The Annals of Statistics
2011, Vol. 39, No. 3, 1526–1550
DOI: 10.1214/10-AOS861

UNIFORM MOMENT BOUNDS OF FISHER'S INFORMATION WITH APPLICATIONS TO TIME SERIES

BY NGAI HANG CHAN[1] AND CHING-KANG ING[2]

Chinese University of Hong Kong and Academia Sinica

In this paper, a uniform (over some parameter space) moment bound for the inverse of Fisher's information matrix is established. This result is then applied to develop moment bounds for the normalized least squares estimate in (nonlinear) stochastic regression models. The usefulness of these results is illustrated using time series models. In particular, an asymptotic expression for the mean squared prediction error of the least squares predictor in autoregressive moving average models is obtained. This asymptotic expression provides a solid theoretical foundation for some model selection criteria.

1. Introduction. Moment inequalities and moment bounds have long been vibrant topics in modern probability and statistics. The celebrated inequalities of Burkholder [3] and Doob [5] offer exemplary illustrations of the importance of moment inequalities. Using moment bounds, the order of magnitude of the spectral norm of the inverse of the Fisher's information matrix can be quantified and consistency and efficiency of least squares estimates of stochastic regression and adaptive control can be established; see, for example, the seminal work of Lai and Wei [15] and the succinct review of Lai and Ying [16]. In this paper, a uniform (over some parameter space) moment bound for the inverse of the Fisher's information matrix is established. This bound is used to investigate the moment properties of least squares estimates and the mean squared prediction error (MSPE) for time series models.

To appreciate the significance of uniform moment bounds, consider the stochastic regression model

$$y_t = g_t(\theta_0) + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{1.1}$$

where $g_t(\cdot)$ is a random function, θ_0 is an unknown parameter and $\{\varepsilon_t\}$ is a martingale difference sequence. There are two important problems related to this model.

Received May 2010; revised August 2010.

[1]Supported in part by the General Research Fund Nos. 400408 and 400410 and the Collaborative Research Fund No. CityU8/CRF/09, all from the Research Grants Council of Hong Kong.

[2]Supported in part by the National Science Council of Taiwan under Grant NSC 94-2118-M-001-013.

MSC2010 subject classifications. Primary 62J02; secondary 62M10, 62F12, 60F25.

Key words and phrases. Fisher's information matrix, least squares estimates, mean squared prediction errors, stochastic regression models, uniform moment bounds.

The first one concerns the mean squared error prediction. In practice, the unknown parameter θ_0 is usually estimated by the least squares estimate $\hat{\theta}_n$, which minimizes $S_n(\theta) = \sum_{t=1}^{n}(y_t - g_t(\theta))^2$. Although the (strong) law of large numbers (LLN) and the central limit theorem (CLT) of $\hat{\theta}_n$ were established under certain assumptions on $g_t(\cdot)$ and ε_t (see among others, Lai [14] and Skouras [19]), relatively little is known about the moment convergence of $\hat{\theta}_n$. Moment convergence of $\hat{\theta}_n$ offers important insight in the pursuit of the mean squared prediction problem. To see this, suppose that $n^{1/2}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $\eta > 0$. Then an immediate question is to pursue

$$\mathrm{E}|n^{1/2}(\hat{\theta}_n - \theta_0)|^q = O(1), \qquad q \geq 1. \tag{1.2}$$

In particular, if (1.2) holds for some $q > 2$, then $\{n(\hat{\theta}_n - \theta_0)^2\}$ is uniformly integrable and consequently, $\lim_{n\to\infty} n\mathrm{E}(\hat{\theta}_n - \theta_0)^2 = \eta$. This result can be applied to develop an asymptotic expression for the mean squared error of $\hat{\theta}_n$ as

$$\mathrm{E}(\hat{\theta}_n - \theta_0)^2 = \frac{\eta}{n} + o(n^{-1})$$

from which asymptotic properties of the MSPE of the least squares predictor $g_{n+1}(\hat{\theta}_n)$ of y_{n+1}, $E(y_{n+1} - g_{n+1}(\hat{\theta}_n))^2$, can be established; see Sections 2 and 3 for further details.

To establish (1.2), consider the Fisher's information number, $n^{-1}\sum_{t=1}^{n}(g_t'(\theta))^2$ of (1.1), where $g_t'(\theta) = dg_t(\theta)/d\theta$. As will be shown in Section 2, it turns out that the uniform negative moment bound for $n^{-1}\sum_{t=1}^{n}(g_t'(\theta))^2$, that is, for any $q \geq 1$,

$$\mathrm{E}\left\{\sup_{\theta\in B_{\delta_1}(\theta_0)}\left(n^{-1}\sum_{t=1}^{n}(g_t'(\theta))^2\right)^{-q}\right\} = O(1) \tag{1.3}$$

plays a crucial role in proving (1.2), where $B_{\delta_1}(\theta_0) = \{\theta : |\theta - \theta_0| < \delta_1\}$ for some $\delta_1 > 0$.

A second but equally important problem in stochastic regression concerns model selection. To understand how the uniform moment bound is related to this issue, consider the case when $g_t(\cdot)$ in (1.1) contains $k > 1$ unknown parameters $\boldsymbol{\theta}_0 \in R^k$. A multiparameter generalization of (1.3) becomes: for any $q \geq 1$,

$$\mathrm{E}\left\{\sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_0)}\lambda_{\min}^{-q}\left(n^{-1}\sum_{t=1}^{n}\nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}}\right)\right\} = O(1), \tag{1.4}$$

where $\lambda_{\min}(L)$ denotes the minimum eigenvalue of the matrix L and $\nabla g_t(\boldsymbol{\theta})$ denotes the gradient vector of $g_t(\boldsymbol{\theta})$. In particular, when $g_t(\boldsymbol{\theta}) = g_t(\theta_1, \ldots, \theta_k) = \theta_1 y_{t-1} + \cdots + \theta_k y_{t-k}$ in (1.1), that is, when y_t is an autoregressive (AR) model of order k, (1.4) reduces to

$$\mathrm{E}\left\{\lambda_{\min}^{-q}\left(n^{-1}\sum_{t=1}^{n}\mathbf{y}_{t-1}(k)\mathbf{y}_{t-1}^{\mathrm{T}}(k)\right)\right\} = O(1), \tag{1.5}$$

where $\mathbf{y}_t(k) = (y_t, \ldots, y_{t-k+1})^{\mathrm{T}}$. By imposing a Lipschitz type condition on the distribution function of ε_t and a stationarity condition on $g_t(\cdot)$, Findley and Wei [7] established (1.5), thereby providing a rigorous mathematical derivation of the AIC model selection criterion for weakly stationary AR processes. However, the proof of (1.4) for a general stochastic regression model is much more involved than (1.5) due to the presence of an "extra" supremum, which is taken over an uncountable set inside the expectation. As a consequence, similar to the AR case, knowledge about negative uniform moment bounds of the Fisher's information matrix (1.4) constitutes an indispensable tool for the model selection problem.

The rest of this paper is organized as follows. In Section 2, we first show in Theorem 2.1 that (1.4) holds under more general situations where $B_{\delta_1}(\boldsymbol{\theta}_0)$ is replaced by a bounded subset $\boldsymbol{\Theta}$ of R^k and $\nabla g_t(\boldsymbol{\theta})$ is replaced by a vector-valued random function $\mathbf{f}_t(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$, satisfying certain assumptions. We then apply Theorem 2.1 to establish the moment convergence of least squares estimates in (nonlinear) stochastic regression models; see Theorem 2.2. Section 3 focuses on the applications of Theorems 2.1 and 2.2 to autoregressive moving average (ARMA) models. In particular, the moment convergence of the least squares estimates and an asymptotic expression (up to terms of order n^{-1}) for the MSPE of the least squares predictor for ARMA models are established. To facilitate the presentation, technical results of Sections 2 and 3 are deferred to Appendices A and B, respectively.

2. Uniform bounds on negative moments. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_t\}$ be an increasing sequence of σ-fields on $(\Omega, \mathcal{F}, P)$. Let $\mathbf{f}_t(\boldsymbol{\theta}), t = 1, \ldots, n$, be r-dimensional $\mathcal{F}_t$-measurable random functions of a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}} \in \boldsymbol{\Theta} \subset R^k$. In the first half this section, we provide sufficient conditions under which the minimum eigenvalue of the normalized matrix $n^{-1}\sum_{t=1}^n \mathbf{f}_t(\boldsymbol{\theta})\mathbf{f}_t^{\mathrm{T}}(\boldsymbol{\theta})$, $\lambda_{\min}(n^{-1}\sum_{t=1}^n \mathbf{f}_t(\boldsymbol{\theta})\mathbf{f}_t^{\mathrm{T}}(\boldsymbol{\theta}))$, satisfies the following uniform moment bound:

$$\mathrm{E}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}^{-q}\left(n^{-1}\sum_{t=1}^{n} \mathbf{f}_t(\boldsymbol{\theta})\mathbf{f}_t^{\mathrm{T}}(\boldsymbol{\theta})\right)\right\} = O(1) \qquad \text{for any } q \geq 1. \tag{2.1}$$

This uniform negative moment bound is applied to investigate the moment properties of least squares estimates in the second half of this section. To begin, assume the following conditions:

(C1) $\mathbf{f}_t(\boldsymbol{\theta})$ is continuous on $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ is a bounded subset of R^k;

(C2) there exist positive integer d and positive numbers δ, α and M such that for any $t > d$, any $0 < s_2 - s_1 \leq \delta$, any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and any $\|\mathbf{a}\| = 1$,

$$P\big(s_1 < \mathbf{a}^{\mathrm{T}}\mathbf{f}_t(\boldsymbol{\theta}) \leq s_2 | \mathcal{F}_{t-d}\big) \leq M(s_2 - s_1)^{\alpha} \qquad \text{a.s.},$$

where $\|\mathbf{a}\|$ denotes the Euclidean norm of vector $\mathbf{a} \in R^r$;

(C3) there exist $\tau > 0$ and nonnegative random variables B_t satisfying $\sup_{t\geq 1} \mathrm{E}(B_t) \leq C_1$ for some $C_1 > 0$ such that for all $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\| < \tau$,

$$\|\mathbf{f}_t(\boldsymbol{\xi}_1) - \mathbf{f}_t(\boldsymbol{\xi}_2)\| \leq B_t \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\| \qquad \text{a.s.;}$$

(C4) there exists $C_2 > 0$ such that $\sup_{t\geq 1} \mathrm{E}(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{f}_t(\boldsymbol{\theta})\|^2) \leq C_2$.

(C1) is a standard assumption for the regression function and its gradient vector in nonlinear regression; see, for example, Lai [14] and Robinson and Hidalgo [17]. (C2) says that given the information (σ-field) whose time index is sufficiently smaller than the current time index t, the conditional distribution of $\mathbf{a}^{\mathrm{T}}\mathbf{f}_t(\boldsymbol{\theta})$ follows a local Lipschitz condition of order α for all points $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all directions $\mathbf{a}$ with $\|\mathbf{a}\| = 1$. In the special case when $\boldsymbol{\Theta}$ contains only one point, (C2) is related to Findley and Wei's [7] *uniform Lipschitz condition over all directions*, which is the key assumption used in deriving the AIC for stationary AR models. Since we need to deal with the supremum over a class of inverses of minimum eigenvalues indexed by $\boldsymbol{\theta}$, a Lipschitz type condition over all points ($\boldsymbol{\theta}$) in all directions ($\mathbf{a}$) is required in this paper. As will be seen in Section 3, (C2) is flexible enough to encompass many time series applications. Conditions like (C3) have been imposed on the regression function by Andrews [2] and Skouras [19] in proving the uniform law of large numbers for random functions associated with $S_n(\theta)$. (C3) can be verified when $\mathbf{f}_t(\boldsymbol{\theta})$ is sufficiently smooth; see (3.26) for more details. (C4) imposes a mild moment condition on $\mathbf{f}_t(\boldsymbol{\theta})$ and appears to be satisfied in many practical situations. Moreover, (C4) can be weakened to $\sup_{t\geq 1} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}(\|\mathbf{f}_t(\boldsymbol{\theta})\|^2) \leq C_2$ for some $C_2 > 0$ at the price of strengthening the conditions on B_t in (C3) to $\sup_{t\geq 1} \mathrm{E}(B_t^2) \leq C_1$ for some $C_1 > 0$.

THEOREM 2.1. *Assume that* (C1)–(C4) *hold. Then inequality* (2.1) *is true.*

PROOF. First, note that the measurability of $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}^{-q}(n^{-1}\sum_{t=1}^{n} \mathbf{f}_t(\boldsymbol{\theta}) \times \mathbf{f}_t^{\mathrm{T}}(\boldsymbol{\theta}))$ is ensured by the continuity of $\mathbf{f}_t(\boldsymbol{\theta})$. Define $n_d = \lfloor (n-d)/d \rfloor$, where $\lfloor a \rfloor$ is the largest integer $\leq a$. Then for n large,

$$\begin{aligned} & n^q \lambda_{\min}^{-q}\left(\sum_{t=1}^{n} \mathbf{f}_t(\boldsymbol{\theta})\mathbf{f}_t^{\mathrm{T}}(\boldsymbol{\theta})\right) \\ & \leq n^q \left\{\sum_{j=1}^{d} \lambda_{\min}\left(\sum_{i=0}^{n_d-1} \mathbf{f}_{(i+1)d+j}(\boldsymbol{\theta})\mathbf{f}_{(i+1)d+j}^{\mathrm{T}}(\boldsymbol{\theta})\right)\right\}^{-q} \qquad (2.2) \\ & \leq \{n/(n_d d)\}^q d^{-1} \sum_{j=1}^{d} n_d^q \lambda_{\min}^{-q}\left(\sum_{i=0}^{n_d-1} \mathbf{f}_{(i+1)d+j}(\boldsymbol{\theta})\mathbf{f}_{(i+1)d+j}^{\mathrm{T}}(\boldsymbol{\theta})\right), \end{aligned}$$

where the first inequality is ensured by the fact that for symmetric matrices E_1 and E_2, $\lambda_{\min}(E_1 + E_2) \geq \lambda_{\min}(E_1) + \lambda_{\min}(E_2)$, and the second one is ensured by

the convexity of x^{-q}, $x > 0$. As a key step for achieving (2.1), we show, by making use of (C2)–(C4), in Appendix A that there exists a positive integer m, depending only on q, r, k and α, such that for all large n, all $0 \le l \le n_d - m$ and all $1 \le j \le d$,

$$\mathrm{E}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}^{-q}\left(\sum_{i=l}^{l+m-1} \mathbf{f}_{(i+1)d+j}(\boldsymbol{\theta})\mathbf{f}_{(i+1)d+j}^{\mathrm{T}}(\boldsymbol{\theta})\right)\right) \le C_3, \tag{2.3}$$

where C_3 is some positive constant independent of l and j. Let $n_{d,m} = \lfloor n_d/m \rfloor$. Then, analogous to (2.2),

$$\begin{aligned} &n_d^q \lambda_{\min}^{-q}\left(\sum_{i=0}^{n_d-1} \mathbf{f}_{(i+1)d+j}(\boldsymbol{\theta})\mathbf{f}_{(i+1)d+j}^{\mathrm{T}}(\boldsymbol{\theta})\right) \\ &\quad \le (n_d/n_{d,m})^q n_{d,m}^{-1} \sum_{s=0}^{n_{d,m}-1} \lambda_{\min}^{-q}\left(\sum_{i=0}^{m-1} \mathbf{f}_{(i+sm+1)d+j}(\boldsymbol{\theta})\mathbf{f}_{(i+sm+1)d+j}^{\mathrm{T}}(\boldsymbol{\theta})\right). \end{aligned}$$

Combining this fact with (2.2) and (2.3) yields for n large and for some positive number C_4,

$$\begin{aligned} &n^q \mathrm{E}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}^{-q}\left(\sum_{t=1}^{n} \mathbf{f}_t(\boldsymbol{\theta})\mathbf{f}_t^{\mathrm{T}}(\boldsymbol{\theta})\right)\right\} \\ &\quad \le \frac{n^q}{(n_{d,m}d)^q d} \\ &\qquad \times \sum_{j=1}^{d} n_{d,m}^{-1} \sum_{s=0}^{n_{d,m}-1} \mathrm{E}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}^{-q}\left(\sum_{i=0}^{m-1} \mathbf{f}_{(i+sm+1)d+j}(\boldsymbol{\theta})\mathbf{f}_{(i+sm+1)d+j}^{\mathrm{T}}(\boldsymbol{\theta})\right)\right\} \\ &\quad \le C_3 C_4 m^q. \end{aligned}$$

Thus, (2.1) follows. □

To see the extent of the usefulness of (2.1), consider a stochastic regression model of the form

$$y_t = g_t(\boldsymbol{\theta}_0) + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{2.4}$$

where $\{\varepsilon_t\}$ is a martingale difference sequence with respective to $\{\mathcal{G}_t\}$, an increasing sequence of σ-fields on $(\Omega, \mathcal{F}, P)$, such that

$$\sup_t \mathrm{E}(\varepsilon_t^2 | \mathcal{G}_{t-1}) < \infty \qquad \text{a.s.}, \tag{2.5}$$

$g_t(\cdot)$ is a $\mathcal{G}_{t-1}$-measurable random function on a compact set $\boldsymbol{\Theta}_1 \subset R^k$ and $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_1$ is unknown coefficient vector. The least squares estimate $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is obtained by minimizing

$$S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} \big(y_t - g_t(\boldsymbol{\theta})\big)^2 \tag{2.6}$$

over $\mathbf{\Theta}_1$. The next theorem provides a set of sufficient conditions under which

$$\mathrm{E}\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q = O(1), \qquad q \geq 1. \tag{2.7}$$

To state the result, denote the gradient vector and the Hessian matrix of a smooth function $h: R^k \to R$ by $\nabla h(\xi_1, \ldots, \xi_k) = (\partial h/\partial \xi_1, \ldots, \partial h/\partial \xi_k)^{\mathrm{T}}$ and $\nabla^2 h(\xi_1, \ldots, \xi_k) = (\partial^2 h/\partial \xi_i\, \partial \xi_j)_{1 \leq i,j \leq k}$, respectively. For $\boldsymbol{\theta} \in R^k$ and $\eta_1 > 0$, define $B_{\eta_1}(\boldsymbol{\theta}) = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\theta}\| < \eta_1\}$.

THEOREM 2.2. *Consider the stochastic regression model* (2.4) *in which* $g_t(\cdot)$ *is* $\mathcal{G}_{t-1}$*-measurable and continuous on* $\mathbf{\Theta}_1$ *and the martingale difference sequence* $\{\varepsilon_t\}$ *satisfies* (2.5). *Suppose that there exists* $\delta_1 > 0$ *such that* $B_{\delta_1}(\boldsymbol{\theta}_0) \subset \mathbf{\Theta}_1$ *and the gradient vector* ∇g_t *is continuously differentiable on* $B_{\delta_1}(\boldsymbol{\theta}_0)$. *Moreover, assume* $\sup_t \mathrm{E}(|\varepsilon_t|^\gamma | \mathcal{G}_{t-1}) < C_5$ *a.s. for some* $\gamma > \max\{q, 2\}$ *and* $C_5 > 0$, *and the following conditions hold*:

(i) (C2)–(C4) *hold for* $\mathbf{\Theta} = B_{\delta_1}(\boldsymbol{\theta}_0)$, $\mathbf{f}_t(\boldsymbol{\theta}) = \nabla g_t(\boldsymbol{\theta})$ *and* $\mathcal{F}_t = \mathcal{G}_{t-1}$. *In addition, there exists* $q_1 > q$ *such that*

$$\max_{1 \leq i,j \leq k} \mathrm{E}\left(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} \left| n^{-1/2} \sum_{t=1}^{n} \varepsilon_t (\nabla^2 g_t(\boldsymbol{\theta}))_{i,j} \right|^{q_1} \right) = O(1), \tag{2.8}$$

$$\max_{1 \leq i,j \leq k, 1 \leq t \leq n} \mathrm{E}\Big(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} |(\nabla^2 g_t(\boldsymbol{\theta}))_{i,j}|^{4q_1} \Big) = O(1), \tag{2.9}$$

$$\max_{1 \leq t \leq n} \mathrm{E}\Big(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} \|\nabla g_t(\boldsymbol{\theta})\|^{4q_1} \Big) = O(1). \tag{2.10}$$

(ii) *For any* $\delta_2 > 0$ *such that* $\mathbf{\Theta}_1 - B_{\delta_2}(\boldsymbol{\theta}_0)$ *is nonempty*, (C2)–(C4) *hold for* $\mathbf{\Theta} = \mathbf{\Theta}_1 - B_{\delta_2}(\boldsymbol{\theta}_0)$, $\mathbf{f}_t(\boldsymbol{\theta}) = g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)$ *and* $\mathcal{F}_t = \mathcal{G}_{t-1}$. *In addition, there exist* $0 < \nu \leq 1/2$ *and* $q_2 > q/(2\nu)$ *such that*

$$\mathrm{E}\left(\sup_{\boldsymbol{\theta} \in \mathbf{\Theta}_1 - B_{\delta_2}(\boldsymbol{\theta}_0)} \left| n^{-1} \sum_{t=1}^{n} \varepsilon_t \big(g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big) \right|^{q_2} \right) = O(n^{-\nu q_2}). \tag{2.11}$$

(iii) *There exists* $\bar{M} > 0$ *such that*

$$P\left(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} \lambda_{\min}^{-1} \left(n^{-1} \sum_{t=1}^{n} \nabla g_t(\boldsymbol{\theta}) (\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}} \right) > \bar{M} \right) = O(n^{-q}), \tag{2.12}$$

$$P\left(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} n^{-1} \sum_{t=1}^{n} \|\nabla g_t(\boldsymbol{\theta})\|^2 > \bar{M} \right) = O(n^{-q}), \tag{2.13}$$

$$\max_{1 \leq i,j \leq k} P\left(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} n^{-1} \sum_{t=1}^{n} (\nabla^2 g_t(\boldsymbol{\theta}))_{i,j}^2 > \bar{M} \right) = O(n^{-q}). \tag{2.14}$$

Then (2.7) *holds*.

Some comments are in order. Conditions (i) and (iii) are needed to prove that the qth moment of $\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\| I_{A_n}$ is asymptotically bounded in (2.15), where A_n is the event $\hat{\boldsymbol{\theta}}_n$ falls into a small ball around $\boldsymbol{\theta}_0$. Equations (2.9) and (2.10) in condition (i) are similar to Condition 13 of [17], but (2.9) and (2.10) require the existence of higher-order moments of $\nabla g_t(\boldsymbol{\theta})$ and $\nabla^2 g_t(\boldsymbol{\theta})$ to establish inequality (2.26), which plays an important role in deriving (2.15). Equation (2.8) in condition (i) can be viewed as a "moment" counterpart to (3.18) of [14] and can be justified by an argument similar to (3.8) of [14], which shows that the supremum of a Hilbert space (H) valued martingale is dominated by its norm in H under certain smoothness conditions. For more details, see (B.5) and (B.7) of Appendix B. Equations (2.12)–(2.14) in condition (iii) may seem less relevant to the typical assumptions made for LLN and CLT of $\hat{\boldsymbol{\theta}}_n$ at the first sight. However, like (2.9) and (2.10), they are needed for the derivation of (2.26). In fact, (2.12) and (2.13) can be simplified into a single assumption that for any $\bar{m} > 0$,

$$P\left(\sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_0)} \left\| n^{-1}\sum_{t=1}^{n} [\nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}} - \mathrm{E}\{\nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}}\}] \right\| > \bar{m}\right) = O(n^{-q}),$$

where $\|D\|^2 = \sup_{\|\mathbf{x}\|=1} \mathbf{x}^{\mathrm{T}} D^{\mathrm{T}} D\mathbf{x}$ for the matrix D. However, we do not want to complicate the proof of Theorem 2.2 by using this assumption. When $g_t(\boldsymbol{\theta})$ is a linear process with coefficient functions satisfying certain smoothness conditions, (2.12)–(2.14) can be justified based on a *uniform* version of the first moment bound theorem of Findley and Wei [6]. Further details can be found in (B.6) and (B.9)–(B.11) of Appendix B. In contrast to conditions (i) and (iii), condition (ii) is required to prove that the qth moment of $\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\| I_{B_n}$ is asymptotically bounded in (2.27), where B_n denotes the event $\hat{\boldsymbol{\theta}}_n$ falls *outside* a small ball around $\boldsymbol{\theta}_0$. Finally, (C2) in condition (ii) provides an identifiability condition for model (2.4), while (2.11) is a moment counterpart to (3.14) of [14] and can be analogously justified as (2.8).

PROOF OF THEOREM 2.2. Let $0 < \delta_1^* < \min\{\delta_1, 3^{-1}k^{-1}\bar{M}^{-2}\}$ and $A_n = \{\hat{\boldsymbol{\theta}}_n \in B_{\delta_1^*}(\boldsymbol{\theta}_0)\}$. We first show that

$$\mathrm{E}\big(\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n}\big) = O(1). \tag{2.15}$$

By the mean value theorem for vector-valued functions, on the set A_n,

$$\mathbf{0} = \nabla S_n(\hat{\boldsymbol{\theta}}_n) = \nabla S_n(\boldsymbol{\theta}_0) + \left\{\int_0^1 \nabla^2 S_n\big(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)\, dr\right\}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \tag{2.16}$$

where $S_n(\cdot)$ is defined in (2.6) and the integral of a matrix is to be understood component-wise. In view of (2.16) and the identities that $\nabla S_n(\boldsymbol{\theta}) = -2\sum_{t=1}^n (y_t - g_t(\boldsymbol{\theta}))\nabla g_t(\boldsymbol{\theta})$ and $\nabla^2 S_n(\boldsymbol{\theta}) = 2\sum_{t=1}^n \nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}} - 2\sum_{t=1}^n (y_t - g_t(\boldsymbol{\theta})) \times \nabla^2 g_t(\boldsymbol{\theta})$, one has

$$\sum_{t=1}^n \varepsilon_t \nabla g_t(\boldsymbol{\theta}_0) = \big(L(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) - Q(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\big)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \qquad \text{on } A_n, \tag{2.17}$$

where $L(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = \int_0^1 \sum_{t=1}^n \nabla g_t(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0))(\nabla g_t(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)))^{\mathrm{T}}\, dr$ and $Q(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = \int_0^1 \sum_{t=1}^n \{y_t - g_t(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0))\}\nabla^2 g_t(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0))\, dr$. A direct algebraic manipulation leads to

$$\lambda_{\min}(L(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)) \geq \inf_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} \lambda_{\min}\left(\sum_{t=1}^n \nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}}\right) \qquad \text{on } A_n, \tag{2.18}$$

which, together with the continuity of $\nabla g_t(\boldsymbol{\theta})$ on $B_{\delta_1}(\boldsymbol{\theta}_0)$, condition (i) and Theorem 2.1, yields that for any $s \geq 1$,

$$\begin{aligned} &\mathrm{E}(\lambda_{\min}^{-s}(n^{-1}L(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0))I_{A_n}) \\ &\quad \leq \mathrm{E}\left(\sup_{\boldsymbol{\theta} \in B_{\delta_1}(\boldsymbol{\theta}_0)} \lambda_{\min}^{-s}\left(n^{-1}\sum_{t=1}^n \nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}}\right)\right) = O(1). \end{aligned} \tag{2.19}$$

With the help of (2.19), we can assume without loss of generality that $L^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)$ exists on A_n, and hence by (2.17),

$$\begin{aligned} &\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\| I_{A_n} \\ &\quad \leq \|nL^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\| \left\| n^{-1/2}\sum_{t=1}^n \varepsilon_t \nabla g_t(\boldsymbol{\theta}_0)\right\| I_{A_n} \\ &\qquad + \|nL^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\| \left\| \int_0^1 n^{-1/2}\sum_{t=1}^n \varepsilon_t \nabla^2 g_t\big(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)\, dr\right\| \\ &\qquad\quad \times \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| I_{A_n} \\ &\qquad + \|nL^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\| \left\| \int_0^1 n^{-1}\sum_{t=1}^n r(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)^{\mathrm{T}} \nabla g_t(\boldsymbol{\theta}_{t,r}^*) \right. \\ &\qquad\qquad \left. \times \nabla^2 g_t\big(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)\, dr\right\| \\ &\qquad \times \|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\| I_{A_n}, \end{aligned} \tag{2.20}$$

where $\boldsymbol{\theta}_{t,r}^*$ satisfies $\|\boldsymbol{\theta}_{t,r}^* - \boldsymbol{\theta}_0\| \le r\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|$. By the Cauchy–Schwarz inequality and Jensen's inequality, it follows that

$$\begin{aligned}&\left\| \int_0^1 n^{-1/2} \sum_{t=1}^n \varepsilon_t \nabla^2 g_t\big(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)\, dr \right\| \\ &\quad \le k \max_{1\le i,j\le k} \sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_0)} \left| n^{-1/2} \sum_{t=1}^n \varepsilon_t (\nabla^2 g_t(\boldsymbol{\theta}))_{i,j} \right| := k W_n\end{aligned} \tag{2.21}$$

and

$$\begin{aligned}&\left\| \int_0^1 n^{-1} \sum_{t=1}^n r(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)^{\mathrm{T}} \nabla g_t(\boldsymbol{\theta}_{t,r}^*) \nabla^2 g_t\big(\boldsymbol{\theta}_0 + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)\, dr \right\| \\ &\quad \le k\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \left(\sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_0)} n^{-1} \sum_{t=1}^n \|\nabla g_t(\boldsymbol{\theta})\|^2 \right)^{1/2} \\ &\qquad \times \left\{ \max_{1\le i,j\le k} \sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_0)} n^{-1} \sum_{t=1}^n (\nabla^2 g_t(\boldsymbol{\theta}))_{i,j}^2 \right\}^{1/2} \\ &\quad := k\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| R_{1n}^{1/2} R_{2n}^{1/2}.\end{aligned} \tag{2.22}$$

Denoting $\sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_0)} \lambda_{\min}^{-1}(n^{-1}\sum_{t=1}^n \nabla g_t(\boldsymbol{\theta})(\nabla g_t(\boldsymbol{\theta}))^{\mathrm{T}})$ by R_n and combining (2.18) and (2.20)–(2.22), we obtain

$$\begin{aligned}&\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n} \\ &\quad \le 3^q \left\{ R_n^q \left\| n^{-1/2} \sum_{t=1}^n \varepsilon_t \nabla g_t(\boldsymbol{\theta}_0) \right\|^q \right. \\ &\qquad \left. + \delta_1^{*q} k^q R_n^q W_n^q + \delta_1^{*q} k^q R_n^q R_{1n}^{q/2} R_{2n}^{q/2} \|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n} \right\} \\ &\quad := 3^q\{\mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)}\}.\end{aligned} \tag{2.23}$$

Applying (2.8), (2.10), (2.19), $\sup_t \mathrm{E}(|\varepsilon_t|^\gamma | \mathcal{G}_{t-1}) < C_5$ a.s., Hölder's inequality and Lemma 2 of Wei [21], it can be shown that for n large and some positive constants C_1^* and C_2^*,

$$\mathrm{E}(\mathrm{(I)}) \le C_1^* \tag{2.24}$$

and

$$\mathrm{E}(\mathrm{(II)}) \le C_2^*; \tag{2.25}$$

see Appendix A of Chan and Ing [4] for more details. In addition, by making use of (2.9), (2.10) and (2.12)–(2.14), we show in Appendix A that for n large,

$$\mathrm{E}(\mathrm{(III)}) \le C_3^* + C_4^* \mathrm{E}\big(\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n}\big), \tag{2.26}$$

where C_3^* and C_4^* are some positive constants with C_4^* satisfying $0 < C_4^* < 3^{-q}$. Consequently, the desired conclusion (2.15) follows from (2.23)–(2.26).

Letting $B_n = \{\hat{\boldsymbol{\theta}}_n \in \tilde{\boldsymbol{\Theta}}_1 = \boldsymbol{\Theta}_1 - B_{\delta_1^*}(\boldsymbol{\theta}_0)\}$, the rest of the proof aims to show that

$$\mathrm{E}\big(\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{B_n}\big) = O(1), \tag{2.27}$$

which, together with (2.15), yields the desired conclusion (2.7).

Since $\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q \le C_5^* n^{q/2}$ for some $C_5^* > 0$, (2.27) follows immediately once we can show that

$$P(B_n) = O(n^{-q/2}). \tag{2.28}$$

By the continuity of $g_t(\cdot)$ on $\boldsymbol{\Theta}_1$, condition (ii) and Theorem 2.1, one has for any $s \ge 1$,

$$\mathrm{E}\left\{\left[\inf_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_1} n^{-1}\sum_{t=1}^{n}\big(g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big)^2\right]^{-s}\right\} = O(1). \tag{2.29}$$

In addition, it is straightforward to see that

$$\begin{aligned} B_n \subseteq \Bigg\{ & 2\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_1}\left|n^{-1}\sum_{t=1}^{n}\varepsilon_t\big(g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big)\right| \\ & \ge \inf_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_1} n^{-1}\sum_{t=1}^{n}\big(g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big)^2\Bigg\}. \end{aligned} \tag{2.30}$$

Since $q_2 > q/(2\nu)$, there exists $\eta_1 > 0$ such that $q_2 = q(1+\eta_1)/(2\nu)$. By (2.29), (2.30), (2.11), Chebyshev's inequality and Hölder's inequality, there exists $C_6^* > 0$ such that for all large n,

$$\begin{aligned} P(B_n) \le C_6^* & \left\{\mathrm{E}\left(\inf_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_1} n^{-1}\sum_{t=1}^{n}\big(g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big)^2\right)^{-q_2/\eta_1}\right\}^{\eta_1/(1+\eta_1)} \\ & \times\left\{\mathrm{E}\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_1}\left|n^{-1}\sum_{t=1}^{n}\varepsilon_t\big(g_t(\boldsymbol{\theta}) - g_t(\boldsymbol{\theta}_0)\big)\right|^{q_2}\right)\right\}^{1/(1+\eta_1)} = O(n^{-q/2}). \end{aligned}$$

Consequently, (2.28) is established and the theorem is proved. □

As mentioned in the Introduction, (2.7) can be used to examine the asymptotic properties of MSPE of $g_{n+1}(\hat{\boldsymbol{\theta}}_n)$, $\mathrm{E}(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\theta}}_n))^2$, which is also known as the *final prediction error* (FPE) for AR models; see Akaike [1]. To see this, note first that under certain mild conditions such as (2.2) and (2.3) of [14], $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s. If one can further show that

$$n^{1/2}(\nabla g_{n+1}(\boldsymbol{\theta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \Rightarrow \mathbf{H} \tag{2.31}$$

and

$$n\{(\nabla g_{n+1}(\boldsymbol{\theta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\}^2 \text{ is uniformly integrable,} \tag{2.32}$$

where $\Rightarrow$ denotes convergence in distribution and $\mathbf{H}$ is a random variable with $\mathrm{E}(\mathbf{H}^2) < \infty$, then

$$\lim_{n\to\infty} n\mathrm{E}\{(\nabla g_{n+1}(\boldsymbol{\theta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\}^2 = \mathrm{E}(\mathbf{H}^2). \tag{2.33}$$

Once (2.33) is established, it can be linked to $\mathrm{E}(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\theta}}_n))^2$ by means of Taylor's expansion as follows. Note that

$$\begin{aligned} &n\{\mathrm{E}(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\theta}}_n))^2 - \mathrm{E}(\varepsilon_{n+1}^2)\} \\ &\quad = n\mathrm{E}\{(\nabla g_{n+1}(\boldsymbol{\theta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\}^2 + \mathrm{E}(\tilde{R}_n) \to \mathrm{E}(\mathbf{H}^2), \end{aligned} \tag{2.34}$$

provided the remainder term $\tilde{R}_n$ satisfies $\mathrm{E}(\tilde{R}_n) = o(1)$. While (2.31) can be established by means of asymptotic distribution results (see Section 3), (2.7) serves as an important device in establishing (2.32) and $\mathrm{E}(\tilde{R}_n) = o(1)$. If one further assumes that $\mathrm{E}(\varepsilon_t^2) = \sigma^2 > 0$ for all $t > 0$, then (2.34) provides an asymptotic expression for $\mathrm{E}(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\theta}}_n))^2$ as

$$\mathrm{E}(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\theta}}_n))^2 = \sigma^2 + \frac{\mathrm{E}(\mathbf{H}^2)}{n} + o(n^{-1}). \tag{2.35}$$

Although the second term in (2.35) is asymptotically negligible compared to σ^2, $\mathrm{E}(\mathbf{H}^2)$ becomes a key quantity. Utilizing (2.7), one can make use of the asymptotic expression in (2.35), in particular $\mathrm{E}(\mathbf{H}^2)$, to construct optimal model selection criteria; see, for example, Akaike [1], Wei [22] and Findley and Wei [7]. See, also, Section 3 for further discussions.

3. Applications to ARMA models. Let $y_1, \ldots, y_n$ be generated from the stochastic regression model,

$$y_t = g_t(\boldsymbol{\eta}_0) + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{\eta}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,p_1}, \beta_{0,1}, \ldots, \beta_{0,p_2})^{\mathrm{T}}$ is an unknown coefficient vector and $g_t(\boldsymbol{\eta}_0)$ has the ARMA representation

$$g_t(\boldsymbol{\eta}_0) = \alpha_{0,1}y_{t-1} + \cdots + \alpha_{0,p_1}y_{t-p_1} - \beta_{0,1}\varepsilon_{t-1} - \cdots - \beta_{0,p_2}\varepsilon_{t-p_2} \tag{3.2}$$

with the initial conditions $y_t = \varepsilon_t = 0$ for all $t \le 0$. Define

$$\hat{\boldsymbol{\eta}}_n = \arg\min_{\boldsymbol{\eta}\in\Pi} \sum_{t=1}^{n} (y_t - g_t(\boldsymbol{\eta}))^2,$$

where $\Pi \subset R^{p_1+p_2}$ is a compact set that includes $\boldsymbol{\eta}_0$ as an interior point and whose elements $\boldsymbol{\eta} = (\alpha_1, \ldots, \alpha_{p_1}, \beta_1, \ldots, \beta_{p_2})^{\mathrm{T}}$ satisfy the following properties:

$$A_{1,\boldsymbol{\eta}}(z) = 1 - \sum_{j=1}^{p_1} \alpha_j z^j \neq 0,$$
$$A_{2,\boldsymbol{\eta}}(z) = 1 - \sum_{j=1}^{p_2} \beta_j z^j \neq 0 \qquad \text{for all } |z| \leq 1; \tag{3.3}$$

$$A_{1,\boldsymbol{\eta}}(z) \text{ and } A_{2,\boldsymbol{\eta}}(z) \text{ have no common zeros;} \tag{3.4}$$

$$|\alpha_{p_1}| + |\beta_{p_2}| > 0. \tag{3.5}$$

In this section, we apply the results obtained in Section 2 to show that

$$\mathrm{E}\|n^{1/2}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\|^q = O(1), \qquad q \geq 1. \tag{3.6}$$

Applications of (3.6) to the investigation of the MSPE of $g_{n+1}(\hat{\boldsymbol{\eta}}_n)$, $\mathrm{E}(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\eta}}_n))^2$, are also given. It should be mentioned that our initial conditions, $y_t = \varepsilon_t = 0$ for all $t \leq 0$, are made for simplicity of the argument only and all results in this section can be straightforwardly extended to the case where (y_t, ε_t) obey the same assumptions for $t \leq 0$ as for $t > 0$.

Let $\boldsymbol{\eta} \in \Pi$. Define $\varepsilon_t(\boldsymbol{\eta}) = 0$ for $t \leq 0$ and define $\varepsilon_t(\boldsymbol{\eta})$ recursively for $t \geq 1$ by

$$\begin{aligned} \varepsilon_t(\boldsymbol{\eta}) &= y_t - g_t(\boldsymbol{\eta}) \\ &= y_t - \alpha_1 y_{t-1} - \cdots - \alpha_p y_{t-p_1} \\ &\quad + \beta_1 \varepsilon_{t-1}(\boldsymbol{\eta}) + \cdots + \beta_{p_2} \varepsilon_{t-p_2}(\boldsymbol{\eta}), \end{aligned} \tag{3.7}$$

noting that $\varepsilon_t(\boldsymbol{\eta}_0) = \varepsilon_t$. As observed in (2.19) and (2.29) of Section 2, to obtain (3.6), it is crucial to verify that for some $\delta_1 > 0$ with $B_{\delta_1}(\boldsymbol{\eta}_0) \subset \Pi$ and any $s \geq 1$,

$$\mathrm{E}\left\{ \sup_{\boldsymbol{\eta} \in B_{\delta_1}(\boldsymbol{\eta}_0)} \lambda_{\min}^{-s} \left[n^{-1} \sum_{t=1}^{n} \nabla \varepsilon_t(\boldsymbol{\eta}) (\nabla \varepsilon_t(\boldsymbol{\eta}))^{\mathrm{T}} \right] \right\} = O(1); \tag{3.8}$$

and for any $\delta_2 > 0$ with $\tilde{\Pi} = \Pi - B_{\delta_2}(\boldsymbol{\eta}_0) \neq \varnothing$ and any $s \geq 1$,

$$\mathrm{E}\left\{ \sup_{\boldsymbol{\eta} \in \tilde{\Pi}} \left[n^{-1} \sum_{t=1}^{n} \big(\varepsilon_t(\boldsymbol{\eta}) - \varepsilon_t(\boldsymbol{\eta}_0)\big)^2 \right]^{-s} \right\} = O(1). \tag{3.9}$$

Denote the ith component of $\nabla \varepsilon_t(\boldsymbol{\eta})$ by $(\nabla \varepsilon_t(\boldsymbol{\eta}))_i$. Straightforward calculations yield that for $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$,

$$(\nabla \varepsilon_t(\boldsymbol{\eta}))_i = -y_{t-i} + \sum_{s=1}^{p_2} \beta_s (\nabla \varepsilon_{t-s}(\boldsymbol{\eta}))_i, \tag{3.10}$$

$$(\nabla \varepsilon_t(\boldsymbol{\eta}))_{p_1+j} = \varepsilon_{t-j}(\boldsymbol{\eta}) + \sum_{s=1}^{p_2} \beta_s (\nabla \varepsilon_{t-s}(\boldsymbol{\eta}))_{p_1+j}. \tag{3.11}$$

For $j < 0$, let $c_j^{(1)}(\boldsymbol{\eta}) = c_j^{(2)}(\boldsymbol{\eta}) = 0$ and for $j \geq 0$, let $c_j^{(1)}(\boldsymbol{\eta})$ and $c_j^{(2)}(\boldsymbol{\eta})$ satisfy

$$\sum_{j=0}^{\infty} c_j^{(1)}(\boldsymbol{\eta}) z^j = \frac{-A_{2,\boldsymbol{\eta}_0}(z)}{A_{2,\boldsymbol{\eta}}(z) A_{1,\boldsymbol{\eta}_0}(z)}, \qquad \sum_{j=0}^{\infty} c_j^{(2)}(\boldsymbol{\eta}) z^j = \frac{A_{1,\boldsymbol{\eta}}(z) A_{2,\boldsymbol{\eta}_0}(z)}{A_{2,\boldsymbol{\eta}}^2(z) A_{1,\boldsymbol{\eta}_0}(z)}. \tag{3.12}$$

In view of (3.3)–(3.5) and the compactness of Π, there exist positive constants K_1 and K_2 such that for all $j \geq 0$ and $i = 1, 2$,

$$\sup_{\boldsymbol{\eta} \in \Pi} |c_j^{(i)}(\boldsymbol{\eta})| \leq K_1 \exp(-K_2 j). \tag{3.13}$$

Define $b_j^{(l)}(\boldsymbol{\eta}) = c_{j-l}^{(1)}(\boldsymbol{\eta})$, $1 \leq l \leq p_1$, and $b_j^{(p_1+l)}(\boldsymbol{\eta}) = c_{j-l}^{(2)}(\boldsymbol{\eta})$, $1 \leq l \leq p_2$. Then it follows from (3.10)–(3.13) that

$$\nabla \varepsilon_t(\boldsymbol{\eta}) = \left(\sum_{j=1}^{t-1} b_j^{(1)}(\boldsymbol{\eta}) \varepsilon_{t-j}, \ldots, \sum_{j=1}^{t-1} b_j^{(p_1+p_2)}(\boldsymbol{\eta}) \varepsilon_{t-j} \right)^{\mathrm{T}} \tag{3.14}$$

and

$$\max_{1 \leq l \leq p_1+p_2} \sup_{\boldsymbol{\eta} \in \Pi} |b_j^{(l)}(\boldsymbol{\eta})| \leq K_1' \exp(-K_2 j) \qquad \text{for some } K_1' > 0. \tag{3.15}$$

Moreover, one has

$$\varepsilon_t(\boldsymbol{\eta}) - \varepsilon_t(\boldsymbol{\eta}_0) = \sum_{i=1}^{t-1} b_i(\boldsymbol{\eta}) \varepsilon_{t-i}, \tag{3.16}$$

where $b_j(\boldsymbol{\eta})$, $j \geq 1$, satisfy $1 + \sum_{j=1}^{\infty} b_j(\boldsymbol{\eta}) z^j = A_{1,\boldsymbol{\eta}}(z) A_{2,\boldsymbol{\eta}_0}(z) / (A_{2,\boldsymbol{\eta}}(z) \times A_{1,\boldsymbol{\eta}_0}(z))$ and

$$\sup_{\boldsymbol{\eta} \in \Pi} |b_j(\boldsymbol{\eta})| \leq K_3 \exp(-K_4 j) \tag{3.17}$$

for some positive constants K_3 and K_4. The next theorem provides sufficient conditions under which

$$\mathrm{E}\left\{ \sup_{\boldsymbol{\eta} \in \Pi} \lambda_{\min}^{-s} \left[n^{-1} \sum_{t=1}^{n} \nabla \varepsilon_t(\boldsymbol{\eta}) (\nabla \varepsilon_t(\boldsymbol{\eta}))^{\mathrm{T}} \right] \right\} = O(1) \qquad \text{for any } s \geq 1. \tag{3.18}$$

This result leads immediately to (3.8).

THEOREM 3.1. *Assume model* (3.1), *with* $g_t(\cdot)$ *defined in* (3.2) *and* ε_t *being independent random variables satisfying* $\mathrm{E}(\varepsilon_t) = 0$ *and* $\mathrm{E}(\varepsilon_t^2) = \sigma^2$ *for all* $t \geq 1$. *Moreover, assume that there exist positive constants* α_1, ξ *and* M_1 *such that for any* $0 < s_2 - s_1 \leq \xi$,

$$\sup_{1 \leq m \leq t < \infty, \|\mathbf{v}\| = 1} |F_{t,m,\mathbf{v}}(s_2) - F_{t,m,\mathbf{v}}(s_1)| \leq M_1 (s_2 - s_1)^{\alpha_1}, \tag{3.19}$$

where $\mathbf{v} \in R^m$ *and* $F_{t,m,\mathbf{v}}(\cdot)$ *denotes the distribution function of* $\mathbf{v}^{\mathrm{T}}(\varepsilon_t, \ldots, \varepsilon_{t+1-m})^{\mathrm{T}}$. *Then*, (C1)–(C4) *hold for* $\boldsymbol{\Theta} = \Pi$, $\mathbf{f}_t(\boldsymbol{\theta}) = \nabla\varepsilon_t(\boldsymbol{\eta})$ *and* $\mathcal{F}_t = \sigma\{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$, *the* σ*-field generated by* $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$. *Hence, by Theorem* 2.1, (3.18) *follows*.

PROOF. According to (3.3) and (3.12), it is easy to see that $\nabla\varepsilon_t(\boldsymbol{\eta})$ is continuous on Π, and hence (C1) follows. Define $\boldsymbol{\Lambda} = \{\mathbf{a} : \mathbf{a} \in R^{\bar{p}}, \|\mathbf{a}\| = 1\}$, where $\bar{p} = p_1 + p_2$. To show (C2), note first that by (3.3)–(3.5), one has for any $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and $\boldsymbol{\eta} \in \Pi$, there exists $\delta_2 = \delta_2(\boldsymbol{\lambda}, \boldsymbol{\eta}) > 0$ such that for all large t,

$$E(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_t(\boldsymbol{\eta}))^2 > \delta_2. \tag{3.20}$$

In addition, it follows from (3.14) and (3.15) that

$$\mathrm{E}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_t(\boldsymbol{\eta}))^2 \text{ converges to } l(\boldsymbol{\lambda}, \boldsymbol{\eta}) \text{ uniformly on } \boldsymbol{\Lambda} \times \Pi, \tag{3.21}$$

where $l(\boldsymbol{\lambda}, \boldsymbol{\eta})$ is some nonnegative function on $\boldsymbol{\Lambda} \times \Pi$. Moreover, since $\mathrm{E}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_t(\boldsymbol{\eta}))^2$ is continuous on $\boldsymbol{\Lambda} \times \Pi$, uniform convergence implies that $l(\boldsymbol{\lambda}, \boldsymbol{\eta})$ is also continuous on $\boldsymbol{\Lambda} \times \Pi$. By (3.20) and the compactness of $\boldsymbol{\Lambda} \times \Pi$, $\inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}, \boldsymbol{\eta}\in\Pi} l(\boldsymbol{\lambda}, \boldsymbol{\eta}) > 0$. This, together with (3.21), yields that there is a positive number ϵ and a positive integer L such that for all $t > L$,

$$\inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}, \boldsymbol{\eta}\in\Pi} \mathrm{E}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_t(\boldsymbol{\eta}))^2 > \epsilon > 0. \tag{3.22}$$

For $t > l_1 \geq 1$, define $\nabla\varepsilon_{t,l_1}(\boldsymbol{\eta}) = (\sum_{i=1}^{l_1} b_i^{(1)}(\boldsymbol{\eta})\varepsilon_{t-i}, \ldots, \sum_{i=1}^{l_1} b_i^{(\bar{p})}(\boldsymbol{\eta})\varepsilon_{t-i})^{\mathrm{T}}$. According to (3.14) and (3.15), there exists a positive integer $L_1(\epsilon)$ such that for all $t > l_1 \geq L_1(\epsilon)$,

$$\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}, \boldsymbol{\eta}\in\Pi} |\mathrm{E}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_t(\boldsymbol{\eta}))^2 - \mathrm{E}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_{t,l_1}(\boldsymbol{\eta}))^2| < \epsilon/2. \tag{3.23}$$

From (3.22) and (3.23), it follows that for all $t > d_1 = \max\{L, L_1(\epsilon)\}$,

$$\inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}, \boldsymbol{\eta}\in\Pi} \mathrm{E}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_{t,d_1}(\boldsymbol{\eta}))^2 > \epsilon/2. \tag{3.24}$$

Denote $\boldsymbol{\lambda}^{\mathrm{T}}(\nabla\varepsilon_{t,d_1}(\boldsymbol{\eta}) - \nabla\varepsilon_t(\boldsymbol{\eta}))$ by $R_t(\boldsymbol{\lambda}, \boldsymbol{\eta})$ and $\sigma^{-1}(\mathrm{var}(\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_{t,d_1}(\boldsymbol{\eta})))^{1/2}$ by $g_t(\sigma, \boldsymbol{\lambda}, \boldsymbol{\eta})$. Since $\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_{t,d_1}(\boldsymbol{\eta})/g_t(\sigma, \boldsymbol{\lambda}, \boldsymbol{\eta})$ can be written as $\sum_{j=1}^{d_1} c_j\varepsilon_{t-j}$ with $\sum_{j=1}^{d_1} c_j^2 = 1$, (3.19) and (3.24) imply that for any $\boldsymbol{\lambda} \times \boldsymbol{\eta} \in \boldsymbol{\Lambda} \times \Pi$ and $t > d_1$,

$$\begin{aligned}
&P\big(s_1 < \boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_t(\boldsymbol{\eta}) \leq s_2 | \mathcal{F}_{t-d_1}\big) \\
&\quad = P\big(s_1 + R_t(\boldsymbol{\lambda}, \boldsymbol{\eta}) < \boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_{t,d_1}(\boldsymbol{\eta}) \leq s_2 + R_t(\boldsymbol{\lambda}, \boldsymbol{\eta}) | \mathcal{F}_{t-d_1}\big) \\
&\quad = P\left(\frac{s_1 + R_t(\boldsymbol{\lambda}, \boldsymbol{\eta})}{g_t(\sigma, \boldsymbol{\lambda}, \boldsymbol{\eta})} < \frac{\boldsymbol{\lambda}^{\mathrm{T}}\nabla\varepsilon_{t,d_1}(\boldsymbol{\eta})}{g_t(\sigma, \boldsymbol{\lambda}, \boldsymbol{\eta})} \leq \frac{s_2 + R_t(\boldsymbol{\lambda}, \boldsymbol{\eta})}{g_t(\sigma, \boldsymbol{\lambda}, \boldsymbol{\eta})} \Big| \mathcal{F}_{t-d_1}\right) \\
&\quad \leq M_1\left(\frac{\sigma(s_2 - s_1)}{\sqrt{\epsilon/2}}\right)^{\alpha_1} \qquad \text{a.s.},
\end{aligned} \tag{3.25}$$

provided $0 < s_2 - s_1 \le (\xi\sqrt{\epsilon/2})/\sigma$. In view of (3.25), (C2) holds with $d = d_1$, $M = M_1(\sigma\sqrt{2/\epsilon})^{\alpha_1}$, $\alpha = \alpha_1$ and $\delta = (\xi\sqrt{\epsilon/2})/\sigma$.

On the other hand, it is shown in Appendix B that there exists $\tau^{**} > 0$ such that for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Pi$, with $\|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\| < \tau^{**}$,

$$\|\nabla\varepsilon_t(\boldsymbol{\eta}_2) - \nabla\varepsilon_t(\boldsymbol{\eta}_1)\| \le \|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\|\tilde{B}_t, \tag{3.26}$$

where $\tilde{B}_t$ are nonnegative random variables satisfying

$$\sup_{t\ge 1} \mathrm{E}(\tilde{B}_t^2) < \infty. \tag{3.27}$$

Combining (3.26) and (3.27), we obtain (C3). Finally, the proof is completed by noting that (C4) is an immediate consequence of (3.26), (3.27), (3.14), (3.15) and the compactness of Π. $\square$

REMARK 1. In the proof of Theorem 3.1, (3.19) plays the same role as that of (C2) in the proof of Theorem 2.1. When ε_t's are normally distributed, (3.19) is satisfied with $M_1 = (2\pi\sigma^2)^{-1/2}$, $\alpha_1 = 1$ and any $\xi > 0$. In addition, when ε_t's are i.i.d. with an integrable characteristic function, (3.19) is satisfied with any $\xi > 0$, $\alpha_1 = 1$ and some $M_1 > 0$. For more details, see Lemma 4 of [11]. An extension of Theorem 3.1 to autoregressive fractionally integrated moving average models (ARFIMA) has also been obtained by the authors. However, since the proof of this extension is quite involved, the details will be reported elsewhere.

THEOREM 3.2. *Under the same assumptions as in Theorem* 3.1, (C1)–(C4) *hold for* $\mathbf{f}_t(\boldsymbol{\theta}) = \varepsilon_t(\boldsymbol{\eta}) - \varepsilon_t(\boldsymbol{\eta}_0)$, $\boldsymbol{\Theta} = \tilde{\Pi}$ *and* $\mathcal{F}_t = \sigma\{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$, *and hence by Theorem* 2.1, (3.9) *follows*.

The proof of Theorem 3.2 is omitted, since it is similar to the proof of Theorem 3.1. Using Theorems 2.2, 3.1 and 3.2 and Lemma B.1 of Appendix B, the next theorem, whose proof is deferred to Appendix B, establishes moment bounds for $n^{1/2}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)$.

THEOREM 3.3. *Assume that the assumptions of Theorem* 3.1 *hold and for some* $q_1 > q \ge 1$,

$$\sup_{t\ge 1} \mathrm{E}|\varepsilon_t|^{4q_1} < \infty. \tag{3.28}$$

Then, (3.6) *follows*.

As an application of Theorem 3.3, an asymptotic expression for the MSPE of $\hat{\boldsymbol{\eta}}_n$, $\mathrm{E}\{y_{n+1} - g_{n+1}(\hat{\boldsymbol{\eta}}_n)\}^2$, is given in Theorem 3.4 below.

THEOREM 3.4. *Assume that the assumptions of Theorem 3.1 hold. Moreover, let ε_t be i.i.d. random variables satisfying for some $q_1 > 18$,*

$$\mathrm{E}|\varepsilon_1|^{q_1} < \infty. \tag{3.29}$$

Then,

$$\lim_{n\to\infty} n[\mathrm{E}\{y_{n+1} - g_{n+1}(\hat{\boldsymbol{\eta}}_n)\}^2 - \sigma^2] = \bar{p}\sigma^2. \tag{3.30}$$

PROOF. Let δ_1 be any positive number such that $B_{\delta_1}(\boldsymbol{\eta}_0) \subset \Pi$ and define $A_n = \{\hat{\boldsymbol{\eta}}_n \in B_{\delta_1}(\boldsymbol{\eta}_0)\}$ and $A_n^c = \{\hat{\boldsymbol{\eta}}_n \in \tilde{\Pi} = \Pi - B_{\delta_1}(\boldsymbol{\eta}_0)\}$. By Taylor's theorem,

$$\begin{aligned} & n^{1/2}\big(y_{n+1} - g_{n+1}(\hat{\boldsymbol{\eta}}_n) - \varepsilon_{n+1}\big) \\ &\quad = n^{1/2}(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)I_{A_n} \\ &\qquad + \frac{n^{1/2}}{2}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)^{\mathrm{T}}\nabla^2\varepsilon_{n+1}(\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)I_{A_n} \\ &\qquad + n^{1/2}\big(\varepsilon_{n+1}(\hat{\boldsymbol{\eta}}_n) - \varepsilon_{n+1}(\boldsymbol{\eta}_0)\big)I_{A_n^c} \\ &\quad := \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)}, \end{aligned} \tag{3.31}$$

where $\|\boldsymbol{\eta}^* - \boldsymbol{\eta}_0\| \le \|\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0\|$. In view of (3.31), (3.30) holds immediately if one can show that

$$\lim_{n\to\infty} \mathrm{E(I)}^2 = \bar{p}\sigma^2, \tag{3.32}$$

$$\lim_{n\to\infty} \mathrm{E(II)}^2 = 0, \tag{3.33}$$

$$\lim_{n\to\infty} \mathrm{E(III)}^2 = 0. \tag{3.34}$$

By utilizing the martingale CLT (cf. [9]) and a truncation argument in [10], it can be shown that

$$n^{1/2}\{(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\}I_{A_n} \Rightarrow \mathbf{F}^{\mathrm{T}}\mathbf{Q}, \tag{3.35}$$

where $\mathbf{Q}$ is distributed as $N(\mathbf{0}, \sigma^2\Gamma^{-1})$ with $\Gamma = \lim_{t\to\infty} \mathrm{E}\{\nabla\varepsilon_t(\boldsymbol{\eta}_0)(\nabla\varepsilon_t(\boldsymbol{\eta}_0))^{\mathrm{T}}\}$, and $\mathbf{F}$, satisfying $\mathrm{E}(\mathbf{F}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{F}\mathbf{F}^{\mathrm{T}}) = \Gamma$, is independent of $\mathbf{Q}$. Let $2 < r \le 18/5$. Then, it follows from Hölder's inequality, Theorem 3.3, (3.15) and (3.29) that

$$\begin{aligned} & \mathrm{E}\{|n^{1/2}(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)|^r\} \\ &\quad \le \mathrm{E}\{\|n^{1/2}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\|^r \|\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\|^r\} \\ &\quad \le \big(\mathrm{E}\|n^{1/2}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\|^{5r/4}\big)^{4/5}(\mathrm{E}\|\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0)\|^{5r})^{1/5} = O(1), \end{aligned}$$

which implies the uniform integrability of $n\{(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\}^2 I_{A_n}$. Combing this with (3.35) yields

$$\lim_{n\to\infty} \mathrm{E}[n\{(\nabla\varepsilon_{n+1}(\boldsymbol{\eta}_0))^{\mathrm{T}}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\}^2 I_{A_n}] = \mathrm{E}(\mathbf{F}^{\mathrm{T}}\mathbf{Q})^2 = \bar{p}\sigma^2,$$

and hence (3.32) follows. Moreover, applying Theorems 3.2 and 3.3, (3.29) and an argument similar to that used to prove (B.8) and (B.12) of Appendix B, it is shown in Appendix B of [4] that (3.33) and (3.34) are also true. Consequently, the desired conclusion (3.30) holds. □

REMARK 2. Note that the moment restriction (3.29) is stronger than necessary for the proofs of (3.32) and (3.34). On the other hand, since (3.33) requires that $E\|n^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)\|^q = O(1)$ holds with $q = 9/2$ (see Appendix B of [4]), it seems that one cannot easily weaken (3.29) because Theorem 3.3 constitutes a key tool in verifying (3.33).

In the special case of $p_2 = 0$ (the pure AR case), equation (3.30) was examined by Fuller and Hasza [8], Kunitomo and Yamamoto [13] and Ing [10]. In addition, for the case $p_2 > 0$, equation (3.30) was also considered in Yamamoto [23], but a rigorous proof of (3.30) is still lacking in the literature. By establishing a set of uniform moment bounds, this paper offers a rigorous proof of (3.30) for the ARMA case.

Equation (3.30) implies that when two competing ARMA models are entertained, the one having fewer estimated parameters also possesses a smaller MSPE, up to terms of order n^{-1}. As a result, the principle of parsimony (e.g., Tukey [20]), which roughly asserts that mathematical models with the smallest number of parameters are preferred, is now endowed with a precise meaning in the context of ARMA modelling. When $p_2 = 0$, (3.30) was established in Akaike [1] using an ad-hoc argument, which immediately led him to develop the final prediction error criterion,

$$\frac{n + \bar{p}}{(n - \bar{p})n} \sum_{t=1}^{n} \bigl(y_t - g_t(\hat{\boldsymbol{\eta}}_n)\bigr)^2$$

that is commonly used for AR model selection with optimal prediction efficiency; see Shibata [18] or Ing and Wei [12]. Under this perspective, a contribution of (3.30) is that it provides a theoretical foundation for the construction of the FPE criterion for ARMA models. The issue of whether the FPE criterion (or its variants) is asymptotically efficient (in the sense of [12] or [18]) in ARMA model selection still remains open, however.

As a final remark, we note that (3.30) is obtained based on Theorems 2.1 and 2.2. Moreover, since these theorems provide a useful device for exploring the moment properties of least squares estimates in (nonlinear) stochastic regression models, their applications to prediction or model selection in models beyond the ARMA case are anticipated.

APPENDIX A: PROOFS OF (2.3) AND (2.26)

PROOF OF (2.3). Let $m = \lfloor \{l_1(r + 2k) + r + k + 2q\}/\alpha \rfloor + 1$ with $l_1 > q$. We only prove (2.3) for the case of $l = 0$ and $j = 1$ since the other cases can be similarly verified. First, define $A(u) = \{\sum_{i=0}^{m-1} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\|^2 \le u^{l_1/q}/r\}$ and

$B(u) = \{\sum_{i=0}^{m-1} B_{(i+1)d+1} \le u^{l_1/q}/k^{1/2}\}$, where B_t are random variables defined in (C3). Then, the left-hand side of (2.3) (with $l = 0$ and $j = 1$) is bounded by

$$
\begin{aligned}
&K_0 + \int_{K_0}^{\infty} P\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left(\inf_{\|\mathbf{y}\|=1}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2\right)^{-q} > u\right\} du \\
&\quad = K_0 + \int_{K_0}^{\infty} P\left\{\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{y}\|=1}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2 < u^{-1/q}\right\} du \\
&\quad \le K_0 + \int_{K_0}^{\infty} P\left\{\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{y}\|=1}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2 < u^{-1/q}, A(u), B(u)\right\} du \qquad \text{(A.1)} \\
&\qquad + \int_{K_0}^{\infty} P(A^c(u))\,du + \int_{K_0}^{\infty} P(B^c(u))\,du \\
&\quad \equiv K_0 + \text{(I)} + \text{(II)} + \text{(III)},
\end{aligned}
$$

where $K_0 = K_0(l_1, \delta, q, k, \tau)$ is a positive number to be specified later and $A^c(u)$ and $B^c(u)$ denote the complements of $A(u)$ and $B(u)$, respectively. Since $l_1 > q$, by (C3), (C4) and Chebyshev's inequality, it follows that for n large,

$$
\text{(II)} \le C_1^* \quad \text{and} \quad \text{(III)} \le C_2^*, \tag{A.2}
$$

where C_1^* and C_2^* are some positive constants depending on $C_1, C_2, \alpha, l_1, r, k, q$ and K_0.

To deal with (I), consider the hypersphere $\mathbf{S}_r = \{\mathbf{y} : \mathbf{y} \in R^r, \|\mathbf{y}\| = 1\}$ and the hypercube $\mathbf{H}^r(u) = [1 - 2u^{-(l_1+1)/2q}(\lfloor u^{(l_1+1)/2q}\rfloor + 1), 1]^r$, $u > 0$. Note first that $\mathbf{S}_r \subseteq \mathbf{H}^r(u)$ for any $u > 0$. Divide $\mathbf{H}^r(u)$ into sub-hypercubes of equal size, each of which has an edge length of $2u^{-(l_1+1)/2q}$ and a circumscribed circle of radius $\sqrt{r}u^{-(l_1+1)/2q}$. Denote these sub-hypercubes by $\tilde{B}_i(u)$, $1 \le i \le m^* = (\lfloor u^{(l_1+1)/2q}\rfloor + 1)^r$. Letting $G_i(u) = \mathbf{S}_r \cap \tilde{B}_i(u)$ and $\{G_{v_i}(u), i = 1, \ldots, m^{**}\}$ denote the collection of nonempty $G_i(u)$'s, it follows that $\mathbf{S}_r = \bigcup_{i=1}^{m^{**}} G_{v_i}(u)$ with $m^{**} \le (\lfloor u^{(l_1+1)/2q}\rfloor + 1)^r$. On the other hand, since $\boldsymbol{\Theta}$ is a bounded subset in R^k, there is a positive integer g such that for any $u > 0$, $\boldsymbol{\Theta} \subseteq \mathbf{H}_g^k(u) = [g - 2gu^{-(l_1+1/2)/q}(\lfloor u^{(l_1+1/2)/q}\rfloor + 1), g]^k$. We can similarly divide $\mathbf{H}_g^k(u)$ into equal-sized sub-hypercubes $\tilde{W}_i(u), i = 1, \ldots, e^*$, where the edge length of $\tilde{W}_i(u)$ is $2u^{-(l_1+1/2)q^{-1}}$ and $e^* = g^k(\lfloor u^{(l_1+1/2)/q}\rfloor + 1)^k$. In addition, it holds that $\boldsymbol{\Theta} = \bigcup_{i=1}^{e^{**}} J_{v_i}(u)$, where with $J_i(u) = \boldsymbol{\Theta} \cap \tilde{W}_i(u)$, $\{J_{v_i}(u), i = 1, \ldots, e^{**}\}$ denotes the collection of nonempty $J_i(u)$'s. By observing

$$
\begin{aligned}
&\left\{\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{y}\|=1}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2 < u^{-1/q}\right\} \\
&\quad = \bigcup_{s=1}^{e^{**}}\bigcup_{j=1}^{m^{**}}\left\{\inf_{\boldsymbol{\theta}\in J_{v_s}(u)}\inf_{\mathbf{y}\in G_{v_j}(u)}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2 < u^{-1/q}\right\},
\end{aligned}
$$

one has

$$P\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{y}\|=1}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2 < u^{-1/q}, A(u), B(u)\right)$$
$$\le \sum_{s=1}^{e^{**}}\sum_{j=1}^{m^{**}} P\left(\bigcap_{i=0}^{m-1} C_i^{(s,j)}(u)\right), \tag{A.3}$$

where

$$C_i^{(s,j)}(u) = \left\{\inf_{\boldsymbol{\theta}\in J_{v_s}(u)}\inf_{\mathbf{y}\in G_{v_j}(u)}\left|\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right| < u^{-1/(2q)},\right.$$
$$\left. B_{(i+1)d+1} \le \frac{u^{l_1/q}}{k^{1/2}}, \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right\| \le \frac{u^{l_1/2q}}{r^{1/2}}\right\}.$$

Let $\mathbf{y}_j \in G_{v_j}(u)$, $j = 1, \ldots, m^{**}$, and $\boldsymbol{\theta}_s \in J_{v_s}(u)$, $s = 1, \ldots, e^{**}$, be arbitrarily chosen. Then, for any $\mathbf{y} \in G_{v_j}(u)$ and $\boldsymbol{\theta} \in J_{v_s}(u)$,

$$\begin{aligned}\left|\mathbf{y}_j^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta}_s)\right| &\le \|\mathbf{y}_j - \mathbf{y}\|\left\|\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta}_s)\right\| \\ &\quad + \|\mathbf{y}\|\left\|\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta}_s) - \mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right\| \\ &\quad + \left|\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right|.\end{aligned}$$

Combining this with (C3) yields that on the set $C_i^{(s,j)}(u)$ with $u > (2k^{1/2}/\tau)^{q/(l_1+1/2)}$,

$$\begin{aligned}&\left|\mathbf{y}_j^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta}_s)\right| \\ &\quad\le 2\sqrt{r}u^{-(l_1+1)/2q}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right\| \\ &\qquad + 2\sqrt{k}u^{-(l_1+1/2)/q}B_{(i+1)d+1} + \inf_{\boldsymbol{\theta}\in J_{v_s}(u)}\inf_{\mathbf{y}\in G_{v_j}(u)}\left|\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right| \\ &\quad\le 5u^{-1/2q}\end{aligned}$$

and hence

$$C_i^{(s,j)}(u) \subseteq D_i^{(s,j)}(u) := \left\{\left|\mathbf{y}_j^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta}_s)\right| \le 5u^{-1/2q}\right\}. \tag{A.4}$$

In view of (A.3) and (A.4), it follows that for $u > (2k^{1/2}/\tau)^{q/(l_1+1/2)}$,

$$P\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\inf_{\|\mathbf{y}\|=1}\sum_{i=0}^{m-1}\left(\mathbf{y}^{\mathrm{T}}\mathbf{f}_{(i+1)d+1}(\boldsymbol{\theta})\right)^2 < u^{-1/q}, A(u), B(u)\right)$$
$$\le \sum_{s=1}^{e^{**}}\sum_{j=1}^{m^{**}} P\left(\bigcap_{i=0}^{m-1} D_i^{(s,j)}(u)\right). \tag{A.5}$$

Observe that

$$P\left(\bigcap_{i=0}^{m-1} D_i^{(s,j)}(u)\right) = \mathrm{E}\left\{\prod_{i=0}^{m-2} I_{D_i^{(s,j)}(u)} P\big(D_{m-1}^{(s,j)}(u)|\mathcal{F}_{(m-1)d+1}\big)\right\},$$

where $I_{D_i^{(s,j)}(u)}$ denotes the indicator function of the set $D_i^{(s,j)}(u)$. This, together with (C2), implies that for $u > (10/\delta)^{2q}$, all $1 \le s \le e^{**}$, all $1 \le j \le m^{**}$ and n large,

$$P\left(\bigcap_{i=0}^{m-1} D_i^{(s,j)}(u)\right) \le M(10)^{\alpha} u^{-\alpha/2q} \mathrm{E}\left\{\prod_{i=0}^{m-2} I_{D_i^{(s,j)}(u)}\right\}.$$

Repeating the same argument $m-1$ times, one has

$$P\left(\bigcap_{i=0}^{m-1} D_i^{(s,j)}(u)\right) \le M^m (10)^{m\alpha} u^{-m\alpha/2q}. \tag{A.6}$$

Taking $K_0 > \max\{(10/\delta)^{2q}, (2k^{1/2}/\tau)^{q/(l_1+1/2)}, 1\}$, it follows from (A.5), (A.6) and $m > \{l_1(r+2k) + r + k + 2q\}/\alpha$ that

$$\begin{aligned}
\text{(I)} &\le \int_{K_0}^{\infty} \sum_{s=1}^{e^{**}} \sum_{j=1}^{m^{**}} P\left(\bigcap_{i=0}^{m-1} D_i^{(s,j)}(u)\right) du \\
&\le 2^{r+k} g^k M^m (10)^{\alpha m} \int_{K_0}^{\infty} u^{-\{1/(2q)\}\{\alpha m - (l_1+1)r - (2l_1+1)k\}} du \qquad \text{(A.7)} \\
&= 2^{r+k} g^k M^m (10)^{\alpha m} \{C(q,\alpha,m,l_1,r,k)\}^{-1} K_0^{-C(q,\alpha,m,l_1,r,k)},
\end{aligned}$$

where $C(q,\alpha,m,l_1,r,k) = \{\alpha m - (l_1+1)r - (2l_1+1)k - 2q\}/2q$. Consequently, (2.3) is ensured by (A.1), (A.2) and (A.7). □

PROOF OF (2.26). Let $C_4^* = \delta_1^{*q} k^q \bar{M}^{2q}$. Since δ_1^*, defined at the beginning of the proof of Theorem 2.2, is smaller than $3^{-1}k^{-1}\bar{M}^{-2}$, it follows that $C_4^* < 3^{-q}$. By the Cauchy–Schwarz inequality and (2.12)–(2.14), one has

$$\begin{aligned}
\mathrm{E(III)} &\le \delta_1^{*2q} k^q n^{q/2} \mathrm{E}\big(R_n^q R_{1n}^{q/2} R_{2n}^{q/2} I_{\{R_n R_{1n}^{1/2} R_{2n}^{1/2} > \bar{M}^2\}}\big) \\
&\quad + C_4^* \mathrm{E}\big(\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n}\big) \\
&\le \delta_1^{*2q} k^q n^{q/2} \{\mathrm{E}(R_n^{2q} R_{1n}^q R_{2n}^q)\}^{1/2} \\
&\quad \times \{P(R_n > \bar{M}) + P(R_{1n} > \bar{M}) + P(R_{2n} > \bar{M})\}^{1/2} \qquad \text{(A.8)} \\
&\quad + C_4^* \mathrm{E}\big(\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n}\big) \\
&= O(1)\{\mathrm{E}(R_n^{2q} R_{1n}^q R_{2n}^q)\}^{1/2} + C_4^* \mathrm{E}\big(\|n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|^q I_{A_n}\big).
\end{aligned}$$

In addition, $\mathrm{E}(R_n^{2q} R_{1n}^q R_{2n}^q) = O(1)$ follows from Hölder's inequality, (2.9), (2.10) and (2.19). Combining this with (A.8) yields (2.26). □

APPENDIX B: PROOFS OF (3.26), (3.27) AND THEOREM 3.3

Throughout this Appendix, $\mathbf{J}(m, \bar{p})$, $1 \le m \le \bar{p}$, denotes the set $\{(j_1, \ldots, j_m)$: $j_1 < \cdots < j_m$, $j_i \in \{1, \ldots, \bar{p}\}$ for $1 \le i \le m\}$, and for $\mathbf{j} = (j_1, \ldots, j_m) \in \mathbf{J}(m, \bar{p})$ and smooth function $w = w(\boldsymbol{\xi}) = w(\xi_1, \ldots, \xi_{\bar{p}})$, $\mathbf{D}_{\mathbf{j}} w$ denotes the partial derivative $\partial^m w / \partial \xi_{j_1}, \ldots, \partial \xi_{j_m}$. Before proving (3.26) and (3.27), we note that according to (3.3)–(3.5), (3.10)–(3.14) and the compactness of Π, $(\nabla^2 \varepsilon_t(\boldsymbol{\eta}))_{i,j} = \sum_{s=1}^{t-2} c_{s,ij}(\boldsymbol{\eta}) \varepsilon_{t-1-s}$, where $c_{s,ij}(\boldsymbol{\eta})$ are continuously differentiable on Π and satisfy, for some $D_1, D_2 > 0$ (independent of i, j and s),

$$\sup_{\boldsymbol{\eta} \in \Pi} |c_{s,ij}(\boldsymbol{\eta})| \le D_1 \exp(-D_2 s). \tag{B.1}$$

Moreover, there exists a small positive number τ^* such that

$$\sup_{\boldsymbol{\eta} \in \Pi^*} |\mathbf{D}_{\mathbf{j}} b_s(\boldsymbol{\eta})| \le D_3 \exp(-D_6 s), \tag{B.2}$$

$$\max_{\mathbf{j} \in \mathbf{J}(m, \bar{p}), 1 \le m \le \bar{p}} \sup_{\boldsymbol{\eta} \in \Pi^*} \left|\mathbf{D}_{\mathbf{j}} b_s^{(l)}(\boldsymbol{\eta})\right| \le D_4 \exp(-D_6 s), \tag{B.3}$$

$$\max_{\mathbf{j} \in \mathbf{J}(m, \bar{p}), 1 \le m \le \bar{p}} \sup_{\boldsymbol{\eta} \in \Pi^*} |\mathbf{D}_{\mathbf{j}} c_{s,ij}(\boldsymbol{\eta})| \le D_5 \exp(-D_6 s), \tag{B.4}$$

where $\Pi^* = \bigcup_{\boldsymbol{\eta} \in \Pi} B_{\tau^*}(\boldsymbol{\eta})$ and $D_3, \ldots, D_6$ are some positive constants independent of i, j, l and s.

PROOFS OF (3.26) AND (3.27). Let $\tau^{**} = \tau^*/2$. For $\|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\| < \tau^{**}$, it follows from the mean value theorem for vector-valued functions that $\|\nabla \varepsilon_t(\boldsymbol{\eta}_2) - \nabla \varepsilon_t(\boldsymbol{\eta}_1)\|^2 \le \|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\|^2 \| \int_0^1 \nabla^2 \varepsilon_t(\boldsymbol{\eta}_1 + v(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1))\, dv\|^2 \le \|\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\|^2 (\tilde{B}_t)^2$, where $\tilde{B}_t = \{\sum_{1 \le i,j \le \bar{p}} \sup_{\boldsymbol{\eta} \in \Pi^{**}} (\nabla^2 \varepsilon_t(\boldsymbol{\eta}))_{i,j}^2\}^{1/2}$, with $\Pi^{**} = \bigcup_{\boldsymbol{\eta} \in \Pi} B_{\tau^{**}}(\boldsymbol{\eta})$. Denoting by $\bar{\Pi}^{**}$ the compact closure of Π^{**}, one has $\bar{\Pi}^{**} \subset \Pi^*$, which further yields $\bar{\Pi}^{**} \subset \bigcup_{r=1}^{\bar{r}} B_{\tau^*}(\boldsymbol{\theta}_r)$, for some $1 \le \bar{r} < \infty$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{\bar{r}} \in \Pi$. Hence, $\mathrm{E}(\tilde{B}_t^2) \le \sum_{1 \le i,j \le \bar{p}} \sum_{r=1}^{\bar{r}} \mathrm{E}\{\sup_{\boldsymbol{\eta} \in B_{\tau^*}(\boldsymbol{\theta}_r)} (\nabla^2 \varepsilon_t(\boldsymbol{\eta}))_{i,j}^2\}$. Moreover, it follows from (B.1), (B.4) and (3.10) of Lai [14] that for all $1 \le i, j \le \bar{p}$, $1 \le r \le \bar{r}$ and $t \ge 3$, $\mathrm{E}\{\sup_{\boldsymbol{\eta} \in B_{\tau^*}(\boldsymbol{\theta}_r)} (\nabla^2 \varepsilon_t(\boldsymbol{\eta}))_{i,j}^2\} < C \sum_{s=1}^{\infty} \{\exp(-2D_2 s) + \exp(-2D_6 s)\}$ for some $C > 0$ (see Appendix B of [4] for more details). Consequently, (3.26) and (3.27) follow. □

The next lemma, Lemma B.1, provides moment bounds for the supremums of some random functions associated with (2.8) and (2.11)–(2.14). Lemma B.1, together with Theorems 3.1 and 3.2, constitutes the major tools for proving Theorem 3.3.

LEMMA B.1. *Let $\boldsymbol{\theta}_a$ be some point in R^k, $k \ge 1$, and δ_1 be some positive number. For $t \ge 2$, define $K_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1} c_i(\boldsymbol{\theta}) \epsilon_{t-i}$ and $Q_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1} d_i(\boldsymbol{\theta}) \epsilon_{t-i}$,*

where ϵ_i are independent random variables with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{E}(\epsilon_i^2) = \sigma_\epsilon^2 > 0$ for all $i \geq 1$, and $c_i(\boldsymbol{\theta})$ and $d_i(\boldsymbol{\theta})$ are real-valued functions on $B_{\delta_1}(\boldsymbol{\theta}_a)$. Assume that for any $i \geq 1$, $\mathbf{j} \in \mathbf{J}(m,k)$ and $1 \leq m \leq k$, $\mathbf{D}_\mathbf{j} c_i(\boldsymbol{\theta})$ are continuous on $B_{\delta_1}(\boldsymbol{\theta}_a)$, and for some $q_1 \geq 2$, $\sup_{i\geq 1} \mathrm{E}|\epsilon_i|^{q_1} < \infty$. Then, there exists $C > 0$ such that for all $n \geq 2$,

$$\begin{aligned}
&\mathrm{E}\left(\sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_a)} \left| \sum_{t=2}^{n} K_t(\boldsymbol{\theta})\epsilon_t \right|^{q_1} \right) \\
&\quad \leq C n^{q_1/2} \Bigg[\left\{ \sum_{i=1}^{n-1} c_i^2(\boldsymbol{\theta}_a) \right\}^{q_1/2} \\
&\qquad + \left\{ \sum_{i=1}^{n-1} \max_{\mathbf{j}\in\mathbf{J}(m,k),1\leq m\leq k} \sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_a)} (\mathbf{D}_\mathbf{j} c_i(\boldsymbol{\theta}))^2 \right\}^{q_1/2} \Bigg].
\end{aligned} \tag{B.5}$$

Moreover, if for any $i, j \geq 1$, $\mathbf{j} \in \mathbf{J}(m,k)$ and $1 \leq m \leq k$, $\mathbf{D}_\mathbf{j}\{c_i(\boldsymbol{\theta}) d_j(\boldsymbol{\theta})\}$ are continuous on $B_{\delta_1}(\boldsymbol{\theta}_0)$, and for some $q_1 \geq 2$, $\sup_{i\geq 1} \mathrm{E}|\epsilon_i|^{2q_1} < \infty$, then there exists $C > 0$ such that for all $n \geq 3$,

$$\begin{aligned}
&\mathrm{E}\left(\sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_a)} \left| \sum_{t=2}^{n} K_t(\boldsymbol{\theta}) Q_t(\boldsymbol{\theta}) - \mathrm{E}(K_t(\boldsymbol{\theta}) Q_t(\boldsymbol{\theta})) \right|^{q_1} \right) \\
&\quad \leq C \Bigg[\left\{ \sum_{j=1}^{n-1} \left(\sum_{l=1}^{n-j} S_{l,l} \right)^2 + \sum_{j=1}^{n-1} \left(\sum_{l=1}^{n-j} V_{l,l} \right)^2 \right\}^{q_1/2} \\
&\qquad + n^{(q_1-2)/2} \sum_{j=2}^{n-1} \Bigg\{ \left(\sum_{i=1}^{j-1} \left(\sum_{l=1}^{n-j} S_{l+j-i,l} \right)^2 \right)^{q_1/2} \\
&\qquad\quad + \left(\sum_{i=1}^{j-1} \left(\sum_{l=1}^{n-j} S_{l,l+j-i} \right)^2 \right)^{q_1/2} \\
&\qquad\quad + \left(\sum_{i=1}^{j-1} \left(\sum_{l=1}^{n-j} V_{l+j-i,l} \right)^2 \right)^{q_1/2} \\
&\qquad\quad + \left(\sum_{i=1}^{j-1} \left(\sum_{l=1}^{n-j} V_{l,l+j-i} \right)^2 \right)^{q_1/2} \Bigg\} \Bigg],
\end{aligned} \tag{B.6}$$

where

$$V_{i,j} = |c_i(\boldsymbol{\theta}_a) d_j(\boldsymbol{\theta}_a)| \quad \text{and} \quad S_{i,j} = \max_{\mathbf{j}\in\mathbf{J}(m,k),1\leq m\leq k} \sup_{\boldsymbol{\theta}\in B_{\delta_1}(\boldsymbol{\theta}_a)} |\mathbf{D}_\mathbf{j}\{c_i(\boldsymbol{\theta}) d_j(\boldsymbol{\theta})\}|.$$

The proof of (B.5), given in Appendix B of [4], is based on (3.8) of [14] and Lemma 2 of [21]. Assuming that $\sup_{i\geq 1}\mathrm{E}|\varepsilon_i|^{q_1}<\infty$ for some $q_1>\max\{q,2\}$ with $q\geq 1$, (B.5) can be used to justify (2.8) for the ARMA case. More precisely, applying (B.5) with $K_t(\boldsymbol{\theta})=(\nabla^2\varepsilon_t(\boldsymbol{\eta}))_{i,j}$ and $\epsilon_t=\varepsilon_t$, in conjunction with (B.1) and (B.4), it follows that for any $\delta_1>0$ with $B_{\delta_1}(\boldsymbol{\eta}_0)\subset\Pi$,

$$\max_{1\leq i,j\leq\bar{p}}\mathrm{E}\left(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}\left|n^{-1/2}\sum_{t=1}^{n}\varepsilon_t(\nabla^2\varepsilon_t(\boldsymbol{\eta}))_{i,j}\right|^{q_1}\right)=O(1). \tag{B.7}$$

In addition, by making use of (B.5) with $K_t(\boldsymbol{\theta})=\varepsilon_t(\boldsymbol{\eta})-\varepsilon_t(\boldsymbol{\eta}_0)$ and $\epsilon_t=\varepsilon_t$, the compactness of $\tilde{\Pi}$, (3.17) and (B.2), we obtain

$$\mathrm{E}\left(\sup_{\boldsymbol{\eta}\in\tilde{\Pi}}\left|n^{-1}\sum_{t=1}^{n}\varepsilon_t\big(\varepsilon_t(\boldsymbol{\eta})-\varepsilon_t(\boldsymbol{\eta}_0)\big)\right|^{q_1}\right)=O(n^{-q_1/2}), \tag{B.8}$$

which gives (2.11) (with $q_2=q_1$ and $\nu=1/2$) for the ARMA case.

On the other hand, (B.6), whose proof is also given in Appendix B of [4], can be viewed as a uniform version of the first moment bound theorem of [6] and plays a key role in verifying (2.12)–(2.14) for the ARMA case. Let $\bar{M}_3$ be any positive number larger than $2D_1^2\sigma^2\sum_{l=1}^{\infty}\exp(-2D_2l)$ and δ_1 be any positive number satisfying $B_{\delta_1}(\boldsymbol{\eta}_0)\subset\Pi$, noting that D_1 and D_2 are defined in (B.1). Assume $\sup_{i\geq 1}\mathrm{E}|\varepsilon_i|^{2q_1}<\infty$ for some $q_1\geq 2q$ with $q\geq 1$. Then, by (B.6) with $K_t(\boldsymbol{\theta})=Q_t(\boldsymbol{\theta})=(\nabla^2\varepsilon_t(\boldsymbol{\eta}))_{i,j}$ and $\epsilon_t=\varepsilon_t$, (B.1), (B.4) and Chebyshev's inequality, one has for any $1\leq i,j\leq\bar{p}$,

$$\begin{aligned}
&P\left(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}n^{-1}\sum_{t=1}^{n}(\nabla^2\varepsilon_t(\boldsymbol{\eta}))_{i,j}^2>\bar{M}_3\right)\\
&\leq P\left(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}\left|n^{-1}\sum_{t=1}^{n}[(\nabla^2\varepsilon_t(\boldsymbol{\eta}))_{i,j}^2-\mathrm{E}\{(\nabla^2\varepsilon_t(\boldsymbol{\eta}))_{i,j}^2\}]\right|^{q_1}>(\bar{M}_3/2)^{q_1}\right) \qquad \text{(B.9)}\\
&=O(n^{-q_1/2})=O(n^{-q}),
\end{aligned}$$

which is (2.14) for the ARMA case. In addition, (2.12) and (2.13) for the ARMA case, that is, for some $\bar{M}_1,\bar{M}_2>0$,

$$P\left(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}\lambda_{\min}^{-1}\left(n^{-1}\sum_{t=1}^{n}\nabla\varepsilon_t(\boldsymbol{\eta})(\nabla\varepsilon_t(\boldsymbol{\eta}))^{\mathrm{T}}\right)>\bar{M}_1\right)=O(n^{-q}), \tag{B.10}$$

$$P\left(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)}n^{-1}\sum_{t=1}^{n}\|\nabla\varepsilon_t(\boldsymbol{\eta})\|^2>\bar{M}_2\right)=O(n^{-q}), \tag{B.11}$$

can also be similarly verified. With the help of these results, we are now in a position to prove Theorem 3.3.

PROOF OF THEOREM 3.3. Since (3.28) is assumed, (B.7)–(B.11) follow. In view of Theorems 2.2, 3.1 and 3.2, it remains to show that for some $q_1 > q \geq 1$ and some small positive number δ_1 with $B_{\delta_1}(\boldsymbol{\eta}) \subset \Pi$,

$$\max_{1\leq i,j\leq \tilde{p},1\leq t\leq n} \mathrm{E}\Big(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)} |(\nabla^2 \varepsilon_t(\boldsymbol{\eta}))_{i,j}|^{4q_1}\Big) = O(1), \tag{B.12}$$

and

$$\max_{1\leq t\leq n} \mathrm{E}\Big(\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)} \|\nabla \varepsilon_t(\boldsymbol{\eta})\|^{4q_1}\Big) = O(1). \tag{B.13}$$

These equations, however, can be verified based on (3.15), (3.28), (B.1), (B.3), (B.4) and an argument similar to (3.10) of [14]. The details are thus omitted here. □

Acknowledgments. We would like to thank the Editor, an Associate Editor and an anonymous referee for helpful comments and suggestions, which lead to an improved version of this paper.

REFERENCES

[1] AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203–217. MR0286233

[2] ANDREWS, D. W. K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica* **55** 1465–1471. MR0923471

[3] BURKHOLDER, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab.* **1** 19–42. MR0365692

[4] CHAN, N. H. and ING, C. K. (2010). Uniform moment bounds of Fisher's information with applications to time series. Technical report, Institute of Statistical Science, Academia Sinica. Available at http://www.stat.sinica.edu.tw/Ing/ChanIng10.pdf.

[5] DOOB, J. L. (1954). Semimartingales and subharmonic functions. *Trans. Amer. Math. Soc.* **77** 86–121. MR0064347

[6] FINDLEY, D. F. and WEI, C. Z. (1993). Moment bounds for deriving time series CLTs and model selection procedures. *Statist. Sinica* **3** 453–480. MR1243396

[7] FINDLEY, D. F. and WEI, C.-Z. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *J. Multivariate Anal.* **83** 415–450. MR1945962

[8] FULLER, W. A. and HASZA, D. P. (1981). Properties of predictors for autoregressive time series. *J. Amer. Statist. Assoc.* **76** 155–161. MR0608187

[9] HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*: *Probability and Mathematical Statistics*. Academic Press, New York. MR0624435

[10] ING, C.-K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory* **19** 254–279. MR1966030

[11] ING, C.-K. and SIN, C.-Y. (2006). On prediction errors in regression models with nonstationary regressors. In *Time Series and Related Topics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **52** 60–71. IMS, Beachwood, OH. MR2427839

[12] ING, C.-K. and WEI, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Ann. Statist.* **33** 2423–2474. MR2211091

[13] KUNITOMO, N. and YAMAMOTO, T. (1985). Properties of predictors in misspecified autoregressive time series models. *J. Amer. Statist. Assoc.* **80** 941–950. MR0819598
[14] LAI, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.* **22** 1917–1930. MR1329175
[15] 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. MR0642726
[16] LAI, T. L. and YING, Z. (2006). Efficient recursive estimation and adaptive control in stochastic regression and ARMAX models. *Statist. Sinica* **16** 741–772. MR2281300
[17] ROBINSON, P. M. and HIDALGO, F. J. (1997). Time series regression with long-range dependence. *Ann. Statist.* **25** 77–104. MR1429918
[18] SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164. MR0557560
[19] SKOURAS, K. (2000). Strong consistency in nonlinear stochastic regression models. *Ann. Statist.* **28** 871–879. MR1792791
[20] TUKEY, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics* **3** 191–219. MR0125733
[21] WEI, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15** 1667–1682. MR0913581
[22] WEI, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20** 1–42. MR1150333
[23] YAMAMOTO, T. (1981). Predictions of multivariate autoregressive-moving average models. *Biometrika* **68** 485–492. MR0626411

DEPARTMENT OF STATISTICS
ROOM 118, LADY SHAW BUILDING
CHINESE UNIVERSITY OF HONG KONG
SHATIN, NEW TERRITORIES
HONG KONG
E-MAIL: nhchan@sta.cuhk.edu.hk

INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI 115
TAIWAN, ROC
E-MAIL: cking@stat.sinica.edu.tw