On the Estimation of the Order of a Markov Chain and Universal Data Compression

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

Abstract—**We focus on the estimation of the order of a finite Markov source based on empirically observed statistics. The following performance criterion is adopted: minimize the probability of underestimating the model order while keeping the overestimation probability exponent at a given prescribed level. A universal asymptotically optimal test, in the sense just defined, is proposed for the case where a given integer is known to upper bound the true order. For the case where such a bound is unavailable, an alternative rule, based on the Lempel–Ziv (LZ) data compression algorithm, is also shown to be asymptotically optimal and computationally more efficient.**

1. INTRODUCTION

THE PROBLEM of selecting the model order of a stochastic process has been widely studied by Akaike [1]–[3], Kayshap [4], Shibata [5], Rissanen [6]–[9], Parzen [10], Hannan [11], Hannan and Quinn [12], Schwarz [13], Tong [14], Wax and Kailath [15], Broersen [16], and others. Most of the methods proposed for order selection are heuristic in the sense that they do not directly minimize a certain measure of the error between the order estimate and the true order. Instead, they define various information criteria [e.g., A information criterion (AIC), Bayesian information criterion (BIC), criterion autoregressive transfer function (CAT), final prediction error (FPE), minimum description length (MDL), etc.], that depend upon the unknown order. The order estimator is usually defined as the minimizing value of this information criterion. However, no optimality results concerning their statistical performance (beyond strong consistency) have been reported for the case where a true order does exist. An exception is Rissanen [6]–[9], whose minimum description length (MDL) rule was shown by Schwarz [13] to be optimal in the Bayesian sense of minimizing the error probability. His results hold for the independent identically distributed (i.i.d.) exponential family (Darmois–Koopman). Rissanen was the first to point out an interesting relation between estimation and coding. His information criterion, the MDL, was motivated by a tight lower bound [9, Theorem 1] on the expected length of a universal lossless code for a given class of probabilistic sources $\{P_\theta\}$ where θ takes values in a compact set $\Omega \subset \mathbb{R}^k$. To approach this lower bound, one should first estimate the parameter vector $\theta = (\theta_1, \cdots, \theta_k)$ from the observed sequence $x = (x_1, \cdots, x_n)$, then encode each component of the estimate $\hat{\theta}_i$ $(i = 1, \cdots, k)$ by $\frac{1}{2}\log n$ bits, and finally, allocate $-\log P_{\hat{\theta}}(x)$ bits for encoding the observation sequence. Hence the total length of the codeword is

$$\mathrm{MDL}(k) \triangleq -\log P_{\hat{\theta}}(x) + \tfrac{1}{2}k \log n.$$

The order estimator $\hat{k}$, proposed by Rissanen, is the value of k which minimizes $\mathrm{MDL}(k)$, and the minimum value $\mathrm{MDL}(\hat{k})$ is a measure of the information treasured in the observed sequence. Notice that the term $\frac{1}{2}k \log n$ plays a role of "penalty" for selecting high orders. Rissanen also proposed predictive versions [9] of the MDL, but these have the same asymptotic behavior as the nonpredictive MDL. Although Rissanen's criterion for the model choice has been shown to be strongly consistent and optimal in the Bayesian sense (as mentioned earlier), his discussion is not limited to the case where a "true" order exists (i.e., the case where the source obeys one of the competing models). When there is no real underlying model for generating the data, there is no longer a meaning for the "performance" of the order estimator in the usual statistical sense but only in the sense of minimizing the description length.

In this paper, we study the order estimation for discrete-time finite-alphabet ergodic Markov chains. In contrast to Rissanen, we limit ourselves to the case where a true order k exists; that is, we assume the data to be actually generated by a kth-order Markov source, and our goal is to estimate the true order k as "accurately" as possible. To measure accuracy, we employ the following performance criterion. Among all estimators $\hat{k}$ for which the overestimation probability $P_k(\hat{k} > k)$ decays faster than $2^{-\lambda n}$ (for a given $\lambda > 0$) *uniformly* for any Markovian probability measure P_k of order k, we wish to find an estimator that minimizes the underestimation probability $P_k(\hat{k} < k)$ uniformly for every P_k. (A more precise definition will be given in Section II.) This criterion, which can be viewed as an extension to the Neyman–Pearson criterion, makes sense for the following two reasons.

1) The overestimation event can be interpreted as a "false alarm" event where "too much effort" is dedicated

Manuscript received November 23, 1987.

N. Merhav was with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel. He is now with AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974.

M. Gutman was with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel. He is now with Codex Corporation, 20 Cabot Boulevard, Mansfield, MA 02048-1193.

J. Ziv is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 8930431.

0018-9448/89/0900-1014$01.00 ©1989 IEEE

to model the given data, while the underestimation case represents an identification failure (missed detection). The statistician would like to guarantee a certain tolerable level of overestimation probability if his computational resources (e.g., memory, computational power) are limited.

2) The trade-off between overestimation and underestimation probability can be balanced by appropriately varying λ. This is different from other existing methods [1]–[15], where the overestimation probability is balanced by adding to the log-likelihood penalty terms which are proportional to the model order.

It should be pointed out that our optimal estimator (in the sense defined earlier), turns out to have an intuitively appealing interpretation related to universal data compression. However, in contrast to Rissanen, who approaches the problem from an information-theoretic point of view *a priori*, here we first define a performance criterion to be optimized, while the relation of our optimal estimator to universal coding is only observed *a posteriori*. (In other words, while Rissanen's rule is efficient for *coding*, our algorithm is optimal in the extended Neyman–Pearson sense described earlier.)

Furthermore, it is shown that every P_k (which can not be reduced to P_{k-1}), some $\lambda > 0$ exists such that both overestimation and underestimation probabilities vanish exponentially fast as $n \to \infty$. On the other hand, it is demonstrated that other existing algorithms yield an overestimation probability larger than $2^{-\epsilon n}$ for any $\epsilon > 0$ and n sufficiently large (i.e., a nonexponential decay). This does not contradict Schwarz [13], who proved the optimality of the MDL rule in the Bayesian sense, because our result is uniformly optimal for any P_k and satisfies the exponential constraint.

Another advantage of the proposed estimator, as compared to existing estimators, concerns its computational complexity and is discussed later.

In the remainder of this paper we propose an optimal order estimator for a case where some given integer k_o is known to upper-bound the true order k. The proposed estimator involves empirical entropies of orders $0 \le j \le k_o$ and is therefore closely related to universal data compression (where the unknown letter probabilities are replaced by their ML estimates). For the case where such an upper bound k_o is unavailable, we derive an alternative estimator, based upon the Lempel–Ziv (LZ) data compression algorithm [17], which is also shown to be asymptotically optimal and computationally more efficient. The techniques applied here are similar to those used by Gutman [18].

II. Problem Formulation

Let $\mathbb{P}_k$ denote the class of all stationary ergodic discrete-time kth-order Markov sources (i.e., $P(x_i|x_{-\infty}^{i-1}) = P(x_i|x_{i-k}^{i-1})$), whereby each random variable x_i takes on values in a finite set A (alphabet) with cardinality $|A| = \alpha$. Denote by $P_k \in \mathbb{P}_k$, a kth-order Markovian probability measure. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n) \in A^n$ be an observed sequence emitted from an unknown kth-order Markovian source $P_k \in \mathbb{P}_k$. We wish to estimate the order k under the following performance criterion. Minimize $P_k(\hat{k}(n) < k)$ for all k and every $P_k \in \mathbb{P}_k$, subject to the following constraint: for every k and every $P_k \in \mathbb{P}_k$,

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log P_k\big(\hat{k}(n) > k\big)\right] > \lambda \tag{1}$$

where $\lambda > 0$ is a given number and $\hat{k}(n) \triangleq \hat{k}$ is an order estimator. That is, we seek a rule which on the one hand guarantees, for every $P_k \in \mathbb{P}_k$, a certain level of overestimation error exponent, and on the other hand, minimizes the underestimation probability, whatever the true underlying probability measure is. We first assume that there is a known integer k_o which upper-bounds the true order k, namely, $0 \le k \le k_o$ ($k = 0$ denotes a memoryless source). This assumption will be later relaxed.

III. Main Results

Let $s_i \triangleq x_{i-k}^{i-1} \equiv (x_{i-1}, x_{i-2}, \cdots, x_{i-k}) \in A^k$, $(1 \le i \le n)$ denotes the state of the Markov source at time i. (It is assumed that the source starts at a fixed state $s_1 = (x_0, x_{-1}, \cdots, x_{-k+1}) \in A^k$). We denote by $\delta(x_i, u, s_i, s)$ the indicator function for $x_i = u$ and $s_i = s$ ($u \in A$, $s \in A^k$). Now let

$$q_{\boldsymbol{x}}^k(u, s) \triangleq \frac{1}{n}\sum_{i=1}^{n}\delta(x_i, u, s_i, s) \tag{2}$$

$$q_{\boldsymbol{x}}^k(s) \triangleq \sum_{u\in A} q_{\boldsymbol{x}}^k(u, s) \tag{3}$$

$$q_{\boldsymbol{x}}^k(u|s) = \begin{cases} q_{\boldsymbol{x}}^k(u, s)/q_{\boldsymbol{x}}^k(s), & q_{\boldsymbol{x}}^k(s) > 0 \\ 0, & q_{\boldsymbol{x}}^k(s) = 0. \end{cases} \tag{4}$$

The $\alpha^k \times \alpha$ matrix whose entries are $q_{\boldsymbol{x}}^k(u, s)$, $u \in A$, $s \in A^k$ will be referred to as the kth-order Markov type of $\boldsymbol{x}$ and denoted by $q_{\boldsymbol{x}}^k$. Clearly, $q_{\boldsymbol{x}}^k$ can be viewed as a kth-order Markovian probability measure for any $\boldsymbol{x}$.

We next define the kth-order empirical entropy and the kth-order divergence as follows:

$$H\big(q_{\boldsymbol{x}}^k\big) \triangleq -\sum_{s\in A^k} q_{\boldsymbol{x}}^k(s)\sum_{u\in A} q_{\boldsymbol{x}}^k(u|s)\log q_{\boldsymbol{x}}^k(u|s) \tag{5}$$

$$D\big(q_{\boldsymbol{x}}^k\|P\big) \triangleq \sum_{s\in A^k} q_{\boldsymbol{x}}^k(s)\sum_{u\in A} q_{\boldsymbol{x}}^k(u|s)\log\frac{q_{\boldsymbol{x}}^k(u|s)}{p(u|s)} \tag{6}$$

where $p(u|s) \triangleq \Pr\{x_i = u|s_i = s\}$, $\log(\cdot) \triangleq \log_2(\cdot)$ and $0\log 0 \triangleq 0$. Let $P_k^n(\boldsymbol{x})$ denote the probability of $\boldsymbol{x} \in A^n$ under P_k. It can now be easily shown that

$$P_k^n(\boldsymbol{x}) = \exp_2\big\{-n\big[H\big(q_{\boldsymbol{x}}^k\big) + D\big(q_{\boldsymbol{x}}^k\|P_k\big)\big]\big\}. \tag{7}$$

Define the following order estimator:

$$k^* = \min\big\{j: H\big(q_{\boldsymbol{x}}^j\big) - H\big(q_{\boldsymbol{x}}^{k_0}\big) \le \lambda\big\}. \tag{8}$$

The following theorem establishes the asymptotic optimality of k^*.

Theorem 1: For any integer k and $P_k \in \mathbb{P}_k$,
a) $\liminf_{n\to\infty}[-(1/n)\log P_k(k^* > k)] \geq \lambda$;
b) for any estimator $\hat{k}$ satisfying (1),

$$\limsup_{n\to\infty}\left[\frac{1}{n}\log P_k(k^* < k)\right] \leq \liminf_{n\to\infty}\left[\frac{1}{n}\log P_k(\hat{k} < k)\right].$$

Notice that the estimator k^* (defined in (8)) has a simple interpretation related to universal data compression. We start from $j = 0$ and seek the first integer j for which $H(q_x^j)$ is sufficiently close to $H(q_x^{k_o})$. The empirical entropy $H(q_x^j)$ is approximately (up to a vanishingly small quantity) equal to the normalized length of Rissanen's universal code for the class $\mathbb{P}_j$. Similarly, $H(q_x^{k_o})$ is related to P_{k_o}. If these two quantities are close enough (with difference less than λ), then there is no significant saving in the codeword length if one increases the memory of the codebook from j to k_o. In that case we therefore decide that j is the order of the Markovian source. Another way is to observe that k^* is asymptotically equivalent to

$$\min\left\{ j; \frac{1}{n}\mathrm{MDL}(j) - \frac{1}{n}\mathrm{MDL}(k_o) < \lambda \right\}$$

where MDL is defined in the Introduction. This then differs from Rissanen's original rule that just minimizes $\mathrm{MDL}(j)$ over the integers $0 \leq j \leq k_o$.

Proof of Theorem 1: Let $M_j \triangleq \{x: H(q_x^j) - H(q_x^{k_o}) \leq \lambda\}$, and let T_x^j be the set of all sequences in A^n with the same jth-order Markov type as x. The cardinality of T_x^j will be denoted by $|T_x^j|$. To prove a) we first establish an auxiliary lemma.

Lemma A: For every $\epsilon > 0$ and n sufficiently large,

$$2^{n[H(q_x^j)-\epsilon]} \leq |T_x^j| \leq 2^{nH(q_x^j)}.$$

The proof appears in Appendix I (Sharper bounds are given in [19]).

Since all sequences in $T_x^{k_o} \subset T_x^k$ are equiprobable, we have

$$P_k(k^* > k) \leq P_k(M_k^c) \leq \max_{P_k \in \mathbb{P}_k} P_k(M_k^c)$$

$$= \sum_{T_x^{k_o} \subset M_k^c} |T_x^{k_o}| 2^{-nH(q_x^k)}$$

(by Lemma A)

$$\leq \sum_{T_x^{k_o} \subset M_k^c} \exp_2\left\{ -n\left[H(q_x^k) - H(q_x^{k_o})\right]\right\}$$

$$\leq (n+1)^{\alpha^{k+1}} 2^{-\lambda n} \leq 2^{-(\lambda-\epsilon)n}$$

for any $\epsilon > 0$ and n sufficiently large. Hence,

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log P_k(k^* > k)\right] \geq \lambda. \tag{9}$$

Part b) follows from the following consideration. Let $\hat{k}$ be an arbitrary order estimator satisfying (1) and induced by the partition of A^n to k_o disjoint decision regions $\{\Omega_j\}_{j=0}^{k_o}$, where Ω_j is the set of x's for which $\hat{k} = j$. As k_o is the maximum possible value of $\hat{k}$, it can be shown that $q_x^{k_o}$ is a sufficient statistic for optimal estimation in the sense of error exponents. Following (1), some $\epsilon > 0$ exists such that for sufficiently large n for any $x' \in \bigcup_{j=k+1}^{k_o} \Omega_j$, and all k:

$$2^{-(\lambda+\epsilon)n} \geq \max_{P_k \in \mathbb{P}_k} P_k(\hat{k} > k) \geq \max_{P_k \in \mathbb{P}_k} \sum_{j=k+1}^{k_o} \sum_{x \in \Omega_j} P_k(x)$$

$$\geq \sum_{x \in T_{x'}^{k_o}} \exp_2\left[-nH(q_x^k)\right]$$

$$= |T_{x'}^{k_o}| \exp_2\left[-nH(q_{x'}^k)\right]$$

(by Lemma A)

$$\geq \exp_2\left[n\left(H(q_{x'}^{k_o}) - \epsilon\right)\right] \exp_2\left[-nH(q_{x'}^k)\right]. \tag{10}$$

As (10) must hold for any $x' \in \bigcup_{j=k+1}^{k_o} \Omega_j$ and for every $P_j \in \mathbb{P}_j$, $j \leq k$, we conclude that

$$\bigcup_{j>k} \Omega_j \subseteq \bigcap_{j\leq k} M_j^c, \qquad 0 \leq k \leq k_o \tag{11}$$

or, equivalently,

$$\bigcup_{j\leq k} \Omega_j = \bigcap_{j>k} \Omega_j^c \supseteq \bigcup_{j\leq k} M_j. \tag{12}$$

Hence, for any $P_k \in \mathbb{P}_k$,

$$P_k(k^* < k) = P_k\left(\bigcup_{j\leq k-1} M_j\right) \leq P_k\left(\bigcup_{j\leq k-1} \Omega_j\right)$$

$$= P_k(\hat{k} < k). \tag{13}$$

This completes the proof of Theorem 1.

Assume now that k_o is unknown, but still k is *bounded*. Clearly, the algorithm k^* as defined in (8) is no longer applicable. We now demonstrate an alternative estimator k^{**}, based on the LZ data compression algorithm [17], which is also shown to be asymptotically optimal.

Let

$$k^{**} \triangleq \min\left\{ j: H(q_x^j) - \frac{1}{n}U_{\mathrm{LZ}}(x) \leq \lambda \right\} \tag{14}$$

where $U_{\mathrm{LZ}}(x)$ is the LZ codeword length of x. That is, the unknown term $H(q_x^{k_o})$ is simply approximated by the normalized LZ codeword length function.

Theorem 2: For any integer k and $P_k \in \mathbb{P}_k$,
a) $\liminf_{n\to\infty}[-1/n \log P_k(k^{**} > k)] \geq \lambda$;
b) for any estimator $\hat{k}$ satisfying (1),

$$\limsup_{n\to\infty}\left[\frac{1}{n}\log P_k(k^{**} < k)\right] \leq \liminf_{n\to\infty}\left[\frac{1}{n}\log P_k(\hat{k} < k)\right].$$

Proof: We use the following inequality (Plotnik and Ziv [20]),

$$\frac{1}{n}U_{\mathrm{LZ}}(x) \leq H(q_x^{k_o}) + \delta(\alpha, n, k_o) \tag{15}$$

where $\delta(\alpha, n, k_o) = O(\log\log n/\log n)$, uniformly for every

$x \in A^n$. To make the paper self-contained the proof of (15) is given in Appendix III.

Now let

$$N_j \triangleq \left\{ x : H(q_x^j) - \frac{1}{n} U_{LZ}(x) \le \lambda \right\}. \tag{16}$$

As for part a), we use Kraft's inequality for uniquely decipherable binary codeword length functions,

$$\begin{aligned} P_k(k^{**} > k) &\le P_k(N_k^c) = \sum_{x \in N_k^c} P_k(x) \le \sum_{x \in N_k^c} \max_{P_k \in \mathbb{P}_k} P_k(x) \\ &= \sum_{x \in N_k^c} 2^{-n[H(q_x^k)]} \le 2^{-\lambda n} \sum_{x \in N_k^c} 2^{-U_{LZ}(x)} \\ &\le 2^{-\lambda n} \sum_{x \in A^n} 2^{-U_{LZ}(x)} \le 2^{-\lambda n}. \end{aligned} \tag{17}$$

It follows from (16) that $N_j \subseteq M_j$ for M_j defined with a threshold $\lambda + \delta(\alpha, n, k_o)$. However, $\delta(\alpha, n, k_o) \to 0$ as $n \to \infty$.

Now, by Theorem 1,

$$\begin{aligned} P_k(k^{**} < k) &= P_k\left(\bigcup_{j<k} N_j \right) \le P_k\left(\bigcup_{j<k} M_j \right) \\ &\le P_k\left(\bigcup_{j<k} \Omega_j \right) = P_k(\hat{k} < k). \end{aligned} \tag{18}$$

This completes the proof of Theorem 2.

The following remarks are appropriate.

1) For any $j < k$, let

$$D(P_k \| P_j) \triangleq \sum_{x_1^k \in A^k} P_k(x_1^k) \sum_{x_{k+1} \in A} P_k(x_{k+1} | x_1^k) \cdot \log \frac{P_k(x_{k+1} | x_1^k)}{P_k(x_{k+1} | x_{k-j+1}^k)}.$$

a) If $\lambda < D(P_k \| P_j)$, $0 \le j \le k-1$, then the underestimation probabilities associated with k^* and k^{**} tend to zero exponentially fast as $n \to \infty$. Let

$$e(\lambda) = \lim_{n \to \infty} \left[-\frac{1}{n} \log P_k(k^* < k) \right] = \lim_{n \to \infty} \left[-\frac{1}{n} \log P_k(k^{**} < k) \right]$$

There exists a value $\lambda = \lambda_o$ (depending on P_k), for which $e(\lambda_o) = \lambda_o$, namely, the errors are symmetric.

b) If $\lambda > D(P_k \| P_j)$, then the probability of underestimation tends to 1 as $n \to \infty$.

The choice of λ is, of course, dictated by the maximum tolerable level of overestimation probability (see (1)). However, in view of a) and b), one should select λ to be small enough to guarantee the exponential decay of both error probabilities, for a large set of probability measures in the class. However, since λ is just a single parameter, it can easily be adjusted empirically to balance appropriately the trade-off between rates of the two kinds of errors. In fact, this is always the case in Neyman–Pearson detection. The detector first computes a test statistic and then compares it to an adjustable threshold.

2) In view of remark 1), it is demonstrated in Appendix II that, for existing order estimators (MDL, AIC, etc.), the overestimation probability vanishes more slowly than any exponent. This result does not contradict Schwarz [13], who proved the optimality of the MDL rule in the Bayesian sense, because the integration over the parameter space (defined by the transition probabilities) includes Markovian measures for which $D(P_k \| P_j) = 0$.

3) The estimation rules k^* and k^{**} are optimal only if the true order is bounded. This bound is either known (k^*) or unknown (k^{**}). These rules are universal in the sense that the optimality holds for every source in the class $\mathbb{P}_k$.

4) The computational complexity involved in applying k^* or k^{**} is smaller than that associated with the existing rules, with probability larger than $1 - 2^{-\lambda n}$, because the first group stops at the first time a threshold is exceeded, while the last group seeks a global minimum among the integers. In addition, k^{**} saves the computational effort associated with the calculation of $H(q_x^{k_o})$, which grows exponentially as k_o increases. In other words, since k_o might be much larger than k, the computational complexity involved with the evaluation of $U_{LZ}(x)$ is, in general, much smaller than that associated with the evaluation of $H(q_x^{k_o})$. (In many cases the complexity of calculating $U_{LZ}(x)$ is smaller than that of $H(q_x^k)$.)

5) The estimators k^* and k^{**} are optimal if no prior information whatsoever exists about the parameters of the true underlying P_k. The reader might argue that some gain in performance can be achieved if one knows a priori that P_j must lie in a smaller subset $\mathbb{F}_j \subset \mathbb{P}_j$ $(0 \le j \le k_o)$. In fact this partial prior knowledge can be utilized to improve performance. The modification needed for k^* and k^{**} would be to decrease the threshold λ by the quantity $\min_{P \in \mathbb{F}_j} D(q_x^j \| P)$, in (8) and (14), respectively.

6) The estimator k^* can be extended in two different directions:

a) estimation of the number of states of a finite-state finite-alphabet unifilar/nonunifilar source (Ziv and Merhav [21]);

b) estimation of the model order in exponential families [22] (by applying the theory of large deviations). This class includes the widely used continuous alphabet models: the Gaussian linear regression model, the autoregressive (AR) model, and several well-known hypothesis testing problems.

7) It should be emphasized that, despite the indirect relation of our algorithm to data compression, it is not optimal for coding, in contrast to Rissanen's rule. However, it minimizes the above defined cost, if a "true" order exists.

ACKNOWLEDGMENT

We wish to thank the anonymous referees for their helpful suggestions.

APPENDIX I
PROOF OF LEMMA A

By (7),

$$1 \geq \max_{P \in \mathbf{P}_j} P\{T_x^j\} = |T_x^j| 2^{-n[H(q_x^j) + \min_{P \in \mathbf{P}_j} D(q_x^j \| P)]}$$

$$= |T_x^j| 2^{-nH(q_x^j)} \tag{A1}$$

and therefore $|T_x^j| \leq 2^{nH(q_x^j)}$. As for the lower bound, suppose there exist some $\epsilon > 0$ and infinitely many values of n such that $|T_x^j| < 2^{n[H(q_x^j)-\epsilon]}$. It follows that

$$2^{-n\epsilon} > 2^{-nH(q_x^j)} |T_x^j| = \max_{P \in \mathbf{P}_j} P\{T_x^j\} \geq P\{T_x^j\}. \tag{A2}$$

As the number of distinct jth-order Markov types is never larger than $(n+1)^{\alpha^{j+1}}$, we obtain by (A2),

$$1 = \sum_{T_x^j \subset A^n} P\{T_x^j\} < \sum_{T_x^j \subset A_n} 2^{-n\epsilon} < (n+1)^{\alpha^{j+1}} 2^{-n\epsilon}. \tag{A3}$$

Since the right side tends to zero as $n \to \infty$, the assumption is contradicted. This completes the proof of Lemma A.

APPENDIX II
THE OVERESTIMATION PROBABILITIES FOR EXISTING RULES

To demonstrate the slow convergence of the overestimation probabilities for existing rules, we focus on the MDL rule (Rissanen [6]–[9]). The considerations are similar for other rules as well (AIC, BIC, FPE, etc.).

Consider the case $\alpha = 2$, $k_o = 1$; that is, we select $\hat{k}$ in accordance with

$$\hat{k} = \arg\min_{j=0,1} \left[-\log \max_{P_j \in \mathbf{P}_j} P_j(x) + \tfrac{1}{2} 2^j \log n \right]$$

$$= \arg\min_{j=0,1} \left[H(q_x^j) + \delta_j(n) \right] \tag{B1}$$

where $\delta_j(n) = 2^{j-1}(\log n)/n$. Thus, we decide in favor of $\hat{k} = 1$ whenever

$$x \in \Omega_1 \triangleq \{ x : H(q_x^0) - H(q_x^1) > \delta(n) \} \tag{B2}$$

where $\delta(n) = \delta_1(n) - \delta_0(n) = \frac{1}{2}(\log n)/n$.

We now lower-bound the overestimation probability,

$$P_0(\hat{k} = 1) = \sum_{x \in \Omega_1} P_0(x) = \sum_{T_x^1 \subset \Omega_1} P_0(T_x^1)$$

(since $T_x^1 \subset T_x^0$)

$$= \sum_{T_x^1 \subset \Omega_1} |T_x^1| P_0(x) \geq 2^{n[H(q_x^1) - (\epsilon/3)]} 2^{-n[H(q_x^0) + D(q_x^0 \| P_0)]}$$

$$= 2^{-n[D(q_x^0 \| P_0) + H(q_x^0) - H(q_x^1) + (\epsilon/3)]} \tag{B3}$$

for any $x \in \Omega_1$, $\epsilon > 0$ and n sufficiently large. Since $\delta(n) \to 0$ as $n \to \infty$, there exists $x \in \Omega_1$ for which $\delta(n) < H(q_x^0) - H(q_x^1) \leq \epsilon/3$ and at the same time $D(q_x^0 \| P_0) \leq \epsilon/3$ for any $\epsilon > 0$ and n sufficiently large. It now follows from (B2) that $P_o(\hat{k} = 1) \geq 2^{-n\epsilon}$ for any $\epsilon > 0$ and n sufficiently large; that is the overestimation probability tends to zero slower than any exponent.

APPENDIX III
PROOF OF INEQUALITY (15)

Let us apply the incremental parsing procedure [17] to the observed sequence x, and let c denote the resulting number of distinct phrases in x. Denote by x_i the ith phrase ($1 \leq i \leq c$). It is easy to show that there exists a Markovian probability measure $P_{k_o} \in \mathbf{P}_{k_o}$ for which $P_{k_o}(x) = 2^{-nH(q_x^{k_o})}$ (by taking the maximum likelihood estimator for the transition probabilities). Thus

$$nH(q_x^{k_o}) = -\log P_{k_o}(x) = -\log \prod_{i=1}^{c} P(x_i | s_i)$$

(where s_i denotes the initial state of phrase i)

$$= -\sum_{l=1}^{l_{\max}} \sum_{s=1}^{S} \sum_{j=1}^{c(l,s)} \log P(x_j | s) \tag{C1}$$

where $c(l, s)$ denotes the number of l-length phrases starting with state s, $S = \alpha^{k_o}$ is the number of states, and $l_{\max}$ is the length of the longest phrase. Let us consider the inner summation:

$$-\sum_{j=1}^{c(l,s)} \log P(x_i | s) = -c(l,s) \sum_{j=1}^{c(l,s)} \frac{1}{c(l,s)} \log P(x_j | s)$$

(by Jensen's inequality)

$$\geq -c(l,s) \log \sum_{j=1}^{c(l,s)} \frac{P(x_i | s)}{c(l,s)} \geq c(l,s) \log c(l,s) \tag{C2}$$

where we used the fact that $\sum P(x_i | s) \leq 1$ for fixed length distinct phrases. Plugging (C2) into (C1) we get

$$nH(q_x^{k_o}) \geq \sum_{l=1}^{l_{\max}} \sum_{s=1}^{S} c(l,s) \log c(l,s)$$

$$= \sum_{l=1}^{l_{\max}} c(l) \sum_{s=1}^{S} \frac{c(l,s)}{c(l)} \left[\log \frac{c(l,s)}{c(l)} + \log c(l) \right] \tag{C3}$$

where $c(l) = \sum_s c(l, s)$.

The first term in the inner sum has the form of an entropy (with the sign changed); thus it is always greater than $-\log S$. We therefore obtain

$$nH(q_x^{k_o}) \geq \sum_{l=1}^{l_{\max}} c(l) \log c(l) - c \log S. \tag{C4}$$

As for the first term on the right-hand side of (C4),

$$\sum_{l=1}^{l_{\max}} c(l) \log c(l) = -c \sum_{l=1}^{l_{\max}} \frac{c(l)}{c} \log \frac{1}{c(l)}$$

$$= -c \sum_{l=1}^{l_{\max}} \frac{c(l)}{c} \log \frac{2^{-l/\bar{l}}}{c(l)} - c$$

(where $\bar{l} \triangleq (1/c) \sum_{l=1}^{l_{\max}} l c(l) = n/c$)

$$\geq -c \log \sum_{l=1}^{l_{\max}} \frac{c(l)}{c} \frac{2^{-l/\bar{l}}}{c(l)} - c \geq -c + c \log c - c \log \sum_{l=1}^{l_{\max}} 2^{-l/\bar{l}}$$

$$\geq -c + c \log c - c \log \frac{2^{-1/\bar{l}}}{1 - 2^{-1/\bar{l}}} \geq -c + c \log c + c \log (2^{-1/\bar{l}} - 1)$$

(by the Taylor expansion for 2^x)

$$\geq -c + c \log c + c \log \left(\frac{\ln 2}{\bar{l}} \right) = c \log c - c(1 - \log \ln 2) - c \log \frac{n}{c}$$

$$= c \log c - c\delta - c \log \frac{n}{c} \tag{C5}$$

where $\delta = 1 - \log\ln 2$.

Thus by (C4)

$$nH(q_x^{k_o}) \geq c\log c - c(\delta + \log S) - c\log\frac{n}{c}. \tag{C6}$$

On the other hand, it is shown in [17] that

$$U_{LZ}(x) \leq c\log c + c\log\alpha + c. \tag{C7}$$

It follows from (C6) and (C7) that

$$\frac{1}{n}U_{LZ}(x) - H(q_x^{k_o}) \leq \frac{c}{n}(1 + \log\alpha + \log S + \delta) + \frac{c}{n}\log\frac{n}{c}$$
$$= \frac{c}{n}\cdot K + \frac{\log(n/c)}{n/c} \tag{C8}$$

where K is a constant.

It is shown in [23, Theorem 2] that $c < n/[(1-\epsilon_n)\log n]$ where $\epsilon_n = 2[1 + \log\log(\alpha n)]/\log n$. Therefore, the first term in (C8) tends to zero as quickly as $1/\log n$. As for the second term, since the function $(\log x)/x$ is monotonically decreasing for $x > e$, it follows that for n large enough $(n/c > e)$

$$\frac{\log(n/c)}{n/c} < \frac{\log[(1-\epsilon_n)\log n]}{(1-\epsilon_n)\log n} = O\left(\frac{\log\log n}{\log n}\right).$$

This completes the proof of inequality (15).

References

[1] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, vol. 21, pp. 243–247, 1969.

[2] ____, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

[3] ____, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[4] R. L. Kayshap, "Inconsistency of the AIC rule for estimating the order of autoregressive models," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 996–998, 1980.

[5] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117–126, 1976.

[6] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[7] ____, "Consistent order estimates of autoregressive processes by shortest description of data," in *Analysis and Optimization of Stochastic Systems*. New York: Academic, 1980.

[8] ____, "Estimation of structure by minimum description of length," *Circuits, Syst., Signal Processing*, vol. 1, nos. 3–4, pp. 395–406, 1982.

[9] ____, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[10] E. Parzen, "Some recent advances in time series modeling," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 723–730, 1974.

[11] E. J. Hannan, "The estimation of the order of an ARMA process," *Ann. Statist.*, vol. 8, pp. 1071–1081, 1980.

[12] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc.*, Ser. B, 41, no. 2, pp. 190–195, 1979.

[13] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[14] H. Tong, "Determination of the order of a Markov chain by Akaike's information criterion," *J. Appl. Prob.*, vol. 12, pp. 488–497, 1975.

[15] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, 1985.

[16] P. M. T. Broersen, "Selecting the order of autoregressive models from small samples," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 4, pp. 874–879, 1986.

[17] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, 1978.

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

[19] ____, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol IT-35, no. 2, pp. 401–408, Mar. 1989.

[20] E. Plotnik and J. Ziv, "On the pointwise convergence of universal data compression algorithms," submitted for publication to the *IEEE Trans. Inform. Theory*.

[21] J. Ziv and N. Merhav, "Estimating the number of states of a finite-state source," submitted for publication.

[22] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, pp. 326–333, Sept. 1989.

[23] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 75–81, Jan. 1976.