# On Universally Efficient Estimation of the First-Order Autoregressive Parameter and Universal Data Compression

NERI MERHAV, MEMBER, IEEE, AND JACOB ZIV, FELLOW, IEEE

*Abstract*—**A universal nearly efficient estimator is proposed for the first order autoregressive (AR) model where the probability distribution of the driving noise is unknown. The proposed estimator has an intuitively appealing relation to universal data compression and to universal tests for randomness.**

*Index Terms*—**asymptotically efficient estimation, universal estimation, Cramér–Rao bound, Fisher information, autoregressive process, universal coding.**

## I. Introduction

CONSIDER A STATIONARY first-order autoregressive (AR) process that satisfies the stochastic difference equation

$$X_i = \theta X_{i-1} + V_i, \tag{1.1}$$

where $\{V_i\}$ are independent identically distributed (i.i.d.) random variables governed by an unknown probability distribution function (PDF) $F(x) \triangleq \Pr\{V_i \le x\}$. The problem of estimating $\theta$ from a finite sample $x_0, x_1, \cdots, x_n$ has been widely studied in the literature, where usually $F$ is assumed to be of known form (in particular, Gaussian). The case of unknown $F$ has been discussed as well by several investigators. Minimum distance (MD) estimators, which do not depend on the underlying PDF have been suggested by Wang [1] and Koul [2], and were shown to be robust in a qualitative sense. Koul [2] assumed $F$ to have a symmetric derivative and demonstrated the existence of an optimal estimator among estimators in a certain class, but no global optimality results have been established. Denby and Martin [3] also discussed some robustness properties of an $M$-estimator (maximum likelihood type estimator) they have proposed. A more comprehensive study concerning robust estimation has been reported by Millar [4]. For the AR model, however, the first step towards *universal* estimation, that is uniformly asymptotically efficient estimation that is independent of the unknown PDF, was taken by Beran [5], who suggested an estimator that is universal with respect to a relatively small class of PDF's. Stronger results have been derived more recently by Kreiss [6], [7] who has proposed universally efficient estimators for the autoregressive moving average (ARMA) model where $F$, the PDF of the driving noise, is completely unknown. In [6] Kreiss assumes, among other regularity conditions, a symmetric density $f = F'$ and in [7] the symmetry requirement was relaxed, but new regularity conditions imposed are considerably demanding.

In this paper, universal nearly efficient estimation of a first order AR parameter is studied under regularity conditions weaker than those in [6] and [7]. The main purpose of this paper, however, is to demonstrate that universal estimators for the AR model can be derived from universal data compression algorithms and universal tests for randomness [9]. In other words, estimators derived appropriately from efficient universal codes, can be expected to inherit good estimation performance under some conditions. The estimator proposed in this paper, as we shall see later, has a simple information theoretic interpretation related to universal coding, which can be easily generalized to the higher-order case and to other parametric models, e.g., the one-sample location model, the two-sample location model, and the linear regression model.

Specifically, fix a real number $t$ and let $Z_i(t) = X_i - X_{i-1}t$, $i = 1, \cdots, n$. Suppose that $L$ divides $n$ and consider the $L$-tuples $[Z_1(t), \cdots, Z_L(t)]$, $[Z_{L+1}(t), \cdots, Z_{2L}(t)], \cdots$, $[Z_{n-L+1}(t), \cdots, Z_n(t)]$. Next, denote by $H_L(t)$ the unnormalized $L$th order empirical entropy associated with these $L$-tuples, where each $Z_i(t)$, $i = 1, \cdots, n$ is quantized to $(k+1)$ levels. Similarly, let $H^i(t)$, $i = 1, \cdots, L$ denote the first order empirical entropy associated with the quantized version of the sequence $\{Z_i(t), Z_{L+i}(t), \cdots, Z_{n-L+i}(t)\}$. (More precise definitions will be given in Section II.) Generally speaking, our estimation $\hat{\theta}_n(k, L)$ is given by

$$\hat{\theta}_n(k, L) = \arg\max_{\theta \in \Theta_n} \left[ H_L(t) - \sum_{i=1}^{L} H^i(t) \right], \tag{1.2}$$

where $\Theta_n \subset (-1, 1)$ is a uniform grid formed by integer

Manuscript received March 7, 1988; revised March 2, 1990. This work was presented in part at the IEEE International Symposium on Information Theory, San Diego, CA, January 14–19, 1990.

N. Merhav was with AT&T Bell Laboratories, Room 2C-543, 600 Mountain Ave., Murray Hill, NJ 07974. He is now with the Department of Electrical Engineering, Technion I.I.T., Technion City, Haifa 32000, Israel.

J. Ziv is with the Department of Electrical Engineering, Technion I.I.T., Technion City, Haifa 32000, Israel.

IEEE Log Number 9038008.

0018-9448/90/1100-1245$01.00 ©1990 IEEE

multiples of $\delta/\sqrt{n}$ for some fixed real number $\delta > 0$. It is shown that the estimation error, when multiplied by $\sqrt{n}$, has a normal asymptotic distribution with mean arbitrarily close to zero and variance arbitrarily close to the Cramér–Rao lower bound, provided that $\delta$ is sufficiently small, $k$ and $L$ are large enough, and the quantization is sufficiently fine. This does *not* imply that the asymptotic mean-squared error is close to the Cramér–Rao bound. Following [8, Theorem 3.2.3], however, such a result implies, under certain regularity conditions, that the proposed estimator is nearly optimal in the sense of maximizing $\lim_{n\to\infty} P_\theta[-a \le \sqrt{n}(\hat{\theta}_n - \theta) \le b]$ for all $a > 0$, $b > 0$ and $-1 < \theta < 1$, in the class of all asymptotically median-unbiased estimators, that is, estimators for which

$$\lim_{n\to\infty} P_\theta\left[\hat{\theta}_n \le \theta\right] = \lim_{n\to\infty} P_\theta\left[\hat{\theta}_n \ge \theta\right] = \frac{1}{2}, \qquad (1.3)$$

where $P_\theta$ denotes probability with respect to $\theta$.

Clearly, the memory requirement and the computational effort associated with the estimator $\hat{\theta}_n(k,L)$ grows exponentially with $L$. For this reason, it is interesting to calculate the asymptotic variance of this estimator as $k \to \infty$ while $L$ is kept fixed. It will be shown that for fixed $L \ge 2$, and $k$ sufficiently large, the estimator $\hat{\theta}_n(k,L)$ has asymptotic variance (in the previous sense) that can be made arbitrarily close to the Cramér–Rao bound times a factor of

$$\left(1 - \frac{1}{L}\frac{1-\theta^{2L}}{1-\theta^2}\right)^{-1}, \qquad (1.4)$$

independently of $F$. Thus, if $L = 2$, for instance, the asymptotic variance is $2/(1-\theta^2)$ times the Cramér–Rao bound for sufficiently large $k$.

The relationship between this estimator and universal data compression is as follows. We seek a value of $t$ for which $Z(t) = \{Z_1(t), \cdots, Z_n(t)\}$ "looks as much as possible" like white noise in the sense that encoding $Z(t)$ in blocks of length $L$, which takes about $H_L(t)$ bits, is essentially equivalent to a letter-by-letter data compression which disregards memory and nearly $\sum_{i=1}^{L} H^i(t)$ bits are required. Note, that by increasing $L$, longer term memory is removed by appropriately adjusting $t$, resulting in a more accurate estimation of $\theta$. This can be considered as a generalization of the decorrelation (whitening) approach, commonly used to remove second order dependencies. While in the Gaussian case, second order (wide sense) whitening is sufficient for estimating $\theta$ efficiently, in the general case, efficient estimation can be approached only by whitening $Z(t)$ in the *strict* sense, namely, complete removal of statistical dependencies (rather than just correlations) among the components of $Z(t)$. A similar idea has been used in [9] for deriving a universal test for randomness which is asymptotically optimal in the Neyman–Pearson sense.

The remaining part of this paper is organized as follows. In Section II, we provide a more precise statement of the main result. Section III discusses this result and demonstrates how the underlying idea can be extended to other parameter estimation problems. Finally, in Section IV, we analyze the asymptotic performance of the proposed estimator and prove the main result. The technique of the proof is essentially the same as in Bhattacharya [10]. For the sake of completeness, however, we cite the necessary auxiliary results from [10].

## II. Statement of Main Result

We first assume several regularity conditions about $F$.

1) The derivatives $f(x) \triangleq (d/dx)F(x)$, $f'(x) \triangleq (d^2/dx^2)F(x)$ exist everywhere and are continuous.
2) The variance of the driving noise $\sigma^2 \triangleq \int_{-\infty}^{\infty} x^2 f(x)\,dx$ is finite and strictly positive.
3) The density $f$ is bounded and strictly positive everywhere.
4) The driving noise has mean zero, i.e., $\int_{-\infty}^{\infty} x f(x)\,dx = 0$.
5) $0 < \int_{-\infty}^{\infty} [f'(x)]^2 / f(x)\,dx < \infty$.

Note, that the integral in 5) is related to the Fisher information of the AR parameter,

$$I(f,\theta) = \frac{\sigma^2}{1-\theta^2}\int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)}\,dx. \qquad (2.1)$$

Now, select $k$ real numbers (quantization levels)

$$-\infty = a_0 < a_1 < \cdots < a_k < a_{k+1} = \infty,$$

and let $\chi_j(\cdot)$, $j = 1, 2, \cdots, k+1$, denote the indicator function of the interval $C_j = (a_{j-1}, a_j]$. Next, fix a real number $t$, let

$$Z_i(t) = X_i - tX_{i-1}, \qquad i = 1, 2, \cdots, n \qquad (2.2)$$

assume that a fixed integer $L$ divides $n$, and define the following quantities. The relative frequency of the $j$th point in each $L$-tuple being in $C_i$,

$$q_i^j(t) = \frac{L}{n}\sum_{r=0}^{n/L-1} \chi_i\left[Z_{Lr+j}(t)\right], \qquad j = 1, \cdots, L, \quad (2.3)$$

the relative frequencies of quantized $L$-tuples,

$$q_{i_1 i_2 \cdots i_L}(t) = \frac{L}{n}\sum_{j=0}^{n/L-1}\prod_{l=1}^{L} \chi_{i_l}\left[Z_{Lj+l}(t)\right], \qquad (2.4)$$

the (marginal) empirical entropy associated with (2.3),

$$H^j(t) = -\sum_{i=1}^{k+1} q_i^j(t)\log q_i^j(t), \qquad j = 1, \cdots, L, \quad (2.5)$$

the $L$th order empirical entropy associated with (2.4),

$$H_L(t) = -\sum_{i_1=1}^{k+1}\cdots\sum_{i_L=1}^{k+1} q_{i_1\cdots i_L}(t)\log q_{i_1\cdots i_L}(t), \quad (2.6)$$

and finally,

$$Q_n(t) = H_L(t) - \sum_{j=1}^{L} H^j(t). \qquad (2.7)$$

Observe that $Q_n(t) \le 0$ for any sample and every $t$, with equality iff the empirical measure $q_{i_1 \cdots i_L}(t)$ factorizes into the product of its marginals $q_{i_1}(t), \cdots, q_{i_L}(t)$, i.e., $q$ is memoryless within a block of length $L$. Hence, maximizing $Q_n(t)$ with respect to $t$ is a reasonable objective if we are interested in finding the value of $t$ for which $Z(t)$ is "as memoryless as possible," namely, an estimate of $\theta$. However, since $Q_n(t)$ is a piecewise constant function of $t$ for any given sample, the asymptotic behavior of $\arg\max_t Q_n(t)$ cannot be analyzed by the usual methods. To this end, fix $\delta > 0$ and confine attention to the grid $\Theta_n = \{r\delta/\sqrt{n}:\ r = 0, \pm 1, \cdots, \pm\lfloor\sqrt{n}/\delta\rfloor\}$ in search of an estimate of $\theta$, where $\lfloor a \rfloor$ denotes the largest integer not exceeding $a$.

Let us now define our estimator $\hat{\theta}_n(k,L)$ as the point $r\delta/\sqrt{n}$ that locally maximizes $Q_n(t)$ in $\Theta_n$, i.e.,

$$Q_n(r\delta/\sqrt{n}) > \max\{Q_n((r-1)\delta/\sqrt{n}), Q_n((r+1)\delta/\sqrt{n})\}, \quad (2.8)$$

provided that such a point exists uniquely. Obviously, (2.8) can either have more than one solution or no solutions at all, but it can be shown (similar to [10, Theorems 1, 2]) that the probabilities of these events (no solution or many solutions) are asymptotically negligible. To make the estimator $\hat{\theta}_n(k,L)$ well defined, however, these situations must be considered as well. To this end, let us make the convention that in these cases $\hat{\theta}_n(k,L)$ will rely on some $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$, say, the least square estimator $\tilde{\theta}_n = \hat{\theta}_{LS} = (\sum_{i=1}^n x_i x_{i-1})/\sum_{i=1}^n x_{i-1}^2$ in the following manner: If (2.8) is satisfied nowhere, then $\hat{\theta}_n(k,L)$ is defined as the nearest neighbor $\bar{\theta}_n \in \Theta_n$ of $\tilde{\theta}_n$. Else, if (2.8) holds at more than one point in $\Theta_n$, then $\hat{\theta}_n(k,L)$ is defined as the one lying closest to $\bar{\theta}_n$.

As mentioned earlier, we would like to analyze the asymptotic distribution of the estimation error scaled by $\sqrt{n}$. Note, however, that once we restrict $\hat{\theta}_n(k,L)$ to $\Theta_n$ only, then the distribution of the scaled estimation error, $\sqrt{n}(\hat{\theta}_n(k,L) - \theta)$), does not converge in general, as $n \to \infty$. This follows from the following consideration: If $\theta$ is allowed to take any real value in $(-1,1)$, then it can be uniquely represented as $\theta = (c_n + r_0\delta)/\sqrt{n}$, where $r_0$ is an integer and $c_n = (\theta\sqrt{n}) \bmod (\delta)$ characterizes the position of $\theta$ relative to its left-hand side neighbor in $\Theta_n$. Hence, the scaled estimation error takes values of the form $(r - r_0)\delta + c_n$ ($r$ integer), namely, integer multiples of $\delta$ shifted by $0 \le c_n < \delta$. But $c_n$ does *not* converge in general as $n$ grows indefinitely, hence the scaled estimation error cannot have an asymptotic distribution for an arbitrary fixed $\theta$.

To alleviate this difficulty, we shall formalize the result with respect to an underlying parameter value $\theta_n$ that is the nearest neighbor of $\theta$ among all members in $\Theta_n$. By doing this, we guarantee that $|\theta_n - \theta| \le \delta/(2\sqrt{n})$, that is $\theta_n \to \theta$, and at the same time, $c_n = 0$ for an $n$, hence avoiding the aforementioned obstacle.

The main result is given in the following theorem.

*Theorem 1:* If conditions 1)–5) are met, then for any integer $s$,

$$\lim_{n\to\infty} P_{\theta_n}\{\sqrt{n}(\hat{\theta}_n(k,L) - \theta_n) \le s\delta\} = \Phi\left(\left(s + \frac{1}{2}\right)\delta \cdot \sqrt{I_k^L(f,\theta)}\right), \quad (2.9)$$

where $\Phi$ is the standard normal PDF, i.e.,

$$\Phi(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du, \quad (2.10)$$

and $I_k^L(f,\theta)$ is the Fisher information associated with $k$ levels of quantization and blocks of length $L$, i.e.,

$$I_k^L(f,\theta) \triangleq \frac{1}{L}\left\{\sum_{i=1}^{k+1} \frac{[f(a_i) - f(a_{i-1})]^2}{F(a_i) - F(a_{i-1})}\right\} \cdot \sum_{p=1}^{L} \sum_{i_1=1}^{k+1} \cdots \sum_{i_L=1}^{k+1} E_\theta^2(X_{p-1} | V_1 \in C_{i_1}, \cdots, V_L \in C_{i_L}) \cdot \Pr\{V_1 \in C_{i_1}, \cdots, V_L \in C_{i_L}\}, \quad (2.11)$$

$E_\theta(\cdot)$ being the mathematical expectation with respect to $\theta$.

The proof is given in Section IV.

The theorem states that the asymptotic distribution of the random variable $\sqrt{n}(\hat{\theta}_n(k,L) - \theta_n)$ is roughly normal with mean $-\delta/2$ and variance $1/I_k^L(f,\theta)$. It is easy to check (see also [10]) that $I_k^L(f,\theta) < I(f,\theta)$ but $I_k^L(f,\theta) \to I(f,\theta)$ as $L \to \infty$, and $k \to \infty$, provided that $a_1 \to -\infty$, $a_k \to \infty$ and $a_j - a_{j-1} \to 0$, $j = 2 \cdots k$, i.e., the high resolution limit. Hence, we observe that the parameters $k$ and $L$ control the asymptotic variance of the estimator while the parameter $\delta$ dictates the asymptotic bias. By choosing $k$ and $L$ sufficiently large, the asymptotic variance can be made arbitrarily close to the Cramér–Rao lower bound, and by selecting $\delta$ sufficiently small, the asymptotic bias can be reduced arbitrarily.

## III. Discussion

Note, that the application of $\hat{\theta}_n(k,L)$ requires computational resources and storage that grow exponentially with $L$. It is of interest, therefore, to observe the asymptotic behavior of $I_k^L(f,\theta)$ as $k \to \infty$ while $L$ is fixed. To this end, recall that the outer summation over $p$ in (2.11) consists of $L$ terms of the form

$$H_p^k \triangleq \sum_{i_1=1}^{k+1} \cdots \sum_{i_L=1}^{k+1} E_\theta^2(X_{p-1} | V_1 \in C_{i_1}, \cdots, V_L \in C_{i_L}) \cdot \Pr\{V_1 \in C_{i_1}, \cdots, V_L \in C_{i_L}\}. \quad (3.1)$$

As for the first term, $H_1^k$, observe that $X_{p-1} = X_0$ is independent of future noise samples $V_1,\cdots,V_L$. Thus,

$$E_\theta(X_0|V_1\in C_{i_1},\cdots,V_L\in C_{i_L})=E_\theta(X_0)=0, \quad (3.2)$$

where the last equality follows from condition 4). Hence, $H_1^k$ is identically zero for all $k$. Now, since

$$X_{p-1}=\theta^{p-1}X_0+\sum_{i=1}^{p-1}\theta^{i-1}V_{p-i}, \quad (3.3)$$

by (1.1), it is easy to see that for $p\geq 2$,

$$H_p\triangleq\lim_{k\to\infty}H_p^k=E\left[E_\theta^2(X_{p-1}|V_1^L)\right]=E\left[\left(\sum_{i=1}^{p-1}\theta^{i-1}V_{p-i}\right)^2\right]=\sigma^2\sum_{j=0}^{p-2}\theta^{2j}=\frac{\sigma^2(1-\theta^{2(p-1)})}{1-\theta^2}. \quad (3.4)$$

Next, the summation of (3.4) over $p$ results in

$$H\triangleq\sum_{p=1}^{L}H_p=\sum_{p=2}^{L}H_p=\frac{\sigma^2}{1-\theta^2}\sum_{p=2}^{L}(1-\theta^{2(p-1)})=\frac{\sigma^2}{1-\theta^2}\left(L-\frac{1-\theta^{2L}}{1-\theta^2}\right). \quad (3.5)$$

Finally, on substituting the asymptotic summation (3.5) into (2.11) and using the fact that

$$\sum_{i=1}^{k+1}[f(a_i)-f(a_{i-1})]^2/[F(a_i)-F(a_{i-1})]$$

tends to

$$\int_{-\infty}^{\infty}\{[f'(x)]^2/f(x)\}\,dx$$

in the high resolution limit, we obtain

$$I^L(f,\theta)\triangleq\lim_{k\to\infty}I_k^L(f,\theta)=\left(1-\frac{1}{L}\frac{1-\theta^{2L}}{1-\theta^2}\right)I(f,\theta). \quad (3.6)$$

Thus, $I^L(f,\theta)$ approaches $I(f,\theta)$ at the rate of $1/L$, essentially.

The idea of deriving nearly efficient estimators from universal coding algorithms can be used in several other parametric models as well, as we shall see in the following examples.

*The One-Sample Location Model:* Let $X_1,\cdots,X_n$ be i.i.d. random variables drawn from a PDF $F(x-\theta)$, where $\theta$ is an unknown parameter to be estimated and $F(x)$ has a symmetric density $f(x)$ about the origin. Namely, we are interested in estimating the center of symmetry of a density. Fix a constant $t$ and form quantized versions, (using a $k$-level symmetric quantizer), say, $Y(t)$ and $Z(t)$, of the sequences $X-t\triangleq(X_1-t,\cdots,X_n-t)$ and $t-X\triangleq(t-X_1,\cdots,t-X_n)$, respectively. Similarly to (2.5) and (2.6), let $H_Y(t)$, $H_Z(t)$ and $H_{YZ}(t)$ denote the entropies associated with the empirical distributions (relative frequencies of letters) of $Y(t)$, $Z(t)$, and the concatenation of $Y(t)$ and $Z(t)$, respectively. Nearly efficient estimation of $\theta$, in the sense of Theorem 1, is now achieved by maximizing (over a grid) the quantity

$$Q_n(t)=H_Y(t)+H_Z(t)-H_{YZ}(t). \quad (3.7)$$

Note, that $Q_n(t)\leq 0$ with equality iff $Y(t)$ and $Z(t)$ have the same empirical distribution, in which case encoding these two sequences jointly, using approximately $H_{YZ}(t)$ bits, is equivalent to encoding them separately, with nearly $H_Y(t)+H_Z(t)$ bits. By the symmetry of $f$, however, $Y(t)$ and $Z(t)$ have the same distribution iff $t=\theta$, hence maximizing (3.7) results in a reasonable estimate. A quantity similar to (3.7), serves as a test statistic of an asymptotically optimal test [11], [12] in the Neyman–Pearson sense, for deciding whether or not two given sequences, say, $Y$ and $Z$, have emerged from the same source.

*The Two-Sample Location Model:* Let $X_1,\cdots,X_n$ be i.i.d. random variables drawn from an unknown PDF $F(x)$ (not necessarily with symmetric density), and let $Y_1,\cdots,Y_m$ be another sequence of i.i.d. random variables, independent of $X_1,\cdots,X_n$ and governed by $F(x-\theta)$, a shifted version of $F$, but by an unknown amount $\theta$ to be estimated. An idea similar to (3.7) can be used to estimate $\theta$ efficiently: Subtract $t$ from one sequence, quantize, and adjust $t$ such that the two sequences have the same empirical distribution in a sense similar to (3.7). Alternatively, in [10] nearly efficient estimation has been achieved by minimizing (with respect to $t$) the informational divergence between these two empirical distributions.

*The Linear Regression Model:* The simplest linear regression model is as follows. Let $Y_i=\theta X_i+V_i$, $i=1,2,\cdots,n$, where $X_1,\cdots,X_n$ and $Y_1,\cdots,Y_n$ are given observation sequences, and $V_1,\cdots,V_n$ are i.i.d. random variables, independent of $X_1,\cdots,X_n$, and drawn from an unknown distribution $F$. In this case, a reasonable idea for estimating $\theta$ is to adjust $t$ such that the sequence $Y_i-tX_i$, $i=1,2,\cdots,n$, "looks" independent of $X_1,\cdots,X_n$. This can achieved by minimizing with respect to $t$ the empirical mutual information associated with quantized versions of these two sequences. Again, a similar idea has been used for universal optimal testing for independence in the Neyman–Pearson sense [9].

To prove that these estimators are asymptotically nearly efficient in the sense of Theorem 1, again, a technique similar to [10] can be used. In the next section we use this technique to prove Theorem 1 for the first order AR model.

## IV. Proof of Theorem 1

We shall adopt the following notation:

$$\alpha_j\triangleq\Pr\{V_1\in C_j\}=\int_{-\infty}^{\infty}\chi_j(v)f(v)\,dv,\qquad j=1,\cdots,k+1, \quad (4.1)$$

$$\beta_j(t)\triangleq P_{\theta_n}\{Z_1(t)\in C_j\}=\int_{-\infty}^{\infty}g_{\theta_n}(x)\,dx\int_{a_{j-1}+x(t-\theta_n)}^{a_j+x(t-\theta_n)}f(v)\,dv,\qquad j=1,\cdots k+1 \quad (4.2)$$

where $g_{\theta_n}(x)$ denotes the marginal density of $X_i$. (Note that $\beta_j(\theta_n)=\alpha_j$.) For every two integers $p$ and $q$ with $p \le q$, let

$$V_p^q \triangleq (V_p, V_{p+1}, \cdots, V_q), \tag{4.3}$$

and similarly,

$$Z_p^q(t) \triangleq (Z_p(t), \cdots, Z_q(t)). \tag{4.4}$$

Next, let

$$\boldsymbol{i} \triangleq (i_1, i_2, \cdots, i_L), \qquad \boldsymbol{j} \triangleq (j_1, j_2, \cdots, j_L), \tag{4.5}$$

$$N \triangleq \{\boldsymbol{i}: i_1, i_2, \cdots, i_L = 1, 2, \cdots, k+1\}, \tag{4.6}$$

$$C_{\boldsymbol{i}} \triangleq C_{i_1} \times C_{i_2} \times \cdots \times C_{i_L}, \tag{4.7}$$

where $\times$ denotes a Cartesian product between sets,

$$\chi_{\boldsymbol{i}}(V_{p+1}^{p+L}) \triangleq \prod_{j=1}^{L} \chi_{i_j}(V_{p+j}), \tag{4.8}$$

$$\chi_{\boldsymbol{i}}(Z_{p+1}^{p+L}(t)) \triangleq \prod_{j=1}^{L} \chi_{i_j}(Z_{p+j}(t)), \tag{4.9}$$

$$\alpha_{\boldsymbol{i}} \triangleq \prod_{j=1}^{L} \alpha_{i_j}, \tag{4.10}$$

$$\begin{aligned}\beta_{\boldsymbol{i}}(t) &\triangleq P_{\theta_n}\{Z_1^L(t) \in C_{\boldsymbol{i}}\} \\ &= \int_{-\infty}^{\infty} g_{\theta_n}(x_0)\, dx_0 \int_{x_1 = a_{i_1-1}+tx_0}^{a_{i_1}+tx_0} dx_1 \\ &\quad \cdot \int_{x_2 = a_{i_2-1}+tx_2}^{a_{i_2}+tx_1} dx_2 \cdots \int_{x_L = a_{i_L-1}+tx_{L-1}}^{a_{i_L}+tx_{L-1}} dx_L \\ &\quad \cdot \prod_{i=1}^{L} f(x_i - \theta_n x_{i-1}),\end{aligned} \tag{4.11}$$

and for any positive integer $d$, let

$$\beta_{\boldsymbol{ij}}^d(t_1,t_2) \triangleq P_\theta\{Z_1^L(t_1) \in C_{\boldsymbol{i}}, Z_{dL+1}^{dL+L}(t_2) \in C_{\boldsymbol{j}}\}. \tag{4.12}$$

The quantity $s\delta/\sqrt{n}$, $s$ integer, will be denoted by $\delta_n(s)$. The function

$$J_r(t) \triangleq \sum_{\boldsymbol{i} \in N} \chi_{\boldsymbol{i}}[Z_{Lr+1}^{Lr+L}(t)] \log \frac{\beta_{i_1}(t) \cdots \beta_{i_L}(t)}{\beta_{\boldsymbol{i}}(t)}, \qquad r = 0, 1, \cdots \frac{n}{L} - 1 \tag{4.13}$$

is closely related to $Q_n(t)$ as we shall see later. Finally, define

$$U_n^r(s) \triangleq J_r(\theta_n + \delta_n(s)), \tag{4.14}$$

$$U_n(s) \triangleq \sum_{r=0}^{n/L-1} U_n^r(s), \tag{4.15}$$

$$W_{1n}(s) \triangleq U_n(s) - U_n(s-1), \tag{4.16}$$

and

$$W_{2n}(s) \triangleq U_n(s) - U_n(s+1). \tag{4.17}$$

Our first goal is to calculate $\lim_{n\to\infty} P_{\theta_n}[W_{1n}(s) > 0, W_{2n}(s) > 0]$ and $\lim_{n\to\infty} P_{\theta_n}[W_{1n}(s) < 0, W_{2n}(s) < 0]$ for a fixed integer $s$. Then, it will be shown that the event $\{W_{1n}(s) > 0, W_{2n}(s) > 0\}$ is asymptotically equivalent to (2.8) for $r = s$. To compute these limiting probabilities, it is necessary to evaluate first $\beta_i(t)$, $\beta_{ij}^d(t_1,t_2)$, $\log \beta_i(t)$ and $\log \beta_{\boldsymbol{i}}(t)$ at $t, t_1, t_2 = \theta + \delta_n(r)$, $r = s-1$, $s$ and $s+1$, respectively, and some moments of $W_{1n}(s)$ and $W_{2n}(s)$. Hence, we need the asymptotic expansions of the previous functions about $t = \theta_n$. These are given in the following lemma.

*Lemma 1:* Let

$$h_i \triangleq f(a_i) - f(a_{i-1})$$

and

$$l_i \triangleq f'(a_i) - f'(a_{i-1}), \qquad i = 1, \cdots k+1.$$

Then, for any fixed $s$, as $n \to \infty$

a) $\beta_i(\theta_n + \delta_n(s)) = \alpha_i + \frac{1}{2}\delta_n^2(s)\frac{\sigma^2}{1-\theta_n^2} l_i + o(n^{-1})$,

b) $\log \beta_i(\theta_n + \delta_n(s)) = \log \alpha_i + \frac{1}{2}\delta_n^2(s)\frac{\sigma^2}{1-\theta_n^2}\frac{l_i}{\alpha_i} + o(n^{-1})$,

c) $\beta_{\boldsymbol{i}}(\theta_n + \delta_n(s)) = \alpha_{\boldsymbol{i}} + \delta_n(s) E_{\boldsymbol{i}} \alpha_{\boldsymbol{i}} + \frac{1}{2}\delta_n^2(s) S_{\boldsymbol{i}} \alpha_{\boldsymbol{i}} + o(n^{-1})$,

where

$$E_{\boldsymbol{i}} \triangleq \sum_{p=1}^{L} \frac{h_{i_p}}{\alpha_{i_p}} E_{\theta_n}(X_{p-1} | V_1^L \in C_{\boldsymbol{i}}), \tag{4.18}$$

and

$$\begin{aligned} S_{\boldsymbol{i}} \triangleq & \sum_{p=1}^{L} \frac{l_{i_p}}{\alpha_{i_p}} E_{\theta_n}(X_{p-1}^2 | V_1^L \in C_{\boldsymbol{i}}) \\ & + 2\frac{\partial}{\partial \theta} \sum_{p=1}^{L} \frac{h_{i_p}}{\alpha_{i_p}} E_\theta(X_{p-1} | V_1^L \in C_{\boldsymbol{i}})|_{\theta=\theta_n} \\ & + 2\sum_{p=1}^{L-1} \sum_{q=p+1}^{L} \theta_n^{q-p} \frac{h_{i_p} h_{i_q}}{\alpha_{i_p}\alpha_{i_p}} E_{\theta_n}(X_{p-1}^2 | V_1^L \in C_{\boldsymbol{i}}) \\ & + 2\sum_{p=1}^{L-1} \sum_{q=p+1}^{L} \theta_n^{q-p-1} \frac{h_{i_p}}{\alpha_{i_p}\alpha_{i_q}} E_{\theta_n}(X_{p-1} | V_1^L \in C_{\boldsymbol{i}}) \\ & + 2\sum_{p=1}^{L-1} \sum_{q=p+1}^{L} \frac{h_{i_p} h_{i_q}}{\alpha_{i_p}\alpha_{i_q}} \\ & \cdot E_{\theta_n}\left(X_{p-1} \sum_{j=p+1}^{q-1} V_j \theta_n^{q-j-1} | V_1^L \in C_{\boldsymbol{i}}\right), \end{aligned} \tag{4.19}$$

d) $\log \beta_{\boldsymbol{i}}(\theta_n + \delta_n(s)) = \log \alpha_{\boldsymbol{i}} + \delta_n(s) E_{\boldsymbol{i}} + \frac{1}{2}\delta_n^2(s)(S_{\boldsymbol{i}} - E_{\boldsymbol{i}}^2) + o(n^{-1})$,

e) $\beta_{\boldsymbol{ij}}^d(\theta_n + \delta_n(s_1), \theta_n + \delta_n(s_2)) = \alpha_{\boldsymbol{i}}\alpha_{\boldsymbol{j}} + \delta_n(s_1) E_{\boldsymbol{i}} \alpha_{\boldsymbol{i}}\alpha_{\boldsymbol{j}} + \delta_n(s_2) E_{\boldsymbol{ij}}^d \alpha_{\boldsymbol{i}}\alpha_{\boldsymbol{j}} + \frac{1}{2}\delta_n^2(s_1) S_{\boldsymbol{i}} \alpha_{\boldsymbol{i}}\alpha_{\boldsymbol{j}} + \delta_n(s_1)\delta_n(s_2) M_{\boldsymbol{ij}}^d \alpha_{\boldsymbol{i}}\alpha_{\boldsymbol{j}} + \frac{1}{2}\delta_n^2(s_2) S_{\boldsymbol{ij}}^d \alpha_{\boldsymbol{i}}\alpha_{\boldsymbol{j}} + o(n^{-1})$, where

$$E_{ij}^d \triangleq \sum_{p=1}^{L} \frac{h_{j_p}}{\alpha_{j_p}} E_{\theta_n} \cdot \left( X_{dL+p-1} | V_1^L \in C_i, V_{dL+1}^{dL+L} \in C_j \right), \quad (4.20)$$

$$M_{ij}^d \triangleq \sum_{p=1}^{L} \sum_{q=dL+1}^{dL+L} \frac{h_{i_p} h_{j_q}}{\alpha_{i_p} \alpha_{j_q}} E_{\theta_n} \cdot \left( X_{p-1} X_{q-1} | V_1^L \in C_i, V_{dL+1}^{dL+L} \in C_j \right), \quad (4.21)$$

and

$$\begin{aligned} S_{ij}^d \triangleq & \sum_{p=1}^{L} \frac{l_{j_p}}{\alpha_{j_p}} E_{\theta_n}\left( X_{dL+p-1}^2 | V_1^L \in C_i, V_{dL+1}^{dL+L} \in C_j \right) \\ & + 2 \frac{\partial}{\partial \theta} \sum_{p=1}^{L} \frac{h_{j_p}}{\alpha_{j_p}} E_{\theta} \cdot \left( X_{dL+p-1} | V_1^L \in C_i, V_{dL+1}^{dL+L} \in C_j \right)|_{\theta = \theta_n} \\ & + 2 \sum_{p=1}^{L-1} \sum_{q=1}^{L} \theta_n^{q-p} \frac{h_{j_p} h_{j_q}}{\alpha_{j_p} \alpha_{j_p}} E_{\theta_n} \cdot \left( X_{dL+p-1}^2 | V_1^L \in C_i, V_{dL+1}^{dL+L} \in C_j \right) \\ & + 2 \sum_{p=1}^{L-1} \sum_{q=p+1}^{L} \theta_n^{q-p-1} \frac{h_{j_p}}{\alpha_{j_p} \alpha_{j_q}} E_{\theta_n} \cdot \left( X_{dL+p-1} | V_1^L \in C_i, V_{dL+1}^{dL+L} \in C_j \right) \\ & + 2 \sum_{p=1}^{L-1} \sum_{q=p+1}^{L} \frac{h_{j_p} h_{j_q}}{\alpha_{j_p} \alpha_{j_q}} \cdot E_{\theta_n} \left( X_{dL+p-1} \sum_{r=dL+p+1}^{dL+q-1} V_r \theta_n^{dL+q-r-1} | V_1^L \in C_i, \cdot V_{dL+1}^{dL+L} \in C_j \right). \end{aligned} \quad (4.22)$$

*Proof of Lemma 1:* To prove a) and b), differentiate (4.2) twice with respect to $t$ at $t = \theta_n$ and use the Taylor expansion. Parts c) and d) can be checked by changing the integration variables of (4.11) using the following transformation:

$$x_0 = z_0, \qquad x_1 = z_1 + z_0 t, \cdots x_L = \sum_{j=0}^{k} z_j t^{k-j}.$$

By doing this, (4.11) can be rewritten as follows:

$$\beta_i(t) = \int_{-\infty}^{\infty} g_{\theta_n}(z_0)\, dz_0 \cdot \int_{C_i} \prod_{k=1}^{L} f\left( z_k + (t - \theta_n) \sum_{j=0}^{k-1} z_j t^{k-j-1} \right) dz_1^L. \quad (4.23)$$

Again, by differentiating (4.23) twice with respect to $t$ at $t = \theta_n$, the Taylor expansions c) and d) are obtained. Part e) is verified similarly by calculating the first- and second-order derivatives $(\partial/\partial t)\beta_{ij}^d(t, \theta_n)|_{t=\theta_n}$, $(\partial/\partial t)\beta_{ij}^d(\theta_n, t)|_{t=\theta_n}$, $(\partial^2/\partial t^2)\beta_{ij}^d(t, \theta_n)|_{t=\theta_n}$, $(\partial^2/\partial t\, \partial \tau)\beta_{ij}^d(t, \tau)|_{t=\tau=\theta_n}$ and $(\partial^2/\partial t^2)\beta_{ij}^d(\theta_n, t)|_{t=\theta_n}$, and substituting these in the bivariate Taylor expansion of $\beta_{ij}^d(t_1, t_2)$ about $t_1 = t_2 = \theta_n$. □

To calculate the required moments of $W_{1n}(s)$ and $W_{2n}(s)$ we first establish an auxiliary result.

*Lemma 2:* Under conditions 1) and 3),

$$\lim_{n \to \infty} \left\{ \frac{\sqrt{n}}{\delta} \left[ U_n^r(s) - U_n^r(s-1) \right] - \dot{J}_r(\theta_n) \right\} = 0, \qquad \text{a.s.},$$

where

$$\dot{J}_r(\theta_n) \triangleq - \sum_{i \in N} \chi_i \left( V_{Lr+1}^{Lr+L} \right) E_i.$$

*Proof of Lemma 2:* Since $J_r(t)$ is differentiable almost everywhere, it is also differentiable at $t = \theta_n$ with probability 1, and the almost sure derivative is given by $\dot{J}_r(\theta_n) \triangleq -\sum_{i \in N} \chi_i(V_{Lr+1}^{Lr+L}) E_i$. Hence,

$$\begin{aligned} & \lim_{n \to \infty} \left\{ \frac{\sqrt{n}}{\delta} \left[ U_n^r(s) - U_n^r(s-1) \right] - \dot{J}_r(\theta_n) \right\} \\ & = \lim_{n \to \infty} \left\{ \frac{J_r(\theta_n + \delta_n(s)) - J_r(\theta_n + \delta_n(s-1))}{\delta / \sqrt{n}} - \dot{J}_r(\theta_n) \right\} \\ & = 0, \end{aligned} \quad (4.24)$$

almost surely by the definition of a derivative. □

The asymptotic moments of $W_{1n}(s)$ and $W_{2n}(s)$ are given in Lemma 3.

*Lemma 3:* For any fixed $s$, as $n \to \infty$,

a) $E[W_{1n}(s)] = -\delta^2(s - \frac{1}{2}) I_k^L(f, \theta) + o(1)$.
b) $E[W_{2n}(s) = \delta^2(s + \frac{1}{2}) I_k^L(f, \theta) + o(1)$.
c) $\mathrm{var}[W_{1n}(s)] = \delta^2 I_k^L(f, \theta) + o(1)$.
d) $\mathrm{var}[W_{2n}(s)] = \delta^2 I_k^L(f, \theta) + o(1)$.
e) $\mathrm{cov}[W_{1n}(s), W_{2n}(s)] = -\delta^2 I_k^L(f, \theta) + o(1)$.

The proof of Lemma 3 appears in the Appendix. Lemma 3 implies immediately by the Chebychev inequality that

$$\lim_{n \to \infty} \left[ W_{1n}(s) + W_{2n}(s) \right] = \delta^2 I_k^L(f, \theta) \quad (4.25)$$

in probability. The next lemma introduces the asymptotic distribution of $W_{1n}(s)$ for a fixed integer $s$.

*Lemma 4:* For fixed $s$, $W_{1n}(s)$ is asymptotically normally distributed with mean $-\delta^2(s - 1/2) I_k^L(f, \theta)$ and variance $\delta^2 I_k^L(f, \theta)$.

*Proof of Lemma 4:* First, observe that

$$W_{1n}(s) = \sum_{r=0}^{n/L-1} [U_n^r(s) - U_n^r(s-1)]$$

$$= \frac{\delta}{\sqrt{n}} \sum_{r=0}^{n/L-1} \frac{J_r(\theta_n + \delta_n(s)) - J_r(\theta_n + \delta_n(s-1))}{\delta/\sqrt{n}}$$

$$= \frac{\delta}{\sqrt{n}} \sum_{r=0}^{n/L-1} \dot{J}_r(\theta_n)$$

$$+ \frac{\delta}{\sqrt{n}} \sum_{r=0}^{n/L-1} \left[ \frac{J_r(\theta_n + \delta_n(s)) - J_r(\theta_n + \delta_n(s-1))}{\delta/\sqrt{n}} - \dot{J}_r(\theta_n) \right]. \tag{4.26}$$

The first term of the last expression is a sum of i.i.d. zero-mean random variables with variances $\delta^2(L/n)I_k^L(f,\theta_n)$. Thus, by the central limit theorem its asymptotic distribution is $N(0,\delta^2 I_k^L(f,\theta_n))$. It remains to show that the second term tends to $-\delta^2(s-\frac{1}{2})I_k^L(f,\theta)$ in probability. To prove that, we rewrite the summand of the second term as

$$\frac{J_r(\theta_n + \delta_n(s)) - J_r(\theta_n + \delta_n(s-1))}{\delta/\sqrt{n}} - \dot{J}_r(\theta_n)$$

$$= s\left[ \frac{J_r(\theta_n + \delta_n(s)) - J_r(\theta_n)}{\delta_n(s)} - \dot{J}_r(\theta_n) \right]$$

$$-(s-1)\left[ \frac{J_r(\theta_n + \delta_n(s-1)) - J_r(\theta_n)}{\delta_n(s-1)} - \dot{J}_r(\theta_n) \right]. \tag{4.27}$$

Since $f$ and $f'$ are continuous, it follows from the stochastic Taylor expansion of $J_r(t)$ that

$$\frac{J_r(\theta_n + \delta_n(s)) - J_r(\theta_n)}{\delta_n(s)} - \dot{J}_r(\theta_n)$$

$$= \frac{1}{2}\ddot{J}_r(\theta_n)\delta_n(s) + o_p(n^{-1/2}), \tag{4.28}$$

where

$$\ddot{J}_r(\theta_n) \triangleq -\sum_{i \in N} \chi_i(V_{Lr+1}^{Lr+L}) \left( S_i - E_i^2 - \frac{\sigma^2}{1-\theta_n^2} \sum_{k=1}^{L} \frac{l_{i_k}}{\alpha_{i_k}} \right) \tag{4.29}$$

is the almost sure second derivative of $J_r(t)$ at $t=\theta_n$. We therefore obtain by (4.27) and (4.28),

$$\frac{\delta}{\sqrt{n}} \sum_{r=0}^{n/L-1} \left[ \frac{J_r(\theta_n + \delta_n(s) - J_r(\theta_n + \delta_n(s-1))}{\delta/\sqrt{n}} - \dot{J}_r(\theta_n) \right]$$

$$= -\frac{\delta}{\sqrt{n}} \sum_{r=0}^{n/L-1} \left[ \frac{s}{2}\ddot{J}_r(\theta_n)\delta_n(s) - \frac{s-1}{2}\ddot{J}_r(\theta_n) \cdot \delta_n(s-1) + o_p(n^{-1/2}) \right]$$

$$= -\frac{1}{n} \sum_{r=0}^{n/L-1} \frac{1}{2}\left[ (s\delta)^2 - ((s-1)\delta)^2 \right] \ddot{J}_r(\theta_n) + o_p(1)$$

$$\xrightarrow[n\to\infty]{P} -\frac{\delta^2}{L}\left(s-\frac{1}{2}\right) E[\ddot{J}_r(\theta)]$$

$$= -\frac{\delta^2}{L}\left(s-\frac{1}{2}\right) \lim_{n\to\infty} \sum_{i\in N} \alpha_i \left( S_i - E_i^2 - \frac{\sigma^2}{1-\theta_n^2} \sum_{k=1}^{L} \frac{l_{ik}}{\alpha_{ik}} \right)$$

$$= -\delta^2\left(s-\frac{1}{2}\right) I_k^L(f,\theta). \tag{4.30}$$

Equation (4.25) together with Lemma 4 suggests that to calculate the asymptotic probabilities of the events $\{W_{1n}(s)>0,\ W_{2n}(s)>0\}$ and $\{W_{1n}(s)<0,\ W_{2n}(s)<0\}$, we may substitute $W_{2n}(s)$ by $\delta^2 I_k^L(f,\theta) - W_{1n}(s)$, and then use the asymptotic distribution of $W_{1n}(s)$ as it was derived in Lemma 4. The following lemma [10, Lemma 7] enables us to use that procedure.

*Lemma 5 (Bhattacharya):* Let $\{A_n\}$ and $\{B_n\}$ be sequences of random variables such that the distribution of $\{A_n\}$ converges to a distribution function $F$ at all points of continuity of $F$, and $A_n + B_n$ converges in probability to a constant $\mu$. If $a+b<\mu$, and if $a$ and $\mu - b$ are points of continuity of $F$, then

$$\lim_{n\to\infty} P[A_n \le a,\quad B_n \le b] = 0, \tag{4.31a}$$

and

$$\lim_{n\to\infty} P[A_n > a,\quad B_n > b] = F(\mu - b) - F(a). \tag{4.31b}$$

The proof is given in [10].

From Lemmas 4 and 5 we now conclude that for any two constants $a_1$ and $a_2$ whose sum is less than $\delta^2 I_k^L(\mathrm{if},\theta)$, we have

$$\lim_{n\to\infty} P_{\theta_n}[W_{1n}(s) \le a_1,\quad W_{2n}(s) \le a_2] = 0, \tag{4.32}$$

and

$$\lim_{n\to\infty} P_{\theta_n}[W_{1n}(s) > a_1,\quad W_{2n}(s) > a_2]$$

$$= \Phi\left( \left[ \delta^2 I_k^L(f,\theta) - a_2 + \delta^2\left(s-\frac{1}{2}\right) I_k^L(f,\theta) \right] \Big/ \delta\sqrt{I_k^L(f,\theta)} \right)$$

$$- \Phi\left( \left[ a_1 + \delta^2\left(s-\frac{1}{2}\right) I_k^L(f,\theta) \right] \Big/ \delta\sqrt{I_k^L(f,\theta)} \right). \tag{4.33}$$

Next, by the stochastic Taylor expansion of $Q_n(t)$ at $t=\theta_n+\delta_n(r)$, $r=s-1,s,s+1$, we note that for any fixed $s$ and as $n\to\infty$,

$$Q_n(\theta_n+\delta_n(s))-Q_n(\theta_n+\delta_n(s-1))=W_{1n}(s)+o_p(1), \quad (4.34)$$

and

$$Q_n(\theta_n+\delta_n(s))-Q_n(\theta_n+\delta_n(s+1))=W_{2n}(s)+o_p(1). \quad (4.35)$$

The last auxiliary lemma states that the situation of multiple solutions to (2.8), in a fixed interval of integers $r$, becomes rare as $n$ grows.

*Lemma 6:* For any $K>0$, the probability that (2.8) holds for more than one point in the interval $(\theta_n-K/\sqrt{n},\theta_n-K/\sqrt{n})$ tends to zero as $n\to\infty$.

The proof of Lemma 6 is a straightforward extension of [10, Theorem 1].

We are now ready to prove the main result. Recall that $\hat{\theta}_n(k,L)$ is defined as follows. Let $\tilde{\theta}_n$ be a $\sqrt{n}$-consistent estimator and let $\bar{s}\triangleq\arg\min_s|\theta_n+\delta_n(s)-\tilde{\theta}_n|$. Then,

if (2.8) holds for a unique member $r\delta/\sqrt{n}$ of $\Theta_n$, then $\hat{\theta}_n(k,L)=r\delta/\sqrt{n}$;
if (2.8) holds nowhere, then $\hat{\theta}_n(k,L)=\theta_n+\delta_n(\bar{s})$;
if (2.8) is satisfied at more than one point, then $\hat{\theta}_n(k,L)$ is the one lying close to $\theta+\delta_n(\bar{s})$.

Let $A_n(s)$ be the set of all samples of size $n$ for which (2.8) is satisfied at $t=\theta_n+\delta_n(s)$, and let $B_n(s)$ denote the set of all samples of size $n$ for which $\hat{\theta}_n(k,L)=\theta_n+\delta_n(s)$. Then,

$$P_{\theta_n}\{\sqrt{n}\,(\hat{\theta}_n(k,L)-\theta_n)\le s\delta\}=P_{\theta_n}\Big\{\bigcup_{r\le s}B_n(r)\Big\}, \quad (4.36)$$

and we want to prove that this probability tends to $\Phi((s+1/2)\delta\sqrt{I_k^L(f,\theta)})$ as $n\to\infty$. We first note the following facts about $A_n(s)$ and $B_n(s)$:

a) For any given $s$,

$$\lim_{n\to\infty}P_{\theta_n}[A_n(s)]=\Phi\left(\left(s+\frac{1}{2}\right)\delta\sqrt{I_k^L(f,\theta)}\right)-\Phi\left(\left(s-\frac{1}{2}\right)\delta\sqrt{I_k^L(f,\theta)}\right),$$

b) $\lim_{n\to\infty}P_{\theta_n}[\bigcup_s A_n(s)]=1$,

c) For any $s_1\ne s_2$, $B_n(s_1)\cup B_n(s_2)=\phi$,

d) The union $\bigcup_{-\infty<s<\infty}B_n(s)$ is the entire sample space of size $n$.

To prove a), note from (4.34) and (4.35) that $\lim_{n\to\infty}P_{\theta_n}[A_n(s)]$ lies between $\lim_{n\to\infty}P_{\theta_n}[W_{1n}(s)\ge\epsilon, W_{2n}(s)\ge\epsilon]$ and $\lim_{n\to\infty}P_{\theta_n}[W_{1n}(s)\ge-\epsilon, W_{2n}(s)\ge-\epsilon]$ for any $\epsilon>0$. Choosing $\epsilon<\delta^2 I_k^L(f\theta)/2$, using (4.33), once for $a_1=a_2=\epsilon$, then for $a_1=a_2=-\epsilon$, and finally letting $\epsilon\to 0$, yields the result. Part b) follows easily from part a) and from Lemma 6, and parts c) and d) simply follow from the fact that $\hat{\theta}_n(k,L)$ is well defined.

Next, from the consistency of $\tilde{\theta}_n$, for any positive integer $S$ we have

$$P_{\theta_n}\{\theta_n+\delta_n(-S)\le\tilde{\theta}_n\le\theta_n+\delta_n(S)\}\ge 1-\epsilon(S), \quad (4.37)$$

where $\epsilon(S)\to 0$ as $S\to\infty$. Denote by $C_n(S)$ the set of all samples of size $n$ for which $-S\le\bar{s}\le S$. Let $D_n(s)$ be the set of samples for which $\bar{s}=s$. Then, by definition of $\hat{\theta}_n(k,L)$ we have for any $s'\ne s$ with $|s'|,|s|\le S$,

$$D_n(s')\cap A_n(s)\supseteq D_n(s')\cap B_n(s)\supseteq D_n(s')\cap A_n(s)\cap\left[\bigcap_{r\in M_s}A_n^c(r)\right], \quad (4.38)$$

where $M_s\triangleq\{r\colon -3S\le r\le 3S,\ r\ne s\}$ and the superscript $c$ denotes the complementary set. Similarly, for all $s$

$$B_n(s)\cap D_n(s)=D_n(s)\cap\left[\Big\{\bigcap_r A_n^c(r)\Big\}\cup A_n(s)\right]. \quad (4.39)$$

Hence, by (4.38) and Lemma 6, for any $\epsilon'>0$, $s\ne s'$ and $n$ sufficiently large

$$P_{\theta_n}[D_n(s')\cap A_n(s)]-\epsilon'\le P_{\theta_n}[D_n(s')\cap B_n(s)]\le P_{\theta_n}[D_n(s')\cap A_n(s)]. \quad (4.40)$$

Similarly, by (4.39) and by b), we get

$$P_{\theta_n}[D_n(s)\cap A_n(s)]\ge P_{\theta_n}[D_n(s)\cap B_n(s)]\ge P_{\theta_n}[D_n(s)\cap A_n(s)]-\epsilon'. \quad (4.41)$$

Thus, (4.40) holds either if $s=s'$ or $s\ne s'$. Next, since $D_n(s_1)\cap D_n(s_2)=\phi$ for any $s_1\ne s_2$ we have

$$|P_{\theta_n}[C_n(S)\cap B_n(s)]-P_{\theta_n}[C_n(S)\cap A_n(s)]|\le\sum_{|s'|\le S}|P_{\theta_n}[D_n(s')\cap B_n(s)]-P_{\theta_n}[D_n(s')\cap A_n(s)]|\le(2S+1)\epsilon'. \quad (4.42)$$

As $P_{\theta_n}[C_n(S)]\ge 1-\epsilon(S-1)$ by (4.37), we also have

$$|P_{\theta_n}[C_n(S)\cap A_n(s)]-P_{\theta_n}[A_n(s)]|\le\epsilon(S-1), \quad (4.43)$$

and

$$|P_{\theta_n}[C_n(S)\cap B_n(s)]-P_{\theta_n}[B_n(s)]|\le\epsilon(S-1), \quad (4.44)$$

implying that

$$|P_{\theta_n}[A_n(s)]-P_{\theta_n}[B_n(s)]|\le 2\epsilon(S-1)+(2S+1)\epsilon'. \quad (4.45)$$

Now fix $\delta>0$ and let $S$ be so large that $\epsilon(S-1)\le\delta/4$. Next, choose $\epsilon'>0$ sufficiently small such that $(2S+1)\epsilon'\le\delta/2$. It follows from (4.45) and fact a) that

$$\lim_{n\to\infty}P_{\theta_n}[B_n(s)]=\Phi\left(\left(s+\tfrac{1}{2}\right)\delta\sqrt{I_k^L(f,\theta)}\right)-\Phi\left(\left(s-\tfrac{1}{2}\right)\delta\sqrt{I_k^L(f,\theta)}\right). \quad (4.46)$$

To complete the proof, note that

$$\lim_{n\to\infty} P_{\theta_n}\left[\sqrt{n}\left(\hat{\theta}_n(k,L)-\theta_n\right)\le s\delta\right] = \lim_{n\to\infty} P_{\theta_n}\left[\bigcup_{r\le s} B_n(r)\right] = \lim_{n\to\infty}\sum_{r\le s} P_{\theta_n}[B_n(r)]. \quad (4.47)$$

For an arbitrary $\epsilon>0$ let us choose $s_1$ and $s_2$ such that

$$\Phi\left(\left(s_1-\frac{1}{2}\right)\delta\sqrt{I_k^L(f,\theta)}\right)\le\epsilon \quad (4.48)$$

and

$$\Phi\left(\left(s_2+\frac{1}{2}\right)\delta\sqrt{I_k^L(f,\theta)}\right)\le 1-\epsilon. \quad (4.49)$$

If $\epsilon$ is sufficiently small then $s_1<s<s_2$, and hence

$$\begin{aligned}\lim_{n\to\infty}\sum_{r\le s} P_{\theta_n}[B_n(r)] &\ge \lim_{n\to\infty}\sum_{r=s_1}^{s} P_{\theta_n}[B_n(r)]\\ &= \sum_{r=s_1}^{s}\lim_{n\to\infty} P_{\theta_n}[B_n(r)]\\ &= \Phi\left((s+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right) - \Phi\left((s_1-\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)\\ &\ge \Phi\left((s+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)-\epsilon. \quad (4.50)\end{aligned}$$

On the other hand,

$$\begin{aligned}\lim_{n\to\infty}\sum_{r\le s} P_{\theta_n}[B_n(r)] &= \lim_{n\to\infty}\sum_{r\le s_2} P_{\theta_n}[B_n(r)] - \sum_{r=s+1}^{s_2}\lim_{n\to\infty} P_{\theta_n}[B_n(r)]\\ &= \lim_{n\to\infty} P_{\theta_n}\left[\bigcup_{r\le s_2} B_n(r)\right] - \Phi\left((s_2+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)\\ &\quad + \Phi\left((s+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)\\ &\le 1-\Phi\left((s_2+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)\\ &\quad + \Phi\left((s+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)\\ &\le \Phi\left((s+\tfrac{1}{2})\delta\sqrt{I_k^L(f,\theta)}\right)+\epsilon. \quad (4.51)\end{aligned}$$

Since $\epsilon>0$ is arbitrary in (4.50) and (4.51), the theorem is proved.

## Acknowledgment

Useful comments made by the anonymous referees are greatly appreciated.

## Appendix
## Proof of Lemma 3

As for parts a) and b), we first need to calculate $E[U_n^r(s)]$ for a fixed $s$:

$$\begin{aligned}E[U_n^r(s)] &= \sum_{i\in N}\beta_i(\theta_n+\delta_n(s))\log\frac{\prod_{k=1}^{L}\beta_{i_k}(\theta_n+\delta_n(s))}{\beta_i(\theta_n+\delta_n(s))}\\ &= L\sum_{i=1}^{k+1}\beta_i(\theta_n+\delta_n(s))\log\beta_i(\theta_n+\delta_n(s))\\ &\quad -\sum_{i\in N}\beta_i(\theta_n+\delta_n(s))\log\beta_i(\theta_n+\delta_n(s)). \quad (A.1)\end{aligned}$$

Now by the asymptotic Taylor expansions of Lemma 1, and the facts that

$$\sum_{i=1}^{k+1} h_i = \sum_{i=1}^{k+1} l_i = 0, \quad (A.2)$$

and consequently,

$$\sum_{i\in N} E_i\alpha_i(1+\log\alpha_i)=0, \quad (A.3)$$

$$\sum_{i\in N} S_i\alpha_i = 0, \quad (A.4)$$

$$\sum_{i\in N} S_i\alpha_i\log\alpha_i = L\sigma^2(1-\theta^2)^{-1}\sum_{i=1}^{k+1} l_i\log\alpha_i, \quad (A.5)$$

and

$$\sum_{i\in N} E_i^2\alpha_i = LI_k^L(f,\theta_n)\to LI_k^L(f,\theta), \quad (A.6)$$

parts a) and b) are easily verified. To prove c), expand $\mathrm{var}[W_{1n}(s)]$ as follows:

$$\begin{aligned}\mathrm{var}[W_{1n}(s)] &= \frac{n}{L}\mathrm{var}[U_n^r(s)-U_n^r(s-1)]\\ &\quad +2\sum_{d=1}^{n/L-1}\left(\frac{n}{L}-d\right)\mathrm{cov}\left[U_n^r(s)-U_n^r(s-1),\right.\\ &\qquad \left. U_n^{r+d}(s)-U_n^{r+d}(s-1)\right]. \quad (A.7)\end{aligned}$$

It can be shown that the sequence of random variables $\{(\sqrt{n}/\delta)[U_n^r(s)-U_n^r(s-1)]\}_{n\ge1}$ is uniformly bounded with probability 1, and therefore from Lemma 2 we can calculate the first term on the right-hand side of (A.7) using the dominated convergence theorem [13, page 257, Theorem 6.5.5],

$$\begin{aligned}\lim_{n\to\infty}\frac{n}{L}\mathrm{var}[U_n^r(s)-U_n^r(s-1)] &= \frac{\delta^2}{L}\lim_{n\to\infty}E\left[\frac{U_n^r(s)-U_n^r(s-1)}{\delta/\sqrt{n}}\right]^2\\ &= \frac{\delta^2}{L}E\left[\lim_{n\to\infty}\frac{U_n^r(s)-U_n^r(s-1)}{\delta/\sqrt{n}}\right]^2\\ &= \frac{\delta^2}{L}E\left[\lim_{n\to\infty}\dot{J}_r(\theta_n)\right]^2\\ &= \frac{\delta^2}{L}\lim_{n\to\infty}\sum_{i\in N}E_i^2\alpha_i = \delta^2 I_k^L(f,\theta). \quad (A.8)\end{aligned}$$

It remains to show that the second term on the right-hand side of (A.7) vanishes as $n \to \infty$. To do this, it is sufficient to demonstrate that for every fixed $s_1$ and $s_2$,

$$\lim_{n\to\infty} \sum_{d=1}^{n/L-1} \operatorname{cov}\left[U_n^r(s_1), U_n^{r+d}(s_2)\right] = 0. \tag{A.9}$$

We first evaluate

$$\begin{aligned}
&\operatorname{cov}\left[U_n^r(s_1), U_n^{r+d}(s_2)\right]\\
&= \sum_{i\in N}\sum_{j\in N}\left[\beta_{ij}^d(\theta_n+\delta_n(s_1), \theta_n+\delta_n(s_2)) - \beta_i(\theta_n+\delta_n(s_1))\beta_j(\theta_n+\delta_n(s_2))\right]\\
&\cdot \log\left[\frac{\prod_{k=1}^{L}\beta_{i_k}(\theta_n+\delta_n(s_1))}{\beta_i(\theta_n+\delta_n(s_1))}\right]\log\left[\frac{\prod_{k=1}^{L}\beta_{j_k}(\theta_n+\delta_n(s_2))}{\beta_j(\theta_n+\delta_n(s_2))}\right]\\
&= \sum_{i\in N}\sum_{j\in N}\Big[\delta_n(s_2)\left(E_{ij}^d - E_j\right)\alpha_i\alpha_j + \delta_n(s_1)\delta_n(s_2)\left(M_{ij}^d - E_iE_j\right)\alpha_i\alpha_j\\
&\quad + \frac{1}{2}\delta_n^2(s_2)\left(S_{ij}^d - S_j\right)\alpha_i\alpha_j + o(n^{-1})\Big]\\
&\cdot\left[-\delta_n(s_1)E_i + o(n^{-1/2})\right]\\
&\cdot\left[-\delta_n(s_2)E_j + o(n^{-1/2})\right].
\end{aligned} \tag{A.10}$$

To examine the behavior of $(E_{ij}^d - E_j)$ as $d$ increases, note that

$$\begin{aligned}
&E_{\theta_n}\left(X_{dL+p-1} | V_1^L \in C_i, V_{dL+1}^{dL+L}\in C_j\right)\\
&= E_{\theta_n}\Bigg(X_0\theta_n^{dL+p-1} + \sum_{m=1}^{L} V_m\theta_n^{dL+p-1-m} + \sum_{m=L+1}^{dL} V_m\theta_n^{dL+p-1-m}\\
&\quad + \sum_{m=dL+1}^{dL+p-1} V_m\theta_n^{dL+p-1-m} | V_1^L\in C_i, V_{dL+1}^{dL+L}\in C_j\Bigg)\\
&= 0 + \sum_{m=1}^{L} E\left(V_m | V_m \in C_{i_m}\right)\theta_n^{dL+p-1-m} + 0 + \sum_{m=dL+1}^{dL+p-1} E\left(V_m | V_m\in C_{j_m}\right)\theta_n^{dL+p-1-m}\\
&= \sum_{m=1}^{L} E\left(V_m | V_m\in C_{i_m}\right)\theta_n^{dL+p-1-m} + E_{\theta_n}\left(X_{p-1} | V_{dL+1}^{dL+L}\in C_j\right).
\end{aligned} \tag{A.11}$$

As $|E(V_1|V_1\in C_i)|$ is bounded by, say, $E_{\max} < \infty$ under assumption in Condition 2), it follows from (A.11) that

$$\begin{aligned}
&|E_{\theta_n}\left(X_{dL+p-1}|V_1^L\in C_i, V_{dL+1}^{dL+L}\in C_j\right) - E_{\theta_n}\left(X_{dL+p-1}|V_{dL+1}^{dL+L}\in C_j\right)|\\
&= \left|\sum_{m=1}^{L} E\left(V_m|V_m\in C_{i_m}\right)\theta_n^{dL+p-1-m}\right|\\
&\le L\cdot E_{\max}(|\theta| + \epsilon)^{(d-1)L},
\end{aligned} \tag{A.12}$$

for any $\epsilon > 0$ and $n$ sufficiently large. Consequently, since $|h_j/\alpha_j|$, $j = 1, \cdots k+1$, are all bounded by some $\lambda < \infty$, then for all large $n$,

$$\begin{aligned}
|E_{ij}^d - E_i| &= \left|\sum_{p=1}^{L}\frac{h_{j_p}}{\alpha_{j_p}}\left[E_{\theta_n}\left(X_{dL+p-1}|V_1^L\in C_i, V_{dL+1}^{dL+L}\in C_j\right) - E_{\theta_n}\left(X_{dL+p-1}|V_{dL+1}^{dL+L}\in C_j\right)\right]\right|\\
&\le \lambda L^2 E_{\max}|\theta_n|^{(d-1)L}\\
&\le \lambda L^2 E_{\max}(|\theta| + \epsilon)^{(d-1)L}.
\end{aligned} \tag{A.13}$$

The terms $(M_{ij}^d - E_iE_j)$ and $(S_{ij}^d - S_j)$ in (A.10) are similarly bounded by exponential functions of $d$. Hence, for all large $n$

$$\left|\operatorname{cov}\left[U_n^r(s_1), U_n^{r+d}(s_2)\right]\right| \le (|\theta| + \epsilon)^{(d-1)L}\cdot O(n^{-3/2}). \tag{A.14}$$

This implies that

$$\begin{aligned}
&\left|\sum_{d=1}^{n/L-1}\left(\frac{n}{L} - d\right)\operatorname{cov}\left[U_n^r(s_1), U_n^{r+d}(s_2)\right]\right|\\
&\le n\cdot\sum_{d=1}^{\infty}|\theta_n|^{(d-1)L}\cdot O(n^{-3/2})\\
&= \left(1 - |\theta_n|^L\right)^{-1}\cdot O(n^{-1/2})\\
&\le \left[1 - (|\theta| + \epsilon)^L\right]^{-1}\cdot O(n^{-1/2}) = o(1),
\end{aligned} \tag{A.15}$$

which completes the proof of part c). Parts d) and e) can be proved in a similar manner. □

## References


[1] C. W. H. Wang, "A minimum distance estimator for first-order autoregressive process," *Ann Statist.*, vol. 14, no. 3, pp. 1180–1193, 1986.

[2] H. L. Koul, "Minimum distance estimation and goodness-of-fit tests in first-order autoregression," *Ann Statist.*, vol. 14, no. 3, pp. 1194–1213, 1986.

[3] L. Denby and R. D. Martin, "Robust estimation of the first-order autoregressive parameter," *J. Amer. Statist. Assoc.*, vol. 74, pp. 140–146, 1979.

[4] P. W. Millar, "Optimal estimation of a general regression function," *Ann. Statist.*, vol. 10, pp. 717–740, 1982.

[5] R. Beran, "Adaptive estimates for autoregressive processes," *Ann. Inst. Statist. Math.*, vol. 28, pp. 77–89, 1976.

[6] J. P. Kreiss, "On adaptive estimation in stationary ARMA processes," *Ann. Statist.*, vol. 15, no. 1, pp. 112–133, 1987.

[7] ——, "On adaptive estimation in autoregressive models when there are nuisance functions," *Statist. and Decisions*, vol. 5, pp. 59–76, 1987.

[8] M. Akahira and K. Takeuchi, *Lecture Notes in Statistics, Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. New York: Springer-Verlag, 1981.

[9] M. Gutman, "On tests for randomness, tests for independence, and universal data compression," submitted to *IEEE Trans. Inform. Theory*.

[10] P. K. Bhattacharya, "Efficient estimation of a shift parameter from grouped data," *Ann. Math. Statist.*, vol. 38, pp. 1770–1787, 1967.

[11] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, pp. 278–286, Mar. 1988.

[12] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, pp. 401–408, Mar. 1989.

[13] R. B. Ash, *Real Analysis and Probability*. New York: Academic Press, 1972.