Order Estimation and Sequential Universal Data Compression of a Hidden Markov Source by the Method of Mixtures

Chuang-Chun Liu and Prakash Narayan, *Senior Member, IEEE*

Abstract—**We consider first the estimation of the order, i.e., the number of states, of a discrete-time finite-alphabet stationary ergodic hidden Markov source (HMS). Our estimator uses a description of the observed data in terms of a uniquely decodable code with respect to a mixture distribution, obtained by suitably mixing a parametric family of distributions on the observation space. This procedure avoids maximum likelihood calculations. The order estimator is shown to be strongly consistent with the probability of underestimation decaying exponentially fast in the number n of observations, while the probability of overestimation does not exceed cn^{-3}, where c is a constant. Next, we present a sequential algorithm for the uniquely decodable universal data compression of the HMS, which performs an on-line estimation of source order followed by arithmetic coding. This code asymptotically attains optimum average redundancy.**

Index Terms—**Hidden Markov source, order, estimator, consistency, mixture distribution, universal data compression, uniquely decodable, pointwise redundancy, average redundancy.**

I. Introduction

A KEY feature of current research in speech processing involves the development of mathematical models for the speech signal. A popular choice of such a model is the hidden Markov model, also referred to as the hidden Markov source (HMS) [21]. This paper is concerned with two important problems arising in the study of the HMS, namely order estimation and sequential (i.e., symbol-by-symbol), noiseless universal data compression. The difficulty encountered by the statistical approach to these problems, viz., the computation of maximum likelihood functions, is circumvented by employing an information-theoretic approach.

Order estimation and noiseless universal data compression are fundamental to the statistical modeling of observed data, i.e., finding a model or class of models that completely capture the salient features of the observed data. A widely-studied approach to modeling data, advocated by several authors (cf., e.g., Rissanen [22]–[25], Merhav *et al.* [16], Barron [1]) involves representing the data by its shortest noiseless universal (data-compression) code. On the other hand, the length of such a codeword for the data in terms of a given model is closely related to the notion of order or complexity of the model. Often, this complexity is determined by the number of parameters that specify the models in a given class. For a discrete alphabet Markov process, the number of parameters is determined by the size of the alphabet and the order (memory) of the process. For an HMS, the number of parameters is determined by its order (number of states of the underlying Markov process), and the size of the observation alphabet.

The problems of estimating the order of a Markov process [7], [15], [16] and that of a finite-state source and an HMS [7], [30] have recently received much attention. Merhav, Gutman, and Ziv [16] have proposed an algorithm (the MGZ algorithm) for estimating the order of a discrete-time discrete-alphabet Markov process. Their approach is also employed to estimate the number of parameters of an independent and identically distributed (IID) exponential family of distributions [17], and the number of states of a finite-state source [30]. It employs a Neyman–Pearson-like criterion, namely, minimizing the probability of underestimation (i.e., selecting an order that is smaller than the true order) while constraining the overestimation probability to decrease exponentially fast in the number of observations. For the Markov case, Merhav, Gutman, and Ziv [16] have also shown that if the exponent governing the overestimation probability is small enough, the optimal order estimator yields an exponentially decreasing underestimation probability and is consistent under this condition. However, when the prescribed overestimation exponent is too large, the estimator becomes inconsistent with the probability of underestimation approaching unity. From the point of view of data compression, this tendency to underestimate can be very serious. Intuitively, if models of higher orders are allowed so that they include the true data-generating distribution, even though the redundancy may not be optimal in encoding the data, its normalized value with respect to the number of observations tends to zero almost surely. On the other hand, a restriction to lower order models may

Manuscript received July 2, 1992; revised August 12, 1993. The work of C.-C. Liu was supported by the Systems Research Center, University of Maryland, College Park, under NSF Grant OIR-85-00108. The work of P. Narayan was supported by the Institute for Systems Research, University of Maryland, College Park, under NSF Grant OIR-85-00108, and by Sonderforschungsbereich 343, Discrete Strukturen in der Mathematik, Universität Bielefeld, Bielefeld, Germany.

C.-C. Liu was with the Department of Electrical Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742. He is now with the IBM Corporation, San Jose, CA 95123.

P. Narayan is with the Department of Electrical Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742.

IEEE Log Number 9403834.

preclude the true distribution so that the average normalized redundancy does not vanish as the number of observations increases.

A variant of the Merhav–Gutman–Ziv (MGZ) method has been employed by Weinberger *et al.* [28] to compress data emitted by an unifilar source (a subclass of the set of finite-state sources), assuming the source state at each time instant to depend on at most a fixed number of past source symbols. Their approach consists of estimating first the states at each time instant and subsequently using these estimates in an arithmetic code. This procedure does not generalize immediately to the compression of (general) finite-state sources or hidden Markov sources. Ziv and Merhav [30] have proposed an estimator for the order of a finite-state source wherein for each possible value of order, a function of the maximum likelihood probability of the observed data is compared with its average Lempel–Ziv data compression length. Their estimator asymptotically attains the minimum probability of underestimation among all estimators with a prescribed exponential decay rate of overestimation probability. On the other hand, this estimator tends to underestimate the source order. We shall show that a slight modification of the Ziv–Merhav [30] approach results in consistency.

Finesso [7] has recently finessed a technique for estimating the order of a Markov source, which involves the minimization of a description length consisting of a likelihood function together with a compensation term. The "smallest" possible compensation terms are obtained via the *law of iterated logarithm* for the maximum likelihood function. A generalization of this approach to an HMS is as yet elusive; the major obstacle is the lack of a law of iterated algorithm for the corresponding maximum likelihood estimate (MLE). Nonetheless, for an HMS, by approximating the maximum likelihood function by model complexity, Finesso [7] has succeeded in choosing compensation terms that ensure the strong consistency of the corresponding estimator. Kieffer [11] has proposed an estimator for the order of a class of sources, including Markov, hidden Markov, and finite-state sources. His estimator, which is strongly consistent, is based on Rissanen's minimum description length (MDL) principle (cf., e.g., [22]).

The modified MGZ algorithm [28], Finesso's approach [7], and Kieffer's estimator [11] all rely on maximum likelihood estimates. In general, the MLE is very difficult to compute exactly; furthermore, there is no known algorithm (including the EM algorithm) that guarantees a convergence of its estimate to the true MLE. Thus, although theoretically sound, the actual estimates in [7], [11], [28] may be intractable in practice.

In this paper, we first consider the estimation of the order of a discrete-time finite-alphabet stationary ergodic HMS. Our estimator employs a description of the observed data in terms of a uniquely decodable code with respect to a mixture distribution, obtained by suitably mixing a parametric family of distributions on the observation space. The mixture distribution for the HMS, proposed by Csiszár [3], was motivated by the work of Davisson *et al.* [6] and Shtar'kov [27]. Our approach avoids computationally burdensome maximum likelihood calculations; however, the evaluation of the mixture distribution is itself arduous. The resulting order estimator is shown to be strongly consistent.

We next propose a uniquely decodable universal code for the sequential data compression of the HMS. This scheme employs the previous estimate of HMS order followed by arithmetic coding. It is shown that our code asymptotically attains optimum average redundancy by dint of the adequacy of the rate of convergence of the HMS order estimator.

The remainder of the paper is organized as follows. Section II describes the problem of HMS order estimation as well as that of a general stationary ergodic source. The HMS order estimator based on mixture distributions is treated in Section III, and the universal data-compression scheme is presented in Section IV. Section V discusses a problem of *inexact* or *mismatched* modeling; the consistent estimation of the order of a general stationary ergodic process is also addressed using a minor modification of the Ziv–Merhav approach [30].

II. Preliminaries

Let $\mathcal{S} = \{1,\cdots,k\}$ be a finite set of integers. Let $\{S_n\}_{n=0}^{\infty}$ be an $\mathcal{S}$-valued first-order, stationary ergodic Markov process, generated by a $k \times k$-stochastic matrix $A = \{a_{uv}\}$, and an initial probability distribution π on $\mathcal{S}$. Here, $a_{uv} \triangleq P(S_n = v | S_{n-1} = u)$, u and v in $\mathcal{S}$, denote the transition probabilities of the process $\{S_n\}_{n=0}^{\infty}$. Throughout this paper, we shall use the notation $\mathbf{s}_m^n$ to refer to the subsequence $(\mathbf{s}_m,\cdots,\mathbf{s}_n)$, $0 \le m < n$, of symbols from $\mathcal{S}$.

Let $\mathcal{X} = \{1,\cdots,q\}$, $q \ge 2$, be a finite set of integers. Let $\{X_n\}_{n=1}^{\infty}$ be an $\mathcal{X}$-valued stochastic process, which is generated by the process $\{S_n\}_{n=0}^{\infty}$ according to the following probabilistic mechanism:

$$\begin{aligned} b_{il} &\triangleq P(X_n = l | S_n = i,\ S_{n-1}\cdots S_0, X_{n-1}\cdots X_1) \\ &= P(X_n = l | S_n = i), \qquad 1 \le i \le k \quad 1 \le l \le q \end{aligned} \tag{2.1}$$

for $n \ge 1$, where $B = \{b_{il}\}$ is a $k \times q$-stochastic matrix. The process $\{X_n\}_{n=1}^{\infty}$ so generated, which is a function of the Markov chain $\{S_n\}_{n=0}^{\infty}$, is commonly referred to as a *hidden Markov source* (HMS). The n-dimensional joint probability distribution of the HMS $\{X_n\}_{n=1}^{\infty}$ is completely determined by an initial distribution π on $\mathcal{S}$, and the stochastic matrices A, B. In particular, we have

$$\begin{aligned} P(\mathbf{X}_1^n = \mathbf{x}_1^n | S_0 = s_0) &= \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} P(\mathbf{X}_1^n = \mathbf{x}_1^n | \mathbf{S}_1^n = \mathbf{s}_1^n, S_0 = s_0) \\ &\quad \cdot P(\mathbf{S}_1^n = \mathbf{s}_1^n | S_0 = s_0) \\ &= \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} \prod_{m=1}^{n} P(X_m = x_m | S_m = s_m) \\ &\quad \cdot P(S_m = s_m | S_{m-1} = s_{m-1}) \\ &= \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} \prod_{m=1}^{n} b_{s_m x_m} a_{s_{m-1} s_m} \end{aligned} \tag{2.2}$$

where, in keeping with previous notation, $\mathbf{X}_m^n \triangleq (X_m, \cdots, X_n)$ and $\mathbf{x}_m^n \triangleq (s_m, \cdots, x_n)$, $1 \le m < n$. Hereafter, the *order* of the HMS $\{X_n\}_{n=1}^{\infty}$ will refer to the cardinality of the state space $\mathcal{S}$ of the underlying Markov process $\{S_n\}_{n=0}^{\infty}$. For IID and Markov processes, order will be defined in terms of "memory." Thus, an $\mathcal{X}$-valued Markov process $\{X_n\}_{n=1}^{\infty}$ with $P(X_n = x_n | \mathbf{X}_1^{n-1} = \mathbf{x}_1^{n-1}) = P(X_n = x_n | \mathbf{X}_{n-k}^{n-1} = \mathbf{x}_{n-k}^{n-1})$, for $n > k$, will be said to have order k, $k \ge 1$; an IID process will have order 0. *We shall assume that the HMS $\{X_n\}_{n=1}^{\infty}$ is observed, and that its order k is unknown, except that it does not exceed a known integer $k_0 \ge 1$; the stochastic matrices A and B are also assumed to be unknown.* Our first objective is to obtain a consistent estimate of the order k based on observations of the process $\{X_n\}_{n=1}^{\infty}$. A second objective is to find uniquely decodable (UD) universal codes for the HMS $\{X_n\}_{n=1}^{\infty}$ achieving minimal redundancy in a suitable sense.

To this end, we begin by distinguishing between three parametric spaces for each k, $1 \le k \le k_0$. First, let Θ^k denote the set of all pairs of stochastic matrices (A, B), where A is a $k \times k$-stochastic matrix and B is a $k \times q$-stochastic matrix. Next, for $\delta > 0$, let $\Theta_\delta^k \subseteq \Theta^k$ denote the set of pairs of such stochastic matrices (A, B), as satisfy $a_{ij} \ge \delta$, $b_{il} \ge \delta$, $1 \le i, j \le k$, $1 \le l \le q$. (Clearly, this requirement cannot be met for all $\delta > 0$.) Finally, $\Theta_0^k \subseteq \Theta^k$ is defined to be the set of pairs of stochastic matrices (A, B), where each matrix A generates a unique stationary distribution that is necessarily ergodic; note that $\Theta_\delta^k \subseteq \Theta_0^k$. Let $\mathcal{X}^\infty$ denote the set of all infinite sequences of symbols from $\mathcal{X}$. Throughout this paper, *hidden Markov measures on $\mathcal{X}^\infty$ will refer to those generated in accordance with* (2.1) *and* (2.2), *with the underlying Markov process having a unique stationary distribution*; for θ in Θ_0^k, $1 \le k \le k_0$, let P_θ denote a (stationary ergodic) hidden Markov measure on $\mathcal{X}^\infty$.

Lemma 2.1: For each θ in Θ_0^k, $1 \le k < k_0$, there exists $\tilde{\theta}$ in Θ_0^{k+1} such that P_θ and $P_{\tilde{\theta}}$ constitute equal measures on $\mathcal{X}^\infty$.

Proof: The claim is perfectly trivial. □

An obvious ambiguity that may arise in the aforementioned estimation problem concerns the possible lack of uniqueness of the "true" order of the HMS: more than one distinct parameter (corresponding to different values of k) may yield the same measure on $\mathcal{X}^\infty$. The mathematical difficulty posed by this identifiability issue is usually circumvented in a standard manner by assuming the HMS $\{X_n\}_{n=1}^{\infty}$ to be regular (cf. [2], [20]): a regular HMS of order k cannot induce the same measure on $\mathcal{X}^\infty$ as any other HMS of a lower order [20]. However, the ambiguity concerning the "true" order can be resolved without recourse to the assumption of regularity by adopting Rissanen's guiding principle of model-building [22], namely that the *simplest* model in the class of models that conforms to the observed data indeed constitutes the best model of the data. Thus, the "true" order of the HMS is the smallest value of k, $1 \le k \le k_0$, for which there exists a parameter θ in Θ_0^k, such that P_θ and the measure on $\mathcal{X}^\infty$ corresponding to the observed process $\{X_n\}_{n=1}^{\infty}$ are equal. We define a set of *minimal models* $\mathcal{M}$ as the set of pairs (k, θ), $1 \le k \le k_0$, θ in Θ_0^k, with the property that for any (k, θ) in $\mathcal{M}$, there exists no pair (k', θ'), $1 \le k' < k$, θ' in $\Theta_0^{k'}$, such that P_θ and $P_{\theta'}$ are equal measures on $\mathcal{X}^\infty$. We shall assume that the *observed process $\{X_n\}_{n=1}^{\infty}$ is generated in accordance with P_θ, θ in Θ_0^k, for some (k, θ) in $\mathcal{M}$*. Our first objective then is to search for an estimator of order that is consistent in the sense that the corresponding estimate converges, for large observed samples, to a value of k associated with a minimal model in $\mathcal{M}$.

We begin with two pertinent technical lemmas for an HMS. The first lemma, stated and proved in [7, Lemma 1.4.3], establishes the relationship between the parametric sets Θ_δ^k, $1 \le k \le k_0$. For the sake of completeness, we repeat the result below.

Lemma 2.2 (Finesso): For each θ in Θ_δ^k, there exists $\tilde{\theta}$ in $\Theta_{\frac{\delta}{2}}^{k+1}$ such that P_θ and $P_{\tilde{\theta}}$ are equal measures on $\mathcal{X}^\infty$.

Proof: The proof is based on a straightforward state-splitting argument. Let $\theta = (A, B)$, where

$$A = \{a_{u,v},\ 1 \le u \le k,\ 1 \le v \le k\},$$

$$B = \{b_{i,l},\ 1 \le i \le k,\ 1 \le l \le q\}.$$

Set $\tilde{\theta} = (A', B')$, where

$$A' = \{a'_{u,v},\ 1 \le u \le k+1,\ 1 \le v \le k+1\},$$

$$B' = \{b'_{i,l},\ 1 \le i \le k+1,\ 1 \le l \le q\}$$

with

$$a'_{u,v} = a_{u,v}, \qquad 1 \le u \le k,\ 1 \le v \le k-1$$

$$a'_{u,v} = \frac{a_{u,v}}{2}, \qquad 1 \le u \le k,\ k \le v \le k+1$$

$$a'_{k+1,v} = a'_{k,v}, \qquad 1 \le v \le k+1.$$

Clearly $\tilde{\theta}$ is in $\Theta_{\frac{\delta}{2}}^{k+1}$, and $P_{\tilde{\theta}} = P_\theta$. □

We define the *Kullback–Leibler distance* or *information divergence* between two stationary ergodic hidden Markov measures P and P' on $\mathcal{X}^\infty$ as

$$D(P\|P') \triangleq \lim_n E_P\left[\log \frac{P(X_{n+1}|\mathbf{X}_1^n)}{P'(X_{n+1}|\mathbf{X}_1^n)}\right] \tag{2.3}$$

where E_P denotes expectation with respect to the measure P. The information divergence in (2.3) is well-defined; a proof of this fact can be found in [7, Theorem 2.3.3].

The following is a key result used in this paper, with independent proofs by Leroux [14, Theorem 2], Finesso [7, Theorem 2.3.3] and Csiszár-Shields [5].

Lemma 2.3 (Csiszár–Shields, Finesso, Leroux): If P_θ and $P_{\theta'}$ are stationary ergodic hidden Markov measures on $\mathcal{X}^\infty$, then

$$\lim_n \frac{1}{n} \log \frac{P_\theta(\mathbf{X}_1^n)}{P_{\theta'}(\mathbf{X}_1^n)} = D(P_\theta\|P_{\theta'}) \qquad P_\theta - \text{a.s.}$$

If P_θ and $P_{\theta'}$ are not equal measures on $\mathcal{X}^\infty$, then $D(P_\theta\|P_{\theta'}) > 0$.

Remark: If P_θ is a "general" stationary ergodic measure on $\mathscr{X}^\infty$, $\lim_n 1/n \log P_\theta(\mathbf{x}_1^n)/P_{\theta'}(\mathbf{x}_1^n)$ exists when $P_{\theta'}$ is a Markov measure of finite order. Although an HMS is usually not a Markov process of finite order, the limit in Lemma 2.3 nonetheless exists (cf. [7, Theorem 2.3.]).

We close this section with a series of technical results for a (general) $\mathscr{X}$-valued stationary ergodic process, which is not necessarily an HMS. These results will be applied in the next section to an HMS in proving the consistency of its order estimator. They are also of independent interest (cf. the comment in Section V concerning the Ziv–Merhav estimator [30] of the order of a general stationary ergodic process).

Let $\{Y_n\}_{n=1}^\infty$ be an $\mathscr{X}$-valued stationary ergodic process. Let $\{\mathbf{P}_k\}_{k=1}^{k_0}$ be a sequence of families of stationary ergodic probability measures on $(\mathscr{X}^\infty, \mathscr{B}^\infty)$, where $\mathscr{B}^\infty$ is the standard σ-field on $\mathscr{X}^\infty$, and $k_0 \geq 1$ is a known integer as in Section II. If the process $\{Y_n\}_{n=1}^\infty$ is generated according to a probability measure in $\mathbf{P}_k$, we refer to k as the *order* of the stationary ergodic process.

Let $\mathbf{Y}$ and $\mathbf{y}$ denote, respectively, the infinite sequence of $\mathscr{X}$-valued random variables $(Y_1, \cdots, Y_n, \cdots)$, and an element $(y_1, \cdots, y_n, \cdots)$ in $\mathscr{X}^\infty$. We assume that $\{\mathbf{P}_k\}_{k=1}^{k_0}$ satisfies the following conditions.

(A1): For each k, $1 \leq k \leq k_0$, $\mathbf{P}_k$ is a parametric family, namely $\mathbf{P}_k = \{P_\theta : \theta \in \prod^k\}$, where $\prod^k$ is a compact subset of a metric space with metric $d(\cdot, \cdot)$. Furthermore, $\mathbf{P}_k \subseteq \mathbf{P}_{k+1}$, $1 \leq k < k_0$.

(A2): For each $n \geq 1$, we assume for θ in $\bigcup_{k=1}^{k_0} \prod^k$ that $P_\theta(\mathbf{y}_1^n) > 0$ for all $\mathbf{y}_1^n$ in $\mathscr{X}^n$. Furthermore, for θ, θ' in $\prod^k$, $1 \leq k \leq k_0$, and $\mathbf{y} = (y_1, \cdots, y_n, \cdots)$ in $\mathscr{X}^\infty$

$$L_{\theta,\theta'}^n(\mathbf{y}_1^n) \triangleq \frac{1}{n} \log \frac{P_\theta(\mathbf{y}_1^n)}{P_{\theta'}(\mathbf{y}_1^n)} \tag{2.4}$$

is assumed to be equicontinuous in θ'.

Next, we define

$$D_{\theta,\theta'}(\mathbf{y}) \triangleq \liminf_n L_{\theta,\theta'}^n(\mathbf{y}_1^n). \tag{2.5}$$

(A3): If θ, θ' in $\prod^k$, $1 \leq k \leq k_0$, are such that P_θ and $P_{\theta'}$ are not equal measures on $\mathscr{X}^\infty$, then $D_{\theta,\theta'}(\mathbf{Y}) > 0$ P_θ − a.s.

Note that assumption (A3) enables a separation of the "true model" of the observed data from an underparametrized model, as is shown in Corollary 2.4 below.

Remark: In view of (A1), we can extend the definition of $D_{\theta,\theta'}$ in (2.5) to the case θ in $\prod^k$, θ' in $\prod^{k'}$, where $k \neq k'$. Furthermore, for $1 \leq k, k' \leq k_0$, θ in $\prod^k$, θ' in $\prod^{k'}$, and $\mathbf{y}$ in $\mathscr{X}^\infty$, $D_{\theta,\theta'}(\mathbf{y})$ is continuous in θ' by virtue of the assumed equicontinuity of $L_{\theta,\theta'}^n(\mathbf{y}_1^n)$.

Corollary 2.4: If θ in $\prod^k$, $1 \leq k \leq k_0$, is such that there exists no θ' in $\prod^{k'}$, $1 \leq k' < k$, for which P_θ and $P_{\theta'}$ are equal, then $\inf_{\theta' \in \prod^{k'}} D_{\theta,\theta'}(\mathbf{Y}) = \min_{\theta' \in \prod^{k'}} D_{\theta,\theta'}(\mathbf{Y}) > 0$ P_θ − a.s.

For each $1 \leq k \leq k_0$, we define the *maximum likelihood estimate* of the parameter θ in $\prod^k$, given the observation sequence $\mathbf{y}_1^n = (y_1, \cdots, y_n)$ in $\mathscr{X}^n$ by

$$\hat{\theta}_k(\mathbf{y}_1^n) = \arg\max_{\theta \in \prod^k} \log P_\theta(\mathbf{y}_1^n)$$

for $n \geq 1$; observe that the previous maximum exists by virtue of assumptions (A1) and (A2). For convenience, we shall hereafter use the abbreviated notation $\hat{\theta}_k^n$ for $\hat{\theta}_k(\mathbf{y}_1^n)$.

Lemma 2.5: For $1 \leq k, k' \leq k_0$, θ in $\prod^k$, we have

$$\liminf_n L_{\theta,\hat{\theta}_k^n}^n(\mathbf{y}_1^n) \geq \min_{\theta' \in \prod^{k'}} D_{\theta,\theta'}(\mathbf{y})$$

for all $\mathbf{y}$ in $\mathscr{X}^\infty$.

Proof: Fix $\epsilon > 0$ and $\mathbf{y}$ in $\mathscr{X}^\infty$. For each θ' in $\prod^{k'}$, let $\delta(\epsilon, \theta', \mathbf{y})$ and $N(\epsilon, \theta', \mathbf{y})$ be chosen as in assumption (A2). Let $O(\theta') \triangleq \{\tilde{\theta} : \tilde{\theta} \in \prod^{k'},\ d(\theta', \tilde{\theta}) < \delta(\epsilon, \theta', \mathbf{y})\}$. Clearly, $\{O(\theta')\}_{\theta' \in \prod^{k'}}$ is an open cover for $\prod^{k'}$; and by the compactness assumption (A1), there exists a finite subcover $\{O(\theta_i')\}_{i=1}^l$ for $\prod^{k'}$. Hence, there exists an element of $\{\theta_i'\}_{i=1}^l$, say θ_j' (where j may depend on n), such that

$$d\left(\hat{\theta}_{k'}^n(\mathbf{y}), \theta_j'\right) < \delta(\epsilon, \theta_j', \mathbf{y}).$$

By (A2), for $n \geq \max_{1 \leq i \leq l} N(\epsilon, \theta_i', \mathbf{y})$, we have

$$\left| L_{\theta,\hat{\theta}_k^n}^n(\mathbf{y}_1^n) - L_{\theta,\theta_j'}^n(\mathbf{y}_1^n) \right| \leq \epsilon$$

whence

$$\begin{aligned} L_{\theta,\hat{\theta}_k^n}^n(\mathbf{y}_1^n) &\geq L_{\theta,\theta_j'}^n(\mathbf{y}_1^n) - \epsilon \\ &\geq \min_{1 \leq i \leq l} L_{\theta,\theta_i'}^n(\mathbf{y}_1^n) - \epsilon. \end{aligned}$$

Thus,

$$\begin{aligned} \liminf_n L_{\theta,\hat{\theta}_k^n}^n(\mathbf{y}_1^n) &\geq \liminf_n \min_{1 \leq i \leq l} L_{\theta,\theta_i'}^n(\mathbf{y}_1^n) - \epsilon \\ &= \min_{1 \leq i \leq l} D_{\theta,\theta_i'}(\mathbf{y}) - \epsilon \\ &\geq \min_{\theta' \in \prod^{k'}} D_{\theta,\theta'}(\mathbf{y}) - \epsilon \end{aligned}$$

and since $\epsilon > 0$ is arbitrary, the assertion of the lemma follows. □

Assumptions (A1)–(A3) above are satisfied by certain classes of (stationary ergodic) Markov processes and (stationary ergodic) hidden Markov sources, as illustrated by Examples 1 and 2 below. Before proceeding to the examples, we present a technical lemma for checking the validity of assumption (A2) for the Markov processes of Example 1.

Lemma 2.6: Let $\{X_n\}_{n=1}^\infty$, $\{Y_n\}_{n=1}^\infty$ be two $\mathscr{X}$-valued first-order, Markov processes generated, respectively, by the $q \times q$-stochastic matrices $A \triangleq \{a_{ij}\}$, $A' \triangleq \{a_{ij}'\}$, both with positive entries. Given $\epsilon > 0$, there exists an integer N such that for every $\mathbf{x}_1^n$ in $\mathscr{X}^n$

$$\left| \frac{1}{n} \log \frac{P(\mathbf{X}_1^n = \mathbf{x}_1^n)}{P(\mathbf{Y}_1^n = \mathbf{x}_1^n)} \right| \leq D(A, A') + \epsilon$$

for all $n \geq N$, where $D(A, A') \triangleq \sum_{i=1,j=1}^q |\log a_{ij}/a_{ij}'|$.

Proof: Let $N(i, j|\mathbf{x}_1^n)$ denote the number of transitions from symbol i to symbol j in the sequence $\mathbf{x}_1^n$. By the assumption that $a_{ij} > 0$, $a_{ij}' > 0$, $1 \leq i, j \leq q$, the processes $\{X_n\}_{n=1}^\infty$ and $\{Y_n\}_{n=1}^\infty$ have unique steady-state

distributions, say π_X and π_Y, respectively, with all probabilities positive. We then have

$$\begin{aligned}\frac{1}{n}\log\frac{P(\mathbf{X}_1^n=\mathbf{x}_1^n)}{P(\mathbf{Y}_1^n=\mathbf{x}_1^n)} &= \frac{1}{n}\left[\log\frac{\pi_X(x_1)}{\pi_Y(x_1)}+\sum_{i=1,\,j=1}^{q}N(i,j|\mathbf{x}_1^n)\log\frac{a_{ij}}{a'_{ij}}\right]\\ &\le \frac{1}{n}\log\frac{\pi_X(x_1)}{\pi_Y(x_1)}+\frac{1}{n}\sum_{i=1,\,j=1}^{q}N(i,j|\mathbf{x}_1^n)\left|\log\frac{a_{ij}}{a'_{ij}}\right|\\ &\le \epsilon + D(A,A')\end{aligned}$$

for $n \ge N_1$, where $N_1 \triangleq 1/\epsilon \max_{x\in\mathscr{X}}|\log \pi_X(x)/\pi_Y(x)|$.

Similarly, we can show

$$\frac{1}{n}\log\frac{P(\mathbf{Y}_1^n=\mathbf{x}_1^n)}{P(\mathbf{X}_1^n=\mathbf{x}_1^n)} \le \epsilon + D(A,A')$$

for $n \ge N_2 \triangleq 1/\epsilon \max_{x\in\mathscr{X}}|\log \pi_Y(x)/\pi_X(x)|$. Setting $N = \max\{N_1, N_2\}$, the proof is completed. □

Example 1: Let $\prod^k$, $0 \le k \le k_0$, be the set of all $q^k \times q$-stochastic matrices $A_k = \{a_{ij}\}$ with entries satisfying $a_{ij} \ge \delta > 0$, where δ is a suitably prespecified constant. Let $\{P_k\}_{k=1}^{k_0}$ be the corresponding sequence of families of $\mathscr{X}$-valued homogeneous stationary ergodic Markov processes, with P_k representing all Markov measures of order k, $0 \le k \le k_0$. Assumption (A1) is readily seen to be satisfied. Assumption (A2) holds by virtue of Lemma 2.6 and the compactness of $\prod^k$, $0 \le k \le k_0$. Finally, the limit infimum defining $D_{\theta,\theta'}(\mathbf{Y})$ in (A3) is, by the law of large numbers, indeed a limit that equals a positive constant P_θ − a.s.

Example 2: Consider the HMS of (2.1)–(2.2), with $|\mathscr{S}| = k$. Fix $\delta > 0$ and for each k, $1 \le k \le k_0$, let $\prod^k$ be chosen to be the set $\Theta_{\delta_k}^k$, where $\delta_k = 2^{-k}\delta$. Lemma 2.2 then guarantees the embedding $\prod^k \subseteq \prod^{k+1}$ of assumption (A1). Assumption (A3) is true by virtue of Lemma 2.3. Finally, note that $[(S_n, X_n)]_{n=1}^{\infty}$ is a Markov process, and since $P_\theta(\mathbf{X}_1^n = \mathbf{y}_1^n) = \sum_{\mathbf{s}_1^n\in\mathscr{S}^n} P_\theta(\mathbf{X}_1^n = \mathbf{y}_1^n, \mathbf{S}_1^n = \mathbf{s}_1^n)$, a result analogous to Lemma 2.6 can be derived by which (A2) is readily verified.

III. Order Estimation of an HMS via Coding of Mixture Distributions

In this section, we present an estimator of the order of an HMS based on the method of coding of mixture distributions introduced in [3], [6], [27], among other literature, in the context of universal data compression. This technique directly yields an estimate of the said order—our real objective—rather than involving also the estimation of the parameter θ in the appropriate parametric family.

We begin with two technical lemmas for the HMS of (2.1)–(2.2), the first of which is an analog of Lemma 2.5, but without the compactness assumption (A1).

Lemma 3.1: For $1 \le k, k' \le k_0$, θ in Θ_0^k, we have

$$\lim_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\theta'\in\Theta_0^{k'}}P_{\theta'}(\mathbf{X}_1^n)} = \inf_{\theta'\in\Theta_0^{k'}} D(P_\theta\|P_{\theta'})\ P_\theta - \text{a.s.}$$

Proof: Fix k, k', $1 \le k, k' \le k_0$, and θ in Θ_0^k. Then, for every θ' in $\Theta_0^{k'}$

$$\limsup_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\tilde\theta\in\Theta_0^{k'}}P_{\tilde\theta}(\mathbf{X}_1^n)} \le \lim_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{P_{\theta'}(\mathbf{X}_1^n)} = D(P_\theta\|P_{\theta'})\ P_\theta - \text{a.s.}$$

so that

$$\limsup_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\theta'\in\Theta_0^{k'}}P_{\theta'}(\mathbf{X}_1^n)} \le \inf_{\theta'\in\Theta_0^{k'}} D(P_\theta\|P_{\theta'})\ P_\theta - \text{a.s.}$$

For every $\delta > 0$ and θ' in $\Theta_0^{k'}$, we can obtian a modified θ'_δ in $\Theta_\delta^{k'}$ by suitably using the maximum entry in the associated stochastic matrices to compensate for those entries less than δ. Correspondingly, $\max_{x\in\mathscr{X}} P_\theta(x|s) = r \ge 1/q$ is reduced by a factor no larger than $r - |\mathscr{X}|\delta/r \ge 1 - q^2\delta$. Similarly, $\max_{s\in\mathscr{S}} P_{\theta'}(s|s')$ is reduced by a factor no larger than $1 - k'^2\delta$. Therefore,

$$\begin{aligned}&\frac{1}{n}\log\max_{\theta'\in\Theta_\delta^{k'}}P_{\theta'}(\mathbf{X}_1^n)\\ &\quad\ge \frac{1}{n}\log\sup_{\theta'\in\Theta_0^{k'}}P_{\theta'}(\mathbf{X}_1^n)(1-q^2\delta)^n(1-k'^2\delta)^n\\ &\quad= \frac{1}{n}\log\sup_{\theta'\in\Theta_0^{k'}}P_{\theta'}(\mathbf{X}_1^n) + \log(1-q^2\delta)(1-k'^2\delta).\end{aligned} \tag{3.1}$$

Note that $-\log(1-q^2\delta)(1-k'^2\delta)$ decreases to 0 with δ. Hence, given any $\epsilon > 0$ there exists $\delta > 0$—for instance, δ satisfying $\epsilon \le -\log(1-q^2\delta)(1-k'^2\delta)$—with

$$\begin{aligned}\liminf_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\theta'\in\Theta_0^{k'}}P_{\theta'}(\mathbf{X}_1^n)} &\ge \liminf_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{\max_{\theta'\in\Theta_\delta^{k'}}P_{\theta'}(\mathbf{X}_1^n)} - \epsilon\\ &\ge \min_{\theta'\in\Theta_\delta^{k'}} D_{\theta,\theta'}(\mathbf{X}) - \epsilon\end{aligned} \tag{3.2}$$

by Lemma 2.5 since $\Theta_\delta^{k'}$ is compact. Since $D_{\theta,\theta'}(\mathbf{X}) = D(P_\theta\|P_{\theta'})\ P_\theta$ − a.s. by (2.4) and (2.5), it follows from (3.2) that

$$\liminf_n \frac{1}{n}\log\frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\theta'\in\Theta_0^{k'}}P_{\theta'}(\mathbf{X}_1^n)} \ge \inf_{\theta'\in\Theta_0^{k'}} D(P_\theta\|P_{\theta'}) - \epsilon \qquad P_\theta - \text{a.s.}$$

Since ϵ was chosen arbitrarily, our proof is completed. □

Lemma 3.2: For (k,θ) in $\mathscr{M}$ and $k' < k$, it holds that $\inf_{\theta'\in\Theta_0^{k'}} D(P_\theta\|P_{\theta'}) > 0$.

Proof: Pick θ' in $\Theta_0^{k'}$. By Lemma 2.3,

$$D(P_\theta \| P_{\theta'}) = \lim_n \frac{1}{n} \log \frac{P_\theta(\mathbf{X}_1^n)}{P_{\theta'}(\mathbf{X}_1^n)} = \lim_n \frac{1}{nl} \log \frac{P_\theta(\mathbf{X}_1^{nl})}{P_{\theta'}(\mathbf{X}_1^{nl})} = \lim_n \frac{1}{nl} E_{P_\theta}\left[\log \frac{P_\theta(\mathbf{X}_1^{nl})}{P_{\theta'}(\mathbf{X}_1^{nl})}\right]$$

for every positive integer l, where the last inequality follows from Finesso [7, Theorem 2.3.3].

Thus,

$$D(P_\theta \| P_{\theta'}) \geq \frac{1}{l} E_{P_\theta}\left[\log P_\theta(\mathbf{X}_1^l)\right] - \lim_n \frac{1}{nl} E_{P_\theta}\left[\log P_{\theta'}(\mathbf{X}_1^{nl})\right] \quad (3.3)$$

since $1/nE_{P_\theta}[\log P_\theta(\mathbf{X}_1^n)]$ is nondecreasing in n [10, Theorem 3.5.1]. Next, by [20, Proposition 2.7], for positive integers m, n, we have that

$$E_{P_\theta}[\log P_{\theta'}(\mathbf{X}_1^{n+m})] \leq E_{P_\theta}[\log P_{\theta'}(\mathbf{X}_1^n)] + E_{P_\theta}[\log P_{\theta'}(\mathbf{X}_1^m)].$$

Consequently,

$$\lim_n \frac{1}{nl} E_{P_\theta}\left[\log P_{\theta'}(\mathbf{X}_1^{nl})\right] \leq \lim_n \frac{1}{nl} \sum_{i=0}^{n-1} E_{P_\theta}\left[\log P_{\theta'}(\mathbf{X}_{il+1}^{il+l})\right] = \frac{1}{l} E_{P_\theta}\left[\log P_{\theta'}(\mathbf{X}_1^l)\right],$$

which, combined with (3.3), yields that

$$D(P_\theta \| P_{\theta'}) \geq \frac{1}{l} E_{P_\theta}\left[\log \frac{P_\theta(\mathbf{X}_1^l)}{P_{\theta'}(\mathbf{X}_1^l)}\right] \quad (3.4)$$

for every positive integer l.

Finally, observe that $P_{\theta'}$ equals $P_{\tilde{\theta}}$ for some $\tilde{\theta}$ in Θ_0^k; however, since (k, θ) belongs to $\mathcal{M}$ and $k' < k$, $P_{\tilde{\theta}}$ is not equal to P_θ. From [18, Theorem 3.2, pp. 61–62], if θ_1, θ_2 are in Θ_0^k, then P_{θ_1} and P_{θ_2} are equal iff for $l \geq 2k$, it holds that $P_{\theta_1}(\mathbf{x}_1^l) = P_{\theta_2}(\mathbf{x}_1^l)$ for all $\mathbf{x}_1^l$ in $\mathcal{X}^l$. Thus, with P_θ and $P_{\theta'}$ (and hence $P_{\tilde{\theta}}$) not being equal, if follows that $E_{P_\theta}[\log P_\theta(\mathbf{X}_1^l)/P_{\theta'}(\mathbf{X}_1^l)] > 0$ for every θ' in $\Theta_0^{k'}$ and $l \geq 2k$. The assertion of the lemma then follows by the semicontinuity of $E_{P_\theta}[\log P_\theta(\mathbf{X}_1^l)/P_{\theta'}(\mathbf{X}_1^l)]$ on $\Theta^{k'}$.

We now introduce the notion of a *mixture distribution* on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$ (cf. Csiszár [3], Davisson *et al.* [6], Shtar'kov [27]), together with some pertinent properties. Let ν_k be a prior distribution on Θ_0^k, $1 \leq k \leq k_0$. The conditional mixture distribution $Q_k(\cdot|s_0)$ on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$, conditioned on an initial state s_0 in $\mathcal{S}$, is defined by

$$Q_k(A|s_0) \triangleq \int_{\Theta_0^k} P_\theta(A|s_0) \nu_k(d\theta)$$

for all A in $\mathcal{B}^\infty$. The mixture distribution $Q_k(\cdot)$ on $(\mathcal{X}^\infty, \mathcal{B}^\infty)$ is then defined by

$$Q_k(A) \triangleq \frac{1}{k} \sum_{s_0 \in \mathcal{S}} Q_k(A|s_0) \quad (3.5)$$

for all A in $\mathcal{B}^\infty$. (The assumption of equiprobable initial states in (3.5) above is convenient, but not necessary. For our purposes, any initial distribution on $\mathcal{S}$ will suffice, which assigns positive mass to every state in $\mathcal{S}$.) For a finite sequence $\mathbf{x}_1^n$ in $\mathcal{X}^n$, the probability $Q_k(\mathbf{x}_1^n)$ is formally the Q_k-measure of the set of all infinite sequences in $\mathcal{X}^\infty$ whose initial segment is $\mathbf{x}_1^n$.

The observed sequence $\mathbf{x}_1^n$ in $\mathcal{X}^n$ can be encoded with respect to the mixture distribution Q_k by a Shannon–Fano (prefix) code (cf. [4, pp. 61–65]) of length $\log 1/Q_k(\mathbf{x}_1^n)$ bits. If P_θ, θ in Θ_0^k, is the "true" distribution generating the observations, then the *pointwise coding redundancy* (up to 1 bit) is

$$R_{P_\theta}(\mathbf{x}_1^n; Q_k) \triangleq \log \frac{1}{Q_k(\mathbf{x}_1^n)} - \log \frac{1}{P_\theta(\mathbf{x}_1^n)} = \log \frac{P_\theta(\mathbf{x}_1^n)}{Q_k(\mathbf{x}_1^n)}.$$

In general, we can define the pointwise coding redundancy for an uniquely decodable code as follows. Consider any UD code for encoding sequences from $\mathcal{X}^n$; without any loss of essential generality, we can assume [3] that the code satisfies Kraft's inequality with equality, and hence is a Shannon–Fano code with respect to some probability distribution Q (not necessarily of the mixture type) on $\mathcal{X}^n$. The pointwise coding redundancy of this code relative to P_θ (θ in Θ_0^k, $1 \leq k \leq k_0$) is defined as

$$R_{P_\theta}(\mathbf{x}_1^n; Q) \triangleq \log \frac{P_\theta(\mathbf{x}_1^n)}{Q(\mathbf{x}_1^n)} \quad (3.6)$$

for $\mathbf{x}_1^n$ in $\mathcal{X}^n$. The pointwise coding redundancy, relative to P_θ, of a Shannon-Fano code on $\mathcal{X}^n$ with respect to the mixture distribution Q_k will then be denoted, as earlier, by $R_{P_\theta}(\mathbf{x}_1^n; Q_k)$, where $\mathbf{x}_1^n$ is in $\mathcal{X}^n$.

It is clear that the *average redundance* of a uniquely decodable code Q on $\mathcal{X}^n$ relative to P_θ, namely $E_{P_\theta}[R_{P_\theta}(\mathbf{X}_1^n; Q)]$ is nonnegative; however, $R_{P_\theta}(\mathbf{x}_1^n; Q)$ could be negative for some $\mathbf{x}_1^n$ in $\mathcal{X}^n$. The next lemma, due to Barron [1] and stated here without proof, asserts that $R_{P_\theta}(\mathbf{X}_1^n; Q)$ is essentially nonnegative for all large n.

Lemma 3.3 (Barron [1, Theorem 3.1]): Let (k, θ) belong to $\mathcal{M}$. For each k' and mixture distribution $Q_{k'}$, $1 \leq k' \leq k_0$, it holds that $R_{P_\theta}(\mathbf{X}_1^n; Q_{k'}) \geq -2 \log n$ eventually P_θ – a.s.[1]

[1]Given a sequence of $\mathbb{R}$-valued random variables $\{Z_n\}_{n=1}^\infty$ and a $\mathbb{R}$-valued sequence $\{a_n\}_{n=1}^\infty$, we say that $Z_n \geq a_n$ eventually a.s. if there exists a $\mathbb{R}$-valued random variable $N = N(\omega)$, which is infinite a.s., and $Z_n \geq a_n$ for all $n \geq N$.

Typically, the pointwise redundancy of a code constructed in ignorance of the "true" distribution P_θ is not only essentially nonnegative, but increases with n to infinity; a good code is one for which this redundancy increases slowly with n. The following lemma, due to Csiszár [3], establishes the existence of such a good code based on a mixture distribution.

Lemma 3.4 (Csiszár [3]): For each k, $1 \le k \le k_0$, there exists a prior distribution ν_k on Θ_0^k such that the corresponding mixture distribution $Q_k(\mathbf{x}_1^n) = 1/k \sum_{s_0 \in \mathcal{S}} \int_{\Theta_0^k} P_\theta(\mathbf{x}_1^n | s_0) \nu_k(d\theta)$ satisfies

$$\log \frac{\sup_{\theta \in \Theta_0^k} P_\theta(\mathbf{x}_1^n)}{Q_k(\mathbf{x}_1^n)} \le \frac{k(k+q-2)}{2} \log n + c_{k,q}$$

for all $\mathbf{x}_1^n$ in $\mathcal{X}^n$, $n \ge N'(k,q)$, where $c_{k,q}$ is a constant depending only on k, q.

Remark: The pointwise coding redundancy of Lemma 3.4 is asymptotically optimal in the following sense. Consider *any* uniquely decodable code Q on $\mathcal{X}^n$. Suppose that we weaken the requirement of a uniformly small pointwise redundancy (i.e., for every $\mathbf{x}_1^n$ in $\mathcal{X}^n$) to that of a small average redundancy, viz., $E_{P_\theta}[R_{P_\theta}(\mathbf{X}_1^n; Q)]$, where θ belongs to Θ_0^k, $1 \le k \le k_0$. Then it follows from Rissanen [22, Theorem 3.1, p. 71)] together with Baum and Petrie [2, Theorem, p. 1562] that the previous average redundancy is, in effect, bounded *below* for all large n by $[k(k+q-2)/2] \log n - A$, where $A = A(k)$ does not depend on n.

Csiszár's proof [3] of Lemma 3.4 above relies on a specific construction of the mixture distribution Q_k using a Dirichlet density as a prior, and is similar to that of Shtar'kov [27] for a mixture of Markov processes. This construction will play an explicit role in Section IV, in the universal coding of the HMS; Csiszár's proof of Lemma 3.4 is, therefore, reproduced in the Appendix. *Hereafter, by mixture distributions Q_k, $1 \le k \le k_0$, we shall refer solely to those constructed in the Appendix.*

Lemmas 3.3 and 3.4 above provide the necessary tools for constructing our estimator of the order of the HMS as follows. Given an observed sequence $\mathbf{x}_1^n$ in $\mathcal{X}^n$, the order estimator $\hat{k}_n$ is defined by

$$\hat{k}_n(\mathbf{x}_1^n) \triangleq \max\left\{1 \le k' \le k_0 : \log Q_{k'}(\mathbf{x}_1^n) - \log Q_{k'-1}(\mathbf{x}_1^n) > \frac{k'(k'+q-2)+5}{2} \log n\right\} \quad (3.7)$$

with the convention $Q_0(\mathbf{x}_1^n) = 1$ for all $\mathbf{x}_1^n$ in $\mathcal{X}^n$. If the set above is empty, we set $\hat{k}_n(\mathbf{x}_1^n) = 1$.

The (strong) consistency of the previous estimator is established by the following

Proposition 3.5: For each (k, θ) in $\mathcal{M}$, $\lim_n \hat{k}_n(\mathbf{X}_1^n) = k$ P_θ − a.s.

Proof: Fix (k, θ) in $\mathcal{M}$ and pick $k' \ge k$. Then, by Lemma 3.3,

$$-2 \log n + \log Q_{k'+1}(\mathbf{X}_1^n) \le \log P_\theta(\mathbf{X}_1^n) \quad (3.8)$$

eventually P_θ − a.s. Also, by Lemma 3.4,

$$\log \frac{P_\theta(\mathbf{X}_1^n)}{Q_{k'}(\mathbf{X}_1^n)} \le \log \frac{\sup_{\tilde{\theta} \in \Theta_0^{k'}} P_{\tilde{\theta}}(\mathbf{X}_1^n)}{Q_{k'}(\mathbf{X}_1^n)} \le \frac{k'(k'+q-2)+1}{2} \log n$$

for all $n \ge N'(k', q)$, so that

$$\log P_\theta(\mathbf{X}_1^n) \le \log Q_{k'}(\mathbf{X}_1^n) + \frac{k'(k'+q-2)+1}{2} \log n \quad (3.9)$$

for all $n \ge N'(k', q)$.

By combining (3.8) and (3.9) and eliminating $\log P_\theta(\mathbf{X}_1^n)$, we get that $\log Q_{k'+1}(\mathbf{X}_1^n) - \log Q_{k'}(\mathbf{X}_1^n) \le [(k'(k'+q-2)+5)/2] \log n$ eventually P_θ − a.s. Hence, $\limsup \hat{k}_n(\mathbf{X}_1^n) \le k$ P_θ − a.s.

The proof is completed by establishing that $\liminf_n \hat{k}_n(\mathbf{X}_1^n) \ge k$ P_θ − a.s. To this end, it suffices to show that $\liminf_n 1/n \log Q_k(\mathbf{X}_1^n)/Q_{k-1}(\mathbf{X}_1^n) > 0$ P_θ − a.s. It can be shown that

$$\log \frac{\sup_{\theta \in \Theta_0^k} P_\theta(\mathbf{x}_1^n)}{Q_k(\mathbf{x}_1^n)} \ge 0$$

for $k = 1, \cdots, k_0$, and for all $\mathbf{x}_1^n$ in $\mathcal{X}^n$ and for all $n \ge \tilde{N}(k)$. Then

$$\begin{aligned}
&\liminf_n \frac{1}{n} \log \frac{Q_k(\mathbf{X}_1^n)}{Q_{k-1}(\mathbf{X}_1^n)} \\
&\quad = \liminf_n \left[\frac{1}{n} \log \frac{Q_k(\mathbf{X}_1^n)}{P_\theta(\mathbf{X}_1^n)} + \frac{1}{n} \log \frac{P_\theta(\mathbf{X}_1^n)}{Q_{k-1}(\mathbf{X}_1^n)}\right] \\
&\quad \ge \lim_n \frac{1}{n} \log \frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\tilde{\theta} \in \Theta_0^{k-1}} P_{\tilde{\theta}}(\mathbf{X}_1^n)} \\
&\qquad P_\theta \text{ − a.s., by Lemma 3.1} \\
&\quad = \inf_{\tilde{\theta} \in \Theta_0^{k-1}} D(P_\theta \| P_{\tilde{\theta}}) \quad P_\theta \text{ − a.s., by Lemma 3.1} \\
&\quad > 0 \quad P_\theta \text{ − a.s., by Lemma 3.2}
\end{aligned}$$

thereby completing the proof. □

IV. Universal Data Compression of an HMS

In this section we address the problem of universal data compression, in a uniquely decodable manner, of an HMS $\{X_n\}_{n=1}^\infty$ of unknown order k, $1 \le k \le k_0$; the codes considered will be shown to be asymptotically optimal in a suitable sense. Note that if the order of the HMS is *known to the encoder* but not the decoder, say $k = \tilde{k}$ (but the $k \times k$ − and $k \times q$ − stochastic matrices A and B (cf. sect. 2) are still *unknown to both*), such an universal code is easily obtained. One possible method is based on Rissanen's minimum description length (MDL) principle [22, sect. 3.6]. Given an observed sequence $\mathbf{x}_1^n$ in $\mathcal{X}^n$, consider a code consisting of a two-stage description of $\mathbf{x}_1^n$ within the given parametric family $\{\Theta_0^k\}_{k=1}^{k_0}$. Such a description comprises a Chaitin (prefix) code [22, sect. 2.2.4]

for the (known) HMS order $\tilde{k}$, of length $L(\tilde{k})$ bits (where $L(\tilde{k}) \le \log k_0 + 2\log\log k_0 + 3$), concatenated with a Shannon–Fano code for $\mathbf{x}_1^n$ with respect to the mixture distribution $Q_{\tilde{k}}$ on $\mathscr{X}^n$. Clearly, this code of length $L(\tilde{k}) + \log 1/Q_{\tilde{k}}(\mathbf{x}_1^n)$ bits will asymptotically possess the minimum pointwise redundancy among all UD universal codes for the HMS of order $\tilde{k}$.

When the HMS order k, $1 \le k \le k_0$, is unknown to both encoder and decoder, Rissanen's scheme above can be modified to encode the observed sequence $\mathbf{x}_1^n$ in an asymptotically optimal manner. This is done by replacing $\tilde{k}$ by the MDL estimate $\hat{k}_n^{\mathrm{MDL}}(\mathbf{x}_1^n)$ of HMS order, where $\hat{k}_n^{\mathrm{MDL}}(\mathbf{x}_1^n)$ is the value of k minimizing the length (in bits) of the two-stage description of $\mathbf{x}_1^n$, viz.,

$$\hat{k}_n^{\mathrm{MDL}}(\mathbf{x}_1^n) \triangleq \arg \min_{1 \le k \le k_0} \left[L(k) + \log \frac{1}{Q_k(\mathbf{x}_1^n)} \right].$$

The following proposition is a simple consequence of Lemma 3.4.

Proposition 4.1: For each (k, θ) in $\mathscr{M}$, the pointwise redundancy of the two-stage code satisfies

$$L(k_n^{\mathrm{MDL}}(\mathbf{x}_1^n)) + \log \frac{1}{Q_{\hat{k}_n^{\mathrm{MDL}}(\mathbf{x}_1^n)}} - \log \frac{1}{P_\theta(\mathbf{x}_1^n)} \le \frac{k_0(k_0+q-2)}{2} \log n + d_{k_0,q}$$

for all $\mathbf{x}_1^n$ in $\mathscr{X}^n$, $n \ge N'(k_0, q)$, where $d_{k_0,q}$ is a constant depending only on k_0, q.

The previous uniquely decodable code for an HMS of unknown order asymptotically achieves, by Proposition 4.1, minimum pointwise redundancy. It is handicapped in a practical sense, however, by delays in encoding and decoding incurred by these operations being performed on blocks of symbols, rather than sequentially on individual symbols. We present below a *sequential code* (SC) for the HMS, which is similar to that used in [28] to encode a unifilar source. Our SC employs a first-in first-out arithmetic code (cf., e.g., [12], [26]) in conjunction with the order estimate in (3.7), and is uniquely decodable. It avoids the aforementioned delays[2] at the possible expense of pointwise asymptotic optimal redundancy. We shall show that (SC) is, however, asymptotically optimal in the sense of achieving minimum *average* redundancy.

Sequential Code (SC): Given the observed sequence $\{x_n\}$, the encoding proceeds as follows.

- Encode the first symbol x_1 by an arithmetic code with respect to the probability value $1/q$.
- Encode the $(n+1)$th symbol $\mathbf{x}_{n+1}$ by an arithmetic code with respect to the conditional probability $Q_{\hat{k}_n(\mathbf{x}_1^n)}(x_{n+1} \mid \mathbf{x}_1^n)$, $n \ge 1$ (cf. (A.27) of Appendix for the computation of the mixture probabilities).

[2]An arithmetic code (and hence SC), unlike a prefix code, need not allow instantaneous decoding. However, for the encoding and decoding of a symbol, only a few adjacent symbols are needed [13].

The decoder, having correctly decoded the received sequence to retrieve $\mathbf{x}_1^n$, can determine $\hat{k}_n(\mathbf{x}_1^n)$, $n \ge 1$, in exactly the same manner as the encoder. This fact, together with the unique decodability of an arithmetic code, renders SC uniquely decodable.

Remarks:

i) As indicated in [28], the finite arithmetic precision employed by arithmetic coding introduces significant redundancy in SC, especially when encoding long observed sequences. Consequently, SC will asymptotically achieve optimal redundancy not in the pointwise sense, but rather in the average sense as shown below in Proposition 4.1.

ii) In order to asymptotically achieve average optimal redundancy, the order estimator $\hat{k}_n$ of (3.7) in SC can be replaced by any other estimator whose probability of incorrect estimation decays to zero rapidly enough with n. This is seen in the proof of Proposition 4.1 below.

Let $L_{\mathrm{SC}}(\mathbf{x}_1^n)$ be the length of the codeword when $\mathbf{x}_1^n$ is encoded using SC, $n \ge 1$. With an abuse of notation, let $L_{\mathrm{SC}}(x_i)$, $i = 1, \cdots, n$, be the length of the corresponding codeword for symbol x_i.

Proposition 4.1: For every (k, θ) in $\mathscr{M}$, the average redundancy of SC is bounded above according to

$$E_{P_\theta}\left[L_{\mathrm{SC}}(\mathbf{X}_1^n) - \log \frac{1}{P_\theta(\mathbf{X}_1^n)} \right] \le \frac{k(k+q-2)}{2} \log n + e$$

for all n large enough, where $e = e(k_0)$ is a constant.

The proof of Proposition 4.1 relies on two technical lemmas establishing upper bounds on the probabilities of overestimation and underestimation of the HMS order estimator of (3.7). We state below these Lemmas 4.2 and 4.3, followed by the proof of Proposition 4.1. This section then concludes with the proofs of Lemmas 4.2 and 4.3.

Lemma 4.2 (Probability of Overestimation): For every (k, θ) in $\mathscr{M}$, the order estimator $\hat{k}_n$ of (3.7) satisfies

$$P_\theta\left(\hat{k}_n(\mathbf{X}_1^n) > k\right) \le k_0 \left(\max_{1 \le k' \le k_0} 2^{c_{k',q}} \right) n^{-3} \qquad P_\theta - \text{a.s.}$$

for all n large enough.

Lemma 4.3 (Probability of Underestimation): For every (k, θ) in $\mathscr{M}$, there exists $\lambda > 0$ such that

$$\limsup_n \frac{1}{n} \log P_\theta\left(\hat{k}_n(\mathbf{X}_1^n) < k\right) \le -\lambda.$$

Proof of Proposition 4.1: Fix (k, θ) in $\mathscr{M}$. Observe that

$$E_{P_\theta}\left[L_{\mathrm{SC}}(\mathbf{X}_1^n) - \log \frac{1}{P_\theta(\mathbf{X}_1^n)} \right] = E_{P_\theta}\left[L_{\mathrm{SC}}(\mathbf{X}_1^n) - \log \frac{1}{Q_k(\mathbf{X}_1^n)} \right] + E_{P_\theta}\left[\log \frac{1}{Q_k(\mathbf{X}_1^n)} - \log \frac{1}{P_\theta(\mathbf{X}_1^n)} \right].$$

By Lemma 3.4, the second term on the right-hand side is bounded by $(k(k+q-2)/2)\log n + c_{k,q}$ for all $n \geq N'(k,q)$. Hence, the proposition is established by showing that

$$E_{P_\theta}\left[L_{\mathrm{SC}}(\mathbf{X}_1^n) - \log \frac{1}{Q_k(\mathbf{X}_1^n)}\right] \leq K \tag{4.1}$$

for all n suitably large, where K is a constant. To this end, observe that $E_{P_\theta}[L_{\mathrm{SC}}(X_1)] = \log q$; further, for $i \geq 2$,

$$E_{P_\theta}[L_{\mathrm{SC}}(X_i)] = E_{P_\theta}\Big[L_{\mathrm{SC}}(X_i)\cdot \mathbf{1}\big(\hat{k}_{i-1}(\mathbf{X}_1^{i-1}) = k\big) + L_{\mathrm{SC}}(X_i)\cdot \mathbf{1}\big(\hat{k}_{i-1}(\mathbf{X}_1^{i-1}) \neq k\big)\Big]$$

where $\mathbf{1}(\cdot)$ denotes indicator function. By the construction of SC, the first term above does not exceed $E_{P_\theta}[\log 1/Q_k(X_i|\mathbf{X}_1^{i-1})]$, while in the second term

$$\begin{aligned} L_{\mathrm{SC}}(x_i) &= \log \frac{1}{Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(x_i|\mathbf{x}_1^{i-1})} \\ &= \log \frac{Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{x}_1^{i-1})}{Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{x}_1^{i})} \\ &= \log \frac{\sum_{\mathbf{s}_1^i} Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{s}_1^i) Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{x}_1^{i-1}|\mathbf{s}_1^i)}{\sum_{\mathbf{s}_1^i} Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{s}_1^i) Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{x}_1^{i} \mid \mathbf{s}_1^i)}. \end{aligned}$$

It is easily shown from (A.19) that

$$\frac{Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{x}_1^{i-1}|\mathbf{s}_1^{i-1})}{Q_{\hat{k}_{i-1}(\mathbf{x}_1^{i-1})}(\mathbf{x}_1^{i}|\mathbf{s}_1^{i})} \leq 1 + \log(q+i)$$

from which it follows that $E_{P_\theta}[L_{\mathrm{SC}}(X_i)] \leq 1 + \log(q+i)$. Thus, for $i \geq 2$,

$$E_{P_\theta}[L_{\mathrm{SC}}(X_i)] \leq E_{P_\theta}\left[\log \frac{1}{Q_k(X_i|\mathbf{X}_1^{i-1})}\right] + [1+\log(q+i)]P_\theta\big(\hat{k}_{i-1}(\mathbf{X}_1^{i-1}) \neq k\big)$$

so that from [26, Theorem 5]

$$\begin{aligned} &E_{P_\theta}[L_{\mathrm{SC}}(\mathbf{X}_1^n)] \\ &= E_{P_\theta}\left[\sum_{i=1}^n L_{\mathrm{SC}}(X_i)\right] + O\left(\frac{1}{n}\right) \\ &\leq \log q + E_{P_\theta}\left[\log \frac{1}{Q_k(X_1)} + \sum_{i=2}^n \log \frac{1}{Q_k(X_i|\mathbf{X}_1^{i-1})}\right] \\ &\quad + \sum_{i=1}^n [1+\log(q+i)]P_\theta\big(\hat{k}_{i-1}(\mathbf{X}_1^{i-1}) \neq k\big) + O\left(\frac{1}{n}\right) \\ &= \log q + E_{P_\theta}\left[\log \frac{1}{Q_k(\mathbf{X}_1^n)}\right] \\ &\quad + \sum_{i=1}^n [1+\log(q+i)]P_\theta\big(\hat{k}_{i-1}(\mathbf{X}_1^{i-1}) \neq k\big) + O\left(\frac{1}{n}\right). \end{aligned} \tag{4.2}$$

Hence,

$$\begin{aligned} &E_{P_\theta}\left[L_{\mathrm{SC}}(\mathbf{X}_1^n) - \log \frac{1}{Q_k(\mathbf{X}_1^n)}\right] \\ &\leq \log q + \sum_{i=1}^n [1+\log(q+i)]P_\theta\big(\hat{k}_{i-1}(\mathbf{X}_1^{i-1}) \neq k\big) \\ &\leq K \end{aligned}$$

for all n suitably large by virtue of Lemmas 4.2 and 4.3 above, where $K = K(k_0)$ is a constant.

This establishes (4.1) and, hence, the proposition. □

Proof of Lemma 4.2: Fix (k,θ) in $\mathcal{M}$ where $k < k_0$. For any k', $k < k' \leq k_0$, we have from (3.7) that

$$P_\theta\big(\hat{k}_n(\mathbf{X}_1^n) = k'\big) \leq \sum_{x_1^n \in \mathcal{X}^n} P_\theta(\mathbf{x}_1^n) \cdot \mathbf{1}\left(\log \frac{Q_{k'}(\mathbf{x}_1^n)}{Q_{k'-1}(\mathbf{x}_1^n)} > \frac{k'(k'+q-2)}{2}\log n\right). \tag{4.3}$$

By Lemma 3.4,

$$\log \frac{P_\theta(\mathbf{x}_1^n)}{Q_{k'-1}(\mathbf{x}_1^n)} \leq \frac{(k'-1)(k'-1+q-2)}{2}\cdot \log n + c_{k'-1,q}$$

for all $n \geq N'(k'-1,q)$, which, when substituted in (4.3), yields

$$\begin{aligned} &P_\theta\big(\hat{k}_n(\mathbf{X}_1^n) = k'\big) \\ &\leq \sum_{\mathbf{x}_1^n \in \mathcal{X}^n} \left(Q_{k'}(\mathbf{x}_1^n) \cdot n^{\frac{-k'(k'+q-2)+5}{2}}\right) n^{\frac{(k'-1)(k'+q-3)}{2}} 2^{c_{k'-1}} \\ &\leq 2^{c_{k'-1}} n^{-(k'+\frac{q}{2}+1)} \text{ since } \sum_{\mathbf{x}_1^n \in \mathcal{X}^n} Q_k(\mathbf{x}_1^n) = 1 \\ &\leq 2^{c_{k'-1}} n^{-3} \end{aligned}$$

for all $n \geq N'(k'-1,q)$. Consequently,

$$\begin{aligned} P_\theta\big(\hat{k}_n(\mathbf{X}_1^n) > k\big) &= \sum_{k'=k+1}^{k_0} P_\theta\big(\hat{k}_n(\mathbf{X}_1^n) = k'\big) \\ &\leq n^{-3} \sum_{k'=1}^{k_0} 2^{c_{k'-1}} \end{aligned}$$

for all n large enough, whence the assertion of the lemma follows. □

Proof of Lemma 4.3: Fix (k,θ) in $\mathcal{M}$. Let $a_k = (k(k+q-2)+5)/2$, and define

$$A_n \triangleq \{\mathbf{x}_1^n \in \mathcal{X}^n : \log Q_k(\mathbf{x}_1^n) - \log Q_{k-1}(\mathbf{x}_1^n) \leq a_k \log n\}.$$

Clearly

$$P_\theta\big(\hat{k}_n(\mathbf{X}_1^n) < k\big) \leq P_\theta(A_n). \tag{4.4}$$

Given any $\mathbf{x}_1^n$ in $\mathscr{X}^n$, it holds by Lemma 3.4 that $\log Q_k(\mathbf{x}_1^n) \geq \log P_\theta(\mathbf{x}_1^n) - b_k \log n - c_k$ for all $n \geq N'(k, q)$, where $b_k = k(k + q - 2)/2$; furthermore, $Q_{k-1}(\mathbf{x}_1^n) \leq \sup_{\theta' \in \Theta_0^{k-1}} P_{\theta'}(\mathbf{x}_1^n)$ for all $n \geq \tilde{N}(k-1)$. Hence, defining

$$B_n \triangleq \left\{ \mathbf{x}_1^n \in \mathscr{X}^n : \log P_\theta(\mathbf{x}_1^n) - \log \sup_{\theta' \in \Theta_0^{k-1}} P_{\theta'}(\mathbf{x}_1^n) \leq (a_k + b_k) \log n + c_k \right\}$$

it follows that $A_n \subseteq B_n$, so that by (4.4)

$$P_\theta\left(\hat{k}_n(\mathbf{X}_1^n) < k\right) \leq P_\theta(B_n) \tag{4.5}$$

for all $n \geq \max\{N'(k, q), \tilde{N}(k-1)\}$. Next, by (3.1), for any ϵ, $0 < \epsilon < \inf_{\theta' \in \Theta_0^{k-1}} D(P_\theta \| P_{\theta'})$, there exists $\delta = \delta(\epsilon, k, q) > 0$ such that

$$\frac{1}{n} \log \max_{\theta' \in \Theta_\delta^{k-1}} P_{\theta'}(\mathbf{x}_1^n) \geq \frac{1}{n} \log \sup_{\theta' \in \Theta_0^{k-1}} P_{\theta'}(\mathbf{x}_1^n) - \frac{\epsilon}{4} \tag{4.6}$$

for all $\mathbf{x}_1^n$ in $\mathscr{X}^n$.

Combining the compactness of Θ_δ^{k-1} with the fact that the HMS $\{X_n\}_{n=1}^\infty$ satisfies assumption (A2) (cf. Example 2 of Section II), we can find a finite cover $\{O(\theta_i)\}_{i=1}^m$ for Θ_δ^{k-1} and positive integers $\{N(\theta_i)\}_{i=1}^m$ with the following property holding for each $i = 1, \cdots, m$: for every θ' in $O(\theta_i)$ and for all $n \geq N(\theta_i)$, we have that

$$\left| \frac{1}{n} \log P_{\theta'}(\mathbf{x}_1^n) - \frac{1}{n} \log P_{\theta_i}(\mathbf{x}_1^n) \right| \leq \frac{\epsilon}{4}$$

for all $\mathbf{x}_1^n$ in $\mathscr{X}_1^n$ (cf. Lemma 2.6). In conjunction with (4.6), this implies for each $n \geq \max_{1 \leq i \leq m} N(\theta_i)$ that there exists $i^* = i^*(n)$ in $\{1, \cdots, m\}$ such that

$$\frac{1}{n} \log \sup_{\theta' \in \Theta_0^{k-1}} P_{\theta'}(\mathbf{x}_1^n) \leq \frac{1}{n} \log P_{\theta_{i^*}}(\mathbf{x}_1^n) + \frac{\epsilon}{2}$$

for all $\mathbf{x}_1^n$ in $\mathscr{X}^n$; so that

$$\frac{1}{n} \log P_\theta(\mathbf{x}_1^n) - \frac{1}{n} \log \sup_{\theta' \in \Theta_0^{k-1}} P_{\theta'}(\mathbf{x}_1^n) \geq \frac{1}{n} \log P_\theta(\mathbf{x}_1^n) - \frac{1}{n} \log P_{\theta_{i^*}}(\mathbf{x}_1^n) - \frac{\epsilon}{2} \tag{4.7}$$

for all $\mathbf{x}_1^n$ in $\mathscr{X}^n$.

Next, for $i = 1, \cdots, m$, define

$$C_n(i) \triangleq \left\{ \mathbf{x}_1^n \in X^n : \log P_\theta(\mathbf{x}_1^n) - \log P_{\theta_i}(\mathbf{x}_1^n) \leq (a_k + b_k) \log n + \frac{\epsilon n}{2} + c_k \right\}.$$

Observe that for any θ' in Θ_0^{k-1}

$$\begin{aligned} \lim_n \frac{1}{n} \log \frac{P_\theta(\mathbf{X}_1^n)}{P_{\theta'}(\mathbf{X}_1^n)} &\geq \lim_n \frac{1}{n} \log \frac{P_\theta(\mathbf{X}_1^n)}{\sup_{\tilde{\theta} \in \Theta_0^{k-1}} P_{\tilde{\theta}}(\mathbf{X}_1^n)} \\ &= \inf_{\tilde{\theta} \in \Theta_0^{k-1}} D(P_\theta \| P_{\tilde{\theta}}) \qquad P_\theta - \text{a.s. by Lemma 3.1} \\ &> \epsilon \qquad P_\theta - \text{a.s.} \end{aligned}$$

by the choice of ϵ. Hence, by [19, Theorem 2.1], for each $i = 1, \cdots, m$, there exists $\delta'(\theta_i, \epsilon) > 0$ such that

$$\limsup_n \frac{1}{n} \log P_\theta(C_n(i)) \leq -\delta'(\theta_i, \epsilon). \tag{4.8}$$

Finally, note that for all $n \geq \max_{1 \leq i \leq m} N(\theta_i)$, it holds that $B_n \subseteq \bigcup_{i=1}^m C_n(i)$, which combined with (4.8), gives

$$\limsup_n \frac{1}{n} \log P_\theta(B_n) \leq -\min_{1 \leq i \leq m} \delta'(\theta_i, \epsilon).$$

The previous inequality, in conjunction with (4.5), yields the assertion of the lemma with $\lambda = \min_{1 \leq i \leq m} \delta'(\theta_i, \epsilon) > 0$. □

V. Discussion

The order estimator of (3.7) is shown, in Lemmas 4.2 and 4.3, to yield an overestimation probability which decays to zero polynomially in the sample size n, while the probability of overestimation decays exponentially in n. We have been unable to characterize precisely this exponent. We show in a forthcoming paper [9], however, that this estimator, when adapted to the problem of Markov order estimation, is indeed asymptotically optimal over the class of strongly consistent order estimators in that it achieves the optimal error exponent in the underestimation probability characterized in [8] and [9].

We have assumed heretofore that the observed HMS can be modeled *exactly* by a member of the hypothesized class of models, i.e., $\{X_n\}_{n=1}^\infty$ is generated by P_θ for some θ in $\bigcup_{k=1}^{k_0} \Theta_0^k$. Accordingly, our definition of the minimal model set $\mathscr{M}$ (cf. Section II) affords the following interpretation of the "true model" (k, θ): if (k, θ) belongs to $\mathscr{M}$, it holds that

$$k = \arg \min_{1 \leq k' \leq k_0} \inf_{\theta' \in \Theta_0^{k'}} D(P_\theta \| P_{\theta'}). \tag{5.1}$$

Thus, from among the models (k', θ'), θ' in $\Theta_0^{k'}$, for each of which $P_{\theta'}$ achieves the minimum Kullback–Leibler distance—here, zero—from the probability measure generating the observed HMS, the true model is the one corresponding to the lowest order.

Quite often the observed data cannot be characterized *exactly* by any member of the class of hypothesized models. This occurs in our context if, for instance, the stationary ergodic measure on $\mathscr{X}^\infty$ generating the observed process $\{X_n\}_{n=1}^\infty$ corresponds to an HMS of order exceeding k_0, or is not an HMS at all. In such situations, it is desirable to approximate the observed data in terms of one of the hypothesized models which is closest to it in a

suitable sense. To be specific, consider the situation in which the observed process $\{X_n\}_{n=1}^{\infty}$ is generated by a stationary ergodic hidden Markov measure P *not belonging* to $\bigcup_{1\le k'\le k_0}\bigcup_{\theta'\in\Theta_0^{k'}} P_{\theta'}$. In analogy with (5.1), a desirable model order k for the observed HMS would satisfy

$$k = \arg \min_{1\le k'\le k_0} \inf_{\theta'\in\Theta_0^{k'}} D(P\|P_{\theta'}). \tag{5.2}$$

Note in (5.2) that k may be less than the maximum allowable order k_0. An interesting class of order estimators would then be one for which the estimates corresponding to increasing sample sizes converge $P-$ a.s. to k. It is unclear whether the estimator of (3.7) possesses this property in general; it may do so in special cases as is illustrated by the following example.

Example 3: The observation process $\{X_n\}_{n=1}^{\infty}$, generated by a stationary ergodic Markov measure P on $\{0,1\}^{\infty}$ of order 3, is a $\{0,1\}$-valued Markov process satisfying the following two conditions: i) X_1, X_2, X_3 are IID random variables with $P(X_i = 0) = 1/2$, $i = 1,2,3$; ii) for $n \ge 1$, $X_{n+3} = X_n + W_n$ modulo 2, where $\{W_n\}_{n=1}^{\infty}$ is a $\{0,1\}$-valued IID process independent of (X_1, X_2, X_3), and with $P(W_n = 0) = \alpha \ne 1/2$, $n \ge 1$.

It is readily verified that $P(X_{n+1}|\mathbf{X}_1^n) = P(X_{n+1}|\mathbf{X}_{n-2}^n)$ for $n \ge 3$. Let $\{P_\theta\}_{\theta\in\Theta_0^k}$ be the set of all stationary ergodic Markov measures on $\{0,1\}^{\infty}$ of order k, $0 \le k \le k_0$. If we choose $k_0 \ge 4$, so that the hypothesized class of models includes the one generating the observed process, it is evident that $k = 4$ in (5.1), and by Proposition 3.5, the estimator of (3.7) obeys $\lim_n \hat{k}_n(\mathbf{X}_1^n) = 4$ $P-$ a.s. On the other hand, if we pick $k_0 = 3$, straightforward but tedious calculations show that $k = 0$ in (5.2), so that an IID model best represents the observed process in the sense of (5.2). For this case, the estimator of (3.7), suitably modified for Markov order estimation, can also be shown to satisfy $\lim_n \hat{k}_n(\mathbf{X}_1^n) = 0$ $P-$ a.s.

We conclude by addressing the problem of consistent estimation of the order of the $\mathscr{X}$-valued (general) stationary ergodic process $\{Y_n\}_{n=1}^{\infty}$ introduced in Section II. As for the HMS, ambiguity about the "true" order is avoided by considering a set of minimal models $\tilde{\mathscr{M}} = \{(k,\theta): 1 \le k \le k_0, \theta \in \prod^k\}$ with the following property: For any (k,θ) in $\tilde{\mathscr{M}}$, there exists no pair (k',θ'), $k' < k$, θ' in $\prod^{k'}$, such that P_θ and $P_{\theta'}$ are equal measures on $\mathscr{X}^{\infty}$. Note that the HMS order estimator of (3.7) now ceases to be appropriate for two reasons. First, the mixture distribution Q_k, $1 \le k \le k_0$, for the process $\{Y_n\}_{n=1}^{\infty}$ no longer admits a convenient form, in contrast to that for the HMS $\{X_n\}_{n=1}^{\infty}$ (cf. (A.27) in the Appendix). Second, although Lemma 3.3 still holds for $\{Y_n\}_{n=1}^{\infty}$, the validity of Lemma 3.4 is unclear.

Consistent estimation of the order of $\{Y_n\}_{n=1}^{\infty}$ can be achieved by a slight modification of the procedure proposed by Merhav *et al.* [16] and Ziv and Merhav [30] for estimating the order of the smaller classes of finite-state and Markov processes. The resulting estimator, described below, does, however, involve cumbersome maximum likelihood computations. Given an observed sequence $\mathbf{y}_1^n$ in $\mathscr{X}^n$, the order estimator $\hat{k}_n^{\mathrm{ZM}}$ is defined by

$$\hat{k}_n^{\mathrm{ZM}}(\mathbf{y}_1^n) = \begin{cases} 1, & \text{if } -\dfrac{1}{n}\log P_{\hat{\theta}_k^n}(\mathbf{y}_1^n) - \dfrac{1}{n}L_{\mathrm{WZ}}(\mathbf{y}_1^n) \le \lambda_n \text{ for } 1 \le k \le k_0 \\ \max\left\{k: 1 \le k \le k_0, \quad -\dfrac{1}{n}\log P_{\hat{\theta}_k^n}(\mathbf{y}_1^n) - \dfrac{1}{n}L_{\mathrm{WZ}}(\mathbf{y}_1^n) > \lambda_n\right\} + 1 & \text{otherwise,} \end{cases} \tag{5.3}$$

where $L_{\mathrm{WZ}}(\mathbf{y}_1^n)$ is the length of the Wyner–Ziv data compression codeword [29] for $\mathbf{y}_1^n$, and the sequence $\{\lambda_n\}_{n=1}^{\infty}$ is chosen so as to satisfy simultaneously the conditions $\lim_n \lambda_n = 0$ and $\lim_n n\lambda_n = \infty$. Using Lemma 2.5 in conjunction with standard techniques, it is shown in the Appendix that under conditions (A1)–(A3) (cf. Section II), the order estimator of (5.3) has the following consistency properties. If $\sum_{n=1}^{\infty} 2^{-n\lambda_n} < \infty$, then for every (k,θ) in $\tilde{\mathscr{M}}$, $\lim_n \hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) = k$ $P_\theta-$ a.s.; otherwise, $\lim_n \hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) = k$ in probability P_θ.

It is not clear how the performance of the Ziv–Merhav estimator of (5.3) compares with that of the order estimator of (3.7) when the data is emitted by an HMS. In particular, it is not known if the former, like the latter, yields an underestimation probability that decays exponentially with sample size. On the other hand, the Ziv–Merhav estimator has the advantage of not requiring an a priori upper bound k_0 on HMS order; the estimator of (3.7) relies on this knowledge of k_0.

Acknowledgments

One of the authors (P.N.) wishes to thank Prof. R. Ahlswede for several stimulating discussions and his hospitality during a visit to the Department of Mathematics, Universität Bielefeld, Bielefeld, Germany, in July 1991. The facilities provided there by Sonderforschungsbereich 343, Discrete Strukturen in der Mathematik, are also gratefully acknowledged. The authors also wish to thank the anonymous referees for their valuable comments.

Appendix

A. Consistency of the Order Estimator in (5.3)

Proof: Assume conditions (A1)–(A3) of Section II to hold. We prove below that if $\sum_{n=1}^{\infty} 2^{-n\lambda_n} < \infty$, then for every (k,θ) in $\tilde{\mathscr{M}}$, $\lim_n \hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) = k$ $P_\theta-$ a.s.; otherwise, $\lim_n \hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) = k$ in probability P_θ.

The proof involves separately overbounding the probabilities of overestimation and underestimation of the order of $\{Y_n\}_{n=1}^{\infty}$. We first obtain an upper bound on the overestimation probability in the manner of [16], [30]. Fix (k,θ) in $\tilde{\mathscr{M}}$. For $1 \le k' \le k_0$, letting

$$\mathscr{N}_{k'} \triangleq \left\{\mathbf{y}_1^n \in \mathscr{X}^n: -\frac{1}{n}\log P_{\hat{\theta}_{k'}^n}(\mathbf{y}_1^n) - \frac{1}{n}L_{\mathrm{WZ}}(\mathbf{y}_1^n) > \lambda_n\right\}$$

we observe that

$$P_\theta(\hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) > k) \le P_\theta\left(\bigcup_{k'=k}^{k_0} \mathscr{N}_{k'}\right) = P_\theta(\mathscr{N}_{k_0}) \tag{A.1}$$

where the previous inequality follows from assumption (A1).

Next, observe that

$$\begin{aligned} P_\theta(\mathcal{N}_{k_0}) &= \sum_{\mathbf{y}_1^n \in \mathcal{N}_{k_0}} P_\theta(\mathbf{y}_1^n) \\ &\le \sum_{\mathbf{y}_1^n \in \mathcal{N}_{k_0}} P_{\hat{\theta}_k^n}(\mathbf{y}_1^n) \\ &= \sum_{\mathbf{y}_1^n \in \mathcal{N}_{k_0}} 2^{-n[-\frac{1}{n}\log P_{\hat{\theta}_k^n}(\mathbf{y}_1^n)]} \\ &\le \sum_{\mathbf{y}_1^n \in \mathcal{Z}^n} 2^{-n[\lambda_n + \frac{1}{n}L_{\mathrm{WZ}}(\mathbf{y}_1^n)]} \\ &= 2^{-n\lambda_n} \sum_{\mathbf{y}_1^n \in \mathcal{Z}^n} 2^{-L_{\mathrm{WZ}}(\mathbf{y}_1^n)} \\ &\le 2^{-n\lambda_n} \end{aligned} \tag{A.2}$$

where the last inequality holds since the UD Wyner–Ziv code [29] satisfies the Kraft inequality: $\sum_{\mathbf{y}_1^n \in \mathcal{Z}^n} 2^{-L_{\mathrm{WZ}}(\mathbf{y}_1^n)} \le 1$. From (A.1) and (A.2), we have that $\lim_n P_\theta(\hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) > k) \le \lim_n 2^{-n\lambda_n} = 0$.

If $\sum_{n=1}^\infty 2^{-n\lambda_n} < \infty$, the Borel–Cantelli lemma yields $P_\theta(\hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) > k \text{ infinitely often}) = 0$, i.e., $\lim_n \sup \hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) \le k$ P_θ – a.s.

We consider next the event of underestimation. First note that if $\mathbf{y}_1^n$ is in $\mathcal{N}_{k-1}$, then $\hat{k}_n^{\mathrm{ZM}}(\mathbf{y}_1^n) \ge k$. If it can be shown that

$$\liminf_n \left[-\frac{1}{n}\log P_{\hat{\theta}_{k-1}^n}(\mathbf{Y}_1^n) - \frac{1}{n}L_{\mathrm{ZM}}(\mathbf{Y}_1^n) \right] > 0 \qquad P_\theta - \text{a.s.} \tag{A.3}$$

then, since $\lim_n \lambda_n = 0$, it follows that

$$\liminf_n \hat{k}_n^{\mathrm{ZM}}(\mathbf{Y}_1^n) \ge k \qquad P_\theta - \text{a.s.}$$

Let $H_\theta \triangleq \lim_n \{-1/nE_{P_\theta}[\log P_\theta(\mathbf{Y}_1^n)]\}$ be the entropy of $\{Y_n\}_{n=1}^\infty$ under P_θ. Then (A.3) is established as follows:

$$\begin{aligned} &\liminf_n \left[-\frac{1}{n}\log P_{\hat{\theta}_{k-1}^n}(\mathbf{Y}_1^n) - \frac{1}{n}L_{\mathrm{ZM}}(\mathbf{Y}_1^n) \right] \\ &= \liminf_n \left[L^n_{\theta,\hat{\theta}^n_{k-1}}(\mathbf{Y}_1^n) - \frac{1}{n}\log P_\theta(\mathbf{Y}_1^n) \right. \\ &\quad \left. - \frac{1}{n}L_{\mathrm{ZM}}(\mathbf{Y}_1^n) \right] \\ &\ge \liminf_n L^n_{\theta,\hat{\theta}^n_{k-1}}(\mathbf{Y}_1^n) \\ &\quad + \liminf_n \left[-\frac{1}{n}\log P_\theta(\mathbf{Y}_1^n) - \frac{1}{n}L_{\mathrm{ZM}}(\mathbf{Y}_1^n) \right] \\ &\ge \min_{\theta' \in \Pi^{k-1}} D_{\theta,\theta'}(\mathbf{Y}) + H_\theta \\ &\quad - \limsup_n \frac{1}{n}L_{\mathrm{ZM}}(\mathbf{Y}_1^n) \text{ (by Lemma 2.5)} \\ &> 0 \qquad P_\theta - \text{a.s.} \end{aligned}$$

where the last inequality results from Corollary 2.4 and [29, Theorem 3(b)]. This completes the proof of consistency of the order estimator of (5.3). □

Proof of Lemma 3.4: To prove Lemma 3.4, it suffices to show for each θ in Θ_0^k that

$$\log \frac{P_\theta(\mathbf{x}_1^n|s_0)}{Q_k(\mathbf{x}_1^n)} \le \frac{k(k+q-2)}{2}\log n + c'_{k,q} \tag{A.4}$$

for all $\mathbf{x}_1^n$ in $\mathcal{Z}^n$, s_0 in $\mathcal{S}$, and $n \ge N'(k,q)$, where $c'_{k,q}$ depends only on k, q. Then, if $s_0^* = s_0^*(\mathbf{x}_1^n) = \arg\max_{s_0 \in \mathcal{S}} P_\theta(\mathbf{x}_1^n|s_0)$, we have from (3.5) for each θ in Θ_0^k that

$$\begin{aligned} \log \frac{P_\theta(\mathbf{x}_1^n)}{Q_k(\mathbf{x}_1^n)} &= \log \frac{P_\theta(\mathbf{x}_1^n)}{\frac{1}{k}\sum_{s_0 \in \mathcal{S}} Q_k(\mathbf{x}_1^n|s_0)} \\ &\le \log \frac{P_\theta(\mathbf{x}_1^n|s_0^*)\cdot}{Q_k(\mathbf{x}_1^n|s_0^*)} + \log k \\ &\le \frac{k(k+q-2)}{2}\log n + c'_{k,q} + \log k \end{aligned} \tag{A.5}$$

where the last inequality follows from (A.4) for all $\mathbf{x}_1^n$ in $\mathcal{Z}^n$ and $n \ge N'(k,q)$. Setting $c_{k,q} = c'_{k,q} + \log k$, the assertion of the Lemma is proved since (A.5) is valid for all θ in Θ_0^k.

We now proceed to establish the claim in (A.4). Note first that the conditional mixture distribution for each s_0 in $\mathcal{S}$ can be written, using (2.2), as

$$\begin{aligned} Q_k(\mathbf{x}_1^n|s_0) &= \int_{\Theta_0^k} P_\theta(\mathbf{x}_1^n|s_0)\nu_k(d\theta) \\ &= \int_{\Theta_0^k} \left(\sum_{\mathbf{s}_1^n \in \mathcal{S}^n} P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n)P_\theta(\mathbf{s}_1^n|s_0) \right) \nu_k(d\theta) \\ &= \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} \left(\int_{\Theta_0^k} P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n)P_\theta(\mathbf{s}_1^n|s_0)\nu_k(d\theta) \right). \end{aligned} \tag{A.6}$$

Fix s_0 in $\mathcal{S}$. For a given $\mathbf{s}_1^n$, let $n_{ij} = n_{ij}(\mathbf{s}_0^n)$ denote the number of continguous occurrences of the symbols i, j in $\mathbf{s}_0^n$, $1 \le i, j \le k$. Let $n_i = n_i(\mathbf{s}_0^n) \triangleq \sum_{j=1}^k n_{ij}$ denote the number of occurrences of the symbol i in $\mathbf{s}_0^{n-1}$, $1 \le i \le k$. (In order to avoid tedious notation, we shall display the dependence of n_{ij} and n_i on $\mathbf{s}_0^n$ only when necessary.) It readily follows that

$$P_\theta(\mathbf{s}_1^n|s_0) = \prod_{i=1}^k \prod_{j=1}^k a_{ij}^{n_{ij}} \tag{A.7}$$

and, further, that

$$P_\theta(\mathbf{s}_1^n|s_0) \le \sup_{\theta \in \Theta^k} P_\theta(\mathbf{s}_1^n|s_0) = \prod_{i=1}^k \prod_{j=1}^k \left(\frac{n_{ij}}{n_i} \right)^{n_{ij}} \tag{A.8}$$

with the convention in (A.8) that if $n_{i'} = 0$ for some i', then $(n_{i'j}/n_{i'}) = 1$ for all j, $1 \le j \le k$.

Next, given $\mathbf{s}_1^n$ in $\mathcal{S}^n$, $\mathbf{x}_1^n$ in $\mathcal{Z}^n$, let $m_{rt} = m_{rt}(\mathbf{s}_1^n, \mathbf{x}_1^n)$ denote the number of pairs of symbols (r,t), $1 \le r \le k$, $1 \le t \le q$, such that $s_l = r$, $x_l = t$ for some l, $1 \le l \le n$. (Thus, m_{rt} is the number of occurrences of the state symbol r and the data symbol t at the same time instant.) Let $m_r = m_r(\mathbf{s}_1^n) \triangleq \sum_{t=1}^q m_{rt}$ denote the number of occurrences of the state symbol r in $\mathbf{s}_1^n$. Then

$$P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n) = \prod_{r=1}^k \prod_{t=1}^q b_{rt}^{m_{rt}} \tag{A.9}$$

and

$$P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n) \le \sup_{\theta \in \Theta^k} P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n) = \prod_{r=1}^k \prod_{t=1}^q \left(\frac{m_{rt}}{m_r} \right)^{m_{rt}}. \tag{A.10}$$

Substituting (A.7) and (A.9) in (A.6), we get

$$\begin{aligned} Q_r(\mathbf{x}_1^n|s_0) = \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} &\int_{\Theta_0^k} \left(\prod_{r=1}^k \prod_{t=1}^q b_{rt}^{m_{rt}(\mathbf{s}_1^n, \mathbf{x}_1^n)} \right) \\ &\cdot \left(\prod_{i=1}^k \prod_{j=1}^k a_{ij}^{n_{ij}(\mathbf{s}_0^n)} \right) \nu_k(d\theta). \end{aligned} \tag{A.11}$$

Now recall from the passage following (2.2) that each θ in Θ_0^k is of the form (A, B), where A and B are $k \times k$ – and $k \times q$ – stochastic matrices, respectively. We pick the prior ν_k on Θ_0^k to be the Dirichlet prior density [3], [6], [27] given by

$$\nu_k(\theta) = \nu_k(A, B) \triangleq \nu_k^S(\{a_{ij}\})\nu_k^X(\{b_{rt}\}) \tag{A.12}$$

where

$$\nu_k^S(\{a_{ij}\}) \triangleq \prod_{i=1}^{k}\left[\frac{\Gamma\left(\frac{k}{2}\right)}{\left[\Gamma\left(\frac{1}{2}\right)\right]^k}\prod_{j=1}^{k} a_{ij}^{-\frac{1}{2}}\right] \tag{A.13}$$

and

$$\nu_k^X(\{b_{rt}\}) \triangleq \prod_{i=1}^{k}\left[\frac{\Gamma\left(\frac{k}{2}\right)}{\left(\Gamma\left(\frac{1}{2}\right)\right)^q}\prod_{t=1}^{q} b_{rt}^{-\frac{1}{2}}\right]. \tag{A.14}$$

In view of (A.12)–(A.14), we can then express (A.11) as

$$Q_k(\mathbf{x}_1^n|s_0) = \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n)Q_k(\mathbf{s}_1^n|s_0) \tag{A.15}$$

where

$$Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n) \triangleq \int_{\{b_{rt}\}}\left(\prod_{r=1}^{k}\prod_{t=1}^{q} b_{rt}^{m_{rt}}\right)\nu_k^X(\{b_{rt}\})\left(\prod_{r=1}^{k}\prod_{t=1}^{q} db_{rt}\right) \tag{A.16}$$

and

$$Q_k(\mathbf{s}_1^n|s_0) \triangleq \int_{\{a_{ij}\}}\left(\prod_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^{n_{ij}}\right)\nu_k^S(\{a_{ij}\})\left(\prod_{i=1}^{k}\prod_{j=1}^{k} da_{ij}\right). \tag{A.17}$$

We proceed with the evaluation of $Q_k(\mathbf{s}_1^n|s_0)$ above. Substituting (A.13) in (A.17), we obtain

$$\begin{aligned} Q_k(\mathbf{s}_1^n|s_0) &= \prod_{i=1}^{k}\left[\int_{\{a_{i1},\ldots,a_{ik}\}}\left(\prod_{j=1}^{k} a_{ij}^{n_{ij}-\frac{1}{2}}\right)\right. \\ &\qquad \left.\cdot\frac{\Gamma\left(\frac{k}{2}\right)}{\left[\Gamma\left(\frac{1}{2}\right)\right]^k} da_{ij}\cdots da_{ik}\right] \\ &= \prod_{i=1}^{k}\left[\frac{\Gamma\left(\frac{k}{2}\right)}{\left[\Gamma\left(\frac{1}{2}\right)\right]^k}\frac{\prod_{j=1}^{k}\Gamma\left(n_{ij}+\frac{1}{2}\right)}{\Gamma\left(n_i+\frac{k}{2}\right)}\right] \\ &= \prod_{i=1}^{k}\left[\left(\prod_{j=1}^{k}\frac{\Gamma\left(n_{ij}+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\right)\frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(n_i+\frac{k}{2}\right)}\right]. \end{aligned} \tag{A.18}$$

In a similar manner, a substitution of (A.14) into (A.16) yields

$$Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n) = \prod_{r=1}^{k}\left[\left(\prod_{t=1}^{q}\frac{\Gamma\left(m_{rt}+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\right)\frac{\Gamma\left(\frac{q}{2}\right)}{\Gamma\left(m_r+\frac{q}{2}\right)}\right]. \tag{A.19}$$

From (A.8) and (A.18), we get

$$\frac{P_\theta(\mathbf{s}_1^n|s_0)}{Q_k(\mathbf{s}_1^n|s_0)} \le \prod_{i=1}^{k}\left[\frac{\prod_{j=1}^{k}\left(\frac{n_{ij}}{n_i}\right)^{n_{ij}}}{\frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(n_i+\frac{k}{2}\right)}\left(\prod_{j=1}^{k}\frac{\Gamma\left(n_{ij}+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}\right)}\right]. \tag{A.20}$$

It can be shown as in [6, eqs. (52)–(61)] that the right-hand side of (A.20) does not exceed

$$\left(\frac{\Gamma\left(n+\frac{k}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}\right)^k$$

which, in turn, by using Stirling's formula for the gamma function, is bounded above by

$$c_k'' n^{\frac{k(k-1)}{2}}$$

for all $n \ge N''(k)$. Consequently, for each s_0 in $\mathcal{S}$, $\mathbf{s}_1^n$ in $\mathcal{S}^n$, we obtain that

$$P_\theta(\mathbf{s}_1^n|s_0) \le Q_k(\mathbf{s}_1^n|s_0)n^{\frac{k(k-1)}{2}}c_k'' \tag{A.21}$$

for all $n \ge N''(k)$.

In a similar manner, it can be shown from (A.10) and (A.19) for each $\mathbf{s}_1^n$ in $\mathcal{S}^n$, $\mathbf{x}_1^n$ in $\mathcal{X}^n$, that

$$P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n) \le Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n)n^{\frac{k(q-1)}{2}}c_{k,q}''' \tag{A.22}$$

for all $n \ge N'''(k, q)$, where $c_{k,q}'''$ depends only on k, q. Then in view of (A.21) and (A.22), we obtain that

$$\begin{aligned} P_\theta(\mathbf{x}_1^n|s_0) &= \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} P_\theta(\mathbf{x}_1^n|\mathbf{s}_1^n)P_\theta(\mathbf{s}_1^n|s_0) \\ &\le \sum_{\mathbf{s}_1^n \in \mathcal{S}^n} Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n)Q_k(\mathbf{s}_1^n|s_0)n^{\frac{k(k+q-2)}{2}}c_k''c_{k,q}''' \end{aligned}$$

for all $n \ge N'(k, q) \triangleq \max\{N''(k), N'''(k, q)\}$.

Finally, recalling the formula for $Q_k(\mathbf{x}_1^n|s_0)$ from (A.15), taking logarithms of both sides above, and setting $c_{k,q}' = \log(c_k''c_{k,q}'')$, we get

$$\log\frac{P_\theta(\mathbf{x}_1^n|s_0)}{Q_k(\mathbf{x}_1^n|s_0)} \le \frac{k(k+q-2)}{2}\log n + c_{k,q}'$$

for all s_0 in $\mathcal{S}$, $\mathbf{x}_1^n$ in $\mathcal{X}^n$, and $n \ge N'(k, q)$, establishing (A.4) and thereby Lemma 3.4. □

B. Computation of Mixture Probabilities $Q_k(\mathbf{x}_1^n)$

We provide below a formula to compute the mixture probabilities $Q_k(\mathbf{x}_1^n)$, $\mathbf{x}_1^n$ in $\mathcal{X}^n$, for use in the *sequential code* (SC) of Section IV.

Given $\mathbf{x}_1^n$ in $\mathscr{X}^n$ and $\mathbf{s}_1^n$ in $\mathscr{S}^n$, we can alternatively express $Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n)$ in (A.19) as

$$Q_k(\mathbf{x}_1^n|\mathbf{s}_1^n) \triangleq \prod_{l=1}^{n} Q_k^X(x_l|\mathbf{x}_1^{l-1}, \mathbf{s}_1^l) \tag{A.23}$$

where

$$Q_k^X(x_l|\mathbf{x}_1^{l-1}, \mathbf{s}_1^l) \triangleq \frac{\alpha_l(r,t) + \frac{1}{2}}{\alpha_l(r) + \frac{q}{2}} \quad \text{if } x_l = t,\ s_l = r \tag{A.24}$$

with $\alpha_l(r,t)$, $1 \le r \le k$, $1 \le t \le q$, being the number of simultaneous occurrences of the state symbol r and data symbol t at the same time instants in $\mathbf{s}_1^l$ and $\mathbf{x}_1^l$, respectively; further $\alpha_l(r) \triangleq \sum_{t=1}^{q} \alpha_l(r,t)$.

In a similar manner, given s_0 in $\mathscr{S}$ and $\mathbf{s}_1^n$ in $\mathscr{S}^n$, we can alternatively express $Q_k(\mathbf{s}_1^n|s_0)$ in (A.18) as

$$Q_k(\mathbf{s}_1^n|s_0) \triangleq \prod_{l=1}^{n} Q_k^{\mathscr{S}}(s_l|\mathbf{s}_0^{l-1}) \tag{A.25}$$

where

$$Q_k^{\mathscr{S}}(s_l|\mathbf{s}_0^{l-1}) \triangleq \frac{\beta_{l-1}(i,j) + \frac{1}{2}}{\beta_{l-1}(i) + \frac{k}{2}} \quad \text{if } s_{l-1} = i,\ s_l = j \tag{A.26}$$

with $\beta_{l-1}(i,j)$ being the number of contiguous occurrences of the symbols i, j in $\mathbf{s}_0^{l-1}$, $1 \le i, j \le k$, and $\beta_{l-1}(i) \triangleq \sum_{j=1}^{k} \beta_{l-1}(i,j)$.

Finally, from (3.5), (A.15), and (A.23)–(A.26), we obtain that

$$Q_k(\mathbf{x}_1^n) = \frac{1}{k} \sum_{s_0 \in \mathscr{S}} \left[\sum_{\mathbf{s}_1^n \in \mathscr{S}^n} \left(\prod_{l=1}^{n} \frac{\alpha_l(r,t) + \frac{1}{2}}{\alpha_l(r) + \frac{q}{2}} \cdot \frac{\beta_{l-1}(i,j) + \frac{1}{2}}{\beta_{l-1}(i) + \frac{k}{2}} \right) \right] \tag{A.27}$$

for all $\mathbf{x}_1^n$ in $\mathscr{X}^n$. Note in (A.27) that $\alpha_l(r,t)$ and $\alpha_l(r)$ depend on $(\mathbf{x}_1^n, \mathbf{s}_1^n)$, while $\beta_{l-1}(i,j)$ and $\beta_{l-1}(i)$ depend on $\mathbf{s}_0^n$. The summation over exponentially many (in n) state sequences in (A.27) above presents a formidable computational burden. This can, however, be alleviated by breaking the sum in (A.27) into partial sums over polynomially many (in n) appropriate "conditional Markov types," and noting that the summand is constant in each such partial sum.

References

[1] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. E.ectr. Eng., Stanford Univ., Aug. 1985.

[2] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 30, pp. 1554–1563, 1966.

[3] I. Csiszár, "Information theoretical methods in statistics," class notes, Univ. Maryland, College Park, Spring 1990.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[5] I. Csiszár and P. Shields, private communication, 1991.

[6] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 269–279, May 1981.

[7] L. Finesso, "Order estimation for functions of Markov chains," Ph.D. dissertation, Dept. Electr. Eng. Univ. Maryland, College Park, Dec. 1990.

[8] L. Finesso, C. Liu and P. Narayan, "The optimal error exponent for Markov order estimation," in *Proc: 1993 IEEE Int. Symp. Inform. Theory*, 1993, p. 186.

[9] ——, "Optimal error exponent and estimators for Markov order estimation," in preparation.

[10] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[11] J. C. Kieffer, "Strongly consistent code-based identification and order estimation for constrained finite-state model classes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 893–902, May 1993.

[12] G. G. Langdon, "An introduction to arithmetic coding," *IBM J. Res. Develop*, vol. 28, pp. 135–149, Mar. 1984.

[13] G. G. Langdon and J. Rissanen, "Compression of black-white images and arithmetic coding," *IEEE Trans. Commun.*, vol. COM-39, pp. 858–867, June 1981.

[14] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Processes, Applicat.*, vol. 40, pp. 127–143, 1992.

[15] C. Liu, "On the estimation of the order of a stationary ergodic Markov source," *Abstr. Papers, Int. Symp. Inform. Theory*, San Diego, CA, Jan. 1990.

[16] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1014–1019, Sept. 1989.

[17] N. Merhav, "The estimation of model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109–1113, Sept. 1989.

[18] A. Paz, *Introduction to Probabilistic Automata*. New York: Academic, 1971.

[19] A. Perez, "Generalization of Chernoff's result on the asymptotic discernibility of two random processes," *Colloquia Mathematica Societies Janos Bolyai*, vol. 9, pp. 619–632, 1972.

[20] T. Petrie, "Probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 40, pp. 97–115, 1969.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[22] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989, Singapore.

[23] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, 1986.

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

[25] ——, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[26] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM J. Res. Develop.*, vol. 23, pp. 149–162, Mar. 1979.

[27] Yu. M. Shtar'kov, "Universal sequential coding of single messages," *Problemi Peredachi Informatsii*, English translation, vol. 23, no. 3, pp. 175–186, 1987.

[28] M. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite memory sources," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1002–1014, May 1992.

[29] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1250–1263, Nov. 1989.

[30] J. Ziv and N. Merhav, "Estimating the number of states of a finite state source," *IEEE Trans. Inform. Theory*, vol. 38, pp. 61–65, Jan. 1992.