

# 1 An overview of almost sure convergence

An attractive features of almost sure convergence (which we define below) is that it often reduces the problem to the convergence of deterministic sequences.

First we go through some definitions (these are not very formal). Let  $\Omega_t$  be the set of all possible outcomes (or realisations) at the point  $t$ , and define the random variable  $Y_t$  as the function  $Y_t : \Omega_t \rightarrow \mathbb{R}$ . Define the set of possible outcomes over all time as  $\Omega = \otimes_{t=1}^{\infty} \Omega_t$ , and the random variables  $X_t : \Omega \rightarrow \mathbb{R}$ , where for every  $\omega \in \Omega$ , with  $\omega = (\omega_0, \omega_2, \dots)$ , we have  $X_t(\omega) = Y_t(\omega_t)$ . Hence we have a sequence of random variables  $\{X_t\}_t$  (which we call a random process). When we observe  $\{x_t\}_t$ , this means there exists an  $\omega \in \Omega$ , such that  $X_t(\omega) = x_t$ . To complete things we have a sigma-algebra  $\mathcal{F}$  whose elements are subsets of  $\Omega$  and a probability measure  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ . But we do not have to worry too much about this.

**Definition 1** *We say that the sequence  $\{X_t\}$  converges almost sure to  $\mu$ , if there exists a set  $M \subset \Omega$ , such that  $\mathbb{P}(M) = 1$  and for every  $\omega \in M$  we have*

$$X_t(\omega) \rightarrow \mu.$$

*In other words for every  $\varepsilon > 0$ , there exists an  $N(\omega)$  such that*

$$|X_t(\omega) - \mu| < \varepsilon, \tag{1}$$

*for all  $t > N(\omega)$ . We denote  $X_t \rightarrow \mu$  almost surely, as  $X_t \xrightarrow{a.s.} \mu$ .*

We see in (1) how the definition is reduced to a nonrandom definition. There is an equivalent definition, which is explicitly stated in terms of probabilities, which we do not give here.

The object of this handout is to give conditions under which an estimator  $\hat{a}_n$  converges almost surely to the parameter we are interested in estimating and to derive its limiting distribution.

## 2 Strong consistency of an estimator

### 2.1 The autoregressive process

A simple example of a sequence of random variables  $\{X_t\}$  which are dependent is the autoregressive process. The AR(1) process satisfies the representation

$$X_t = aX_{t-1} + \epsilon_t, \tag{2}$$

where  $\{\epsilon_t\}_t$  are iid random variables with  $\mathbb{E}(\epsilon_t) = 0$  and  $\mathbb{E}(\epsilon_t^2) = 1$ . It has the unique causal solution

$$X_t = \sum_{j=0}^{\infty} a^j \epsilon_{t-j}.$$

From the above we see that  $\mathbb{E}(X_t^2) = (1 - a^2)^{-1}$ . Let  $\sigma^2 := \mathbb{E}(X_0^2) = (1 - a^2)^{-1}$ .

**Question** What can we say about the limit of

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n X_j^2.$$

- If  $\mathbb{E}(\varepsilon_t^4) < \infty$ , then we show that  $\mathbb{E}(\hat{\sigma}_n - \sigma^2)^2 \rightarrow 0$  (*Exercise*).
- What can we say about almost sure convergence?

## 2.2 The strong law of large numbers

Recall the strong law of large numbers; Suppose  $\{X_t\}_t$  is an iid sequence, and  $\mathbb{E}(|X_0|) < \infty$  then by the SLLN we have that

$$\frac{1}{n} \sum_{j=1}^n X_j \xrightarrow{\text{a.s.}} \mathbb{E}(X_0),$$

for the proof see Grimmet and Stirzaker (1994).

- What happens when the random variables  $\{X_t\}$  are dependent?
- In this case we use the notion of *ergodicity* to obtain a similar result.

## 2.3 Some ergodic tools

Suppose  $\{Y_t\}$  is an *ergodic* sequence of random variables (which implies that  $\{Y_t\}_t$  are identically distributed random variables). For the definition of ergodicity and a full treatment see, for example, Billingsley (1965).

- A simple example of an ergodic sequence is  $\{Z_t\}$ , where  $\{Z_t\}_t$  are iid random variables.

**Theorem 1** *One variation of the ergodic theorem: If  $\{Y_t\}$  is an ergodic sequence, where  $\mathbb{E}(|g(Y_t)|) < \infty$ . Then we have*

$$\frac{1}{n} \sum_{i=1}^n g(Y_i) \xrightarrow{\text{a.s.}} \mathbb{E}(g(Y_0)).$$

We give some sufficient conditions for a process to be ergodic. The theorem below is a simplified version of Stout (1974), Theorem 3.5.8.

**Theorem 2** *Suppose  $\{Z_t\}$  is an ergodic sequence (for example iid random variables) and  $g : \mathbb{R}^\infty \rightarrow \mathbb{R}$  is a continuous function. Then the sequence  $\{Y_t\}_t$ , where*

$$Y_t = g(Z_t, Z_{t-1}, \dots),$$

*is an ergodic process.*

- (i) **Example I** Let us consider the sequence  $\{X_t\}$ , which satisfies the AR(1) representation. We will show by using Theorem 2, that  $\{X_t\}$  is an ergodic process. We know that  $X_t$  has the unique (casual) solution

$$X_t = \sum_{j=0}^{\infty} a^j \epsilon_{t-j}.$$

The solution motivates us to define the function

$$g(x_0, x_1, \dots) = \sum_{j=0}^{\infty} a^j x_j.$$

We have that

$$\begin{aligned} |g(x_0, x_1, \dots) - g(y_0, y_1, \dots)| &= \left| \sum_{j=0}^{\infty} a^j x_j - \sum_{j=0}^{\infty} a^j y_j \right| \\ &\leq \sum_{j=0}^{\infty} a^j |x_j - y_j|. \end{aligned}$$

Therefore if  $\max_j |x_j - y_j| \leq |1 - a|\varepsilon$ , then  $|g(x_1, x_2, \dots) - g(y_1, y_2, \dots)| \leq \varepsilon$ . Hence the function  $g$  is continuous (under the sup-norm), which implies, by using Theorem 2, that  $\{X_t\}$  is an ergodic process.

**Application** Using the ergodic theorem we have that

$$\frac{1}{n} \sum_{t=1}^n X_t^2 \xrightarrow{\text{a.s.}} \sigma^2.$$

- (ii) **Example II** A stochastic process often used in finance is the ARCH process.  $\{X_t\}$  is said to be an ARCH(1) process if it satisfies the representation

$$X_t = \sigma_t Z_t \quad \sigma_t^2 = a_0 + a_1 X_{t-1}^2.$$

It has almost surely the solution

$$X_t^2 = a_0 \sum_{j=0}^{\infty} a_1^j \prod_{i=0}^j Z_{t-i}^2.$$

Using the arguments above we can show that  $\{X_t^2\}_t$  is an ergodic process.

### 3 Likelihood estimation

Our object here is to evaluate the maximum likelihood estimator of the AR(1) parameter and to study its asymptotic properties. We recall that the maximum likelihood estimator is the parameter which maximises the joint density of the observations. Since the log-likelihood

often has a simpler form, we often maximise the log density rather than the density (since both the maximum likelihood estimator and maximum log likelihood estimator yield the same estimator).

Suppose we observe  $\{X_t; t = 1, \dots, n\}$  where  $X_t$  are observations from an AR(1) process. Let  $F_\epsilon$   $f_\epsilon$  be the distribution function and the density function of  $\epsilon$  respectively. We first note that the AR(1) process is Markovian, that is

$$\begin{aligned} \mathbb{P}(X_t \leq x | X_{t-1}, X_{t-2}, \dots) &= \mathbb{P}(X_t \leq x | X_{t-1}) \\ \Rightarrow f_a(X_t | X_{t-1}, \dots) &= f_a(X_{t-1} | X_{t-1}). \end{aligned} \tag{3}$$

- The Markov property is where the probability of  $X_t$  given the past is the same as the probability of  $X_t$  given  $X_{t-1}$ .
- To prove (3) we see that

$$\begin{aligned} \mathbb{P}(X_t \leq x | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}) &= \mathbb{P}(aX_{t-1} + \epsilon_t \leq x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}) \\ &= \mathbb{P}(\epsilon_t \leq x_t - ax_{t-1} | X_{t-2} = x_{t-2}) \\ &= \mathbb{P}_\epsilon(\epsilon_t \leq x_t - ax_{t-1}) = \mathbb{P}(X_t \leq x_t | X_{t-1} = x_{t-1}), \end{aligned}$$

hence the process satisfies the Markov property.

By using the above we have

$$\begin{aligned} \mathbb{P}(X_t \leq x | X_{t-1}) &= \mathbb{P}_\epsilon(\epsilon \leq x - aX_{t-1}) \\ \Rightarrow \mathbb{P}(X_t \leq x | X_{t-1}) &= F_\epsilon(x - aX_{t-1}) \\ f_a(X_t | X_{t-1}) &= f_\epsilon(X_t - aX_{t-1}). \end{aligned} \tag{4}$$

Evaluating the joint density and using (4) we see that it satisfies

$$\begin{aligned} f_a(X_1, X_2, \dots, X_n) &= f_a(X_1) \prod_{t=2}^n f_a(X_t | X_{t-1}, X_{t-1}, \dots) \quad (\text{by Bayes theorem}) \\ &= f_a(X_1) \prod_{t=2}^n f(X_t | X_{t-1}) \quad (\text{by the Markov property}) \\ &= f_a(X_1) \prod_{t=2}^n f_\epsilon(X_t - aX_{t-1}) \quad (\text{by (4)}). \end{aligned}$$

Therefore the log likelihood is

$$\log f_a(X_1, X_2, \dots, X_n) = \underbrace{\log f(X_1)}_{\text{often ignored}} + \underbrace{\sum_{k=2}^n \log f_\epsilon(X_k - aX_{k-1})}_{\text{conditional likelihood}}.$$

Usually we ignore the initial distribution  $f(X_1)$  and maximise the conditional likelihood to obtain the estimator.

Hence we use  $\hat{a}_n$  as an estimator of  $a$ , where

$$\hat{a}_n = \arg \max_{a \in \Theta} \sum_{k=2}^n \log f_{\epsilon}(X_t - aX_{t-1}),$$

and  $\Theta$  is the parameter space we do the maximisation in. We say that  $\hat{a}_n$  is the conditional likelihood estimator.

### 3.1 Likelihood function when the innovations are Gaussian

We now consider the special case that the innovations  $\{\epsilon_t\}_t$  are Gaussian. In this case we have that

$$\log f_{\epsilon}(X_t - aX_{t-1}) = -\frac{1}{2} \log 2\pi - (X_t - aX_{t-1})^2.$$

Therefore if we let

$$\mathcal{L}_n(a) = -\frac{1}{n-1} \sum_{t=2}^n (X_t - aX_{t-1})^2,$$

since  $\frac{1}{2} \log 2\pi$  is a constant we have

$$\mathcal{L}_n(a) \propto \sum_{k=2}^n \log f_{\epsilon}(X_t - aX_{t-1}).$$

Therefore the maximum likelihood estimator is

$$\Rightarrow \hat{a}_n = \arg \max_{a \in \Theta} \mathcal{L}_n(a).$$

In the derivations here we shall assume  $\Theta = [-1, 1]$ .

**Definition 2** *An estimator  $\hat{\alpha}_n$  is said to be a consistent estimator of  $\alpha$ , if there exists a set  $M \subset \Omega$ , where  $\mathbb{P}(M) = 1$  and for all  $\omega \in M$  we have*

$$\hat{\alpha}_n(\omega) \rightarrow \alpha.$$

**Question:** Is the likelihood estimator  $\hat{a}_n$  strongly consistent?

- In the case of least squares for AR processes,  $\hat{a}_n$  has the explicit form

$$\hat{a}_n = \frac{\frac{1}{n-1} \sum_{t=2}^n X_t X_{t-1}}{\frac{1}{n-1} \sum_{t=1}^{n-1} X_t^2}.$$

Now by just applying the ergodic theorem to the numerator and denominator we get almost sure convergence of  $\hat{a}_n$  (*exercise*).

- However we will tackle the problem in a rather artificial way and assume that it does not have an explicit form and instead assume that  $\hat{a}_n$  is obtained by minimising  $\mathcal{L}_n(a)$  using a numerical routine. In general this is the most common way of minimising a likelihood function (usually explicit solutions do not exist).
- In order to derive the sampling properties of  $\hat{a}_n$  we need to directly study the likelihood function  $\mathcal{L}_n(a)$ . We will do this now in the least squares case.

Most of the analysis of  $\mathcal{L}_n(a)$  involves repeated application of the ergodic theorem.

- The first clue to solving the problem is to let

$$\ell_t(a) = -(X_t - aX_{t-1})^2.$$

By using Theorem 2 we have that  $\{\ell_t(a)\}_t$  is an ergodic sequence. Therefore by using the ergodic theorem we have

$$\mathcal{L}_n(a) = \frac{1}{n-1} \sum_{t=2}^n \ell_t(a) \xrightarrow{\text{a.s.}} \mathbb{E}(\ell_0(a)).$$

- In other words for every  $a \in [-1, 1]$  we have that  $\mathcal{L}_n(a) \xrightarrow{\text{a.s.}} \mathbb{E}(\ell_0(a))$ . This is what we call almost sure pointwise convergence.

**Remark 1** *It is interesting to note that the least squares/likelihood estimator  $\hat{a}_n$  can also be used even if the innovations do not come from a Gaussian process.*

### 3.2 Strong consistency of a general estimator

We now consider the general case where  $\mathcal{B}_n(a)$  is a ‘criterion’ which we maximise (or minimise). We note that  $\mathcal{B}_n(a)$  includes the likelihood function, least squares criterion etc.

We suppose it has the form

$$\mathcal{B}_n(a) = \frac{1}{n-1} \sum_{j=2}^n g_j(a), \tag{5}$$

where for each  $a \in [c, d]$ ,  $\{g_t(a)\}_t$  is a ergodic sequence. Let

$$\tilde{\mathcal{B}}(a) = \mathbb{E}(g_t(a)), \tag{6}$$

we assume that  $\tilde{\mathcal{B}}(a)$  is continuous and has a unique maximum in  $[c, d]$ .

We define the estimator  $\hat{a}_n$  where

$$\hat{a}_n = \arg \max_{a \in [c, d]} \mathcal{B}_n(a).$$

**Object** To show under what conditions  $\hat{\alpha}_n \xrightarrow{\text{a.s.}} \alpha$ , where  $\alpha = \arg \max \mathcal{B}(a)$ .

Now suppose that for each  $a \in [c, d]$   $\mathcal{B}_n(a) \xrightarrow{\text{a.s.}} \tilde{\mathcal{B}}(a)$ , then this is called almost sure pointwise convergence. That is for each  $a \in [c, d]$  we can show that there exists a set  $M_a$  such that  $M_a \subset \Omega$  where  $P(M_a) = 1$ , and for each  $\omega \in M_a$  and every  $\varepsilon > 0$

$$|\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \leq \varepsilon,$$

for all  $n > N_a(\omega)$ . But the  $N_a(\omega)$  depends on the  $a$ , hence the rate of convergence is not uniform (the rate depends on  $a$ ).

**Remark 2** We will assume that the set  $M_a$  is common for all  $a$ , that is  $M_a = M_{a'}$  for all  $a, a' \in [c, d]$ . We will prove this result in Lemma 1 under the assumption of equicontinuity (defined later), however I think the same result can be shown under the weaker assumption that  $\tilde{\mathcal{B}}(a)$  is continuous.

Returning to the estimator, by using the pointwise convergence we observe that

$$\mathcal{B}_n(a) \leq \mathcal{B}_n(\hat{a}_n) \xrightarrow{\text{a.s.}} \tilde{\mathcal{B}}(\hat{a}_n) \leq \tilde{\mathcal{B}}(a), \quad (7)$$

where  $\hat{a}_n$  is kept fixed in the limit. We now consider the difference  $|\mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(a)|$ , if we can show that  $|\mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(a)| \xrightarrow{\text{a.s.}} 0$ , then  $\hat{a}_n \xrightarrow{\mathcal{P}} a$  almost surely.

Studying  $(\mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(a))$  and using (7) we have

$$\mathcal{B}_n(a) - \tilde{\mathcal{B}}(a) \leq \mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(a) \leq \mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(\hat{a}_n).$$

Therefore we have

$$|\mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(a)| \leq \max \left\{ |\mathcal{B}_n(a) - \tilde{\mathcal{B}}(a)|, |\mathcal{B}_n(\hat{a}_n) - \tilde{\mathcal{B}}(\hat{a}_n)| \right\}.$$

To show that the RHS of the above converges to zero we require not only pointwise convergence but uniform convergence of  $\mathcal{B}_t(a)$ , which we define below.

**Definition 3**  $\mathcal{B}_n(a)$  is said to almost surely converge uniformly to  $\tilde{\mathcal{B}}(a)$ , if

$$\sup_{a \in [c, d]} |\mathcal{B}_n(a) - \tilde{\mathcal{B}}(a)| \xrightarrow{\text{a.s.}} 0.$$

In other words there exists a set  $M \subset \Omega$  where  $P(M) = 1$  and for every  $\omega \in M$ ,

$$\sup_{a \in [c, d]} |\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \rightarrow 0.$$

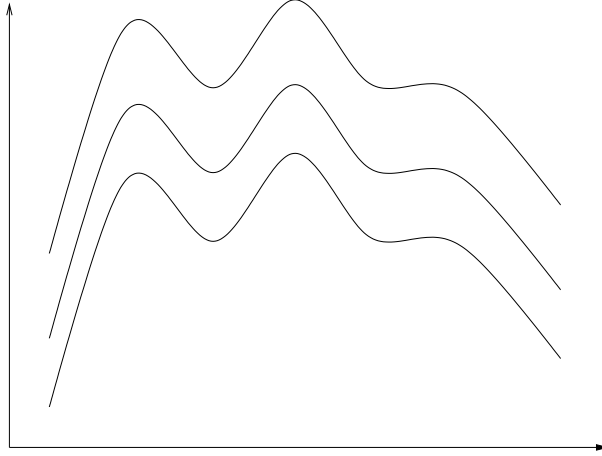
Therefore returning to our problem, if  $\mathcal{B}_n(a) \xrightarrow{\text{a.s.}} \tilde{\mathcal{B}}(a)$  uniformly, then we have the bound

$$|\mathcal{B}_n(\hat{\alpha}_n) - \tilde{\mathcal{B}}(\alpha)| \leq \sup_{a \in [c, d]} |\mathcal{B}_n(a) - \tilde{\mathcal{B}}(a)| \xrightarrow{\text{a.s.}} 0.$$

This implies that  $\mathcal{B}_n(\hat{\alpha}_n) \xrightarrow{\text{a.s.}} \tilde{\mathcal{B}}(\alpha)$  and since  $\tilde{\mathcal{B}}(\alpha)$  has a unique minimum we have that  $\hat{\alpha}_n \xrightarrow{\text{a.s.}} \alpha$ . Therefore if we can show almost sure uniform convergence of  $\mathcal{B}_n(a)$ , then we have strong consistency of the estimator  $\hat{\alpha}_n$ .

**Comment:** Pointwise convergence is relatively easy to show, but how to show uniform convergence?

Figure 1: The middle curve is  $\mathcal{B}(\tilde{a}, \omega)$ . If the sequence  $\{\mathcal{B}_n(a, \omega)\}$  converges uniformly to  $\mathcal{B}(\tilde{a}, \omega)$ , then  $\mathcal{B}_n(a, \omega)$  will lie inside these boundaries, for all  $n > N(\omega)$ .



### 3.3 Uniform convergence and stochastic equicontinuity

We now define the concept of stochastic equicontinuity. We will prove that stochastic equicontinuity together with almost sure pointwise convergence (and a compact parameter space) imply uniform convergence.

**Definition 4** *The sequence of stochastic functions  $\{\mathcal{B}_n(a)\}_n$  is said to be stochastically equicontinuous if there exists a set  $M \in \Omega$  where  $P(M) = 1$  and for every and  $\varepsilon > 0$ , there exists a  $\delta$  and such that for every  $\omega \in M$*

$$\sup_{|a_1 - a_2| \leq \delta} |\mathcal{B}_n(\omega, a_1) - \mathcal{B}_n(\omega, a_2)| \leq \varepsilon,$$

for all  $n > N(\omega)$ .

**Remark 3** *A sufficient condition for stochastic equicontinuity is that there exists an  $N(\omega)$ , such that for all  $n > N(\omega)$   $\mathcal{B}_n(\omega, a)$  belongs to the Lipschitz class  $C(L)$  (where  $L$  is the same for all  $\omega$ ). In general we verify this condition to show stochastic equicontinuity.*

In the following lemma we show that if  $\{\mathcal{B}_n(a)\}_n$  is stochastically equicontinuous and also pointwise convergent, then there exists a set  $M \subset \Omega$ , where  $\mathbb{P}(M) = 1$  and for every  $\omega \in M$ , there is pointwise convergence of  $\mathcal{B}_n(\omega, a) \rightarrow \tilde{\mathcal{B}}(a)$ . This lemma can be omitted on first reading (it is mainly technical).

**Lemma 1** *Suppose the sequence  $\{\mathcal{B}_n(a)\}_n$  is stochastically equicontinuous and also pointwise convergent (that is  $\mathcal{B}_n(a)$  converges almost surely to  $\tilde{\mathcal{B}}(a)$ ), then there exists a set  $M \in \Omega$  where  $P(M) = 1$  and for every  $\omega \in M$  and  $a \in [c, d]$  we have*

$$|\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \rightarrow 0.$$

PROOF. Enumerate all the rationals in the interval  $[c, d]$  and call this sequence  $\{a_i\}_i$ . Then for every  $a_i$  there exists a set  $M_{a_i}$  where  $P(M_{a_i}) = 1$ , such that for every  $\omega \in M_{a_i}$  we have  $|\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i)| \rightarrow 0$ . Define  $M = \cap M_{a_i}$ , since the number of sets is countable  $P(M) = 1$  and for every  $\omega \in M$  and  $a_i$  we have  $\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i) \rightarrow 0$ . Suppose we have equicontinuity for every realisation in  $\tilde{M}$ , and  $\tilde{M}$  is such that  $P(\tilde{M}) = 1$ . Let  $\bar{M} = \tilde{M} \cap \{\cap M_{a_i}\}$ . Let  $\omega \in \bar{M}$ , then for every  $\varepsilon/3 > 0$ , there exists a  $\delta > 0$  such that

$$\sup_{|a_1 - a_2| \leq \delta} |\mathcal{B}_n(\omega, a_1) - \mathcal{B}_n(\omega, a_2)| \leq \varepsilon/3,$$

for all  $n > N(\omega)$ . Now for any given  $a$ , choose a rational  $a_j$  such that  $|a - a_j| \leq \delta$ . By pointwise continuity we have

$$|\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i)| \leq \varepsilon/3,$$

where  $n > N'(\omega)$ . Then we have

$$|\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \leq |\mathcal{B}_n(\omega, a) - \mathcal{B}_n(\omega, a_i)| + |\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i)| + |\tilde{\mathcal{B}}(a) - \tilde{\mathcal{B}}(a_i)| \leq \varepsilon,$$

for  $n > \max(N(\omega), N'(\omega))$ . To summarise for every  $\omega \in \tilde{M}$  and  $a \in [c, d]$ , we have  $|\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \rightarrow 0$ . Hence we have pointwise convergence for every realisation in  $\tilde{M}$ .  $\square$

We now show that equicontinuity implies uniform convergence.

**Theorem 3** *Suppose the set  $[a, b]$  is a compact interval and for every  $a \in [c, d]$   $\mathcal{B}_n(a)$  converges almost surely to  $\tilde{\mathcal{B}}(a)$ . Furthermore assume that  $\{\mathcal{B}_n(a)\}$  is almost surely equicontinuous. Then we have*

$$\sup_{a \in [c, d]} |\mathcal{B}_n(a) - \tilde{\mathcal{B}}(a)| \xrightarrow{a.s.} 0.$$

PROOF. Let  $\bar{M} = M \cap \tilde{M}$  where  $M$  is the set where we have uniform convergence and  $\tilde{M}$  the set where all  $\omega \in \tilde{M}$ ,  $\{\mathcal{B}_n(\omega, a)\}_n$  converge pointwise, it is clear that  $P(\bar{M}) = 1$ . Choose  $\varepsilon/3 > 0$  and let  $\delta$  be such that for every  $\omega \in \tilde{M}$  we have

$$\sup_{|a_1 - a_2| \leq \delta} |\mathcal{B}_n(\omega, a_1) - \mathcal{B}_n(\omega, a_2)| \leq \varepsilon/3,$$

for all  $n > N(\omega)$ . Since  $[c, d]$  is compact it can be divided into a finite number of intervals. Therefore let  $c = \rho_0 \leq \rho_1 \leq \dots \leq \rho_p = d$  and  $\sup_i |\rho_{i+1} - \rho_i| \leq \delta$ . Since  $p$  is finite, there exists a  $\tilde{N}(\omega)$  such that

$$\max_{1 \leq i \leq p} |\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i)| \leq \varepsilon/3,$$

for all  $n > \tilde{N}(\omega)$  (where  $N_i(\omega)$  is such that  $|\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i)| \leq \varepsilon/3$ , for all  $n \geq N_i(\omega)$  and  $\tilde{N}(\omega) = \max_i(N_i(\omega))$ ). For any  $a \in [c, d]$ , choose the  $i$ , such that  $a \in [a_i, a_{i+1}]$ , then by stochastic equicontinuity we have

$$|\mathcal{B}_n(\omega, a) - \mathcal{B}_n(\omega, a_i)| \leq \varepsilon/3,$$

for all  $n > N(\omega)$ . Therefore we have

$$|\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \leq |\mathcal{B}_n(\omega, a) - \mathcal{B}_n(\omega, a_i)| + |\mathcal{B}_n(\omega, a_i) - \tilde{\mathcal{B}}(a_i)| + |\tilde{\mathcal{B}}(a) - \tilde{\mathcal{B}}(a_i)| \leq \varepsilon,$$

for all  $n \geq \max(N(\omega), \tilde{N}(\omega))$ . Since  $\bar{N}(\omega)$  does not depend on  $a$ , then  $\sup_a |\mathcal{B}_n(\omega, a) - \tilde{\mathcal{B}}(a)| \rightarrow 0$ . Furthermore it is true for all  $\omega \in \bar{M}$  and  $\mathbb{P}(\bar{M}) = 1$  hence we have almost sure uniform convergence.  $\square$

The following theorem summarises all the sufficient conditions for almost sure consistency.

**Theorem 4** *Suppose  $\mathcal{B}_n(n)$  and  $\tilde{\mathcal{B}}(a)$  are defined as in (5) and (6) respectively. Let*

$$\hat{\alpha}_n = \arg \max_{a \in [c, d]} \mathcal{B}_n(a) \tag{8}$$

$$\text{and } \alpha = \arg \max_{a \in [c, d]} \tilde{\mathcal{B}}(a). \tag{9}$$

*Assume that  $[c, d]$  is a compact subset and that  $\tilde{\mathcal{B}}(a)$  has a unique maximum in  $[c, d]$ . Furthermore assume that for every  $a \in [c, d]$   $\mathcal{B}_n(a) \xrightarrow{a.s.} \tilde{\mathcal{B}}(a)$  and the sequence  $\{\mathcal{B}_n(a)\}_n$  is stochastically equicontinuous. Then we have*

$$\hat{\alpha}_n \xrightarrow{a.s.} \alpha.$$

### 3.4 Strong consistency of the least squares estimator

We now verify the conditions in Theorem 4 to show that the least squares estimator is strongly consistent.

We recall that

$$\hat{a}_n = \arg \max_{a \in \Theta} \mathcal{L}_n(a), \tag{10}$$

where

$$\mathcal{L}_n(a) = -\frac{1}{n-1} \sum_{t=2}^n (X_t - aX_{t-1})^2.$$

We have already shown that for every  $a \in [-1, 1]$  we have  $\mathcal{L}_n(a) \xrightarrow{a.s.} \tilde{\mathcal{L}}(a)$ , where  $\tilde{\mathcal{L}}(a) := \mathbb{E}(g_0(a))$ . Recalling Theorem 3, since the parameter space  $[-1, 1]$  is compact, to show strong consistency we need only to show that  $\mathcal{L}_n(a)$  is stochastically equicontinuous.

By expanding  $\mathcal{L}_n(a)$  and using the mean value theorem we have

$$\mathcal{L}_n(a_1) - \mathcal{L}_n(a_2) = \nabla \mathcal{L}_n(\bar{a})(a_1 - a_2), \tag{11}$$

where  $\bar{a} \in [\min[a_1, a_2], \max[a_1, a_2]]$  and

$$\nabla \mathcal{L}_n(\alpha) = \frac{-2}{n-1} \sum_{t=2}^n X_{t-1}(X_t - \alpha X_{t-1}).$$

Because  $\alpha \in [-1, 1]$  we have

$$|\nabla \mathcal{L}_n(\alpha)| \leq \mathcal{D}_n,$$

where

$$\mathcal{D}_n = \frac{2}{n-1} \sum_{t=2}^n (|X_{t-1}X_t| + X_{t-1}^2).$$

Since  $\{X_t\}_t$  is an ergodic process, then  $\{|X_{t-1}X_t| + X_{t-1}^2\}$  is an ergodic process. Therefore, if  $\text{var}(\epsilon_0) < \infty$ , by using the ergodic theorem we have

$$\mathcal{D}_n \xrightarrow{\text{a.s.}} 2\mathbb{E}(|X_{t-1}X_t| + X_{t-1}^2).$$

Let  $\mathcal{D} := 2\mathbb{E}(|X_{t-1}X_t| + X_{t-1}^2)$ . Therefore there exists a set  $M \in \Omega$ , where  $\mathbb{P}(M) = 1$  and for every  $\omega \in M$  and  $\varepsilon > 0$  we have

$$|\mathcal{D}_n(\omega) - \mathcal{D}| \leq \varepsilon,$$

for all  $n > N(\omega)$ . Substituting the above into (11) we have

$$\begin{aligned} |\mathcal{L}_n(\omega, a_1) - \mathcal{L}_n(\omega, a_2)| &\leq \mathcal{D}_n(\omega)|a_1 - a_2| \\ &\leq (\mathcal{D} + \varepsilon)|a_1 - a_2|, \end{aligned}$$

for all  $n \geq N(\omega)$ . Therefore for every  $\varepsilon > 0$ , there exists a  $\delta := \varepsilon/(\mathcal{D} + \varepsilon)$  such that

$$|\mathcal{L}_n(\omega, a_1) - \mathcal{L}_n(\omega, a_2)| \leq (\mathcal{D} + \varepsilon)|a_1 - a_2|,$$

for all  $n \geq N(\omega)$ . Since this is true for all  $\omega \in M$  we see that  $\{\mathcal{L}_n(a)\}$  is stochastically equicontinuous.

**Theorem 5** *Let  $\hat{a}_n$  be defined as in (10). Then we have*

$$\hat{a}_n \xrightarrow{\text{a.s.}} a.$$

PROOF. Since  $\{\mathcal{L}_n(a)\}$  is almost sure equicontinuous, the parameter space  $[-1, 1]$  is compact and we have pointwise convergence of  $\mathcal{L}_n(\alpha) \xrightarrow{\text{a.s.}} \tilde{\mathcal{L}}(\alpha)$ , by using Theorem 4 we have that  $\hat{a}_n \xrightarrow{\text{a.s.}} \alpha$ , where  $\alpha = \max \tilde{\mathcal{L}}(\alpha)$ . Finally we need to show that  $\alpha \equiv a$ . Since

$$\tilde{\mathcal{L}}(\alpha) = \mathbb{E}(\ell_0(a)) = -\mathbb{E}(X_1 - \alpha X_0)^2,$$

we see by differentiating  $\tilde{\mathcal{L}}(\alpha)$  that it is maximised when  $\alpha = \mathbb{E}(X_0 X_1)/\mathbb{E}(X_0^2)$ . Inspecting the AR(1) process we see that

$$\begin{aligned} X_t &= aX_{t-1} + \epsilon_t \\ \Rightarrow X_t X_{t-1} &= aX_{t-1}^2 + \epsilon_t X_{t-1} \\ \Rightarrow \mathbb{E}(X_t X_{t-1}) &= a\mathbb{E}(X_{t-1}^2) + \underbrace{\mathbb{E}(\epsilon_t X_{t-1})}_{=0}. \end{aligned}$$

Therefore  $a = \mathbb{E}(X_0 X_1)/\mathbb{E}(X_0^2)$ , hence  $\alpha \equiv a$ , thus we have shown strong consistency of the least squares estimator.  $\square$

- (i) This is the general method for showing almost sure convergence of a whole class of estimators. Using a similar technique one can show strong consistency of maximum likelihood estimator of the ARCH(1) process, see Berkes, Horváth, and Kokoszka (2003) for details.
- (ii) Described here was strong consistency where we showed  $\hat{\alpha}_n \xrightarrow{\text{a.s.}} \alpha$ . Using similar tools one can show weak consistency where  $\hat{\alpha}_n \xrightarrow{\mathcal{P}} \alpha$  (convergence in probability), which requires a much weaker set of conditions. For example, rather than show almost sure pointwise convergence we would show pointwise convergence in probability. Rather than stochastic equicontinuity we would show equicontinuity in probability, that is for every  $\epsilon > 0$  there exists a  $\delta$  such that

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \sup_{|a_1 - a_2| \leq \delta} |\mathcal{B}_n(a_1) - \mathcal{B}_n(a_2)| > \epsilon \right) \rightarrow 0.$$

Some of these conditions are easier to verify than almost sure conditions.

## 4 Central limit theorems

Recall once again the AR(1) process  $\{X_t\}$ , where  $X_t$  satisfies

$$X_t = aX_{t-1} + \epsilon_t.$$

It is of interest to check if there actually is dependency in  $\{X_t\}$ , if there is no dependency then  $a = 0$  (in which case  $\{X_t\}$  would be white noise). Of course in most situations we only have an estimator of  $a$  (for example,  $\hat{\alpha}_n$  defined in (10)).

- (i) Given the estimator  $\hat{\alpha}_n$ , how to see if  $a = 0$ ?
- (ii) Usually we would do a hypothesis test, with  $H_0: a = 0$  against the alternative of say  $a \neq 0$ . However in order to do this we require the distribution of  $\hat{\alpha}_n$ .
- (iii) Usually evaluating the exact sample distribution is extremely hard or close to impossible. Instead we would evaluate the limiting distribution which in general is a lot easier.
- (iv) In this section we shall show asymptotic normality of  $\sqrt{n}(\hat{\alpha}_n - a)$ . The reason for normalising by  $\sqrt{n}$ , is that  $(\hat{\alpha}_n - a) \xrightarrow{\text{a.s.}} 0$  as  $n \rightarrow \infty$ , hence in terms of distributions it converges towards the point mass at zero. Therefore we need to increase the magnitude of the difference  $\hat{\alpha}_n - a$ . We can show that  $(\hat{\alpha}_n - a) = O(n^{-1/2})$ , therefore  $\sqrt{n}(\hat{\alpha}_n - a) = O(1)$ . Multiplying  $(\hat{\alpha}_n - a)$  by anything larger would mean that its limit goes to infinity, multiplying  $(\hat{\alpha}_n - a)$  by something smaller in magnitude would mean its limit goes to zero, so  $n^{1/2}$ , is the happy median.

We often use  $\nabla \mathcal{L}_n(a)$  to denote the derivative of  $\mathcal{L}_n(a)$  with respect to  $a$  ( $\frac{\partial \mathcal{L}_n(a)}{\partial a}$ ). Since  $\hat{a}_n = \arg \max \mathcal{L}_n(a)$ , we observe that  $\nabla \mathcal{L}_n(\hat{a}_n) = 0$ . Now expanding  $\nabla \mathcal{L}_n(\hat{a}_n)$  about  $a$  (the true parameter) we have

$$\begin{aligned}\nabla \mathcal{L}_n(\hat{a}_n) - \nabla \mathcal{L}_n(a) &= \nabla^2 \mathcal{L}_n(\hat{a}_n - a), \\ \Rightarrow -\nabla \mathcal{L}_n(a) &= \nabla^2 \mathcal{L}_n(\hat{a}_n - a),\end{aligned}\tag{12}$$

where  $\nabla \mathcal{L}_n(\alpha) = \frac{\partial^2 \mathcal{L}_n}{\partial \alpha^2}$ ,

$$\begin{aligned}\nabla \mathcal{L}_n(a) &= \frac{-2}{n-1} \sum_{t=2}^n X_{t-1}(X_t - aX_{t-1}) = \frac{-2}{n-1} \sum_{t=2}^n X_{t-1}\epsilon_t \\ \text{and } \nabla^2 \mathcal{L}_n &= \frac{2}{n-1} \sum_{t=2}^n X_{t-1}^2.\end{aligned}$$

Therefore by using (12) we have

$$(\hat{a}_n - a) = - (\nabla^2 \mathcal{L}_n)^{-1} \nabla \mathcal{L}_n(a).\tag{13}$$

Since  $\{X_t^2\}$  are ergodic random variables, by using the ergodic theorem we have  $\nabla^2 \mathcal{L}_n \xrightarrow{\text{a.S.}} 2\mathbb{E}(X_0^2)$ . This with (13) implies

$$\sqrt{n}(\hat{a}_n - a) = - \underbrace{(\nabla^2 \mathcal{L}_n)^{-1}}_{\text{a.S.} \cdot (2\mathbb{E}(X_0^2))^{-1}} \sqrt{n} \nabla \mathcal{L}_n(a).\tag{14}$$

To show asymptotic normality of  $\sqrt{n}(\hat{a}_n - a)$ , will show asymptotic normality of  $\sqrt{n} \nabla \mathcal{L}_n(a)$ .

We observe that

$$\nabla \mathcal{L}_n(a) = \frac{-2}{n-1} \sum_{t=2}^n X_{t-1}\epsilon_t,$$

is the sum of martingale differences, since  $\mathbb{E}(X_{t-1}\epsilon_t|X_{t-1}) = X_{t-1}\mathbb{E}(\epsilon_t|X_{t-1}) = X_{t-1}\mathbb{E}(\epsilon_t) = 0$ . In order to show asymptotic of  $\nabla \mathcal{L}_n(a)$  we will use the martingale central limit theorem.

#### 4.0.1 Martingales and the conditional likelihood

First a quick overview.  $\{Z_t; t = 1, \dots, \infty\}$  are called martingale differences if

$$\mathbb{E}(Z_t|Z_{t-1}, Z_{t-2}, \dots) = 0.$$

An example is the sequence  $\{X_{t-1}\epsilon_t\}_t$  considered above. Because  $\epsilon_t$  and  $X_{t-1}, X_{t-2}, \dots$  are independent then  $\mathbb{E}(X_{t-1}\epsilon_t|X_{t-1}, X_{t-2}, \dots) = 0$ .

The stochastic sum  $\{\mathcal{S}_n\}_n$ , where

$$\mathcal{S}_n = \sum_{k=1}^n Z_k$$

is called a martingale if  $\{Z_t\}$  are martingale differences.

First we show that the gradient of the conditional log likelihood, defined as

$$\mathcal{C}_n(\theta) = \sum_{t=2}^n \frac{\partial \log f_\theta(X_t|X_{t-1}, \dots, X_1)}{\partial \theta},$$

at the true parameter  $\theta_0$  is the sum of martingale differences. By definition if  $\mathcal{C}_n(\theta_0)$  is the sum of martingale differences then

$$\mathbb{E} \left( \frac{\partial \log f_\theta(X_t|X_{t-1}, \dots, X_1)}{\partial \theta} \Big|_{\theta=\theta_0} \Big| X_{t-1}, X_{t-2}, \dots, X_1 \right) = 0.$$

Rewriting the above in terms of integrals and exchanging derivative with integral we have

$$\begin{aligned} & \mathbb{E} \left( \frac{\partial \log f_\theta(X_t|X_{t-1}, \dots, X_1)}{\partial \theta} \Big|_{\theta=\theta_0} \Big| X_{t-1}, X_{t-2}, \dots, X_1 \right) \\ &= \int \frac{\partial \log f_\theta(x_t|X_{t-1}, \dots, X_1)}{\partial \theta} \Big|_{\theta=\theta_0} f_{\theta_0}(x_t|X_{t-1}, \dots, X_1) dx_t \\ &= \int \frac{1}{f_{\theta_0}(x_t|X_{t-1}, \dots, X_1)} \frac{\partial f_\theta(x_t|X_{t-1}, \dots, X_1)}{\partial \theta} \Big|_{\theta=\theta_0} f_{\theta_0}(x_t|X_{t-1}, \dots, X_1) dx_t \\ &= \frac{\partial}{\partial \theta} \left( \int f_\theta(x_t|X_{t-1}, \dots, X_1) dx_t \right) \Big|_{\theta=\theta_0} = 0. \end{aligned}$$

Therefore  $\left\{ \frac{\partial \log f_\theta(X_t|X_{t-1}, \dots, X_1)}{\partial \theta} \Big|_{\theta=\theta_0} \right\}_t$  are a sequence of martingale differences and  $\mathcal{C}_t(\theta_0)$  is the sum of martingale differences (hence it is a martingale).

## 4.1 The martingale central limit theorem

Let us define  $\mathcal{S}_n$  as

$$\mathcal{S}_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n Y_t, \tag{15}$$

where  $\mathcal{F}_t = \sigma(Y_t, Y_{t-1}, \dots)$ ,  $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = 0$  and  $\mathbb{E}(Y_t^2) < \infty$ . In the following theorem adapted from Hall and Heyde (1980), Theorem 3.2 and Corollary 3.1, we show that  $\mathcal{S}_n$  is asymptotically normal.

**Theorem 6** *Let  $\{\mathcal{S}_n\}_n$  be defined as in (15). Further suppose*

$$\frac{1}{n} \sum_{t=1}^n Y_t^2 \xrightarrow{\mathcal{P}} \sigma^2, \tag{16}$$

where  $\sigma^2$  is a finite constant, for all  $\varepsilon > 0$ ,

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon \sqrt{n}) | \mathcal{F}_{t-1}) \xrightarrow{\mathcal{P}} 0, \tag{17}$$

(this is known as the conditional Lindeberg condition) and

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 | \mathcal{F}_{t-1}) \xrightarrow{\mathcal{P}} \sigma^2. \quad (18)$$

Then we have

$$\mathcal{S}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2). \quad (19)$$

## 4.2 Asymptotic normality of the least squares estimator

We now use Theorem 6 to show that  $\sqrt{n} \nabla \mathcal{L}_n(a)$  is asymptotically normal, which means we have to verify conditions (16)-(18). We note in our example that  $Y_t := X_{t-1} \epsilon_t$ , and that the series  $\{X_{t-1} \epsilon_t\}_t$  is an ergodic process. Furthermore, since for any function  $g$ ,  $\mathbb{E}(g(X_{t-1} \epsilon_t) | \mathcal{F}_{t-1}) = \mathbb{E}(g(X_{t-1} \epsilon_t) | X_{t-1})$ , where  $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \dots)$  we need only to condition on  $X_{t-1}$  rather than the entire sigma-algebra  $\mathcal{F}_{t-1}$ .

**C1** : By using the ergodicity of  $\{X_{t-1} \epsilon_t\}_t$  we have

$$\frac{1}{n} \sum_{t=1}^n Y_t^2 = \frac{1}{n} \sum_{t=1}^n X_{t-1}^2 \epsilon_t^2 \xrightarrow{\mathcal{P}} \mathbb{E}(X_{t-1}^2) \underbrace{\mathbb{E}(\epsilon_t^2)}_{=1} = \sigma^2.$$

**C2** : We now verify the conditional Lindeberg condition.

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon \sqrt{n}) | \mathcal{F}_{t-1}) = \frac{1}{n} \sum_{t=1}^n \mathbb{E}(X_{t-1}^2 \epsilon_t^2 I(|X_{t-1} \epsilon_t| > \varepsilon \sqrt{n}) | X_{t-1})$$

We now use the Cauchy-Schwartz inequality for conditional expectations to split  $X_{t-1}^2 \epsilon_t^2$  and  $I(|X_{t-1} \epsilon_t| > \varepsilon)$ . We recall that the Cauchy-Schwartz inequality for conditional expectations is  $\mathbb{E}(X_t Y_t | \mathcal{G}) \leq [\mathbb{E}(X_t^2 | \mathcal{G}) \mathbb{E}(Y_t^2 | \mathcal{G})]^{1/2}$  almost surely. Therefore

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon \sqrt{n}) | \mathcal{F}_{t-1}) \\ & \leq \frac{1}{n} \sum_{t=1}^n \left\{ \mathbb{E}(X_{t-1}^4 \epsilon_t^4 | X_{t-1}) \mathbb{E}(I(|X_{t-1} \epsilon_t| > \varepsilon \sqrt{n})^2 | X_{t-1}) \right\}^{1/2} \\ & \leq \frac{1}{n} \sum_{t=1}^n X_{t-1}^2 \mathbb{E}(\epsilon_t^4)^{1/2} \left\{ \mathbb{E}(I(|X_{t-1} \epsilon_t| > \varepsilon \sqrt{n})^2 | X_{t-1}) \right\}^{1/2}. \end{aligned} \quad (20)$$

We note that rather than use the Cauchy-Schwartz inequality we can use a generalisation of it called the Hölder inequality. The Hölder inequality states that if  $p^{-1} + q^{-1} = 1$ , then  $\mathbb{E}(XY) \leq \{\mathbb{E}(X^p)\}^{1/p} \{\mathbb{E}(Y^q)\}^{1/q}$  (the conditional version also exists). The advantage of using this inequality is that one can reduce the moment assumptions on  $X_t$ .

Returning to (20), and studying  $\mathbb{E}(I(|X_{t-1}\epsilon_t| > \varepsilon)^2|X_{t-1})$  we use that  $\mathbb{E}(I(A)) = \mathbb{P}(A)$  and the Chebyshev inequality to show

$$\begin{aligned} & \mathbb{E}(I(|X_{t-1}\epsilon_t| > \varepsilon\sqrt{n})^2|X_{t-1}) = \mathbb{E}(I(|X_{t-1}\epsilon_t| > \varepsilon\sqrt{n})|X_{t-1}) \\ &= \mathbb{E}(I(|\epsilon_t| > \varepsilon\sqrt{n}/X_{t-1})|X_{t-1}) \\ &= \mathbb{P}_\varepsilon(|\epsilon_t| > \frac{\varepsilon\sqrt{n}}{X_{t-1}}) \leq \frac{X_{t-1}^2 \text{var}(\epsilon_t)}{\varepsilon^2 n}. \end{aligned} \quad (21)$$

Substituting (21) into (20) we have

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon\sqrt{n}) | \mathcal{F}_{t-1}) \\ & \leq \frac{1}{n} \sum_{t=1}^n X_{t-1}^2 \mathbb{E}(\epsilon_t^4)^{1/2} \left\{ \frac{X_{t-1}^2 \text{var}(\epsilon_t)}{\varepsilon^2 n} \right\}^{1/2} \\ & \leq \frac{\mathbb{E}(\epsilon_t^4)^{1/2}}{\varepsilon n^{3/2}} \sum_{t=1}^n |X_{t-1}|^3 \mathbb{E}(\epsilon_t^4)^{1/2} \\ & \leq \frac{\mathbb{E}(\epsilon_t^4)^{1/2}}{\varepsilon n^{1/2}} \frac{1}{n} \sum_{t=1}^n |X_{t-1}|^3. \end{aligned}$$

If  $\mathbb{E}(\epsilon_t^4) < \infty$ , then  $\mathbb{E}(X_t^4) < \infty$ , therefore by using the ergodic theorem we have  $\frac{1}{n} \sum_{t=1}^n |X_{t-1}|^3 \xrightarrow{\text{a.S.}} \mathbb{E}(|X_0|^3)$ . Since almost sure convergence implies convergence in probability we have

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon\sqrt{n}) | \mathcal{F}_{t-1}) \leq \underbrace{\frac{\mathbb{E}(\epsilon_t^4)^{1/2}}{\varepsilon n^{1/2}}}_{\rightarrow 0} \underbrace{\frac{1}{n} \sum_{t=1}^n |X_{t-1}|^3}_{\xrightarrow{\mathbb{P}} \mathbb{E}(|X_0|^3)} \\ & \xrightarrow{\mathbb{P}} 0. \end{aligned}$$

Hence condition (17) is satisfied.

**C3** : We need to verify that

$$\frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 | \mathcal{F}_{t-1}) \xrightarrow{\mathbb{P}} \sigma^2.$$

Since  $\{X_t\}_t$  is an ergodic sequence we have

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^n \mathbb{E}(Y_t^2 | \mathcal{F}_{t-1}) = \frac{1}{n} \sum_{t=1}^n \mathbb{E}(X_{t-1}^2 \varepsilon^2 | X_{t-1}) \\ &= \frac{1}{n} \sum_{t=1}^n X_{t-1}^2 \mathbb{E}(\varepsilon^2 | X_{t-1}) = \mathbb{E}(\varepsilon^2) \underbrace{\frac{1}{n} \sum_{t=1}^n X_{t-1}^2}_{\xrightarrow{\text{a.S.}} \mathbb{E}(X_0^2)} \\ & \xrightarrow{\mathbb{P}} \mathbb{E}(\varepsilon^2) \mathbb{E}(X_0^2) = \sigma^2, \end{aligned}$$

hence we have verified condition (18).

Altogether conditions C1-C3 imply that

$$\sqrt{n}\nabla\mathcal{L}_n(a) = \frac{1}{\sqrt{n}} \sum_{t=1}^n X_{t-1}\epsilon_t \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2). \quad (22)$$

Recalling (14) and that  $\sqrt{n}\nabla\mathcal{L}_n(a) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$  we have

$$\sqrt{n}(\hat{a}_n - a) = - \underbrace{(\nabla^2\mathcal{L}_n)^{-1}}_{\text{a.s.} \cdot (2\mathbb{E}(X_0^2))^{-1}} \underbrace{\sqrt{n}\nabla\mathcal{L}_n(a)}_{\xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)}. \quad (23)$$

Using that  $\mathbb{E}(X_0^2) = \sigma^2$ , this implies that

$$\sqrt{n}(\hat{a}_n - a) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{4}(\sigma^2)^{-1}\right). \quad (24)$$

Thus we have derived the limiting distribution of  $\hat{a}_n$ .

## References

- Berkes, I., Horváth, L., & Kokoskza, P. (2003). GARCH processes: Structure and estimation. *Bernoulli*, 9, 201-2017.
- Billingsley, P. (1965). *Ergodic theory and information*.
- Grimmet, G., & Stirzaker. (1994). *Probability and random processes*.
- Hall, P., & Heyde, C. (1980). *Martingale Limit Theory and its Application*. New York: Academic Press.
- Stout, W. (1974). *Almost Sure Convergence*. New York: Academic Press.