The Optimal Error Exponent for Markov Order Estimation

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

***Abstract*— We consider the problem of estimating the order of a stationary ergodic Markov chain. Our focus is on estimators which satisfy a generalized Neyman–Pearson criterion of optimality. Specifically, the optimal estimator minimizes the probability of underestimation among all estimators with probability of overestimation not exceeding a given value. Our main result identifies the best exponent of asymptotically exponential decay of the probability of underestimation. We further construct a consistent estimator, based on Kullback–Leibler divergences, which achieves the best exponent. We also present a consistent estimator involving a recursively computable statistic based on appropriate mixture distributions; this estimator also achieves the best exponent for underestimation probability.**

***Index Terms*—Markov order, error exponent, hypothesis testing, order estimation.**

I. Introduction

A WIDE variety of approaches [1]–[6], have been developed over the years to estimate the order of dependence of a finite Markov chain. In the early literature, this problem was treated as one of multiple hypothesis testing; Billingsley [1] provides a systematic presentation of this approach and related results. Another approach involves an extension to finite-state systems [5], [6] of penalized likelihood estimators, introduced by Akaike and Rissanen for estimating the order of ARMA and state-space models. Recently, Kieffer [3] has proposed a strongly consistent order estimator, based on Rissanen's Minimum Description Length (MDL) Principle [5], [7], for a class of processes including Markov, hidden Markov, and finite-state cases.

Yet another approach, leading to recent contributions [4], [8], [9], employs information-theoretic notions and techniques to address the problem of optimal estimators in the sense of a generalized Neyman–Pearson criterion. An information-theoretic approach had been proposed in the earlier work of Chatfield [2], who had considered an estimator of the order based on empirical entropies H_k, with k being the order of dependence under test. The method suggested in [2] is based on the differences $H_k - H_{k+1}$, but is purely heuristic; in particular, the thresholds with which these differences are to be compared remain unspecified. The independent work of Merhav–Gutman–Ziv [4] is also based, in its simplest form, on a comparison of the differences $H_k - H_{k_0}$ with a prespecified threshold, where k_0 is a known upper bound on the order. The analysis in [4] shows the optimality of the resulting estimator in a given class. However, this estimator may, under adverse specifications, be inconsistent with the probability of underestimation of the order approaching unity. Such a tendency to underestimate is undesirable in several applications, e.g., universal data compression based on models for the data. If the estimate exceeds the "true" model order, models of higher orders will be permitted in describing the observed data, and these will include the true data-generating mechanism. Then, even though the redundancy of the resulting code for the data may not be optimal (as a consequence of order overestimation), its normalized value with respect to the number of observations tends to zero with probability one. On the other hand, order underestimation would lead to a restriction to lower order models in describing the data. Since the true distribution is now precluded from consideration, the average normalized redundancy does not vanish with increasing data size. It is, therefore, often desirable to seek consistent order estimators, which additionally afford minimal probability of underestimation.

Our results are in the spirit of Merhav–Gutman–Ziv [4] and provide, under more general conditions, a complete characterization of the consistency properties and error exponent of a class of estimators. By considering a wider class of estimators than in [4], we identify consistent estimators, based on Kullback–Leibler divergences, for which the probability of underestimation is additionally guaranteed to decay exponentially with the optimal exponent. These consistency and optimality properties are shown to be shared by estimators based on mixture distributions (cf., e.g., [10], [11]). The latter estimators offer a computational advantage in that the statistic involved can be updated recursively. Furthermore, they exhibit interesting connections with the MDL estimator [5], [7].

In Section II, relevant results are compiled on the cardinalities and probabilities of Markov types. The order estimator is proposed in Section III, where its overestimation and underestimation probabilities are overbounded in Theorem 1, which also addresses the issue of consistency. The optimality of the estimator, with respect to a generalized Neyman–Pearson criterion, is proved in Theorem 2. Section III concludes by establishing a connection with the results of [4]. In Section IV, we present the mixture-based estimator, prove its optimality and consistency in Theorem 3, and conclude with remarks on its relationship with the MDL estimator [5], [7].

Manuscript received November 15, 1994; revised December 15, 1995. This work was supported by the National Science Foundation under Grant NSF D CDR 8803012, and by Consiglio Nazionale delle Ricerche, Rome, Italy. The material in this paper was presented in part at the 1993 IEEE International Symposium on Information Theory, San Antonio, TX, January 17–22, 1993.

L. Finesso is with LADSEB-CNR, Corso Stati Uniti 4, 35127 Padova, Italy.

C.-C. Liu is with IBM Almaden Research Center, San Jose, CA CA 95120 USA.

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

Publisher Item Identifier S 0018-9448(96)06313-4.

II. Preliminaries

Let the finite set $\mathcal{X} = \{1, \cdots, r\}$ and a constant $k_0 > 1$ be given. For each $k \in \{1, \cdots, k_0\}$, the set $\{p(a|s),\ a \in \mathcal{X},\ s \in \mathcal{X}^k\}$ defines a transition probability matrix (t.p.m.) of *memory* k. Let Ψ_k be the set of all *strictly positive* t.p.m.'s of memory k. For each $p \in \Psi_k$, we define a Markov measure P of memory k on the set $\mathcal{X}^\infty$ of infinite sequences from $\mathcal{X}$ as follows. We assume throughout that any observed sequence $x^n := (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ is preceded by k_0 *fixed* initial samples $(x_{-k_0+1}, \cdots, x_0)$, and

$$P(x^n) := \prod_{t=1}^{n} p(x_t | x_{t-k}^{t-1}) \tag{1}$$

where $x_{t-k}^{t-1} := (x_{t-k}, \cdots, x_{t-1})$.

Let $\mathcal{M}_k$ be the set of all Markov measures of memory k thus generated. Note that this construction yields an increasing sequence $\mathcal{M}_k$ of sets of measures.

Define the *order* of $P \in \mathcal{M}_{k_0}$ as the smallest constant $k \in \{1, \cdots, k_0\}$ such that, for $n > k$

$$P(x_n | x_1^{n-1}) = P(x_n | x_{n-k}^{n-1}) \quad \text{for all } x^n \in \mathcal{X}^n.$$

It is convenient to define a mutually disjoint sequence of sets $\mathcal{P}_k$ as follows:

$$\begin{aligned} \mathcal{P}_1 &:= \mathcal{M}_1 \\ \mathcal{P}_k &:= \mathcal{M}_k \backslash \mathcal{M}_{k-1}, \quad 1 < k \le k_0. \end{aligned}$$

The set of Markov measures of order k coincides with $\mathcal{P}_k$. We denote by Θ_k the subset of Ψ_k which is in one-to-one correspondence with the elements of $\mathcal{P}_k$. Observe that Θ_k is open in Ψ_k. This is seen as follows: Ψ_k is the interior, in the Euclidean topology, of the unit simplex in $\mathbb{R}^{r^k \times r}$ and, hence, is open. The subset of Ψ_k corresponding to $\Psi_{k'}$ (with $k' < k$) is closed in Ψ_k. Since

$$\Theta_k = \Psi_k \backslash \bigcup_{l=1}^{k-1} \Psi_l$$

the observation is immediate.

In the order estimation problem, we observe a stochastic process $\{X_t,\quad t \ge 1\}$ with values in $\mathcal{X}$ and generated by an *unknown* measure

$$P \in \bigcup_{k=1}^{k_0} \mathcal{P}_k$$

whose order we seek to estimate.

A key tool involves the notion of Markov types which is described below. Given a sequence $x^n \in \mathcal{X}^n$, $\quad n > 1$, we define its k_0th Markov type (cf. [12]) as the empirical distribution on $\mathcal{X}^{k_0} \times \mathcal{X}$ given by

$$Q := \{q_{sa},\ s \in \mathcal{X}^{k_0},\ a \in \mathcal{X}\}$$

with

$$q_{sa} := \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}(x_{t-k_0}^{t-1} = s,\ x_t = a)$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. (Note that $Q = Q^{(n)}$. For notational convenience, we shall, however, suppress this dependence on n, except where explicitly needed.) Let

$$q_s := \sum_{a \in \mathcal{X}} q_{sa}$$

be the marginal distribution on $\mathcal{X}^{k_0}$ corresponding to q_{sa}. Finally, denote by $\mathcal{Q}^{(n)}$ the set of all k_0th Markov types with denominator n.

We define the (conditional) entropy of Q to be

$$H(Q) := -\sum_{s \in \mathcal{X}^{k_0}} \sum_{a \in \mathcal{X}} q_{sa} \log \frac{q_{sa}}{q_s}$$

with the convention that $q_{sa}/q_s = 0$ if $q_s = 0$. All logarithms and exponentials are with respect to the base 2. For

$$p \in \bigcup_{k=1}^{k_0} \Theta_k$$

we define the conditional (Kullback–Leibler) divergence of Q and p as

$$D(Q \| p) := \sum_{s \in \mathcal{X}^{k_0}} \sum_{a \in \mathcal{X}} q_{sa} \log \frac{q_{sa}/q_s}{p(a|s)}.$$

Note that if $p \in \Theta_k$ for some $k \le k_0$, $p(a|s)$ will depend only on the latest k components of $s \in \mathcal{X}^{k_0}$.

Let $\mathcal{T}_Q$ be the set of sequences $x^n \in \mathcal{X}^n$ of (common) k_0th Markov type Q, i.e.,

$$\mathcal{T}_Q := \{\overline{x}^n\colon\ \ \overline{q}_{sa} = q_{sa}, \quad s \in \mathcal{X}^{k_0},\ a \in \mathcal{X}\}.$$

Let $|\mathcal{T}_Q|$ denote the cardinality of $\mathcal{T}_Q$. Further, let

$$P(\mathcal{T}_Q) := \sum_{\mathcal{T}_Q} P(x^n).$$

The following bounds are obtained from [13].

Lemma 1 (Gutman [13]): For every k_0th Markov type Q

$$\begin{aligned} n^{-r^{k_0}} (n+1)^{-(r^{2k_0+1})} \exp[nH(Q)] &\le |\mathcal{T}_Q| \\ &\le r^{k_0} \exp[nH(Q)]. \end{aligned}$$

Moreover, for $p \in \bigcup_{k=1}^{k_0} \Theta_k$, and the corresponding P

$$\begin{aligned} (n+1)^{-r^{k_0+1}} \exp[-nD(Q\|p)] &\le P(\mathcal{T}_Q) \\ &\le r^{k_0} \exp[-nD(Q\|p)] \end{aligned}$$

where P is defined by (1). □

We remark that our choice of the notion of Markov type as defined above enables us to readily exploit the bounds in [13]. An alternative notion of Markov type, namely circular type (cf., [12]), exists and will indeed be briefly used in the proofs of Lemma 2 and Theorem 2 below.

Next, for $1 \le k' \le k \le k_0$, we define the (conditional) divergence between $p' \in \Theta_{k'}$ and $p \in \Theta_k$ as follows. Let p'_e be the embedding of p' in Ψ_k, i.e., for $a \in \mathcal{X}$, $s_1^k \in \mathcal{X}^k$, $p'_e(a|s_1^k) = p'(a|s_{k-k'+1}^k)$. Note that $p'_e \in \Psi_k \backslash \Theta_k$ and will have some identical rows. Then

$$D(p' \| p) := \sum_{s \in \mathcal{X}^k} \pi'_e(s) \sum_{a \in \mathcal{X}} p'_e(a|s) \log \frac{p'_e(a|s)}{p(a|s)} \tag{2}$$

with π'_e being the (unique) invariant measure associated with p'_e. For $p' \in \overline{\Theta}_{k'}$, where $\overline{\Theta}_{k'}$ denotes the Euclidean closure of $\Theta_{k'}$, the uniqueness of the invariant measure of p'_e is generally lost; in this case we define, with a slight abuse of notation,

$$D(p'||p) := \min_{\substack{\pi'_e \\ \pi'_e = \pi'_e p'_e}} D(p'||p).$$

It is seen in a standard manner that $D(\cdot||p)$ is continuous on $\Theta_{k'}$ and lower semicontinuous (l.s.c.) on $\overline{\Theta}_{k'}$.

III. The Order Estimator and its Optimality

In this section, we present an estimator which satisfies a generalized Neyman–Pearson criterion of optimality. Namely, the estimator minimizes the probability of underestimation among all estimators whose probability of overestimation lies below a prespecified level.

We observe the process $\{X_t, \ t \geq 1\}$ generated by an *unknown* measure $P \in \mathcal{P}_k, \ 1 \leq k \leq k_0$. Based on a sample $x^n \in \mathcal{X}^n$, we wish to construct an estimator, namely a decision rule, $\hat{k}_n$, of the order of P. This is equivalent to solving the following multiple *composite* hypothesis testing problem:

$$H_k : P \in \mathcal{P}_k, \quad k = 1, \cdots, k_0.$$

For a given $\alpha \in (0, 1)$, consider the class of estimators $k_n : \mathcal{X}^n \to \{1, \cdots, k_0\}$ for which, for $P \in \mathcal{P}_k, \ 1 \leq k \leq k_0$

$$P(k_n(X^n) > k) < \alpha, \quad n \geq N(\alpha, p). \tag{3}$$

We seek in this class an estimator $\hat{k}_n$ such that for each $P \in \mathcal{P}_k$, $1 \leq k \leq k_0$

$$\overline{\lim_n} \frac{1}{n} \log P(\hat{k}_n(X^n) < k) \leq \underline{\lim_n} \frac{1}{n} \log P(k_n(X^n) < k). \tag{4}$$

We remark that there are basic differences between this approach and that in [4]. These are best discerned in the light of the optimality properties of the estimator $\hat{k}_n$ which is described next.

Definition: Given $x^n \in \mathcal{X}^n$, $n > 1$, let Q denote the k_0th Markov type of the sample.

$\hat{k}_n(x^n) = k$ iff both (5a) and (5b) hold.

$$D(Q||p') > \epsilon_n, \quad \forall p' \in \bigcup_{l=1}^{k-1} \Theta_l \tag{5a}$$

$$D(Q||p) \leq \epsilon_n, \quad \text{for some } p \in \Theta_k \tag{5b}$$

$\hat{k}_n(x^n) = k_0$ if either condition above is not satisfied; where

$$\epsilon_n := (r^{k_0+1} + \delta)(\log n/n)$$

and $\delta > 0$ is any positive constant. □

The decision rule above is motivated as follows. Observe that if the sample is generated by a Markov measure of order k corresponding to a t.p.m. $p \in \Theta_k$, its k_0th Markov type Q (as also its kth Markov type, defined analogously) is eventually trapped in a "divergence neighborhood" of p. This basic fact is reflected in rule (5b), where the choice of a *shrinking* neighborhood is essential for enabling the eventual exclusion of Q from shrinking neighborhoods of lower order measures, thereby rendering underestimation unlikely. On the other hand, the choice of a neighborhood of constant size in (5) would, in some cases, lead to underestimation as discussed at the end of this section. Rule (5a) controls the likelihood of overestimation. Overestimation would require the eventual exclusion of Q from the neighborhood of p, which is contrary to our earlier observation. Indeed, a neighborhood whose size is either constant or shrinking slowly *at an appropriate rate* in (5), will result in a diminishing overestimation probability.

As will be seen below, it turns out that $\hat{k}_n$ can be expressed in an alternative form which is a slight modification of the estimator proposed in [4].

The estimator $\hat{k}_n$ of (5) is a solution to the multiple composite hypothesis testing problem stated above as shown in Theorems 1 and 2 below.

Theorem 1: Fix $\delta > 0$. Then for each $P \in \mathcal{P}_k$, $1 \leq k \leq k_0$

a) $P(\hat{k}_n(X^n) > k) \leq r^{k_0} n^{-(1+\delta)}, \quad n \geq N(\delta)$.

b) For each $\eta > 0$, and for $n \geq N(\eta, p)$

$$P(\hat{k}_n(X^n) < k) \leq \exp\left[-n\left(\min_{k'<k} D(\overline{\Theta}_{k'}||p) - \eta\right)\right]$$

where

$$D(\overline{\Theta}_{k'}||p) := \min_{p' \in \overline{\Theta}_{k'}} D(p'||p).$$

Corollary: The estimator $\hat{k}_n$ is strongly consistent for every $\delta > 0$. □

The proof of Theorem 1 requires the following technical lemma which generalizes a result of Anantharam [14]. Let $\mathcal{B}^n_{k'} \subset \mathcal{Q}^{(n)}$ be the set of all k_0th Markov types satisfying (5b), i.e.,

$$\mathcal{B}^n_{k'} := \{Q \in \mathcal{Q}^{(n)}; \ D(Q||p') \leq \epsilon_n \ \text{ for some } \ p' \in \overline{\Theta}_{k'}\}.$$

Lemma 2: Fix $1 \leq k' < k \leq k_0$ and $p \in \Theta_k$. Then

a) $\lim_n \min_{\mathcal{B}^n_{k'}} D(Q||p) = D(\overline{\Theta}_{k'}||p)$;

b) $D(\overline{\Theta}_{k'}||p) > 0$. □

Proof of Lemma 2:

a) It is convenient to work with circular types

$$\tilde{Q} := \{\tilde{q}_{sa}, \ s \in \mathcal{X}^{k_0}, \ a \in \mathcal{X}\}$$

where the circular type of x^n is obtained as the k_0th type of

$$x^n x^{k_0} := (x_1, \cdots, x_n, x_1, \cdots, x_{k_0})$$

with the counting procedure commencing at the first symbol. Denote the set of all k_0th circular types with denominator n by $\tilde{\mathcal{Q}}^{(n)}$. Let

$$\tilde{\mathcal{B}}^n_{k'} := \{\tilde{Q} \in \tilde{\mathcal{Q}}^{(n)}; \ D(\tilde{Q}||p') \leq \epsilon_n \quad \text{for some } \ p' \in \overline{\Theta}_{k'}\}.$$

For a given sequence x^n, the corresponding types Q and $\tilde{Q}$ satisfy $|\tilde{Q} - Q| = O(n^{-1})$ from which it follows in a standard manner that

$$\lim_n \left| \min_{\mathcal{B}^n_{k'}} D(Q||p) - \min_{\tilde{\mathcal{B}}^n_{k'}} D(\tilde{Q}||p) \right| = 0.$$

Thus it suffices to prove that

$$\lim_n \min_{\tilde{\mathcal{B}}^n_{k'}} D(\tilde{Q}\|p) = D(\overline{\Theta}_{k'}\|p).$$

Let us denote by $\mathcal{M}$ the set of measures on $\mathcal{X}^{k_0}\times\mathcal{X}$. We introduce the decreasing sequence of sets

$$\mathcal{D}_n := \{\tilde{p}\in\mathcal{M};\quad D(\tilde{p}\|p')\le\epsilon_n \quad\text{for some } p'\in\overline{\Theta}_{k'}\}$$

(the divergence being computed with respect to the initial probabilities $\tilde{p}(s) := \Sigma_a \tilde{p}(s, a)$). Then, for all n we have

$$\tilde{\mathcal{Q}}^{(n)}\cap\mathcal{M}(\overline{\Theta}_{k'})\subset\tilde{\mathcal{B}}^n_{k'}\subset\mathcal{D}_n\subset\overline{\mathcal{D}}_n$$

where $\mathcal{M}(\overline{\Theta}_{k'})$ denotes the subset of $\mathcal{M}$ corresponding to the t.p.m.'s in $\overline{\Theta}_{k'}$. Then

$$\min_{\overline{\mathcal{D}}_n} D(\tilde{p}\|p)\le\min_{\tilde{\mathcal{B}}^n_{k'}} D(\tilde{Q}\|p)\le\min_{\tilde{\mathcal{Q}}^{(n)}\cap\mathcal{M}(\overline{\Theta}_{k'})} D(\tilde{Q}\|p).$$

It is, therefore, enough to prove that

$$\overline{\lim_n}\min_{\tilde{\mathcal{Q}}^{(n)}\cap\mathcal{M}(\overline{\Theta}_{k'})} D(\tilde{Q}\|p)\le D(\overline{\Theta}_{k'}\|p)\le\lim_n\min_{\overline{\mathcal{D}}_n}D(\tilde{p}\|p).$$

The left inequality is seen as follows. Let $(\tilde{\pi}, \tilde{p})$ achieve $D(\overline{\Theta}_{k'}\|p)$. Clearly, there exists

$$\hat{Q}^{(n)}\in\tilde{\mathcal{Q}}^{(n)}\bigcap\mathcal{M}(\overline{\Theta}_{k'})$$

whose variational distance from $(\tilde{\pi}, \tilde{p})$ is $O(n^{-1})$. Hence

$$|D(\hat{Q}^{(n)}\|p)-D(\overline{\Theta}_{k'}\|p)|=O\left(\frac{\log n}{n}\right)$$

so that given any $\nu>0$, we have

$$D(\overline{\Theta}_{k'}\|p)\ge D(\hat{Q}^{(n)}\|p)-\nu$$

for all n large. To establish the right inequality first observe that a standard compactness argument shows that

$$\lim_n\min_{\overline{\mathcal{D}}_n}D(\tilde{p}\|p)=D(\tilde{p}_*\|p)$$

for some $\tilde{p}_*\in\cap_n\overline{\mathcal{D}_n}$. Therefore, there exists a $p'_*\in\overline{\Theta}_{k'}$ such that $D(\tilde{p}_*\|p'_*)=0$. By the remarks at the end of Section II it follows that

$$D(\tilde{p}_*\|p)\ge D(p'_*\|p):=\min_{\substack{\pi'\\ \pi'=\pi'p'_*}} D(p'_*\|p)\ge D(\overline{\Theta}_{k'}\|p).$$

b) The l.s.c. of $D(\cdot\|p)$ on $\overline{\Theta}_{k'}$ yields the existence of $(\tilde{\pi}', \tilde{p}')$, $\tilde{p}'\in\overline{\Theta}'_k$, and $\tilde{\pi}'=\tilde{\pi}'\tilde{p}'$, such that the corresponding embedding $(\tilde{\pi}'_e, \tilde{p}'_e)$ achieves $D(\overline{\Theta}_{k'}\|p)$. Now, assume $D(\overline{\Theta}_{k'}\|p)=0$. A standard calculation shows that

$$\tilde{\pi}'_e(s)\tilde{p}'_e(a|s)=\tilde{\pi}'_e(s)p(a|s),\qquad a\in\mathcal{X},\ s\in\mathcal{X}^k$$

from which it follows that $\tilde{\pi}'_e p=\tilde{\pi}'_e$. Since the t.p.m. p is strictly positive, so must be $\tilde{\pi}'_e$, which by the equality above, implies that $\tilde{p}'_e=p$. This is clearly impossible by the remarks at the end of Section II. Hence $D(\overline{\Theta}_{k'}\|p)>0$. □

Proof of Theorem 1: The proof of a) is standard and is relegated to the Appendix. We provide the proof of b). Fix $k'<k$ and $P\in\mathcal{P}_k$. Then

$$\begin{aligned}P(\hat{k}_n(X^n)=k')&\le P\left(\bigcup_{\mathcal{B}^n_{k'}}\mathcal{T}_Q\right)=\sum_{\mathcal{B}^n_{k'}}P(\mathcal{T}_Q)\\&\le\sum_{\mathcal{B}^n_{k'}} r^{k_0}\exp[-nD(Q\|p)],\quad\text{by Lemma 1}\\&\le\binom{n+r^{k_0+1}-1}{r^{k_0+1}-1} r^{k_0}\exp\left[-n\min_{\mathcal{B}^n_{k'}}D(Q\|p)\right]\\&\le\exp[-n(D(\overline{\Theta}_{k'}\|p)-\eta)],\quad n\ge N'(\eta, p)\end{aligned}$$

by Lemma 2. Finally

$$\begin{aligned}P(\hat{k}_n(X^n)<k)&=\sum_{k'<k}P(\hat{k}_n(X^n)=k')\\&\le(k-1)\exp\left[-n\left(\min_{k'<k}D(\overline{\Theta}_{k'}\|p)-\eta\right)\right]\end{aligned}$$

for $n\ge N''(\eta, p)$, from which the desired result follows. □

The proof of the Corollary is straightforward. The rates of decay of the overestimation and underestimation probabilities are the critical quantities. Strong consistency follows by an application of the Borel–Cantelli lemma.

The exponent in the underestimation probability in Theorem 1 is optimal as is shown in

Theorem 2: Let $0<\alpha<1$ be given. Let k_n be an estimator such that for all $P\in\mathcal{P}_k$, $1\le k\le k_0$

$$P(k_n(X^n)>k)<\alpha,\quad n\ge N_1(\alpha, p).\tag{6}$$

Then for every $\epsilon>0$, for all $P\in\mathcal{P}_k$, $1<k\le k_0$, and for $n\ge N_2(\alpha, \epsilon, p)$

$$P(k_n(X^n)<k)\ge\exp\left[-n\left(\min_{k'<k}D(\overline{\Theta}_{k'}\|p)+\epsilon\right)\right].\tag{7}$$

Remarks:

i) We emphasize that the conclusion of Theorem 2 holds only if the hypothesis is met for all $P\in\bigcup_{k\le k_0}\mathcal{P}_k$; the hypothesis is not required to be uniform in P. However, Theorem 2 does not imply that if (6) holds *only* for a particular P, then (7) is valid for the same P.

ii) Theorems 1 and 2 constitute, in effect, an extension of Stein's Lemma (cf. e.g., [15]) to the Markov order estimation problem.

Proof: Loosely speaking, the proof consists of two steps. The first step involves a *Claim* that any estimator k_n which meets the hypothesis (6) on the probability of overestimation must possess the following property: For a fixed k, and for every $k'<k$, and for every t.p.m. p' of order k', there must exist at least one Markov type $Q_{p'}$ which is "close" to p', and is such that sequences x^n of type $Q_{p'}$ occupy a sizable fraction of the decision region $\{x^n: k_n(x^n)=k'\}$. The second step then uses the assertion of the *Claim* to bound below the underestimation probability in terms of the probabilities of these fractions of the decision regions. Formally, assume for the time being the following

Claim: Fix k and $\delta > 0$. Then: For each $p' \in \overline{\Theta}_{k'}$, $k' < k$, there exist:

- a type $Q_{p'} = Q_{p'}^{(n)}$ with $D(Q_{p'}^{(n)} || \pi') \leq \epsilon_n$ and $D(Q_{p'}^{(n)} || p') \leq \epsilon_n$ where π' achieves $D(p' || p)$ and
$$\epsilon_n = (r^{k_0+1} + \delta)\,(\log n/n);$$
- $\eta(\alpha, k') > 0$;
- $N(\alpha, k') > 0$;

such that for all $n \geq N(\alpha, k')$
$$|\mathcal{T}_{Q_{p'}^{(n)}} \cap [k_n < k]| \geq \eta(\alpha, k') |\mathcal{T}_{Q_{p'}^{(n)}}| \tag{8}$$
where
$$[k_n < k] := \bigcup_{l<k} \{x^n : k_n(x^n) = l\}. \qquad \square$$

Then for each $p' \in \overline{\Theta}_{k'}$, $k' < k$
$$\begin{aligned} P(k_n(X^n) < k) &= P([k_n < k]) \\ &\geq P([k_n < k] \cap \mathcal{T}_{Q_{p'}^{(n)}}) \end{aligned}$$
where $Q_{p'}^{(n)}$ is the type in the *Claim*. Since every sequence in $\mathcal{T}_{Q_{p'}^{(n)}}$ has the same P probability, it follows from (8) and Lemma 1 that for $n \geq N(\alpha, k')$
$$P(k_n(X^n) < k) \geq \frac{\eta(\alpha, k')}{(n+1)^{r^{k_0+1}}} \exp\left[-nD\left(Q_{p'}^{(n)} || p\right)\right].$$

Note that $N(\alpha, k')$ *does not* depend on p'. Since the previous bound holds for all $p' \in \overline{\Theta}_{k'}$ and $k' < k$, we obtain that
$$\begin{aligned} P(k_n(X^n) < k) \geq &\left[\min_{k'<k} \eta(\alpha, k')\right] (n+1)^{-r^{k_0+1}} \\ &\cdot \exp\left[-n \min_{k'<k} \min_{p' \in \overline{\Theta}_{k'}} D\left(Q_{p'}^{(n)} || p\right)\right] \end{aligned} \tag{9}$$
for all
$$n \geq \max_{k'<k} [N(\alpha, k')].$$

To obtain the desired exponent, it suffices to establish in (9) that
$$\min_{p' \in \overline{\Theta}_{k'}} D\left(Q_{p'}^{(n)} || p\right) \leq D(\overline{\Theta}_{k'} || p) + \epsilon \tag{10}$$
for all $n \geq N_2(\alpha, \epsilon, p)$. To this end, first note that there exists $p_* \in \overline{\Theta}_{k'}$ such that
$$D(p_* || p) = D(\overline{\Theta}_{k'} || p). \tag{11}$$
Next, the *Claim* provides the existence of a type $Q_{p_*}^{(n)}$ with $D(Q_{p_*}^{(n)} || p_*) \leq \epsilon_n$. Furthermore, since
$$\min_{p' \in \overline{\Theta}_{k'}} D\left(Q_{p'}^{(n)} || p\right) \leq D\left(Q_{p_*}^{(n)} || p\right) \tag{12}$$
observe that by showing for all n sufficiently large (depending only on p, ϵ) that
$$D\left(Q_{p_*}^{(n)} || p\right) \leq D(p_* || p) + \epsilon \tag{13}$$
where (10) follows from (11)–(13).

In order to show (13), note that there exists a circular type $\tilde{Q}_{p_*}^{(n)}$ (cf. proof of Lemma 2) with
$$D\left(Q_{p_*}^{(n)} || p\right) \leq D\left(\tilde{Q}_{p_*}^{(n)} || p\right) + \frac{\epsilon}{2} \tag{14}$$
for all n sufficiently large (depending only on p, ϵ). Furthermore, $D(\tilde{Q}_{p_*}^{(n)} || p_*) \to 0$ which implies that $\tilde{Q}_{p_*}^{(n)} \to \pi_* p_*$, where π_* is the p_*-invariant measure which achieves $D(p_* || p)$. Consequently
$$D\left(\tilde{Q}_{p_*}^{(n)} || p\right) \leq D(p_* || p) + \frac{\epsilon}{2}$$
for all n large (again depending only on p, ϵ), which, together with (14), establishes (13).

It only remains to establish the *Claim*. The proof is by contradiction. Assume the negation of the *Claim*, i.e., there exist $k' < k$ and $p' \in \overline{\Theta}_{k'}$ such that

- for all $Q^{(n)}$ with $D(Q^{(n)} || p') \leq \epsilon_n$
- for all $\eta > 0$
- for all $N > 0$

there is an $n \geq N$ satisfying
$$|\mathcal{T}_{Q^{(n)}} \cap [k_n < k]| < \eta |\mathcal{T}_{Q^{(n)}}|. \tag{15}$$
The previous assumptions yield that infinitely often in n (i.o. (n))
$$\begin{aligned} P'&\left(\bigcup_{Q^{(n)}: D(Q^{(n)}||p') \leq \epsilon_n} \mathcal{T}_{Q^{(n)}} \cap [k_n < k]\right) \\ &= \sum_{Q^{(n)}: D(Q^{(n)}||p') \leq \epsilon_n} P'(\mathcal{T}_{Q^{(n)}} \cap [k_n < k]) \\ &\leq \eta \sum_{Q^{(n)}: D(Q^{(n)}||p') \leq \epsilon_n} P'(\mathcal{T}_{Q^{(n)}}) \leq \eta \end{aligned}$$
by (15). It then follows that i.o. (n)
$$P'([k_n < k]^c) \geq 1 - \eta - P'\left(\left(\bigcup_{Q^{(n)}: D(Q^{(n)}||p') \leq \epsilon_n} \mathcal{T}_{Q^{(n)}}\right)^c\right) \tag{16}$$
where superscript c denotes complement. Next
$$\begin{aligned} P'&\left(\left(\bigcup_{Q^{(n)}: D(Q^{(n)}||p') \leq \epsilon_n} \mathcal{T}_{Q^{(n)}}\right)^c\right) \\ &= P'\left(\bigcup_{Q^{(n)}: D(Q^{(n)}||p') > \epsilon_n} \mathcal{T}_{Q^{(n)}}\right) \\ &= \sum_{Q^{(n)}: D(Q^{(n)}||p') > \epsilon_n} P'(\mathcal{T}_{Q^{(n)}}) \\ &\leq \binom{n + r^{k_0+1} - 1}{r^{k_0+1} - 1} r^{k_0} \exp[-n\epsilon_n] \\ &= \binom{n + r^{k_0+1} - 1}{r^{k_0+1} - 1} r^{k_0} \\ &\quad \cdot \exp\left[-n(r^{k_0+1} + \delta)\frac{\log n}{n}\right] \\ &\leq n^{-\delta} \end{aligned}$$

for all n large (not depending on δ). Substitution in (16) yields i.o. (n)

$$P'([k_n \geq k]) \geq 1 - \eta - n^{-\delta}$$

whereby

$$P'([k_n > k']) \geq P'([k_n \geq k]) \geq 1 - \eta - n^{-\delta}$$

which contradicts the hypothesis of the theorem that $P'([k_n > k']) < \alpha$ for all n sufficiently large. □

Remarks: i) For a given $\alpha \in (0,\, 1)$, the class of estimators defined by (3) includes consistent as well as inconsistent estimators. Theorem 1a) ensures that the estimator $\hat{k}_n$ of (5) belongs to this class. (By the Corollary of Theorem 1, the estimator $\hat{k}_n$ of (5) is additionally strongly consistent.) Theorem 2 identifies the best exponent of underestimation probability among all estimators in this class. Since the estimator $\hat{k}_n$ of (5) possesses, by Theorem 1b), an exponent of underestimation probability which coincides with the best exponent prescribed by Theorem 2, it is optimal in the sense of (4) over the class of estimators given by (3).

ii) Our approach based on the method of Markov types (as an extension of "i.i.d. types" (cf. [16])) provides polynomial terms (cf. Lemma 1 above) which, together with the choice of the threshold ϵ_n in (5), enables us to simultaneously control the overestimation probability of $\hat{k}_n$ in Theorem 1a), and determine the best underestimation error exponent in Theorem 2. In particular, the substantiation of the *Claim* in the proof of Theorem 2 relies critically on the exact behavior of the polynomial terms (cf. (16) and the subsequent analysis). Large deviations theorems for general Markov processes (cf. e.g., [17]) do not provide these crucial polynomial terms.

The following Proposition 1 and its corollary show that the overestimation probability of an order estimator cannot decay exponentially for $P \in \mathcal{P}_k$ without rendering it inconsistent for *some* $\overline{P} \in \mathcal{P}_{\overline{k}}$, $\overline{k} > k$.

Proposition 1: Let $k_0 = 2$. Let $0 < \beta < 1$ be given. Let k_n be an estimator such that

$$\overline{\lim_{n}}\, \frac{1}{n} \log P(k_n(X^n) = 2) \leq -\lambda(P) \tag{17}$$

for *some* $P \in \mathcal{P}_1$, where $\lambda(P) > 0$. Then there exists $\overline{P} \in \mathcal{P}_2$ such that

$$\overline{P}(k_n(X^n) = 1) > \beta \quad \text{i.o.}\,(n). \tag{18}$$

Corollary: Let k_n be an estimator such that, for a given k, for all $P \in \mathcal{P}_k$

$$\overline{\lim_{n}}\, \frac{1}{n} \log P(k_n(X^n) > k) \leq -\lambda(P)$$

where $\lambda(P) > 0$ for all $P \in \mathcal{P}_k$. Then there exists $\overline{P} \in \mathcal{P}_{\overline{k}}$, $k < \overline{k} \leq k_0$ such that

$$\lim_{n}\, \overline{P}(k_n(X^n) < \overline{k}) = 1.$$

Proof: It is easily seen from our construction of $\mathcal{P}_1$, $\mathcal{P}_2$ (or equivalently Θ_1, Θ_2), that for $P \in \mathcal{P}_1$ (corresponding to $p \in \Theta_1$), there exists $\overline{P} \in \mathcal{P}_2$ (corresponding to $\overline{p} \in \Theta_2$) such that $D(\overline{p}\|p) < \lambda(P)$.

Consider the simple binary hypothesis testing problem of deciding between P and $\overline{P}$ with the decision rule which coincides with k_n, i.e., the decision for a given sample x^n is P (resp., $\overline{P}$) iff $k_n(x^n) = 1$ (resp., 2). Now suppose the negation of (18), i.e., that there exists $\beta' \in (0,\, 1)$ such that

$$\overline{P}(k_n(X^n) = 1) \leq \beta', \quad n \geq N(\beta',\, \overline{p}).$$

Then

$$\underline{\lim_{n}}\, \frac{1}{n} \log P(k_n(X^n) = 2) \geq -\, D(\overline{p}\|p) > -\lambda(P)$$

where the first inequality is a consequence of Stein's lemma (cf. [15, p. 81, Exercise 3.4.18], with a simple modification to take into account the different orders of P and $\overline{P}$) for the problem of testing P versus $\overline{P}$ with the decision rule above. This contradicts (17), completing the proof of the Proposition.

The corollary follows in a straightforward manner. □

Remark: It is evident from the proof of the Proposition that (18) is, indeed, valid for every $\overline{P} \in \mathcal{P}_2$ (corresponding to $\overline{p} \in \Theta_2$) which satisfies $D(\overline{p}\|p) < \lambda(P)$. The subset of Θ_2 with this property has positive Lebesgue measure in Ψ_2.

We conclude this section with a comparison of $\hat{k}_n$ with the order estimator of Merhav–Gutman–Ziv [4]. To this end, the rules (5) specifying $\hat{k}_n$ can be equivalently expressed in the form

$$\hat{k}_n = \min \left\{ 1 \leq k \leq k_0; \quad \inf_{p \in \Theta_k} D(Q\|p) \leq \epsilon_n \right\}.$$

Next, observe that

$$\begin{aligned} \inf_{p \in \Theta_k} D(Q\|p) &= -\, H(Q) - \sup_{p \in \Theta_k} \sum_{s,\,a} q_{sa} \log p(a|s) \\ &= H(Q_k) - H(Q) \end{aligned} \tag{19}$$

where Q_k is the kth Markov type of x^n defined in analogy with the k_0th Markov type Q. We thus have the alternative expression

$$\hat{k}_n = \min\, \{1 \leq k \leq k_0; \quad H(Q_k) - H(Q) \leq \epsilon_n\}. \tag{20}$$

Merhav–Gutman–Ziv [4] seek an estimator which minimizes $P(k'_n(X^n) < k)$ for all $P \in \mathcal{P}_k$, $1 \leq k \leq k_0$, over the class of estimators k'_n which satisfy for each $P \in \mathcal{P}_k$, $1 \leq k \leq k_0$

$$\overline{\lim_{n}}\, \frac{1}{n} \log\, (P(k'_n(X^n) > k) < -\lambda \tag{21}$$

where $\lambda > 0$ is a given number. They propose as a solution the estimator k_n^* [4, p. 1015, eq. (8)] which, in our notation, is expressed as

$$k_n^* = \min\, \{1 \leq k \leq k_0; \quad H(Q_k) - H(Q) \leq \lambda\}. \tag{22}$$

The only apparent difference between $\hat{k}_n$ and k_n^* is in the choice of the thresholds; the constant threshold λ in (22) is replaced by a decaying threshold ϵ_n in (20). However, this difference in thresholds leads to significant differences in the behavior of the two estimators. In [4, p. 1016, Theorem 1], the authors assert that k_n^* is optimal in that

$$\overline{\lim_n} \frac{1}{n} \log P(k_n^*(X^n) < k) \le \underline{\lim_n} \frac{1}{n} \log P(k'_n(X^n) < k) \tag{23}$$

for all $P \in \mathcal{P}_k$, $1 \le k \le k_0$, in the class of estimators k'_n satisfying (21). However, a scrutiny of the proof reveals that they have instead established this optimality for the much more restricted class of estimators k'_n for which

$$\overline{\lim_n} \max_{1 \le k \le k_0} \sup_{P \in \mathcal{P}_k} \frac{1}{n} \log P(k'_n(X^n) > k) < -\lambda. \tag{24}$$

Next, since k_n^* satisfies (21), note that the Corollary to Proposition 1 above renders it inconsistent, as observed in [4, p. 1017, Remark 1(b)]. The authors further state in [4, p. 1017, Remark 1(a)] that there exists $P \in \mathcal{P}_k$ (depending on λ), $1 \le k \le k_0$, for which $P(k_n^*(X^n) < k)$ decays exponentially; the exponent is not explicitly given. Note that the previous remark does not contradict our Proposition 1.

Thus it appears that a restriction to the class of estimators specified in [4] by (21) (or, accurately, (24)) leads to an estimator k_n^*, which is optimal in the sense of (23); however, in general, k_n^* will underestimate with probability one. By widening our search over the larger class of estimators specified by (3), it is possible to find a consistent estimator $\hat{k}_n$, as a slight modification of k_n^*, which additionally has an optimally decaying probability of underestimation.

IV. Mixtures for Optimal Order Estimation

In this section, we present an alternative optimal estimator based on mixture distributions. This approach is appealing in that the statistic involved can be updated recursively. Furthermore, it affords a connection with the Minimum Description Length (MDL) [5], [7] and other penalized maximum-likelihood techniques.

We introduce the notion of a mixture distribution for $\mathcal{P}_k$, $1 \le k \le k_0$, as follows. Let ν_k be the density on $\overline{\Psi}_k$ (the set of *all* $r^k \times r$ t.p.m.'s) obtained as the product of r^k independent Dirichlet priors applied to the rows of each t.p.m. Namely, for $p = \{p(a|s),\ s \in \mathcal{X}^k,\ a \in \mathcal{X}\} \in \overline{\Psi}_k$

$$\nu_k(p) := \prod_{s \in \mathcal{X}^k} \left[\frac{\Gamma(r/2)}{(\Gamma(1/2))^r} \prod_{a \in \mathcal{X}} (p(a|s))^{-1/2} \right] \tag{25}$$

where Γ denotes the gamma function. The corresponding mixture distribution M_k on $\mathcal{X}^n$, $n \ge 1$, is defined by

$$M_k(x^n) := \int_{\overline{\Psi}_k} P(x^n) \nu_k(p)\, dp$$

where P is as defined in Section II. Observe that $M_k(x^n) > 0$ for all $x^n \in \mathcal{X}^n$.

As stated above, the mixture distribution M_k can be updated recursively. Namely, as shown in [10], [11], [18], the mixtures M_k, $1 \le k \le k_0$, can be expressed as

$$M_k(x^n) := \prod_{t=1}^{n} M_k(x_t | x^{t-1})$$

where

$$M_k(x_t | x^{t-1}) := \frac{q_{sa}^{(t-1)} + \frac{1}{2}}{q_s^{(t-1)} + \frac{r}{2}} \quad \text{if } x_{t-k}^{t-1} = s,\ x_t = a$$

and $q_{sa}^{(t-1)}$ is obtained from the kth Markov type of x^{t-1}.

The close relationship between the mixture M_k and the penalized likelihood for $\overline{\Psi}_k$ is indicated by the following lemma. Let

$$\hat{P}_k(x^n) := \sup_{p \in \overline{\Psi}_k} P(x^n) \tag{26}$$

be the maximum likelihood of $x^n \in \mathcal{X}^n$.

Lemma 3 ([10], [11], [18]): For $1 \le k \le k_0$, the mixture M_k satisfies for each $x^n \in \mathcal{X}^n$

$$\log \frac{1}{\hat{P}_k(x^n)} \le \log \frac{1}{M_k(x^n)} \le \log \frac{1}{\hat{P}_k(x^n)} + \frac{r^k(r-1)}{2} \log n + c_k$$

for all $n \ge N(k)$, where the constant c_k depends only on k. □

The left inequality in the lemma is trivial. The right inequality, whose proof is omitted, follows in a straightforward manner from results in [10], [11], [18], which were used to develop bounds on the pointwise redundancy of a universal code for noiseless data compression. Observe that the right-hand side in Lemma 3 corresponds to the description length to be minimized in Rissanen's MDL principle (cf. [7]); namely, consisting of the negative maximum loglikelihood penalized by $\{[r^k(r-1)]/2\} \log n$. This motivates the selection of the following estimator based on the mixture distributions M_k, $1 \le k \le k_0$:

$$k_n^M(x^n) := \min \arg \min_{1 \le k \le k_0} \left[\log \frac{1}{M_k(x^n)} + \frac{r^k(r-1)}{2} \log n \right]. \tag{27}$$

Theorem 3 below shows that the previous estimator is similar in behavior to the divergence based estimator $\hat{k}_n$ of (5) with the exception that its overestimation probability may decay at a slower than polynomial rate in n. However, k_n^M retains the optimality properties of $\hat{k}_n$ in the sense of satisfying (3) and (4), and achieving the optimal exponent of decay for the probability of underestimation. This is shown by the following

Theorem 3: For each $P \in \mathcal{P}_k$, $1 \le k \le k_0$

a) $P(\overline{\lim}_n k_n^M(X^n) > k) = 0$.

b) For each $\eta > 0$, and $n \ge N(\eta, p)$

$$P(k_n^M(X^n) < k) \le \exp\left[-n \left(\min_{k' < k} D(\overline{\Theta}_{k'} \| p) - \eta \right) \right].$$

Corollary: The estimator k_n^M is strongly consistent.

Proof of Theorem 3:

a) Fix $k' > k$ and $P \in \mathcal{P}_k$. Then

$$\begin{aligned} &P(k_n^M(X^n) = k') \\ &\quad \le P\left(\log \frac{M_{k'}(X^n)}{M_k(X^n)} > \frac{(r^{k'} - r^k)(r-1)}{2} \log n\right) \\ &\quad \le P\left(\log \frac{\hat{P}_{k'}(X^n)}{P(X^n)} > \frac{(r^{k'} - 2r^k)(r-1)}{2} \log n\right) \end{aligned}$$

by Lemma 3

$$\le P\left(\frac{1}{\log n} \log \frac{\hat{P}_{k'}(X^n)}{P(X^n)} > (r^{k'} - r^k)(r-1)\right). \quad (28)$$

From [19] it is known that for $k' > k$

$$P\left(\overline{\lim_n} \frac{1}{\log \log n} \log \frac{\hat{P}_{k'}(X^n)}{P(X^n)} \le \frac{r^{k'}(r-1)}{2}\right) = 1$$

whence

$$P\left(\lim_n \frac{1}{\log n} \log \frac{\hat{P}_{k'}(X^n)}{P(X^n)} = 0\right) = 1. \quad (29)$$

Hence, from (28) and (29), it holds for each $k' > k$ that

$$P\left(\overline{\lim_n}\, k_n^M(X^n) = k'\right) = 0$$

whence

$$P\left(\overline{\lim_n}\, k_n^M(X^n) > k\right) = 0.$$

b) Fix $k' < k$ and $P \in \mathcal{P}_k$. Then

$$\begin{aligned} &P(k_n^M(X^n) = k') \\ &\quad \le P\left(\log \frac{M_k(X^n)}{M_{k'}(X^n)} \le \frac{(r^k - r^{k'})(r-1)}{2} \log n\right) \\ &\quad \le P\left(\log \frac{\hat{P}_k(X^n)}{\hat{P}_{k'}(X^n)} < \frac{(2r^k - r^{k'})(r-1)}{2} \log n\right) \\ &\quad = P\left(\frac{1}{n} \log \frac{\hat{P}_k(X^n)}{\hat{P}_{k'}(X^n)} \le \frac{(2r^k - r^{k'})(r-1)}{2} \frac{\log n}{n}\right) \\ &\quad = P\left(\frac{1}{n} \log \frac{\hat{P}_k(X^n)}{\hat{P}_{k'}(X^n)} \le \rho_n\right) \end{aligned} \quad (30)$$

for all $n \ge N(k)$ by Lemma 3, where

$$\rho_n := \frac{(2r^k - r^{k'})(r-1)}{2} \frac{\log n}{n} > 0.$$

The remainder of the proof is similar to that of Theorem 1b). To see this, first note that $(1/n) \log \hat{P}_k(x^n) = -H(Q_k)$ where Q_k is the kth Markov type of x^n defined in a manner analogous to the k_0th type in Section II. Thus

$$\frac{1}{n} \log \frac{\hat{P}_k(x^n)}{\hat{P}_{k'}(x^n)} = H(Q_{k'}) - H(Q_k) = \inf_{p' \in \Theta_{k'}} D(Q_k \| p') \quad (31)$$

where the second equality follows from (19). In analogy with the set $\mathcal{B}_{k'}^n$ used in the proof of Theorem 1b), we define

$$\mathcal{C}_{k'}^n := \{Q_k \in \mathcal{Q}^{(n)};\quad D(Q_k \| p') \le \rho_n \text{ for some } p' \in \overline{\Theta}_{k'}\}. \quad (32)$$

Then (30)–(32) yield, for $n \ge N'(\eta, p)$, that

$$\begin{aligned} P(k_n^M(X^n) = k') &\le P\left(\bigcup_{\mathcal{C}_{k'}^n} \mathcal{T}_{Q_k}\right) \\ &\le \exp\left[-n(D(\overline{\Theta}_{k'} \| p) - \eta)\right] \end{aligned}$$

where the previous inequality follows similarly as in the proof of Theorem 1b). Finally

$$\begin{aligned} P(k_n^M(X^n) < k) &= \sum_{k' < k} P(k_n^M(X^n) = k') \\ &\le (k-1) \exp\left[-n\left(\min_{k' < k} D(\overline{\Theta}_{k'} \| p) - \eta\right)\right] \end{aligned}$$

for $n \ge N''(\eta, p)$, from which the desired result follows.

The Corollary follows by a standard application of the Borel–Cantelli lemma. □

Remarks:

i) As suggested by Lemma 3, k_n^M is closely linked with the penalized maximum likelihood (or MDL [7]) estimator k_n^{MDL} defined as follows:

$$k_n^{\mathrm{MDL}}(x^n) := \min \arg \min_{1 \le k \le k_0} \left[\log \frac{1}{\hat{P}_k(x^n)} + \frac{r^k(r-1)}{2} \log n\right].$$

As is evident from its proof, Theorem 3 and its corollary apply to k_n^{MDL} as well.

ii) It is worth noting the effect, on the asymptotic performance of k_n^M, of the additive compensation term $\{[r^k(r-1)]/2\} \log n$ in (27). First a deletion of this term from both sides of (27) does not affect the asymptotic probability of underestimation, i.e., Theorem 3b) still holds. Next, a polynomial decay (in n) of overestimation probability can be achieved by suitably augmenting the penalty term in (27). Specifically, consider the estimator $\tilde{k}_n^M$ defined by modifying (27) with an increased penalty term of $((r^k(r-1)/2) + f(k)) \log n$, where f satisfies $f(k+1) - f(k) = \delta > 0$, $1 \le k < k_0$. Theorem 4 below shows that the overestimation probability of $\tilde{k}_n^M$ decays as $n^{-\delta}$; its underestimation probability, of course, is given by Theorem 3b). (Clearly, remark ii) applies to k_n^{MDL} as well.)

iii) There exist alternative mixture-based estimators which are comparable in performance to $\tilde{k}_n^M$. For instance, consider the estimator proposed in [8, eq. (3.7)] which, given a sample $x^n \in \mathcal{X}^n$, yields the estimate k iff k is the largest integer in $\{1, \cdots, k_0\}$ such that

$$\log \frac{M_k(x^n)}{M_{k-1}(x^n)} > \left(\frac{r^k(r-1)}{2} + \delta\right) \log n$$

(with the convention $M_0(x^n) = 1$). This estimator asymptotically performs similarly to $\tilde{k}_n^M$ above.

We conclude this section by considering the overestimation probability for the estimator $\tilde{k}_n^M$ of remark ii) above.

Theorem 4: For each $P \in \mathcal{P}_k$, $1 \leq k < k_0$, it holds that

$$P(\tilde{k}_n^M(X^n) > k) \leq n^{-\delta}, \quad n \geq N(k).$$

Proof: Fix $P \in \mathcal{P}_k$ and $k' > k$. Then

$$\begin{aligned} &P\left(\tilde{k}_n^M(X^n) = k'\right) \\ &\leq \sum_{x^n} P(x^n) \cdot \mathbf{1}\left(\frac{M_{k'}(x^n)}{M_k(x^n)} > n^{\frac{(r^{k'}-r^k)(r-1)}{2}+f(k')-f(k)}\right) \\ &\leq \sum_{x^n} M_k(x^n)^{\frac{r^k(r-1)}{2}} \\ &\quad \cdot \mathbf{1}\left(\frac{M_{k'}(x^n)}{M_k(x^n)} > n^{\frac{(r^{k'}-r^k)(r-1)}{2}+f(k')-f(k)}\right) \\ &\leq \sum_{x^n} M_{k'}(x^n) n^{-\left[\frac{(r^{k'}-r^k)(r-1)}{2}+f(k')-f(k)\right]} n^{r^k(r-1)/2} \\ &= n^{-\left[\frac{((r^{k'}-2r^k)(r-1)}{2}+f(k')-f(k)\right]} \end{aligned}$$

for all $n \geq N(k)$, where the second inequality is a consequence of Lemma 3.

Since

$$\frac{(r^{k'}-2r^k)(r-1)}{2} + f(k') - f(k) \geq \delta \qquad \text{for all } r > 2,\ 1 \leq k < k_0$$

the assertion of the theorem readily follows. □

V. Discussion

The estimator $\hat{k}_n$ of (5) relies in an essential way on the Markov order being bounded by a *known* integer k_0. If the order is bounded but k_0 is *unknown*, our approach is not directly applicable. For this case, Merhav–Gutman–Ziv propose an alternative estimator k_n^{**} [4, p. 1016, eq. (14)] defined similarly as k_n^* but with $H(Q)$ in (22) replaced by the normalized Lempel–Ziv codeword length. It is shown in [4] that k_n^{**} shares the asymptotic properties of k_n^*. It can be additionally shown that replacing the constant threshold λ in k_n^{**} by the decaying threshold ϵ_n of $\hat{k}_n$ in (5) will render the former consistent.

If the maximum allowable order k_0 is not *fixed* but allowed to vary with the sample size n, we conjecture that a suitable modification of $\hat{k}_n$ will still yield the asymptotic performance given by Theorem 1, provided $k_0(n)$ grows no faster than $O(\log n)$.

Next, we turn to the problem of estimating consistently the order of a hidden Markov source (HMS). In [8], a strongly consistent estimator based on mixture distributions (cf. Remark iii) of Section IV) is presented with an exponentially decaying underestimation probability, but the corresponding exponent is not explicitly determined. The optimal error exponent for HMS order is as yet unknown, and the approach of this paper, relying on Markov types, does not directly extend to the HMS. However, recent work by Ziv–Merhav [9] in developing the notion of "generalized types" for finite-state processes, offers some hope in this direction.

Appendix
Proof of Theorem 1a)

Fix k and $P \in \mathcal{P}_k$. Let

$$\mathcal{A}_k^n := \left\{ Q \in \mathcal{Q}^{(n)}; \quad D(Q||p') > \epsilon_n \quad \forall p' \in \bigcup_{l=1}^{k} \Theta_l \right\}.$$

Then

$$\begin{aligned} P(\hat{k}_n(X^n) > k) &\leq P\left(\bigcup_{\mathcal{A}_k^n} \mathcal{T}_Q\right) \\ &\leq P\left(\bigcup_{Q \in \mathcal{A}_k^n:\, D(Q||p) > \epsilon_n} \mathcal{T}_Q\right) \\ &= \sum_{Q \in \mathcal{A}_k^n:\, D(Q||p) > \epsilon_n} P(\mathcal{T}_Q) \\ &\leq \sum_{Q \in \mathcal{A}_k^n:\, D(Q||p) > \epsilon_n} r^{k_0} \exp\left[-nD(Q||p)\right] \\ &\leq \begin{pmatrix} n + r^{k_{0+1}} & -1 \\ r^{k_{0+1}} & -1 \end{pmatrix} r^{k_0} \exp\left[-n\epsilon_n\right]. \end{aligned}$$

Since

$$\begin{pmatrix} n + r^{k_0+1} & 1 \\ r^{k_0+1} & -1 \end{pmatrix} < n^{r^{k_0}}$$

for all n sufficiently large, we can continue the bounding above as

$$P(\hat{k}_n(X^n) > k) \leq n^{r^{k_0}} r^{k_0} \exp\left[-n\epsilon_n\right]$$

for all n large. By our choice of ϵ_n, we finally get

$$P(\hat{k}_n(X^n) > k) \leq r^{k_0} n^{-(1+\delta)}, \quad n \geq N(\delta). \qquad \square$$

Acknowledgment

The authors wish to thank S. Khudanpur for several helpful comments and for an observation which led to a sharpening of Theorem 1.

References

[1] P. Billingsley, "Statistical methods in Markov chains," *Ann. Math. Statist.*, vol. 32, pp. 12–40; Correction: p. 1343, 1961.

[2] C. Chatfield, "Statistical inference regarding Markov chain models," *Appl. Statist.*, vol. 22, pp. 7–20, 1973.

[3] J. C. Kieffer, "Strongly consistent code-based identification and order estimation for constrained finite-state model classes," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 893–902, May 1993.

[4] 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, 1989.

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

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

[7] J. Rissanen, *Stochastic Complexity in Statistical Enquiry.* Singapore: World Scientific, 1989.

[8] C. Liu and P. Narayan, "Order estimation and sequential universal data compression of a hidden Markov source via the method of mixtures," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1167–1180, July 1994.

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

[10] 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, 1981.

[11] Y. M. Shtar'kov, "Universal sequential coding of separate messages," *Probl. Pered. Inform.*, vol. 23, pp. 3–17, 1987.

[12] L. D. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 431–438, 1981.

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

[14] V. Anantharam, "A large deviation approach to error exponents in source coding and hypothesis testing," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 938–943, 1981.

[15] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications.* Boston, MA: Jones and Bartlett, 1993.

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

[17] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time I–III," *Comm. Proc. App. Math.*, vol. 28, pp. 1–47, 279–301, and vol. 29, pp. 389–461, 1975/1976.

[18] I. Csiszár, "Information theoretical methods in statistics," Class notes, University of Maryland, College Park, MD, Spring 1990.

[19] L. Finesso, "Estimation of the order of a finite Markov chain," in *Recent Advances in the Mathematical Theory of Systems, Control, Networks Signals, Proc. MTNS-91*, H. Kimura and S. Kodama, Eds. Mita Press, 1992, pp. 643–645.