INFORMATICA, 1992, Vol.3, No.1, 98-118

ON INVARIANCE PRINCIPLES FOR DISTRIBUTED PARAMETER IDENTIFICATION ALGORITHMS

George YIN[1] and Ben G. FITZPATRICK[2]

Department of Mathematics
Wayne State University, Detroit, MI 48202

Department of Mathematics
University of Tennessee, Knoxville, TN 37996

Abstract. We consider a class of identification algorithms for distributed parameter systems. Utilizing stochastic optimization techniques, sequences of estimators are constructed by minimizing appropriate functionals. The main effort is to develop weak and strong invariance principles for the underlying algorithms. By means of weak convergence methods, a functional central limit theorem is established. Using the Skorohod imbedding, a strong invariance principle is obtained. These invariance principles provide very precise rates of convergence results for parameter estimates, yielding important information for experimental design.

Key words: identification, distributed parameter system, stochastic optimization, invariance principle.

1. Introduction. In a wide range of applications, various problems have been formulated by using partial differential equations with appropriate boundary and initial conditions. Quite frequently, the underlying systems involve some unknown parameters, typically in the form of coefficients in the equation. As a consequence, distributed parameter identification, in which parameters are estimated from observed data, has witnessed rapid progress in recent years. To illustrate, we consider the following examples.

[1] Research of this author was supported in part by the National Science Foundation under grant DMS-9022139.

[2] Research of this author was supported in part by the Air Force Office of Scientific Research under grant AFOSR-91-0021.

EXAMPLE 1. The following differential equation models fluid transport in cat brain tissue:

$$u_t = (\mathcal{V}u)_x + (\mathcal{D}u)_{xx}.$$

Here u represents fluid concentration, and the parameters $\mathcal{V}$ and $\mathcal{D}$ are convection and diffusion coefficients, respectively. Banks and Kareiva (1983) (see also Banks and Fitzpatrick, 1989) used least squares techniques to fit this model to observed data, and went on to apply ANOVA-type hypothesis tests for $\mathcal{V} = 0$, in order to verify conjectures concerning the role of convection in grey and white matter.

EXAMPLE 2. The determination of damping terms in flexible structures is crucial to modeling and control objectives. Banks et. al. (1987) applied an Euler-Bernoulli model with viscous and Kelvin-Voigt damping terms:

$$u_{tt} + \frac{\partial^2}{\partial x^2}[EIu_{xx} + c_D I u_{xxt}] + \gamma u_t = f,$$

where EI is the stiffness, f is the forcing function, and $c_D I$ and γ are the Kelvin-Voigt and viscous damping coefficients, respectively. The function u represents displacement of the beam. Accelerometer data (u_{tt}) is used to determine the parameters. In Banks and Fitzpatrick (1989) statistical tests based on asymptotic normality results were applied to examine the importance of the viscous and Kelvin-Voigt damping models.

EXAMPLE 3. This model describes the predator-prey interactions in size structured populations:

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x}(\mathcal{V}(t,x,u_i)u_i) = \frac{\partial}{\partial x}\left(\mathcal{D}(t,x,u_i)\frac{\partial u_i}{\partial x}\right) + f_i(\lambda, \underline{u}).$$

In most experimental situations, one has observations of the number and sizes of individuals at various times (that is, one observes the solution of the differential equation) rather than values for the parameters $\mathcal{V}, \mathcal{D}$ and λ. Using the model to predict population size for resource management requires fitting data to determine the parameters.

To recover or identify the parameters in any of these examples, one needs to use observations. More often than not, such observations are corrupted with noise. In Fitzpatrick (1988) and Banks and Fitzpatrick (1990), a general nonlinear least squares type of algorithm was proposed for the distributed parameter identification problems. In Banks and Fitzpatrick (1989), Fitzpatrick (1988), Banks and Fitzpatrick (1990), the effects of noisy observations on the class of stochastic optimization and parameter estimation procedures were analyzed. In particular, consistency and asymptotic normality were established, with a primary objective of developing appropriate statistics for hypothesis tests.

This work complements the papers of Fitzpatrick (1988), Banks and Fitzpatrick (1990) by developing weak and strong functional invariance principles of the least squares algorithms for distributed parameter identification. Our main concerns are to investigate further the asymptotic properties and to develop rate of convergence results. The importance of these results for applications is obvious: the amount of data required to achieve some specified estimation accuracy would be very helpful information for designing experiments.

Functional central limit theorems and functional laws of iterated logarithms have played important roles in statistical estimation theory involving large samples. In (Heyde, 1981), Heyde gives an extensive survey on the usefulness and recent progress in these invariance theorems, which both use and extend the interplay between statistical estimation and stochastic processes.

The results to be presented in the sequel deal with the convergence of functions constructed out of the sequence of least squares estimators (suitably scaled), and provide portmanteau forms from which other limit theorems may be obtained. A wide range of limit distribution results involving functionals of the sequence of estimators can be inferred by employing the weak invariance principle, and the "with probability one" convergence rate of the algorithm can be derived by virtue of the strong invariance theorem. These results provide us with further insight on the behavior of the nonlinear least squares type of stochastic optimization and identification algorithms.

The rest of the paper is organized as follows. In the next section, we set up the notations, and summarize some previous results. Section 3 is devoted to the weak convergence issue. Under suitable conditions, we show an appropriately scaled sequence converges weakly to a Brownian motion. Exploiting this function space setting further, we derive an almost sure estimate on the error bound in Section 4. As a consequence, the functional law of iterated logarithm holds.

To proceed, a brief explanation about the notations is in order. We shall use "$\prime$" to denote the transpose of a matrix and use K to denote a generic positive constant; its value may change from time to time. The short hand notion "w.p.1" is meant to be "with probability one".

2. The general least squares problem. We begin this section by setting up the least squares identification problem. Let X be a compact subset of $\mathbf{R}^m$, $g : X \to \mathbf{R}$ be an unknown continuous function. We make a sequence of observations $\{Y_k\}$ with

$$Y_k = g(x_k) + \varepsilon_k, \quad 1 \leqslant k \leqslant n, \tag{2.1}$$

where $\{x_k\}$ (with $x_k \in X$, $1 \leqslant k \leqslant n$) is a collection of settings at which the measurements are made. Also, we have a parameterized function $f(x,q)$ (called the model function) to which we wish to fit our observations. As in the examples above, this parameterized function often arises from a parameter dependent differential equation of a system under consideration. In what follows, we assume $q \in Q_{ad}$, where $Q_{ad} \subset Q \subset \mathbf{R}^r$. The set Q_{ad} is referred to as the admissible parameter set: it incorporates various constraints on the parameters, such as maintaining positive diffusion coefficients in Examples 1 and 3.

To carry out the identification task, we set this as a stochastic optimization problem and use the following mean square objective function

$$J_n(q) = \frac{1}{n}\sum_{k=1}^{n}(Y_k - f(x_k, q))^2. \tag{2.2}$$

This objective function is of the long run average type. The identification task can be stated as follows: find a sequence of estimates $\{q_n\}$ that converges to q^* (the true parameter) and yields the "best" approximation of $g(\cdot)$ via the function $f(\cdot)$. What we mean by "best" will become clear below.

Define another functional $J_n^0(\cdot)$ as

$$J_n^0(q) = \sigma^2 + \frac{1}{n}\sum_{k=1}^{n}(g(x_k) - f(x_k, q))^2. \tag{2.3}$$

In many situations, especially those in parameter identification problems for distributed systems, $f(\cdot)$ is not a function we can compute explicitly. Thus, approximating $f(\cdot)$ is necessary. We shall denote an approximating sequence of $f(\cdot)$ by $\{f^N(\cdot)\}$ in the sequel. Corresponding to this sequence, we define $J_n^N(q)$ and $J_n^{0,N}(q)$ by replacing $f(\cdot)$ in (2.2) and (2.3) by $f^N(\cdot)$, respectively. Along the same line, $\{q_n^N\}$ and $\{q_n^{0,N}\}$ are the corresponding minimizers for J^N and $J^{0,N}$, respectively.

Unless otherwise indicated, we shall make use of the following assumptions throughout this paper.

(A1) The sequence $\{\varepsilon_k\}$ is composed of independent and identically distributed random variables with $E\varepsilon_k = 0$, $E\varepsilon_k^2 = \sigma^2 < \infty$.

(A2) The functions f, f^N: $Q \rightarrow C(X)$, are continuous, where $C(X)$ denotes the space of continuous functions defined on X and $Q \subset \mathbf{R}^r$. The set Q_{ad} is a compact subset of Q. For each x, $f(x,\cdot)$, $f^N(x,\cdot)$ are twice continuously differentiable. The function g: $X \rightarrow \mathbf{R}$ is continuous. Moreover, $f^N \rightarrow f$, $\partial^2 f^N/\partial q^2 \rightarrow \partial^2 f/\partial q^2$ as $N \rightarrow \infty$ and the convergence is uniform on any compact subset of Q.

(A3) The sequence $\{x_k\}$ is taken in X in such a way that there exists a finite measure μ on X, such that for each bounded and continuous function h,

$$\frac{1}{n}\sum_{k=1}^{n} h(x_k) \overset{n\to\infty}{\longrightarrow} \int_X h\,d\mu.$$

(A4) The functional

$$J^*(q) = \sigma^2 + \int_X (g(x) - f(x,q))^2 d\mu \tag{2.4}$$

has a unique minimizer $q^* \in \text{Int}\, Q_{ad} \subset \text{Int}\, Q$, where Int G denotes the interior of the set G. $J_n^0(q)$ defined in (2.3) has a unique minimizer $q_n^0 \in Q_{ad}$. In addition,

$$\mathcal{T} = \partial^2 J^*(q^*)/\partial q^2 \text{ and } V = \sigma^2 \int_X \frac{\partial f(x,q^*)}{\partial q} \frac{\partial f'(x,q^*)}{\partial q} d\mu(x) \tag{2.5}$$

are positive definite.

Remark: The above conditions are essentially those employed in Banks and Fitzpatrick (1989), Fitzpatrick (1988), Banks and Fitzpatrick (1990). (A1) requires $\{\varepsilon_k\}$ to be an i.i.d. (independent and identically distributed) sequence with mean zero and finite variance. In the literature, this is often referred to as 'white' noise. (A2) can be viewed as a regularity condition on f, g, $\{f^N\}$ and (A3) is an ergodic type of assumption. The assumption (A4) gives the meaning of the phrases "true parameter" and "best fit of g via f." The requirement of $q^* \in \text{Int}\, Q_{ad}$ is essential. As pointed out in Fitzpatrick (1988), without this assumption, the desired asymptotic properties cannot be obtained. In fact, several counterexamples were given in Fitzpatrick (1988).

We are now in a position to state some limit theorems obtained previously in Banks and Fitzpatrick (1989), Fitzpatrick (1988), Banks and Fitzpatrick (1990). These results will be used in the subsequent development.

Theorem 2.1. *Under the above conditions,*

(1) *for each $q \in Q$, $P(\lim_n J_n(q) = J^*(q)) = 1$ and the convergence is uniform on each compact subset of Q.*

(2) $P(\lim_n q_n = q^*) = 1$.

(3) *$P(\lim_{n,N} J_n^N = J^*) = 1$ and the convergence is uniform on any compact subset of Q.*

(4) $P(\lim_{n,N} q_n^N = q^*) = 1$.

Theorem 2.2. *Under the above conditions,*

(1) $\sqrt{n}\partial J_n^N(q_n^{0,N})/\partial q \xrightarrow{n} N(0, V)$ *in distribution.*

(2) $\sqrt{n}(q_n^N - q_n^{0,N}) \xrightarrow{n,N} N(0, 4T^{-1}VT^{-1})$ *in distribution. In the above* $N(0, S)$ *denotes a normal distribution with mean 0 and covariance* S.

REMARK: In proving each of the statement in the above theorems, only a subset of the assumptions is needed. For the ease of presentation, we stated the results in an integrated fashion here, however.

Theorem 2.1 is a convergence or consistency result. It indicates that under suitable conditions, the algorithm for the minimization task is a strongly consistent one. Theorem 2.2 derives the asymptotic normality and provides a ground for further work such as testing hypothesis and 'parameter space reduction' (cf. Fitzpatrick (1988) for more details).

In what follows, we will devote our attention to the investigation of the asymptotic properties of the estimators obtained by this least squares procedure.

3. A weak invariance principle. In Fitzpatrick (1988), Banks and Fitzpatrick (1990) asymptotic normality was obtained. In view of Theorem 2.2, we have

$$\sqrt{n}(q_n^N - q_n^{0,N}) = O_p(1).$$

The notation $Z_n = O_p(1)$ is meant to be 'bounded in probability', i.e., for every $\eta > 0$, there is a $K_1(\eta)$ and a $K_2(\eta)$ such that

$$P(|Z_n| \leqslant K_1(\eta)) > 1 - \eta \quad \text{for all} \quad n > K_2(\eta).$$

Thus, it gives us an error bound in probability.

This section is an extension of Theorem 2.2. We shall obtain a functional invariance principle (or functional central limit theorem) for the estimators from the identification algorithm discussed in Section 1. The main technical framework is the method of weak convergence.

In order to study the least squares algorithm with the approximating sequence of model functions $\{f^N\}$, the objective function J_n^N is needed. In the sequel, we shall let $N = N_n$, i.e., N is a function of n and $N_n \xrightarrow{n} \infty$. However, for notational simplicity, throughout the paper, we will still write N instead. By virtue of the convergence of f^N to f, (2.3) and Theorem 2.1, it is easily seen that $J_n^{0,N} \xrightarrow{n} J^*$ and that the convergence is uniform on any compact subset of Q. In addition, $q_n^{0,N} \xrightarrow{n} q^*$ as $n \to \infty$.

For each $T < \infty$, define

$$M_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} \frac{\partial f(x_k, q_{[nt]}^{0,N})}{\partial q} \varepsilon_k, \quad t \in [0,T]. \tag{3.1}$$

It is readily seen that $M_n(\cdot) \in D^r[0,\infty)$, where $D^r[0,\infty)$ denotes the space of $\mathbb{R}^r$-valued functions that are right continuous and have left-hand limits, endowed with the Skorohod topology (cf. Ethier and Kurtz (1986) and the references therein for definitions and further details). With the above definition, the first thing we want to show is:

Lemma 3.1. *Under the conditions* (A1) - (A4), *the following holds.*

$$M_n(t) = \tilde{M}_n(t) + o(1), \tag{3.2}$$

where

$$\tilde{M}_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} \frac{\partial f(x_k, q^*)}{\partial q} \varepsilon_k, \quad t \in [0,T] \tag{3.3}$$

and $o(1) \xrightarrow{n} 0$ *in probability uniformly in* $t \in [0,T]$.

Proof. To establish this assertion, we examine the difference $M_n(t) - \tilde{M}_n(t)$. In view of the definitions of $M_n(\cdot)$ and $\tilde{M}_n(\cdot)$, and the orthogonality of $\{\varepsilon_k\}$, the compactness of X and the convergence of $q_{[nt]}^{0,N}$ to q^* imply that for all $t \in [0,T]$,

$$\begin{aligned}
&E|M_n(t)-\tilde{M}_n(t)|^2\\
=&E\Bigg(\frac{1}{n}\sum_{j,k=1}^{[nt]}\left(\frac{\partial f(x_j,q_{[nt]}^{0,N})}{\partial q}-\frac{\partial f(x_j,q^*)}{\partial q}\right)'\\
&\times\left(\frac{\partial f(x_k,q_{[nt]}^{0,N})}{\partial q}-\frac{\partial f(x_k,q^*)}{\partial q}\right)\varepsilon_j\varepsilon_k\Bigg)\\
=&\frac{\sigma^2}{n}\sum_{k=1}^{[nt]}\Bigg(\left(\frac{\partial f(x_k,q_{[nt]}^{0,N})}{\partial q}-\frac{\partial f(x_k,q^*)}{\partial q}\right)'\\
&\times\left(\frac{\partial f(x_k,q_{[nt]}^{0,N})}{\partial q}-\frac{\partial f(x_k,q^*)}{\partial q}\right)\Bigg)\\
=&\frac{\sigma^2[nt]}{n}\frac{1}{[nt]}\sum_{k=1}^{[nt]}\Bigg(\left(\frac{\partial f(x_k,q_{[nt]}^{0,N})}{\partial q}-\frac{\partial f(x_k,q^*)}{\partial q}\right)'\\
&\times\left(\frac{\partial f(x_k,q_{[nt]}^{0,N})}{\partial q}-\frac{\partial f(x_k,q^*)}{\partial q}\right)\Bigg)\xrightarrow{n}0.
\end{aligned}$$

Thus, (3.2) holds. The lemma is proved.

A very similar argument gives us

$$\sup_n E|\tilde{M}_n(t)|^2<\infty. \tag{3.4}$$

The lemma above indicates that in order to study the asymptotic behavior of $M_n(\cdot)$, it is enough to consider $\tilde{M}_n(\cdot)$. In this sense, they are "asymptotically equivalent".

Let $\mathcal{F}_n$ denote the σ-algebra generated by $\{\varepsilon_k, k\leqslant n\}$. To proceed, define

$$\begin{aligned}
A_n(t)&=\frac{1}{n}\sum_{k=1}^{[nt]}\frac{\partial f(x_k,q^*)}{\partial q}\frac{\partial f'(x_k,q^*)}{\partial q}E(\varepsilon_k^2|\mathcal{F}_{k-1})\\
&=\frac{\sigma^2}{n}\sum_{k=1}^{[nt]}\frac{\partial f(x_k,q^*)}{\partial q}\frac{\partial f'(x_k,q^*)}{\partial q},
\end{aligned}\tag{3.5}$$

and

$$\Delta\tilde{M}_n(t)=\tilde{M}_n(t)-\tilde{M}_n(t^-)=\tilde{M}_n(t)-\lim_{s\to t^-}\tilde{M}_n(s). \tag{3.6}$$

The second line in (3.5) follows from the i.i.d. assumption on $\{\varepsilon_k\}$.

In view of the assumptions (A3) and (A4), $A_n(t) \xrightarrow{n} A(t) = tV$, where V is given by (2.5). Observe that $A_n(t) - A_n(s)$ is nonnegative definite for $t > s \geqslant 0$.

Lemma 3.2. *Under the hypotheses of Lemma* 3.1,

$$\lim_n E\Big(\sup_{0\leqslant t\leqslant T} |\Delta \tilde{M}_n(t)|^2\Big) = 0. \tag{3.7}$$

Proof. Due to the fact that $\{\varepsilon_k\}$ is a sequence of i.i.d. random variables with 0 mean, $\tilde{M}_n(t)$ is a martingale and it is square integrable. By using a familiar martingale inequality, we have

$$E\Big(\sup_{0\leqslant t\leqslant T} |\Delta \tilde{M}_n(t)|^2\Big) \leqslant 4E|\Delta \tilde{M}_n(T)|^2. \tag{3.8}$$

Thus, to verify the lemma, we need only show that the right-hand side of (3.8) tends to 0 as $n \to \infty$. By virtue of the L_2 boundedness of $\{\varepsilon_k\}$ together with (A3), direct computation yields that

$$\begin{aligned}
&E|\Delta \tilde{M}_n(T)|^2\\
&= \frac{1}{n} \sum_{k=[nT^-]}^{[nT]} \frac{\partial f'(x_k, q^*)}{\partial q} \frac{\partial f(x_k, q^*)}{\partial q} E\varepsilon_k^2\\
&= \frac{\sigma^2 [nT]}{n} \frac{1}{[nT]} \sum_{k=1}^{[nT]} \frac{\partial f'(x_k, q^*)}{\partial q} \frac{\partial f(x_k, q^*)}{\partial q}\\
&\quad - \frac{\sigma^2 [nT^-]}{n} \frac{1}{[nT^-]} \sum_{k=1}^{[nT^-]} \frac{\partial f'(x_k, q^*)}{\partial q} \frac{\partial f(x_k, q^*)}{\partial q} \xrightarrow{n} 0.
\end{aligned}$$

The proof is concluded.

Lemma 3.3. *Under the same conditions as in Lemma* 3.1, *the following hold:*

(1) $\lim_n(\sup_{0\leqslant t\leqslant T} |A_n^{ij}(t) - A_n^{ij}(t^-)|) = 0$, $i, j = 1, \ldots, r$.

(2) $\tilde{M}_n^i(t)\tilde{M}_n^j(t) - A_n^{ij}(t)$ *are martingales, where* $\tilde{M}^i(\cdot)$, $\tilde{M}_n^i(\cdot)$ *and* $A^{ij}(\cdot)$, $A_n^{ij}(\cdot)$ *denote the i-th component of* $\tilde{M}(\cdot)$, $\tilde{M}_n(\cdot)$ *and the ij-th entry of* $A(\cdot)$, $A_n(\cdot)$, *respectively.*

Proof. (1) follows from the convergence of $A_n(\cdot)$ and the continuity of $A(\cdot)$. As for (2), we note for $s \leqslant t$,

$$\begin{aligned}
&E((\tilde{M}_n^i(t)\tilde{M}_n^j(t) - A_n^{ij}(t)) \mid \mathcal{F}_{[ns]})\\
=&\frac{1}{n}\sum_{k=1}^{[nt]}\Big(\frac{\partial f(x_k,q^*)}{\partial q}\Big)^i\Big(\frac{\partial f(x_k,q^*)}{\partial q}\Big)^j E(\varepsilon_k^2 \mid \mathcal{F}_{[ns]})\\
&-\frac{\sigma^2}{n}\sum_{k=1}^{[nt]}\Big(\frac{\partial f(x_k,q^*)}{\partial q}\Big)^i\Big(\frac{\partial f(x_k,q^*)}{\partial q}\Big)^j\\
=&\frac{1}{n}\sum_{k=1}^{[ns]}\Big(\frac{\partial f(x_k,q^*)}{\partial q}\Big)^i\Big(\frac{\partial f(x_k,q^*)}{\partial q}\Big)^j(\varepsilon_k^2-\sigma^2)\\
=&\tilde{M}_n^i(s)\tilde{M}_n^j(s) - A_n^{ij}(s).
\end{aligned}$$

The martingale property is verified.

We now recall a general result concerning weak convergence to a Gaussian limit, from which we can obtain our weak invariance principle.

Proposition 3.4. (Ethier and Kurtz, 1986) *Let $m_n(\cdot)$ be a martingale with sample paths in $D^r[0,\infty)$ and $m(0) = .0$. Let $C_n(\cdot)$ be a symmetric $r \times r$ matrix-valued process such that $C_n^{ij}(\cdot)$ has sample paths in $D^r[0,\infty)$ and $C_n(t) - C_n(s)$ is nonnegative definite for $t > s \geqslant 0$. Assume that for each $T > 0$, and $i, j = 1, 2, \dots, r$,*

$$\lim_n E\Big(\sup_{0\leqslant t\leqslant T} |C_n^{ij}(t) - C_n^{ij}(t^-)|\Big) = 0,$$

$$\lim_n E\Big(\sup_{0\leqslant t\leqslant T} |m_n(t) - m_n(t^-)|^2\Big) = 0,$$

and for $i, j = 1, 2, \dots, r$, $t \geqslant 0$,

$$m_n^i(t)m_n^j(t) - C_n^{ij}(t)$$

are martingales. Suppose that for each $t \geqslant 0$, $C_n(t) \xrightarrow{n} C_n(t)$ in probability such that $C(t)$ is symmetric nonnegative definite and *having continuous paths. Then, $m_n(\cdot)$ converges to a process $m(\cdot)$ with independent Gaussian increments and continuous paths.*

REMARK: Combining Lemma 3.2 and Lemma 3.3, all conditions in the above proposition are satisfied. By virtue of Proposition 3.4, $\tilde{M}_n(\cdot)$ converges weakly to a process $M(\cdot)$ which has independent Gaussian increments and continuous paths. In view of Lemma 3.1, the same conclusion also holds for the process $M_n(\cdot)$.

Theorem 3.5. *Assume* (A1) - (A4) *are satisfied. Then, $M_n(\cdot)$ converges weakly to a Brownian motion $M(\cdot)$ with convariance matrix V given by* (2.5).

Proof. In view of Proposition 3.4 and the above remark, we need only show that the increments of $M(\cdot)$ are stationary. Since they are Gaussian, the stationarity will be implied if we can show that the covariance function $\Sigma(t,s)$ satisfies (cf. Breiman (1968) Proposition 11.17)

$$\Sigma(t,s) = \Sigma(t-s,0) \quad \text{for all} \quad t > s \geqslant 0, \tag{3.9}$$

i.e., the underlying process is wide sense (or second order) stationary. Equation (3.9) can be verified fairly easily along the same line as the proofs of Lemmas 3.2 and 3.3.

Theorem 3.5 is our weak invariance result. To illustrate its utility, we define

$$\hat{M}_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} \frac{\partial f^N(x_k, q_{[nt]}^{0,N})}{\partial q} \varepsilon_k. \tag{3.10}$$

With an argument analogous to that of Lemma 3.1, we have

$$\hat{M}_n(t) = \tilde{M}_n(t) + o(1), \tag{3.11}$$

where $o(1) \xrightarrow{n} 0$ in probability. Theorem 3.5 then implies that $\hat{M}_n(\cdot)$ converges weakly to a Brownian motion $M(\cdot)$. In addition, as in Fitzpatrick (1988), by using (3.11),

$$\begin{aligned} \sqrt{n}\frac{\partial J_n^N(q_n^{0,N})}{\partial q} &= -2\hat{M}_n(1) + \sqrt{n}\frac{\partial J_n^{0,N}(q_n^{0,N})}{\partial q} \\ &= -2\tilde{M}_n(1) + \sqrt{n}\frac{\partial J_n^{0,N}(q^{0,N})}{\partial q} + o(1), \end{aligned} \tag{3.12}$$

where $o(1) \xrightarrow{n} 0$ in probability. Note that the interior condition in (A4) gives us that $\frac{\partial J_n^{0,N}(q^{0,N})}{\partial q} \equiv 0$ for sufficiently large n. Taking a truncated Taylor's expansion with $\bar{q}_n^N$ denoting the vector that has components between the corresponding components of q_n^N and $q_n^{0,N}$, we have

$$\sqrt{n}(q_n^N - q_n^{0,N}) = 2(\mathcal{T}_n^N)^{-1}\tilde{M}_n(1) + \sqrt{n}(\mathcal{T}_n^N)^{-1}\frac{\partial J_n^N(q_n^N)}{\partial q} - \sqrt{n}(\mathcal{T}_n^N)^{-1}\frac{\partial J_n^{0,N}(q_n^{0,N})}{\partial q} + o(1), \tag{3.13}$$

where $\mathcal{T}_n^N = \partial^2 J_N^{0,N}(\bar{q}_n^N)/\partial q^2$. The assumption $q^* \in \text{Int } Q_{ad}$ yields that $q_n \in \text{Int } Q_{ad}$ and hence for sufficiently large n, the second and third terms on the right side above are 0.

In view of the above paragraph, in particular (3.12) and (3.13), we have

$$\sqrt{n}(q_{[nt]}^N - q_{[nt]}^{0,N}) = 2(\mathcal{T}_{[nt]}^N)^{-1}\tilde{M}_n(t) + o(1), \tag{3.14}$$

where $o(1) \xrightarrow{n} 0$ in probability. Now, define

$$Q_n(t) = \sqrt{n}(q_{[nt]}^N - q_{[nt]}^{0,N}). \tag{3.15}$$

In lieu of looking at the sequence $\sqrt{n}(q_n^N - q_n^{0,N})$, we emphasize the stochastic process aspects of the problem. As a result, more far reaching "dynamical" properties of the estimation procedures are obtained.

Theorem 3.6. *Assume* (A1) - (A4) *are satisfied. Then,* $Q_n(\cdot)$ *converges weakly to a Brownian motion* $Q(\cdot)$ *such that* $Q(t) = 2\mathcal{T}^{-1}M(t)$, *where* $M(\cdot)$ *is given by Theorem* 3.5.

As a consequence, we rediscover the asymptotic normality of Theorem 2.2.

Corollary 3.7. $\sqrt{n}(q_n^N - q_n^{0,N}) \sim N(0, 4\mathcal{T}^{-1}V\mathcal{T}^{-1})$ with V and $\mathcal{T}$ defined by (2.5).

Proof. Simply set $t = 1$ in Theorem 3.6.

We remark that the weak functional invariance results in Theorem 3.5 and 3.6 are a generalization of the central limit theorem to a function space and gives more details about the process involved. Using this theorem, one can generalize the asymptotic results further in many ways. For example, one could incorporate random numbers of observations (that is, to replace n by a random variable) into the formulation. For a related problem in stochastic approximation, see the work Yin (1990) and the references therein.

Furthermore, the scaling and properties of the Brownian motion give the rates of convergence, and much interesting information. We treat the behavior of the asymptotic part of $\{q_n^N - q_n^{0,N}\}$ as a dynamical process. This enables us to exploit the (stochastic) structure of the least squares algorithm. It seems that the techniques used here can easily be generalized to other noise processes. It should also be pointed out that the dynamic behavior of the iterates cannot be obtained by using the ordinary central limit theorem type of results alone. Thus, the invariance principles are necessary and indispensable in this regard. Furthermore, to the best of our knowledge, for the distributed parameter identification problem via use of the least squares algorithm, relatively little has been known concerning the 'stochastic process' aspects.

4. A strong invariance theorem. In this section, we derive almost sure error bounds and obtain a functional law of the iterated logarithm, which gives us rates of convergence in the almost sure sence. Interest in the law of iterated logarithm and related asymptotic fluctuation results has renewed since the mid-70's, due primarily to the celebrated work of Strassen (1964). In his paper, Strassen derived a functional law of iterated logarithm by using the Skorohod imbedding or Skorohod representation (cf. Skorohod, 1956). In this work, we shall closely follow the approach presented in Philipp and Stout (1975), which is a refinement of Strassen's invariance principle. The main idea is of "martingale approximation;" the block sums of dependent random variables constitute approximately a martingale difference sequence, to which Skorohod imbedding can be applied. This is especially suited in our case due to the fact that the underlying sums are themselves martingales.

We remark that, since $\{\varepsilon_k\}$ is a sequence of real-valued random variables, it appears to be simplest to work componentwise, treating each component of $q_n^N - q_n^{0,N}$ as a real-valued stochastic process. To deal with vector-valued processes, a different approach has to be taken. For example, in dealing with prediction error estimators in Heunis (1988), the framework developed in Berkes and Philipp (1979) for vector-valued processes was employed.

To begin, we observe that by (3.13),

$$q_n^N - q_n^{0,N} = \frac{2}{n}(\mathcal{T}_n^N)^{-1}R_n + o(n^{-1/2}) \tag{4.1}$$

for sufficiently large n, where $R_n = \sqrt{n}\tilde{M}_n(1)$, $o(n^{-1/2})$ is in the sence of "in probability" as before.

By means of the Skorohod imbedding, corresponding to the original probability space $(\Omega, \mathcal{F}, P)$, there is another one $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ with $\tilde{q}_n^N$, $\tilde{\mathcal{T}}_n^N$, $\tilde{R}_n$ defined on it such that $\tilde{q}_n^N$, $\tilde{\mathcal{T}}_n^N$, $\tilde{R}_n$ have the same distribution as $q_n^N, \mathcal{T}_n^N, R_n$ and

$$\tilde{q}_n^N - q_n^{0,N} = \frac{2}{n}(\tilde{\mathcal{T}}_n^N)^{-1}\tilde{R}_n + o_s(n^{-1/2})$$

where $o_s(n^{-1/2})$ is in the sence of "with probability one" (with respect to the probability measure $\tilde{P}$). For notational simplicity, and without loss of generality, we shall drop the symbol "tilde" and the phrase "with respect to the probability measure $\tilde{P}$" in the sequel. When we say 'on a richer probability space,' we are indicating the use of the Skorohod imbedding, changing the probability space but preserving the distributions of the random variables involved.

The desired asymptotic behavior will be based on the sequence R_n. To proceed, we quote a theorem from Philipp and Stout (1975), which is developed for nonstationary mixing processes.

Proposition 4.1. (Philipp and Stout, 1975) *Let $\{z_k\}$ be a sequence of random variables satisfying $\sup_k E|z_k|^{2+\delta} < \infty$ for some $0 < \delta \leqslant 2$. Suppose that there exists a monotonically increasing function $F(\cdot)$ such that*

$$\sum_{k=p+1}^{p+n} E^{\frac{1}{2+\delta}}|z_k|^{2+\delta} \leqslant F\left(E\left(\sum_{k=p+1}^{p+n} z_k\right)^2\right), \ \forall p, \ \forall n. \tag{4.2}$$

Moreover, assume that

$$S_n^2 = E\Big(\sum_{k=1}^{n} z_k\Big)^2 \xrightarrow{n} \infty \tag{4.3}$$

and $\{S_n^2\}$ is strictly increasing; $\{z_k\}$ is a mixing sequence with mixing rate

$$\varphi(k) = O\big(k^{-300(1+\frac{2}{\delta})}\big). \tag{4.4}$$

Define $S(t) = \Sigma_{k\leqslant t} z_k$, $t \geqslant 0$. We can take $\{S(t);\ t \geqslant 0\}$ on a richer probability space, together with a Brownian motion $\{w(t);\ t \geqslant 0\}$, such that

$$S(t) - w(t) = O\big(t^{\frac{1}{2}-\varpi}\big) \ w.p.1,$$

for each $\varpi < \delta/588$.

Now, we come back to our identification problem. Let R_n^i be the i-th component of R_n. Then, we have

$$R_n^i = \sum_{k=1}^{n} \left(\frac{\partial f(x_k, q^*)}{\partial q}\right)^i \varepsilon_k. \tag{4.5}$$

In addition to (A1) – (A4), we shall assume that $E|\varepsilon_k|^{2+\delta} < \infty$, for some $0 < \delta \leqslant 2$. We identify $\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i \varepsilon_k$ with z_k as in Proposition 4.1. Since $\{\varepsilon_k\}$ is a sequence of i.i.d. random variables with 0 mean, the mixing condition is automatically satisfied. It is easily seen that

$$E(R_n^i)^2 = \sum_{k=1}^{n} \left(\left(\frac{\partial f(x_k, q^*)}{\partial q}\right)^i\right)^2 E\varepsilon_k^2$$

by the orthogonality. Clearly, for each i, $\{E(R_n^i)^2\}$ is an increasing sequence in n. It either tends to a finite limit as $n \to \infty$ or grows without bound. In view of (A4), the first alternative cannot happen. Hence, (4.3) is verified. Due to the orthogonality again, we have that

$$E\left(\sum_{k=p+1}^{p+n} \left(\frac{\partial f(x_k, q^*)}{\partial q}\right)^i \varepsilon_k\right)^2 = \sigma^2 \sum_{k=p+1}^{p+n} \left(\left(\frac{\partial f(x_k, q^*)}{\partial q}\right)^i\right)^2.$$

Choose $\varkappa_0$ such that

$$\varkappa_0\sigma^2\left(\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\right)^2 - \left|\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\right| \geqslant 0.$$

This can be done by the continuity of $\frac{\partial f(x_k,q^*)}{\partial q}$ and the compactness of $X \times Q_{ad}$. Let

$$\varkappa = \varkappa_0 E^{\frac{1}{2+\delta}}|\varepsilon_1|^{2+\delta} \quad\text{and}\quad F(x) = \varkappa x.$$

Under this choice of $F(\cdot)$, we have

$$\begin{aligned}
&\sum_{k=p+1}^{p+n}\left|\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\right| E^{\frac{1}{2+\delta}}|\varepsilon_k|^{2+\delta} \\
&\leqslant \varkappa_0 E^{\frac{1}{2+\delta}}|\varepsilon_1|^{2+\delta}\sigma^2 \sum_{k=p+1}^{p+n}\left(\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\right)^2 \\
&= \varkappa\sigma^2 \sum_{k=p+1}^{p+n}\left(\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\right)^2 \\
&= F\left(E\left(\sum_{k=p+1}^{p+n}\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\varepsilon_k\right)^2\right).
\end{aligned}$$

Thus, (4.2) is also verified. Therefore, all conditions of Proposition 4.1 are fulfilled. We have:

Theorem 4.2. *Suppose the assumptions* (A1) - (A4) *are satisfied and* $E|\varepsilon_k|^{2+\delta} < \infty$ *for some* $0 < \delta \leqslant 2$. *For* $i = 1, 2, \ldots, r$, *define*

$$S^i(t) = \sum_{k\leqslant t}\left(\frac{\partial f(x_k,q^*)}{\partial q}\right)^i\varepsilon_k, \quad t \geqslant 0.$$

Without changing the distribution, we can define $\{S^i(t);\ t \geqslant 0\}$ *on a richer probability space, together with a Brownian motion* $\{W^i(t);\ t \geqslant 0\}$, *such that for each* $i = 1, 2, \ldots, r$,

$$S^i(t) - W^i(t) = O(t^{\frac{1}{2}-\varpi})\ w.p.1,$$

for each $\varpi < \delta/588$.

Next, let

$$S(t) = (S^1(t), S^2(t), \ldots, S^r(t))'$$
$$W(t) = (W^1(t), W^2(t), \ldots, W^r(t))'.$$

REMARK: $W(\cdot)$ is a Brownian motion. Due to the fact that $\{\varepsilon_k\}$ is a sequence of real-valued random variables, roughly, $W(\cdot)$ is an one-dimensional Brownian motion multiplied by a deterministic vector. This can be seen by noting the weak convergence of last section and the fact

$$S(t) = \sum_{k \leqslant t} \left(\left(\frac{\partial f(x_k, q^*)}{\partial q} \right)^1, \ldots, \left(\frac{\partial f(x_k, q^*)}{\partial q} \right)^r \right)' \varepsilon_k.$$

In view of Theorem 4.2, the following holds.

COROLLARY 4.3. $S(t) - W(t) = O(t^{\frac{1}{2}-\varpi})$ *w.p.1 for each $\varpi < \delta/588$.*
Proof. Owing to Theorem 4.2,

$$|S(t) - W(t)| \leqslant \sum_{i=1}^{r} |S^i(t) - W^i(t)| = O(t^{\frac{1}{2}-\varpi}) \text{ w.p.1.}$$

Let $q(t)$, $q^0(t)$ and $\mathcal{T}(t)$ be the piecewise constant interpolations of q_n^N, $q_n^{0,N}$ and $\mathcal{T}_n^N$, respectively; that is, for $n \leqslant t < n+1$, we define

$$q(t) = q_n^N, \quad q^0(t) = q_n^{0,N}, \quad \mathcal{T}(t) = \mathcal{T}_n^N(\bar{q}_n^N).$$

Theorem 4.4. *Suppose the conditions of Theorem 4.2 are satisfied. Without changing the distribution, we can redefine $\{q(t);\ t \geqslant \}$ on a richer probability space, together with the Brownian motion $W(t)$ (as in Theorem 4.3), such that*

$$P\left(\lim_{t \to \infty} \frac{t(q(t) - q^0(t)) - 2\mathcal{T}^{-1} W(t)}{(t \log\log t)^{\frac{1}{2}}} = 0 \right) = 1. \tag{4.6}$$

Proof. We use (4.1) with the understanding that the equation holds with probability one. We have

$$\begin{aligned}\frac{t(q(t)-q^0(t))-2\mathcal{T}^{-1}W(t)}{(t\log\log t)^{\frac{1}{2}}} &= \frac{(t-[t])(q(t)-q^0(t))}{(t\log\log t)^{\frac{1}{2}}}\\ &+\frac{2((\mathcal{T}(t))^{-1}-\mathcal{T}^{-1})W(t)}{(t\log\log t)^{\frac{1}{2}}}+\frac{2(\mathcal{T}(t))^{-1}(S(t)-W(t))}{(t\log\log t)^{\frac{1}{2}}}\\ &+\frac{o(t^{1/2})}{(t\log\log t)^{\frac{1}{2}}}.\end{aligned} \tag{4.7}$$

Due to the convergence of q_n^N and $q_n^{0,N}$, the interpolations $q(t)$ and $q^0(t)$ are bounded. The first term on right-hand side of (4.7) thus tends to 0 w.p.1. Since $W(t)$ is a Brownian motion, by virtue the law of iterated logarithm for $W(\cdot)$,

$$P\left(\frac{W(t)}{(t\log\log t)^{\frac{1}{2}}}=O(1)\right)=1. \tag{4.8}$$

In view of the assumptions (A3) and (A4), and Theorem 2.1, we have that $(\mathcal{T}(t))^{-1}\to\mathcal{T}^{-1}$ w.p.1., so that the second term on the right-hand side of (4.7) also goes to 0 w.p.1. In addition, the third term approaches 0 by the almost sure boundedness of $(\mathcal{T}(t))^{-1}$ and Corollary 4.3. Finally, the last term is of the order $o(1/(\log\log t)^{\frac{1}{2}})$ w.p.1. The theorem thus follows.

REMARK: The above theorem provides a bound (in the sense of w.p.1) on the order of magnitude of the estimation error by the well-known Brownian motion. It exploits the intrinsic behavior of the estimation procedure and relates the normalized sequence to a standard random process which we know quite a bit about.

We have that $q_n^N-q_n^{0,N}\xrightarrow{n}0$ w.p.1. The following question is in order. How fast does the convergence take place? What is the rate of convergence in the sense of w.p.1? Owing to the strong consistency and the asymptotic normality, it is possible to derive that as $n\to\infty$, for $\nu<1/2$,

$$q_n^N-q_n^{0,N}=o(n^{-\nu})\text{ w.p.1.}$$

This, however, is only a rather coarse estimate. Is it possible to obtain sharper bounds? This question can be answered by employing the above theorem. In fact, the solution is a by-product or corollary of Theorem 4.4.

In view of (4.6) and (4.8) and noticing the interpolation, we have

$$P\left(\lim_{n\to\infty}\frac{\sqrt{n}(q_n^N - q_n^{0,N})}{(\log\log n)^{1/2}} = O(1)\right) = 1. \tag{4.9}$$

As a consequence, we have as $n \to \infty$,

$$q_n^N - q_n^{0,N} = O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right) \text{ w.p.1.} \tag{4.10}$$

Eq. (4.10) gives precise order of the speed of convergence and places a tighter bound on the estimation error. By computing the estimates for $n = n_1$ and $n = n_2 > n_1$, one can use this rate of convergence result to estimate the error of the estimator, and determine, for example, if more data is needed.

REFERENCES

Banks, H.T., and P.Kareiva (1983). Parameter estimation techniques for transport equations with applications to population dispersal and tissue bulk flow models. *J. Math. Bio.*, **17**, 253–272.

Banks, H.T., and B.G.Fitzpatrick (1989). Inverse problems for distributed systems: statistical test and ANOVA. *Proc. International Symposium on Mathematical Approaches to Environmental and Ecological Problems*, Lecture Notes in Biomath, Springer-Verlag, New York.

Banks, H.T., Y.Wang, D.Inman, and H.Cudney (1987). Parameter identification techniques for the estimation of damping in flexible structure experiments. *Proc. 26th Conf. Decision Control*, 1392–1395.

Fitzpatrick, B.G. (1988). Statistical methods in parameter identification and model selection. Ph. D. Dissertation, Division of Applied Mathematics, Brown University, Providence, RI.

Banks, H.T., and B.G. Fitzpatrick (1990). Statistical tests for model comparison in parameter estimation problems for distributed systems. *J. Math. Bio.*, **28**, 501–527.

Heyde, C.C. (1981). Invariance principles in statistics. *Intern. Statistical Review*, **49**, 143–152.

Ethier, S.N., and T.G.Kurtz (1986). *Markov Processes: Characterization and Convergence.* Wiley, New York.

Breiman, L. (1968). *Probability.* Addison-Wesley, Reading, MA.

Yin, G. (1990). A stopping rule for the Robbins-Monro method. *J. Optim. Theory Appl.*, **67**, 151–173.

Strassen, V. (1964). An invariance principle for the law of the iterated logorithm. *Z. Wahrsch. Verw. Gebiete,* **3**, 211–226.

Skorohod, A.V. (1956). Limit theorems for stochastic processes. *Theory Probab. Appl.*, **1**, 262–290.

Philipp, W., and W. Stout (1975). Almost sure invariance principles for partial sums of weakly dependent random variables. *Mem. Amer. Math. Soc.*, 161.

Heunis, A.J. (1988). Asymptotic properties of prediction error estimations in approximate system identification. *Stochastics*, **24**, 1–43.

Berkes, I., and W. Philipp (1979). Approximation theorems for independent and weakly dependent random vectors. *Ann. Probab.*, **7**, 29–54.

Received September 1991

George Yin received his B.S. degree in mathematics, from the University of Delaware in 1983, M.S. degree in Electrical Engineering and Ph.D. degree in Applied Mathematics, from Brown University in 1987. Since then, he has been with Wayne State University, where he is currently an assistant professor of mathematics. His research interests include applied probability, stochastic processes and problems in control, optimization, and signal processing.

Ben G. Fitzpatrick received the B.S. degree in applied mathematics 1981, and the Master of Probability and Statistics degree in 1983, from Auburn University. He received the Ph.D. in applied mathematics from Brown University in 1988. Since 1989, he has been an assistant professor of mathematics at the University of Tennessee, Knoxville. His research interests include numerical and statistical aspects of identification and control problems.