1. Introduction

Despite the many significant and elegant theoretical developments of the past several decades, the art of statistical inference on time series is, from the applied point of view, in its infancy. An important class of problems, which has been relatively neglected, arises from the fact that there are always computations associated with statistical procedures; a procedure which is "optimal" in the decision theoretic sense can be somewhat less than optimal from a practical point of view if the associated computations are prohibitively lengthy. This dilemma is compounded when we consider a time series as a *flow* of data. In "space age" applications, it is especially important that statistical procedures keep pace with the incoming data so that, at any instant, all of the available information has already been processed. The acquisition of new observations merely serves to update the current state of knowledge.

In this monograph we will investigate nonlinear regression from that point of view. Let

$$\{Y_n : n = 1, 2, \cdots\}$$

be a stochastic process whose mean-value sequence is a member of a family of known sequences, that is to say,

$$\mathscr{E} Y_n = F_n(\theta),$$

where θ is a vector parameter which is not known and must be estimated.

We will explore the asymptotic (large n) properties of recursive estimation schemes for $\boldsymbol{\theta}$ of the form

$$\mathbf{t}_{n+1} = \mathbf{t}_n + \mathbf{a}_n[Y_n - F_n(\mathbf{t}_n)], \tag{1.1}$$

where $\mathbf{t}_{n+1}$ is the estimate of $\boldsymbol{\theta}$ based upon the first n observations and $\{\mathbf{a}_n\}$ is a suitably chosen sequence of "smoothing vectors."

Without question, estimators of the type of Equation 1.1 are computationally appealing, provided the smoothing sequence is chosen reasonably. After each observation, we compute the prediction error $Y_n - F_n(\mathbf{t}_n)$ and correct $\mathbf{t}_n$ by adding to it the vector $[Y_n - F_n(\mathbf{t}_n)]\mathbf{a}_n$. Such recursions are sometimes called "differential correction" procedures.

In contrast, maximum-likelihood and least-squares estimation methods, although often efficient in the purely statistical sense, require the solution of systems of simultaneous nonlinear normal equations. If we want "running" values of these estimates, the computational problems are often great.

Of course, the choice of the weights $\mathbf{a}_n$ critically affects the computational simplicity and statistical properties of the recursive estimate (Equation 1.1). The main purpose of this monograph is to relate the large-sample statistical behavior of the estimates to the properties of the regression function and the choice of smoothing vectors.

Estimation schemes of the type of Equation 1.1 find their origins in Newton's method for finding the root of a nonlinear function. Suppose that $G(\cdot)$ is a monotone differentiable function of a real variable, and we wish to find the root θ of the equation

$$G(x) = 0.$$

If t_1 were known to be a reasonably good estimate of (i.e., is close to) θ, then

$$0 = G(\theta) \approx G(t_1) + (\theta - t_1)\dot{G}(t_1), \tag{1.2}$$

where the dot denotes differentiation. This equation says that $G(\theta)$ takes on nearly the same values as the line L which passes through the point $(t_1, G(t_1))$ with slope $\dot{G}(t_1)$ [i.e., is tangent to the curve $y = G(x)$ at $x = t_1$], provided that θ is not too far from t_1. Solving Equation 1.2 for θ, we see that

$$\theta \approx t_1 - \frac{G(t_1)}{\dot{G}(t_1)},$$

so that a potentially better estimator for θ might be (see Figure 1.1)

$$t_2 = t_1 - \frac{G(t_1)}{\dot{G}(t_1)}. \tag{1.3}$$

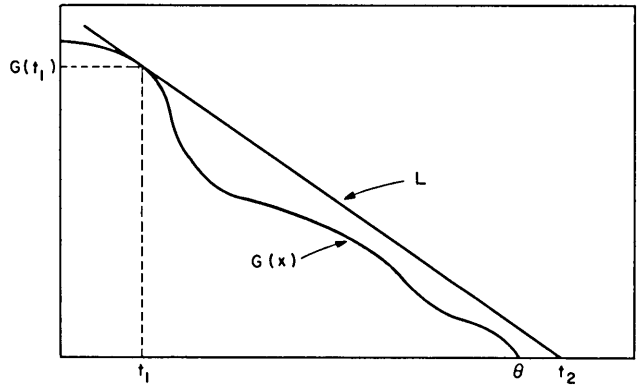

Figure 1.1 Graphical interpretation of Newton's method.

In turn, t_2 could be "improved" in the same way, and Equation 1.3 suggests that an ever-improving sequence of estimators for θ can be obtained by means of the recursion

$$t_{n+1} = t_n - \frac{G(t_n)}{\dot{G}(t_n)} \qquad (n \geq 1). \tag{1.4}$$

It would appear, though, that the first guess t_1 must be close to θ in order that the linear approximation, Equation 1.2, should be accurate. This is not essential if $|\dot{G}|$ is bounded above and away from zero:

$$0 < b \leq |\dot{G}(x)| \leq d < \infty.$$

We choose a number a to satisfy

$$0 < a \leq \frac{b}{d},$$

and we modify the recursion, Equation 1.4, to read

$$t_{n+1} = t_n - a_n G(t_n), \qquad a_n = \frac{a}{\dot{G}(t_n)}. \tag{1.5}$$

It is easy to show that t_n converges to θ as $n \to \infty$. Indeed, by the mean-value theorem, we obtain

$$G(t_n) = \dot{G}(u_n)(t_n - \theta), \tag{1.6}$$

where u_n lies between θ and t_n. Thus, by Equations 1.5 and 1.6, it follows that

$$t_{n+1} - \theta = [1 - a_n \dot{G}(u_n)](t_n - \theta) = \left\{\prod_{j=1}^{n}\left[1 - a\frac{\dot{G}(u_j)}{\dot{G}(t_j)}\right]\right\}(t_1 - \theta). \tag{1.7}$$

But

$$0 \leq 1 - a\frac{d}{b} \leq 1 - a\frac{\dot{G}(u_j)}{\dot{G}(t_j)} \leq 1 - a\frac{b}{d} < 1,$$

so that

$$|t_{n+1} - \theta| \leq \left(1 - a\frac{b}{d}\right)^n |t_1 - \theta| \to 0$$

as $n \to \infty$.

Let us now complicate matters by letting G vary with n. There is a sequence of monotone differentiable functions, G_n, all having a common root θ:

$$G_n(\theta) = 0 \qquad (n = 1, 2, \cdots).$$

Again, we estimate θ by sequences of the form

$$t_{n+1} = t_n - a_n G_n(t_n).$$

In precisely the same way, in place of Equation 1.7 we obtain

$$t_{n+1} - \theta = [1 - a_n \dot{G}_n(u_n)](t_n - \theta) = \left\{\prod_{j=1}^{n} [1 - a_j \dot{G}_j(u_j)]\right\}(t_1 - \theta).$$

Now assuming that

$$0 < b_n < |\dot{G}_n(x)| \leq Mb_n < \infty$$

for all n and all x, we choose a_n so that

1. a_n has the same sign as $\dot{G}_n$,
2. $|a_n| \leq \dfrac{1}{Mb_n}$,
3. $\sum\limits_n |a_n b_n| = \infty$.

Then we have

$$0 \leq \prod_{j=1}^{n} [1 - a_j \dot{G}(u_j)] \leq \prod_{j=1}^{n} (1 - |a_j b_j|) \to 0,$$

and $|t_{n+1} - \theta|$ tends once again to zero as $n \to \infty$.

This technique can be applied to the problem of discrete-time curve fitting: Suppose $Y_1, Y_2, \cdots$ is a sequence of numbers, and it is known that this sequence is one of a family of sequences, $\{F_n(\theta)\}$, indexed by a real parameter θ. Here θ is not known, and we wish to find that value of θ for which

$$Y_n = F_n(\theta) \qquad (n = 1, 2, \cdots).$$

If we let

$$G_n(x) = F_n(x) - Y_n,$$

the desired parameter value is that value of x which makes $G_n(x)$ vanish identically in n.

Now let noise be introduced, so that the sequence of observations, Y_n, are corrupted versions of $F_n(\theta)$:

$$Y_n = F_n(\theta) + W_n \qquad (n = 1, 2, \cdots),$$

where W_n is (zero mean) noise. Motivated by the previous discussion, we consider estimation schemes of the form

$$t_{n+1} = t_n + a_n[Y_n - F_n(t_n)], \tag{1.8}$$

which can be rewritten as

$$t_{n+1} = t_n - |a_n| Z_n(t_n). \tag{1.8a}$$

For every x, we can regard $Z_n(x)$ as an observable random variable with expectation equal to

$$G_n(x) = \operatorname{sgn} \dot{F}_n[F_n(x) - F_n(\theta)] = |\dot{F}_n(u_n)|(x - \theta), \tag{1.9}$$

where $u_n = u_n(x, \theta)$ lies between x and θ. Thus,

$$t_{n+1} - \theta = (1 - |a_n \dot{F}_n(u_n)|)(t_n - \theta) + a_n W_n, \tag{1.10}$$

and we are led to the study of certain first-order nonlinear difference equations with stochastic driving terms.

This brings to mind the literature associated with stochastic approximation, which dates back to a paper by Robbins and Monro (1951). That paper concerns itself with the problem of estimating the root, say α, of an unknown (time-homogeneous) regression function $G(x)$, which is the mean value of an observable random variable $Z(x)$. The distribution of the latter depends on a scalar parameter, x, which can be controlled by the experimenter. They proposed that α be estimated recursively by Equation 1.8*a*, where $Z(t_n)$ is the value of an observation taken at the "level" $x = t_n$, and $\{a_n\}$ is any nonsummable null sequence of scalars with $\sum_n a_n^2 < \infty$. The success of the Robbins–Monro procedure (it converges to α with probability one and in mean square under a wide range of conditions) encourages us to believe in the reasonableness of Equation 1.8.

Burkholder (1956) has studied processes of the form of Equation 1.8*a* in detail. In fact, he considers the more general situation where the root of G_n depends upon n but converges to a limit θ as $n \to \infty$. (This is not just an academic generalization, for such a result is needed in the treatment of the Kiefer–Wolfowitz procedure for locating the minimum of a

time-homogeneous regression function.) Consequently, there will be some overlap between his work and Chapters 2 through 4 of the present work. In fact, after appropriate reinterpretation of the symbols, we obtain some results that are significantly stronger than those given by Burkholder.

If we view the stochastic-approximation literature as a study in the asymptotic behavior of the solutions to a certain class of nonlinear first-order difference equations with stochastic driving terms, then the results of this monograph (particularly Chapters 3 and 4) serve to extend and complement many of the results in that literature, and accounts for our choice of title. However, our primary consideration is nonlinear regression *per se* and, for this reason, we often fail to state theorems with the weakest possible hypotheses; we want to keep their statements and proofs relatively simple.

We will treat the scalar-parameter case, Equation 1.8, and the general vector case, Equation 1.1, separately. For the vector-parameter case, we will treat the topics of strong consistency (probability-one convergence) and mean-square convergence. In the scalar-parameter case, we also treat the questions of convergence rates, asymptotic distribution theory, and efficiency. A wide class of gain sequences are examined. Some are deterministic, and some depend on the actual data which have been observed. Examples are sprinkled throughout the body of the monograph, and Chapter 8 is devoted exclusively to applications.

The techniques we use are, by now, standard to those who are familiar with the literature of stochastic approximation, but for the sake of the nonspecialist we have tried to keep our treatment self-contained. In all cases, we seek the asymptotic properties of the solutions to the intrinsically nonlinear difference equations of the type 1.1. We accomplish this by studying the asymptotic properties of certain linear difference equations which, in a sense, dominate the original ones.

Now a word about notation. In Chapters 6 through 9, we do not adhere to the convention which reserves lower-(resp. upper-)case boldface symbols for vectors (resp. matrices). The reader must keep in mind not only this point but also the orders of the various vectors and matrices involved. The symbol $a_n = O(b_n)$ means that $|a_n/b_n|$ has a finite limit superior as n tends to infinity, while $a_n = o(b_n)$ means the ratio tends to zero. The balance of the abbreviations are standard and are defined when they are first used.

We begin by studying the problems of probability-one and mean-square convergence in the scalar case.