Estimation of the innovation density in nonlinear autoregressive models with applications

Kengo Kato [*]

September 6, 2009

Abstract

This paper addresses the problem of estimating the innovation density in nonlinear autoregressive models. Specifically, we establish the convergence rate of the supremum distance between the residual-based kernel density estimator and the kernel density estimator using the unobservable actual innovation variables. The proof of the main theorem relies on empirical process theory instead of the conventional Taylor expansion approach. As applications, we obtain the exact rate of weak uniform consistency on the whole line, pointwise asymptotic normality of the residual-based kernel density estimator and the asymptotic distribution of a Bickel-Rosenblatt type global measure statistic related to it. We also examine the conditions of the main theorem for some specific time series model.

Key words: empirical process theory; density estimation; nonlinear autoregressive model.
AMS subject classifications: 62G07, 62M10.
Running headline: Estimation of innovation density.

1 Introduction

This paper addresses the problem of estimating the innovation density in nonlinear autoregressive models. A nonlinear autoregressive model of order p is defined as

$$X_t = m(X_{t-1}, \dots, X_{t-p}; \boldsymbol{\theta}) + e_t,\ t = 0, \pm 1, \pm 2, \dots, \tag{1.1}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)$ is a vector of unknown parameters, Θ is a parameter space which is a Borel measurable subset of $\mathbb{R}^q$, $m : \mathbb{R}^p \times \Theta \to \mathbb{R}$ is called an autoregression function, $\{e_t\}$ is a sequence of iid random variables and e_t is independent of $\{X_{t-k}, k \geq 1\}$ for all t. Let f denote the density of e_t, which we want to estimate.

[*] Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan. Email: `kkato@hiroshima-u.ac.jp`

Suppose first that e_t's are observable. In this case, the kernel density estimator of f is defined as

$$f_n(u_0) = \frac{1}{n} \sum_{t=1}^{n} K_{h_n}(e_t - u_0),$$

where u_0 is an arbitrary point in $\mathbb{R}$, $K(\cdot)$ is a kernel function, $h_n > 0$ is a bandwidth and $K_{h_n}(u) = h_n^{-1} K(u/h_n)$. There is much literature on properties of kernel density estimators. The history of density estimation for observable data is well summarized in Section 5.8 of Fan & Yao (2005). In practice, of course, e_t's are not observable. A natural approach to estimate f is to replace e_t's by residuals in the definition of f_n. Suppose we have a $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Throughout the paper, the sample is $\{X_{-p+1}, \ldots, X_n\}$. The residuals for the model (1.1) are defined as $\hat{e}_t = X_t - m(\mathbf{X}_{t-1}; \hat{\boldsymbol{\theta}}),\ t = 1, \ldots, n$, where $\mathbf{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})$. Then, the residual-based kernel density estimator of f is defined as

$$\hat{f}_n(u_0) = \frac{1}{n} \sum_{t=1}^{n} K_{h_n}(\hat{e}_t - u_0).$$

Since we use $\hat{e}_t$ as a proxy of e_t, we expect that $\hat{f}_n$ behaves like f_n. A natural question is: How "close" is $\hat{f}_n$ to f_n? The question motivates us to study the convergence rate of the supremum distance between $\hat{f}_n$ and f_n, that is, $\|\hat{f}_n - f_n\|_\infty := \sup_{u_0 \in \mathbb{R}} |\hat{f}_n(u_0) - f_n(u_0)|$. Although the statement of the problem is clear, there are some difficulties in analyzing the convergence rate. The difficulties are listed as follows: (1) the data are dependent; (2) the estimated parameter of possibly unknown form appears inside the kernel; (3) the supremum is taken over the whole line. It is worth pointing out that a simple expansion approach does not provide a useful result. To account for this, let us consider a linear autoregressive model of order 1: $X_t = \theta_0 + \theta_1 X_{t-1} + e_t$ where $|\theta_1| < 1$. Assuming that $K(\cdot)$ is Lipschitz continuous, we may infer that $|\hat{f}_n(u_0) - f_n(u_0)| \leq L h_n^{-2} (|\hat{\theta}_0 - \theta_0| + |\hat{\theta}_1 - \theta_1| n^{-1} \sum_{t=1}^{n} X_{t-1})$, where $L > 0$ is a Lipschitz constant of $K(\cdot)$. Under standard assumptions, the right hand side is of order $O_p(n^{-1/2} h_n^{-2})$. However, this rate is far from sharp. For example, if we want to establish the rate of uniform consistency of $\hat{f}_n$, we expect that $\|\hat{f}_n - f_n\|_\infty$ is of order $o_p\{(nh_n)^{-1/2} \log h_n^{-1}\}$. In this case, the above rate is not enough.

For the supremum distance on a compact subset of $\mathbb{R}$, Liebscher (1999) established the $o\{(nh_n)^{-1/2}\}$ rate of almost sure convergence under slightly restrictive conditions. He used the second order Taylor expansion in conjunction with a sophisticated truncation argument and an exponential inequality for mixing processes established in Liebscher (1996) to obtain the rate. Müller et al. (2005) also considered a problem similar to ours. Based on the Taylor expansion approach, they established the convergence rate of the weighted L_1 norm between $\hat{f}_n$ and f_n. However, their results do not lead to the convergence rate of $\|\hat{f}_n - f_n\|_\infty$. To the best of our knowledge, the convergence rate of the supremum distance between $\hat{f}_n$ and f_n on the whole line is still an open problem.

The main purpose of this paper is to establish a sharp convergence rate of $\|\hat{f}_n - f_n\|_\infty$. The result of this paper improves upon that of Liebscher (1999) in several aspects. Our result enables us to determine the condition on h_n under which the pre-specified rate of convergence of $\|\hat{f}_n - f_n\|_\infty$ is ensured. For example, if we want to make $\|\hat{f}_n - f_n\|_\infty =$

$o_p\{(nh_n)^{-1/2}\log h_n^{-1}\}$, the main theorem states that if f is Lipschitz continuous, it is enough to set h_n such that $h_n \to 0$ and $n^{1/2}h_n \log h_n^{-1}/\log n \to \infty$. It is worth pointing out that the convergence rate can be even of order $O_p(n^{-1/2})$, which is faster than the convergence rate of kernel estimators; see Corollary 2.1 below. As applications, we obtain the exact rate of uniform consistency and pointwise asymptotic normality of $\hat{f}_n$. Also, we establish the asymptotic distribution of the maximum deviation of $\hat{f}_n$ from $\mathrm{E}[f_n(u_0)]$ on an interval where f is bounded away from 0 and ∞, which is an extension of Theorem 3.1 in Bickel & Rosenblatt (1973) to the residual-based kernel density estimator. The last application is of practical importance since it provides an asymptotically distribution free goodness-of-fit test for the innovation density; see Proposition 2.3 below.

From a methodological point of view, there is an important difference between the present paper and the past two papers. The proofs in the past two paper rely on the Taylor expansion approach. Whereas, the proof of Proposition 4.1 in the present paper, which is a key to the main theorem, relies on empirical process techniques. Empirical process techniques are now basic tools for studying uniform asymptotic behaviors of kernel density estimators. Pollard (1984) demonstrated how empirical process techniques can be used to prove uniform consistency of kernel density estimators. Yu (1993) established rates of uniform consistency of kernel density estimators for β-mixing processes. She introduced the blocking technique to modify Pollard's methods to β-mixing processes. The blocking technique plays an important role in the proof of the main theorem. Recent developments in this fields include Deheuvels (2000), Einmahl & Mason (2000, 2005) and Giné & Guillou (2002), to name only a few. The contribution of this paper is to demonstrate how empirical process techniques are incorporated to studying asymptotic behaviors of residual-based kernel density estimators for possibly nonlinear time series models.

The organization of this paper is as follows. Section 2 presents the main theorem which establishes the convergence rate of the supremum distance between $\hat{f}_n$ and f_n. The latter part of Section 2 includes some applications of the main theorem. In Section 3, the conditions of the main theorem are examined for some specific nonlinear time series models. All the technical proofs are provided in Section 4.

We introduce some notations used in the present paper. Let $I(A)$ denote the indicator of an event A. For $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$. For $a \in \mathbb{R}$, $[a]$ denotes the greatest integer not exceeding a. Let $\|g\|_\infty$ denote the supremum norm of a generic function g. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ when there exists a positive constant M such that $M^{-1}b_n \le a_n \le Mb_n$ for every n. Throughout the paper, all vectors are row vectors.

2 Main results

2.1 Convergence rate

In this section, we establish the convergence rate of $\|\hat{f}_n - f_n\|_\infty$.

To study the convergence rate, we assume that the process $\{X_t\}$ is stationary and geometrically β-mixing. Let $(\Omega, \mathcal{A}, \mathrm{P})$ denote an underlying probability space. Let $\mathcal{A}_i^j$ denote the σ-field generated by $\{X_t, i \leq t \leq j\}$ $(i < j)$. The β-mixing coefficient of the process $\{X_t\}$ is defined as

$$\beta(k) = \mathrm{E}\left[\sup\{|\mathrm{P}(B|\mathcal{A}_{-\infty}^0) - \mathrm{P}(B)| : B \in \mathcal{A}_k^\infty\}\right], \ k \geq 1.$$

From Volkonskii & Rozanov (1959), it is shown that $\beta(k)$ has an alternative expression

$$\begin{aligned}\beta(k) = \frac{1}{2}\sup\Big\{ \textstyle\sum_{i=1}^I \sum_{j=1}^J |\mathrm{P}(A_i \cap B_j) - \mathrm{P}(A_i)\mathrm{P}(B_j)| : \\ \{A_i\}_{i=1}^I \text{ is a finite partition in } \mathcal{A}_{-\infty}^0, \\ \{B_j\}_{j=1}^J \text{ is a finite partition in } \mathcal{A}_k^\infty\Big\}. \qquad (2.1)\end{aligned}$$

For a general treatment on various mixing conditions, we refer to Section 2.6 of Fan & Yao (2005) and references therein.

We now state our regularity conditions. Throughout the paper, the true parameter $\boldsymbol{\theta} \in \Theta$ is fixed.

(A1) The process $\{X_t\}$ is stationary and β-mixing with exponential decaying coefficients.

(A2) There exists a closed ball B in $\mathbb{R}^q$ centered at $\boldsymbol{\theta}$ such that $B \subset \Theta$.

(A3) The map $(\mathbf{x}, \boldsymbol{\vartheta}) \mapsto m(\mathbf{x}; \boldsymbol{\vartheta})$ is Borel measurable; there exists a Borel measurable function $M(\mathbf{x})$ such that $|m(\mathbf{x}; \boldsymbol{\vartheta}) - m(\mathbf{x}; \boldsymbol{\theta})| \leq M(\mathbf{x})\|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|$ for $\boldsymbol{\vartheta}$ in a neighborhood of $\boldsymbol{\theta}$ in Θ and $\mathrm{E}[M^2(\mathbf{X}_{t-1})] < \infty$.

(A4) The innovation density f is bounded and λ-th Hölder continuous with $\lambda \in (0, 1]$, that is, $|f(u) - f(v)| \leq \text{const.} \times |u - v|^\lambda$ for all $u, v \in \mathbb{R}$.

(A5) The kernel function $K(\cdot)$ is a Lipschitz continuous density function with

$$\int_{-\infty}^{\infty} |uK'(u)| du < \infty. \qquad (2.2)$$

(A6) The estimator $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$; that is, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + O_p(n^{-1/2})$.

We state some remarks on the conditions. There are several sufficient conditions for stationarity and geometric β-mixing property of the process $\{X_t\}$. See, for example, Theorems 2.2 and 2.4 in Fan & Yao (2005). Remember that stationary geometrically ergodic Markov chains are β-mixing with exponentially decaying coefficients; see equation (2.58) in Fan & Yao (2005) and Theorem 2.1 of Nummelin & Tuominen (1982). Liebscher (1999) imposed the condition adopted in Masry & Tjøstheim (1995) to ensure geometric ergodicity of the process; see Condition $\mathcal{G}$ in his paper. In truth, the geometric β-mixing condition is stronger than needed; however, we put it for a technical convenience. Condition (A3) is weaker than Condition $\mathcal{M}$ in Liebscher (1999). Actually, we do not assume the second order differentiability of the map $\boldsymbol{\vartheta} \mapsto m(\mathbf{x}; \boldsymbol{\vartheta})$ and require the weaker moment condition on

$M(\mathbf{X}_{t-1})$. Condition (A5) allows for the Gaussian, the triangular and the Epanchnikov kernels. Equation (2.2) implies that $K(\cdot)$ is of bounded variation on $\mathbb{R}$. Clearly, a compactly supported Lipschitz continuous density satisfies condition (A5). It is noted that condition (A5) is weaker than Condition $\mathcal{K}(2)$ in Liebscher (1999). Note that Condition $\mathcal{K}(2)$ in Liebscher (1999) does not imply the Lipschitz continuity of $K(\cdot)$. However, he used the property in the proof of his Theorem 3.1; see equation (3.5) in his paper. There is vast literature on the estimation of parameters in nonlinear autoregressive models. Among them, Klimko & Nelson (1978) established asymptotic normality of conditional least squares estimators for smooth nonlinear autoregressive models. Tjøstheim (1986) developed general conditions under which consistency and asymptotic normality of M-estimators hold for nonlinear time series models. More recently, Koul (1996) established asymptotic normality of M-, R- and minimum distance estimators for nonlinear autoregressive models.

We now present the main result of this paper. The proof of Theorem 2.1 is relegated to Section 4.

Theorem 2.1. *Assume that conditions (A1)-(A6) are satisfied. Let*

$$h_n \to 0, \ r_n \to \infty, \ \frac{n^{3/2} h_n^2}{r_n^2 \log n} \to \infty. \tag{2.3}$$

Then, we have $\|\hat{f}_n - f_n\|_\infty = o_p(r_n^{-1}) + O_p(n^{-1/2} h_n^{\lambda-1} \wedge 1)$.

Remark 2.1. It is possible to develop an a.s. version of Theorem 2.1 if we replace condition (A6) by

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log\log n}} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \text{const., a.s.} \tag{2.4}$$

Liebscher (1999) adopted (2.4) as an initial condition on the estimator. Sufficient conditions under which (2.4) holds are found in Liebscher (2003) and references therein.

We shall compare Theorem 2.1 with Theorem 3.1 of Liebscher (1999). He assumed that the estimator $\hat{\boldsymbol{\theta}}$ satisfies (2.4) and the bandwidth h_n is such that $h_n \to 0$ and $h_n \geq \text{const.} \times n^{-1/5}$. With some additional conditions including the Lipschitz continuity of f, he showed in Theorem 3.1 that for any compact subset D of $\mathbb{R}$, $\sup_{u_0 \in D} |\hat{f}_n(u_0) - f_n(u_0)| = o\{(nh_n)^{-1/2}\}$, almost surely. Theorem 2.1 relaxes the condition on the bandwidth, removes the restriction on compact subsets and gives a sharper result on the convergence rate.

We explain an intuition behind the proof of Theorem 2.1. Put $\Delta(\mathbf{X}_{t-1}; \boldsymbol{\vartheta}) = m(\mathbf{X}_{t-1}; \boldsymbol{\vartheta}) - m(\mathbf{X}_{t-1}; \boldsymbol{\theta})$. Let us define $f_n(u_0, \boldsymbol{\vartheta}) = n^{-1} \sum_{t=1}^n K_{h_n}(e_t - u_0 - \Delta(\mathbf{X}_{t-1}, \boldsymbol{\vartheta}))$. The difference $\hat{f}_n(u_0) - f_n(u_0)$ is decomposed as

$$\begin{aligned} \hat{f}_n(u_0) - f_n(u_0) = [\{f_n(u_0, \hat{\boldsymbol{\vartheta}}) - \mathrm{E}[f_n(u_0, \boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta} = \hat{\boldsymbol{\theta}}}\} - \{f_n(u_0, \boldsymbol{\theta}) - \mathrm{E}[f_n(u_0, \boldsymbol{\theta})]\}] \\ + \{\mathrm{E}[f_n(u_0, \boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta} = \hat{\boldsymbol{\theta}}} - \mathrm{E}[f_n(u_0, \boldsymbol{\theta})]\}. \end{aligned}$$

Using the fact that $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$, we may handle the first term by means of empirical process techniques. The second term depends only on the smoothness of the non-random map $\boldsymbol{\vartheta} \mapsto \mathrm{E}[f_n(u_0, \boldsymbol{\vartheta})]$ and can be handled more easily. In Section 4, we will establish that the first term is $o_p(r_n^{-1})$ and the second term is $O_p(n^{-1/2}h_n^{\lambda-1} \wedge 1)$ uniformly over $u_0 \in \mathbb{R}$.

It is worth pointing out that the Hölder continuity of f is not used in establishing the uniform convergence rate of the first part. With a suitable choice of h_n, the uniform convergence rate of the first term can be even faster than $O_p(n^{-1/2})$. In this case, the second term dominates the first term. Under some additional conditions, we can obtain a ramification of Theorem 2.1, which is stated as a corollary. Put $\dot{m}(\mathbf{x};\boldsymbol{\vartheta}) = \partial m(\mathbf{x};\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}$. We introduce another set of conditions that are stronger than conditions (A3)-(A5).

(A3') The map $(\mathbf{x},\boldsymbol{\vartheta}) \mapsto m(\mathbf{x};\boldsymbol{\vartheta})$ is Borel measurable; the map $\boldsymbol{\vartheta} \mapsto m(\mathbf{x};\boldsymbol{\vartheta})$ is continuously differentiable for each $\mathbf{x}$. There exists a Borel measurable function $M(\mathbf{x})$ such that $\|\dot{m}(\mathbf{x};\boldsymbol{\vartheta})\| \leq M(\mathbf{x})$ for all $\mathbf{x}$ and for all $\boldsymbol{\vartheta}$ in a neighborhood of $\boldsymbol{\theta}$ and $\mathrm{E}[M^3(\mathbf{X}_{t-1})] < \infty$.

(A4') The innovation density f is bounded and twice continuously differentiable with bounded first and second derivatives.

(A5') The kernel function $K(\cdot)$ is a Lipschitz continuous density function with

$$\int_{-\infty}^{\infty} |u^2 K'(u)| du < \infty.$$

Corollary 2.1. *Assume that conditions (A1)-(A2), (A3')-(A5') and (A6) are satisfied. If* $h_n \to 0$ *and* $n^{1/2}h_n^2/\log n \to \infty$*, then we have*

$$\|\hat{f}_n(\cdot) - f_n(\cdot) - f'(\cdot)\mathrm{E}[\dot{m}(\mathbf{X}_{t-1};\boldsymbol{\theta})](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\|_\infty = o_p(n^{-1/2}).$$

Müller et al. (2005) obtained a result similar to Corollary 2.1; see Theorem 3.2 in their paper. Under suitable regularity conditions, they showed that if $nh_n^{(50+20q)/(14+5q)} \to \infty$,

$$\int_{-\infty}^{\infty} |\hat{f}_n(u_0) - f_n(u_0) - f'(u_0)\mathrm{E}[\dot{m}(\mathbf{X}_{t-1};\boldsymbol{\theta})](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top| V(u_0) du_0 = o_p(n^{-1/2}),$$

where $V(\cdot)$ is an appropriate weight function. As mentioned earlier, however, their result does not imply our Corollary 2.1. We note that in order to apply the second order Taylor expansion to the kernel density estimator, they assumed twice continuous differentiability of the kernel $K(\cdot)$, which rules out some important kernels such as the Epanchnikov and the triangular kernels.

2.2 Applications

In this section, we present three applications of Theorem 2.1. Specifically, we consider the exact rate of weak uniform consistency and pointwise asymptotic normality of the residual-based kernel density estimator. Also, we establish the asymptotic distribution of the maximum deviation of $\hat{f}_n$ from $\mathrm{E}[f_n(u_0)]$ on a compact interval.

Silverman (1978), Stute (1984), Deheuvels (2000), Einmahl & Mason (2000, 2005) and Giné & Guillou (2002) studied rates of uniform consistency of kernel density estimators for observable data. Among them, Giné & Guillou (2002) used empirical process theory to establish the exact rate of strong uniform consistency of f_n under very mild assumptions. Assume that $K(\cdot)$ is a bounded, compactly supported density function, the functional class $\mathcal{K} = \{e \mapsto K((e - u_0)/h) : u_0 \in \mathbb{R}, h > 0\}$ is Euclidean (see Definition 4.2 in Section 4) and f is bounded and uniformly continuous on $\mathbb{R}$. Theorem 3.3 of Giné & Guillou (2002) shows that if

$$h_n \downarrow 0,\ nh_n \uparrow \infty,\ \frac{nh_n}{|\log h_n|} \to \infty,\ \frac{|\log h_n|}{\log\log n} \to \infty,$$

we have

$$\lim_{n\to\infty} \sqrt{\frac{nh_n}{2\log h_n^{-1}}} \|f_n - \mathrm{E}[f_n(\cdot)]\|_\infty = \|K\|_2 \|f\|_\infty^{1/2},\ \text{a.s.}, \tag{2.5}$$

where $\|K\|_2^2 = \int_{-\infty}^{\infty} K^2(u)du$. In view of (2.5), we shall seek conditions under which equation (2.6) below holds:

$$\operatorname*{plim}_{n\to\infty} \sqrt{\frac{nh_n}{2\log h_n^{-1}}} \|\hat{f}_n - \mathrm{E}[f_n(\cdot)]\|_\infty = \|K\|_2 \|f\|_\infty^{1/2}. \tag{2.6}$$

To make $\|\hat{f}_n - f_n\|_\infty$ negligible, take $r_n = \sqrt{nh_n/\log h_n^{-1}}$. The last part of (2.3) is satisfied if $n^{1/2}h_n \log h_n^{-1}/\log n \to \infty$. Since $K(\cdot)$ is of bounded variation, the class $\mathcal{K}$ is shown to be Euclidean under condition (A5); see Lemma 22 in Nolan & Pollard (1987). Therefore, we obtain the next proposition.

Proposition 2.1. *Assume that conditions (A1)-(A6) are satisfied with $\lambda \in [1/2, 1]$ in (A4). If $K(\cdot)$ is compactly supported and if*

$$h_n \downarrow 0,\ nh_n \uparrow \infty,\ \frac{|\log h_n|}{\log\log n} \to \infty,\ \frac{n^{1/2}h_n \log h_n^{-1}}{\log n} \to \infty,$$

then (2.6) holds.

Remark 2.2. To establish the rate only, the compactness of the support of $K(\cdot)$ is not required. See Theorem 2.3 in Giné & Guillou (2002).

It is rather easy to find sufficient conditions under which pointwise asymptotic normality of $\hat{f}_n$ holds. Fix $u_0 \in \mathbb{R}$. If $K(\cdot)$ is a bounded density function and f is bounded on $\mathbb{R}$ and continuous at u_0, the Lyapunov central limit theorem shows that

$$\sqrt{nh_n}\{f_n(u_0) - \mathrm{E}[f_n(u_0)]\} \xrightarrow{d} N\{0, f(u_0)\|K\|_2^2\}.$$

Take $r_n = (nh_n)^{1/2}$. The last part of (2.3) is satisfied if $n^{1/2}h_n/\log n \to \infty$. Thus, we obtain the next proposition:

Proposition 2.2. *Fix $u_0 \in \mathbb{R}$. Assume that conditions (A1)-(A6) are satisfied with $\lambda \in (1/2, 1]$ in (A4). If $h_n \to 0$ and $n^{1/2}h_n/\log n \to \infty$, then we have*

$$\sqrt{nh_n}\{\hat{f}_n(u_0) - \mathrm{E}[f_n(u_0)]\} \xrightarrow{d} N\{0, f(u_0)\|K\|_2^2\}.$$

A replacement of $\mathrm{E}[f_n(u_0)]$ by $f(u_0)$ is a routine problem in density estimation theory. For example, if f is twice continuously differentiable and $K(\cdot)$ is symmetric with $\mu_2(K) := \int_{-\infty}^{\infty} u^2 K(u) du < \infty$, $\mathrm{E}[f_n(u_0)]$ may be expanded as

$$\mathrm{E}[f_n(u_0)] = f(u_0) + \frac{h_n^2 f''(u_0)}{2}\mu_2(K) + o(h_n^2).$$

In this case, we can calculate the asymptotic mean squared error (AMSE) of $\hat{f}_n(u_0)$. Minimizing the AMSE yields the optimal bandwidth $h_n^{\mathrm{opt}} \asymp n^{-1/5}$. Clearly, the optimal bandwidth for $\hat{f}_n(u_0)$ is same as that for $f_n(u_0)$.

The last application is to derive the asymptotic distribution of the maximum deviation of $\hat{f}_n$ from $\mathrm{E}[f_n(u_0)]$ on a compact interval. Assume that f is continuous, positive and bounded. For an arbitrary closed interval $[a, b]$ with $-\infty < a < b < \infty$, define $\hat{M}_n = \{(b-a)nh_n\}^{1/2} \sup_{u_0 \in [a,b]} |\hat{f}_n(u_0) - \mathrm{E}[f_n(u_0)]|/\sqrt{f(u_0)}$ and $M_n = \{(b-a)nh_n\}^{1/2} \sup_{u_0 \in [a,b]} |f_n(u_0) - \mathrm{E}[f_n(u_0)]|/\sqrt{f(u_0)}$. Bickel & Rosenblatt (1973) showed in Theorem 3.1 that under their conditions A1-(b), A2 and A3, if $h_n = \text{const.} \times n^{-\delta}$ with $0 < \delta < 1/2$,

$$\mathrm{P}\left((-2\log h_n)^{1/2}\left(\frac{M_n}{\|K\|_2} - d_n\right) < x\right) \to \exp\{-2\exp(-x)\}, \tag{2.7}$$

where

$$d_n = (-2\log h_n)^{1/2} + \frac{1}{(-2\log h_n)^{1/2}}\left(\log \frac{1}{2\pi}\frac{\|K'\|_2}{\|K\|_2}\right).$$

In order to make $\hat{M}_n - M_n = o_p\{(-\log h_n)^{-1/2}\}$, take $r_n = (-nh_n \log h_n)^{1/2}$ in Theorem 2.1. The last part of (2.3) is satisfied for $h_n = \text{const.} \times n^{-\delta}$ with $0 < \delta < 1/2$. Condition (A5) implies condition A1-(b) in Bickel & Rosenblatt (1973). The first part of condition A3 in Bickel & Rosenblatt (1973) is satisfied if f is Lipschitz continuous and $f'/f^{1/2}$ is bounded in absolute value. Therefore, we obtain the next proposition.

Proposition 2.3. *Let $[a, b]$ be an arbitrary closed interval with $-\infty < a < b < \infty$. Assume that conditions (A1)-(A6) are satisfied with $\lambda = 1$ in (A4), that is, f is Lipschitz continuous. Assume further that f is positive, $f'/f^{1/2}$ is bounded in absolute value and $u^2K(u), u^2K'(u)$ are integrable. If $h_n = \text{const.} \times n^{-\delta}$ with $0 < \delta < 1/2$, then (2.7) holds for every x with M_n replaced by $\hat{M}_n$.*

Cheng (2005), putting more restrictive conditions on h_n, established the same conclusion of Proposition 2.3 for a linear autoregressive model of order 1. Proposition 2.3 is an extension of Cheng's Theorem 4.1 to a general nonlinear autoregressive model. Applications of Proposition 2.3 include the construction of uniform confidence bands and goodness-of-fit tests for the innovation density. For example, let us consider to test the null hypothesis $H_0 : f = f_0$ where f_0 is some known density. In this case, it is natural to reject H_0 if $\hat{M}_n$ is large. Proposition 2.3 enables us to tabulate approximate critical values of $\hat{M}_n$. It is worth pointing out that the test based on $\hat{M}_n$ is asymptotically distribution free (ADF), that is, the asymptotic null distribution of the normalized statistic of $\hat{M}_n$ does not depend on f_0 and $\boldsymbol{\theta}$. To construct a non-trivial ADF goodness-of-fit test for the error distribution is not an easy problem since the asymptotic distribution of a functional of the residual empirical process generally depends on both the underlying distribution and the unknown parameter; see Koul (1996). For example, the residual-based Kolmogorov-Smirnov test is not ADF. The dependence of the asymptotic distribution on the unknown parameter is apparently undesirable when calculating critical values of the test. Recently, Khmaladze & Koul (2004) obtained ADF goodness-of-fit tests for the error distribution in nonlinear regression models. The crux of their approach is a certain martingale type transform of the residual empirical process. Proposition 2.3 shows that we need no such transform when using $\hat{M}_n$ as a goodness-of-fit test statistic for the innovation distribution.

3 Examples

In this section, we examine the conditions of Theorem 2.1 for some specific nonlinear time series models. We deal with well known linear autoregressive (AR) models, threshold autoregressive (TAR) models (Tong & Lim, 1980), exponential autoregressive (EXPAR) models (Ozaki, 1980; Haggan & Ozaki, 1981) and log transformed squared autoregressive conditional heteroscedasticity (ARCH) processes (Engle, 1982).

Example 3.1 (Linear AR model)**.** The first example is a linear AR model of order p:

$$X_t = \theta_1 X_{t-1} + \cdots + \theta_p X_{t-p} + e_t,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a vector of unknown parameters and $\{e_t\}$ is a sequence of iid random variables with mean zero and finite variance. In this case, $q = p$, $m(\mathbf{x}; \boldsymbol{\vartheta}) = \vartheta_1 x_1 + \cdots + \vartheta_p x_p$. The stationarity (causality) condition of the AR process is found in Theorem 3.1.1 of Brockwell & Davies (1991); if the characteristic polynomial $\theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^p$ has no zero in $\{z \in \mathbb{C} : |z| \leq 1\}$, the AR equation has the unique stationarity solution $X_t = \sum_{j=0}^{\infty} a_j e_{t-j}$ with $a_j \to 0$ exponentially fast as $j \to \infty$. If further e_t has a positive density, the process $\{X_t\}$ is β-mixing with exponentially decaying coefficients. Condition (A3) is satisfied with $M(\mathbf{x}) = \|\mathbf{x}\|$. Concerning (A6), the maximum likelihood, the least squares and the Whittle estimators are $\sqrt{n}$-consistent for $\boldsymbol{\theta}$; see Section 10 in Brockwell & Davies (1991).

Example 3.2 (TAR model)**.** The second example is a TAR model with known thresholds. For simplicity, we consider a TAR model of order 1:

$$X_t = \theta_1 X_{t-1} I(X_{t-1} \leq 0) + \theta_2 X_{t-1} I(X_{t-1} > 0) + e_t,$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is a vector of unknown parameters with $\theta_1 \neq \theta_2$ and $\{e_t\}$ is a sequence of iid random variables with mean zero and finite variance. In this case, $p = 1, q = 2$ and $m(x; \boldsymbol{\vartheta}) = \vartheta_1 x I(x \leq 0) + \vartheta_2 x I(x > 0)$. The process $\{X_t\}$ satisfies condition (A1) if $\theta_1 < 1, \theta_2 < 1, \theta_1 \theta_2 < 1$ and e_t has a positive density; see Example 3.7 in An & Huang (1996). Condition (A3) is satisfied with $M(x) = |x|$. It is not difficult to see that, for example, the conditional least squares estimator is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$ under suitable regularity conditions.

Example 3.3 (EXPAR model)**.** Again, for simplicity, we consider an EXPAR model of order 1:

$$X_t = \{\theta_1 + \theta_2 \exp(-\theta_3 X_{t-1}^2)\} X_{t-1} + e_t,$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ is a vector of unknown parameters with $\theta_3 > 0$, $\{e_t\}$ is a sequence of iid random variables with mean zero and finite variance. In this case, $p = 1, q = 3$ and $m(x; \boldsymbol{\vartheta}) = \{\vartheta_1 + \vartheta_2 \exp(-\vartheta_3 x^2)\}x$. The process $\{X_t\}$ satisfies condition (A1) if $|\theta_1| < 1$ and e_t has a positive density; see Example 3.2 in An & Huang (1996). Condition (A3) is satisfied with $M(x) = C|x|$ for some constant $C > 0$. Concerning condition (A6), Tjøstheim (1986) showed in his Theorem 4.1 that if further $\mathrm{E}[e_t^6] < \infty$, the conditional least squares estimator is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$.

Example 3.4 (Log transformed squared ARCH process)**.** The last example is motivated by Peng & Yao (2003) who studied least absolute deviation (LAD) type estimators for GARCH models. An ARCH model of order p is defined as

$$Y_t = \sigma_t \epsilon_t, \ \sigma_t^2 = \theta_0 + \sum_{j=1}^{p} \theta_j Y_{t-j}^2,$$

where $\theta_j > 0$ $(j = 0, \ldots, p)$ are unknown coefficients, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)$, $\{\epsilon_t\}$ is a sequence of iid random variables with mean zero and finite variance. Often ϵ_t is standardized such that $\mathrm{E}[\epsilon_t^2] = 1$ and the Gaussian quasi maximum likelihood estimator (QMLE) is used. Asymptotic theory of Gaussian QMLEs was studied by Weiss (1986) for ARCH models, Hall & Yao (2003) for GARCH models. On the other hand, Peng & Yao (2003) introduced another standardization of ϵ_t. They standardized ϵ_t such that the median of ϵ_t^2 is 1. With this standardization, they proposed the LAD-type estimator

$$\hat{\boldsymbol{\theta}}_{\mathrm{LAD}} = \arg\min_{\boldsymbol{\vartheta}} \sum_{t=1}^{n} |\log Y_t^2 - \log(\vartheta_0 + \textstyle\sum_{j=1}^{p} \vartheta_j Y_{t-j}^2)|.$$

The LAD estimator is advantageous over the Gaussian QMLE in the sense that asymptotic normality of the LAD estimator holds under the weaker moment condition on ϵ_t than the Gaussian QMLE. In fact, the LAD estimator is always asymptotically normal provided that $\mathrm{E}[\epsilon_t^2] < \infty$. On the other hand, when $\mathrm{E}[|\epsilon_t|^d] = \infty$ with $2 < d < 4$, the asymptotic distribution of the Gaussian QLME is no longer normal with a convergence rate slower than $n^{1/2}$. The asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{LAD}} - \boldsymbol{\theta})$ is obtained as

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{LAD}} - \boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}/\{4f^2(0)\}],$$

where f is the density of $\log \epsilon_t^2$ and $\boldsymbol{\Sigma}$ is given in Peng & Yao (2003). Thus, in order to conduct statistical inference using the LAD estimator, we have to estimate $f(0)$. The LAD estimator is the maximum likelihood estimator when $\log \epsilon_t^2$ has a Laplace distribution. So, it is of practical interest to test whether the distribution of $\log \epsilon_t^2$ is close to the Laplace distribution; see also Huang et al. (2008).

It is not difficult to see that $e_t = \log \epsilon_t^2$ is the innovation variable of the nonlinear AR model (1.1) with $X_t = \log Y_t^2$ and $m(\mathbf{x}; \boldsymbol{\vartheta}) = \log\{\vartheta_0 + \sum_{j=1}^{p} \vartheta_j \exp(x_j)\}$. Thus, the results of Section 2 are applicable to the inference on f. The original process $\{Y_t\}$ is stationary and β-mixing with exponential decaying coefficients if $(\mathrm{E}[\epsilon_t^2])^{1/2}\sum_{j=1}^p \theta_j < 1$ and the density of ϵ_t is positive in an interval containing 0; see Carrasco & Chen (2002). Thus, under the same condition, the process $\{X_t\}$ satisfies condition (A1). Condition (A3) is satisfied with $M(\mathbf{x}) = \text{const}$.

4 Proofs

4.1 Proof of Theorem 2.1

In this section, we provide a proof of Theorem 2.1. Throughout this section, we assume all the conditions of Theorem 2.1. We begin with introducing some terminologies related to empirical process theory.

Definition 4.1. Let $\epsilon > 0$ and let $\mathcal{G}$ be a functional class equipped with semimetric ρ.

(a) Every finite collection $g_1, \dots, g_N \in \mathcal{G}$ with the property that for every $g \in \mathcal{G}$, there exists a $j \in \{1, \dots, N\}$ such that $\rho(g, g_j) < \epsilon$ is called *an ϵ-cover of $\mathcal{G}$ with respect to ρ.*

(b) Let $\mathcal{N}(\epsilon, \mathcal{G}, \rho)$ be the size of the smallest ϵ-cover of $\mathcal{G}$ with respect to ρ. Take $\mathcal{N}(\epsilon, \mathcal{G}, \rho) = \infty$ if no finite ϵ-cover exists. Then $\mathcal{N}(\epsilon, \mathcal{G}, \rho)$ is called *an ϵ-covering number of $\mathcal{G}$ with respect to ρ.*

The present definition of a cover of a functional class $\mathcal{G}$ requires that the cover is a subset of $\mathcal{G}$. Compare the definition in Pollard (1984), pp. 25. However, this requirement does not lose any generality. Suppose that there is a enlarged class of functions equipped with some semimetric. $\mathcal{G}$ is a subset of the enlarged class. Then, it is not difficult to see that if there is an ϵ-cover of $\mathcal{G}$ which is not subset of $\mathcal{G}$, there is a (2ϵ)-cover of $\mathcal{G}$ with the property that it is a subset of $\mathcal{G}$ and has the same size as the ϵ-cover.

Following Nolan & Pollard (1987), we introduce *an Euclidean class* of functions, which already appeared in Section 2.2. Analogously, we introduce the notion of *a uniformly Euclidean family* of functional classes, which will be used in the proof of Proposition 4.1.

Definition 4.2. Let d be a positive integer.

(a) Let $\mathcal{G}$ be a class of Borel measurable functions on $\mathbb{R}^d$ with Borel measurable envelope G. The class $\mathcal{G}$ is said to be *Euclidean* with envelope G if there exists positive constants A and V such that for every probability measure Q on $\mathbb{R}^d$ with $0 < \|G\|_{1,Q} := \int G dQ < \infty$,

$$\mathcal{N}(\epsilon\|G\|_{1,Q}, \mathcal{G}, \rho_{1,Q}) \leq A\epsilon^{-V},\ 0 < \epsilon < 1,$$

where $\rho_{1,Q}$ is the L_1 semimetric with respect to Q.

(b) Let S be an arbitrary index set. For each $s \in S$, let $\mathcal{G}(s)$ be a class of Borel measurable functions on $\mathbb{R}^d$. Suppose that there exists a Borel measurable function G such that $\sup_{s\in S}\sup_{g\in\mathcal{G}(s)}|g| \leq G$. The family $\{\mathcal{G}(s), s \in S\}$ is said to be *uniformly Euclidean* with envelope G if there exists positive constants A and V such that for every probability measure Q on $\mathbb{R}^d$ with $0 < \|G\|_{1,Q} < \infty$,

$$\sup_{s\in S}\mathcal{N}(\epsilon\|G\|_{1,Q}, \mathcal{G}(s), \rho_{1,Q}) \leq A\epsilon^{-V},\ 0 < \epsilon < 1.$$

Section 5 of Nolan & Pollard (1987) summarizes some basic facts on Euclidean classes. An example of a uniformly Euclidean family is $\{\mathcal{G}(s), s \in S\}$ such that each $\mathcal{G}(s)$ is a Vapnik-Červonenkis (VC) subgraph class with VC index less than some constant independent of s; use Theorem 2.6.7 of van der Vaart & Wellner (1996, abbreviated as vdVW hereafter) to check this. For the definitions of a VC subgraph class and a VC index, we refer to Section 2.6 of vdVW.

We now turn to the proof of Theorem 2.1. Recall that $\Delta(\mathbf{X}_{t-1};\boldsymbol{\vartheta}) = m(\mathbf{X}_{t-1};\boldsymbol{\vartheta}) - m(\mathbf{X}_{t-1};\boldsymbol{\theta})$. Define the stochastic process on $\mathbb{R}\times\Theta$

$$W_n(u_0,\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{t=1}^{n}\{K_{h_n}(e_t - u_0 - \Delta(\mathbf{X}_{t-1};\boldsymbol{\vartheta})) - K_{h_n}(e_t - u_0)\},\ (u_0,\boldsymbol{\vartheta}) \in \mathbb{R}\times\Theta.$$

As explained in Section 2, the proof of Theorem 2.1 is divided into two steps. The first step is to show that $W_n(u_0,\hat{\boldsymbol{\theta}}) - \mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\theta}}} = o_p(r_n^{-1})$ uniformly over $u_0 \in \mathbb{R}$. The second step is to show $\mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\theta}}} = O_p(n^{-1/2})$ uniformly over $u_0 \in \mathbb{R}$. Since $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$, Propositions 4.1 and 4.2 below will suffice for the first and second steps, respectively.

Proposition 4.1. *For every $l > 0$ and $\eta > 0$,*

$$\lim_{n\to\infty}\mathrm{P}\left(\sup_{(u_0,\boldsymbol{\vartheta})\in\mathbb{R}\times\Theta_n}\left|W_n(u_0,\boldsymbol{\vartheta}) - \mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]\right| > r_n^{-1}\eta\right) = 0, \tag{4.1}$$

where $\Theta_n = \{\boldsymbol{\vartheta} \in \Theta : \|\boldsymbol{\vartheta} - \boldsymbol{\theta}\| \leq ln^{-1/2}\}$.

The proof of Proposition 4.1 below relies on the blocking technique used in Yu (1993, 1994) and Arcones & Yu (1994). The blocking technique enables us to employ the symmetrization technique and an exponential inequality available in the iid case. Before the proof, we introduce some notations and a key lemme related to the blocking technique. Put $\xi_t = (e_t, \mathbf{X}_{t-1})$. Divide the n-sequence $\{1,\dots,n\}$ into blocks of length a_n with $a_n \to \infty$ and $a_n = o(n)$, one after the other:

$$\begin{aligned}H_k &= \{t : 2(k-1)a_n + 1 \leq t \leq (2k-1)a_n\},\\ T_k &= \{t : (2k-1)a_n + 1 \leq t \leq 2ka_n\},\end{aligned}$$

for $1 \leq k \leq \mu_n$, where $\mu_n = [n/(2a_n)]$. The exact form of a_n will be specified later. Put $\Xi_k = (\xi_t, t \in H_k)$, $1 \leq k \leq \mu_n$. With a slight abuse of notation, for a function

$g : \mathbb{R}^{p+1} \to \mathbb{R}$, we write $g(\Xi_k) = \sum_{t \in H_k} g(\xi_t)$. Let $\tilde{\Xi}_k = (\tilde{\xi}_t, t \in H_k)$, $1 \le k \le \mu_n$ be independent blocks such that each $\tilde{\Xi}_k$ has the same distribution as Ξ_1. The next lemma, which is a key to the blocking technique, is due to Volkonskii & Rozanov (1959) and Eberlein (1984). Lemma 4.1 is deduced from the second expression (2.1) of the β-mixing coefficient and the induction.

Lemma 4.1. *Work with the same notations as above. Assume that $n \ge 2a_n$ and $a_n \ge p+1$. For every Borel measurable subset A of $\mathbb{R}^{(p+1)a_n\mu_n}$, we have*

$$\left|\mathrm{P}\big((\Xi_1, \dots, \Xi_{\mu_n}) \in A\big) - \mathrm{P}\big((\tilde{\Xi}_1, \dots, \tilde{\Xi}_{\mu_n}) \in A\big)\right| \le \mu_n \beta(a_n - p).$$

The next lemma will be used in the proofs of Propositions 4.1 and 4.2.

Lemma 4.2. *For $a \in \mathbb{R}$, we have*

$$\int_{-\infty}^{\infty} |K(u+a) - K(u)| du \le C_K |a|, \ \int_{-\infty}^{\infty} |u|^{\lambda} |K(u+a) - K(u)| du \le C_K |a|(1 + |a|),$$

where the constant C_K depends only on $K(\cdot)$.

Proof. Let $a \ge 0$. Because of (A5), Fubini's theorem implies that

$$\int_{-\infty}^{\infty} |K(u+a) - K(u)| du = \int_{-\infty}^{\infty} \left| \int_u^{u+a} K'(v) dv \right| du$$
$$\le \int_{-\infty}^{\infty} \left\{ \int_{v-a}^{v} du \right\} |K'(v)| dv = a \int_{-\infty}^{\infty} |K'(v)| dv.$$

Similarly,

$$\int_{-\infty}^{\infty} |u|^{\lambda} |K(u+a) - K(u)| du = \int_{-\infty}^{\infty} |u|^{\lambda} \left| \int_u^{u+a} K'(v) dv \right| du$$
$$\le \int_{-\infty}^{\infty} \left\{ \int_{v-a}^{v} |u|^{\lambda} du \right\} |K'(v)| dv \le a \int_{-\infty}^{\infty} \{1 \vee (a + |v|)\} |K'(v)| dv.$$

The same argument applies for the $a < 0$ case. Therefore, the lemma holds with

$$C_K = 1 + \int_{-\infty}^{\infty} |K'(v)| dv + \int_{-\infty}^{\infty} |vK'(v)| dv.$$

□

We are now in position to prove Proposition 4.1.

Proof of Proposition 4.1. Without loss of generality, we may assume that $\{\boldsymbol{\vartheta} \in \mathbb{R}^q : \|\boldsymbol{\vartheta} - \boldsymbol{\theta}\| \le ln^{-1/2}\} \subset \Theta$. Also, we may assume that the inequality $|\Delta(\mathbf{x}; \boldsymbol{\vartheta})| \le M(\mathbf{x}) \|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|$ holds for all $\mathbf{x}$ and all $\boldsymbol{\vartheta} \in \Theta_n$. Otherwise, take n large enough; see conditions (A2) and (A3). Put $\kappa = \|K\|_{\infty}$. Define the functional class $\mathcal{G}_n = \{g_{u_0, \boldsymbol{\vartheta}, h_n} : u_0 \in \mathbb{R}, \boldsymbol{\vartheta} \in \Theta_n\}$, where

$$g_{u_0, \boldsymbol{\vartheta}, h}(e, \mathbf{x}) = K((e - u_0 - \Delta(\mathbf{x}; \boldsymbol{\vartheta}))/h) - K((e - u_0)/h), \ u_0 \in \mathbb{R}, \boldsymbol{\vartheta} \in \Theta, h > 0.$$

Since $(e, \mathbf{x}, u_0, \boldsymbol{\vartheta}) \mapsto g_{u_0,\boldsymbol{\vartheta},h_n}(e, \mathbf{x})$ is jointly Borel measurable, the class $\mathcal{G}_n$ is image admissible Suslin. Thus, the measurability problem will not occur throughout the proof; see Chapter 5.3 in Dudley (1999) and Appendix in Yu (1994). Put $\delta_n = h_n r_n^{-1} \eta$. It is not difficult to see that (4.1) is equivalent to

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{t=1}^{n} \{ g(\xi_t) - \mathrm{E}[g(\xi_t)] \} \right| > \delta_n \right) = 0. \tag{4.2}$$

We divide the proof of (4.2) into six steps.

Step 1. (Symmetrization) Let $\sigma_1, \ldots, \sigma_{\mu_n}$ be independent and uniformly distributed over $\{-1, 1\}$ and independent of $\tilde{D}_n := \{\tilde{\Xi}_k, 1 \le k \le \mu_n\}$. Take $a_n = [(\log n)^{-1} n^{1/4}]$. We shall show that

$$\limsup_{n\to\infty} \mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{t=1}^{n} \{ g(\xi_t) - \mathrm{E}[g(\xi_t)] \} \right| > \delta_n \right) \le \limsup_{n\to\infty} 8\mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{16} \right).$$

Observe that

$$\begin{aligned}
&\mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{t=1}^{n} \{ g(\xi_t) - \mathrm{E}[g(\xi_t)] \} \right| > \delta_n \right) \\
&\le \mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{t=1}^{2a_n\mu_n} \{ g(\xi_t) - \mathrm{E}[g(\xi_t)] \} \right| > \frac{\delta_n}{2} \right) \\
&\quad + \mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{t=2a_n\mu_n+1}^{n} \{ g(\xi_t) - \mathrm{E}[g(\xi_t)] \} \right| > \frac{\delta_n}{2} \right) \\
&\le 2\mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \{ g(\Xi_k) - \mathrm{E}[g(\Xi_k)] \} \right| > \frac{\delta_n}{4} \right) \\
&\quad + \mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{t=2a_n\mu_n+1}^{n} \{ g(\xi_t) - \mathrm{E}[g(\xi_t)] \} \right| > \frac{\delta_n}{2} \right).
\end{aligned} \tag{4.3}$$

Since $\sup_{g\in\mathcal{G}_n} \|g\|_\infty \le 2\kappa$, the second term of the right hand side of (4.3) is zero for large n. On the other hand, in view of Lemma 4.1, for large n, we may bound the first term of the right hand side of (4.3) by

$$2\mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \{ g(\tilde{\Xi}_k) - \mathrm{E}[g(\tilde{\Xi}_k)] \} \right| > \frac{\delta_n}{4} \right) + 2\mu_n \beta(a_n - p).$$

Since $\beta(m) \to 0$ exponentially fast as $m \to \infty$, $\mu_n \beta(a_n - p) \to 0$ as $n \to \infty$. To bound the first term, we use the symmetrization technique. Because $\tilde{\Xi}_k$ $(1 \le k \le \mu_n)$ are iid blocks, Lemma 2.3.7 of vdVW implies that

$$\zeta_n \mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \{ g(\tilde{\Xi}_k) - \mathrm{E}[g(\tilde{\Xi}_k)] \} \right| > \frac{\delta_n}{4} \right) \le 2\mathrm{P}\left(\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{16} \right),$$

where $\zeta_n = 1-(64\mu_n/(n\delta_n)^2)\sup_{g\in\mathcal{G}_n}\mathrm{E}[\{g(\tilde{\Xi}_1)\}^2]$. We will show that $\sup_{g\in\mathcal{G}_n}\mathrm{E}[\{g(\tilde{\Xi}_1)\}^2] = O(a_n n^{-1/2})$. By stationarity,

$$\mathrm{E}[\{g(\tilde{\Xi}_1)\}^2] = a_n \mathrm{E}[g^2(\xi_1)] + 2a_n \sum_{i=1}^{a_n-1}(1-i/a_n)\mathrm{E}[g(\xi_1)g(\xi_{1+i})].$$

By the usual change of variables and Lemma 4.2, $\mathrm{E}[|g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_1)|]$ is bounded by a constant times $\|\boldsymbol{\vartheta}-\boldsymbol{\theta}\|$, which, together with the fact that $\sup_{g\in\mathcal{G}_n}\|g\|_\infty \le 2\kappa$, implies that $\mathrm{E}[g^2_{u_0,\boldsymbol{\vartheta},h_n}(\xi_1)] = O(n^{-1/2})$ uniformly over $(u_0,\boldsymbol{\vartheta}) \in \mathbb{R}\times\Theta_n$. On the other hand, invoke that $\mathrm{E}[|g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_1)g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_{1+i})|] = \mathrm{E}[|g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_1)|\cdot\mathrm{E}[|g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_{1+i})||\mathbf{X}_i,\xi_1]]$ and $\mathrm{E}[|g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_{1+i})||\mathbf{X}_i,\xi_1] \le$ const. $\times\, M(\mathbf{X}_i)\|\boldsymbol{\vartheta}-\boldsymbol{\theta}\|$. Truncating $M(\mathbf{X}_i)$ at $n^{1/4}$ and using the fact that $\mathrm{E}[M^2(\mathbf{X}_i)] < \infty$, we may show that $\mathrm{E}[|g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_1)g_{u_0,\boldsymbol{\vartheta},h_n}(\xi_{1+i})|] = O(n^{-3/4})$ uniformly over $(u_0,\boldsymbol{\vartheta}) \in \mathbb{R}\times\Theta_n$ and $i \ge 1$. Since $a_n = o(n^{1/4})$, we conclude that $\sup_{g\in\mathcal{G}_n}\mathrm{E}[\{g(\tilde{\Xi}_1)\}^2] = O(a_n n^{-1/2})$. Thus, we have $\zeta_n = 1-O(n^{-3/2}h_n^{-2}r_n^2) = 1-o(1)$, which implies that $\zeta_n \ge 1/2$ for large n. Therefore, for large n,

$$\mathrm{P}\left(\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{k=1}^{\mu_n}\{g(\tilde{\Xi}_k)-\mathrm{E}[g(\tilde{\Xi}_k)]\}\right| > \frac{\delta_n}{4}\right) \le 4\mathrm{P}\left(\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k g(\tilde{\Xi}_k)\right| > \frac{\delta_n}{16}\right),$$

which leads to the conclusion of Step 1.

Step 2. Let $\mathcal{G}_n(\boldsymbol{\vartheta})$ be the section of $\mathcal{G}_n$ at $\boldsymbol{\vartheta}\in\Theta_n$, that is, $\mathcal{G}_n(\boldsymbol{\vartheta}) = \{g_{u_0,\boldsymbol{\vartheta},h_n} : u_0 \in \mathbb{R}\}$. Clearly, $\mathcal{G}_n = \bigcup_{\boldsymbol{\vartheta}\in\Theta_n}\mathcal{G}_n(\boldsymbol{\vartheta})$. We may cover Θ_n by a finite number of open balls $B_{n,j}$ in $\mathbb{R}^q$ centered at $\boldsymbol{\vartheta}_{n,j}$ in such a way that

$$\Theta_n \subset \bigcup_{j=1}^{m_n} B_{n,j},\ \sup_{\boldsymbol{\vartheta}\in B_{n,j}}\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_{n,j}\| \le \frac{h_n\delta_n}{\log n},\ m_n = O\left\{\left(\frac{\log n}{n^{1/2}h_n\delta_n}\right)^q \vee 1\right\}.$$

We shall show that

$$\begin{aligned}\limsup_{n\to\infty}\mathrm{P}&\left(\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k g(\tilde{\Xi}_k)\right| > \frac{\delta_n}{16}\right)\\ &\le \limsup_{n\to\infty}\left[\sum_{j=1}^{m_n}\mathrm{P}\left(\sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta}_{n,j})}\left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k g(\tilde{\Xi}_k)\right| > \frac{\delta_n}{32}\right)\right]. \quad (4.4)\end{aligned}$$

Suppose for a moment that $\boldsymbol{\vartheta}\in\Theta_n$ is arbitrarily fixed. By definition, there exists a $j \in \{1,\dots,m_n\}$ such that $\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_{n,j}\| \le h_n\delta_n/\log n$. For each $g = g_{u_0,\boldsymbol{\vartheta},h_n}\in\mathcal{G}_n(\boldsymbol{\vartheta})$, pick $\bar{g} = g_{u_0,\boldsymbol{\vartheta}_{n,j},h_n}\in\mathcal{G}_n(\boldsymbol{\vartheta}_{n,j})$. Because of the Lipschitz continuity of $K(\cdot)$, $|g(e,\mathbf{x})-\bar{g}(e,\mathbf{x})| \le Lh_n^{-1}|m(\mathbf{x};\boldsymbol{\vartheta})-m(\mathbf{x};\boldsymbol{\vartheta}_{n,j})| \le Lh_n^{-1}M(\mathbf{x})\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_{n,j}\| \le L\delta_n M(\mathbf{x})/\log n$, where $L>0$ is a Lipschitz constant of $K(\cdot)$. Thus, we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k g(\tilde{\Xi}_k)\right| &\le \left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k \bar{g}(\tilde{\Xi}_k)\right| + \frac{1}{n}\sum_{k=1}^{\mu_n}|g(\tilde{\Xi}_k)-\bar{g}(\tilde{\Xi}_k)|\\ &\le \left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k \bar{g}(\tilde{\Xi}_k)\right| + \frac{L\delta_n}{2\log n}\cdot\frac{1}{\mu_n}\sum_{k=1}^{\mu_n}\bar{M}_{n,k},\end{aligned}$$

where $\bar{M}_{n,k} = a_n^{-1} \sum_{t \in H_k} M(\tilde{\mathbf{X}}_{t-1})$.

Because $\mathcal{G}_n = \bigcup_{\boldsymbol{\vartheta} \in \Theta_n} \mathcal{G}_n(\boldsymbol{\vartheta})$, the preceding argument shows that

$$\begin{aligned}
&\mathrm{P}\left(\sup_{g \in \mathcal{G}_n} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{16} \right) \\
&\quad \le \mathrm{P}\left(\max_{1 \le j \le m_n} \sup_{g \in \mathcal{G}_n(\boldsymbol{\vartheta}_{n,j})} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{32} \right) + \mathrm{P}\left(\frac{L\delta_n}{2\log n} \cdot \frac{1}{\mu_n} \sum_{k=1}^{\mu_n} \bar{M}_{n,k} > \frac{\delta_n}{32} \right) \\
&\quad \le \sum_{j=1}^{m_n} \mathrm{P}\left(\sup_{g \in \mathcal{G}_n(\boldsymbol{\vartheta}_{n,j})} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{32} \right) + \mathrm{P}\left(\frac{1}{\mu_n} \sum_{k=1}^{\mu_n} \bar{M}_{n,k} > \frac{\log n}{16L} \right).
\end{aligned}$$

By Markov's inequality,

$$\mathrm{P}\left(\frac{1}{\mu_n} \sum_{k=1}^{\mu_n} \bar{M}_{n,k} > \frac{\log n}{16L} \right) \le \frac{16L\mathrm{E}[\bar{M}_{n,k}]}{\log n} = \frac{16L\mathrm{E}[M(\mathbf{X}_0)]}{\log n} \to 0.$$

Therefore, we obtain (4.4).

Step 3. Fix $\boldsymbol{\vartheta} \in \Theta_n$. Let $\tilde{\rho}_{1,n}$ be the L_1 semimetric with respect to the empirical distribution on $\mathbb{R}^{p+1}$ that assigns probability $1/(a_n\mu_n)$ to each $\tilde{\xi}_t$. We shall show that for each $c > 0$,

$$\begin{aligned}
\mathrm{P}\left(\sup_{g \in \mathcal{G}_n(\boldsymbol{\vartheta})} \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{32} \right) &\le 2n^{-c}\mathrm{E}[\mathcal{N}(\delta_n/32, \mathcal{G}_n(\boldsymbol{\vartheta}), \tilde{\rho}_{1,n})] \\
&\quad + \mathrm{P}\left(\sup_{g \in \mathcal{G}_n(\boldsymbol{\vartheta})} \frac{1}{n} \sum_{k=1}^{\mu_n} \{g(\tilde{\Xi}_k)\}^2 > \frac{n\delta_n^2}{2 \cdot 64^2 c \log n} \right). \qquad (4.5)
\end{aligned}$$

Let $\mathcal{G}_{n,\delta_n/32}$ be a minimal $(\delta_n/32)$-cover of $\mathcal{G}_n(\boldsymbol{\vartheta})$ with respect to $\tilde{\rho}_{1,n}$. For every $g \in \mathcal{G}_n(\boldsymbol{\vartheta})$, there exists a $\bar{g} \in \mathcal{G}_{n,\delta_n/32}$ such that

$$\frac{1}{n} \sum_{k=1}^{\mu_n} |g(\tilde{\Xi}_k) - \bar{g}(\tilde{\Xi}_k)| \le \frac{a_n\mu_n}{n} \tilde{\rho}_{1,n}(g, \bar{g}) < \frac{\delta_n}{64},$$

which leads to

$$\begin{aligned}
\left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| &\le \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k \bar{g}(\tilde{\Xi}_k) \right| + \frac{1}{n} \sum_{k=1}^{\mu_n} |g(\tilde{\Xi}_k) - \bar{g}(\tilde{\Xi}_k)| \\
&\le \left| \frac{1}{n} \sum_{k=1}^{\mu_n} \sigma_k \bar{g}(\tilde{\Xi}_k) \right| + \frac{\delta_n}{64}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathrm{P}\left(\sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta})} \left| \frac{1}{n}\sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{32} \,\middle|\, \tilde{D}_n \right) \\
&\quad\le \mathrm{P}\left(\max_{g\in\mathcal{G}_{n,\delta_n/32}} \left| \frac{1}{n}\sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{64} \,\middle|\, \tilde{D}_n \right) \\
&\quad\le \mathcal{N}(\delta_n/32, \mathcal{G}_n(\boldsymbol{\vartheta}), \tilde{\rho}_{1,n}) \sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta})} \mathrm{P}\left(\left| \frac{1}{n}\sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{64} \,\middle|\, \tilde{D}_n \right) \wedge 1. \qquad (4.6)
\end{aligned}$$

It remains to bound the right hand side of (4.6). By Hoeffding's inequality, we have

$$\sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta})} \mathrm{P}\left(\left| \frac{1}{n}\sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{64} \,\middle|\, \tilde{D}_n \right) \le 2\exp\left(-\frac{n\delta_n^2}{2\cdot 64^2 w_n} \right),$$

where $w_n = \sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta})}[n^{-1}\sum_{k=1}^{\mu_n}\{g(\tilde{\Xi}_k)\}^2]$. Define the event

$$A_n = \left\{ w_n > \frac{n\delta_n^2}{2\cdot 64^2 c\log n} \right\}.$$

Taking the expectation of both sides of (4.6), we have

$$\begin{aligned}
&\mathrm{P}\left(\sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta})} \left| \frac{1}{n}\sum_{k=1}^{\mu_n} \sigma_k g(\tilde{\Xi}_k) \right| > \frac{\delta_n}{32} \right) \\
&\quad\le 2\mathrm{E}\left[\mathcal{N}(\delta_n/32, \mathcal{G}_n(\boldsymbol{\vartheta}), \tilde{\rho}_{1,n}) \exp\left(-\frac{n\delta_n^2}{2\cdot 64^2 w_n} \right) I(A_n^c) \right] + \mathrm{P}(A_n) \\
&\quad\le 2n^{-c}\mathrm{E}[\mathcal{N}(\delta_n/32, \mathcal{G}_n(\boldsymbol{\vartheta}), \tilde{\rho}_{1,n})] + \mathrm{P}(A_n),
\end{aligned}$$

which leads to the conclusion of Step 3.

Step 4. Fix $c > 0$. Take $(\delta_n')^2 = n^2\delta_n^2/(2\cdot 64^3 c\kappa^2 a_n^2 \mu_n \log n)$. We shall show that there exists a positive integer n_0 such that for $n \ge n_0$, the inequality

$$\mathrm{P}\left(\sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta})} \frac{1}{n}\sum_{k=1}^{\mu_n}\{g(\tilde{\Xi}_k)\}^2 > \frac{n\delta_n^2}{2\cdot 64^2 c\log n} \right) \le 4e^{-\mu_n(\delta_n')^2}\mathrm{E}[\mathcal{N}(\kappa(\delta_n')^2, \mathcal{G}_n(\boldsymbol{\vartheta}), \tilde{\rho}_{1,n})] \qquad (4.7)$$

holds for every $\boldsymbol{\vartheta} \in \Theta_n$. We note that n_0 depends only on c and η.

To show this, we make use of Lemma II 33 of Pollard (1984), which is sometimes referred to as the "square root trick". Let

$$\mathcal{H}_n(\boldsymbol{\vartheta}) = \{(\xi_1, \dots, \xi_{a_n}) \mapsto \textstyle\sum_{j=1}^{a_n} g(\xi_j)/(2a_n\kappa) : g \in \mathcal{G}_n(\boldsymbol{\vartheta})\}.$$

Let $\tilde{\rho}_{2,\mu_n}$ be the L_2 semimetric with respect to the empirical distribution on $\mathbb{R}^{(p+1)a_n}$that assigns probability $1/\mu_n$ to each $\tilde{\Xi}_k$. Since $\tilde{\Xi}_k$ $(1 \le k \le \mu_n)$ are iid blocks and $\sup_{\varphi\in\mathcal{H}_n(\boldsymbol{\vartheta})}\|\varphi\|_\infty \le 1$, Lemma II 33 of Pollard (1984) shows that

$$\begin{aligned}
&\delta \ge \sup_{\varphi\in\mathcal{H}_n(\boldsymbol{\vartheta})} (\mathrm{E}[\varphi^2(\tilde{\Xi}_1)])^{1/2} \\
&\qquad \Rightarrow \mathrm{P}\left(\sup_{\varphi\in\mathcal{H}_n(\boldsymbol{\vartheta})} \frac{1}{\mu_n}\sum_{k=1}^{\mu_n} \varphi^2(\tilde{\Xi}_k) > 64\delta^2 \right) \le 4\mathrm{E}[\mathcal{N}(\delta, \mathcal{H}_n(\boldsymbol{\vartheta}), \tilde{\rho}_{2,\mu_n})e^{-\mu_n\delta^2} \wedge 1]. \qquad (4.8)
\end{aligned}$$

Observe that for $\varphi_i(\xi_1,\ldots,\xi_{a_n}) = \sum_{j=1}^{a_n} g_i(\xi_j)/(2a_n\kappa),\ g_i \in \mathcal{G}_n,\ i=1,2,$

$$\begin{aligned}\{\tilde{\rho}_{2,\mu_n}(\varphi_1,\varphi_2)\}^2 &= \frac{1}{4a_n^2\mu_n\kappa^2}\sum_{k=1}^{\mu_n}\{g_1(\tilde{\Xi}_k)-g_2(\tilde{\Xi}_k)\}^2 \\ &\le \frac{1}{a_n\mu_n\kappa}\sum_{k=1}^{\mu_n}|g_1(\tilde{\Xi}_k)-g_2(\tilde{\Xi}_k)| \\ &\le \frac{1}{\kappa}\tilde{\rho}_{1,n}(g_1,g_2). \end{aligned} \tag{4.9}$$

From (4.8) and (4.9), we may infer that

$$\begin{aligned}&\delta \ge \sup_{\varphi\in\mathcal{H}_n(\boldsymbol{\vartheta})}(\mathrm{E}[\varphi^2(\tilde{\Xi}_1)])^{1/2} \\ &\quad\Rightarrow \mathrm{P}\left(\sup_{\varphi\in\mathcal{H}_n(\boldsymbol{\vartheta})}\frac{1}{\mu_n}\sum_{k=1}^{\mu_n}\varphi^2(\tilde{\Xi}_k) > 64\delta^2\right) \le 4\mathrm{E}[\mathcal{N}(\kappa\delta^2,\mathcal{G}_n(\boldsymbol{\vartheta}),\tilde{\rho}_{1,n})e^{-\mu_n\delta^2}\wedge 1]. \end{aligned}\tag{4.10}$$

The probability in (4.10) is equal to the left hand side of (4.7) when $\delta = \delta_n'$. Since $(\delta_n')^2 \asymp nh_n^2/(r_n^2 a_n \log n)$ and $(\delta_n'')^2 := \sup_{\boldsymbol{\vartheta}\in\Theta_n}\sup_{\varphi\in\mathcal{H}_n(\boldsymbol{\vartheta})}\mathrm{E}[\varphi^2(\tilde{\Xi}_1)] = O(a_n^{-1}n^{-1/2})$, there exists a positive integer n_0 such that $\delta_n' \ge \delta_n''$ for $n \ge n_0$; use (2.3). Therefore, we obtain the conclusion of Step 4.

Step 5. There exist positive constants A and V such that for some $\epsilon_0 > 0$

$$\sup_{\boldsymbol{\vartheta}\in\Theta_n}\mathcal{N}(\epsilon,\mathcal{G}_n(\boldsymbol{\vartheta}),\tilde{\rho}_{1,n}) \le A\epsilon^{-V},\ 0<\epsilon<\epsilon_0,\ n\ge 1.$$

For each $\boldsymbol{\vartheta}\in\Theta$, define the functional class $\mathcal{G}(\boldsymbol{\vartheta}) = \{g_{u_0,\boldsymbol{\vartheta},h} : u_0 \in \mathbb{R}, h>0\}$. Since $\mathcal{G}_n(\boldsymbol{\vartheta})$ is a subset of $\mathcal{G}(\boldsymbol{\vartheta})$, it suffices to show that the family $\{\mathcal{G}(\boldsymbol{\vartheta}),\boldsymbol{\vartheta}\in\Theta\}$ is uniformly Euclidean with some constant envelope. We first look that for each fixed $\boldsymbol{\vartheta}\in\Theta$ the functional class $\{(e,\mathbf{x})\mapsto(e-u_0-\Delta(\mathbf{x};\boldsymbol{\vartheta}))/h : u_0\in\mathbb{R}, h>0\}$ is a subset of a vector space of functions on $\mathbb{R}^{p+1}$ spanned by 1 and $e-\Delta(\mathbf{x};\boldsymbol{\vartheta})$. By Lemma 2.6.15 of vdVW, the class is a VC subgraph class with VC index smaller than or equal to 4. Because of (A5), $K(\cdot)$ is of bounded variation and thus can be written as the difference of two bounded, non-decreasing functions, $K(\cdot)=\phi(\cdot)-\psi(\cdot)$, say. By Lemma 2.6.18 (viii) of vdVW, the functional classes $\{(e,\mathbf{x})\mapsto\phi((e-u_0-\Delta(\mathbf{x};\boldsymbol{\vartheta}))/h) : u_0\in\mathbb{R},h>0\}$, $\{(e,\mathbf{x})\mapsto\psi((e-u_0-\Delta(\mathbf{x};\boldsymbol{\vartheta}))/h) : u_0\in\mathbb{R},h>0\}$ are VC subgraph classes with VC indices smaller than or equal to 4. Therefore, Theorem 2.6.7 of vdVW and Lemma 16 of Nolan & Pollard (1987) imply that the family $\{\mathcal{G}(\boldsymbol{\vartheta}),\boldsymbol{\vartheta}\in\Theta\}$ is uniformly Euclidean with some constant envelope.

Step 6. (Conclusion) From Steps 1 and 2, it remains to show that the right hand side of (4.4) is zero. Because of (2.3), there exist positive constants C_1 and α such that $m_n \le C_1 n^\alpha$. Similarly, in view of Step 5, there exist positive constants C_2 and β such that $\sup_{\boldsymbol{\vartheta}\in\Theta_n}\mathcal{N}(\delta_n/32,\mathcal{G}_n(\boldsymbol{\vartheta}),\tilde{\rho}_{1,n}) \le C_2 n^\beta$ in Step 3. Take $c=\alpha+\beta+1$ in Steps 3 and 4. Because of Step 5, there exist positive constants C_3 and γ such that

$\sup_{\boldsymbol{\vartheta}\in\Theta_n}\mathcal{N}(\kappa(\delta_n')^2,\mathcal{G}_n(\boldsymbol{\vartheta}),\tilde{\rho}_{1,n}) \le C_3 n^\gamma$ in Step 4. With this choice of c, for large n, we have

$$\sum_{j=1}^{m_n} \mathrm{P}\left(\sup_{g\in\mathcal{G}_n(\boldsymbol{\vartheta}_{n,j})}\left|\frac{1}{n}\sum_{k=1}^{\mu_n}\sigma_k g(\tilde{\Xi}_k)\right| > \frac{\delta_n}{32}\right) \le 2C_1C_2n^{-1} + 4C_1C_3n^{\alpha+\gamma}e^{-\mu_n(\delta_n')^2}. \quad (4.11)$$

Since $\mu_n(\delta_n')^2/\log n \asymp n^{3/2}h_n^2/r_n^2$ and $n^{3/2}h_n^2/r_n^2 \to \infty$, the second term of the right hand side of (4.11) goes to zero as $n \to \infty$. Therefore, we complete the proof. □

Remark 4.1. If the functional class $\{g_{u_0,\boldsymbol{\vartheta},h} : u_0 \in \mathbb{R}, \boldsymbol{\vartheta} \in \Theta, h > 0\}$ is Euclidean, we may apply the arguments in Steps 3 and 4 directly to $\mathcal{G}_n$ instead of $\mathcal{G}_n(\boldsymbol{\vartheta})$. In this case, Step 2 may be skipped and the Lipschitz continuity of $K(\cdot)$ may be dropped. The proof could be shortened if we put this assumption. However, it restricts the class of $m(\mathbf{x};\boldsymbol{\vartheta})$ in a less explicit manner. That is why we do not put such an assumption.

Proposition 4.2. *For every $l > 0$, $\mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})] = O(n^{-1/2}h_n^{\lambda-1} \wedge 1)$ uniformly over $(u_0,\boldsymbol{\vartheta}) \in \mathbb{R}\times\Theta_n$, where $\Theta_n = \{\boldsymbol{\vartheta}\in\Theta : \|\boldsymbol{\vartheta}-\boldsymbol{\theta}\| \le ln^{-1/2}\}$.*

Proof. Since $K(\cdot)$ is a density function,

$$\begin{aligned}\mathrm{E}[K_{h_n}(e_t - u_0 - \Delta(\mathbf{X}_{t-1};\boldsymbol{\vartheta})) \mid \mathbf{X}_{t-1}] - f(u_0)&\\ = \int_{-\infty}^{\infty} K(u - h_n^{-1}\Delta(\mathbf{X}_{t-1};\boldsymbol{\vartheta}))\{f(u_0+uh_n) - f(u_0)\}du.& \quad (4.12)\end{aligned}$$

Since f is λ-th Hölder continuous, the absolute value of $\mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]$ is bounded by

$$\text{const.} \times h_n^\lambda \mathrm{E}\left[\int_{-\infty}^{\infty}|u|^\lambda|K(u-h_n^{-1}\Delta(\mathbf{X}_{t-1};\boldsymbol{\vartheta})) - K(u)|du\right]. \quad (4.13)$$

Because of Lemma 4.2, (4.13) is of order $O(n^{-1/2}h_n^{\lambda-1})$ uniformly over $(u_0,\boldsymbol{\vartheta}) \in \mathbb{R}\times\Theta_n$. On the other hand, it is not difficult to see that $|\mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]| \le 2\|f\|_\infty$. Therefore, we obtain the desired result. □

Proof of Theorem 2.1. Observe that

$$\hat{f}_n(u_0) - f_n(u_0) = \{W_n(u_0,\hat{\boldsymbol{\theta}}) - \mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\theta}}}\} + \mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\theta}}}. \quad (4.14)$$

Recall that $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$. By Proposition 4.1 and 4.2, the first term of the right hand side of (4.14) is $o_p(r_n^{-1})$ and the second term is $O_p(n^{-1/2}h_n^{\lambda-1}\wedge 1)$ uniformly over $u_0 \in \mathbb{R}$. Therefore, we obtain the desired result. □

4.2 Proof of Corollary 2.1

We follow the notations used in the proof of Theorem 2.1. We may take $r_n = n^{1/2}$ in Proposition 4.1. From (4.14), we have

$$\sqrt{n}\{\hat{f}_n(u_0) - f_n(u_0)\} = \sqrt{n}\mathrm{E}[W_n(u_0,\boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\theta}}} + o_p(1),$$

where the equality holds uniformly over $u_0 \in \mathbb{R}$. It remains to show that the equality

$$\sqrt{n}\mathrm{E}[W_n(u_0, \boldsymbol{\vartheta})]|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\theta}}} = \sqrt{n} f'(u_0)\mathrm{E}[\dot{m}(\mathbf{X}_{t-1}; \boldsymbol{\theta})](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top + o_p(1)$$

holds uniformly over $u_0 \in \mathbb{R}$. Because of (A4'),

$$|f(u_0 + uh) - f(u_0) - f'(u_0)uh_n| \le C_f(uh_n)^2,$$

where $C_f = \|f''\|_\infty/2$. Using the identity (4.12), we have

$$\begin{aligned}\left|\mathrm{E}[W_n(u_0, \boldsymbol{\vartheta})] - f'(u_0)h_n\mathrm{E}\left[\int_{-\infty}^{\infty} u\{K(u - h_n^{-1}\Delta(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})) - K(u)\}du\right]\right| \\ \le C_f h_n^2 \mathrm{E}\left[\int_{-\infty}^{\infty} |u^2\{K(u - h_n^{-1}\Delta(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})) - K(u)\}|du\right].\end{aligned}$$

For each $l > 0$, the right hand side is shown to be of order $O(n^{-1/2}h_n)$ uniformly over $\boldsymbol{\vartheta} \in \Theta_n := \{\boldsymbol{\vartheta} \in \Theta : \|\boldsymbol{\vartheta} - \boldsymbol{\theta}\| \le ln^{-1/2}\}$; use the similar inequality as that in Lemma 4.2. On the other hand, a direct calculation shows that

$$\begin{aligned}&h_n \int_{-\infty}^{\infty} u\{K(u - h_n^{-1}\Delta(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})) - K(u)\}du \\ &= -\Delta(\mathbf{X}_{t-1}; \boldsymbol{\vartheta}) \int_{-\infty}^{\infty} uK'(u)du - \frac{\Delta^2(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})}{2h_n} \int_{-\infty}^{\infty} K'(u)du \\ &= \Delta(\mathbf{X}_{t-1}; \boldsymbol{\vartheta}).\end{aligned}$$

Because of (A3'), the map $\boldsymbol{\vartheta} \mapsto \mathrm{E}[m(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})]$ is continuously differentiable in a neighborhood of $\boldsymbol{\theta}$ with $\partial\mathrm{E}[m(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})]/\partial\boldsymbol{\vartheta} = \mathrm{E}[\dot{m}(\mathbf{X}_{t-1}; \boldsymbol{\vartheta})]$. Therefore, uniformly over $(u_0, \boldsymbol{\vartheta}) \in \mathbb{R} \times \Theta_n$,

$$\mathrm{E}[W_n(u_0, \boldsymbol{\vartheta})] = f'(u_0)\mathrm{E}[\dot{m}(\mathbf{X}_{t-1}; \boldsymbol{\theta})](\boldsymbol{\vartheta} - \boldsymbol{\theta})^\top + o(n^{-1/2}).$$

Since $\hat{\boldsymbol{\theta}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}$, we obtain the desired result. □

References

An, H.Z. & Huang, F.C. (1996). The geometric ergodicity of nonlinear autoregressive models. *Statist. Sinica* **6**, 943-956.

Arcones, M.A. & Yu, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *J. Theoret. Probab.* **7**, 47-71.

Bickel, P.J. & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1**, 1071-1095. [Corrections: **3**, (1975) 1370.]

Brockwell, P.J. & Davies, R.A. (1991). *Time Series: Theory and Methods*, 2nd eds. Springer-Verlag, New York.

Carrasco, M. & Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17-39.

Cheng, F. (2005). Asymptotic distributions of error density estimators in first-order autoregressive models. *Sankhya* **67**, 553-567.

Deheuvels, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals. In *Recent Advaces in Reliability Theory: Methodology, Practice and Inference* (eds. N. Limnios & M. Nikulin), 477-492, Birkhauser, Boston.

Dudley, R.M. (1999). *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge.

Eberlein, E. (1984). Weak convergence of partial sums of absolutely regular sequences. *Statist. Probab. Lett.* **2**, 291-293.

Einmahl, U. & Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel type function estimators. *J. Theoret. Probab.* **13**, 1-37.

Einmahl, U. & Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33**, 1380-1403.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* **50**, 987-1008.

Fan, J. & Yao, Q. (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer-Verlag, New York.

Giné, E. & Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **38**, 907-921.

Haggan, V. & Ozaki, T. (1981). Modelling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189-196.

Hall, P. & Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71**, 285-317.

Huang, D., Wang, H. & Yao, Q. (2008). Estimating GARCH models: when to use what? *Econometrics J.* **11**, 27-38.

Khmaladze, E.V. & Koul, H.L. (2004). Martingale transforms goodness-of-fit tests in regression models. *Ann. Statist.* **32**, 995-1034.

Klimko, L.A. & Nelson, P.L. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* **6**, 629-643.

Koul, H.L. (1996). Asymptotics of some estimators and sequential residual empiricals in nonlinear autoregressive models. *Ann. Statist.* **24**, 380-404.

Liebscher, E. (1996). Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stoch. Process. Appl.* **65**, 69-80.

Liebscher, E. (1999). Estimating the density of the residuals in autoregressive models. *Stat. Inference Stoch. Process.* **2**, 105-117.

Liebscher, E. (2003). Strong convergence of estimators in nonlinear autoregressive models. *J. Multivariate Anal.* **84**, 247-261.

Masry, E. & Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory* **11**, 258-289.

Müller, U.U., Schick, A. & Wefelmeyer, W. (2005). Weighted residual-based density estimators for nonlinear autoregressive models. *Statist. Sinica* **15**, 177-195.

Nolan, D. & Pollard, D. (1987). *U*-processes: rates of convergence. *Ann. Statist.* **15**, 780-799.

Nummelin, E. & Tuominen, P. (1982). Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. *Stochastic Process. Appl.* **12**, 187-202.

Ozaki, T. (1980). Non-linear time series models for nonlinear random variables. *J. Appl. Probab.* **17**, 84-93.

Peng, L. & Yao, Q. (2003). Least absolute deviations for ARCH and GARCH models. *Biometrika* **90**, 967-975.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

Silverman, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177-184.

Stute, W. (1984). The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab.* **22**, 361-379.

Tjøstheim, D. (1986). Estimation in nonlinear time series models. *Stochastic Process. Appl.*. **21**, 251-273.

Tong, H. & Lim, K.S. (1980). Threshold autoregression, limit cycles and cyclical data (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **42**, 245-292.

van der Vaart, A.W. & Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.

Volkonskii, V.A. & Rozanov, Y.A. (1959). Some limit theorems for random functions I. *Theory Probab. Appl.* **4**, 178-197.

Weiss, A.A. (1986). Asymptotic theory of ARCH models: estimation and testing. *Econometric Theory* **2**, 107-131.

Yu, B. (1993). Density estimation in the L^∞ norm for dependent data with applications to the Gibbs sampler. *Ann. Statist.* **21**, 711-735.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.* **22**, 94-116.