Statistica Sinica 1(1991), 309-334

ADAPTIVE PREDICTION IN NON-LINEAR AUTOREGRESSIVE MODELS AND CONTROL SYSTEMS

Tze Leung Lai and Guangrui Zhu

Stanford University

Abstract: In non-linear autoregressive models, minimum variance multi-step ahead predictors involve knowledge of both the system parameters and the probability distribution of the unobservable random disturbances. Herein we study adaptive versions of these optimal predictors when neither the system parameters nor the underlying error distribution are known in advance and have to be estimated from the data. Under certain assumptions, we show that the cumulative squared difference $\sum_{t=n_0}^{n}(\hat{y}_{t+d}-\tilde{y}_{t+d})^2$ between the optimal predictors $\tilde{y}_{t+d}$ and their adaptive versions $\hat{y}_{t+d}$ is of the order of $\log n$, generalizing previous results on least squares adaptive prediction in linear stochastic systems.

Key words and phrases: Linear and non-linear ARX models, adaptive prediction, strong consistency, linear and non-linear stochastic regression.

1. Introduction

A widely studied model in the time series and control systems literature is the linear ARX model (autoregressive model with exogenous inputs) defined by the linear stochastic difference equation

$$y_n = a_1 y_{n-1} + \cdots + a_p y_{n-p} + b_0 u_{n-\Delta} + \cdots + b_k u_{n-\Delta-k} + \epsilon_n, \tag{1.1}$$

where $\{y_n\}$, $\{u_n\}$ and $\{\epsilon_n\}$ denote the output, input and disturbance sequences, respectively, and $\Delta \geq 1$ represents the delay. When the input terms u_t are absent, (1.1) reduces to the classical autoregressive AR(p) model. The random disturbances ϵ_n are often assumed to be i.i.d. with mean 0 and variance σ^2. We shall also assume throughout the sequel that $E|\epsilon_1|^\alpha < \infty$ for some $\alpha > 2$ and that for every n, ϵ_{n+1} is independent of $y_n, u_n, \ldots, y_1, u_1$. Let $1 \leq d \leq \Delta$, and let $\mathcal{F}_n$ denote the σ-field generated by the current and past outputs and inputs $y_n, u_n, \ldots, y_1, u_1$. When the parameter vector $\theta = (a_1, \ldots, a_p, b_0, \ldots, b_k)'$ is known, the minimum variance d-step ahead predictor $\tilde{y}_{n+d} = E(y_{n+d}|\mathcal{F}_n)$ can be determined as follows. Define $f_1, \ldots, f_{d-1}, g_1, \ldots, g_p$ by the identity

$$(1-a_1 z - \cdots - a_p z^p)(1+f_1 z+\cdots+f_{d-1} z^{d-1}) + z^d(g_1 + g_2 z+\cdots+g_p z^{p-1}) = 1. \tag{1.2}$$

For $j = 0, \ldots, k+d-1$, let $\gamma_j = \sum_{t+s=j} f_t b_s$ $(f_0 = 1)$, so $\gamma_0 = b_0$, $\gamma_1 = b_1 + b_0 f_1$, etc. Then (1.1) can be written as $y_{n+d} = \widetilde{y}_{n+d} + \eta_{n+d}$, where

$$\widetilde{y}_{n+d} = g_1 y_n + \cdots + g_p y_{n-p+1} + \gamma_0 u_{n-\Delta+d} + \cdots + \gamma_{k+d-1} u_{n-\Delta-k+1}, \tag{1.3}$$

$$\eta_{n+d} = \epsilon_{n+d} + f_1 \epsilon_{n+d-1} + \cdots + f_{d-1} \epsilon_{n+1}, \tag{1.4}$$

cf. Åström (1970). In practice, the parameter vector θ is usually unknown, and one has to "adapt" the optimal predictor (1.3) by substituting the unknown entities in (1.3) by their estimates, leading to an adaptive predictor $\widehat{y}_{n+d}$. A review of the adaptive prediction problem in linear stochastic systems and some new unifying results have recently been given by Lai and Ying (1991).

The linear stochastic difference equation (1.1) is a special case of the general ARX model of the form

$$y_n = f_\theta(y_{n-1}, \ldots, y_{n-p}, u_{n-\Delta}, \ldots, u_{n-\Delta-k}) + \epsilon_n, \tag{1.5}$$

where θ is a ν-dimensional parameter. Even when θ is known, the minimum variance d-step ahead predictor $\widetilde{y}_{n+d}$ for non-linear ARX models is much more complicated than (1.3) for the linear case when $d > 1$. For $j = 1, \ldots, d$ $(\leq \Delta)$, define inductively

$$\begin{aligned} y_{n,1}(w;\theta) &= f_\theta(y_n, \ldots, y_{n-p+1}, u_{n-\Delta+1}, \ldots, u_{n-\Delta-k+1}) + w, \\ y_{n,j}(w_1, \ldots, w_j;\theta) &= f_\theta(y_{n,j-1}(w_1, \ldots, w_{j-1};\theta), \ldots, y_{n,j-p}(w_1, \ldots, w_{j-p};\theta), \\ &\qquad u_{n-\Delta+j}, \ldots, u_{n-\Delta-k+j}) + w_j, \quad \text{if} \quad j > p, \\ &= f_\theta(y_{n,j-1}(w_1, \ldots, w_{j-1};\theta), \ldots, y_{n,1}(w_1;\theta), y_n, \ldots, y_{n-p+j}, \\ &\qquad u_{n-\Delta+j}, \ldots, u_{n-\Delta-k+j}) + w_j, \quad \text{if} \quad j \leq p. \end{aligned} \tag{1.6}$$

Note that $y_{n+1} = y_{n,1}(\epsilon_{n+1};\theta)$, $y_{n+2} = y_{n,2}(\epsilon_{n+1}, \epsilon_{n+2};\theta)$, etc. Letting H denote the common distribution function of the i.i.d. ϵ_i and noting that $\int w_d dH(w_d) = 0$, it then follows that the minimum variance d-step ahead predictor $\widetilde{y}_{n+d}$ of y_{n+d} is given by

$$\begin{aligned} \widetilde{y}_{n+d} &= f_\theta(y_n, \ldots, y_{n-p+1}, u_{n-\Delta+1}, \ldots, u_{n-\Delta-k+1}), \quad d = 1, \\ &= \int \cdots \int y_{n,d}(w_1, \ldots, w_{d-1}, 0;\theta) dH(w_1) \cdots dH(w_{d-1}), \quad d \geq 2. \end{aligned} \tag{1.7}$$

For the linear ARX model (1.1), since $f_\theta(y_{n-1}, \ldots, y_{n-p}, u_{n-\Delta}, \ldots, u_{n-\Delta-k})$ is a linear function of $(\theta', y_{n-1}, \ldots, y_{n-p}, u_{n-\Delta}, \ldots, u_{n-\Delta-k})$, and since $\int w dH(w) = 0$, (1.7) reduces to the form (1.3) which does not involve H. However, for $d \geq 2$ in non-linear ARX models, the minimum variance d-step ahead predictor (1.7) requires knowledge of both θ and H and involves multiple integration.

In practice θ and H are usually unknown. For $d \geq 2$ in non-linear ARX models, an obvious way to "adapt" the optimal predictor (1.7) is to first replace θ by a consistent estimator $\widehat{\theta}_n$ based on the current and past observations $y_n, u_n, y_{n-1}, u_{n-1}, \ldots, y_1, u_1$, and then to replace

$$dF(w_1, \ldots, w_{d-1}) = dH(w_1) \cdots dH(w_{d-1})$$

by $d\widehat{F}_n(w_1, \ldots, w_{d-1})$, where $\widehat{F}_n$ is the empirical distribution function of

$$\{(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}) : n - d + 1 \geq i \geq m = \max(p, k + \Delta)\}$$

and $\widehat{\epsilon}_{n,i} = y_i - f_{\widehat{\theta}_n}(y_{i-1}, \ldots, y_{i-p}, u_{i-\Delta}, \ldots, u_{i-\Delta-k})$ denote the residuals. This leads to the adaptive d-step ahead predictor

$$\widehat{y}_{n+d} = (n - d - m + 2)^{-1} \sum_{i=m}^{n-d+1} y_{n,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \widehat{\theta}_n). \tag{1.8}$$

A commonly used estimator $\widehat{\theta}_n$ in the non-linear time series literature is the least squares estimate, which does not require knowledge of H for its implementation. In Section 2 we first review several basic results on the strong consistency of least squares estimates in linear and non-linear regression models in which the regressors are sequentially determined random vectors. In this connection, the results of Lai and Wei (1982a) on adaptive one-step ahead prediction based on least squares estimates in linear stochastic regression models will be extended to the case of non-linear ARX models.

Section 3 studies d-step ahead prediction in the non-linear ARX model (1.5) with $d \geq 2$. Numerical results on the performance of these adaptive predictors are presented for some non-linear time series, including first-order exponential autoregressive models for which d-step ahead prediction assuming known system parameters has recently been studied by Al-Qassam and Lane (1989). We provide a numerical comparison of the adaptive predictor (1.8) with the optimal predictor (1.7) evaluated by direct numerical integration and with some simple approximations thereof considered by Al-Qassam and Lane (1989). We also develop a partial generalization of the asymptotic theory of adaptive predictors in linear stochastic system, cf. Lai and Ying (1991), for the d-step ahead prediction problem in a non-linear ARX model with unknown parameter vector θ and error distribution H.

In Section 4 we extend the ideas underlying the construction of the adaptive predictors (1.8) to develop strongly consistent estimators of the variance and other functionals of the conditional distribution of y_{n+d} given $\mathcal{F}_n$ (i.e., the d-step ahead predictive distribution of y_{n+d}). Note that (1.7) is simply the mean of the predictive distribution.

2. Least Squares Estimates and the Associated Adaptive One-Step Ahead Predictors

The ARX model (1.5) can be written as a regression model of the form

$$y_n = f_\theta(\mathbf{x}_n) + \epsilon_n, \quad \text{where} \quad \mathbf{x}_n = (y_{n-1}, \cdots, y_{n-p}, u_{n-\Delta}, \cdots, u_{n-\Delta-k})'. \tag{2.1}$$

The regressors $\mathbf{x}_n$ in (2.1) are $\mathcal{F}_{n-1}$-measurable random vectors. In the case of a linear function $f_\theta(\mathbf{x}_n) = \theta'\mathbf{x}_n$, (2.1) reduces to the linear ARX model (1.1). More generally, if $f_\theta(\mathbf{x}_n)$ can be expressed as a linear function of θ, i.e., if there exists a vector-valued function ψ of $\mathbf{x}_n$ such that

$$f_\theta(\mathbf{x}_n) = \theta'\psi_n \qquad \text{with} \quad \psi_n = \psi(\mathbf{x}_n), \tag{2.2}$$

then (2.1) can be expressed in the form of a linear stochastic regression model

$$y_n = \theta'\psi_n + \epsilon_n \tag{2.3}$$

studied by Lai and Wei (1982a), who proved that $\widehat{\theta}_n = (\sum_1^n \psi_i\psi_i')^{-1}(\sum_1^n \psi_i y_i)$, the least squares estimate, converges to θ a.s. if

$$\lambda_{\min}\Big(\sum_1^n \psi_i\psi_i'\Big) \to \infty \text{ and } \Big\{\log \lambda_{\max}\Big(\sum_1^n \psi_i\psi_i'\Big)\Big\}\Big/\lambda_{\min}\Big(\sum_1^n \psi_i\psi_i'\Big) \to 0 \text{ a.s.,} \tag{2.4}$$

under the assumptions that $\{\epsilon_n\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$ such that $\sup_n E(|\epsilon_n|^\alpha|\mathcal{F}_{n-1}) < \infty$ a.s. for some $\alpha > 2$ and that ψ_n is $\mathcal{F}_{n-1}$-measurable for every n. Here and in the sequel we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the maximum and minimum eigenvalues of a symmetric matrix A.

As pointed out by Lai and Wei (1982a), the assumption (2.4) on the stochastic regressors ψ_i in the linear stochastic regression model (2.3) is in some sense weakest possible. In the case of the linear ARX model (1.1), since $\psi_i = \mathbf{x}_i$, (2.4) holds if

$$y_n^2 + u_n^2 = O(n^b) \text{ a.s. for some } b > 0 \text{ and } \lambda_{\min}\Big(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\Big)\Big/\log n \to \infty \text{ a.s.} \tag{2.5}$$

This is much weaker than the usual "persistent excitation" condition in the control systems literature:

$$n^{-1}\sum_1^n \mathbf{x}_i\mathbf{x}_i' \quad \text{converges a.s. to a positive definite matrix.} \tag{2.6}$$

More generally, if $f_\theta(\mathbf{x}_n) = \theta'\psi(\mathbf{x}_n)$ and $\psi_n = \psi(\mathbf{x}_n)$, then (2.4) holds if

$$\|\psi(\mathbf{x}_n)\|^2 = O(n^b) \text{ a.s. for some } b > 0 \text{ and } \lambda_{\min}\Big(\sum_1^n \psi_i\psi_i'\Big)\Big/\log n \to \infty \text{ a.s.} \tag{2.7}$$

For an example of such non-linear stochastic systems that are linear in the parameter vector, consider the open loop threshold autoregressive system

$$\begin{aligned} y_n &= a_0 + a_1 y_{n-1} + \cdots + a_p y_{n-p} + b u_{n-\Delta} + \epsilon_n \quad \text{if } u_{n-\Delta} \in S \\ &= \alpha_0 + \alpha_1 y_{n-1} + \cdots + \alpha_q y_{n-q} + \beta u_{n-\Delta} + \epsilon_n \quad \text{if } u_{n-\Delta} \notin S, \end{aligned} \tag{2.8}$$

where the inputs u_t are independent random variables that are independent of $\{\epsilon_n\}$ and S is a given interval, cf. Tong (1983). Let $\theta = (a_0, \ldots, a_p, b, \alpha_0, \ldots, \alpha_q, \beta)'$, $\mathbf{x}_n = (y_{n-1}, \cdots, y_{n-p}, u_{n-\Delta})'$ and $\psi_n = \psi(\mathbf{x}_n)$, where $\psi(\mathbf{x}_n)$ is the vector

$$\big(I_{\{u_n-\Delta\in S\}}, y_{n-1}I_{\{u_n-\Delta\in S\}}, \cdots, u_{n-\Delta}I_{\{u_n-\Delta\in S\}},\\ I_{\{u_n-\Delta\notin S\}}, y_{n-1}I_{\{u_n-\Delta\notin S\}}, \cdots, u_{n-\Delta}I_{\{u_n-\Delta\notin S\}}\big)'.$$

Here and in the sequel, we use I_E to denote the indicator function of an event E. Assume that σ^2 $(= \operatorname{Var} \epsilon_i) > 0$ and that

$$\begin{aligned} &\liminf_{n\to\infty} P\{u_n \in S\} > 0, \quad \liminf_{n\to\infty} P\{u_n \notin S\} > 0, \\ &\sup_n E|u_n|^\rho < \infty \text{ for some } \rho > 2. \end{aligned} \tag{2.9}$$

Suppose that the zeros of the characteristic polynomials $z^p - a_1 z^{p-1} - \cdots - a_p$ and $z^q - \alpha_1 z^{q-1} - \cdots - \alpha_q$ lie on or inside the unit circle. Note that since some of the roots may be on the unit circle, the system (2.8) need not be stable. By a modification of the proof of Corollary 1 of Lai and Wei (1982b), it can be shown that

$$\begin{aligned} &\sum_1^n u_i^2 = O(n) \text{ a.s.}, \quad \sum_1^n y_i^2 = O(n^b) \text{ a.s. for some } b > 1 \text{ and} \\ &\liminf_{n\to\infty} n^{-1}\lambda_{\min}\Big(\sum_1^n \psi_i\psi_i'\Big) > 0 \text{ a.s.} \end{aligned} \tag{2.10}$$

From (2.10) it follows that (2.7) and therefore (2.4) also are satisfied.

For the general ARX model (2.1) in which $f_\theta(\mathbf{x}_n)$ need not be linear in θ, the method of least squares estimates θ by $\widehat{\theta}_n$ which is the value of ϕ that minimizes $S_n(\phi) = \sum_{i\le n}[y_i - f_\phi(\mathbf{x}_i)]^2$ over a given region Θ that contains θ. Throughout the sequel we shall let $Df_\phi(\mathbf{x}) = (\partial f_\phi(\mathbf{x})/\partial\phi_1, \cdots, \partial f_\phi(\mathbf{x})/\partial\phi_\nu)'$, $D^2 f_\phi(\mathbf{x}) = (\partial^2 f_\phi(\mathbf{x})/\partial\phi_i\partial\phi_j)_{1\le i,j\le\nu}$. We shall use $Df_\theta(\mathbf{x})$ to denote the value of

$Df_\phi(\mathbf{x})$ at $\phi = \theta$. We shall also let $\|D^2 f_\phi(\mathbf{x})\| = \max_{1\le i,j\le \nu} |\partial^2 f_\phi(\mathbf{x})/\partial\phi_i\partial\phi_j|$. Suppose that

$$f_\phi(\mathbf{x}) \text{ is twice continuously differentiable in } \phi \text{ belonging to some neighborhood } U \text{ of } \theta \text{ and } \sum_1^n \sup_{\phi\in U} (\|Df_\phi(\mathbf{x}_i)\|^2 + \|D^2 f_\phi(\mathbf{x}_i)\|^2) = O(n) \text{ a.s.}, \tag{2.11}$$

$$\sup_{\|\phi-\theta\|\le\delta} n^{-1} \sum_{i=1}^n \|D^2 f_\phi(\mathbf{x}_i) - D^2 f_\theta(\mathbf{x}_i)\|^2 = O(\delta) \text{ a.s. as } n\to\infty \text{ and } \delta\to 0, \tag{2.12}$$

$$n^{-1} \sum_{i=1}^n (Df_\theta(\mathbf{x}_i))(Df_\theta(\mathbf{x}_i))' \text{ converges (as } n\to\infty) \text{ to a positive definite matrix a.s.} \tag{2.13}$$

Then we can apply the consistency theorem of Klimko and Nelson (1978) for least squares estimates in nonlinear stochastic regression models to conclude that with probability 1, there exists for sufficiently large n a solution $\phi = \theta_n$ to the equation $\partial S_n(\phi)/\partial\phi_j = 0$ $(j = 1,\dots,\nu)$ such that $\theta_n \to \theta$. Note, however, that the least squares estimate $\widehat{\theta}_n$ that attains the global minimum of $S_n(\phi)$ may be different from θ_n that gives a local minimum of $S_n(\phi)$. In the linear case $f_\theta(\mathbf{x}) = \theta'\mathbf{x}$, condition (2.13) reduces to the persistent excitation condition (2.6), condition (2.12) is trivially satisfied, while condition (2.11) reduces to $\sum_1^n \|\mathbf{x}_i\|^2 = O(n)$ a.s.

Assuming Θ to be compact and $f_\phi(\mathbf{x})$ to have continuous partial derivatives $\partial f_\phi(\mathbf{x})/\partial\phi_i$, $\partial^2 f_\phi(\mathbf{x})/\partial\phi_i\partial\phi_j$ $(i\ne j)$, $\dots, \partial^\nu f_\phi(\mathbf{x})/\partial\phi_1\cdots\partial\phi_\nu$ for every $\mathbf{x}$, Lai (1990a) recently showed that the least squares estimate $\widehat{\theta}_n$ is strongly consistent if for every $\lambda \ne \theta$ there exist $1 < p_\lambda < 2$ and an open ball $B(\lambda)$ in Θ, centered at λ, such that

$$\inf_{\phi\in B(\lambda)} \sum_{i=1}^n [f_\phi(\mathbf{x}_i) - f_\theta(\mathbf{x}_i)]^2 \to \infty \text{ a.s.}, \tag{2.14a}$$

$$\max_{\substack{1\le r\le\nu\\ 1\le j_1<\cdots<j_r\le\nu}} \sum_{i=1}^n \int_{\phi\in B(\lambda;j_1,\dots,j_r)} [\partial^r f_\phi(\mathbf{x}_i)/\partial\phi_{j_1}\cdots\partial\phi_{j_r}]^2 d\phi_{j_1}\cdots d\phi_{j_r} + \sum_1^n [f_\lambda(\mathbf{x}_i) - f_\theta(\mathbf{x}_i)]^2 = O\left(\left\{\inf_{\phi\in B(\lambda)} \sum_1^n [f_\phi(\mathbf{x}_i) - f_\theta(\mathbf{x}_i)]^2\right\}^{p_\lambda}\right) \text{ a.s.}, \tag{2.14b}$$

where $B(\lambda; j_1,\dots,j_r)$ denotes the r-dimensional sphere $\{\phi\in B(\lambda) : \phi_j = \lambda_j$ for $j \notin \{j_1,\dots,j_r\}\}$. Note that in the linear case $f_\theta(\mathbf{x}) = \theta'\mathbf{x}$, (2.14a) and (2.14b) reduce to

$$\begin{aligned}&\lambda_{\min}\Big(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\Big) \to \infty \text{ and}\\ &\sum_1^n \|\mathbf{x}_i\|^2 = O\Big(\Big\{\lambda_{\min}\Big(\sum_1^n \mathbf{x}_i\mathbf{x}_i'\Big)\Big\}^p\Big) \text{ a.s. for some } 1 < p < 2,\end{aligned} \tag{2.15}$$

which is much weaker than the persistent excitation condition (2.6) but stronger than the condition (2.4).

For the ARX model (2.1), the minimum variance 1-step ahead predictor of y_{t+1} is $\widetilde{y}_{t+1} = f_\theta(\mathbf{x}_{t+1})$ when θ is known, and the corresponding least squares predictor when θ is unknown is $\widehat{y}_{t+1} = f_{\widehat{\theta}_t}(\mathbf{x}_{t+1})$. If one carries out the adaptive prediction procedure after an initial learning period n_0, the overall squared error of the predicted values up to stage n ($> n_0$) is $\sum_{t=n_0}^{n-1}(\widehat{y}_{t+1} - y_{t+1})^2$. Since $y_{t+1} = \widetilde{y}_{t+1} + \epsilon_{t+1}$ by (2.1) with $E\epsilon_{t+1}^2 = \sigma^2$, it follows that $E\{\sum_{t=n_0}^{n-1}(\widehat{y}_{t+1} - y_{t+1})^2\} = E(R_n) + (n - n_0)\sigma^2$, where

$$R_n = \sum_{t=n_0}^{n-1}(\widehat{y}_{t+1} - \widetilde{y}_{t+1})^2 \tag{2.16}$$

is the cumulative squared difference between the optimal predictor $\widetilde{y}_{t+1}$ and the adaptive predictor $\widehat{y}_{t+1}$. We shall call R_n the "regret" of $\{\widehat{y}_{t+1}, n_0 \le t < n\}$. If $f_\theta(\mathbf{x}_n) = \theta'\psi(\mathbf{x}_n)$, then (2.1) can be expressed as a linear stochastic regression model (2.3) with $\psi_n = \psi(\mathbf{x}_n)$, and in this case,

$$\begin{aligned}&R_n = \sum_{i=n_0}^{n-1}[(\widehat{\theta}_i - \theta)'\psi_{i+1}]^2 = O\Big(\log\lambda_{\max}\Big(\sum_1^n \psi_i\psi_i'\Big)\Big) \text{ a.s.}\\ &\text{on } \Big\{\limsup_{n\to\infty} \psi_n\Big(\sum_1^n \psi_i\psi_i'\Big)^{-1}\psi_n < 1\Big\},\end{aligned} \tag{2.17}$$

cf. Lai and Wei (1982a), Lai and Ying (1991). Therefore if (2.13) also holds, then $R_n = O(\log n)$ a.s. The following theorem extends this logarithmic order of the regret to adaptive 1-step ahead predictors in the general ARX model (2.1) under the assumptions (2.11)-(2.14).

Theorem 1. *Suppose that in the general* ARX *model* (2.1) *the parameter space* Θ *is compact with* θ *belonging to its interior and that* $f_\phi(\mathbf{x})$ *satisfies conditions* (2.11)-(2.14). *Define* R_n *by* (2.16), *in which* $\widetilde{y}_{t+1} = f_\theta(\mathbf{x}_{t+1})$ *and* $\widehat{y}_{t+1} = f_{\widehat{\theta}_t}(\mathbf{x}_{t+1})$, *where* $\widehat{\theta}_t$ *is the least squares estimate* (*of* θ) *that minimizes* $S_t(\phi) = \sum_{i\le t}[y_i - f_\phi(\mathbf{x}_i)]^2$ *over* Θ. *Then* $R_n \sim \sigma^2\nu\log n$ a.s.

Proof. In view of (2.14), $\widehat{\theta}_n \to \theta$ a.s. by Theorem 1 of Lai (1990a). By Taylor's expansion about the true parameter value θ,

$$\begin{aligned}0 = -DS_n(\widehat{\theta}_n)/2 &= \sum_{i=1}^n \epsilon_i Df_{\hat{\theta}_n}(\mathbf{x}_i) + \sum_{i=1}^n Df_{\hat{\theta}_n}(\mathbf{x}_i)\big(f_\theta(\mathbf{x}_i) - f_{\hat{\theta}_n}(\mathbf{x}_i)\big) \\ &= \sum_{i=1}^n \epsilon_i Df_\theta(\mathbf{x}_i) + \Big\{ \sum_{i=1}^n \epsilon_i \big[D^2 f_{\theta_n}(\mathbf{x}_i) - D^2 f_\theta(\mathbf{x}_i)\big] + \sum_{i=1}^n \epsilon_i D^2 f_\theta(\mathbf{x}_i) \\ &\quad - \sum_{i=1}^n \big(Df_\theta(\mathbf{x}_i)\big)\big(Df_\theta(\mathbf{x}_i)\big)' - \sum_{i=1}^n D^2 f_{\theta_n}(\mathbf{x}_i)(\widehat{\theta}_n - \theta)\big(Df_\theta(\mathbf{x}_i)\big)' \\ &\quad - \sum_{i=1}^n Df_{\hat{\theta}_n}(\mathbf{x}_i)(\theta_n^{**} - \theta)' D^2 f_{\theta_n^*}(\mathbf{x}_i)\Big\}(\widehat{\theta}_n - \theta), \qquad (2.18)\end{aligned}$$

where θ_n, θ_n^*, θ_n^{**} lie between θ and $\widehat{\theta}_n$. By (2.11) and Lemma 2(iii) of Lai and Wei (1982a),

$$\Big\| \sum_{i=1}^n \epsilon_i Df_\theta(\mathbf{x}_i) \Big\| + \Big\| \sum_{i=1}^n \epsilon_i D^2 f_\theta(\mathbf{x}_i) \Big\| = o(n^{\frac{1}{2}+\delta}) \text{ a.s.} \qquad (2.19)$$

for every $\delta > 0$. By the Schwarz inequality,

$$\begin{aligned}&\Big\| \sum_{i=1}^n \epsilon_i [D^2 f_{\theta_n}(\mathbf{x}_i) - D^2 f_\theta(\mathbf{x}_i)] \Big\| \\ &\le \Big(\sum_{i=1}^n \epsilon_i^2\Big)^{1/2} \Big(\sum_{i=1}^n \|D^2 f_{\theta_n}(\mathbf{x}_i) - D^2 f_\theta(\mathbf{x}_i)\|^2\Big)^{1/2} = o(n) \text{ a.s.}, \qquad (2.20)\end{aligned}$$

in view of (2.12) and the strong law of large numbers since $\theta_n \to \theta$ a.s. Moreover, by the Schwarz inequality and (2.11),

$$\Big\| \sum_{i=1}^n D^2 f_{\theta_n}(\mathbf{x}_i)(\widehat{\theta}_n - \theta)\big(Df_\theta(\mathbf{x}_i)\big)' \Big\| = O(n\|\widehat{\theta}_n - \theta\|) = o(n) \text{ a.s.}, \qquad (2.21)$$

$$\Big\| \sum_{i=1}^n Df_{\hat{\theta}_n}(\mathbf{x}_i)(\theta_n^{**} - \theta)' D^2 f_{\theta_n^*}(\mathbf{x}_i) \Big\| = O(n\|\theta_n^{**} - \theta\|) = o(n) \text{ a.s.} \qquad (2.22)$$

Combining (2.18)–(2.22) with (2.13) yields

$$\|\theta_n - \theta\| \le \|\widehat{\theta}_n - \theta\| = O\Big(n^{-1} \Big\| \sum_{i=1}^n \epsilon_i Df_\theta(\mathbf{x}_i) \Big\|\Big) = O(n^{-1/2+\delta}) \text{ a.s.} \qquad (2.23)$$

for every $\delta > 0$. Putting (2.23) in (2.20) and (2.12) gives

$$\Big\| \sum_{i=1}^n \epsilon_i [D^2 f_{\theta_n}(\mathbf{x}_i) - D^2 f_\theta(\mathbf{x}_i)] \Big\| = O(n^{3/4+\delta}) \text{ a.s.} \qquad (2.24)$$

for every $\delta > 0$. Moreover, putting (2.23) in (2.21) and (2.22) gives

$$\Big\| \sum_{i=1}^{n} D^2 f_{\theta_n}(\mathbf{x}_i)(\widehat{\theta}_n - \theta)\big(Df_\theta(\mathbf{x}_i)\big)' \Big\| + \Big\| \sum_{i=1}^{n} Df_{\hat{\theta}_n}(\mathbf{x}_i)(\theta_n^{**} - \theta)' D^2 f_{\theta_n^*}(\mathbf{x}_i) \Big\| \\ = o(n^{1/2+\delta}) \text{ a.s.} \tag{2.25}$$

for every $\delta > 0$. From (2.18), (2.19), (2.24) and (2.25), it follows that with probability 1,

$$\begin{aligned} \widehat{\theta}_n - \theta &= \Big\{ \sum_{i=1}^{n} \big(Df_\theta(\mathbf{x}_i)\big)\big(Df_\theta(\mathbf{x}_i)\big)' + O(n^{3/4+\delta}) \Big\}^{-1} \sum_{i=1}^{n} \epsilon_i Df_\theta(\mathbf{x}_i) \\ &= \Big\{ \sum_{i=1}^{n} \big(Df_\theta(\mathbf{x}_i)\big)\big(Df_\theta(\mathbf{x}_i)\big)' \Big\}^{-1} \sum_{i=1}^{n} \epsilon_i Df_\theta(\mathbf{x}_i) + O(n^{-3/4+2\delta}) \end{aligned} \tag{2.26}$$

for $1/8 > \delta > 0$. Let $\mathbf{z}_i = Df_\theta(\mathbf{x}_i)$ and $\widetilde{\theta}_n = \theta + (\sum_1^n \mathbf{z}_i\mathbf{z}_i')^{-1} \sum_1^n \epsilon_i \mathbf{z}_i$. By (2.13), $\|\mathbf{z}_n\|^2/\lambda_{\min}(\sum_1^n \mathbf{z}_i\mathbf{z}_i') \to 0$ a.s., and Theorem 3 of Wei (1987) is applicable to show that

$$\sum_{i=n_0}^{n-1} [(\widetilde{\theta}_i - \theta)'\mathbf{z}_{i+1}]^2 \sim \sigma^2 \log\det\Big(\sum_1^n \mathbf{z}_i\mathbf{z}_i'\Big) \sim \sigma^2 \nu \log n \quad \text{a.s.} \tag{2.27}$$

Since $\sum_1^\infty n^{-3/2+4\delta} < \infty$ for $\delta < 1/8$, combining (2.26) with (2.27) yields

$$\sum_{i=n_0}^{n-1} [(\widehat{\theta}_i - \theta)'\mathbf{z}_{i+1}]^2 \sim \sigma^2 \nu \log n \quad \text{a.s.} \tag{2.28}$$

Moreover, since $\widehat{\theta}_t \to \theta$ a.s.,

$$f_{\hat{\theta}_t}(\mathbf{x}_{t+1}) - f_\theta(\mathbf{x}_{t+1}) = (\widehat{\theta}_t - \theta)' Df_\theta(\mathbf{x}_{t+1}) + O(\|\widehat{\theta}_t - \theta\|^2 \sup_{\phi\in U} \|D^2 f_\phi(\mathbf{x}_{t+1})\|) \text{ a.s.} \tag{2.29}$$

By taking $\delta < 1/4$ in (2.23), it turns out that, with probability 1,

$$\begin{aligned} \sum_{t=1}^{\infty} \|\widehat{\theta}_t - \theta\|^4 \sup_{\phi\in U} \|D^2 f_\phi(\mathbf{x}_{t+1})\|^2 &= \sum_{i=0}^{\infty} \sum_{t=2^i}^{2^{i+1}-1} O(t^{-2+4\delta}) \sup_{\phi\in U} \|D^2 f_\phi(\mathbf{x}_{t+1})\|^2 \\ &\le \sum_{i=0}^{\infty} O(2^{-(2-4\delta)i}) \sum_{t=1}^{2^{i+1}} \sup_{\phi\in U} \|D^2 f_\phi(\mathbf{x}_t)\|^2 = \sum_{i=0}^{\infty} O(2^{-(1-4\delta)i}) < \infty, \end{aligned} \tag{2.30}$$

by (2.11). From (2.28), (2.29) and (2.30), it follows that

$$R_n = \sum_{t=n_0}^{n-1} [f_{\hat{\theta}_t}(\mathbf{x}_{t+1}) - f_\theta(\mathbf{x}_{t+1})]^2 \sim \sigma^2 \nu \log n \quad \text{a.s.}$$

3. Adaptive d-Step Ahead Prediction in Non-Linear ARX Models

In this section we shall let $d \geq 2$. To begin with, consider the minimum variance d-step ahead predictor (1.7) when both θ and H are known. Direct numerical integration to evaluate (1.7) is often quite complicated. Instead of direct numerical integration, one can also evaluate (1.7) by Monte Carlo (MC) methods, generating i.i.d. random variables $\epsilon_1^*, \ldots, \epsilon_N^*$ having distribution H and approximating (1.7) by

$$y_{n,d}^{(N)} = (N-d+2)^{-1} \sum_{i=0}^{N-d+1} y_{n,d}(\epsilon_{i+1}^*, \ldots, \epsilon_{i+d-1}^*, 0; \theta). \tag{3.1}$$

Note that as $N \to \infty$, $y_{n,d}^{(N)} \to \int \cdots \int y_{n,d}(w_1, \ldots, w_{d-1}, 0; \theta) dH(w_1) \cdots H(w_{d-1})$ a.s. by the strong law of large numbers for d-dependent random variables.

In the case of a first-order exponential autoregressive model

$$y_i = f_\theta(y_{i-1}) + \epsilon_i, \quad \text{where } \theta = (a,b)' \text{ and } f_\theta(x) = (a + be^{-x^2})x, \tag{3.2}$$

with i.i.d. normal random disturbances ϵ_i, Al-Qasssam and Lane (1989) developed an approximation to (1.7) based on the assumption of approximately normal forecast errors (NFE). In Tables 1 and 2 we compare the values of (1.7) obtained by direct numerical integration (NUMI) and by the NFE approximation for the exponential autoregressive model (3.2), with $a = -0.3$, $b = -0.8$ and normal $N(0, \sigma^2)$ errors ϵ_i reported by Al-Qassam and Lane (1989), with the values given by the Monte Carlo method (3.1) based on 1000 normal $N(0, \sigma^2)$ random variables ϵ_i^*, for the problem of predicting y_{n+d} given that $y_n = 0.555$. The results indicate good agreement between NUMI, NFE and the Monte Carlo algorithm (3.1). The NFE approximation, however, may not be appropriate when the ϵ_i are not normal. For example, for the case $d = 2$ in Table 2, if $\epsilon_i + 1$ is exponentially distributed with density e^{-t} $(t > 0)$, so that ϵ_i still has mean 0 and variance 1, then numerical integration of (1.7) gives $\tilde{y}_{n+2} = 0.2930$, while the NFE predictor still remains 0.2178.

Now suppose that $\epsilon_1^*, \ldots, \epsilon_N^*$ are not directly observable and that one observes instead $y_1^*, \ldots, y_N^*$ with $y_i^* = f_\theta(y_{i-1}^*) + \epsilon_i^*$. Suppose also that $\theta = (a,b)'$ is unknown. It is natural to first estimate θ by least squares (i.e., let $\widehat{\theta}_N$ minimize $S_N(\theta) = \sum_1^{N-1} \{y_{i+1}^* - f_\theta(y_i^*)\}^2$), and then to replace the unobservable ϵ_i^* by

$\widehat{\epsilon}_{N,i} = y_i^* - f_{\widehat{\theta}_N}(y_{i-1}^*)$ for $i = 2,\dots,N$. This leads to the adaptive Monte Carlo (AMC) predictor

$$\widehat{y}_{n,d}^{(N)} = (N-d+1)^{-1}\sum_{i=1}^{N-d+1} y_{n,d}(\widehat{\epsilon}_{N,i+1},\dots,\widehat{\epsilon}_{N,i+d-1},0;\widehat{\theta}_N). \tag{3.3}$$

In Tables 1 and 2 we also give the values of $\widehat{y}_{n,d}^{(N)}$ that are based on the same 1000 normal $N(0,\sigma^2)$ random variables as in $y_{n,d}^{(N)}$, and these values are adequate approximations to those obtained by direct numerical integration.

Table 1. Values of d-step ahead predictors calculated by four different methods for normal disturbances with $\sigma^2 = 0.01$

d	NUMI: (1.7)	NFE	MC: (3.1) $N = 1000$	AMC: (3.3) $N = 1000$
2	0.4495	0.4495	0.4479	0.4395
3	−0.4183	−0.4183	−0.4185	−0.4078
4	0.3944	0.3944	0.3923	0.3796
5	−0.3751	−0.3751	−0.3747	−0.3602
6	0.3587	0.3589	0.3560	0.3399
7	−0.3443	−0.3448	−0.3436	−0.3258
8	0.3314	0.3222	0.3289	0.3096
9	−0.3196	−0.3207	−0.3190	−0.2983
10	0.3086	0.3101	0.3063	0.2842
15	−0.2612	−0.2640	−0.2617	−0.2335
20	0.2222	0.2226	0.2208	0.1879
30	0.1610	0.1445	0.1565	0.1221
40	0.1166	0.0816	0.1047	0.0754
50	0.0845	0.0423	0.0672	0.0422

Table 2. Values of d-step ahead predictors calculated by four different methods for normal disturbances with $\sigma^2 = 1$

d	NUMI: (1.7)	NFE	MC: (3.1) $N = 1000$	AMC: (3.3) $N = 1000$
2	0.2178	0.2178	0.2284	0.2388
3	−0.0950	−0.0925	−0.0862	−0.0928
4	0.0414	0.0390	0.0472	0.0510
5	−0.0180	−0.0164	−0.0083	−0.0101

The preceding considerations suggest the adaptive predictor (1.8) of y_{n+d}

based on $y_1, u_1, \cdots, y_n, u_n$, in which we let $(\epsilon_1, \cdots, \epsilon_n)$ play the role of $(\epsilon_1^*, \cdots, \epsilon_N^*)$ in the adaptive Monte Carlo predictor (3.3) (with $N = n$) and use the method of least squares to estimate θ. The following theorem, whose proof will be given at the end of this section, shows that under the assumptions (2.12)–(2.14) and a stronger form of (2.11), we can in fact extend the logarithmic order of the regret in Theorem 1 on adaptive 1-step ahead prediction to the adaptive d-step ahead predictors (1.8). As in (2.1), we let $\mathbf{x}_n = (y_{n-1}, \cdots, y_{n-p}, u_{n-\Delta}, \cdots, u_{n-\Delta-k})'$. Let $h = \min(p, d-1)$ and assume that for some compact neighborhood U of θ,

$f_\phi(\mathbf{x})$ is twice continuously differentiable in ϕ belonging to U and
$$\sum_{i=1}^{n} \sup_{\phi \in U} (\|Df_\phi(\mathbf{x}_i)\|^2 + \|D^2 f_\phi(\mathbf{x}_i)\|^2) = O(n) \text{ a.s.}, \tag{3.4a}$$

$\partial^2 f_\phi / \partial x_i \partial \phi_j$ exists and is bounded for $(\phi, \mathbf{x}) \in U \times \mathbf{R}^{p+k+1}$, $1 \le i \le h$, $1 \le j \le \nu$, (3.4b)

$\partial^r f_\phi / \partial x_1^{i_1} \cdots \partial x_h^{i_h}$ exists and is bounded for $(\phi, \mathbf{x}) \in U \times \mathbf{R}^{p+k+1}$, for every $1 \le r \le d-1$ and all h-tuples $(i_1, \ldots, i_h)$ with $i_1 + \cdots + i_h = r$, $0 \le i_j \le r$.(3.4c)

Note that (3.4b) and (3.4c) are clearly satisfied by the function $f_{a,b}(x) = (a + be^{-x^2})x$ in the definition (3.2) of the exponential autoregressive model; moreover (3.4a) holds if $|a| < 1$.

Theorem 2. *Suppose that in the general* ARX *model* (2.1) *the parameter space* Θ *is compact with* θ *belonging to its interior and that* $f_\phi(\mathbf{x})$ *satisfies conditions* (3.4), (2.12) *and* (2.13). *Assume also that for every* $\lambda \ne \theta$ *there exist* $1 < p_\lambda < 2$ *and an open ball* $B(\lambda)$ *centered at* λ *such that* (2.14a) *and* (2.14b) *are satisfied. Let* $m = \max(p, k+\Delta)$. *For* $n > m$, *let* $\widehat{\theta}_n$ *be the least squares estimate of* θ *and let* $\widehat{\epsilon}_{n,i} = y_i - f_{\widehat{\theta}_n}(\mathbf{x}_i)$ *for* $m < i \le n$. *For* $\Delta \ge d \ge 2$ *define the adaptive d-step ahead predictor* $\widehat{y}_{n+d}$ *of* y_{n+d} *by* (1.8), *in which* $y_{n,d}$ *is defined by* (1.6). *Let* $\widetilde{y}_{n+d}$ *denote the minimum variance predictor of* y_{n+d} *given by* (1.7), *assuming knowledge of* θ *and the common distribution function* H *(with* $\int_{-\infty}^{\infty} x dH(x) = 0$ *and* $\int_{-\infty}^{\infty} |x|^\alpha dH(x) < \infty$ *for some* $\alpha > 2$*) of the* ϵ_i.

(i) *Define the regret* $R_n = \sum_{t=n_0}^{n-d} (\widehat{y}_{t+d} - \widetilde{y}_{t+d})^2$, *in which* n_0 *denotes the stage at which adaptive prediction begins after an initial learning period. Then*

$$R_n = O(\log n) \quad \text{a.s.} \tag{3.5}$$

(ii) *Suppose furthermore that* $\sup_{\phi \in U} \|Df_\phi(\mathbf{x}_n)\| = O(1)$ a.s. *and that* H

has bounded support. Then

$$\sup_{t\geq 1}\left|(n-d-m+2)^{-1}\sum_{i=m}^{n-d+1} y_{t,d}(\widehat{\epsilon}_{n,i+1},\dots,\widehat{\epsilon}_{n,i+d-1},0;\widehat{\theta}_n)-\widetilde{y}_{t+d}\right|$$
$$=O\big((n^{-1}\log\log n)^{1/2}\big)\ \text{ a.s.} \tag{3.6}$$

For the exponential model (3.2) in which $|a|<1$ and the ϵ_i are i.i.d. random variables with a common absolutely continuous distribution function having mean 0 and finite absolute moment of some order >2, (3.4) and (2.12) are satisfied. Moreover, since $f_{a,b}(x)=ax+bxe^{-x^2}$ is linear in a,b, (2.13) and (2.14) follow from the fact that

$$n^{-1}\sum_{1}^{n}(y_i,y_ie^{-y_i^2})'(y_i,y_ie^{-y_i^2})\ \text{ converges a.s. to a positive definite matrix,}$$

which in turn can be proved by a standard argument using the ergodicity of $\{y_n\}$, cf. Tong (1990). Hence the assumptions of Theorem 2(i) are satisfied in this model.

The result (3.6) in Theorem 2(ii) is related to the following modification of the adaptive predictors (1.8) to facilitate computation when we perform d-step ahead prediction sequentially over time. Instead of updating the least squares estimate at every stage t and using a new set of residuals $\widehat{\epsilon}_{t,i}=y_i-f_{\widehat{\theta}_t}(\mathbf{x}_i)$ for every t, as in the procedure (1.8), we update the estimates and the corresponding set of residuals only at stages $t=n_k$ and predict y_{t+d} with the adaptive predictor

$$\widehat{y}_{t+d}=(n_k-d-m+2)^{-1}\sum_{i=m}^{n_k-d+1} y_{t,d}(\widehat{\epsilon}_{n_k,i+1},\dots,\widehat{\epsilon}_{n_k,i+d-1},0;\widehat{\theta}_{n_k}),\quad n_k\leq t<n_{k+1}. \tag{3.7}$$

In particular, if $n_k\sim n_0c^k$ for some integer $c>1$, then (3.6) implies that the adaptive predictors (3.7) have regret

$$R_N=\sum_{t=n_0}^{N-d}(\widehat{y}_{t+d}-\widetilde{y}_{t+d})^2=\sum_{k:n_k\leq N-d}O\big((n_k-n_{k-1})n_{k-1}^{-1}\log\log n_{k-1}\big)$$
$$=O\big((\log N)(\log\log N)\big)\ \text{ a.s.} \tag{3.8}$$

We next report a simulation study of the performance of the d-step ahead adaptive predictors (3.7) in the case of a sinusoidal autoregressive model

$$y_n=5\sin(\theta y_{n-1})+\epsilon_n \tag{3.9}$$

with initial state $y_0 = 2$. Suppose that θ is known to belong to the interval $\Theta = [0,2]$. In particular, suppose that $\theta = 1$ and the common distribution H of the ϵ_i is uniform on $(-\pi,\pi)$. Note that the assumptions of Theorem 2(ii) are satisfied in this case. The minimum variance d-step ahead predictor (1.7) assuming knowledge of θ and H is given by

$$\begin{aligned} \widetilde{y}_{n+d} &= 5\sin\theta y_n, && \text{if } d = 1, \\ &= 0, && \text{if } d \geq 2, \end{aligned} \tag{3.10}$$

since $\int_{-\pi}^{\pi} \sin(x+w)dw = 0$ for all x. Without assuming θ and H to be known, consider the adaptive predictors (3.7) with $m = 1$, $n_0 = 100$ and $n_k = 3n_{k-1}$ $(k \geq 1)$. Table 3 gives the mean values, over 100 simulation runs, of the regret R_N (defined in (3.8)) for $N = 300$. Since $n_1 = 3n_0 = 300$, the rule (3.7) only uses the initial least squares estimate and the initial set of residuals based on $n_0 = 100$ observations for predicting y_{t+d} when $t < N = 300$. Also given are the mean values (over the 100 simulations) of the total squared prediction error $V_N = \sum_{t=n_0}^{N-d}(\widehat{y}_{t+d} - y_{t+d})^2$ for the adaptive rule (3.7) and of the total squared prediction error $\sum_{t=n_0}^{N-d}(\widetilde{y}_{t+d} - y_{t+d})^2$ for the optimal rule (3.10) that assumes knowledge of θ and H. When θ is assumed known, Al-Qassam and Lane (1989) also consider the following simple procedure, which they call the "extrapolation method" (EXM), for d-step ahead prediction. The EXM method discards the unobservable random disturbances in forming the d-step ahead predictor

$$y^*_{n+d} = y_{n,d}(0,\dots,0;\theta). \tag{3.11}$$

In Table 3 we also give the mean values (over 100 simulations) of the regret $\sum_{t=n_0}^{N-d}(y^*_{t+d} - \widetilde{y}_{t+d})^2$ and of the squared prediction error $\sum_{t=n_0}^{N-d}(y^*_{t+d} - y_{t+d})^2$ for the EXM method.

The results in Table 3 are consistent with the logarithmic order (3.8) for the regret of the adaptive rule (3.7). They also show the considerable price to be paid by "switching off" the noise for d-step ahead prediction in the EXM method when $d \geq 2$. Note also the good agreement of the mean values of these 100 simulations with the identity $E(R_N) + E\{\sum_{t=n_0}^{N-d}(\widetilde{y}_{t+d} - y_{t+d})^2\} = E(V_N)$ for the adaptive rule (3.7) and also for the EXM rule (3.11).

Table 3. Expected regret and total squared prediction error of d-step ahead predictors

	ER_N (Regret)		EV_N (Total squared prediction error)		
d	Adaptive rule	EXM rule	Optimal rule	Adaptive rule	EXM rule
1	5.95	0	661.2	668.1	661.2
2	8.98	3114.9	3145.2	3154.1	6321.3
3	8.53	3522.8	3131.2	3139.6	6678.3

We now give the proof of Theorem 2, which is prefaced by the following three lemmas.

Lemma 1. *Let* $\{\epsilon_n\}$ *be a martingale difference sequence with respect to an increasing sequence of* σ*-fileds* $\{\mathcal{F}_n\}$ *such that* $\sup_n E(|\epsilon_n|^\alpha|\mathcal{F}_{n-1}) < \infty$ a.s. *for some* $\alpha > 2$. *Let* $\mathbf{z}_n$ *be an* $\mathcal{F}_{n-1}$*-measurable* $\nu \times 1$ *vector for every* n *such that* $n^{-1}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i'$ *converges* a.s. *to a positive definite matrix. Then* $n^{-1/2}\|\mathbf{z}_n\| \to 0$ a.s., *and for every fixed* $r = 1, 2, \ldots,$

$$\sum_{t=1}^n \Big[\mathbf{z}_{t+r}'\Big(\sum_{i=1}^t \mathbf{z}_i\mathbf{z}_i'\Big)^{-1}\Big(\sum_{i=1}^t \epsilon_i\mathbf{z}_i\Big)\Big]^2 = O(\log n) \text{ a.s.,} \tag{3.12}$$

$$\sum_{t=1}^n \Big\|\Big(\sum_{i=1}^t \mathbf{z}_i\mathbf{z}_i'\Big)^{-1}\Big(\sum_{i=1}^t \epsilon_i\mathbf{z}_i\Big)\Big\|^2 = O(\log n) \quad \text{a.s.,} \tag{3.13}$$

$$\Big\|\sum_{t=1}^n \epsilon_t\mathbf{z}_t\Big\| = O\big((n\log\log n)^{1/2}\big) \quad \text{a.s.} \tag{3.14}$$

Proof. For every $0 < \delta < 1$, since $n^{-1}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i'$ converges a.s. to a positive definite matrix $\mathbf{A}$,

$$\|\mathbf{z}_n\|^2 \le \text{tr}\Big(\sum_{n\ge i\ge(1-\delta)n} \mathbf{z}_i\mathbf{z}_i'\Big) \le (\delta + o(1))n\text{tr}(\mathbf{A}) \text{ a.s.}$$

Since δ can be arbitrarily small, $\|\mathbf{z}_n\|^2 = o(n)$ a.s. From Corollary 1.1 of Stout (1973) and a standard truncation argument, (3.14) follows.

Let $\mathbf{A}_t = \sum_{i=1}^t \mathbf{z}_i\mathbf{z}_i'$ and note that $\mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1} = O(\|\mathbf{z}_{t+1}\|^2/t) \to 0$ a.s. By Corollary 1 of Lai and Wei (1982a), (3.12) holds for $r = 1$. Since $\mathbf{A}_{t+1}^{-1} = \mathbf{A}_t^{-1} - \mathbf{A}_t^{-1}\mathbf{z}_{t+1}\mathbf{z}_{t+1}'\mathbf{A}_t^{-1}/(1 + \mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1})$, cf. (1.4b) of Lai and Wei (1982a), we have

$$\begin{aligned}\mathbf{z}_{t+2}'\mathbf{A}_{t+1}^{-1}\Big(\sum_{i=1}^{t+1}\epsilon_i\mathbf{z}_i\Big) &= \mathbf{z}_{t+2}'\mathbf{A}_t^{-1}\Big(\sum_{i=1}^{t}\epsilon_i\mathbf{z}_i\Big) + \mathbf{z}_{t+2}'\mathbf{A}_t^{-1/2}\mathbf{A}_t^{-1/2}\mathbf{z}_{t+1}\epsilon_{t+1}\\ &- (1+\mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1})^{-1}(\mathbf{z}_{t+2}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1})\Big[\mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\Big(\sum_{i=1}^{t}\epsilon_i\mathbf{z}_i\Big) + \mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1}\epsilon_{t+1}\Big].\end{aligned} \tag{3.15}$$

Since $\sum_{t=1}^n[\mathbf{z}_{t+2}'\mathbf{A}_{t+1}^{-1}(\sum_{i=1}^{t+1}\epsilon_i\mathbf{z}_i)]^2 = O(\log n)$ a.s. and since $\mathbf{z}_{t+2}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1} \to 0$ a.s., it follows from (3.15) that

$$\begin{aligned}\sum_{t=1}^n\Big[\mathbf{z}_{t+2}'\mathbf{A}_t^{-1}\Big(\sum_{i=1}^t\epsilon_i\mathbf{z}_i\Big)\Big]^2 &= O(\log n) + O\Big(\sum_{t=1}^n\|\mathbf{A}_t^{-1/2}\mathbf{z}_{t+2}\|^2(\mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1}\epsilon_{t+1}^2)\Big)\\ &+ O\Big(\sum_{t=1}^n(\mathbf{z}_{t+1}'\mathbf{A}_t^{-1}\mathbf{z}_{t+1})^2\epsilon_{t+1}^2\Big) \quad \text{a.s.}\end{aligned} \tag{3.16}$$

By Lemma 2(ii)-(iii) of Lai and Wei (1982a),

$$\sum_{t=1}^{n} \mathbf{z}'_{t+1}\mathbf{A}_t^{-1}\mathbf{z}_{t+1}\epsilon_{t+1}^2 = O\Big(\sum_{t=1}^{n}(\mathbf{z}'_{t+1}\mathbf{A}_t^{-1}\mathbf{z}_{t+1})\Big) = O(\log n) \text{ a.s.} \tag{3.17}$$

Noting that $\|\mathbf{A}_t^{-1/2}\mathbf{z}_{t+2}\|^2 + \mathbf{z}'_{t+1}\mathbf{A}_t^{-1}\mathbf{z}_{t+1} \to 0$ a.s., we obtain from (3.16) and (3.17) that (3.12) holds for $r = 2$. Proceeding inductively in this way, we can then establish (3.12) for $r = 3, 4, \ldots$.

Since $0 < \lim \lambda_{\min}(n^{-1}\mathbf{A}_n) \le \lim \lambda_{\max}(n^{-1}\mathbf{A}_n) < \infty$ a.s., it suffices for the proof of (3.13) to show that for fixed $j = 1, \ldots, \nu$,

$$\sum_{t=1}^{n}\Big[\Big(\sum_{i=1}^{t}\epsilon_i z_{i,j}\Big)^2/t^2\Big] = O(\log n) \text{ a.s.,} \tag{3.18}$$

where the $z_{i,j}$ are the components of $\mathbf{z}_i$. Let $\tilde{z}_{i,j} = z_{i,j}I_{\{|z_{i,j}|\ge 1\}} + I_{\{|z_{i,j}|<1\}}$. Then $|\tilde{z}_{i,j}| = \max(|z_{i,j}|, 1)$, so $t \le \sum_{i=1}^{t}\tilde{z}_{i,j}^2 = O(t)$ a.s. By Corollary 1 of Lai and Wei (1982a),

$$\sum_{t=1}^{n}\Big[\Big(\sum_{i=1}^{t}\epsilon_i\tilde{z}_{i,j}\Big)^2\Big/t^2\Big] \le \sum_{t=1}^{n}\Big[\tilde{z}_{t+1,j}^2\Big(\sum_{i=1}^{t}\epsilon_i\tilde{z}_{i,j}\Big)^2/t^2\Big] = O(\log n) \text{ a.s.} \tag{3.19}$$

Let $z_{i,j}^* = z_{i,j} - \tilde{z}_{i,j} = (z_{i,j}-1)I_{\{|z_{i,j}|<1\}}$. Then $|z_{i,j}^*| < 2$. Applying Corollary 1 of Lai and Wei (1982a) to the martingale difference sequence $\{\epsilon_i z_{i,j}^*\}$ and regressors $x_i \equiv 1$ yields

$$\sum_{t=1}^{n}\Big[\Big(\sum_{i=1}^{t}\epsilon_i z_{i,j}^*\Big)^2/t^2\Big] = O(\log n) \text{ a.s.} \tag{3.20}$$

From (3.19) and (3.20), (3.18) follows.

Lemma 2. *With the same notation and assumptions as in Theorem* 2(i), *for every* $1 \le r \le d-1$ *and all r-tuples* $(j_1, \ldots, j_r)$ *of positive integers with* $j_1 < \cdots < j_r \le d-1$,

$$\sup_{t\ge 1}\Big\{\sup_{\substack{\phi\in U\\(w_1,\ldots,w_{d-1})\in\mathbf{R}^{d-1}}}\Big|\frac{\partial^r}{\partial w_{j_1}\cdots\partial w_{j_r}}y_{t,d}(w_1,\ldots,w_{d-1},0;\phi)\Big|\Big\} < \infty \text{ a.s.} \tag{3.21}$$

Moreover, there exist (*scalar*) *random variables* $\alpha_{n,t,\phi}^{(1)}, \ldots, \alpha_{n,t,\phi}^{(h)}$ *and* $\nu \times 1$ *random vectors* $\mathbf{z}_{n,t,\phi}^{(1)}, \ldots, \mathbf{z}_{n,t,\phi}^{(h)}$ *such that* $\sup_{t\ge 1, n\ge d+m, \phi\in U}(|\alpha_{n,t,\phi}^{(j)}| + \|\mathbf{z}_{n,t,\phi}^{(j)}\|) < \infty$ a.s. *for* $j = 1, \ldots, h$, *where* $h = \min(p, d-1)$, *and*

$$\begin{aligned}\sup_{t\geq 1,\phi\in U}\Bigg\|&(n-d-m+2)^{-1}\sum_{i=m}^{n-d+1} Dy_{t,d}(\widehat{\epsilon}_{n,i+1},\ldots,\widehat{\epsilon}_{n,i+d-1},0;\phi)-Df_{\phi}(\mathbf{x}_{t+d})\\ &-\sum_{j=1}^{h}\alpha_{n,t,\phi}^{(j)}Df_{\phi}(\mathbf{x}_{t+d-j})-\sum_{j=1}^{h}\{f_{\phi}(\mathbf{x}_{t+d-j})-f_{\theta}(\mathbf{x}_{t+d-j})-\epsilon_{t+d-j}\}\mathbf{z}_{n,t,\phi}^{(j)}\Bigg\|\\ =&\,O(1)\ \text{ a.s.,}\end{aligned}\tag{3.22}$$

where D denotes the gradient vector $(\partial/\partial\phi_1,\ldots,\partial/\partial\phi_\nu)'$.

Proof. Let $\Delta_{t,\phi}=f_{\phi}(\mathbf{x}_t)-f_{\theta}(\mathbf{x}_t)$. First consider the case $d=2$. Since $y_{t,2}(w_1,w_2;\phi)=f_{\phi}(f_{\phi}(\mathbf{x}_{t+1})+w_1,y_t,\ldots,u_{t-\Delta-k+2})+w_2$, $\partial y_{t,2}/\partial w_2=1$ and $\partial y_{t,2}/\partial w_1=\partial f_{\phi}/\partial x_1$, so (3.21) follows from (3.4c). Moreover, $f_{\phi}(\mathbf{x}_{t+1})+\widehat{\epsilon}_{n,i+1}=y_{t+1}+\widehat{\epsilon}_{n,i+1}-\epsilon_{t+1}+\Delta_{t+1,\phi}$ and

$$\sum_{i=m}^{n-1} Dy_{t,2}(\widehat{\epsilon}_{n,i+1},0;\phi)=\sum_{i=m}^{n-1} Df_{\phi}(\mathbf{x}_{n,t,2}^{(i)})+\sum_{i=m}^{n-1}\frac{\partial f_{\phi}}{\partial x_1}(\mathbf{x}_{n,t,2}^{(i)})Df_{\phi}(\mathbf{x}_{t+1}),\tag{3.23}$$

where $\mathbf{x}_{n,t,2}^{(i)}=(f_{\phi}(\mathbf{x}_{t+1})+\widehat{\epsilon}_{n,i+1},y_t,\ldots,u_{t-\Delta-k+2})'=\mathbf{x}_{t+2}+(\widehat{\epsilon}_{n,i+1}-\epsilon_{t+1}+\Delta_{t+1,\phi},0,\ldots,0)'$. By Taylor's theorem,

$$Df_{\phi}(\mathbf{x}_{n,t,2}^{(i)})=Df_{\phi}(\mathbf{x}_{t+2})+(\widehat{\epsilon}_{n,i+1}-\epsilon_{t+1}+\Delta_{t+1,\phi})\Big(\frac{\partial^2 f_{\phi}}{\partial x_1\partial\phi_j}(\widehat{\mathbf{x}}_{n,t,2}^{(i)})\Big)'_{1\leq j\leq\nu},\tag{3.24}$$

where $\widehat{\mathbf{x}}_{n,t,2}^{(i)}$ lies between $\mathbf{x}_{t+2}$ and $\mathbf{x}_{n,t,2}^{(i)}$. Since $\widehat{\epsilon}_{n,i}-\epsilon_i=f_{\theta}(\mathbf{x}_i)-f_{\hat{\theta}_n}(\mathbf{x}_i)$ and $\sum_1^n\sup_{\phi\in U}\|Df_{\phi}(\mathbf{x}_i)\|\leq n^{1/2}\{\sum_1^n\sup_{\phi\in U}\|Df_{\phi}(\mathbf{x}_i)\|^2\}^{1/2}$, it follows from (3.4a) that

$$\sum_{i=m}^{n}|\widehat{\epsilon}_{n,i}-\epsilon_i|=O(n\|\widehat{\theta}_n-\theta\|)\ \text{ a.s.}\tag{3.25}$$

Since $\sum_{i=m}^n|\epsilon_i|=O(n)$ a.s., it follows from (3.25) and (3.4b) that

$$\sum_{i=m}^{n-1}|\partial^2 f_{\phi}(\widehat{\mathbf{x}}_{n,t,2}^{(i)})/\partial x_1\partial\phi_j|\,|\widehat{\epsilon}_{n,i+1}|=O(n)\ \text{a.s. for } j=1,\ldots,\nu.\tag{3.26}$$

In view of (3.23), (3.24) and (3.26), (3.22) holds for the case $d=2$ with

$$\alpha_{n,t,\phi}^{(1)}=(n-m)^{-1}\sum_{i=m}^{n-1}\partial f_{\phi}(\mathbf{x}_{n,t,,2}^{(i)})/\partial x_1,$$

$$\mathbf{z}_{n,t,\phi}^{(1)}=(n-m)^{-1}\sum_{i=m}^{n-1}(\partial^2 f_{\phi}(\widehat{\mathbf{x}}_{n,t,,2}^{(i)})/\partial x_1\partial\phi_j)'_{1\leq j\leq\nu}.$$

Moreover, by (3.4b) and (3.4c), $\sup_{t\geq 1, n\geq d+m, \phi\in U}(\|\mathbf{z}^{(1)}_{n,t,\phi}\| + |\alpha^{(1)}_{n,t,\phi}|) < \infty$ a.s.

We next consider the case $d = 3$. First suppose that $p \geq 2$, so $h = 2$. Since $y_{t,3}(w_1, w_2, w_3; \phi) = f_\phi(y_{t,2}(w_1, w_2; \phi), y_{t,1}(w_1; \phi), y_t, \dots, u_{t-\Delta-k+3}) + w_3$, $\partial y_{t,3}/\partial w_3 = 1$, $\partial y_{t,3}/\partial w_2 = \partial f_\phi/\partial x_1$, $\partial y_{t,3}/\partial w_1 = (\partial f_\phi/\partial x_1)(\partial y_{t,2}/\partial w_1) + \partial f_\phi/\partial x_2$, $\partial^2 y_{t,3}/\partial w_1 \partial w_2 = (\partial^2 f_\phi/\partial x_1^2)(\partial y_{t,2}/\partial w_1) + \partial^2 f_\phi/\partial x_1 \partial x_2$, and therefore (3.21) follows from (3.4c) and the corresponding result for the case $d = 2$. Note that $y_{t,1}(\widehat{\epsilon}_{n,i+1}; \phi) = y_{t+1} + \widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi}$ as mentioned above. A slight modification of (3.24) gives

$$f_\phi(\mathbf{x}^{(i)}_{n,t,2}) = f_\phi(\mathbf{x}_{t+2}) + (\widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi})\partial f_\phi(\widetilde{\mathbf{x}}^{(i)}_{n,t,2})/\partial x_1,$$

where $\mathbf{x}^{(i)}_{n,t,2} = \mathbf{x}_{t+2} + (\widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi}, 0, \dots, 0)'$ as before and $\widetilde{\mathbf{x}}^{(i)}_{n,t,2}$ lies between $\mathbf{x}_{t+2}$ and $\mathbf{x}^{(i)}_{n,t,2}$. Therefore

$$\begin{aligned} & y_{t,2}(\widehat{\epsilon}_{n,i+1}, \widehat{\epsilon}_{n,i+2}; \phi) \\ &= y_{t+2} + \widehat{\epsilon}_{n,i+2} - \epsilon_{t+2} + \Delta_{t+2,\phi} + (\widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi})\frac{\partial f_\phi}{\partial x_1}(\widetilde{\mathbf{x}}^{(i)}_{n,t,2}). \end{aligned} \tag{3.27}$$

Let $\mathbf{x}^{(i)}_{n,t,3} = (y_{t,2}(\widehat{\epsilon}_{n,i+1}, \widehat{\epsilon}_{n,i+2}; \phi), y_{t,1}(\widehat{\epsilon}_{n,i+1}; \phi), y_t, \dots, u_{t-\Delta-k+3})'$. Then $\mathbf{x}^{(i)}_{n,t,3} - \mathbf{x}_{t+3}$ is equal to

$$\Big(\widehat{\epsilon}_{n,i+2} - \epsilon_{t+2} + \Delta_{t+2,\phi} + (\widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi})\frac{\partial f_\phi}{\partial x_1}(\widetilde{\mathbf{x}}^{(i)}_{n,t,2}), \\ \widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi}, 0, \dots, 0\Big)'.$$

Hence, analogous to (3.24), we now have for $j = 1, \dots, \nu$

$$\begin{aligned} &\frac{\partial}{\partial \phi_j} f_\phi(\mathbf{x}^{(i)}_{n,t,3}) = \frac{\partial}{\partial \phi_j} f_\phi(\mathbf{x}_{t+3}) \\ &\quad + \Big\{\widehat{\epsilon}_{n,i+2} - \epsilon_{t+2} + \Delta_{t+2,\phi} + (\widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi})\frac{\partial}{\partial x_1} f_\phi(\widetilde{\mathbf{x}}^{(i)}_{n,t,2})\Big\} \\ &\quad \times \frac{\partial^2}{\partial x_1 \partial \phi_j} f_\phi(\widehat{\mathbf{x}}^{(i)}_{n,t,3}) + (\widehat{\epsilon}_{n,i+1} - \epsilon_{t+1} + \Delta_{t+1,\phi})\frac{\partial^2}{\partial x_2 \partial \phi_j} f_\phi(\widehat{\widehat{\mathbf{x}}}^{(i)}_{n,t,3}), \end{aligned} \tag{3.28}$$

where $\widehat{\mathbf{x}}^{(i)}_{n,t,3}$, $\widehat{\widehat{\mathbf{x}}}^{(i)}_{n,t,3}$ lie between $\mathbf{x}_{t+3}$ and $\mathbf{x}^{(i)}_{n,t,3}$. Moreover, since $y_{t,3} = f_\phi(y_{t,2}, y_{t,1}, y_t, \dots, u_{t-\Delta-k+3}) + w_3$, we have analogous to (3.23)

$$\begin{aligned} &\sum_{i=m}^{n-2} D y_{t,3}(\widehat{\epsilon}_{n,i+1}, \widehat{\epsilon}_{n,i+2}, 0; \phi) = \sum_{i=m}^{n-2} D f_\phi(\mathbf{x}^{(i)}_{n,t,3}) \\ &+ \sum_{i=m}^{n-2} (\partial f_\phi(\mathbf{x}^{(i)}_{n,t,3})/\partial x_1) D y_{t,2}(\widehat{\epsilon}_{n,i+1}, 0; \phi) + \sum_{i=m}^{n-2} (\partial f_\phi(\mathbf{x}^{(i)}_{n,t,3})/\partial x_2) D f_\phi(\mathbf{x}_{t+1}), \end{aligned} \tag{3.29}$$

noting that $Dy_{t,2}(w_1, w_2; \phi) = Dy_{t,2}(w_1, 0; \phi)$ and that $Dy_{t,1}(w; \phi) = Df_\phi(\mathbf{x}_{t+1})$. In view of (3.28) and (3.29) together with (3.25), (3.4b) and (3.4c), we obtain from an obvious modification of the corresponding result for the case $d = 2$ that (3.22) also holds for the case $d = 3$ and $p \geq 2$ for suitably defined $\alpha^{(1)}_{n,t,\phi}$, $\alpha^{(2)}_{n,t,\phi}$, $\mathbf{z}^{(1)}_{n,t,\phi}$, $\mathbf{z}^{(2)}_{n,t,\phi}$ with $\sup_{t\geq 1, n\geq d+m, \phi\in U}(|\alpha^{(1)}_{n,t,\phi}| + |\alpha^{(2)}_{n,t,\phi}| + \|\mathbf{z}^{(1)}_{n,t,\phi}\| + \|\mathbf{z}^{(2)}_{n,t,\phi}\|) < \infty$ a.s.

For the case $d = 3$ and $p = 1$, note that $y_{t,3}(w_1, w_2, 0; \phi) = f_\phi(y_{t,2}(w_1, w_2; \phi), u_{t-\Delta+3}, \cdots, u_{t-\Delta-k+3})$, and the argument is therefore completely analogous to that above. Proceeding inductively in this way, we can then establish the desired conclusions for $d = 4, 5, \ldots$.

Lemma 3. *With the same notation and assumptions as in Theorem* 2(i),

$$\sum_{n\leq N} \left\{(n-d-m+2)^{-1} \sum_{i=m}^{n-d+1} y_{n,d}(\epsilon_{i+1}, \cdots, \epsilon_{i+d-1}, 0; \theta) - \tilde{y}_{n+d}\right\}^2$$

$$= O(\log N) \text{ a.s.,} \tag{3.30}$$

$$\sup_{t\geq 1} \left|(n-d-m+2)^{-1} \sum_{i=m}^{n-d+1} y_{t,d}(\epsilon_{i+1}, \cdots, \epsilon_{i+d-1}, 0; \theta) - \tilde{y}_{t+d}\right|$$

$$= O((n^{-1} \log\log n)^{1/2}) \text{ a.s.} \tag{3.31}$$

Proof. Introduce the empirical distribution function

$$F_n(w_1, \cdots, w_{d-1}) = (n-d-m+2)^{-1} \sum_{i=m}^{n-d+1} I_{\{\epsilon_{i+1}\leq w_1, \cdots, \epsilon_{i+d-1}\leq w_{d-1}\}}.$$

Let F be the distribution function of $(\epsilon_1, \cdots, \epsilon_{d-1})$, so $dF(w_1, \cdots, w_{d-1}) = dH(w_1) \cdots dH(w_{d-1})$. Note that for $t \geq 1$,

$$(n-d-m+2)^{-1} \sum_{i=m}^{n-d+1} y_{t,d}(\epsilon_{i+1}, \cdots, \epsilon_{i+d-1}, 0; \theta) - \tilde{y}_{t+d}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} y_{t,d}(w_1, \cdots, w_{d-1}, 0; \theta) d(F_n - F). \tag{3.32}$$

For notational simplicity, we shall focus on the case $d = 3$. Denote $y_{t,3}(w_1, w_2, 0; \theta)$ simply by $y_t(w_1, w_2)$. Let $G_n = F_n - F$, which we shall also regard as a signed measure. Expressing $y_t(w_1, w_2) - y_t(0, 0)$ as

$$\int_0^{w_1} \int_0^{w_2} \frac{\partial^2 y_t}{\partial v_1 \partial v_2}(v_1, v_2) dv_1 dv_2 + \int_0^{w_1} \frac{\partial y_t}{\partial v_1}(v_1, 0) dv_1 + \int_0^{w_2} \frac{\partial y_t}{\partial v_2}(0, v_2) dv_2,$$

$$\text{if } w_1 \geq 0,\ w_2 \geq 0, \tag{3.33}$$

and as

$$-\int_{w_1}^{0}\int_{0}^{w_2}\frac{\partial^2 y_t}{\partial v_1\partial v_2}(v_1,v_2)dv_1dv_2-\int_{w_1}^{0}\frac{\partial y_t}{\partial v_1}(v_1,0)dv_1+\int_{0}^{w_2}\frac{\partial y_t}{\partial v_2}(0,v_2)dv_2,$$
$$\text{if } w_1<0,\ w_2\geq 0,$$

with similar expressions for the cases $w_1 \geq 0$, $w_2 < 0$ and $w_1 < 0$, $w_2 < 0$, and noting that $y(0,0)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dG_n = 0$, we obtain by Fubini's theorem that

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_t(w_1,w_2)dG_n(w_1,w_2)=\\
&\int_{0}^{\infty} G_n([v_1,\infty)\times\mathbf{R})\frac{\partial y_t}{\partial v_1}(v_1,0)dv_1-\int_{-\infty}^{0} G_n((-\infty,v_1)\times\mathbf{R})\frac{\partial y_t}{\partial v_1}(v_1,0)dv_1\\
&+\int_{0}^{\infty} G_n(\mathbf{R}\times[v_2,\infty))\frac{\partial y_t}{\partial v_2}(0,v_2)dv_2-\int_{-\infty}^{0} G_n(\mathbf{R}\times(-\infty,v_2))\frac{\partial y_t}{\partial v_2}(0,v_2)dv_2\\
&+\int\int_{\substack{v_1\geq 0\\ v_2\geq 0}} G_n([v_1,\infty)\times[v_2,\infty))\frac{\partial^2 y_t}{\partial v_1\partial v_2}(v_1,v_2)dv_1dv_2\\
&-\int\int_{\substack{v_1<0\\ v_2\geq 0}} G_n((-\infty,v_1)\times[v_2,\infty))\frac{\partial^2 y_t}{\partial v_1\partial v_2}(v_1,v_2)dv_1dv_2\\
&-\int\int_{\substack{v_1\geq 0\\ v_2<0}} G_n([v_1,\infty)\times(-\infty,v_2))\frac{\partial^2 y_t}{\partial v_1\partial v_2}(v_1,v_2)dv_1dv_2\\
&+\int\int_{\substack{v_1<0\\ v_2<0}} G_n((-\infty,v_1)\times(-\infty,v_2))\frac{\partial^2 y_t}{\partial v_1\partial v_2}(v_1,v_2)dv_1dv_2. \qquad (3.34)
\end{aligned}$$

It therefore suffices for the proof of (3.30) and (3.31) to establish similar results for each of the eight summands above.

Recalling that $E|\epsilon_1|^\alpha < \infty$ for some $\alpha > 2$, we can choose $0 < \beta < \frac{1}{2}$ such that $\int_{-\infty}^{0} H^\beta(u)dv + \int_0^\infty (1-H(u))^\beta dv < \infty$. Let $g(v) = \max(H^\beta(v), e^{-|v|})$ for $v < 0$ and let $A = \int_{-\infty}^{0} g(v)dv$ (> 0). Consider, in particular, the integral

$$\int_{-\infty}^{0}\int_{-\infty}^{0} G_n(v_1,v_2)\frac{\partial^2 y_t}{\partial v_1\partial v_2}dv_1dv_2=A^2\int_{-\infty}^{0}\int_{-\infty}^{0}\frac{G_n(v_1,v_2)}{g(v_1)g(v_2)}\frac{\partial^2 y_t}{\partial v_1\partial v_2}d\lambda(v_1,v_2), \qquad (3.35)$$

where λ is a probability measure on $(-\infty,0)\times(-\infty,0)$ defined by $d\lambda(v_1,v_2) = A^{-2}g(v_1)g(v_2)dv_1dv_2$. Since $\sup_{t,v_1,v_2}|\partial^2 y_t/\partial v_1\partial v_2| < \infty$ a.s. by Lemma 2 and since $\sup_{v_1\leq 0,v_2\leq 0}|G_n(v_1,v_2)/g(v_1)g(v_2)| = O((n^{-1}\log\log n)^{1/2})$ a.s. by a multivariate analog of James' (1975) law of the iterated logarithm for weighted empirical processes, we have by (3.35) that

$$\sup_{t\geq 1}\Big|\int_{-\infty}^{0}\int_{-\infty}^{0} G_n(v_1,v_2)\frac{\partial^2 y_t}{\partial v_1\partial v_2}dv_1dv_2\Big| = O\big((n^{-1}\log\log n)^{1/2}\big) \text{ a.s.} \qquad (3.36)$$

We next proceed to show that

$$\sum_{n\leq N}\Big\{\int_{-\infty}^{0}\int_{-\infty}^{0}[G_n(v_1,v_2)/g(v_1)g(v_2)][\partial^2 y_n/\partial v_1\partial v_2]d\lambda\Big\}^2$$
$$\leq \sum_{n\leq N}\int_{-\infty}^{0}\int_{-\infty}^{0}\frac{G_n^2(v_1,v_2)}{g^2(v_1)g^2(v_2)}\Big(\frac{\partial^2 y_t}{\partial v_1\partial v_2}\Big)^2 d\lambda = O(\log N) \text{ a.s.} \qquad (3.37)$$

The first inequality in (3.36) is a consequence of the Schwarz inequality. Since $\sup_{v_1,v_2}|\partial^2 y_n/\partial v_1\partial v_2| = O(1)$ a.s. by Lemma 2, it remains to show that

$$\int_{-\infty}^{0}\int_{-\infty}^{0}\sum_{n=1}^{N}\Big(n^{-1}\sum_{i=1}^{n}X_i(v_1,v_2)\Big)^2 d\lambda(v_1,v_2) = O(\log N) \text{ a.s.}, \qquad (3.38)$$

where $X_i(v_1,v_2) = \{I_{\{\epsilon_{i+1}\leq v_1,\epsilon_{i+2}\leq v_2\}} - H(v_1)H(v_2)\}/\{g(v_1)g(v_2)\}$. Note that $\{X_1,X_3,\dots\}$ and $\{X_2,X_4,\dots\}$ are two i.i.d. sequences of random functions and that $(\sum_{i=1}^{n}X_i)^2 \leq 2(\sum_{1\leq j\leq n/2}X_{2j})^2 + 2(\sum_{0\leq j<n/2}X_{2j+1})^2$. Moreover, since $(2\beta)^{-1} > 1$,

$$E\Big\{\int_{-\infty}^{0}\int_{-\infty}^{0}X_i^2(v_1,v_2)d\lambda\Big\}^{(2\beta)^{-1}} \leq E\int_{-\infty}^{0}\int_{-\infty}^{0}|X_i(v_1,v_2)|^{1/\beta}d\lambda$$
$$\leq 2^{1/\beta}\int_{-\infty}^{0}\int_{-\infty}^{0}[H(v_1)H(v_2) + (H(v_1)H(v_2))^{1/\beta}][g(v_1)g(v_2)]^{-1/\beta}d\lambda < \infty,$$

noting that $(g(v))^{-1/\beta} \leq 1/H(v)$. Hence (3.38) is a special case of a more general result of Lai (1990b).

The other summands in (3.34) can be analyzed similarly, defining $g(v) = \max((1-H(v))^\beta, e^{-v})$ for $v \geq 0$. In view of (3.32), this proves the lemma in the case $d = 3$. The case $d = 2$ is even simpler, for which (3.34) takes the form

$$\int_{-\infty}^{\infty} y_{t,2}(v,0;\theta)dG_n(v)$$
$$= \int_{0}^{\infty} G_n([v,\infty))\frac{\partial y_{t,2}}{\partial v}(v,0;\theta)dv - \int_{-\infty}^{0} G_n((-\infty,v))\frac{\partial y_{t,2}}{\partial v}(v,0;\theta)dv.$$

As shown in Lai (1990a), the representation (3.33), and therefore (3.34) also, can be generalized to more than two variables, and therefore the same arguments can be extended to prove the lemma for general d.

Proof of Theorem 2. Let $\mathbf{z}_i = Df_\theta(\mathbf{x}_i)$. By (2.26) together with (3.14), (3.13) and (3.12) of Lemma 1, for $r = 1, 2, \ldots,$

$$\widehat{\theta}_n - \theta = O\big((n^{-1}\log\log n)^{\frac{1}{2}}\big) \quad \text{a.s.,} \tag{3.39}$$

$$\begin{aligned}\sum_{n=1}^{N}\|\widehat{\theta}_n - \theta\|^2 &= \sum_{n=1}^{N}\Big\{O\Big(\Big\|\Big(\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'\Big)^{-1}\Big(\sum_{i=1}^{n}\epsilon_i\mathbf{z}_i\Big)\Big\|^2\Big) + O(n^{-3/2+4\delta})\Big\}\\ &= O(\log N) \quad \text{a.s.,}\end{aligned} \tag{3.40}$$

$$\begin{aligned}\sum_{n=1}^{N}[\mathbf{z}_{n+r}'(\widehat{\theta}_n - \theta)]^2 &= O(\log N) + O\Big(\sum_{n=1}^{N}\|\mathbf{z}_{n+r}\|^2 n^{-3/2+4\delta}\Big)\\ &= O(\log N) + O\Big(\sum_{i:2^i\le N} 2^{-3i/2+4\delta i}\sum_{n=2^{i-1}}^{2^i}\|\mathbf{z}_{n+r}\|^2\Big) = O(\log N) \quad \text{a.s.,}\end{aligned} \tag{3.41}$$

in view of (3.4a), taking $\delta < 1/8$. By Taylor's theorem,

$$\begin{aligned}&\sum_{i=m}^{n-d+1} y_{t,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \widehat{\theta}_n) - \sum_{i=m}^{n-d+1} y_{t,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \theta)\\ &\quad = (\widehat{\theta}_n - \theta)'\sum_{i=m}^{n-d+1} Dy_{t,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \theta^*),\end{aligned} \tag{3.42}$$

where θ^* lies between $\widehat{\theta}_n$ and θ. Let $\Delta_{n,\phi} = f_\phi(\mathbf{x}_n) - f_\theta(\mathbf{x}_n)$. Combining (3.42) with (3.22) of Lemma 2 yields

$$\begin{aligned}&\sum_{n=n_0}^{N}\Big\{(n-d-m+2)^{-1}\sum_{i=m}^{n-d+1}[y_{n,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \widehat{\theta}_n)\\ &\qquad\qquad - y_{n,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \theta)]\Big\}^2\\ &= O\Big(\sum_{n=n_0}^{N}\Big\{[(\widehat{\theta}_n - \theta)'Df_{\theta^*}(\mathbf{x}_{n+d})]^2 + \sum_{j=1}^{h}[(\widehat{\theta}_n - \theta)'Df_{\theta^*}(\mathbf{x}_{n+d-j})]^2\Big\}\Big)\\ &\quad + O\Big(\sum_{j=1}^{h}\sum_{n=n_0}^{N}\epsilon_{n+d-j}^2\|\widehat{\theta}_n - \theta\|^2\Big) + O\Big(\sum_{j=1}^{h}\sum_{n=n_0}^{N}\Delta_{n+d-j,\theta^*}^2\|\widehat{\theta}_n - \theta\|^2\Big)\\ &\quad + O\Big(\sum_{n=n_0}^{N}\|\widehat{\theta}_n - \theta\|^2\Big) = O(\log N) \quad \text{a.s.}\end{aligned} \tag{3.43}$$

The last relation above follows from (3.40) and the following bounds for fixed $j \geq 1$:

$$\sum_{n=n_0}^{N} \epsilon_{n+j}^2 \|\widehat{\theta}_n - \theta\|^2 \sim \sigma^2 \sum_{n=n_0}^{N} \|\widehat{\theta}_n - \theta\|^2, \text{ by Lemma 2(iii) of Lai and Wei (1982a),}$$

$$= O(\log N) \text{ a.s., by (3.40);}$$

$$\sum_{n=n_0}^{N} \Delta_{n+j,\theta^*}^2 \|\widehat{\theta}_n - \theta\|^2 = O\Big(\sum_{n=n_0}^{N} \|\widehat{\theta}_n - \theta\|^4 \sup_{\phi \in U} \|Df_\phi(\mathbf{x}_{n+j})\|^2 \Big)$$

$$= O\Big(\sum_{i:2^i \leq N} 2^{-2i} (\log i)^2 \sum_{n=2^{i-1}}^{2^i} \sup_{\phi \in U} \|Df_\phi(\mathbf{x}_{n+j})\|^2 \Big) = O(1) \text{ a.s., by (3.39) and (3.4a);}$$

$$\sum_{n=n_0}^{N} [(\widehat{\theta}_n - \theta)' Df_{\theta^*}(\mathbf{x}_{n+j})]^2$$

$$= \sum_{n=n_0}^{N} [(\widehat{\theta}_n - \theta)' Df_\theta(\mathbf{x}_{n+j})]^2 + O\Big(\sum_{n=n_0}^{N} \|\widehat{\theta}_n - \theta\|^4 \sup_{\phi \in U} \|D^2 f_\phi(\mathbf{x}_{n+j})\|^2 \Big)$$

$$= O(\log N) \text{ a.s., by (3.41), (3.40) and (3.4a).}$$

By Lemma 2 and (3.25),

$$\sup_{t \geq 1} \sum_{i=m}^{n-d+1} |y_{t,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \theta) - y_{t,d}(\epsilon_{i+1}, \ldots, \epsilon_{i+d-1}, 0; \theta)|$$

$$= O\Big(\sum_{i=m}^{n} |\widehat{\epsilon}_{n,i} - \epsilon_i| \Big) = O(n \|\widehat{\theta}_n - \theta\|) \text{ a.s.} \tag{3.44}$$

From (3.44) and (3.40), it follows that

$$\sum_{n=n_0}^{N} \Big\{ (n - d - m + 2)^{-1} \sum_{i=m}^{n-d+1} [y_{n,d}(\widehat{\epsilon}_{n,i+1}, \ldots, \widehat{\epsilon}_{n,i+d-1}, 0; \theta) - y_{n,d}(\epsilon_{i+1}, \ldots, \epsilon_{i+d-1}, 0; \theta)] \Big\}^2$$

$$= O\Big(\sum_{n=n_0}^{N} \|\widehat{\theta}_n - \theta\|^2 \Big) = O(\log N) \text{ a.s.} \tag{3.45}$$

Combining (3.43) and (3.45) with (3.30) of Lemma 3, we obtain the desired conclusion (3.5).

Suppose furthermore that H has bounded support and that $\sup_{\phi \in U} \|Df_\phi(\mathbf{x}_n)\| = O(1)$ a.s. To prove (3.6), first note that by (3.44) and (3.39),

$$\begin{aligned}&\sup_{t\geq 1}(n-d-m+2)^{-1}\sum_{i=m}^{n-d+1}|y_{t,d}(\widehat{\epsilon}_{n,i+1},\ldots,\widehat{\epsilon}_{n,i+d-1},0;\theta)\\&\qquad\qquad\qquad\qquad\qquad\qquad-y_{t,d}(\epsilon_{i+1},\ldots,\epsilon_{i+d-1},0;\theta)|\\&=O((n^{-1}\log\log n)^{1/2})\ \text{a.s.}\end{aligned}\tag{3.46}$$

Moreover, since $|\epsilon_t|+\sup_{\phi\in U}\|Df_\phi(\mathbf{x}_t)\|=O(1)$ a.s., it follows from (3.42) and (3.22) that

$$\begin{aligned}&\sup_{t\geq 1}(n-d-m+2)^{-1}\sum_{i=m}^{n-d+1}|y_{t,d}(\widehat{\epsilon}_{n,i+1},\ldots,\widehat{\epsilon}_{n,i+d-1},0;\widehat{\theta}_n)\\&\qquad\qquad\qquad\qquad\qquad\qquad-y_{t,d}(\widehat{\epsilon}_{n,i+1},\ldots,\widehat{\epsilon}_{n,i+d-1},0;\theta)|\\&=O(\|\widehat{\theta}_n-\theta\|)=O((n^{-1}\log\log n)^{1/2})\ \text{a.s., by (3.39).}\end{aligned}\tag{3.47}$$

Combining (3.46) and (3.47) with (3.31), we obtain the desired conclusion (3.6).

4. Estimation of the Variance and Other Functionals of d-Step Ahead Predictive Distributions

The minimum variance d-step ahead predictor (1.7) is the mean of the predictive (conditional) distribution of y_{n+d} given the current and past outputs and inputs $y_n,u_n,\ldots,y_1,u_1$. Therefore the adaptive predictor (1.8) can be interpreted as an estimate of the mean of this predictive distribution when θ and H are both unknown. The variance of the predictive distribution is

$$\begin{aligned}&\mathrm{Var}(y_{n+d}|\mathcal{F}_n)\\&=\sigma^2(=\mathrm{Var}\,\epsilon_1),\quad\text{if }\ d=1,\\&=\sigma^2+\int\cdots\int y^2_{n,d}(w_1,\ldots,w_{d-1},0;\theta)dH(w_1)\cdots dH(w_{d-1})-\widetilde{y}^2_{n+d},\ \text{if } d\geq 2,\end{aligned}\tag{4.1}$$

where $y_{n,d}$ is defined in (1.6). Note that $\mathrm{Var}(y_{n+d}|\mathcal{F}_n)$ is the conditional mean squared prediction error $E[(y_{n+d}-\widetilde{y}_{n+d})^2|\mathcal{F}_n]$ of the minimum variance predictor $\widetilde{y}_{n+d}$. When σ^2 and H are not known in advance, let $\widehat{\theta}_n$ be the least squares estimate of θ and let $\widehat{\epsilon}_{n,i}=y_i-f_{\widehat{\theta}_n}(\mathbf{x}_i)$, $i\leq n$, be the residuals, as in Section 3. An estimate of (4.1) based on the current and past observations $y_1,u_1,\ldots,y_n,u_n$ is

$$\begin{aligned}&\widehat{\mathrm{Var}}(y_{n+d}|\mathcal{F}_n)\\&=\widehat{\sigma}^2_n:=(n-m+1)^{-1}\sum_{i=m}^{n}\widehat{\epsilon}^2_{n,i},\quad d=1,\end{aligned}$$

$$= \hat{\sigma}_n^2 + (n-m-d+2)^{-1} \sum_{i=m}^{n-d+1} y_{n,d}^2(\hat{\epsilon}_{n,i+1}, \dots, \hat{\epsilon}_{n,i+d-1}, 0; \hat{\theta}_n) - \hat{y}_{n+d}^2, \quad d \geq 2,$$

where $\hat{y}_{n+d}$ is defined in (1.8) and $m = \max(p, k+\Delta)$. Under the assumptions of Theorem 2(ii) and assuming that $\sup_n |y_n| < \infty$ a.s., it can be shown by a straightforward modification of the preceding proof that

$$\widehat{\mathrm{Var}}(y_{n+d}|\mathcal{F}_n) - \mathrm{Var}(y_{n+d}|\mathcal{F}_n) = O((n^{-1}\log\log n)^{\frac{1}{2}}) \text{ a.s.}$$

The same ideas can be used to estimate other moments of the predictive distribution of y_{n+d} given $\mathcal{F}_n$. Under the assumptions of Theorem 2(ii) and assuming H to be continuous, we can also obtain uniformly strongly consistent (cf. (4.2) below) estimates of the predictive distribution function $G_{n,d}(t) = P\{y_{n+d} \leq t|\mathcal{F}_n\}$ by

$$\hat{G}_{n,d}(t) = (n-m-d+1)^{-1} \sum_{i=m}^{n-d} I_{\{y_{n,d}(\hat{\epsilon}_{n,i+1},\dots,\hat{\epsilon}_{n,i+d};\hat{\theta}_n) \leq t\}}.$$

Since $G_{n,d}(t) = \int \cdots \int P\{y_{n,d}(w_1, \dots, w_d; \theta) \leq t\} dH(w_1) \cdots dH(w_{d-1}) dH(w_d)$, it can be shown by a modification of the proof of Theorem 2(ii) together with a Glivenko-Cantelli-type argument that

$$\sup_t |\hat{G}_{n,d}(t) - G_{n,d}(t)| \to 0 \text{ a.s. as } n \to \infty. \tag{4.2}$$

Using the quantiles of $\hat{G}_{n,d}$, we can also obtain strongly consistent estimates of the quantiles of the predictive distribution.

Acknowledgement

This research is supported by the National Science Foundation, the National Security Agency and the Air Force Office of Scientific Research.

References

Al-Qassam, M. S. and Lane, J. A. (1989). Forecasting exponential autoregressive models of order 1. *J. Time Ser. Anal.* **10**, 95–113.

Åström, K. J. (1970). *Introduction to Stochastic Control Theory.* Academic Press, New York.

James, B. R. (1975). A functional law of the iterated logarithm for weighted empirical distributions. *Ann. Probab.* **3**, 762–772.

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

Lai, T. L. (1990a). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. Technical Report, Department of Statistics, Stanford University.

Lai, T. L. (1990b). Covariance operators and limit theorems of Hilbert space valued martingales with applications to adaptive prediction and control. Technical Report, Department of Statistics, Stanford University.

Lai, T. L. and Wei, C. Z. (1982a). 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. (1982b). Asymptotic properties of projections with applications to stochastic regression problems. *J. Multivariate Anal.* **12**, 346–370.

Lai, T. L. and Ying, Z. (1991). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM J. Control & Optimization* **29**, in press.

Stout, W. F. (1973). Maximal inequalities and the law of the iterated logarithm. *Ann. Probab.* **1**, 322–328.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis.* Springer-Verlag, New York-Berlin-Heidelberg-Tokyo.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach.* Oxford University Press.

Wei, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15**, 1667-1682.

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

(Received January 1990; accepted January 1991)