The Annals of Statistics
1994, Vol. 22, No. 4, 1917–1930

ASYMPTOTIC PROPERTIES OF NONLINEAR LEAST SQUARES ESTIMATES IN STOCHASTIC REGRESSION MODELS[1]

BY TZE LEUNG LAI

Stanford University

Stochastic regression models of the form $y_i = f_i(\theta) + \varepsilon_i$, where the random disturbances ε_i form a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{G}_i\}$ and f_i is a random $\mathcal{G}_{i-1}$-measurable function of an unknown parameter θ, cover a broad range of nonlinear (and linear) time series and stochastic process models. Herein strong consistency and asymptotic normality of the least squares estimate of θ in these stochastic regression models are established. In the linear case $f_i(\theta) = \theta^T\psi_i$, they reduce to known results on the linear least squares estimate $(\Sigma_1^n\psi_i\psi_i^T)^{-1}\Sigma_1^n\psi_i y_i$ with stochastic $\mathcal{G}_{i-1}$-measurable regressors ψ_i.

1. Introduction. Consider a general stochastic regression model of the form

$$y_n = f_n(\theta) + \varepsilon_n, \tag{1.1}$$

where $\{\varepsilon_n\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{G}_n$ such that

$$\sup_n E\left(\varepsilon_n^2 \mid \mathcal{G}_{n-1}\right) < \infty \quad \text{a.s.}, \tag{1.2}$$

and $f_n(\theta)$ is a $\mathcal{G}_{n-1}$-measurable random function of a parameter vector $\theta = (\theta_1, \ldots, \theta_k)^T$. The y_n typically represent the observed outputs of a stochastic system while the ε_n represent unobservable random disturbances. The parameter θ is not assumed to be known and has to be estimated at stage n from the data $x_i, y_i, i \le n$, where the x_i represent covariates such that x_i is $\mathcal{G}_i$-measurable, and f_n is a given function that may depend on $y_1, \ldots, y_{n-1}, x_1, \ldots, x_{n-1}$. An example is the NARX model (nonlinear autoregressive model with exogenous inputs)

$$y_n = f(y_{n-1}, \ldots, y_{n-p}, x_{n-d}, \ldots, x_{n-d-q}; \theta) + \varepsilon_n, \tag{1.3}$$

in which $d \ge 1$ represents the delay and x_i is the input at stage i.

When $f_n(\theta) = \psi_n^T\theta$ is linear in θ with $\mathcal{G}_{n-1}$-measurable coefficient vector ψ_n, the least squares estimate $\hat{\theta}_n = (\Sigma_1^n\psi_i\psi_i^T)^{-1}\Sigma_1^n\psi_i y_i$ has been shown by Lai and

Received June 1990; revised March 1992.

[1]Research supported by the National Science Foundation, the National Security Agency and the Air Force Office of Scientific Research.

AMS 1991 *subject classifications.* Primary 62J02; secondary 62M10, 62F12, 60F15.

Key words and phrases. Stochastic regressors, nonlinear autoregressive models, control systems, optimal experimental design, strong consistency, asymptotic normality, martingales in Hilbert spaces.

Wei (1982) to be strongly consistent if

$$\lambda_{\min}\left(\sum_1^n \psi_i\psi_i^T\right) \to \infty \quad \text{and}$$
$$\left\{\log \lambda_{\max}\left(\sum_1^n \psi_i\psi_i^T\right)\right\}^{\rho} \Big/ \lambda_{\min}\left(\sum_1^n \psi_i\psi_i^T\right) \to 0 \quad \text{a.s.} \tag{1.4}$$

for some $\rho > 1$ [or for $\rho = 1$ if it is also assumed that $\sup_n E(|\varepsilon_n|^r \mid \mathcal{G}_{n-1}) < \infty$ a.s. for some $r > 2$], where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues of a symmetric matrix A. Moreover, under some additional assumptions, Lai and Wei (1982) also showed that $(\Sigma_1^n \psi_i\psi_i^T)^{1/2}(\hat{\theta}_n - \theta)$ has a limiting normal distribution. Earlier, assuming that $\lambda_{\min}(\Sigma_1^n \psi_i\psi_i^T) \to \infty$ a.s., Anderson and Taylor (1979) proved the strong consistency of $\hat{\theta}_n$ under the condition $\lambda_{\max}(\Sigma_1^n \psi_i\psi_i^T) = O(\lambda_{\min}(\Sigma_1^n \psi_i\psi_i^T))$ a.s., while Christopeit and Helmes (1980) weakened this condition to

$$\lambda_{\max}\left(\sum_1^n \psi_i\psi_i^T\right) = O\left(\lambda_{\min}^{\rho}\left(\sum_1^n \psi_i\psi_i^T\right)\right) \quad \text{for some } 1 < \rho < 2. \tag{1.5}$$

As pointed out by Lai and Wei (1982, 1987), condition (1.4) with $\rho = 1$ is in some sense weakest possible and plays an important role in the asymptotic solution of the adaptive control problem of choosing the inputs x_i sequentially so that the outputs y_i are as close as possible to some target value y^* in the linear ARX model (1.3) in which f is linear in $\theta = (\alpha_1, \ldots, \alpha_p, \beta_0, \ldots, \beta_q)^T$ and $\psi_n = (y_{n-1}, \ldots, y_{n-p}, x_{n-d}, \ldots, x_{n-d-q})^T$, without assuming prior knowledge of the parameter vector θ.

When f_n is nonlinear in θ, the least squares estimate $\hat{\theta}_n$ that minimizes

$$S_n(\xi) = \sum_{i=1}^n \big(y_i - f_i(\xi)\big)^2 \tag{1.6}$$

does not have a simple closed form and is typically computed by iterative solution of the estimating equation

$$\nabla S_n(\xi) = -2\sum_{i=1}^n \big(y_i - f_i(\xi)\big)\nabla f_i(\xi) = 0, \tag{1.7}$$

assuming the f_i to be differentiable. Here and in the sequel we use $\nabla g = (\partial g/\partial \xi_1, \ldots, \partial g/\partial \xi_k)^T$ to denote the gradient vector and $\nabla^2 g = (\partial^2 g/\partial \xi_i \partial \xi_j)_{1 \le i,j \le k}$ to denote the Hessian matrix of a smooth function g: $\mathbf{R}^k \to \mathbf{R}$. We shall also let $\|\xi\|^2 = \xi^T\xi$ for $\xi \in \mathbf{R}^k$ and let $\|A\| = \sup_{\|x\|=1} \|Ax\|$ for $k \times k$ matrices A. Note that $\max_{i,j}|a_{ij}| \le \|A\| \le k \max_{i,j}|a_{ij}|$ for a matrix $A = (a_{ij})_{1 \le i,j \le k}$. Klimko and Nelson (1978) studied the consistency properties of solutions of (1.7) via a quadratic approximation to (1.6) in a small neighborhood of θ. Specifically, they showed that with probability 1, there exists for sufficiently large n a solution

$\xi = \theta_n$ of (1.7) such that $\theta_n \to \theta$, under the assumptions that the functions f_i are twice continuously differentiable in some neighborhood of θ and that

$$\text{(1.8a)} \quad \limsup_{n\to\infty,\ \delta\to 0} (n\delta)^{-1} \sup_{\|\xi-\theta\|\le\delta} \left\|\nabla^2 S_n(\xi) - \nabla^2 S_n(\theta)\right\| < \infty \quad \text{a.s.},$$

$$\text{(1.8b)} \quad (2n)^{-1}\nabla^2 S_n(\theta) \to V \quad \text{(positive definite and nonrandom)} \quad \text{a.s.},$$

$$\text{(1.8c)} \quad n^{-1}\nabla S_n(\theta) \to 0 \quad \text{a.s.};$$

cf. Theorem 2.1 of Klimko and Nelson (1978). Since

$$\nabla^2 S_n(\theta)/2 = -\sum_{i=1}^n \varepsilon_i \nabla^2 f_i(\theta) + \sum_{i=1}^n \left(\nabla f_i(\theta)\right)\left(\nabla f_i(\theta)\right)^T,$$

and since

$$\sum_{i=1}^n \varepsilon_i \nabla^2 f_i(\theta) = o\left(\sum_{i=1}^n \left\|\nabla^2 f_i(\theta)\right\|^2\right) + O(1) \quad \text{a.s.}$$

by the martingale strong law [cf. Lemma 2(iii) of Lai and Wei (1982)], it follows that (1.8b) can be replaced by the simpler assumption

$$\text{(1.9)} \quad \begin{aligned} &n^{-1}\sum_{i=1}^n \left(\nabla f_i(\theta)\right)\left(\nabla f_i(\theta)\right)^T \to V \quad \text{(positive definite and nonrandom)}, \\ &\sum_{i=1}^n \left\|\nabla^2 f_i(\theta)\right\|^2 = O(n) \quad \text{a.s.}, \end{aligned}$$

as has been noted by Klimko and Nelson (1978) in their remark following (2.2).

In the linear case $f_i(\xi) = \psi_i^T \xi$, condition (1.8c) reduces to the convergence of $n^{-1}\Sigma_1^n \psi_i \psi_i^T$ to a positive-definite matrix V, which is much stronger than (1.4) or (1.5). Furthermore, the least squares estimate $\widehat{\theta}_n$ that attains the global minimum of $S_n(\xi)$ may be different from the θ_n in the Klimko–Nelson theorem which only relates to a local minimum of $S_n(\xi)$ near θ.

Assuming θ to be in some given compact subset Θ of $\mathbf{R}^k$, and assuming the f_i to be nonrandom continuous functions on Θ and the ε_i to be i.i.d., Wu (1981) showed that the least squares estimate $\widehat{\theta}_n$ is strongly consistent if for every $\lambda \ne \theta$ there exists an open ball $B(\lambda)$ centered at λ such that for some $M > 0$ and $1 < \rho < 2$,

$$\text{(1.10a)} \quad \inf_{\xi\in B(\lambda)} \sum_{i=1}^n \left[f_i(\xi) - f_i(\theta)\right]^2 \to \infty,$$

$$\text{(1.10b)} \quad \sum_{i=1}^n \sup_{\xi\in B(\lambda)} \left[f_i(\xi) - f_i(\theta)\right]^2 = O\left(\left\{\inf_{\xi\in B(\lambda)} \sum_{i=1}^n \left[f_i(\xi) - f_i(\theta)\right]^2\right\}^\rho\right),$$

$$\text{(1.10c)} \quad \begin{aligned} &\sup_{\xi,\xi'\in B(\lambda),\,\xi\ne\xi'} |f_i(\xi) - f_i(\xi')|/\|\xi - \xi'\| \\ &\quad \le M \sup_{\xi\in B(\lambda)} |f_i(\xi) - f_i(\theta)| \quad \text{for all } i. \end{aligned}$$

Under (1.10b), condition (1.10a) is satisfied if we simply assume that $\Sigma_{i=1}^{n}[f_i(\xi) - f_i(\theta)]^2 \to \infty$ for all $\xi \neq \theta$, as pointed out by Wu (1981). In view of (1.10c), the function ϕ_i defined on $B(\lambda)$ by $\phi_i(\xi) = f_i(\xi) - f_i(\theta)$ belongs to the Banach space of Lipschitz continuous functions on $B(\lambda)$ with the Lipschitz norm. This enables Wu to apply strong laws and probability bounds for sums of independent random variables taking values in a type 2 Banach space to analyze $\sup_{\xi \in B(\lambda)}|\Sigma_{i=1}^{n}\phi_i(\xi)\varepsilon_i|$, since he assumes the f_i (and therefore the ϕ_i also) to be nonrandom and the ε_i to be independent. The argument, however, cannot be extended to the more general case where the ε_i form a martingale difference sequence and f_i is $\mathcal{G}_{i-1}$-measurable, as will be explained in Section 2.

Instead of embedding the f_i in a Banach space of type 2, we shall work with suitably chosen Hilbert spaces H so that H-valued martingale strong laws can be applied to give an analog of Wu's strong consistency theorem for nonlinear least squares estimates in stochastic regression models. In addition, by making use of martingale central limit theorems and probability bounds for H-valued martingales, the asymptotic normality of $\hat{\theta}_n$ is established. These results, which are stated in Section 2 and proved in Section 3, reduce to the consistency results of Christopeit and Helmes (1980) and the asymptotic normality results of Lai and Wei (1982) in the linear regression case $f_n(\theta) = \psi_n^T\theta$ with $\mathcal{G}_{n-1}$-measurable regressors ψ_n.

2. Main results and some applications. While Wu (1981) assumes the f_i to be nonrandom and Lipschitz continuous on Θ for his consistency result mentioned above and further assumes f_i to be twice continuously differentiable in some neighborhood of θ for his asymptotic normality theorem, we shall drop the assumption that f_i be nonrandom but require the existence and continuity of the partial derivatives $\partial f_i/\partial\xi_j$, $\partial^2 f_i/\partial\xi_j\partial\xi_h (j \neq h)$, $\ldots$, $\partial^k f_i/\partial\xi_1\cdots\partial\xi_k$ in Θ. For $1 \leq m \leq k$, let

$$J(m,k) = \{(j_1,\ldots,j_m): j_1 < \cdots < j_m,\ j_i \in \{1,\ldots,k\} \text{ for } 1 \leq i \leq m\}. \tag{2.1}$$

For $\mathbf{j} = (j_1,\ldots,j_m) \in J(m,k)$, the notation $D_{\mathbf{j}}f = \partial^m f/\partial\xi_{j_1}\cdots\partial\xi_{j_m}$ will be used in the sequel. Moreover, if $B(\lambda) = \{\xi \in \Theta: \|\xi - \lambda\| < r\}$ is an open ball in Θ, we shall let $B(\lambda;\mathbf{j})$ denote the m-dimensional sphere $\{(\xi_{j_1},\ldots,\xi_{j_m}): (\lambda_1,\ldots,\lambda_{j_1-1},\xi_{j_1},\lambda_{j_1+1},\ldots,\lambda_{j_2-1},\xi_{j_2},\ldots,\lambda_{j_m-1},\xi_{j_m},\lambda_{j_m+1},\ldots,\lambda_k) \in B(\lambda)\}$. In particular, $B(\lambda;(1,\ldots,k)) = B(\lambda)$.

THEOREM 1. *Consider the stochastic regression model* (1.1) *in which the* ε_n *form a martingale difference sequence with respect to an increasing sequence of* σ*-fields* $\{\mathcal{G}_n\}$ *such that* (1.2) *holds, and* f_n *is* $\mathcal{G}_{n-1}$*-measurable. Suppose that* θ *belongs to a compact subset* Θ *of* $\mathbf{R}^k$ *and that the* f_i *have continuous partial derivatives* $D_{\mathbf{j}}f_i$ *on* Θ, *for every* $\mathbf{j} \in J(m,k)$ *and* $m = 1,\ldots,k$, *where* $J(m,k)$ *is defined in* (2.1). *Moreover, assume that for every* $\lambda \neq \theta$ *there exist* $1 < \rho_\lambda < 2$

and an open ball $B(\lambda)$ in Θ centered at λ such that

$$\inf_{\xi \in B(\lambda)} \sum_{i=1}^{n} \left[f_i(\xi) - f_i(\theta) \right]^2 \to \infty \quad a.s., \tag{2.2}$$

$$\max_{1 \le m \le k,\ \mathbf{j} \in J(m,k)} \sum_{i=1}^{n} \int_{B(\lambda;\mathbf{j})} (D_{\mathbf{j}} f_i)^2 \, d\xi_{j_1} \cdots d\xi_{j_m} + \sum_{i=1}^{n} \left[f_i(\lambda) - f_i(\theta) \right]^2 = O\left(\left\{ \inf_{\xi \in B(\lambda)} \sum_{i=1}^{n} \left[f_i(\xi) - f_i(\theta) \right]^2 \right\}^{\rho_\lambda} \right) \quad a.s. \tag{2.3}$$

Then the least squares estimate $\hat{\theta}_n$, defined as the minimizer of (1.6) *in Θ, converges a.s. to θ.*

THEOREM 2. *With the same notation and assumptions as in Theorem* 1, *suppose that θ belongs to the interior of Θ and that there exist nonrandom, symmetric, positive-definite matrices C_n and an open ball $B(\theta)$ centered at θ such that*

$$\lambda_{\min}(C_n) \xrightarrow{P} \infty, \qquad C_n^{-1} \left\{ \sum_{i=1}^{n} \left(\nabla f_i(\theta) \right) \left(\nabla f_i(\theta) \right)^T \right\}^{1/2} \xrightarrow{P} I, \tag{2.4}$$

$$\max_{1 \le i \le n} \left\| C_n^{-1} \nabla f_i(\theta) \right\| \xrightarrow{P} 0, \qquad \sup_{\xi \in B(\theta)} \sum_{i=1}^{n} \left\| \nabla^2 f_i(\xi) \right\|^2 = O_p\left(\lambda_{\min}^2(C_n) \right), \tag{2.5}$$

and for every $m, a, b \in \{1, \ldots, k\}$ and $\mathbf{j} \in J(m,k)$, the partial derivatives $D_{\mathbf{j}}(\partial^2 f_i / \partial \xi_a \, \partial \xi_b)$ are continuous on $B(\theta)$ for all i and

$$\left\{ \sum_{i=1}^{n} \int_{B(\theta;\mathbf{j})} \left[D_{\mathbf{j}} \left(\partial^2 f_i / \partial \xi_a \, \partial \xi_b \right) \right]^2 d\xi_{j_1} \cdots d\xi_{j_m} \right\} \Big/ \lambda_{\min}^4(C_n) \xrightarrow{P} 0. \tag{2.6}$$

Suppose that $E(\varepsilon_n^2 \mid \mathcal{G}_{n-1}) \xrightarrow{P} \sigma^2$ (nonrandom). Then $\{\Sigma_{i=1}^n (\nabla f_i(\theta))(\nabla f_i(\theta))^T\}^{1/2} (\hat{\theta}_n - \theta)$ converges in distribution as $n \to \infty$ to a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I$.

In the linear case $f_n(\xi) = \psi_n^T \xi$, $\nabla f_n(\xi) = \psi_n$ and the higher-order partial derivatives of f_n are 0. Therefore, while condition (2.2) of Theorem 1 reduces to $\lambda_{\min}(\Sigma_1^n \psi_i \psi_i^T) \to \infty$ a.s., the other condition (2.3) reduces to the Christopeit–Helmes condition (1.5). Conditions (2.4) and (2.5) of Theorem 2 reduce simply to

$$C_n^{-1} \left(\sum_{i=1}^{n} \psi_i \psi_i^T \right)^{1/2} \xrightarrow{P} I \quad \text{and} \quad \max_{i \le i \le n} \left\| C_n^{-1} \psi_i \right\| \xrightarrow{P} 0, \tag{2.7}$$

which has been assumed in Theorem 3 of Lai and Wei (1982) on asymptotic normality of $\hat{\theta}_n$ in the linear stochastic regression model $y_i = \psi_i^T \theta + \varepsilon_i$, $i = 1, 2, \ldots$.

Moreover, condition (2.6) clearly holds since $\nabla^2 f_i = 0$ for linear f_i. Hence Theorem 3 of Lai and Wei (1982) is a special case of Theorem 2.

Condition (2.3) is analogous to (1.10b) and (1.10c) assumed by Wu (1981). Instead of the sup-norm and Lipschitz norm in (1.10b) and (1.10c), (2.3) uses certain integral norms for sufficiently smooth functions. Likewise (2.6) is analogous to conditions (4.1) and (4.2) in Section 4 of Wu's paper, but uses integral norms instead of the sup-norm and Lipschitz norm that appear in Wu's conditions.

Condition (2.4) is much weaker than (1.9) assumed by Klimko and Nelson (1978). Wu (1981) requires C_n in (2.4) and (2.5) to be of the form $C_n = \tau_n I$ with scalar $\tau_n \uparrow \infty$, and his theory requires the f_n to be nonrandom and the ε_n to be independent since it is based on certain strong laws and probability bounds for sums of independent random variables taking values in a separable Banach space of type 2. The strong laws and probability bounds he used cannot be generalized to martingales unless the Banach space has a much smoother norm than that of a type 2 space. As pointed out by Hoffmann-Jørgensen and Pisier (1976), the Banach space has to be isomorphic to a uniformly 2-smooth space (an example of which is a Hillbert space) for such strong laws and probability bounds to hold for general martingales. This excludes the space of Lipschitz continuous functions considered by Wu.

By assuming the f_i to be sufficiently smooth as in Theorems 1 and 2, we shall represent f_i in Section 3 as a sum of several components that are elements of different Hilbert spaces. This is the basic idea behind conditions (2.3) and (2.6), which are more demanding on the smoothness of the functions f_i than Wu's conditions and which exploit such smoothness to study the asymptotic properties of least squares estimates via Hilbert space-valued martingales.

The conditions of Theorems 1 and 2 are satisfied by many NARX systems for which the function f in (1.3) is sufficiently smooth in θ. In particular, applications of Theorem 1 to consistent parameter estimation and to adaptive h-step-ahead prediction of the outputs y_{n+h} are discussed by Lai and Zhu (1991) for NARX systems and by Zhu (1992) for other nonlinear time series models. In the remainder of this section we consider some applications of Theorems 1 and 2 to the construction of asymptotically optimal adaptive designs and to inference from sequential designs in nonlinear regression models.

Although the theory of optimal experimental design for least squares estimation in linear regression models can in principle be extended to nonlinear models of the form

$$y_n = f(x_n, \theta) + \varepsilon_n, \tag{2.8}$$

with i.i.d. ε_n such that $E\varepsilon_n = 0$ and $E\varepsilon_n^2 = \sigma^2$, the extension has serious practical difficulties since typically an optimal design measure involves the unknown vector θ. To circumvent these difficulties, it has been proposed that designs be constructed sequentially, using observations made to date to estimate θ and choosing the next design point by replacing the unknown θ in the optimal design with the estimate; cf. Fedorov (1972), page 188. Since the classical asymptotic

theory of least squares estimates assumes the x_n to be nonrandom, it cannot be applied to the present setting in which the x_n are sequentially determined random vectors. Ford and Silvey (1980) pointed out these difficulties in asymptotic inference from such adaptive designs, and asked whether these designs would indeed be asymptotically equivalent in some sense to the optimal design that assumes θ to be known. They studied this problem in the particular linear regression model with $f(x_n, \theta) = (\theta_1, \theta_2)x_n$ and $x_n = (u_n, u_n^2)^T$, where the design levels u_n are chosen adaptively from the interval $[-1, 1]$ by replacing the unknown θ in the optimal design with the least squares estimate based on all available data at every stage. They showed that the least squares estimates are indeed strongly consistent in this example.

Subsequently, Ford, Titterington and Wu (1985) and Wu (1985) considered more general linear regression models, and made use of the results in Lai and Wei (1982) on linear stochastic regression models to show that the usual asymptotic inference based on least squares estimates is still valid for certain sequential designs in these linear models. Clearly Theorems 1 and 2, which are extensions of the consistency and asymptotic normality results in Lai and Wei (1982) to the general stochastic regression model (1.1), can be used to address the questions raised by Ford and Silvey (1980) concerning (i) the behavior of adaptive designs that substitute the unknown θ in an optimal design by the least squares estimate based on all available data at every stage, and (ii) the behavior of the least squares estimates from sequential designs in nonlinear regression models.

To illustrate the usefulness of Theorems 1 and 2 in addressing these issues, we consider the problem of sequential design and estimation in the Michaelis–Menton model

$$y_n = \theta_1 x_n/(\theta_2 + x_n) + \varepsilon_n, \tag{2.9}$$

in which the parameters θ_1, θ_2 and the design levels x_i are all positive. Let $\theta = (\theta_1, \theta_2)^T$. In practice one usually has prior knowledge of positive lower and upper bounds for θ_1 and θ_2, say $(0<)a_j < \theta_j < A_j$, $j = 1, 2$, giving a compact parameter space $\Theta = [a_1, A_1] \times [a_2, A_2]$. The random errors ε_n in (2.9) are assumed to form a martingale difference sequence satisfying (1.2) and such that $E(\varepsilon_n^2 \mid \mathcal{G}_{n-1}) \xrightarrow{P} \sigma^2$. Suppose that the design levels are to be chosen from the interval $(0, x^*]$. If θ were known, then the design

$$x_n = \begin{cases} x^*, & \text{if } n \text{ is odd,} \\ \theta_2 x^*/(2\theta_2 + x^*), & \text{if } n \text{ is even,} \end{cases} \tag{2.10}$$

would be D-optimal in the sense of minimizing the determinant of the asymptotic covariance matrix of the least squares estimate [cf. Bates and Watts (1988), pages 125 and 126].

Let $\hat{\theta}_t = (\hat{\theta}_{t,1}, \hat{\theta}_{t,2})^T$ be the least squares estimate based on $x_1, y_1, \ldots, x_t, y_t$. Instead of simply substituting in (2.10) the unknown θ_2 by $\hat{\theta}_{n-1,2}$, we make a slight modification to ensure strong consistency of $\hat{\theta}_t$ via Theorem 1 by redefin-

ing x_n at stages $n \in \{n_1, n_2, \ldots\}$ so that

$$x_{n_i} \sim c/\log n_i \quad \text{and} \quad n_i \sim i^\alpha \quad \text{as } i \to \infty, \tag{2.11}$$

for some $c > 0$ and $1 < \alpha < 2$. Specifically, we shall consider the adaptive design defined by

$$x_n = \begin{cases} x^*, & \text{if } n \text{ is odd and } n \notin \{n_1, n_2, \ldots\}, \\ \widehat{\theta}_{n-1,2} x^* / (2\widehat{\theta}_{n-1,2} + x^*), & \text{if } n \text{ is even and } n \notin \{n_1, n_2, \ldots\}, \\ c/(1+\log n), & \text{if } n \in \{n_1, n_2, \ldots\}. \end{cases} \tag{2.12}$$

Thus, the x_n are sequentially determined random variables such that x_n is $\mathcal{G}_{n-1}$-measurable. Let $f_n(\theta) = \theta_1 x_n/(\theta_2 + x_n)$ and note that f_n has continuous partial derivatives of all orders on Θ.

To show that the assumptions of Theorem 1 are satisfied by the f_n thus defined, take any $\lambda = (\lambda_1, \lambda_2)^T \in \Theta$ such that $\lambda \neq \theta$. Since $f_i(\theta) = \theta_1 x_i/(\theta_2 + x_i)$, it follows that

$$f_i(\lambda) - f_i(\theta) = (\lambda_1 - \theta_1)\frac{x_i^2}{(\lambda_2 + x_i)(\theta_2 + x_i)} + (\lambda_1\theta_2 - \lambda_2\theta_1)\frac{x_i}{(\lambda_2 + x_i)(\theta_2 + x_i)}. \tag{2.13}$$

Since $\lambda \neq \theta$, either (i) $\lambda_1\theta_2 - \lambda_2\theta_1 \neq 0$ or (ii) $\lambda_1\theta_2 = \lambda_2\theta_1$ and $\lambda_1 \neq \theta_1$. Let $N_n = \{n_t : n_t \leq n\}$. For case (i), it follows from (2.13) and (2.11) that for some $\delta > 0$ and for a sufficiently small open ball $B(\lambda)$ in Θ,

$$\begin{aligned} \inf_{\xi \in B(\lambda)} \sum_{i \in N_n} \left[f_i(\xi) - f_i(\theta)\right]^2 &\sim \inf_{\xi \in B(\lambda)} (\xi_1\theta_2 - \xi_2\theta_1)^2 \sum_{i \in N_n} \left(x_i/\xi_2\theta_2\right)^2 \\ &\geq \delta n^{1/\alpha}/(\log n)^2. \end{aligned} \tag{2.14}$$

For case (ii), it follows from (2.13) and (2.12) that

$$\begin{aligned} &\inf_{\xi \in B(\lambda)} \sum_{i \notin N_n,\, i \leq n \text{ and } i \text{ odd}} \left[f_i(\xi) - f_i(\theta)\right]^2 \\ &\qquad \geq (n/4) \inf_{\xi \in B(\lambda)} \left[(\xi_1 - \theta_1)^2 x^{*4} / \left\{(\xi_2 + x^*)(\theta_2 + x^*)\right\}^2\right] \end{aligned} \tag{2.15}$$

for all large n, by choosing $B(\lambda)$ sufficiently small. Since the left-hand side of condition (2.3) is $O(n)$, it follows from (2.14) and (2.15) that (2.3) holds with $\rho_\lambda = \alpha/2 + 1$. Noting that $\alpha < \rho_\lambda < 2$, we can apply Theorem 1 to conclude that $\widehat{\theta}_n \to \theta$ a.s.

To show that the assumptions of Theorem 2 are satisfied, note that

$$\nabla f_n(\xi) = \left(x_n/(\xi_2 + x_n), -\xi_1 x_n/(\xi_2 + x_n)^2\right)^T, \qquad \xi = (\xi_1, \xi_2)^T. \tag{2.16}$$

From (2.16) and (2.12), it follows that conditions (2.4) and (2.5) are satisfied with $C_n = (n\Psi)^{1/2}$, where

$$\begin{aligned} \Psi &= \left(\psi\psi^T + \widetilde{\psi}\,\widetilde{\psi}^T\right)/2, \qquad \psi = \left(x^*/(\theta_2 + x^*), -\theta_1 x^*/(\theta_2 + x^*)^2\right)^T, \\ \widetilde{\psi} &= \left(\widetilde{x}/(\theta_2 + \widetilde{x}), -\theta_1\widetilde{x}/(\theta_2 + \widetilde{x})^2\right)^T \quad \text{with } \widetilde{x} = \theta_2 x^*/(2\theta_2 + x^*). \end{aligned} \tag{2.17}$$

Note in this connection that $\det(\Psi) > 0$ [cf. Bates and Watts (1988), pages 125 and 126]. Moreover, the numerator in (2.6) is $O(n)$ while the denominator is $\lambda_{\min}^4(C_n) = n^2\lambda_{\min}^2(\Psi)$, and therefore condition (2.6) is also satisfied. Hence we can apply Theorem 2 to conclude that $\widehat{\theta}_n$ is asymptotically normal with mean θ and covariance matrix $(n\Psi)^{-1}$, which agrees with the asymptotic distribution of the least squares estimate, constrained to lie inside a ball centered at θ with radius $n^{-1/2}\log n$, from the fixed design (2.10) that assumes knowledge of θ.

3. Martingales taking values in a Hilbert space and the proof of Theorems 1 and 2. To prove Theorems 1 and 2, we make use of martingales taking values in Hilbert spaces of the following type. Let B be a ball in $\mathbf{R}^m$ centered at $(\gamma_1, \ldots, \gamma_m)^T$. Let $L_2(B)$ denote the Hilbert space of square integrable real-valued functions on B. For $g \in L_2(B)$, define the function $\widetilde{g}\colon B \to \mathbf{R}$ by

$$\widetilde{g}(x) = \int_{\gamma_m}^{x_m} \cdots \int_{\gamma_1}^{x_1} g(t_1, \ldots, t_m)\, dt_1 \cdots dt_m, \qquad (x_1, \ldots, x_m)^T (= x) \in B. \tag{3.1}$$

Consider the Hilbert space $H = \{\widetilde{g}\colon g \in L_2(B)\}$ with norm $\|\cdot\|_H$ and inner product

$$\langle \widetilde{f}, \widetilde{g} \rangle_H = \int \cdots \int_B f(t_1, \ldots, t_m) g(t_1, \ldots, t_m)\, dt_1 \ldots dt_m, \qquad f, g \in L_2(B). \tag{3.2}$$

Let X_n be H-valued random variables such that $E\|X_n\|_H^2 < \infty$ for every n and $\{X_n\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{G}_n\}$. Then it is well known that $E\|\Sigma_1^n X_i\|_H^2 = \Sigma_1^n E\|X_i\|_H^2 = E\{\Sigma_1^n E(\|X_i\|_H^2 \mid \mathcal{G}_{i-1})\}$ and

$$\left\| \sum_{i=1}^n X_i \right\|_H^2 = O_p\left(\sum_{i=1}^n E\big(\|X_i\|_H^2 \mid \mathcal{G}_{i-1}\big) \right); \tag{3.3}$$

moreover, for every $\delta > 0$,

$$\left\| \sum_{i=1}^n X_i \right\|_H = o\left(\left\{ \sum_{i=1}^n E\big(\|X_i\|_H^2 \mid \mathcal{G}_{i-1}\big) \right\}^{(1+\delta)/2} \right) \quad \text{a.s.}; \tag{3.4}$$

cf. Lemma 3.1 of Morrow and Philipp (1982) and Lemma 2(iii) of Lai and Wei (1982).

Let $\{\varepsilon_n, \mathcal{G}_n, n \geq 1\}$ be a real-valued martingale difference sequence satisfying (1.2), as assumed in Theorems 1 and 2, and let $\widetilde{g}_n$ be $\mathcal{G}_{n-1}$-measurable H-valued random variables. Choose nonrandom constants a_n sufficiently large so that $P\{\|\widetilde{g}_n\|_H^2 + E(\varepsilon_n^2 \mid \mathcal{G}_{n-1}) > a_n\} \leq n^{-2}$. By the Borel–Cantelli lemma,

$$\begin{aligned} &P\{X_n = \varepsilon_n \widetilde{g}_n \text{ for all large } n\} = 1 \\ &\text{where } X_n = \varepsilon_n \widetilde{g}_n I_{\{\|\widetilde{g}_n\|_H^2 \leq a_n \text{ and } E(\varepsilon_n^2 \mid \mathcal{G}_{n-1}) \leq a_n\}}. \end{aligned} \tag{3.5}$$

Note that $\{X_n, \mathcal{G}_n, n \geq 1\}$ is a martingale difference sequence taking values in H with $E\|X_n\|_H^2 < \infty$ for all n. Hence by (3.3)–(3.5),

$$\left\| \sum_{i=1}^n \varepsilon_i \tilde{g}_i \right\|_H^2 = O_p\left(\sum_{i=1}^n \int \cdots \int_B g_i^2(t_1, \ldots, t_m)\, dt_1 \ldots dt_m \right), \tag{3.6}$$

$$\left\| \sum_{i=1}^n \varepsilon_i \tilde{g}_i \right\|_H^2 = o\left(\left\{ \sum_{i=1}^n \int \cdots \int_B g_i^2(t_1, \ldots, t_m)\, dt_1 \ldots dt_m \right\}^{1+\delta} \right) \tag{3.7}$$

a.s. for every $\delta > 0$,

where g_i is defined from $\tilde{g}_i$ via (3.1). Moreover, by the Schwarz inequality,

$$\begin{aligned} \sup_{x \in B} \left| \sum_{i=1}^n \varepsilon_i \tilde{g}_i(x) \right|^2 &= \sup_{x \in B} \left| \int_{\gamma_m}^{x_m} \cdots \int_{\gamma_1}^{x_1} \sum_{i=1}^n \varepsilon_i g_i(t_1, \ldots, t_m)\, dt_1 \ldots dt_m \right|^2 \\ &\leq \left\{ \int \cdots \int_B dt_1 \ldots dt_m \right\} \\ &\quad \times \left\{ \int \cdots \int_B \left[\sum_{i=1}^n \varepsilon_i g_i(t_1, \ldots, t_m) \right]^2 dt_1 \ldots dt_m \right\} \\ &= \mathrm{vol}(B) \left\| \sum_{i=1}^n \varepsilon_i \tilde{g}_i \right\|_H^2, \end{aligned} \tag{3.8}$$

which bounds the sup-norm of $\Sigma_1^n \varepsilon_i \tilde{g}_i$ by its H-norm.

The key idea in the proof of Theorem 1 is to apply (3.7) and (3.8) to the representation of $f_i(x) - f_i(\lambda)$ given in (3.10) below. This yields the asymptotic bound (3.11) which is then combined with condition (2.3) to complete the proof.

For the case $k = 1$, if f_i has continuous derivatives in an interval $B(\lambda)$ centered at λ, then $f_i(x) - f_i(\lambda) = \int_\lambda^x (df_i/d\xi)\, d\xi$ for all $x \in B(\lambda)$, so (3.1) holds with $\tilde{g}_i(x) = f_i(x) - f_i(\lambda)$ and $g_i = df_i/dt$. For the case $k = 2$, if f_i has continuous partial derivatives $\partial^2 f/\partial\xi_1 \partial\xi_2$, $\partial f/\partial\xi_1$ and $\partial f/\partial\xi_2$ in a disk $B(\lambda)$ centered at λ, then for all $x \in B(\lambda)$,

$$\begin{aligned} f_i(x) - f_i(\lambda) &= \int_{\lambda_2}^{x_2} \int_{\lambda_1}^{x_1} \frac{\partial^2 f_i}{\partial\xi_1 \partial\xi_2}(\xi_1, \xi_2)\, d\xi_1\, d\xi_2 \\ &\quad + \int_{\lambda_2}^{x_2} \frac{\partial f_i}{\partial\xi_2}(\lambda_1, \xi_2)\, d\xi_2 + \int_{\lambda_1}^{x_1} \frac{\partial f_i}{\partial\xi_1}(\xi_1, \lambda_2)\, d\xi_1. \end{aligned} \tag{3.9}$$

Letting $g_i = \partial^2 f_i/\partial\xi_1 \partial\xi_2$, $h_{i,1}(\xi_2) = (\partial/\partial\xi_2) f_i(\lambda_1, \xi_2)$, $h_{i,2}(\xi_1) = (\partial/\partial\xi_1) f_i(\xi_1, \lambda_2)$, it follows from (3.1) that we can rewrite (3.9) as

$$f_i(x) - f_i(\lambda) = \tilde{g}_i(x_1, x_2) + \tilde{h}_{i,1}(x_2) + \tilde{h}_{i,2}(x_1). \tag{3.9'}$$

Note that $\tilde{g}_i, \tilde{h}_{i,1}, \tilde{h}_{i,2}$ belong to different Hilbert spaces H_0, H_1, H_2 of functions of the type (3.1) defined on $\{(x_1, x_2)\colon (x_1 - \lambda_1)^2 + (x_2 - \lambda_2)^2 < r^2\}$, $\{t\colon |t - \lambda_2| < r\}$ and $\{t\colon |t - \lambda_1| < r\}$, respectively, and therefore we can apply (3.6)–(3.8) separately to $\Sigma_1^n \varepsilon_i \tilde{g}_i$, $\Sigma_1^n \varepsilon_i \tilde{h}_{i,1}$ and $\Sigma_1^n \varepsilon_i \tilde{h}_{i,2}$. An induction argument shows that for general $k \geq 1$, if f_i has continuous partial derivatives $D_{\mathbf{j}} f_i$ in a ball $B(\lambda)$ centered at λ for every $\mathbf{j} \in J(m,k)$ and every $m \leq k$, then for all $x \in B(\lambda)$,

$$f_i(x) - f_i(\lambda) = \sum_{m=1}^{k} \sum_{\mathbf{j} \in J(m,k)} \int_{\lambda_{j_m}}^{x_{j_m}} \cdots \int_{\lambda_{j_1}}^{x_{j_1}} D_{\mathbf{j}} f_i|_{\xi_j = \lambda_j \ \forall_j \notin \{j_1, \ldots, j_m\}} \, d\xi_{j_1} \ldots d\xi_{j_m}, \tag{3.10}$$

where $J(m,k)$ is defined in (2.1). Applying (3.8) and (3.7) to (3.10) then shows that for every $\delta > 0$,

$$\sup_{x \in B(\lambda)} \left| \sum_{i=1}^{n} \varepsilon_i \big(f_i(x) - f_i(\lambda)\big) \right| \leq \sum_{m=1}^{k} \sum_{\mathbf{j} \in J(m,k)} o\left(\left\{ \sum_{i=1}^{n} \int_{B(\lambda;\, \mathbf{j})} (D_{\mathbf{j}} f_i)^2 \, d\xi_{j_1} \cdots d\xi_{j_m} \right\}^{(1+\delta)/2} \right) \quad \text{a.s.} \tag{3.11}$$

To prove Theorem 2, we shall replace f_i in (3.10) by $\partial^2 f_i / \partial \xi_a \, \partial \xi_b$ for every fixed $a, b \in \{1, \ldots, k\}$. Moreover, we shall use (3.6) instead of (3.7) and consider the case $\lambda = \theta$. Therefore, analogous to (3.11), we have

$$\sup_{x \in B(\theta)} \left| \sum_{i=1}^{n} \varepsilon_i \left\{ \frac{\partial^2 f_i(x)}{\partial \xi_a \, \partial \xi_b} - \frac{\partial^2 f_i(\theta)}{\partial \xi_a \, \partial \xi_b} \right\} \right|^2 \leq \sum_{m=1}^{k} \sum_{\mathbf{j} \in J(m,k)} O_p\left(\sum_{i=1}^{n} \int_{B(\theta;\, \mathbf{j})} \left[D_{\mathbf{j}} \left(\partial^2 f_i / \partial \xi_a \, \partial \xi_b \right) \right]^2 d\xi_{j_1} \cdots d\xi_{j_m} \right). \tag{3.12}$$

PROOF OF THEOREM 1. Take any $\delta > 0$. Since Θ is compact, $\Theta_\delta := \{\lambda \in \Theta\colon \|\lambda - \theta\| \geq \delta\}$ is also compact and is therefore covered by a finite number of open balls $B(\lambda)$ in Θ, centered at $\lambda \neq \theta$, that satisfy (2.2) and (2.3) for some $1 < \rho_\lambda < 2$. It therefore suffices to show that for each of these (finitely many) balls,

$$\lim_{n \to \infty} \inf_{\xi \in B(\lambda)} \big(S_n(\xi) - S_n(\theta)\big) = \infty \quad \text{a.s.} \tag{3.13}$$

Since $S_n(\xi) - S_n(\theta) = \Sigma_{i=1}^n [f_i(\xi) - f_i(\theta)]^2 - 2\Sigma_{i=1}^n \varepsilon_i (f_i(\xi) - f_i(\lambda)) - 2\Sigma_{i=1}^n \varepsilon_i (f_i(\lambda) - f_i(\theta))$ and since $\Sigma_{i=1}^n \varepsilon_i (f_i(\lambda) - f_i(\theta)) = o(\{\Sigma_{i=1}^n [f_i(\lambda) - f_i(\theta)]^2\}^{(1+\delta)/2})$ a.s. for every $\delta > 0$, we need only show, in view of (2.2) and (2.3), that

$$\sup_{\xi \in B(\lambda)} \left| \sum_{i=1}^{n} \varepsilon_i \big(f_i(\xi) - f_i(\lambda)\big) \right| = o\left(\inf_{\xi \in B(\lambda)} \sum_{i=1}^{n} \big[f_i(\xi) - f_i(\theta)\big]^2 \right) \quad \text{a.s.} \tag{3.14}$$

From (3.11) and (2.3), the desired conclusion (3.14) follows. □

PROOF OF THEOREM 2. By Theorem 1, with probability 1, $\widehat{\theta}_n \to \theta$ and therefore $\widehat{\theta}_n \in B(\theta)$ for all large n. By (1.7),

$$\begin{aligned} 0 &= -\nabla S_n(\widehat{\theta}_n)/2 = \sum_{i=1}^n \varepsilon_i \nabla f_i(\widehat{\theta}_n) + \sum_{i=1}^n \nabla f_i(\widehat{\theta}_n)\big(f_i(\theta) - f_i(\widehat{\theta}_n)\big) \\ &= \sum_{i=1}^n \varepsilon_i \nabla f_i(\theta) + \left\{ \sum_{i=1}^n \varepsilon_i \nabla^2 f_i(\theta) - \sum_{i=1}^n \nabla f_i(\theta)\big(\nabla f_i(\theta)\big)^T + R_n \right\}(\widehat{\theta}_n - \theta), \end{aligned} \tag{3.15}$$

where, by the mean value theorem,

$$\begin{aligned} R_n = &\sum_{i=1}^n \varepsilon_i \left(\frac{\partial^2 f_i(\theta_{n,a})}{\partial \xi_a \, \partial \xi_b} - \frac{\partial^2 f_i(\theta)}{\partial \xi_a \, \partial \xi_b} \right)_{1 \le a,b \le k} \\ &- \sum_{i=1}^n \left(\frac{\partial^2 f_i(\theta^*_{n,a})}{\partial \xi_a \, \partial \xi_b} \right)_{1 \le a,b \le k} (\widehat{\theta}_n - \theta)\big(\nabla f_i(\theta)\big)^T \\ &- \sum_{i=1}^n \nabla f_i(\theta)(\theta_n - \theta)^T \left(\frac{\partial^2 f_i(\theta^{**}_{n,a})}{\partial \xi_a \, \partial \xi_b} \right)^T_{1 \le a,b \le k} \\ &- \sum_{i=1}^n \left(\frac{\partial^2 f_i(\theta^*_{n,a})}{\partial \xi_a \, \partial \xi_b} \right)_{1 \le a,b \le k} (\widehat{\theta}_n - \theta)(\theta_n - \theta)^T \left(\frac{\partial^2 f_i(\theta^{**}_{n,a})}{\partial \xi_a \, \partial \xi_b} \right)^T_{1 \le a,b \le k}, \end{aligned} \tag{3.16}$$

with $\theta_n, \theta_{n,a}, \theta^*_{n,a}$ and $\theta^{**}_{n,a}$ lying between θ and $\widehat{\theta}_n$. From (3.15) it follows that

$$\begin{aligned} \sum_{i=1}^n \varepsilon_i \nabla f_i(\theta) = C_n \Bigg\{ &\sum_{i=1}^n \big(C_n^{-1} \nabla f_i(\theta)\big)\big(C_n^{-1} \nabla f_i(\theta)\big)^T \\ &- C_n^{-1} \sum_{i=1}^n \varepsilon_i \nabla^2 f_i(\theta) C_n^{-1} - C_n^{-1} R_n C_n^{-1} \Bigg\} C_n(\widehat{\theta}_n - \theta). \end{aligned} \tag{3.17}$$

By (3.12) and (2.6), with probability 1, for all large n,

$$\begin{aligned} &\left\| C_n^{-1} \sum_{i=1}^n \varepsilon_i \left(\frac{\partial^2 f_i(\theta_{n,a})}{\partial \xi_a \, \partial \xi_b} - \frac{\partial^2 f_i(\theta)}{\partial \xi_a \, \partial \xi_b} \right)_{1 \le a,b \le k} C_n^{-1} \right\|^2 \\ &\le k^2 \|C_n^{-1}\|^4 \sup_{x \in B(\theta),\, 1 \le a,b \le k} \left| \sum_{i=1}^n \varepsilon_i \left\{ \frac{\partial^2 f_i(x)}{\partial \xi_a \, \partial \xi_b} - \frac{\partial^2 f_i(\theta)}{\partial \xi_a \, \partial \xi_b} \right\} \right|^2 \xrightarrow{P} 0, \end{aligned} \tag{3.18}$$

noting that $\|C_n^{-1}\| = \lambda_{\max}(C_n^{-1}) = 1/\lambda_{\min}(C_n)$. Analogous to (3.3), we also have

$$\left\| C_n^{-1} \left\{ \sum_{i=1}^n \varepsilon_i \nabla^2 f_i(\theta) \right\} C_n^{-1} \right\|^2 \le \|C_n^{-1}\|^4 O_p\left(\sum_{i=1}^n \|\nabla^2 f_i(\theta)\|^2 \right) \xrightarrow{P} 0 \tag{3.19}$$

by (2.5). Noting that $\|\Sigma_{i=1}^n x_i y_i^T\|^2 \le (\Sigma_{i=1}^n \|x_i\|^2)(\Sigma_{i=1}^n \|y_i\|^2)$ by the Schwarz inequality for $k \times 1$ vectors x_i and y_i, we obtain that

$$\begin{aligned}
&\left\| C_n^{-1} \sum_{i=1}^n \left(\frac{\partial^2 f_i(\theta_{n,a}^*)}{\partial \xi_a \partial \xi_b} \right)_{1 \le a,b \le k} (\hat{\theta}_n - \theta)\big(C_n^{-1} \nabla f_i(\theta)\big)^T \right\|^2 \\
&\le \|C_n^{-1}\|^2 \left\{ \sum_{i=1}^n \|C_n^{-1} \nabla f_i(\theta)\|^2 \right\} \\
&\quad \times \left\{ \sum_{i=1}^n k^2 \max_{1 \le a,b \le k} \left| \frac{\partial^2 f_i(\theta_{n,a}^*)}{\partial \xi_a \partial \xi_b} \right|^2 \|\hat{\theta}_n - \theta\|^2 \right\} \xrightarrow{P} 0
\end{aligned} \tag{3.20}$$

by (2.4) and (2.5), since $\|C_n^{-1}\|^2 = 1/\lambda_{\min}^2(C_n)$. Similarly it can be shown that

$$\left\| \sum_{i=1}^n C_n^{-1} \nabla f_i(\theta) \left[\left(\frac{\partial^2 f_i(\theta_{n,a}^{**})}{\partial \xi_a \partial \xi_b} \right)_{1 \le a,b \le k} (\theta_n - \theta) \right]^T C_n^{-1} \right\|^2 \xrightarrow{P} 0, \tag{3.21}$$

$$\begin{aligned}
&\left\| C_n^{-1} \sum_{i=1}^n \left[\left(\frac{\partial^2 f_i(\theta_{n,a}^*)}{\partial \xi_a \partial \xi_b} \right)_{1 \le a,b \le k} (\hat{\theta}_n - \theta) \right] \right. \\
&\quad \left. \times \left[\left(\frac{\partial^2 f_i(\theta_{n,a}^{**})}{\partial \xi_a \partial \xi_b} \right)_{1 \le a,b \le k} (\theta_n - \theta) \right]^T C_n^{-1} \right\|^2 \xrightarrow{P} 0.
\end{aligned} \tag{3.22}$$

Combining (3.17) with (3.16) and (3.18)–(3.22), we obtain that

$$\begin{aligned}
&\left\{ \sum_{i=1}^n \big(\nabla f_i(\theta)\big)\big(\nabla f_i(\theta)\big)^T \right\}^{1/2} (\hat{\theta}_n - \theta) \\
&= \left\{ \left[\sum_{i=1}^n \big(\nabla f_i(\theta)\big)\big(\nabla f_i(\theta)\big)^T \right]^{1/2} C_n^{-1} \right\} \{C_n(\hat{\theta}_n - \theta)\} \\
&= \big(1 + o_p(1)\big) C_n^{-1} \sum_{i=1}^n \varepsilon_i \nabla f_i(\theta).
\end{aligned} \tag{3.23}$$

Since $E(\varepsilon_n^2 \mid \mathcal{G}_{n-1}) \xrightarrow{P} \sigma^2$ and since $\max_{1 \le i \le n} \|C_n^{-1} \nabla f_i(\theta)\| \xrightarrow{P} 0$ while $C_n^{-1}\{\Sigma_{i=1}^n (\nabla f_i(\theta))(\nabla f_i(\theta))^T\} C_n^{-1} \xrightarrow{P} I$, $C_n^{-1} \Sigma_{i=1}^n \varepsilon_i \nabla f_i(\theta)$ converges in distribution to $N(0, \sigma^2 I)$ by the martingale central limit theorem [cf. Hall and Heyde (1980), pages 58 and 175], and therefore the desired conclusion follows from (3.23). □

REFERENCES

ANDERSON, T. W. and TAYLOR, J. (1979). Strong consistency of least squares estimators in dynamic models. *Ann. Statist.* **7** 484–489.

BATES, D. M. and WATTS, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

CHRISTOPEIT, N. and HELMES, K. (1980). Strong consistency of least squares estimators in linear regression models. *Ann. Statist.* **8** 778–788.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments*. Academic, New York.

FORD, I. and SILVEY, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika* **67** 381–388.

FORD, I., TITTERINGTON, D. M. and WU, C. F. J. (1985). Inference and sequential design. *Biometrika* **72** 545–551.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic, New York.

HOFFMANN-JØRGENSEN, J. and PISIER, G. (1976). The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probab.* **4** 587–599.

KLIMKO, L. A. and NELSON, P. I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* **6** 629–642.

LAI, T. L. and WEI, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10** 154–166.

LAI, T. L. and WEI, C. Z. (1987). Asymptotically efficient self-tuning regulators. *SIAM J. Control Optim.* **25** 466–481.

LAI, T. L. and ZHU, G. (1991). Adaptive prediction in non-linear autoregressive models and control systems. *Statist. Sinica* **1** 309–334.

MORROW, G. and PHILIPP, W. (1982). An almost sure invariance principle for Hilbert space valued martingales. *Trans. Amer. Math. Soc.* **273** 231–251.

WU, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9** 501–513.

WU, C. F. J. (1985). Asymptotic inference from sequential design in a nonlinear situation. *Biometrika* **72** 553–558.

ZHU, G. (1992). Least squares estimation and adaptive prediction in non-linear stochastic regression models with applications to time series and stochastic systems. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305