The Annals of Statistics
2000, Vol. 28, No. 3, 871–879

STRONG CONSISTENCY IN NONLINEAR STOCHASTIC REGRESSION MODELS

BY K. SKOURAS

University College London

The class of nonlinear stochastic regression models includes most of the linear and nonlinear models used in time series, stochastic control and stochastic approximation schemes. The consistency of least squares estimators was established first by Lai. We present another set of sufficient conditions for consistency, which avoid the use of partial derivatives and are closer in spirit to the conditions presented by Wu for non-stochastic regression models with independent errors.

1. Introduction. Let $(\Omega, \mathscr{F}, (\mathscr{F}_t), P)$ be a filtered probability space and Θ a compact subset of $\mathbb{R}^p$. We consider the general stochastic regression model:

$$Y_t = f_t(\theta) + \varepsilon_t, \tag{1.1}$$

where $Y_t \in \mathbb{R}$ is $\mathscr{F}_t$-measurable, $f_t(\theta)$ is a $\mathscr{F}_{t-1}$-measurable real function of $\theta \in \Theta$, and (ϵ_t) is a martingale difference sequence, such that with probability one

$$\sup_t E_t(\varepsilon_t^2) < \infty, \tag{1.2}$$

where by $E_t(\cdot)$ we denote the conditional expectation $E(\cdot|\mathscr{F}_{t-1})$.

The class of models described by (1.1) is very wide, and includes many linear and nonlinear regression models commonly used. For example, the $f_t(\theta)$'s can be linear or nonlinear functions of past observations $Y^{t-1} := (Y_1, Y_2, \ldots, Y_{t-1})$, and any other covariates $x^t := (x_1, x_2, \ldots, x_t)$ such that x_t is $\mathscr{F}_{t-1}$-measurable. Important classes of models that can be described by (1.1) are nonlinear time series models (with or without exogenous inputs), stochastic control models, stochastic approximation schemes and sequential designs.

If by θ_0 we denote the (unknown) true value of θ, i.e. $E_t(Y_t) = f_t(\theta_0)$ for all $t \geq 1$, then we may estimate θ_0 using the least squares estimator, which is defined as the parameter value $\hat{\theta}_T$ which minimizes the sum of squared errors:

$$S_T(\theta) := \sum_{t=1}^{T} \{Y_t - f_t(\theta)\}^2.$$

Observe that $S_T(\theta)$ is the cumulative predictive squared error loss, when one is predicting the observations Y_t using $f_t(\theta)$. The method of least squares is based on the rationale that the sequence of true predictions $\{f_t(\theta_0)\}$ is expected to beat eventually any other sequence of predictions generated by the

Received November 1998; revised December 1999.
AMS 1991 *subject classifications*. Primary 62J02; secondary 62M10, 62F12, 60F15.
Key words and phrases. Consistency, least squares estimator, martingale, stochastic regression.

other false models, thus allowing us to identify eventually the true parameter value θ_0. Using this approach, we present in this paper a set of sufficient conditions, on the functions $f_t(\theta)$, which guarantee the consistency of the least squares estimator.

In contrast to the conditions of Lai (1994), ours do not assume differentiability of $f_t(\theta)$, but only the weaker smoothness condition (3.10) below. However, as our example 2 below shows, our conditions can be more restrictive in other respects. In conclusion, no set of conditions is weaker than the other, and each one can be useful in different situations.

The rest of the paper is organized as follows. In Section 2 we review the existing sets of sufficient conditions for consistency of the least squares estimator in nonlinear regression models, and in Section 3 we present our result together with some examples. The proof can be found in Section 4.

2. Review of existing results. The strong consistency of the estimator $\hat{\theta}_T$ for regression models is an important problem, especially for identification and control, and has been studied extensively [see Anderson and Taylor (1979), Christopeit and Helmes (1980), Wu (1981), Lai and Wei (1982), White and Domowitz (1984), Gallant and White (1988), Lai (1994) and references therein]. Hu (1996, 1997, 1998) has studied the consistency of Bayesian estimators for linear and nonlinear stochastic regression models, but his results for the general nonlinear model are confined to the case of independent errors and discrete parameter sets. In order to avoid overwhelming the reader with an extensive review of the sufficient conditions that different authors have presented, we focus our presentation only on papers closely related to model (1.1).

Lai and Wei (1982) have shown the consistency of least squares estimators, under weak conditions, for linear stochastic regression models. The breakthrough for the nonlinear case came from Lai (1994), who presented sufficient conditions for consistency of least squares estimators for the model described by (1.1) and (1.2). Lai's conditions are the following:

CONDITION L1. For $1 \le m \le p$, let

$$J(m, p) = \{(j_1, \ldots, j_m) : j_1 < \cdots < j_m, j_i \in \{1, \ldots, p\} \text{ for } 1 \le i \le m\},$$

and for $\mathbf{j} = (j_1, \ldots, j_m) \in J(m, p)$, let $D_{\mathbf{j}} f_t := \partial^m f_t / \partial \theta_{j_1} \ldots \theta_{j_m}$. Assume that, for every t, $f_t(\theta)$ has continuous partial derivatives $D_{\mathbf{j}}$ on Θ, for every $\mathbf{j} \in J(m, p)$ and $m = 1, \ldots, p$.

CONDITION L2. For every θ_1, θ_2 in Θ let

$$A_T(\theta_1, \theta_2) := \sum_{t=1}^{T} \{f_t(\theta_1) - f_t(\theta_2)\}^2.$$

For every $\lambda \neq \theta_0$ there exists $1 < r_\lambda < 2$, and an open ball $B(\lambda)$ centered at λ, such that with probability one,

$$A_T := \inf_{\theta \in B(\lambda)} A_T(\theta_0, \theta) \to \infty \tag{2.3}$$

and

$$\max_{1\le m\le p\ ,\ \mathbf{j}\in J(m,p)} \sum_{t=1}^{T} \int_{B(\lambda,\mathbf{j})} \{D_{\mathbf{j}} f_t\}^2 d\theta_{j_1} \cdots d\theta_{j_m} + A_T(\theta_0, \lambda) = O\left(A_T^{r_\lambda}\right), \tag{2.4}$$

where

$$B(\lambda, \mathbf{j}) := \{(\theta_{j_1}, \ldots, \theta_{j_m}) : (\lambda_1, \ldots, \lambda_{j_1-1}, \theta_{j1}, \lambda_{j_1+1}, \ldots, \lambda_{j_m-1}, \theta_{j_m}, \lambda_{j_m+1}, \ldots, \lambda_p) \in B(\lambda)\}.$$

REMARK 1. In Condition L2, equation (2.3), the stochastic quantity $A_T(\theta_0, \theta)$ is the accumulated one-step ahead squared predictive bias when one is predicting Y_t using $f_t(\theta)$ instead of the optimum prediction $f_t(\theta_0)$. In that sense, A_T (which depends on λ, although we omit this dependence from the notation for simplicity) is a measure of the information available in the data for separation of θ_0 from the subset $B(\lambda)$. In order to be able to identify θ_0 with probability one, this information should tend to infinity, as is described in equation (2.3). Condition L3, equation (2.4), describes the permitted rate of growth for $A_T(\theta_0, \lambda)$ and for the integrated squared partial derivatives of $f_t(\theta)$. See Remark 3 for a discussion of why such a condition is needed.

When the $f_t(\theta)$ are non-stochastic continuous functions, and (ε_t) is a sequence of independent identically distributed errors, Wu (1981) showed that consistency can be established under the following conditions (using the same notation as in Lai's conditions):

CONDITION W1. For every $\lambda \neq \theta_0$ there exist an open ball $B(\lambda)$ centered at λ, such that for some $M > 0$ and $1 < r_\lambda < 2$:

$$A_T = \inf_{\theta\in B(\lambda)} A_T(\theta_0, \theta) \to \infty, \tag{2.5}$$

$$\sum_{t=1}^{T} \sup_{\theta\in B(\lambda)} \{f_t(\theta) - f_t(\theta_0)\}^2 = O\left(A_T^{r_\lambda}\right) \tag{2.6}$$

and, for all t,

$$\sup_{\theta,\theta'\in B(\lambda),\theta\neq\theta'} \frac{|f_t(\theta) - f_t(\theta')|}{\|\theta - \theta'\|} \le M \sup_{\theta\in B(\lambda)} |f_t(\theta) - f_t(\theta_0)|. \tag{2.7}$$

The two sets of conditions look similar, for example conditions (2.3) and (2.5) are the same, but there is an important difference in the smoothness condition that the two authors use. Lai uses partial derivatives, while Wu uses Condition (2.7) which allows him to use probabilistic results for Lipschitz continuous functions. As Lai pointed out [Lai (1994), Section 2] this approach can not be extended directly to the case where (ε_t) is a martingale difference sequence and $f_t(\theta)$ are $\mathcal{F}_{t-1}$-measurable and a stronger smoothness condition is needed. He overcame this problem using Condition L2, which allows the embedding of the functions $f_t(\theta)$ in a suitably chosen Hilbert space, and therefore results

for martingales taking values in Hilbert spaces can be used to establish consistency. Our approach uses a modified version of the smoothness condition (2.7), based on stochastic Lipschitz upper bounds.

3. Main result. We introduce the following condition:

CONDITION C1. For every $\lambda \neq \theta_0$ there exists $1 < r_\lambda < 2$, and an open ball $B(\lambda)$ centered at λ, such that with probability one:

$$A_T = \inf_{\theta \in B(\lambda)} A_T(\theta_0, \theta) \to \infty \tag{3.8}$$

and

$$\sum_{t=1}^{T} \sup_{\theta \in B(\lambda)} \{f_t(\theta) - f_t(\theta_0)\}^2 = O\left(A_T^{r_\lambda}\right). \tag{3.9}$$

Also, there is a sequence of $\mathcal{F}_{t-1}$-measurable positive variables $M_t(\lambda)$ such that for all θ_1, θ_2 in $B(\lambda)$

$$|f_t(\theta_1) - f_t(\theta_2)| \leq h(\|\theta_1 - \theta_2\|)\, M_t(\lambda), \tag{3.10}$$

and with probability one,

$$\frac{1}{A_T} \sum_{t=1}^{T} M_t(\lambda) = O(1), \tag{3.11}$$

where $h(\cdot)$ is a non-random function such that $h(y) \downarrow h(0) = 0$, as $y \downarrow 0$.

THEOREM 1. *Assume that Condition* C1 *holds. Then, with probability one,* $\|\hat{\theta}_T - \theta_0\| \to 0$.

REMARK 2. Equations (3.8) and (3.9) are the same as Wu's conditions (2.5) and (2.6), the only difference being that since $f_t(\theta)$ are stochastic in our case, these conditions should hold with probability one. Although we have managed to keep most of Wu's conditions, our approach is not a direct extension of his method. Our condition described by equations (3.10) and (3.11) is a modification (with random norming) of the Lipschitz-L_1 condition used by Andrews (1987) [see also Gallant and White (1988)]. Using this condition, together with martingale arguments, we can establish a uniform law of large numbers (cf. Lemma 1), which is the necessary tool for the proof of the main result.

REMARK 3. Observe that in all sets of conditions discussed or presented in this paper, there is one condition that asks that the information A_T should tend to infinity, and a smoothing condition that sets a restriction on the rate of growth for some upper bound either on the partial derivatives [equation (2.4)], or on the Lipschitz bound [equations (2.6), (2.7) and (3.9)–(3.11)]. These two conditions serve two different aims in the proof. The first condition makes sure that any two models in the family are "far" enough apart, in terms of their predictive performance, so that when one of them is the true model one

can identify it. The second condition makes sure that in small neighborhoods all models are "close" enough, so that when all of them are wrong one can establish that their predictive performances are uniformly inferior to that of the true model.

EXAMPLE 1. Let $Y_0 = 0$, and $\mathcal{F}_t = \sigma(Y_0, \ldots, Y_t)$. Consider the model

$$Y_t = \begin{cases} \alpha(\theta - Y_{t-1}) + \varepsilon_t & \theta \geq Y_{t-1}, \\ \beta(\theta - Y_{t-1}) + \varepsilon_t & \theta < Y_{t-1}, \end{cases}$$

where $\alpha > \beta > 0$ are known constants, $\Theta = [-R, R]$, and (ε_t) is a martingale difference sequence satisfying condition (1.2). Since for each t, $f_t(\theta)$ has no derivative at $\theta = Y_{t-1}$, Lai's conditions cannot be used. The conditions cannot be modified in a way that avoids the problem of nonexistence of the derivative at a single point, because Y_{t-1} is stochastic and can take values anywhere in $\mathbb{R}$.

For every $\theta_1 \geq \theta_2 \in \Theta$ we have

$$|f_t(\theta_1) - f_t(\theta_2)| = \begin{cases} \alpha(\theta_1 - \theta_2), & Y_{t-1} < \theta_2, \\ \beta(Y_{t-1} - \theta_2) + \alpha(\theta_1 - Y_{t-1}), & \theta_2 \leq Y_{t-1} < \theta_1, \\ \beta(\theta_1 - \theta_2), & \theta_1 \leq Y_{t-1}. \end{cases} \tag{3.12}$$

We can now study if Condition C1 holds. Let $\lambda \neq \theta_0$. We study only the case $\lambda > \theta_0$, as the other case can be studied similarly. Let $B(\lambda) = (\lambda_1, \lambda_2)$ such that $\theta_0 < \lambda_1 < \lambda < \lambda_2$. Then, using equation (3.12), we can show that

$$\inf_{\theta \in B(\lambda)} |f_t(\theta_0) - f_t(\theta)| \geq \beta(\lambda_1 - \theta_0)$$

and

$$\sup_{\theta \in B(\lambda)} |f_t(\theta_0) - f_t(\theta)| \leq (\alpha + \beta)(\lambda_2 - \theta_0),$$

and therefore conditions (3.8) and (3.9) hold, with $A_T = T$. Now, for every θ_1, θ_2 in $B(\lambda)$,

$$|f_t(\theta_1) - f_t(\theta_2)| \leq (\alpha + \beta)|\theta_1 - \theta_2|,$$

and it is easily shown that conditions (3.10) and (3.11) hold with $h(|\theta_1 - \theta_2|) = |\theta_1 - \theta_2|$ and $M_t(\lambda) = (\alpha + \beta)$. Since Condition C1 holds, then the least squares estimator is consistent.

EXAMPLE 2. Consider the model $Y_t = \exp(-\theta x_t) + \varepsilon_t$, where $\theta \in [\alpha, \beta]$, with $\alpha > 0$, and (x_t) is a sequence of bounded positive regressors, such that x_t is $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_{t-1}$ is the σ-algebra generated by $(x_1, Y_1, \ldots, x_{t-1}, Y_{t-1})$. For every θ_1, θ_2 there exist constants $c_1, c_2 > 0$ so that

$$c_1 |\theta_1 - \theta_2| x_t \leq |\exp(-\theta_1 x_t) - \exp(-\theta_2 x_t)| \leq c_2 |\theta_1 - \theta_2| x_t, \tag{3.13}$$

and therefore for every open interval $B(\lambda)$, which does not include θ_0, there is a constant $C > 0$ such that

$$A_T = \inf_{\theta \in B(\lambda)} \sum_{t=1}^{T} \{\exp(-\theta x_t) - \exp(-\theta_0 x_t)\}^2 \geq C \sum_{t=1}^{T} x_t^2 \tag{3.14}$$

and

$$\sum_{t=1}^{T} \sup_{\theta \in B(\lambda)} \{\exp(-\theta x_t) - \exp(-\theta_0 x_t)\}^2 = O\left(\sum_{t=1}^{T} x_t^2\right). \tag{3.15}$$

Equations (3.13), (3.14) and (3.15) show that the least squares estimator is consistent if

$$\sum_{t=1}^{T} x_t^2 \to \infty \quad \text{and} \quad \sum_{t=1}^{T} x_t = O\left(\sum_{t=1}^{T} x_t^2\right). \tag{3.16}$$

Lai's conditions in this example boil down to the single condition $\sum x_t^2 \to \infty$, which is weaker than (3.16) which excludes cases where $|x_t| \to 0$. From examples 1 and 2, we see that no set of conditions is weaker than the other, and each one can be useful in different cases.

EXAMPLE 3. Consider the power curve model $Y_t = (t + \theta)^d + \varepsilon_t$, where $d \geq 1/2$ is a known constant, and $\Theta = [\alpha, \beta] \subseteq \mathbb{R}$. Wu (1981) considered the same model, but with independent errors. For every $B(\lambda)$ which does not include θ_0, it is easy to show that there exists a constant $C > 0$ such that

$$A_T = \inf_{\theta \in B(\lambda)} \sum_{t=1}^{T} \left\{(t + \theta)^d - (t + \theta_0)^d\right\}^2 \geq C\, T^{2d-1}, \tag{3.17}$$

$$\sum_{t=1}^{T} \sup_{\theta \in B(\lambda)} \left\{(t + \theta)^d - (t + \theta_0)^d\right\}^2 = O(T^{2d-1}) \tag{3.18}$$

and also that there are constants c_1 and c_2 such that

$$|(t + \theta)^d - (t + \theta_0)^d| \leq c_1 |\theta - \theta_0| (t + c_2)^{d-1}. \tag{3.19}$$

The fact that

$$\sum_{t=1}^{T} (t + c_2)^{d-1} = O(T^d),$$

combined with equations (3.17), (3.18) and (3.19), implies that Condition C1 holds, and therefore the least squares estimator is consistent.

4. Proof. First we need to prove the following lemma:

LEMMA 1. *Assumption* C1 *implies that, with probability one,*

$$\frac{1}{A_T} \sup_{\theta \in B(\lambda)} \left| \sum_{t=1}^{T} \varepsilon_t \{f_t(\theta) - f_t(\theta_0)\} \right| \to 0.$$

PROOF. The technique we use is an extension, to martingale error differences, of the approach used by Andrews (1987). Let $B(\theta,\rho) := \{s \in B(\lambda) : \|\theta - s\| < \rho\}$, and $l_t(\theta) := \{f_t(\theta) - f_t(\theta_0)\}$. Observe that for every $\theta \in B(\lambda)$ and $\rho > 0$,

$$\begin{aligned}
&\frac{1}{A_T}\sum_{t=1}^{T} E_t\left\{\sup_{s\in B(\theta,\rho)} \varepsilon_t l_t(s) - \inf_{s\in B(\theta,\rho)} \varepsilon_t l_t(s)\right\} \\
&\quad = \frac{1}{A_T}\sum_{t=1}^{T}\left[E_t\left\{\sup_{s\in B(\theta,\rho)} \varepsilon_t l_t(s) - \varepsilon_t l_t(\lambda)\right\} + E_t\left\{\varepsilon_t l_t(\lambda) - \inf_{s\in B(\theta,\rho)} \varepsilon_t l_t(s)\right\}\right] \\
&\quad \le 2\, h(\rho)\, \left\{\sup_t E_t(|\varepsilon_t|)\right\}\, \left\{\frac{1}{A_T}\sum_{t=1}^{T} M_t(\lambda)\right\}.
\end{aligned}$$

Since $\sup_t E_t(\varepsilon_t^2) < \infty$, then $\sup_t E_t(|\varepsilon_t|) < \infty$, which implies that with probability one,

$$\lim_{\rho\to 0}\left[\limsup_{T\to\infty} \frac{1}{A_T}\sum_{t=1}^{T} E_t\left\{\sup_{s\in B(\theta,\rho)} \varepsilon_t l_t(s) - \inf_{s\in B(\theta,\rho)} \varepsilon_t l_t(s)\right\}\right] \to 0. \tag{4.20}$$

Let $\delta > 0$. Using result (4.20) for every θ there exists an event $F(\theta)$, with $P\{F(\theta)\} = 1$, such that for all ω in $F(\theta)$ we can choose ρ_θ so that for all $T \ge 1$,

$$\limsup_{T\to\infty} \frac{1}{A_T}\sum_{t=1}^{T} E_t\left\{\sup_{s\in B(\theta,\rho_\theta)} \varepsilon_t l_t(s) - \inf_{s\in B(\theta,\rho_\theta)} \varepsilon_t l_t(s)\right\} < \delta.$$

The collection of balls $\{B(\theta,\rho_\theta), \theta \in B(\lambda)\}$ is an open cover of $B(\lambda)$, and therefore, since $B(\lambda)$ is bounded, there is a finite subcover $\{B(\theta_j,\rho_{\theta_j}), j = 1,\dots,J\}$. For ease of notation let $B_j = B(\theta_j,\rho_{\theta_j})$.

If $F_0 = \bigcap_{j=1}^{J} F(\theta_j)$, then for all $\omega \in F_0$ and any $\theta \in B(\lambda)$, we have (for T sufficiently large),

$$\begin{aligned}
\frac{1}{A_T}\sum_{t=1}^{T} \varepsilon_t l_t(\theta) &\le \max_j \frac{1}{A_T}\sum_{t=1}^{T}\left[\sup_{s\in B_j} \varepsilon_t l_t(s) - E_t\left\{\sup_{s\in B_j} \varepsilon_t l_t(s)\right\}\right] \\
&\quad + \max_j \frac{1}{A_T}\sum_{t=1}^{T} E_t\left\{\sup_{s\in B_j} \varepsilon_t l_t(s) - \inf_{s\in B_j} \varepsilon_t l_t(s)\right\} \\
&\le \max_j \frac{1}{A_T}\sum_{t=1}^{T}\left[\sup_{s\in B_j} \varepsilon_t l_t(s) - E_t\left\{\sup_{s\in B_j} \varepsilon_t l_t(s)\right\}\right] + \delta.
\end{aligned} \tag{4.21}$$

In the same way we can show that

$$\frac{1}{A_T}\sum_{t=1}^{T} \varepsilon_t l_t(\theta) \ge \min_j \frac{1}{A_T}\sum_{t=1}^{T}\left[\inf_{s\in B_j} \varepsilon_t l_t(s) - E_t\left\{\inf_{s\in B_j} \varepsilon_t l_t(s)\right\}\right] - \delta. \tag{4.22}$$

Let $V_t(\cdot)$ denote conditional variance given $\mathcal{F}_{t-1}$. For every j,

$$V_t\left\{\sup_{s\in B_j}\varepsilon_t l_t(s)\right\} \le E_t\left\{\sup_{s\in B_j}\varepsilon_t l_t(s)\right\}^2 \le E_t(\varepsilon_t^2)\left\{\sup_{s\in B_j} l_t^2(s)\right\},$$

which implies, using the fact that $\sup_t E_t(\varepsilon_t^2) < \infty$ and (3.9), that with probability one,

$$\sum_{t=1}^{T} V_t\left\{\sup_{s\in B_j}\varepsilon_t l_t(s)\right\} = O\left(\sum_{t=1}^{T}\sup_{s\in B_j} l_t^2(s)\right) = O(A_T^{r_\lambda}). \tag{4.23}$$

It is known [Lai and Wei (1982)] that for any $r > 1/2$,

$$\sum_{t=1}^{T}\left[\sup_{s\in B_j}\varepsilon_t l_t(s) - E_t\left\{\sup_{s\in B_j}\varepsilon_t l_t(s)\right\}\right] = o\left(\left[\sum_{t=1}^{T} V_t\left\{\sup_{s\in B_j}\varepsilon_t l_t(s)\right\}\right]^r\right),$$

which implies, using result (4.23), that

$$\frac{1}{A_T}\sum_{t=1}^{T}\left[\sup_{s\in B_j}\varepsilon_t l_t(s) - E_t\left\{\sup_{s\in B_j}\varepsilon_t l_t(s)\right\}\right] \to 0. \tag{4.24}$$

Using a similar method we can also show that for every j,

$$\frac{1}{A_T}\sum_{t=1}^{T}\left[\inf_{s\in B_j}\varepsilon_t l_t(s) - E_t\left\{\inf_{s\in B_j}\varepsilon_t l_t(s)\right\}\right] \to 0. \tag{4.25}$$

From equations (4.24) and (4.25) we deduce that the upper and lower limits in (4.22) and (4.22) converge to $+\delta$ and -δ respectively. Since δ was arbitrary, the result follows. □

PROOF OF THEOREM 1. Let $\delta > 0$. Since Θ is bounded, then the set $B^c(\delta) := \{\theta \in \Theta : \|\theta - \theta_0\| > \delta\}$ can be covered by a finite number of balls $B(\lambda)$ (with radius ρ_λ), $\lambda \in B^c(\delta)$, such that assumption C1 holds for each one of them. Since there is a finite number of balls that cover $B^c(\delta)$, it is sufficient, in order to establish consistency of the least squares estimator $\hat{\theta}_T$, to focus on one of the balls B_λ, and to show that $\inf_{\theta\in B_\lambda}\{S_T(\theta) - S_T(\theta_0)\} \to \infty$.

The least squares estimator $\hat{\theta}_T$ minimizes $S_T(\theta)$, and therefore it can equivalently be defined as the parameter value which minimizes $S_T(\theta) - S_T(\theta_0)$. We define (for every θ in $B(\lambda)$ and also θ_0)

$$L_T(\theta) := \frac{1}{A_T}\{S_T(\theta) - S_T(\theta_0)\}$$

and

$$L_T^*(\theta) := \frac{1}{A_T}\sum_{t=1}^{T} E_t\Big[\{Y_t - f_t(\theta)\}^2 - \{Y_t - f_t(\theta_0)\}^2\Big].$$

Observe that θ_0 minimizes the function $L^*(\theta)$ and for every T and $\theta \neq \theta_0$,

$$L_T^*(\theta) - L_T^*(\theta_0) = L_T^*(\theta) = \frac{1}{A_T} \sum_{t=1}^{T} \{f_t(\theta) - f_t(\theta_0)\}^2 \geq 1.$$

In order to show that $\inf_{\theta \in B_\lambda} \{S_T(\theta) - S_T(\theta_0)\} \to \infty$, it is sufficient, since $A_T \to \infty$, to show that with probability one,

$$\sup_{\theta \in B(\lambda)} \left| L_T(\theta) - L_T^*(\theta) \right| = o(1),$$

or equivalently that with probability one,

$$\sup_{\theta \in B(\lambda)} \frac{1}{A_T} \left| \sum_{t=1}^{T} \varepsilon_t \{f_t(\theta) - f_t(\theta_0)\} \right| = o(1).$$

The result follows from Lemma 1. □

Acknowledgment. The author is grateful to the Editor for his valuable comments.

REFERENCES

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

ANDREWS, D. W. K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica* **55** 1465–1471.

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

GALLANT, A. R. and WHITE, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Blackwell, Basil.

HU, I. (1996). Strong consistency of Bayes estimates in stochastic regression models. *J. Multivariate Anal.* **57** 215–227.

HU, I. (1997). Strong consistency in stochastic regression models via posterior covariance matrices. *Biometrika* **84** 744–749.

HU, I. (1998). Strong consistency of Bayes estimates in nonlinear stochastic regression models. *J. Statist. Plann. Inf.* **67** 155–163.

LAI, T. and Wei, C. (1982). Least squares estimates in stochastic regression models with appications to identification and control systems. *Ann. Statist.* **10** 154–166.

LAI, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.* **22** 1917–1930.

WHITE, H. and DOMOWITZ, I. (1984). Nonlinear regression with dependent observations. *Econometrica* **52** 143–161.

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

DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
1-19 TORRINGTON PLACE
LONDON WC1E 6BT
UNITED KINGDOM
E-MAIL: skouras@stats.ucl.ac.uk