Parametric Specification Test for Nonlinear Autoregressive Models

June 8, 2014

By Kun Ho Kim\textsuperscript{a}, Ting Zhang\textsuperscript{b} and Wei Biao Wu\textsuperscript{c}

\textsuperscript{a}Hanyang University, \textsuperscript{b}Boston University and \textsuperscript{c}University of Chicago

Abstract: The paper considers testing parametric assumptions on the conditional mean and variance functions for nonlinear autoregressive models. To this end, we compare the kernel density estimate of the marginal density of the process with a convolution-type density estimate. It is shown that, interestingly, the latter estimate has a parametric ($\sqrt{n}$) rate of convergence, thus substantially improving the classical kernel density estimates whose rates of convergence are much inferior. Our results are confirmed by a simulation study for threshold autoregressive processes and autoregressive conditional heteroskedastic processes.

Key words: ARCH process, Bahadur representation, Convolution, Density estimation, Ergodicity, Functional central limit theorem, Gaussian process, Kernel estimator, Martingale, Nonlinear time series, Specification test.

1 Introduction

Consider the nonlinear autoregressive model

$$X_i = \mu(X_{i-1}) + \sigma(X_{i-1})\epsilon_i, \quad i \in \mathbb{Z},$$

where $\mu : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ are unknown conditional mean and variance functions respectively, and $\epsilon_k, k \in \mathbb{Z}$, are independent and identically distributed (iid) innovations. Assuming certain parametric forms on $\mu(\cdot)$ and $\sigma(\cdot)$, model (1) covers many popular time series models as special cases including the linear autoregressive (AR) process $X_i = aX_{i-1} + \epsilon_i$, the threshold autoregressive (TAR) process $X_i = a\max(X_{i-1}, 0) + b\min(X_{i-1}, 0) + \epsilon_i$ by Tong (1990), the autoregressive conditional heteroscedastic (ARCH) process $X_i =$
\( \sqrt{a^2 + b^2 X_{i-1}^2} \epsilon_i \) by Engle (1982), the exponential autoregressive (EAR) process \( X_i = \{a + b \exp(-cX_{i-1})\}X_{i-1} + \epsilon_i \) by Jones (1976) among others. Although the aforementioned parametric models have been widely used in the literature due to their ease of interpretation and prediction, they may suffer from misspecification leading to erroneous conclusions. The primary goal of the paper is to develop a rigorous test for parametric assumptions on the conditional mean and variance functions in (1), namely we are interested in testing the null hypothesis

\[ H_0 : \{\mu(\cdot), \sigma(\cdot)\} = \{\mu_\theta(\cdot), \sigma_\theta(\cdot)\} \text{ holds for some } \theta \in \Theta, \tag{2} \]

where \( \mu_\theta(\cdot) \) and \( \sigma_\theta(\cdot) \) have a known parametric form, \( \theta \) is a finite dimensional parameter vector and \( \Theta \) is a parameter space. Throughout the paper we denote by \( \theta_0 \) the true parameter and assume that the process \((X_i)\) is stationary.

Aït-Sahalia (1996) proposed a parametric specification test by comparing the nonparametric kernel density estimate of the marginal density of \( X_i \) with its closed form solution under the null hypothesis. Specifically, suppose we observe \( X_1, \ldots, X_n \), then the marginal density of \( X_i \) can be estimated nonparametrically by

\[ \hat{f}(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K \left( \frac{x - X_i}{b_n} \right), \tag{3} \]

where \( K \) is a kernel function and \( b_n \) is a bandwidth sequence satisfying \( b_n \to 0 \) and \( nb_n \to \infty \). Nonparametric density estimation has been extensively studied in statistics; see for example Robinson (1983), Silverman (1986), Scott (1992) and Fan and Yao (2005) among others. We remark that the estimate \( \hat{f}(x) \) in (3) does not require specific structural assumptions of the process \((X_i)\). On the other hand, however, if \( \mu(\cdot) \) and \( \sigma(\cdot) \) in (1) falls into certain parametric forms so that the marginal density of \( X_i \) has a closed form solution \( f_\theta(\cdot) \), then one can plug in the parametric estimate \( \hat{\theta} \) of the unknown parameter \( \theta_0 \) and obtain a \( \sqrt{n} \)-consistent density estimate \( f_{\hat{\theta}}(\cdot) \). Aït-Sahalia (1996) designed a testing procedure for the null hypothesis (2) by comparing \( f_{\hat{\theta}}(\cdot) \) with the kernel density estimate.

The test of Aït-Sahalia (1996) relies critically on the existence of a closed form marginal density under the null hypothesis. Nevertheless, it is not always guaranteed that such a closed form expression exists for the nonlinear autoregressive process (1), especially when the innovations are non-Gaussian. This makes the testing procedure of Aït-Sahalia (1996) not directly applicable, and it seems desirable if we can have an alternative testing procedure to handle the case where the marginal density does not have a closed form solution under the null. Here we shall utilize the structure in (1) and the independence of \( X_{i-1} \) and \( \epsilon_i \), and propose a density estimate which can play the role of the closed form marginal density as in Aït-Sahalia’s (1996) setting. Hence, the test of the null hypothesis (2) can be similarly carried out. To illustrate the idea, consider the special case of (1) with \( \sigma(\cdot) \equiv 1: \)

\[
X_i = m_\theta(X_{i-1}) + \epsilon_i, \tag{4}
\]

where \( \theta \in \mathbb{R}^d \) is a \( d \)-dimensional parameter vector, \( m_\theta(\cdot) \) is a measurable function, and \( \epsilon_k, k \in \mathbb{Z}, \) are iid innovations with density function \( h(\cdot). \) Assume at the outset that the innovation density \( h \) is known and \( Z_i = m_{\theta_0}(X_i) \) are observed. Since \( Z_{i-1} \) and \( \epsilon_i \) are independent, we can use the following convolution estimate

\[
\hat{f}(x) = \int_{\mathbb{R}} h(x - y)\hat{g}(y)dy, \quad \text{where } \hat{g}(y) = \frac{1}{nb_n} \sum_{i=1}^{n} K \left( \frac{y - Z_{i-1}}{b_n} \right). \tag{5}
\]

Note that \( \hat{g}(\cdot) \) is the usual kernel density estimate of \( g(\cdot), \) the density of \( Z_i. \) Various types of convolution estimates have been discussed in the literature; see for example Bickel and Ritov (1988), Frees (1994), Saavedra and Cao (2000), Schick and Weifelmeyer (2004, 2007) and Giné and Mason (2007), where it is shown that convolution estimates can have parametric, namely \( \sqrt{n}, \) rate of convergence. In a more realistic situation, however, both the parameter \( \theta \) and the innovation density \( f \) are unknown. Assume that \( \hat{\theta} \) is a consistent
estimate of $\theta_0$. Let

$$
\hat{h}_\theta(x) = \frac{1}{n b_n} \sum_{i=1}^{n} K \left[ \frac{x - \{ X_i - m_\theta(X_{i-1}) \}}{b_n} \right], \quad \text{(6)}
$$

and

$$
\hat{h}_\theta(x) = \frac{1}{n b_n} \sum_{i=1}^{n} K \left\{ \frac{x - \epsilon_i + m_\theta(X_{i-1}) - m_{\theta_0}(X_{i-1})}{b_n} \right\}. \quad \text{(7)}
$$

Then $\hat{h}_\theta(x)$ provides a natural nonparametric estimate of $h$ based on the estimated residuals $\hat{\epsilon}_i = X_i - m_{\hat{\theta}}(X_{i-1}) = X_i - \hat{Z}_{i-1}, 1 \leq i \leq n$. As in (5), we can estimate $f$ by

$$
\tilde{f}(x) = \int_R \hat{h}_\theta(x-y) \hat{g}_\theta(y) dy, \quad \text{where } \hat{g}_\theta(y) = \frac{1}{n b_n} \sum_{i=1}^{n} K \left\{ \frac{y - m_\theta(X_{i-1})}{b_n} \right\}. \quad \text{(8)}
$$

Under suitable conditions, we expect that $\tilde{f}$ has a $\sqrt{n}$-convergence rate. Note that the construction of $\tilde{f}$ is based on the independence of $m_\theta(X_{i-1})$ and $\epsilon_i$ in (4). Hence we can test the hypothesis $H_0 : \mu(\cdot) = m_\theta(\cdot)$ for some $\theta \in \Theta$ by comparing the kernel density estimate (3) and the convolution estimate (8). The latter plays the role of the closed form marginal density as in Aït-Sahalia’s (1996) setting.

The rest of the paper is organized as follows. Section 2 presents asymptotic properties of the convolution estimates. We prove that they achieve $\sqrt{n}$-consistency. In Section 3 we shall apply the convolution estimate and devise a model specification test. Simulation studies are provided in Section 5. Proofs are given in Section 7, the Appendix.

## 2 Asymptotics of convolution estimates

If not otherwise specified, throughout this section we shall work with the process (4) and consider estimating the marginal density of $(X_i)$. Note that $(X_i)$ can be nonlinear. The case of linear processes is covered in Schick and Wefelmeyer (2007), where a similar convolution-type estimate is proposed. Due to the nonlinearity, the involved derivations are much more complicated and the techniques developed for linear processes are no longer directly useful. Hence we need to resort to other theoretical frameworks and technical tools. It turns out that, by using the dependence concept introduced in Wu (2005), we can show that under
proper regularity conditions the estimate \( \hat{f}(x) \) can have a \( \sqrt{n} \)-convergence rate. A central limit theorem for \( \sqrt{n}[\hat{f}(x) - f(x)] \) can also be obtained. Such \( \sqrt{n} \)-consistent density estimates can be used for specification test for inferring the underlying data-generating mechanisms for time series whose marginal densities may not have a closed form expression; see Section 3. Section 2.1 deals with the case that the innovation density \( h \) in (4) is known, while Section 2.2 is for unknown \( h \).

Now we introduce some notation. For a random variable \( W \), write \( W \in L^p, p > 0 \), if \( \|W\|_p := \mathbb{E}(|W|^p)^{1/p} < \infty \), and write \( \|W\| = \|W\|_2 \). We define the projection operator \( P \) as
\[
P_i \cdot \equiv E[\cdot |F_i] - E[\cdot |F_{i-1}],
\]
where \( F_i = (\epsilon_i, \epsilon_{i-1}, \ldots) \).

### 2.1 Case 1: innovation density is known.

If \( h \), the density of \( \epsilon_i \), is known, based on \( X_0, X_1, \cdots, X_n \), we can estimate \( f \) by
\[
\hat{f}(x) = \int_{\mathbb{R}} h(x - y) \hat{g}_\theta(y) dy,
\]
where \( \hat{\theta} \) is an estimate of \( \theta_0 \) and
\[
\hat{g}_\theta(y) = \frac{1}{nb_n} \sum_{i=1}^{n} K \left\{ \frac{y - m_\theta(X_{i-1})}{b_n} \right\}.
\]
Note that \( \hat{g}_\theta(\cdot) \) estimates the density of \( m_\theta(X_{i-1}) \). The following assumptions are needed for the derivation of asymptotic properties of \( \hat{f} \).

**Assumption 1.** Let the kernel function \( K \) be bounded, symmetric, with bounded support \([-A, A] \), \( K \in \mathcal{C}^1[-A, A], K(\pm A) = 0 \) and \( \sup_u |K'(u)| < \infty \).

**Assumption 2.** \( \rho := \sup_{x \neq x'} |m_{\theta_0}(x) - m_{\theta_0}(x')|/|x - x'| < 1 \), and \( \epsilon_i \in \mathcal{L}^p, p > 0 \).

**Assumption 3.** \( A_0 := \sup_x [h(x) + |h'(x)| + |h''(x)|] < \infty \), and as \( |x| \to \infty, h(x) = O(|x|^{-\beta}) \) for some \( \beta > 0 \).

**Assumption 4.** \( \dot{m}_\theta(x) = \partial m_\theta(x)/\partial \theta \) exists, \( \mathbb{E}[|\dot{m}_\theta(X_0)|] < \infty \), and \( \dot{m}_\theta(X_0) \) is stochastic continuous in the sense that \( \mathbb{E}[\sup_{|\theta - \theta_0| \leq \delta} |\dot{m}_\theta(X_0) - \dot{m}_{\theta_0}(X_0)|] \to 0 \) as \( \delta \to 0 \).
Assumption 5. \( \hat{\theta}_n \) is an estimate of \( \theta_0 \) such that

\[
\hat{\theta}_n - \theta_0 = n^{-1} \sum_{i=1}^{n} R_i + o_P(n^{-1/2}),
\]

where \( R_i = R(\epsilon_i, \epsilon_{i-1}, \ldots) \) satisfies the short-range dependence condition

\[
\sum_{i=0}^{\infty} \| P_0 R_i \| < \infty.
\]

Comments on these regularity assumptions are as follows. Assumption 1 allows many popular kernels including Parzen and Epanechnikov kernels among others. For many non-linear time series models, the innovation densities satisfy Assumption 3. For example, it holds for the normal and \( t \)-distributions. Assumption 4 is not the weakest possible. However, for the sake of presentational clarity, we decide to use the current form. Assumptions 2 and 5 are discussed in Remarks 1 and 2 below.

**Remark 1.** Assumption 2 is a contraction condition and it ensures that the process \((X_i)\) defined by (4) has a stationary and ergodic solution of the form \( X_i = G(F_i) \), where \( F_i = (\epsilon_i, \epsilon_{i-1}, \ldots) \). So we can assume that \((X_i)\) is a stationary and ergodic process, and it is also causal. Moreover, as argued in Wu and Shao (2004), \((X_i)\) satisfies the following geometric moment contraction property: let \( \epsilon_0 \) and \( \epsilon_i, i \in \mathbb{Z} \), be iid, then

\[
\| X_i - X_i^* \|_p = O(\rho^i), \quad \text{where} \quad X_i^* = G(F_i^*), \quad F_i^* = (\epsilon_i, \ldots, \epsilon_1, \epsilon_0', \epsilon_{-1}, \ldots).
\]

Here \( X_i^* \) is a coupled process of \( X_i \) with \( \epsilon_0 \) replaced by an iid copy \( \epsilon_0' \).

**Remark 2.** To obtain a central limit theorem for an estimate \( \hat{\theta}_n \), (11) is usually an important intermediate step. In certain situations (11) is called Bahadur representation (Bahadur, 1966). In the context of ARCH and GARCH processes, Hall and Yao (2003) obtained a similar expansion for quasi-maximum likelihood estimates; see equation (5.9) in the latter paper. See also Hall and Heyde (1980), Amemiya (1985), Klimko and Nelson (1978) and Section 5.5 in Tong (1990) for similar expansions for method of moment estimators, least squares estimators and maximum likelihood estimators of nonlinear time series models.
Theorem 1. Assume $\sqrt{n}b_n^2 \to 0$ and $nb_n \to \infty$. (i) Under Assumptions 1–5, we have

$$\sqrt{n}[\hat{f}(x) - f(x)] \Rightarrow N[0, \|D_0(x)\|^2],$$

where, letting $c_0(x) = \mathbb{E}[h'(x - m_{\theta_0}(X_0))m_{\theta_0}(X_0)]$,

$$D_0(x) = \sum_{i=0}^{\infty} P_0[h(x - m_{\theta_0}(X_i)) - c_0(x)R_i].$$

(ii) If additionally $\sup_x |h'''(x)| < \infty$, then there exists a mean zero Gaussian process \{W(x)\} with covariance function $\text{cov}[W(x), W(x')] = \mathbb{E}[D_0(x)D_0(x')]$ such that, for any compact interval $I$,

$$\{\sqrt{n}[\hat{f}(x) - f(x)], x \in I\} \Rightarrow \{W(x), x \in I\}.$$

Remark 3. For the classical kernel density estimate $\hat{f}(\cdot)$, it is well-known that, at different points $x_1, \ldots, x_k$, $\sqrt{nb_n}[\hat{f}(x_i) - \mathbb{E}\hat{f}(x_i)], 1 \leq i \leq n$, are asymptotically independent. Namely, the limiting distribution of $\{\sqrt{nb_n}[\hat{f}(x) - \mathbb{E}\hat{f}(x)], x \in \mathbb{R}\}$ is a white noise process. Besides the faster $\sqrt{n}$-convergence, the convolution type estimate $\tilde{f}(x)$ has another appealing property that the limiting process $W(x)$ is differentiable since $D_0'(x)$ exists. Thanks to the smoothness, the convolution type estimate is visually attractive.

Remark 4. If $\theta_0$ is also known, then we can use the estimate $\hat{f}_0(x) = \int_{\mathbb{R}} h(x - y)\hat{g}_{\theta_0}(y)dy$, which is obtained by replacing $\hat{\theta}$ in (9) by the true value $\theta_0$. Following the argument in the proof of Theorem 1, we have the central limit theorem

$$\sqrt{n}[\hat{f}_0(x) - f(x)] \Rightarrow N[0, \|D_0^\circ(x)\|^2],$$

where $D_0^\circ(x) = \sum_{i=0}^{\infty} P_0 h\{x - m_{\theta_0}(X_i)\}$. Interestingly, it could happen that $\|D_0^\circ(x)\| > \|D_0(x)\|$, indicating that the asymptotic variance of $\hat{f}$ can be reduced if we use the estimated parameters instead of the true ones.
2.2 Case 2: innovation density is unknown.

Theorem 1 requires that $h$ is known. It is considerably more challenging to establish a $\sqrt{n}$-norming central limit theorem for $\hat{f}(x)$ defined in (8) for $h$ with an unknown form. In this case $h$ itself needs to be estimated nonparametrically. Extra conditions are needed so that we can have the $\sqrt{n}$-consistency for $\hat{f}(x)$.

**Assumption 6.** (i) $Z_i = m_{\theta_0}(X_i) \in L^p$, $p > 1$; (ii) $Z_i$ has density $g \in C^2(\mathbb{R})$ satisfying $\sup_z |g''(z)| < \infty$ and $\int_{\mathbb{R}} |g''(z)| dz < \infty$; (iii) Let $g(z|\mathcal{F}_i)$ be the conditional density of $Z_{i+1}$ at $z$ given $\mathcal{F}_i$. Assume that $\sum_{i=0}^{\infty} \int_{\mathbb{R}} \|P_0 g(z|\mathcal{F}_i)\| \, dz < \infty$.

In Assumption 6 (iii), following Wu (2005), the quantity $\int_{\mathbb{R}} \|P_0 g(z|\mathcal{F}_i)\| \, dz$ measures the contribution of $\epsilon_0$ in predicting $Z_{i+1}$. Hence Assumption 6 (iii) suggests that the overall contribution of $\epsilon_0$ in predicting future values $(Z_{i+1})_{i \geq 0}$ is finite.

**Theorem 2.** Let the bandwidth sequence $b_n$ satisfy $(\log n)^2 = o(nb_n^3)$ and $b_n = o(n^{-1/4})$.

(i) Under Assumptions 1–6, we have

$$\sqrt{n}[\hat{f}(x) - f(x)] \Rightarrow N[0, \|D_0^*(x)\|^2],$$

where

$$D_0^*(x) = \sum_{i=0}^{\infty} P_0 \{h(x - m_{\theta_0}(X_i)) + g(x - \epsilon_i) + \gamma_0(x)R_i\}$$

and $\gamma_0(x) = -\text{cov}[h'(x - m_{\theta_0}(X_i)), m_{\theta_0}(X_i)]$. (ii) If additionally $\sup_x (|g'''(x)| + |h'''(x)|) < \infty$, then for any compact interval $\mathcal{I}$ we have the functional convergence

$$\{\sqrt{n}[\hat{f}(x) - f(x)], x \in \mathcal{I}\} \Rightarrow \{W^*(x), x \in \mathcal{I}\},$$

where $W^*$ is a mean zero Gaussian process with covariance function $E[W^*(x)W^*(x')] = E[D_0^*(x)D_0^*(x')]$. 


3 Model Specification Test

In this section we shall apply the convolution-type estimate and devise a test for the null hypothesis (2). Section 3.1 deals with the relatively easy case in which $\sigma(\cdot)$ is a constant, while the case for non-constant variance functions is treated in Section 3.2. Aït-Sahalia (1996) used the $L^2$ distance of $\hat{f}(x)$ in (3) and $f_\theta(\cdot)$ when a closed form marginal density is obtainable. Here we shall use the normalized $L^\infty$ or maximal distance (cf. (17) and (23)). Theorems 3 and 4 suggest that our test statistics are asymptotically pivotal in the sense that the limiting distribution does not depend on the unknown densities. This interesting property allows us to design a simulation-based procedure which can have a very good finite-sample performance; see Section 5.2 for a more detailed description. On the other hand, Pritsker (1998) remarked that Aït-Sahalia’s (1996) $L^2$ based test has a very slow convergence rate and thousands of years of continuously sampled data are needed. In comparison, with only about 500 data points, our testing procedure shows very accurate size; see Section 5.2.

For our testing procedures, a distinguished feature and important advantage is that we allow that the innovation density can be unknown. In the study of model specification test, it is usually assumed that the density function under the null can be computed with an arbitrary accuracy; see for example Aït-Sahalia (1999), Chen et al. (2008), Hong and Li (2005) and Bhardwaj et al. (2008) for tests based on transition densities. Bai (2003) and Corradi and Swanson (2006) considered specification tests for cases in which the innovation density is known, and Zhao (2011) considered the problem of nonparametric model validation for hidden Markov models. Nevertheless, none of the techniques in the aforementioned papers can be directly applicable here since we allow unknown innovation densities.
3.1 Constant volatility functions

If $\sigma(\cdot)$ is a constant, without loss of generality we assume that it is 1 and hence we have the time series model

$$X_i = \mu(X_{i-1}) + \epsilon_i,$$

where $\mu$ is an unknown function and $\epsilon_k, k \in \mathbb{Z}$, are iid innovations. We want to test whether the process (15) is of form (4). In other words, we test the hypothesis

$$H_0 : \mu(\cdot) = m_\theta(\cdot) \text{ holds for some } \theta \in \Theta,$$

where $\Theta$ is a parametric space. Recall (3) for $\hat{f}(x)$ and (8) for $\tilde{f}(x)$. For an interval $\mathcal{I} = [l, u]$, where $l < u$ are fixed, let

$$\Xi_n = \sup_{l \leq x \leq u} \frac{\sqrt{n b_n} |\hat{f}(x) - \tilde{f}(x)|}{\sqrt{\int \tilde{f}(x) \int K^2(u)du}}.$$  

Theorem 3. Suppose $H_0$ is true. Then under conditions of Theorem 2, we have the Gumbel convergence that for every $z \in \mathbb{R}$,

$$\mathbb{P}((2 \log \bar{b}_n^{-1})^{1/2} \Xi_n - 2 \log \bar{b}_n^{-1} - C_K \leq z) \to e^{-2e^{-z}},$$

where $C_K = \frac{1}{2} \log \{ \int_{-A} A [K'(u)]^2 du / \int_{-A} A K^2(u)du \} - \log(2\pi)$ and $\bar{b}_n = b_n/(u-l)$.

The convergence in (18) can be quite slow. However, since $\Xi_n$ is asymptotically pivotal, its distribution can be approximated by extensive simulations of $\Xi_n$; see Section 5.2.

Remark 5. A careful check of the proofs of Theorems 3 and 4 below suggests that Assumption 5 on $\hat{\theta}_n$ can actually be replaced by the weaker one: $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$. We emphasize that in many situations one obtains asymptotic normality for $\hat{\theta}_n$ via the representation (11) and hence the condition $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$ is necessarily satisfied. $\diamond$
3.2 Non-constant volatility functions

If $\sigma(\cdot)$ in model (1) is a non-constant function, then the two components $\mu(X_{i-1})$ and $\sigma(X_{i-1})\epsilon_i$ are no longer independent. Note that the conditional density of $X_i = \mu(X_{i-1}) + \sigma(X_{i-1})\epsilon_i$ given $F_{i-1}$ is $h[\cdot - \mu(X_{i-1})]/\sigma(X_{i-1})/\sigma(X_{i-1})$. So we can estimate $f$ by

$$f_{|\theta}^*(y) = \int_{\mathbb{R}} \hat{h}[\{y - \mu_{\hat{\theta}}(x)\}/\sigma_{\hat{\theta}}(x)]\hat{f}(x)/\sigma_{\hat{\theta}}(x)dx,$$

where $\hat{h}$ is an estimate of $h$ and $\hat{f}$ is given in (3). However, it turns out that it is quite tedious and intractable to derive an asymptotic theory for $f_{|\theta}^*$. Hence we consider the modified version of estimating the density of $Y_i := X_i/\sigma(X_{i-1})$.

Here $m(x) = \mu(x)/\sigma(x)$. In this subsection we let $f_{|\theta}^*$ be the density of $Y_i = X_i/\sigma(X_{i-1})$. Under the null hypothesis (2), let $\theta_0$ be the true parameter and let $\hat{\theta}$ be a $\sqrt{n}$-consistent estimate. Note that $f_{|\theta_0}^*$, the density of $X_i/\sigma_{\theta_0}(X_{i-1})$, can be estimated by $\hat{f}_{|\theta}^*(\cdot)$, where

$$\hat{f}_{|\theta}^*(x) = \frac{1}{nb_n} \sum_{i=1}^n \int_{\mathbb{R}} \hat{h}(x - y)\hat{g}(y)dy.$$

As in Section 3.1, we can test hypothesis (2) by comparing $\hat{f}_{|\theta}^*$ and $\tilde{f}_{|\theta}^*$. For the process (1) to have a stationary ergodic solution, we need to replace Assumption 2 by the following (Wu and Shao, 2004) Assumption 7, which implies (13):
Assumption 7. Both \( \mu_{\theta_0}(\cdot), \sigma_{\theta_0}(\cdot) \in \mathcal{C}(\mathbb{R}), \epsilon_i \in \mathcal{L}^p, p > 0, \) and \( \rho := \sup_{x \neq x'} \| \mu_{\theta_0}(x) + \epsilon_0 \sigma_{\theta_0}(x) \| / |x - x'| < 1. \)

Assumption 8. All of \( \mu_{\theta}(\cdot), \sigma_{\theta}(\cdot) \) and \( m_{\theta}(\cdot) = \mu_{\theta}(\cdot)/\sigma_{\theta}(\cdot) \) satisfy Assumption 4, and there exists \( \delta_0 > 0 \) such that \( \inf_x \inf_{|\theta - \theta_0| \leq \delta} \sigma_{\theta}(x) > 0, \) and \( \mathbb{E} \{ [\tilde{\sigma}_{\theta}(\delta, X_t)/\sigma_{\theta_0}(X_t)]^2 \} < \infty, \) where the local maximal function \( \tilde{\sigma}_{\theta}(\delta, x) = \sup_{|\theta - \theta_0| \leq \delta} |\tilde{\sigma}_{\theta}(x)|. \)

Theorem 4. Assume \((\log n)^2 = o(nb_n^3)\) and \(b_n = o(n^{-1/4})\). Under Assumptions 1, 3, 5–8, we have (18) with \( \Xi_n \) therein replaced by
\[
\Xi_n^* = \sup_{i \leq x \leq u} \frac{\sqrt{nb_n} |\hat{f}_{\theta_i}^1(x) - \tilde{f}_{\theta_i}^1(x)|}{\hat{f}_{\theta_0}^1(x) \int K^2(u) du}.
\]

Example 1. Consider the AR(1)-ARCH(1) model
\[
X_t = \theta_0 X_{t-1} + \sqrt{\theta_1^2 + \theta_2^2 X_{t-1}^2} \epsilon_t,
\]
where \( \theta = (\theta_0, \theta_1, \theta_2) \) is the unknown parameter vector and \( \epsilon_t \) are iid with density \( h \) satisfying Assumption 3. Such models were discussed in Borkovec (2000) and Weiss (1986) among others. Assume \( \theta_1 > 0 \) and \( \sup_{-1 \leq u \leq 1} \| \theta_0 + \theta_2 u \epsilon_t \|_p < 1 \) for some \( p > 0. \) Then Assumption 7 holds. Elementary calculations show that Assumption 8 holds as well. By Ling (2004, 2007) and Chan and Peng (2005), both the QMLE and the least absolute deviation estimate satisfy the central limit theorem. Hence our Theorem 4 is applicable. A simple variant of (24) is the threshold AR-ARCH model with \( \theta_0 X_{t-1} \) therein replaced by \( \theta_0 |X_{t-1}|; \) see Cline and Pu (2004) for stability properties of such processes. For such processes, the regularity conditions can be similarly verified.

4 Extension to Continuous-Time Diffusion Models

In econometrics, the nonlinear autoregressive model (1) can be viewed as the discrete-time version of the diffusion model
\[
dX_t = \mu(X_t)dt + \sigma(X_t)d\mathcal{B}(t), \quad t \geq 0,
\]
where \( \mathcal{B}(t) \) is the Brownian motion.
where $\mu(\cdot)$ and $\sigma(\cdot)$ are the drift and diffusion functions respectively, and $\mathcal{B}(\cdot)$ is the standard Brownian motion. For the continuous-time model (25), various parametric forms on $\mu(\cdot)$ and $\sigma(\cdot)$ have been proposed in the literature to explain the underlying data-generating mechanism. For example, Black and Scholes (1973) assumed that $\{\mu(x), \sigma(x)\} = (ax, bx)$, while Vasicek (1977) suggested using $\{\mu(x), \sigma(x)\} = (a + bx, c)$; see also Cox, Ingersoll and Ross (1985), Courtadon (1982), Chan, Karolyi, Longstaff and Sanders (1992), Constantinides (1992), Aït-Sahalia (1996), Duffie and Kan (1996) and Zhao (2008) for other parametric specifications. Bandi and Phillips (2003) considered nonparametric estimation of the drift and diffusion functions in (25) for stationary and nonstationary recurrent processes, while a parametric estimation method is given by Bandi and Phillips (2007); see also the survey paper by Bandi and Phillips (2009). We shall here focus on stationary processes and consider a rigorous specification test.

For the continuous-time diffusion model (25), since $\mathcal{B}(\cdot)$ is a standard Brownian motion, by the forward Kolmogorov equation, the marginal density of (25) has a closed form expression
\begin{equation}
 f(x) = \frac{C}{\sigma^2(x)} \exp \left\{ \int_{x_0}^{x} \frac{2\mu(u)}{\sigma^2(u)} du \right\},
 \end{equation}
where $x_0$ is a point in the state space of $X_t$ and $C = C(\mu, \sigma, x_0)$ is a normalizing constant such that $\int f(x) dx = 1$. Under the null hypothesis (2), assume that based on the sample $X_1, \ldots, X_n$ one can have a $\sqrt{n}$-consistent estimate $\hat{\theta}$ for the unknown parameter $\theta_0$. Then in (26) we plug in $\mu(\cdot)$ and $\sigma(\cdot)$ by $\mu_\theta(\cdot)$ and $\sigma_\theta(\cdot)$, respectively, then the resulting density, denoted by $f_\theta(\cdot)$, is also $\sqrt{n}$-consistent. Therefore, when forming the test statistic (17), we can simply replace the convolution estimate $\tilde{f}(\cdot)$ given in (8) by $f_\theta(\cdot)$. A careful check of the proof reveals that Theorem 3 will continue to hold which can be used to test parametric assumptions for continuous-time diffusion models (25). Note that the current approach is $\mathcal{L}^\infty$-based, while the one by Aït-Sahalia (1996) is $\mathcal{L}^2$-based.
5 A Simulation Study

In Section 5.1 we shall study the numeric performance of the traditional kernel and convolution-type density estimators. A simulation study for model specification test in given in Section 5.2.

5.1 Comparison of Estimators

Recall (3) for the kernel density estimator $\hat{f}$, (5) for the convolution estimator $\tilde{f}$ with known $\theta$, (9) for the feasible convolution estimator $\bar{f}$ with $\theta$ estimated, and (8) for the nonparametric convolution estimator $\tilde{f}$ with both $f$ and $\theta$ estimated. The superior performance of the convolution estimators in relation to the classical kernel estimator in this section can be explained by the asymptotic results that we showed earlier.

We consider the threshold autoregressive (TAR) processes (Tong, 1990):

$$X_i = \theta |X_{i-1}| + \sqrt{1 - \theta^2} \epsilon_i, \quad i = 1, \ldots, n,$$  \hspace{1cm} (27)

where the parameter $\theta \in (-1, 1)$ and $\epsilon_i, i \in \mathbb{Z}$, are iid random variables. If $\epsilon_i$ are standard normal, then interestingly the marginal density of $X_i$ has a closed form with $f(u) = 2\phi(u)\Phi(\delta u)$, where $\delta = \theta/\sqrt{1 - \theta^2}$ and $\phi = \Phi'$ is the standard normal density function; see Andel, Netuka and Svara (1984).

Throughout this section, we use the Epanechnikov kernel $K(u) = 0.75(1 - u^2), \quad u \in [-1, 1]$. Choose $\theta = 0.6$ and $n = 500$. For an estimator $\hat{f}(\cdot)$, our criterion function is the mean integrated squared errors (MISE) defined as $E \int_{\mathbb{R}} |\hat{f}(u) - f(u)|^2 du$. In our simulation we approximate it by averaging 1,000 realizations of the integrated squared errors $\int_{\mathbb{R}} |\hat{f}(u) - f(u)|^2 du$. The parameter $\theta$ in (27) is estimated by the LSE $\hat{\theta}_n = \sum_{i=1}^{n} X_i |X_{i-1}| / \sum_{i=1}^{n} X_{i-1}^2$. Two types of innovations are considered: $\epsilon_i \sim N(0, 1)$ and $\epsilon_i \sim t_5$. In the latter case, since the true marginal density does not have a closed form, we approximate it by the kernel estimator $\hat{f}$ with $n = 10^5$, and expect it to be very close to the true density. The results are summarized in Table 1. We also provide in Figure 1 plots of convolution and kernel
estimators along with true densities. Observe that convolution estimates with known error
densities (\(\hat{f}\) and \(\tilde{f}\)) are uniformly better than traditional kernel estimates \(\hat{f}\). If the marginal
density of the innovation is unknown, convolution estimators also outperform traditional
kernel estimators if the bandwidth \(b \leq 0.25\) is used. Given the conditions in Theorem 2, a
smaller bandwidth is preferred for the convolution density estimator due to the necessity
of reducing the bias term. In practice, we recommend using \(b_n = cn^{-0.3}\), where \(c > 0\) is a
constant, which meets the conditions \((\log n)^2 = o(nb_n^3)\) and \(b_n = o(n^{-1/4})\) in Theorem 2.
Note that the optimal bandwidth for the traditional kernel density estimate is in the form
of \(b_n = cn^{-0.2}\). As a rule of thumb, we suggest using the automatic bandwidth selector for
the traditional kernel density estimate and denote the selected bandwidth as \(\hat{b}_n\), then we
use \(\tilde{b}_n = (n^{-0.3}/n^{-0.2})\hat{b}_n\) for the convolution density estimate.

### 5.2 Model Specification Tests

We now consider the hypothesis testing problem discussed in Section 3. Consider model
(27) with \(n = 500, \theta = 0.6\) and \((l, u) = (-0.5, 1.5)\), and \(\epsilon_i\) can be standard normal or
\(t_5\) innovations. We want to test the hypothesis \(H_0 : \mu(x) = \theta|x|\) for some \(\theta \in \mathbb{R}\). Note
that the Gumbel convergence in Theorem 3 can be quite slow. We shall here adopt a

<table>
<thead>
<tr>
<th>(b)</th>
<th>(\hat{f}_X)</th>
<th>(\hat{f}_X)</th>
<th>(\hat{f}_X)</th>
<th>(\hat{f}_X)</th>
<th>(\hat{f}_X)</th>
<th>(\hat{f}_X)</th>
<th>(\hat{f}_X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>9.70</td>
<td>0.12</td>
<td>0.48</td>
<td>1.58</td>
<td>9.92</td>
<td>0.23</td>
<td>0.53</td>
</tr>
<tr>
<td>0.10</td>
<td>5.00</td>
<td>0.12</td>
<td>0.46</td>
<td>1.36</td>
<td>5.16</td>
<td>0.21</td>
<td>0.49</td>
</tr>
<tr>
<td>0.15</td>
<td>3.34</td>
<td>0.16</td>
<td>0.51</td>
<td>1.34</td>
<td>3.46</td>
<td>0.19</td>
<td>0.49</td>
</tr>
<tr>
<td>0.20</td>
<td>2.58</td>
<td>0.26</td>
<td>0.60</td>
<td>1.59</td>
<td>2.51</td>
<td>0.14</td>
<td>0.43</td>
</tr>
<tr>
<td>0.25</td>
<td>2.26</td>
<td>0.47</td>
<td>0.80</td>
<td>2.21</td>
<td>2.02</td>
<td>0.17</td>
<td>0.45</td>
</tr>
<tr>
<td>0.30</td>
<td>2.29</td>
<td>0.85</td>
<td>1.17</td>
<td>3.37</td>
<td>1.83</td>
<td>0.32</td>
<td>0.58</td>
</tr>
<tr>
<td>0.35</td>
<td>2.62</td>
<td>1.43</td>
<td>1.73</td>
<td>5.12</td>
<td>1.80</td>
<td>0.59</td>
<td>0.83</td>
</tr>
<tr>
<td>0.40</td>
<td>3.22</td>
<td>2.28</td>
<td>2.57</td>
<td>7.52</td>
<td>1.93</td>
<td>0.90</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Table 1: Simulated MISE (\(\times 1000\)) for various estimators with different bandwidths \(b\).
Figure 1: Density estimates for the TAR process (27) with both normal and $t_5$ innovations. In both graphs, solid (resp. dashed, or dotted dashed) curves stand for true densities (resp. convolution estimates, or the classical kernel density estimators). The true density in the right hand panel is approximated by the kernel estimator with $n = 10^5$.

A simulation-based procedure that can significantly improve the finite sample performance. Let $\epsilon^*_i, i \in \mathbb{Z}$, be iid random variables with density $h$, we form $X^*_i = m_\theta(X^*_{i-1}) + \epsilon^*_i$ and, as in (17), compute the corresponding test statistic $\Xi^*_n$. By Theorem 3, $\Xi_n$ and $\Xi^*_n$ have the same asymptotic distribution. Thus, the cutoff value, $(1 - \alpha)$-th quantile of $\Xi_n$, can be approximated by the sample $(1 - \alpha)$-th quantile of many simulated $\Xi^*_n$. Note that $f(x)$ can be estimated by $\hat{f}_n(x, \hat{\theta})$. Let $\alpha = 10\%$ or $5\%$. For each run, we simulate 5,000 realizations and calculate their corresponding test statistics. For each realization, 5,000 $\Xi_n$’s are generated and we compute their $(1 - \alpha)$-th quantile, denoted by $\hat{q}_{1-\alpha}$. We reject the null hypothesis if $\Xi_n > \hat{q}_{1-\alpha}$. The results are summarized in Table 2. We can see that the empirical acceptance percentages are quite close to their nominal levels 90\% or 95\%. Pritsker (1998) commented that Aït-Sahalia’s (1996) test procedure does not perform well unless one has an astronomically large sample. Table 2 suggests that our simulation-based procedure has a very satisfactory finite sample performance.
As another numeric example, we consider the TAR-ARCH (Threshold AR with ARCH errors) model
\[ X_i = \theta_0 |X_{i-1}| + \sigma_\theta(X_{i-1})\epsilon_i, \quad i = 1, \ldots, n, \quad (28) \]
where \( \sigma_\theta(x) = (\theta_1 + \theta_2 x^2)^{1/2} \). Note that model (27) has a constant volatility function.

Choose \( n = 500, \ \theta = (\theta_0, \theta_1, \theta_2) = (1/2, 1/4, 1/4) \) and \( (l, u) = (-1, 1.5) \). Also two types of innovations are considered: \( \epsilon_i \sim N(0, 1) \) and \( \epsilon_i \sim (3/5)^{1/2}t_5 \). For both cases \( \text{var}(\epsilon_i) = 1 \).

The parameter vector \( \theta \) can be estimated by the least squares method. Following the procedure above, we report the empirical levels in Table 2. We can see that they are also quite close to their nominal levels for a wide range of bandwidths.

### Table 2: Empirical acceptance percentages with different bandwidths \( b \) and error distributions for model (27) and (28).

<table>
<thead>
<tr>
<th>( b )</th>
<th>( 90% )</th>
<th>( 95% )</th>
<th>( 90% )</th>
<th>( 95% )</th>
<th>( 90% )</th>
<th>( 95% )</th>
<th>( 90% )</th>
<th>( 95% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>89.2</td>
<td>94.8</td>
<td>90.0</td>
<td>95.2</td>
<td>90.8</td>
<td>95.3</td>
<td>90.4</td>
<td>95.1</td>
</tr>
<tr>
<td>0.10</td>
<td>89.5</td>
<td>94.6</td>
<td>90.6</td>
<td>95.5</td>
<td>90.9</td>
<td>95.7</td>
<td>90.7</td>
<td>95.6</td>
</tr>
<tr>
<td>0.15</td>
<td>90.5</td>
<td>94.9</td>
<td>90.3</td>
<td>95.2</td>
<td>90.3</td>
<td>94.9</td>
<td>91.0</td>
<td>95.9</td>
</tr>
<tr>
<td>0.20</td>
<td>89.5</td>
<td>94.7</td>
<td>87.8</td>
<td>93.6</td>
<td>89.8</td>
<td>94.7</td>
<td>89.4</td>
<td>94.7</td>
</tr>
<tr>
<td>0.25</td>
<td>88.5</td>
<td>94.0</td>
<td>91.3</td>
<td>95.7</td>
<td>90.3</td>
<td>95.5</td>
<td>89.9</td>
<td>94.2</td>
</tr>
<tr>
<td>0.30</td>
<td>88.6</td>
<td>93.9</td>
<td>90.2</td>
<td>95.0</td>
<td>90.0</td>
<td>94.8</td>
<td>90.0</td>
<td>94.6</td>
</tr>
<tr>
<td>0.35</td>
<td>86.5</td>
<td>93.0</td>
<td>90.4</td>
<td>95.2</td>
<td>89.1</td>
<td>94.2</td>
<td>89.2</td>
<td>94.1</td>
</tr>
<tr>
<td>0.40</td>
<td>90.2</td>
<td>95.3</td>
<td>88.1</td>
<td>93.9</td>
<td>90.2</td>
<td>95.2</td>
<td>90.5</td>
<td>95.2</td>
</tr>
</tbody>
</table>

5.3 Application to Continuous-Time Diffusion Models

As discussed in Section 4, the proposed test can be extended to continuous-time diffusion models, in which case the convolution density estimate is replaced by the parametric density estimate. Aït-Sahalia (1996) considered an \( L^2 \)-based specification test for diffusion processes, while we shall here apply the proposed \( L^\infty \)-based test. For the diffusion process...
(25), we consider testing the null hypothesis \( H_0 : \mu(x) = 0.5(\gamma - x) \) and \( \sigma(x) = \sigma \) for some \( \gamma \in \mathbb{R} \) and \( \sigma > 0 \). Given a discrete sample \( X_i, i = 1, \ldots, n \), the marginal distribution under the null hypothesis is \( N(\gamma, \sigma^2) \) and thus the parameters can be naturally estimated by \( \hat{\gamma}_n = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \) and \( \hat{\sigma}^2_n = (n - 1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \). To improve the finite sample performance, we propose a simulation-based testing procedure similar to the one discussed in Section 5.2 for nonlinear autoregressive models. In particular, the cut-off values are obtained by simulating diffusion processes using the specification under the null with parameter estimates \((\hat{\gamma}, \hat{\sigma}^2)\) plugged in. The results are summarized in Table 3 for different choices of bandwidth and sampling frequency. It can be seen from Table 3 that the proposed simulation-based testing procedure performs reasonably well as the empirical acceptance probabilities are quite close to their nominal levels (90% and 95%). In contrast, the test by Aït-Sahalia (1996) does not have a good finite sample performance due to its slow convergence rate; see also the discussion in Section 3.

6 Acknowledgment

We thank the Editor, the Co-Editor and two anonymous referees for their helpful comments and suggestions. Kim gratefully acknowledges funding from Hanyang University (HY-2013-G). Wu is grateful for the support of NSF grants DMS-0906073 and DMS-1106790.

References


<table>
<thead>
<tr>
<th>b</th>
<th>daily sampling (i)</th>
<th>daily sampling (ii)</th>
<th>weekly sampling (i)</th>
<th>weekly sampling (ii)</th>
<th>monthly sampling (i)</th>
<th>monthly sampling (ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>95%</td>
<td>90%</td>
<td>95%</td>
<td>90%</td>
<td>95%</td>
</tr>
<tr>
<td>0.05</td>
<td>90.6</td>
<td>94.7</td>
<td>100.0</td>
<td>100.0</td>
<td>89.8</td>
<td>95.4</td>
</tr>
<tr>
<td>0.10</td>
<td>90.8</td>
<td>95.5</td>
<td>99.9</td>
<td>100.0</td>
<td>90.6</td>
<td>95.3</td>
</tr>
<tr>
<td>0.15</td>
<td>90.3</td>
<td>95.2</td>
<td>99.9</td>
<td>100.0</td>
<td>89.8</td>
<td>95.5</td>
</tr>
<tr>
<td>0.20</td>
<td>90.4</td>
<td>95.2</td>
<td>100.0</td>
<td>100.0</td>
<td>90.5</td>
<td>95.0</td>
</tr>
<tr>
<td>0.25</td>
<td>90.5</td>
<td>95.5</td>
<td>99.9</td>
<td>100.0</td>
<td>90.2</td>
<td>95.1</td>
</tr>
<tr>
<td>0.30</td>
<td>89.7</td>
<td>95.0</td>
<td>99.9</td>
<td>100.0</td>
<td>89.0</td>
<td>94.2</td>
</tr>
<tr>
<td>0.35</td>
<td>90.5</td>
<td>95.6</td>
<td>100.0</td>
<td>100.0</td>
<td>89.6</td>
<td>94.4</td>
</tr>
<tr>
<td>0.40</td>
<td>87.8</td>
<td>94.7</td>
<td>99.7</td>
<td>99.9</td>
<td>90.1</td>
<td>95.5</td>
</tr>
</tbody>
</table>

Table 3: Empirical acceptance percentages for (i) the proposed simulation-based testing procedure; and (ii) the test by Aït-Sahalia (1996) with different choices of bandwidth $b$ and sampling frequency for continuous-time diffusion models.


7 Appendix

In this section we shall provide proofs for results in Sections 2 and 3. A careful check of the proofs implies that our asymptotic results also hold for the nonlinear AR(p) process

\[ X_i = R(X_{i-1}, \ldots, X_{i-p}; \epsilon_i) := \mu(X_{i-1}, \ldots, X_{i-p}) + \sigma(X_{i-1}, \ldots, X_{i-p})\epsilon_i, \]

with the straightforward modification that the function \( m_{\theta}(X_{i-1}) \) in Assumptions 4 and 8 should now be replaced by \( m_{\theta}(X_{i-1}, \ldots, X_{i-p}) = \mu(X_{i-1}, \ldots, X_{i-p})/\sigma(X_{i-1}, \ldots, X_{i-p}) \), and Assumption 7 is replaced by the following: for all \( y_1, \ldots, y_p, y'_1, \ldots, y'_p \in \mathbb{R} \),

\[ \| R(y_1, \ldots, y_p; \epsilon_i) - R(y'_1, \ldots, y'_p; \epsilon_i) \|_\alpha \leq \sum_{i=1}^p a_i |y_i - y'_i|^{\alpha'}, \]

where \( a_i \) are nonnegative constants with \( \sum_{i=1}^p a_i < 1 \), \( \epsilon_i \in \mathcal{L}^\alpha \), \( \alpha > 0 \) and \( \alpha' = \min(1, \alpha) \).

Shao and Wu (2007) showed that the above condition also implies the geometric moment contraction property (13). In our proofs for ease of reading we consider AR(1) models.

Proof. (Theorem 1) Let \( \hat{S}_n(x) = \sum_{i=1}^n h(x-m_{\hat{\theta}}(X_{i-1})) \) and \( S_n(x) = \sum_{i=1}^n h(x-m_{\theta_0}(X_{i-1})) \).

By Assumption 3, \( \sup_x |\hat{S}_n''(x)| = O(n) \). By Taylor’s expansion, since \( \int_{\mathbb{R}} uK(u)du = 0 \),

\[
|n\hat{f}(x) - \hat{S}_n(x)| = \left| \int_{\mathbb{R}} K(u)\hat{S}_n(x - ub_n)du - \hat{S}_n(x) \right| \\
\leq \int_{\mathbb{R}} |K(u)||\hat{S}_n(x - ub_n) - \hat{S}_n(x) + ub_n\hat{S}_n'(x)|du \\
= \int_{\mathbb{R}} |K(u)|(ub_n)^2O(n)du = O(nb_n^2). \tag{29}
\]

So \( \sup_x |n\hat{f}(x) - \hat{S}_n(x)| = o(\sqrt{n}) \) since \( \sqrt{n}b_n^2 \to 0 \). Let

\[ T_n(x, \theta) = \sum_{i=1}^n h'(x - m_{\theta}(X_{i-1}))\hat{m}_{\theta}(X_{i-1}). \]

By Assumptions 3 and 4 and the Lebesgue Dominated Convergence Theorem,

\[
\lim_{\delta \to 0} \mathbb{E} \left[ \sup_{|\theta - \theta_0| \leq \delta} |h'(x - m_{\theta}(X_{i-1})) - h'(x - m_{\theta_0}(X_{i-1}))||\hat{m}_{\theta_0}(X_{i-1})| \right] = 0.
\]

24
Therefore, again by Assumptions 3 and 4,
\[
\lim_{\delta \to 0} \frac{1}{n} \mathbb{E} \left[ \sup_{\|x\| \leq \delta} |T_n(x, \theta) - T_n(x, \theta_0)| \right] = 0. \tag{30}
\]
By Assumption 5, \( \hat{\theta}_n - \theta_0 = O_P(n^{-1/2}) \). By Taylor’s expansion and (30),
\[
\sup_x |\hat{S}_n(x) - S_n(x) - (\hat{\theta}_n - \theta_0)T_n(x, \theta_0)| = |\hat{\theta}_n - \theta_0|o_P(n) = o_P(n^{1/2}). \tag{31}
\]
By the ergodic Theorem, \( \lim_{n \to \infty} T_n(x, \theta_0)/n \to c_0(x) \) almost surely. The latter convergence can be made uniform over \( x \in \mathbb{R} \) since \( c_0(\cdot) \) is uniformly continuous. By (29)-(31),
\[
\sup_x \left| n\tilde{f}(x) - S_n(x) + nc_0(x) \sum_{i=1}^n R_i \right| = o_P(\sqrt{n}). \tag{32}
\]
Let \( p' = \min(1, p) \). By Assumptions 2 and 3, and Remark 1,
\[
\|P_0 (h - m_{\theta_0}(X_i))\| = \|E[h(x - m_{\theta_0}(X_i)) - h(x - m_{\theta_0}(X_i^*))|\mathcal{F}_0]\| \\
\leq \|E[h(x - m_{\theta_0}(X_i)) - h(x - m_{\theta_0}(X_i^*))]\| \\
\leq \|A_0 \min[1, |m_{\theta_0}(X_i) - m_{\theta_0}(X_i^*)|]\| \\
\leq A_0 \|E(|X_i - X_i^*|^{p'})\|^{1/2} \\
\leq A_0 \|X_i - X_i^*\|^{p'/2} = O(\rho^{p'/2}). \tag{33}
\]
Since \( \sum_{i=0}^\infty \|P_0 R_i\| < \infty \), we have
\[
\sum_{i=0}^\infty \|P_0[h(x - m_{\theta_0}(X_i)) - c_0(x)R_i]\| < \infty, \tag{34}
\]
by Theorem 1 in Hannan (1973), (i) follows.

To prove (ii), since the finite dimensional convergence easily follows from (34), it remains to verify the tightness of \( W_n(x) := \sqrt{n}|\hat{f}(x) - f(x)| \); see Billingsley (1968). Since \( \mathcal{I} \) is compact, by (32), it suffices to show that \( E[\sup_{x \in \mathcal{I}} |S_n'(x) - n f'(x)|] = O(\sqrt{n}) \). Let \( U_n(x) = |S_n'(x) - n f'(x)|/\sqrt{n} \). By Remark 1 and Assumption 3, as (33), we have
\[
\sup_x ||P_0 h'(x - m_{\theta_0}(X_i))|| = O(\rho^{p'/2}).
\]
So \( \sup_x ||U_n'(x)|| \leq \sum_{i=0}^\infty \sup_x ||P_0 h'(x-m_{\theta_0}(X_i))|| = O(1) \), and \( E[\sup_{x \in \mathcal{I}} |U_n(x) - U_n(0)|] \leq \int_\mathcal{I} E[U_n'(x)]dx = O(1) \). Thus the tightness follows. \( \Box \)
Lemma 1. Recall (8) for $\hat{g}_n(y)$. Assume $nb_n^2 \to \infty$ and $b_n \to 0$. Under Assumptions 1, 4, 5 and 6, we have

$$\int \mathbb{E}|\hat{g}_n(y) - g(y)|dy = O_{\mathbb{P}}((\sqrt{nb_n})^{-1} + b_n^2).$$  \hspace{1cm} (35)

Proof. For $\delta > 0$ let $\hat{m}_*(\delta, x) = \sup_{\theta - \theta_0 \leq \delta} |\hat{m}_\theta(x)|$. By Assumption 4, there exists $\delta_0 > 0$ such that $\mathbb{E}[\hat{m}_*(\delta_0, X_0)] < \infty$. Choose $k_0 > 0$ such that $\int_{\mathbb{R}} |K(u) - K(u + \delta)|du \leq |\delta|k_0$.

Then

$$\int \mathbb{R} |\hat{g}_\theta(y) - \hat{g}_\theta_0(y)|dy \leq \frac{1}{nb_n} \sum_{i=1}^n \int_{\mathbb{R}} \left|K\left(y - \frac{m_\theta(X_{i-1})}{b_n}\right) - K\left(y - \frac{m_{\theta_0}(X_{i-1})}{b_n}\right)\right|dy$$

$$\leq \frac{1}{nb_n} \sum_{i=1}^n |m_\theta(X_{i-1}) - m_{\theta_0}(X_{i-1})|k_0 = O_{\mathbb{P}}((\sqrt{nb_n})^{-1}).$$  \hspace{1cm} (36)

In the last step, we have applied the inequality $|m_\theta(x) - m_{\theta_0}(x)| \leq |\theta - \theta_0|\hat{m}_*(|\theta - \theta_0|, x)$, the Ergodic Theorem and the fact that $\hat{\theta} - \theta_0 = O_{\mathbb{P}}(n^{-1/2})$.

Without loss of generality let $1 < p \leq 2$. Let $\lambda(y) = (1 + |y|)^p$. Since $Z_i \in \mathcal{L}^p$,

$$\int_{\mathbb{R}} \|K[(y - Z_0)/b_n]\|^2 \lambda(y)dy = b_n \int_{\mathbb{R}} K^2(u)du \int_{\mathbb{R}} g(y - b_nu)\lambda(y)dy$$

$$\leq b_n \int_{\mathbb{R}} K^2(u)(1 + |b_nu|)^pdu \int_{\mathbb{R}} g(y)\lambda(y)dy$$

$$= O(b_n).$$

Let

$$g^0_n(y) = (nb_n)^{-1} \sum_{i=1}^n \mathbb{E}\{K[(y - Z_{i-1})/b_n]|\mathcal{F}_{i-2}\}.$$  \hspace{1cm} (37)

By the orthogonality of martingale differences and Schwarz’s inequality,

$$\mathbb{E} \int_{\mathbb{R}} |\hat{g}_\theta_0(y) - g^0_n(y)|dy \leq \int_{\mathbb{R}} \|\hat{g}_\theta_0(y) - g^0_n(y)\|dy$$

$$\leq (nb_n)^{-1} \int_{\mathbb{R}} \sqrt{\mathbb{E}}\|K[(y - Z_0)/b_n]\|dy$$

$$\leq \frac{1}{\sqrt{nb_n}} \left[\int_{\mathbb{R}} \|K[(y - Z_0)/b_n]\|^2 \lambda(y)dy\right]^{1/2} \left[\int_{\mathbb{R}} \lambda^{-1}(y)dy\right]^{1/2}$$

$$= O((\sqrt{nb_n})^{-1}).$$
Note that $\mathbb{E}\{K[(y - Z_{i-1})/b_n]|\mathcal{F}_{i-2}\} = b_n \int_{\mathbb{R}} K(u)f_Z(y - b_n u|\mathcal{F}_{i-2})du$. Let
$$H_n(u) = \sum_{i=1}^{n} [f_Z(u|\mathcal{F}_{i-2}) - f_Z(u)].$$

By Theorem 2 in Wu (2005), $\|H_n(u)\| \leq \sqrt{n} \sum_{i=0}^{\infty} \|P_0 f_Z(u|\mathcal{F}_i)\|$, and Assumption 6(iii),
$$\mathbb{E} \int_{\mathbb{R}} |\hat{g}_n^0(y) - \mathbb{E}[\hat{g}_n^0(y)]|dy \leq \frac{1}{n} \int_{\mathbb{R}} \int_{\mathbb{R}} |K(u)||H_n(y - b_n u)||dydu \leq \frac{1}{n} \int_{\mathbb{R}} \int_{\mathbb{R}} |K(u)|\sqrt{n} \sum_{i=0}^{\infty} \|P_0 f_Z(y - b_n u|\mathcal{F}_i)\||dydu = O(n^{-1/2}). \quad (38)$$

By the inequality $\int_{\mathbb{R}} |g(y + t) - g(y) - tg'(y)|dy \leq (t^2/2) \int_{\mathbb{R}} |g''(y)|dy$, we have
$$\int_{\mathbb{R}} |\mathbb{E}[\hat{g}_n^0(y)] - g(y)|dy \leq \int_{\mathbb{R}} \int_{\mathbb{R}} |K(u)||g(y - b_n u) - g(y) + b_n u g'(y)|dydu = O(b_n^2) \quad (39)$$
by Assumption 6(ii). By (36)-(39), the lemma follows. \diamond

**Lemma 2.** Assume $(\log n)^2 = o(nb_n)$ and $b_n \to 0$. Under Assumptions 1–5, we have
$$\sup_{x \in \mathbb{R}} |\hat{h}_n(x) - h(x)| = O_P((nb_n)^{-1/2} \log n + b_n^2). \quad (40)$$

**Proof.** Let $\delta_n \to 0$ be a positive sequence such that $\sqrt{n}\delta_n \to \infty$; let
$$\hat{g}_n^0(x) = \frac{1}{nb_n} \sum_{i=1}^{n} \mathbb{E}\{K[(x + m_\theta(X_{i-1}) - m_{\theta_0}(X_{i-1}) - \epsilon_i)/b_n]|\mathcal{F}_{i-1}\}$$
$$= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K(u)h(x + m_\theta(X_{i-1}) - m_{\theta_0}(X_{i-1}) - b_n u)du.$$ 

Let $W_i(x, \theta) = K[(x + m_\theta(X_{i-1}) - m_{\theta_0}(X_{i-1}) - \epsilon_i)/b_n]$, $h_* = \sup_x h(x)$, $K_* = \sup_u |K(u)|$ and $K_2 = \int K^2(u)du$. Then
$$\mathbb{E}[W_i^2(x, \theta)|\mathcal{F}_{i-1}] = b_n \int_{\mathbb{R}} K^2(u)h(x - b_n u + m_\theta(X_{i-1}) - m_{\theta_0}(X_{i-1}))du \leq b_n h_* K_2.$$

27
Note that \( W_i(x, \theta) - E[W_i(x, \theta) | \mathcal{F}_{i-1}] \), \( i = 1, 2, \ldots \), form (triangular) stationary martingale differences bounded by \( 2K_* \). Using Freedman’s (1975) exponential inequality for bounded martingale differences,

\[
P \left[ |\hat{h}_\theta(x) - \hat{h}_\theta^0(x)| \geq \frac{\log n}{(nb_n)^{1/2}} \right] \leq 2 \exp \left[ -\frac{nb_n \log^2 n}{4K_*(nb_n)^{1/2} \log n + 4nb_nh_*K_2} \right]. \tag{41}
\]

As in the proof of Lemma 1, let \( m_\theta(x, \delta) = \sup_{|\theta - \theta_0| \leq \delta} |m_\theta(x)| \) and choose \( \delta_0 > 0 \) such that \( E[m_\theta(x, 0)] < \infty \). Let \( N = n^4 \) and define \( |x|_N = |x|_N/N \) and, for the \( d \)-dimensional vector \( \theta = (\theta_1, \ldots, \theta_d) \), \( |\theta|_N = (|\theta_1|_N/N, \ldots, |\theta_d|_N/N) \). By Assumptions 1 and 4,

\[
|\hat{h}_\theta(x) - \hat{h}_{\theta | N}(|x|_N)| \leq \sum_{i=1}^{n} k_0 b_n \left( N^{-1} + N^{-1}m_\theta(\delta_0, X_{i-1}) \right)
\]

if \( |\theta - \theta_0| \leq \delta_0 \). Here \( k_0 = \sup_n |K'(u)| \). Hence

\[
E \left[ \sup_{x \in \mathbb{R}} \sup_{|\theta - \theta_0| \leq \delta_0} |\hat{h}_\theta(x) - \hat{h}_{\theta | N}(|x|_N)| \right] = o(n^{-2}). \tag{42}
\]

Clearly the above relation (42) also holds if we replace \( \hat{h}_\theta(x) \) therein by \( \hat{h}_\theta^0(x) \). By Assumptions 2 and 4, there exists a constant \( \tau > 0 \) such that \( \sup_{|\theta - \theta_0| \leq \delta} |m_\theta(X_{i-1}) - m_\theta_0(X_{i-1}) - \varepsilon_i| \in \mathcal{L}^\tau \). Recall Assumption 3 for \( \beta \) and let the interval \( \mathcal{I}_n = [-M, M] \), where \( M = n^{8/3+8/\tau+8} \). Note that the cardinality \#\{\((|\theta|_N, |x|_N) : x \in \mathcal{I}_n, |\theta - \theta_0| \leq \delta_0\} = O(2MN^{1+d}) \). In (41), since \( (\log n)^2 = o(nb_n) \), we obtain

\[
P \left[ \sup_{x \in \mathbb{R}} \sup_{|\theta - \theta_0| \leq \delta_0} |\hat{h}_{\theta | N}(|x|_N) - \hat{h}_\theta^0(\theta, x)| \right] \geq \frac{\log n}{(nb_n)^{1/2}} \leq 2K_*(nb_n)^{1/2} \log n + 4nb_nh_*K_2 \leq o(n^{-1}). \tag{43}
\]

By (42) and (43), we obtain

\[
\sup_{x \in \mathbb{R}} \sup_{|\theta - \theta_0| \leq \delta_0} |\hat{h}_\theta(x) - \hat{h}_\theta^0(x)| = O_P((nb_n)^{-1/2} \log n). \tag{44}
\]

By the choice of \( M \) and since \( K \) has bounded support, we have \( \sup_{|x| \geq M} \sup_{|\theta - \theta_0| \leq \delta_0} |\hat{h}_\theta(x)| = 0 \) in probability. By Assumption 3, \( \sup_{|x| \geq M} h(x) = o(n^{-1}) \). Hence

\[
\sup_{|x| \geq M} \sup_{|\theta - \theta_0| \leq \delta_0} |\hat{h}_\theta(x) - h(x)| = o_P(n^{-1}). \tag{45}
\]
On the other hand, let
\[ I_\theta(x) = I_{\theta,n}(x) = \sum_{i=1}^{n} h(x + m_\theta(X_{i-1}) - m_\theta(X_{i-1})). \]  
(46)

As (36), by the ergodic theorem, \( \sup_x \sup_{|\theta - \theta_0| \leq \delta_n} |I_{\theta,n}(x)/n - h(x)| = O_\delta(\delta_n) \). Hence
\[ \sup_{x \in \mathbb{R}} \sup_{|\theta - \theta_0| \leq \delta_n} |\hat{h}_\theta(x) - \int_{\mathbb{R}} K(u)h(x - b_n u) du| = O_\delta(\delta_n). \]  
(47)

Since the speed \( \sqrt{n} \delta_n \to \infty \) can be made arbitrarily slow, by (44) and (47), we have
\[ \sup_{x \in I_n} |\hat{h}_\theta(x) - \int_{\mathbb{R}} K(u)h(x - b_n u) du| = O_p((nb_n)^{-1/2} \log n). \]  
(48)

Since \( \sup_x |h''(x)| < \infty \), \( \sup_x |\int_{\mathbb{R}} K(u)[h(x - b_n u) - h(x) + b_n uh'(x)]du| = O(b_n^2) \). Note that \( \int_{\mathbb{R}} uK(u) du = 0 \), Lemma 2 then follows from (45) and (48).

\[ \Diamond \]

Proof. (Theorem 2) (i) Since \( (\log n)^2 = o(nb_n^3) \) and \( b_n = o(n^{-1/4}) \), simple calculations show that
\[ [(\sqrt{nb_n})^{-1} + b_n^2][(nb_n)^{-1/2} \log n + b_n^2] = o(n^{-1/2}). \]

By Lemmas 1 and 2, we have
\[ \sup_x \int_{\mathbb{R}} |[\hat{h}_\theta(x - y) - h(x - y)][\hat{g}_\theta(y) - g(y)]|dy = o_p(n^{-1/2}). \]  
(49)

Let
\[ \hat{W}_n(x) = \sum_{i=1}^{n} [h(x - m_\theta(X_{i-1})) + g(x - \hat{e}_i)]. \]

Since \( \sup_x |h''(x)| + |g''(x)| < \infty \),
\[ \hat{W}_n(x - b_n u) - 2nf(x - b_n u) = \{\hat{W}_n(x) - 2nf(x)\} - b_n u\{\hat{W}_n'(x) - 2nf'_X(x)\} + b_n^2 u^2 O(n). \]

Therefore, by Assumption 1, we have \( \int_{\mathbb{R}} K(u)f(x - b_n u) = f(x) + O(b_n^2) \) and
\[ \Delta_n(x) := \int_{\mathbb{R}} [\hat{h}_\theta(x - y) - h(x - y)][\hat{g}_\theta(y) - g(y)]dy \]

29
\[
\begin{align*}
&\hat{f}(x) + f(x) - \int_{\mathbb{R}} \left[ \hat{h}_\theta(x-y)g(y) + h(x-y)\hat{g}_\theta(y) \right] dy \\
&= \hat{f}(x) + f(x) - n^{-1} \int_{\mathbb{R}} K(u)\hat{W}_n(x-b_n u) du \\
&= \hat{f}(x) + f(x) - 2 \int_{\mathbb{R}} K(u)f(x-b_n u) du - \frac{\hat{W}_n(x) - 2n f(x)}{n} + O(b_n^2) \\
&= \hat{f}(x) - f(x) - \frac{W_n(x) - 2n f(x)}{n} + O(b_n^2).
\end{align*}
\]

Let \( W_n(x) = \sum_{i=1}^n [h(x-m_{\theta_0}(X_{i-1})) + g(x-\epsilon_i)] \) and \( V_n(x, \theta) = \sum_{i=1}^n \hat{m}_{\theta_i}(X_{i-1})[g'(x-\epsilon_i + m_{\theta_i}(X_{i-1}) - m_{\theta_0}(X_{i-1})) - h'(x-m_{\theta_0}(X_{i-1}))]. \)

Similarly as (31), we have
\[
\sup_x |\hat{W}_n(x) - W_n(x) - (\hat{\theta}_n - \theta_0) V_n(x, \theta_0)| = o_P(n^{1/2}).
\]

Note that \( \mathbb{E}\{\hat{m}_{\theta_0}(X_0)[g'(x-\epsilon_1) - h'(x-m_{\theta_0}(X_0))]\} = \gamma_0(x). \)

By the ergodic Theorem, since \( \sup_x [|h''(x)| + |g''(x)|] < \infty, \) for any compact interval \( \mathcal{I}, \) we have
\[
\lim_{n \to \infty} \sup_{x \in \mathcal{I}} \mathbb{E}\left| \frac{V_n(x, \theta_0)}{n} - \gamma_0(x) \right| = 0.
\]

Since \( \sqrt{n}b_n^2 \to 0, \) we have
\[
\sup_{x \in \mathcal{I}} |\hat{f}_X(x) - f(x) - n^{-1}W_n(x) - (\hat{\theta}_n - \theta_0)\gamma_0(x)| = o_P(n^{-1/2}). \tag{50}
\]

Then similarly as the argument in (34), we have (14).

(ii) It can be similarly proved by using the argument in the proof of Theorem 1(ii). \( \diamond \)

**Proof.** (Theorem 3) We shall apply Theorem 2.1 in Liu and Wu (2010). The bandwidth condition \( (\log n)^2 = o(nb_n^2) \) and \( b_n = o(n^{-1/4}) \) in our Theorem 2 implies Condition (C1) of Theorem 2.1 in Liu and Wu (2010). Condition (C3) in the latter paper is just our Assumption 3, and (C2) follows from the geometric contraction (13). Since \( b_n = o(n^{-1/4}), \)
\[
\sqrt{n}b_n \mathbb{E}[\hat{f}(x) - f(x)] = O(n^{-1/8}).
\]

By Theorem 2.1 in Liu and Wu (2010), (18) holds for \( \Xi_n^o \) with
\[
\Xi_n^o = \sup_{x \in [a,b]} \sqrt{n}b_n [\hat{f}(x) - f(x)] / \sqrt{f(x) \int K^2(u) du}.
\]
Then (18) results from Theorem 2 in view of \( \sup_{t \leq x \leq u} |\tilde{f}(x) - f(x)| = O_p(n^{-1/2}). \)

**Proof.** (Theorem 4) Under Assumption 7, by Wu and Shao (2004), the process \( (X_i) \) satisfies the geometric moment contraction property (13). By Assumption 8, with elementary manipulations, the process \( \mu(X_{i-1})/\sigma(X_{i-1}) \) also satisfies (13). Since \( X_i/\sigma(X_{i-1}) = \epsilon_i + \mu(X_{i-1})/\sigma(X_{i-1}) \), Theorem 2.1 in Liu and Wu (2010) entails (18) with

\[
\Xi_n := \sup_{i \leq x \leq u} \frac{\sqrt{n|b_n|} |\tilde{f}_{\theta_0}^1(x) - \tilde{f}_{\theta_0}^1(x)|}{\sqrt{\int_{\theta_0} f_{\theta_0}^1(x) \int K^2(u) du}}. \tag{52}
\]

Then it remains to show that \( \tilde{f}_{\theta_0}^1(x) \) (resp. \( \tilde{f}_{\theta_0}^1(x) \)) is close to \( \tilde{f}_{\theta_0}^1(x) \) (resp. \( f_{\theta_0}^1(x) \)) in an appropriate sense. We shall first deal with the difference \( \tilde{f}_{\theta_0}^1(x) - f_{\theta_0}^1(x) \). By Lemma 1, under Assumptions 1, 5, 6 and 8, we have (35). We now show that (40) holds for the estimate \( \hat{h}_\delta(\cdot) \) defined in (21). Let \( |\theta - \theta_0| \leq \delta_n \). Note that the conditional density of \( X_i = \mu(X_{i-1}) + \epsilon_i \sigma(X_{i-1}) \) given \( F_{i-1} \) is \( h[(\cdot - \mu(X_{i-1})/\sigma(X_{i-1})]/\sigma(X_{i-1}) \). Then there exists a constant \( \lambda \), independent of \( x, \theta \) and \( n \), such that \( E[W_i^2(x, \theta)|F_{i-1}] \leq b_n \lambda \). Then following the proof of Lemma 2, we obtain (40), by noting that the function \( I_\delta(x) \) in (46) shall now be replaced by

\[
I_\delta(x) = \sum_{i=1}^n h \left( \frac{x \sigma(X_{i-1}) + \mu(X_{i-1}) - \mu(X_{i-1})}{\sigma(X_{i-1})} \right) \frac{\sigma(X_{i-1})}{\sigma(X_{i-1})}. \]

Hence, with (35) and (40), by the argument in the proof of Theorem 2, we obtain a similar version of (50), which implies that

\[
\sup_{t \leq x \leq u} |\tilde{f}_{\theta_0}^1(x) - \tilde{f}_{\theta_0}^1(x)| = O_p(n^{-1/2}). \tag{53}
\]

We shall now establish a similar bound for \( \tilde{f}_{\theta_0}^1(x) - \tilde{f}_{\theta_0}^1(x) \). Let \( I = [l, u] \); let \( (\delta_n)_{n \geq 0} \) be a decreasing sequence satisfying \( \sqrt{n}\delta_n \to \infty \) and \( \sqrt{n}\delta_n = o(\log n) \). Let

\[
W_i^1(x, \theta) = K \left( \frac{x - X_i/\sigma(X_{i-1})}{b_n} \right) = K \left( \frac{x \sigma(X_{i-1}) - \mu(X_{i-1}) - \epsilon_i \sigma(X_{i-1})}{b_n \sigma(X_{i-1})} \right). \]

31
As in the proof of Lemma 2, let \( h_\ast = \sup_x h(x) \), \( K_\ast = \sup_u |K(u)| \) and \( K_2 = \int K^2(u)du \).

By Assumption 1, there exists a constant \( C_K \), only depending on the kernel \( K \), such that

\[
\mathbb{E}[|W^\dagger_i(x, \theta) - W^\dagger_i(x, \theta_0)||\mathcal{F}_{i-1}| \leq C_K(|x| + b_n) \left| 1 - \frac{\sigma_\theta(X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} \right| \frac{\sigma_\theta(X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} h_\ast. \tag{54}
\]

By Assumption 8, \( |\sigma_\theta(X_{i-1}) - \sigma_{\theta_0}(X_{i-1})| \leq \delta_n \hat{s}_\ast(\delta_0, X_{i-1}) \). Hence, for

\[
V_n := \sum_{i=1}^n \frac{\hat{s}_\ast(\delta_0, X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} \left[ 1 + \delta_0 \frac{\hat{s}_\ast(\delta_0, X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} \right],
\]

we have by (54) that, for \( \lambda_0 = 2K_\ast C_K(|l| + |u| + \sup_n b_n)h_\ast \), we have

\[
\sum_{i=1}^n \mathbb{E}[|W^\dagger_i(x, \theta) - W^\dagger_i(x, \theta_0)|^2|\mathcal{F}_{i-1}| \leq \lambda_0 \delta_n V_n. \tag{55}
\]

By Assumption 8, \( v_1 = \mathbb{E}V_1 < \infty \). Let

\[
G_n(x, \theta) = \sum_{i=1}^n \{W^\dagger_i(x, \theta) - \mathbb{E}[W^\dagger_i(x, \theta)|\mathcal{F}_{i-1}]\}
\]

Note that the summands of \( G_n(x, \theta) \) are martingale differences bounded by \( 2K_\ast \). Again by Freedman’s (1975) inequality and (55), let \( t_n = \sqrt{n\delta_n} \log n \), we have

\[
\mathbb{P} \left[ |G_n(x, \theta) - G_n(x, \theta_0)| \geq t_n, V_n \leq 2nv_1 \right] \leq 2 \exp \left[ -\frac{t_n^2}{4K_\ast t_n + 4\lambda_0 \delta_n v_1} \right]. \tag{56}
\]

Since \( \sqrt{n\delta_n} \to \infty \) and \( \sqrt{n\delta_n} = o(\log n) \), letting \( N = n^4 \), as in (43), we have

\[
\mathbb{P} \left[ \sup_{x \in \mathcal{X}} \sup_{|\theta - \theta_0| \leq \delta_0} |G_n([x]_N, [\theta]_N) - G_n([x]_N, \theta_0)| \geq t_n, V_n \leq 2nv_1 \right] = o(n^{-1}),
\]

which, by the argument in (42) and the fact that \( \mathbb{P}(V_n > 2nv_1) \to 0 \), implies

\[
\sup_{x \in \mathcal{X}} \sup_{|\theta - \theta_0| \leq \delta_0} |G_n(x, \theta) - G_n(x, \theta_0)| = O_P(t_n). \tag{57}
\]

We shall next deal with \( \sum_{i=1}^n \mathbb{E}[W^\dagger_i(x, \theta)|\mathcal{F}_{i-1}] \). Analogous to (46), we now define

\[
I_\theta(x) = \sum_{i=1}^n h \left( x \frac{\sigma_\theta(X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} - \frac{\mu_\theta(X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} \right) \frac{\sigma_\theta(X_{i-1})}{\sigma_{\theta_0}(X_{i-1})} . \tag{58}
\]
Then
\[
\sum_{i=1}^{n} \mathbb{E}[W_i^\dagger(x, \theta) - W_i^\dagger(x, \theta_0)|\mathcal{F}_{i-1}] = b_n \int_{\mathbb{R}} K(v) [I_{\theta}(x - b_n v) - I_{\theta_0}(x - b_n v)] dv.
\]

By the Ergodic Theorem and Assumption 8, we have \( \mathbb{E}[\sup_{x \in I} |\hat{I}_\theta(x)|] = O(n) \). Then
\[
\sup_{x \in I} \sup_{|\theta - \theta_0| \leq \delta_n} \left| \sum_{i=1}^{n} \mathbb{E}[W_i^\dagger(x, \theta) - W_i^\dagger(x, \theta_0)|\mathcal{F}_{i-1}] \right| = b_n \delta_n O_P(n). \tag{59}
\]

By (57) and (59),
\[
\sup_{x \in I} \sup_{|\theta - \theta_0| \leq \delta_n} |\hat{f}_\theta(x) - \hat{f}_{\theta_0}(x)| = \frac{O_P(t_n + nb_n \delta_n)}{nb_n} = \frac{O_P(1)}{(nb_n)^{1/2} \log^2 n} \tag{60}
\]

since \( b_n = o(n^{-1/4}) \), \( \sqrt{n} \delta_n \to \infty \) and \( \sqrt{n} \delta_n = o(\log n) \). Hence \( \Xi_n^\dagger - \Xi_n^\diamond = O(\log^{-2} n) \) and the theorem follows.
\( \diamond \)