 (1965).

Kashyap, R. L. and Blaydon, C. C., Recovery of functions from noisy measurements taken at randomly selected points and its application to pattern classification. *Proc. IEEE 54*, pp. 1127–1129 (1966).

Kesten, H., Accelerated stochastic approximation methods. *Ann. Math. Stat. 29*, No. 1 (1958).

Kiefer, J. and Wolfowitz, J., Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat. 23*, No. 3 (1952).

Loginov, N. V., Stochastic approximation methods. *Automation and Remote Control 27*, No. 4 (1966).

Newbold, P. M., A stochastic approximation scheme with accelerated convergence properties. Tech. Rept. No. 545, Division of Engineering and Applied Physics, Harvard Univ. Cambridge, Mass., 1967.

Robbins, H. and Monro, S., A stochastic approximation method. *Ann. Math. Stat. 22*, No. 1 (1951).

Schmetterer, L., Stochastic approximation. *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, Vol. I (1961).

Venter, J. H., On Dvoretzky stochastic approximation theorems. *Ann. Math. Stat. 37*, No. 4 (1966).

Venter, J. H., An extension of the Robbins–Munro procedure. *Ann. Math. Stat. 38*, No. 2 (1967a).

Venter, J. H., On the convergence of the Kiefer–Wolfowitz approximation procedure. *Ann. Math. Stat. 38*, No. 4, pp. 1031–1036 (1967b).

Wagner, T. J., The rate of convergence of an algorithm for recovering functions from noisy measurements taken at randomly selected points. *IEEE Trans. System Science and Cybernetics*, July (1968).