Estimating the Number of States of a Finite-State Source

Jacob Ziv, *Fellow, IEEE,* and Neri Merhav, *Member, IEEE*

Abstract—**The problem of estimating the number of states of a finite-alphabet, finite-state source is investigated. An estimator is developed that asymptotically attains the minimum probability of underestimating the number of states, among all estimators with a prescribed exponential decay rate of overestimation probability. The proposed estimator relies on the Lempel-Ziv data compression algorithm in an intuitively appealing manner.**

Index Terms—**Model order estimation, finite-state sources, hidden Markov models, universal data compression, Lempel-Ziv algorithm.**

I. Introduction

In [1], the estimation of the order k of a finite-alphabet Markov source was studied. An order estimator k^* was developed and shown to be asymptotically optimal in the sense of having an underestimation probability $\Pr\{k^* < k\}$ smaller than that of any estimator $\hat{k}$ for which the overestimation probability $\Pr\{\hat{k} > k\}$ decays faster than $2^{-\lambda n}$ for some given $\lambda > 0$, where n is the sample size. This is a generalized version of the Neyman-Pearson criterion.

In this paper, the results of [1] are extended to the estimation of the number of states of a finite-alphabet, finite-state (FS) source. Specifically, let $\boldsymbol{x} = x_1, x_2, \cdots, x_n$ be a sequence of observable random variables taking on values in a finite set X of size $|X| = X$. Similarly, let $\boldsymbol{s} = s_1, s_2, \cdots, s_n$ be another sequence of random variables (states), corresponding to $\boldsymbol{x}$, which take on values in another finite set S_M of size $|S_M| = M$. An information source P is said to be *finite-state* (with M states) if the joint probability of $\boldsymbol{x}$ and $\boldsymbol{s}$ is given by

$$P(\boldsymbol{x}, \boldsymbol{s}) = \prod_{i=1}^{n} p(x_i, s_i \mid s_{i-1}), \tag{1}$$

where the initial state $s_0 \in S_M$ is assumed fixed and known, and $p(x_i, s_i \mid s_{i-1})$ is the joint probability of a letter x_i and a state s_i at time instant i given the previous state s_{i-1} at time instant $i-1$. The state sequence $\boldsymbol{s}$ is not apparent in general (in contrast to the Markovian case [1]). Let $\mathbb{P}_M$ denote the class of all FS sources with at most M states. We are interested in an estimator $M^* = M^*(\boldsymbol{x})$ for the number of states M, that is asymptotically optimal in the following sense. Minimize $\Pr\{\hat{M} < M\}$ uniformly for all M and every $P_M \in \mathbb{P}_M$, subject to the constraint

$$\liminf_{n\to\infty} \left[-\frac{1}{n}\log \Pr\{\hat{M} > M\}\right] > \lambda, \qquad \forall P \in \mathbb{P}_M, \tag{2}$$

where $\lambda > 0$ is a given number and logarithms are taken to the base 2 unless specified otherwise.

The main difficulty in generalizing the result of [1] from the class of Markov sources to the more general class of FS sources is that here the data cannot be summarized by a finite dimensional vector of sufficient statistics which allows one to focus on relatively simple classes of estimators without sacrificing optimality. While the proof in [1] relies heavily on the fact that Markov types are sufficient statistics in the Markovian case, here more powerful techniques are required. It is pointed out, on the other hand, that for some important subclasses of FS sources, e.g., hidden Markov sources, unifilar FS sources, Markov sources, we are able to improve the performance of our estimator by utilizing more prior knowledge about the true underlying model.

II. Main Result

Define the following estimator for the number of states M.

$$M^* = \min\left\{j : -\frac{1}{n}\log \max_{P\in\mathbb{P}_j} P(\boldsymbol{x}) - \frac{1}{n}U_{\mathrm{LZ}}(\boldsymbol{x}) < \lambda\right\}, \tag{3}$$

where $U_{\mathrm{LZ}}(\boldsymbol{x})$ is the length (in bits) of the Lempel-Ziv (LZ) codeword [2] for $\boldsymbol{x}$ and $P(\boldsymbol{x}) = \sum_s P(\boldsymbol{x}, \boldsymbol{s})$, with $P(\boldsymbol{x}, \boldsymbol{s})$ being defined as in (1). The maximization of $P(\boldsymbol{x})$ over $\mathbb{P}_j$ is usually carried out by iterative techniques, e.g., the EM algorithm [3], which merely guarantee convergence to a local rather than a global maximum of $P(\boldsymbol{x})$. An alternative approach, which is computationally unattractive, is an exhaustive search over a dense grid of sources in $\mathbb{P}_j$, which may grow polynomially fast with n. Observe that M^* is a generalized version of the estimator proposed in [1] for the Markovian case. It has the following intuitive interpretation. We seek the smallest model order j for encoding $\boldsymbol{x}$, such that the codeword length $-\log P(\boldsymbol{x})$ will be sufficiently close (difference less than λn) to the codeword length associated with the LZ algorithm, which in turn, serves as an estimate of the source entropy. Our main result is the following.

Manuscript received April 26, 1989; revised February 5, 1991. This work was supported in part by the U.S.-Israel Binational Science Foundation. This work was presented in part at the IEEE International Symposium on Information Theory, San Diego, CA, January 14-19, 1991.

The authors are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9103708.

0018-9448/92$03.00 © 1992 IEEE

Theorem 1: The estimator M^* satisfies the following conditions.

a) $\liminf_{n\to\infty}\left[-\frac{1}{n}\, log\, Pr\{M^* > M\}\right] \geq \lambda, \quad \forall P_M \in \mathbb{P}_M.$

b) For any competing estimator $\hat{M}$ that satisfies (2), for every M-state source $P \in \mathbb{P}_M$, and for all large n,

$$\Pr\{M^* < M\} \leq \left(1 + \frac{1}{n}\right)\Pr\{\hat{M} < M\}.$$

The theorem tells us that if the underestimation probability happens to decay exponentially with n (see, e.g., [1, p. 1017, Remark 1]), then the asymptotic underestimation error exponent of M^* is optimal. If, however, the overestimation probability does not decay exponentially, then still it decays at the highest possible rate or it tends to the minimum value attainable. The term $1/n$ on the right-hand side of b) is somewhat arbitrary and can be replaced, more generally, by any positive α_n that decays with n in a subexponential rate, i.e., $n^{-1}\log \alpha_n \to 0$ as $n \to \infty$. The choice of $1/n$ is for the sake of simplicity and convenience. Note that M^* does not necessarily satisfy (2) with strict inequality. In a sense, this means that M^* is asymptotically ϵ-optimal rather than asymptotically optimal, as the strict inequality (2) is satisfied if λ is replaced by $(\lambda - \epsilon)$ for arbitrarily small $\epsilon > 0$.

Proof of Theorem 1: As for Part a), define

$$N_j \triangleq \left\{x : -\frac{1}{n}\log \max_{P\in\mathbb{P}_j} P(x) - \frac{1}{n}U_{\mathrm{LZ}}(x) < \lambda\right\}, \qquad j = 1, 2, \cdots. \quad (4)$$

Then we have

$$\begin{aligned}\Pr\{M^* > M\} &= \Pr\left\{\bigcap_{j=1}^{M} N_j^c\right\} \leq \Pr\{N_M^c\}\\ &\leq \sum_{x\in N_M^c}\max_{P\in\mathbb{P}_M} P(x)\\ &\leq \sum_{x\in N_M^c} 2^{-\lambda n - U_{\mathrm{LZ}}(x)}\\ &\leq 2^{-\lambda n}\sum_{x\in X^n} 2^{-U_{\mathrm{LZ}}(x)} \leq 2^{-\lambda n}, \quad (5)\end{aligned}$$

where X^n is the nth Cartesian power of X. The last step in (5) follows from the Kraft inequality [4] for binary, uniquely decipherable codeword length functions. This completes the proof of Part a).

To prove Part b), select an arbitrary competing estimator $\hat{M}$ that satisfies (2). Assume further that $\hat{M}$ relies on knowledge of an integer M_0 that upper bounds the true number of states M. It will be shown that the estimator M^* is no worse than $\hat{M}$ in spite of the fact that the former does not require knowledge of M_0. Let $\{\Omega_j\}_{j=1}^{M_0}$ denote the partition of X^n induced by $\hat{M}$, that is, $\Omega_j = \{x : \hat{M} = j\}$, $j = 1, 2, \cdots, M_0$. It follows from (2) that for any $j \leq M$, $\epsilon > 0$ and n sufficiently large,

$$2^{-(\lambda+\epsilon)n} \geq \max_{P\in\mathbb{P}_j}\sum_{j>M}\sum_{x\in\Omega_j} P(x). \quad (6)$$

Next assume that l divides n and sparse x into n/l nonoverlapping l-blocks $x_i = (x_{(i-1)l+1}, \cdots, x_{il})$, $i = 1, 2, \cdots, n/l$. Let s^l denote the sequence of initial states of the resulting blocks x_i, i.e.,

$$s^l = s_0^l, s_1^l, s_2^l, \cdots, s_{n/l}^l, \quad (7)$$

where $s_i^l = s_{(i-1)l}$, $i = 1, 2, \cdots, n/l + 1$. Henceforth, the sequence s^l will be referred to as the *sparse* state sequence. Given s^l, let $K(x \mid s^l)$ denote the set of all n-vectors x' generated by permuting phrases x_i with phrases x_j (of x) for which $s_i^l = s_j^l$ and $s_{i+1}^l = s_{j+1}^l$, namely, permuting phrases with the same initial and final states. Since $P(x, s^l) = \Pi_{i=1}^{n/l} P(x_i, s_{i+1}^l \mid s_i^l)$ and products are unaffected by permutations, it follows that for any $x' \in K(x \mid s^l)$

$$P(x', s^l) = P(x, s^l), \qquad \forall P \in \mathbb{P}_j, \quad j = 1, 2, \cdots. \quad (8)$$

Hence, $K(x \mid s^l)$ can be thought of as a conditional type of x given s^l. We now generate from $\{\Omega_j\}_{j=1}^{m_0}$ an auxiliary partition $\{\tilde{\Omega}_j\}_{j=1}^{M_0}$ of the sample space of pairs (x, s^l) such that each decision region $\tilde{\Omega}_j$ contains a sufficiently large fraction of vectors x from the same conditional type $K(x \mid s^l)$. This will be useful later when we apply a lower bound on $|K(x \mid s^l)|$ (see Lemma 1). Specifically, we use the following rules to create the auxiliary partition.

a) For every pair (x, s^l) where $x \in \Omega_i$ and $|K(x \mid s^l) \cap \Omega_i| > (nM_0)^{-1}|K(x \mid s^l)|$, let $(x, s^l) \in \tilde{\Omega}_i$, $i = 1, 2, \cdots, M_0$.

b) For every (x, s^l) where $x \in \Omega_i$ and $|K(x \mid s^l) \cap \Omega_i| \leq (nM_0)^{-1}|K(x \mid s^l)|$, let $(x, s^l) \in \tilde{\Omega}_j$, where j is the smallest integer that maximizes $|K(x \mid s^l) \cap \Omega_j|$.

Since $\max_j |K(x \mid s^l) \cap \Omega_j| \geq M_0^{-1}|K(x \mid s^l)|$, it follows from the previous construction that $|K(x \mid s^l) \cap \tilde{\Omega}_j(s^l)| \leq (1 + 1/n)|K(x \mid s^l) \cap \Omega_j|$, where $\tilde{\Omega}_j(s^l)$ is defined as the collection of sequences x such that $(x, s^l) \in \tilde{\Omega}_j$. Next, observe that $K(x \mid s^l)$ and $K(x' \mid s^l)$ are disjoint whenever $x' \notin K(x \mid s^l)$. Hence, for every $1 \leq j \leq M_0$, we have

$$\begin{aligned}\Pr\{\tilde{\Omega}_j\} &= \sum_{s^l}\Pr\left\{\bigcup_{K(x\mid s^l)\subset X^n} K(x \mid s^l) \cap \tilde{\Omega}_j(s^l)\right\}\\ &= \sum_{s^l}\sum_{K(x\mid s^l)\subset X^n} P(x, s^l)\,|K(x \mid s^l) \cap \tilde{\Omega}_j(s^l)|\\ &\leq \left(1 + \frac{1}{n}\right)\sum_{s^l}\sum_{K(x\mid s^l)\subset X^n} P(x, s^l)\\ &\quad \cdot |K(x \mid s^l) \cap \Omega_j|\\ &= \left(1 + \frac{1}{n}\right)\Pr\{\Omega_j\}. \quad (9)\end{aligned}$$

Let $\tilde{M}$ denote the estimator of M induced by $\{\tilde{\Omega}_j\}_{j\geq 1}$, i.e., $\tilde{M}(x, s^l) = j$ iff $(x, s^l) \in \tilde{\Omega}_j$. By construction of $\{\tilde{\Omega}_j\}_{j\geq 1}$, we have $|K(\boldsymbol{x} \mid \boldsymbol{s}^l) \cap \tilde{\Omega}_j(\boldsymbol{s}^l)| \geq (nM_0)^{-1} |K(\boldsymbol{x} \mid \boldsymbol{s}^l)|$ for every $1 \leq j \leq M_0$. Thus, from (6) and (9), for every $P \in \mathbb{P}_j$ and every $j \leq M$,

$$\begin{aligned} 2^{-(\lambda+\epsilon)n} &\geq \max_{P \in \mathbb{P}_j} \Pr\{\hat{M} > M\} \\ &\geq \left(1 + \frac{1}{n}\right)^{-1} \max_{P \in \mathbb{P}_j} \Pr\{\tilde{M} > M\} \\ &= \left(1 + \frac{1}{n}\right)^{-1} \\ &\quad \cdot \max_{P \in \mathbb{P}_j} \sum_{j=M+1}^{M_0} \sum_{\hat{s}^l} \sum_{K(\hat{x} \mid \hat{s}^l) \subset X^n} |K(\hat{\boldsymbol{x}} \mid \hat{\boldsymbol{s}}^l) \\ &\quad \cap \tilde{\Omega}_j(\hat{\boldsymbol{s}}^l)| \, P(\hat{\boldsymbol{x}} \mid \hat{\boldsymbol{s}}^l) \\ &\geq \left(1 + \frac{1}{n}\right)^{-1} \max_{P \in \mathbb{P}_j} P(\boldsymbol{x}, \boldsymbol{s}^l) \frac{1}{nM_0} |K(\boldsymbol{x} \mid \boldsymbol{s}^l)|, \end{aligned} \tag{10}$$

for every $(\boldsymbol{x}, \boldsymbol{s}^l) \in \cup_{j>M} \tilde{\Omega}_j$. Unfortunately, the right-most side of (10) contains two quantities, $K(\boldsymbol{x} \mid \boldsymbol{s}^l)$ and $\max_{P \in \mathbb{P}_j} P(\boldsymbol{x}, \boldsymbol{s}^l)$, that depend on the unavailable state sequence $\boldsymbol{s}^l$. However, since $\boldsymbol{s}^l$ is sparse for large l, then intuitively, it carries very little information. Indeed, the following two lemmas provide lower bounds on these two quantities, that are independent of $\boldsymbol{s}^l$ and hence will be useful for deriving M^*, which in turn does not allow dependence on $\boldsymbol{s}^l$. The first lemma is, in fact, a generalized version of the well-known fact that the cardinality of a type is exponentially underbounded in terms of its associated empirical entropy [5, p. 30, Lemma 2.3], which in turn is further underestimated by the LZ codeword length function [6].

Lemma 1: For every $\boldsymbol{x} \in X^n$ and $\boldsymbol{s}^l \in S_{M_0}^{n/l+1}$,

$$|K(\boldsymbol{x} \mid \boldsymbol{s}^l)| \geq 2^{U_{\mathrm{LZ}}(\boldsymbol{x}) - n\epsilon_1(n,l)}, \tag{11}$$

where

$$\epsilon_1(n, l) = \frac{C_1}{l} + \frac{C_2 l}{\log n} + \frac{C_3 l X^l}{n} \tag{12}$$

for some positive constants C_1, C_2, and C_3 depending only on X and M_0.

The proof of the Lemma 1, which is based on techniques similar to those developed in [6], [7] can be found in [8, Appendix A] (see (A.15), (A.16) therein). The next lemma tells us that $\max_{P \in \mathbb{P}_j} P(\boldsymbol{x})$ and $\max_{P \in \mathbb{P}_j} P(\boldsymbol{x}, \boldsymbol{s}^l)$ are exponentially equivalent for large l. This result allows us to underbound $\max_{P \in \mathbb{P}_j} P(\boldsymbol{x}, \boldsymbol{s}^l)$ of (10) in terms of $\max_{P \in \mathbb{P}_j} P(\boldsymbol{x})$, which is in turn independent of the unavailable $\boldsymbol{s}^l$.

Lemma 2: For every $\boldsymbol{x} \in X^n$, $\boldsymbol{s}^l \in S_j^{n/l+1}$ and $j \leq M$,

$$\max_{P \in \mathbb{P}_j} P(\boldsymbol{x}, \boldsymbol{s}^l) \geq 2^{-n\epsilon_2(l)} \max_{P \in \mathbb{P}_j} P(\boldsymbol{x}), \tag{13}$$

where

$$\epsilon_2(l) = \frac{1}{l} \log\left[\frac{e}{4}(l+2)^2 M^5 X^4\right]. \tag{14}$$

The proof of Lemma 2 appears in the Appendix.

Combining (10), (11), and (13), we get

$$2^{-(\lambda+\epsilon)n} \geq \max_{P \in \mathbb{P}_j} P(\boldsymbol{x}) 2^{U_{\mathrm{LZ}}(\boldsymbol{x}) - n\epsilon_3(n,l)}, \tag{15}$$

where

$$\epsilon_3(n, l) = \epsilon_1(n, l) + \epsilon_2(l) + \frac{1}{n}\log(nM_0) + \frac{1}{n}\log\left(1 + \frac{1}{n}\right). \tag{16}$$

By letting $l = l_n$ grow slowly with n in an appropriate rate, e.g., $l_n = O(\sqrt{\log n})$, the sequence $\{\epsilon_3(n, l_n)\}_{n \geq 1}$ will vanish as n grows indefinitely. Hence, for sufficiently large n, $\epsilon_3(n, l_n) \leq \epsilon$, and we conclude from (15) that for every $(\boldsymbol{x}, \boldsymbol{s}^{l_n}) \in \cup_{j>M} \tilde{\Omega}_j$ we have

$$-\log \max_{P \in \mathbb{P}_j} P(\boldsymbol{x}) - U_{\mathrm{LZ}}(\boldsymbol{x}) \geq \lambda n, \tag{17}$$

or, in other words, $(\boldsymbol{x}, \boldsymbol{s}^{l_n}) \in N_j^c$, where the superscript c denotes the complementary set. This means that $N_j \times S_{M_0}^{n/l_n+1} \subseteq \cup_{j \leq M} \tilde{\Omega}_j$ for every $j \leq M$ where $1 \leq M \leq M_0$. Hence,

$$\begin{aligned} \Pr\{M^* < M\} &= \Pr\{N_{M-1} \times S_{M_0}^{n/l_n+1}\} \leq \Pr\left\{\bigcup_{j \leq M-1} \tilde{\Omega}_j\right\} \\ &= \Pr\{\tilde{M} < M\} \leq \left(1 + \frac{1}{n}\right) \Pr\{\hat{M} < M\}, \end{aligned} \tag{18}$$

where the last step follows from (9). This completes the proof of Part b). □

III. Discussion

A slightly different version of Theorem 1 could have been obtained if we replaced the overestimation constraint (2) by

$$\Pr\{\hat{M} > M\} \leq 2^{-(\lambda+\epsilon)n}, \tag{19}$$

for all n, and some $\epsilon > 0$, which is a constraint somewhat stronger than (2) and hence, defines a smaller class of competing estimators. In this case, Part b) of Theorem 1 would have been reformulated in a slightly stronger manner as follows: For any competing estimator $\hat{M}$ that satisfies (19), for every M-state source $P \in \mathbb{P}_M$ and for all large n, $\Pr(M^* < M) \leq \Pr\{\hat{M} < M\}$. In other words, the factor $(1 + 1/n)$ in the original version of Theorem 1, Part b) would have been removed. The explanation of this fact is as follows: Consider an auxiliary problem of testing the hypothesis $H_0: M \leq j$ against the alternative $H_1: M > j$, where j is a given integer. Again, the Neyman–Pearson criterion is adopted, i.e., minimize Pr {accept $H_0 \mid H_1$} for a *given* $P \in \mathbb{P}_M$, $M > j$, subject to the false alarm constraint $\max_{P \in \mathbb{P}_j}$ Pr {reject $H_0 \mid H_0$} $\leq 2^{-(\lambda+\epsilon)n}$. Assume, temporarily, that the sparse state sequence $\boldsymbol{s}^l$ of (7) is available to the observer. Since the false alarm constraint is exponentially equivalent to

$$\sum_{(x, s^l)\in\Lambda_1} \max_{P'\in\mathbb{P}_j} P'(\boldsymbol{x}, \boldsymbol{s}^l) \leq 2^{-(\lambda+\epsilon)n}, \qquad (20)$$

where Λ_1 is the decision region for H_1, then by the Neyman-Pearson theorem [9, Theorem 1], [10], the optimal test compares the likelihood ratio n^{-1} log $[P(\boldsymbol{x}, \boldsymbol{s}^l)/\max_{P'\in\mathbb{P}_j} P'(\boldsymbol{x}, \boldsymbol{s}^l)]$ to a threshold function $T(\lambda)$. Observe that this optimal test depends on $(\boldsymbol{x}, \boldsymbol{s}^l)$ only through the conditional type $K(\boldsymbol{x} \mid \boldsymbol{s}^l)$, which can be thought of as sufficient statistics. This allows one to confine attention to universal tests that depend solely on $K(\boldsymbol{x} \mid \boldsymbol{s}^l)$ without loss of optimality and hence avoid the use of the modification $\{\tilde{\Omega}_j\}_{j=1}^{M_0}$ (see Section III), which in turn introduces the factor of $(1 + 1/n)$ in the original version of Theorem 1. Similarly to (10), (11), and (13), one obtains N_j as an asymptotically optimal acceptance region for H_1. Finally, observe that M^* implements the asymptotically optimal test of the above auxiliary hypothesis testing problem *simultaneously* for all positive integers j and hence minimizes the underestimation probability for all large n.

It should be pointed out that if the FS source is known to lie in some subclass $\mathbb{Q}_M \subseteq \mathbb{P}_M$ of FS sources, then $\mathbb{P}_j$, in the definition of M^* (see (3)), can be replaced by $\mathbb{Q}_j$, resulting in a smaller underestimation probability. Furthermore, prior knowledge of $\mathbb{Q}_j$ may considerably reduce computational complexity. Several important examples of $\mathbb{Q}_j$ are the following.

1) *Hidden Markov Sources* are FS sources where $p(x_i, s_i \mid s_{i-1})$ factorizes into a product $p(x_i \mid s_{i-1})$ $p(s_i \mid s_{i-1})$. The number of free parameters in this subclass is $j(X + j - 2)$, which may be considerably smaller than the $j(jX - 1)$ free parameters of the more general class $\mathbb{P}_j$. This saves a significant amount of computations when applying the EM algorithm [3] in order to calculate $\max_{P\in\mathbb{Q}_j} P(\boldsymbol{x})$. The hidden Markov model is used extensively in speech recognition applications (see, e.g., [11] and references therein).
2) *Unifilar FS sources* are FS sources where s_i is given by a deterministic *next state* function $g(x_i, s_{i-1})$. In this case, the underlying state sequence $\boldsymbol{s}$ can be determined recursively from $\boldsymbol{x}$ and s_0 and the resulting empirical joint distribution of letters and states, i.e., $q_j^g(x, s \quad) = n^{-1} \sum_{i=1}^n 1(x_i = x, s_i = s)$, $x\in\boldsymbol{X}$, $s\in\boldsymbol{S}_j$, (where $1(\cdot)$ denotes an indicator function,) serves as sufficient statistics. Specifically, here $-n^{-1}$ log $\max_{P\in\mathbb{Q}_j} P(\boldsymbol{x})$ is given by the conditional empirical entropy $\hat{H}_j^g(x \mid s)$ associated with $\{q_j^g(x, s)\}_{x\in\boldsymbol{X}, s\in\boldsymbol{S}_j}$, which can be calculated relatively easily. If the next state function $g(\cdot, \cdot)$ is unknown, then $\hat{H}_j^g(x \mid s)$ can be replaced by $\min_g \hat{H}_j^g(x \mid s)$, where the minimum is taken over all j^{jX} next state functions associated with j-state unifilar FS sources.
3) *FSMX sources* are unifilar FS sources where the current state s_i depends on no more than the k most recent source letters $x_{i-k+1}, x_{i-k+2}, \cdots, x_i$. In other words, an M-state FSMX source is characterized by a suffix set S_M (with M elements) such that any k-tuple w has a unique suffix suf (w) in S_M and $s_i =$ suf $(x_{i-k+1}, x_{i-k+2}, \cdots, x_i)$. The optimal estimator M^* relies on empirical entropies induced by the respective state sequence similarly to Example 2. For FSMX sources the present estimation approach is extended in [12], where an estimator is proposed for the states themselves, rather than just the number of states M, under a similar optimality criterion in the Neyman-Pearson spirit. This state estimator is then employed by a sequential universal data compression scheme and shown to asymptotically minimize the redundancy of the code.
4) *Markov sources* of order k are FSMX sources where the current state s_i depends exactly on the k most recent source letters, i.e., $s_i = (x_{i-k+1}, \cdots, x_i)$. Again, the estimation of k is similar to that described for the case of unifilar FS sources, where $\hat{H}_j^g(x \mid s)$ is replaced by the jth order empirical conditional entropy $\hat{H}(x \mid x^j)$ of a letter x given its j preceding letters (see [1] for more details).

In [1], it has been shown that if an upper bound k_0 on the true Markov order k is available to the observer, then the LZ codeword length $U_{\mathrm{LZ}}(\boldsymbol{x})$ can be replaced by the empirical conditional entropy of order k_0, $\hat{H}(x \mid x^{k_0})$, namely, an alternative asymptotically optimal estimator of k is given by

$$\begin{aligned} k^* &= \min\{j \colon \hat{H}(x \mid x^j) - \hat{H}(x \mid x^{k_0}) < \lambda\} \\ &= \min\left\{ j \colon \frac{1}{n} \log \frac{\max_{P\in\mathbb{Q}_{k_0}} P(\boldsymbol{x})}{\max_{P\in\mathbb{Q}_j} \colon P(\boldsymbol{x})} < \lambda \right\}, \end{aligned} \qquad (21)$$

where $\mathbb{Q}_j$ is the subclass of Markov sources. An interesting open problem is whether a similar result holds for general FS sources, in other words, is the estimator

$$M^{**} = \min\left\{ j \colon \frac{1}{n} \log \frac{\max_{P\in\mathbb{P}_{M_0}} P(\boldsymbol{x})}{\max_{P\in\mathbb{P}_j} P(\boldsymbol{x})} < \lambda \right\}, \qquad (22)$$

asymptotically optimal if M_0 is a given upper bound on M? This question, which is discussed in more depth in [13], is important as it may serve as a first step towards an extension of the above result to sources with continuous valued observations, where the LZ algorithm is not directly applicable but probability mass functions in (22) can be naturally substituted by probability density functions.

Appendix

Proof of Lemma 2: Fix $0 < \delta \leq (jX)^{-2}$ and let $\mathbb{P}_j^\delta \subset \mathbb{P}_j$ be the set of all j-state sources for which $p(x, s \mid \sigma) \triangleq \Pr\{x_i = x, s_i = s \mid s_{i-1} = \sigma\} \geq \delta$ for all $x\in\boldsymbol{X}$, s, $\sigma\in\boldsymbol{S}_j$. Let $\hat{P} = \{\hat{p}(x, s \mid \sigma)\}_{x,s,\sigma}$ be a source that maximizes $P(\boldsymbol{x})$ over $\mathbb{P}_j$. We first show that

$$\begin{aligned} \max_{P\in\mathbb{P}_j^\delta} P(\boldsymbol{x}) &\geq \left(1 - \delta j^2 X^2\right)^n \max_{P\in\mathbb{P}_j} P(\boldsymbol{x}) \\ &= \left(1 - \delta j^2 X^2\right)^n \hat{P}(\boldsymbol{x}). \end{aligned} \qquad (\mathrm{A}.1)$$

To see that (A.1) holds, consider a source $P' = \{p'(x, s \mid \sigma)\}_{x,s,\sigma} \in \mathbb{P}_j^\delta$ that is derived from $\hat{P}$ as follows. First, index all pairs $(x, s) \in X \times S_j$ by integers $1, 2, \cdots, jX$. Then, for every $\sigma \in S_j$ repeat the following procedure: For every pair $(x, s) \in X \times S_j$, if $\hat{p}(x, s \mid \sigma) < \delta$, let $p'(x, s \mid \sigma) = \delta$. For every other (x, s), set $p'(x, s \mid \sigma) = \hat{p}(x, s \mid \sigma)$, except for (x^*, s^*), the pair with the smallest index that attains $\max_{(x,s)} \hat{p}(x, s \mid \sigma)$, for which $p'(x^*, s^*) = 1 - \sum_{(x,s) \neq (x^*, s^*)} p'(x, s \mid \sigma)$. It follows from this procedure that $p'(x, s \mid \sigma) \geq \hat{p}(x, s \mid \sigma)$ for all (x, s) except for (x^*, s^*) where

$$\begin{aligned} p'(x^*, s^* \mid \sigma) &\geq \hat{p}(x^*, s^* \mid \sigma) - \delta j X \\ &= \hat{p}(x^*, s^* \mid \sigma)\left(1 - \frac{\delta j X}{\hat{p}(x^*, s^* \mid \sigma)}\right) \\ &= \hat{p}(x^*, s^* \mid \sigma)\left(1 - \delta j^2 X^2\right), \end{aligned} \tag{A.2}$$

where we have used the fact that $\hat{p}(x^*, s^* \mid \sigma) \geq (jX)^{-1}$. It now follows that

$$\begin{aligned} \max_{P \in \mathbb{P}_j^\delta} P(\boldsymbol{x}) &\geq P'(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{s}} \prod_{i=1}^{n} p'(x_i, s_i \mid s_{i-1}) \\ &\geq \sum_{\boldsymbol{s}} \prod_{i=1}^{n} \left(1 - \delta j^2 X^2\right) \hat{p}(x_i, s_i \mid s_{i-1}) \\ &= \left(1 - \delta j^2 X^2\right)^n \hat{P}(\boldsymbol{x}). \end{aligned} \tag{A.3}$$

Next, observe that for every $P \in \mathbb{P}_j^\delta$,

$$\begin{aligned} P(\boldsymbol{x}_i, s_{i+1}^l \mid s_i^l) &= P(\boldsymbol{x}_i, s_{il} \mid s_{(i-1)l}) \\ &= \sum_{s_{(i-1)l+1}^{il-1}} \prod_{t=(i-1)l+1}^{il} p(x_t, s_t \mid s_{t-1}) \\ &= \sum_{s_{(i-1)l+1}} p(x_{(i-1)l+1}, s_{(i-1)l+1} \mid s_{(i-1)l}) \cdot \\ &\qquad \sum_{s_{(i-1)l+2}^{il-2}} \prod_{t=(i-1)l+2}^{il-1} p(x_t, s_t \mid s_{t-1}) \cdot \\ &\qquad \sum_{s_{il-1}} p(x_{il}, s_{il} \mid s_{il-1}) \\ &\geq \delta^2 \sum_{s_{(i-1)l+1}^{il-1}} \prod_{t=(i-1)l+2}^{il-1} p(x_t, s_t \mid s_{t-1}), \end{aligned} \tag{A.4}$$

where s_k^m denotes the segment $(s_k, s_{k+1}, \cdots, s_m)$ for $m \geq k$. Similarly, since $p(x, s \mid \sigma) \leq 1$, we have

$$P(\boldsymbol{x}_i, s_{i+1}^l \mid s_i^l) \leq \sum_{s_{(i-1)l+1}^{il-1}} \prod_{t=(i-1)l+2}^{il-1} p(x_t, s_t \mid s_{t-1}). \tag{A.5}$$

Since the expression

$$\sum_{s_{(i-1)l+1}^{il-1}} \prod_{t=(i-1)l+2}^{il-1} p(x_t, s_t \mid s_{t-1})$$

does not depend on $s_{(i-1)l}$ and s_{il}, it follows that any $P \in \mathbb{P}_j^\delta$ we must have

$$\delta^2 \leq \frac{P(\boldsymbol{x}_i, s_{i+1}^l \mid s_i^l)}{P(\boldsymbol{x}_i, \hat{s}_{i+1}^l \mid \hat{s}_i^l)} \leq \frac{1}{\delta^2}, \tag{A.6}$$

for any $s_i^l, \hat{s}_i^l, s_{i+1}^l \hat{s}_{i+1}^l \in S_j$. Hence,

$$\begin{aligned} P(\boldsymbol{x}) = \sum_{\hat{\boldsymbol{s}}^l} P(\boldsymbol{x}, \hat{\boldsymbol{s}}^l) &= P(\boldsymbol{x}, \boldsymbol{s}^l) \sum_{\hat{\boldsymbol{s}}^l} \frac{P(\boldsymbol{x}, \hat{\boldsymbol{s}}^l)}{P(\boldsymbol{x}, \boldsymbol{s}^l)} \\ &= P(\boldsymbol{x}, \boldsymbol{s}^l) \sum_{\hat{\boldsymbol{s}}^l} \prod_{i=1}^{n/l} \frac{P(\boldsymbol{x}_i, \hat{s}_{i+1}^l \mid \hat{s}_i^l)}{P(\boldsymbol{x}_i, s_{i+1}^l \mid s_i^l)} \\ &\leq P(\boldsymbol{x}, \boldsymbol{s}^l) \sum_{\hat{\boldsymbol{s}}^l} \left(\frac{1}{\delta^2}\right)^{n/l} \\ &= P(\boldsymbol{x}, \boldsymbol{s}^l) j^{n/l} \left(\frac{1}{\delta^2}\right)^{n/l} \\ &= P(\boldsymbol{x}, \boldsymbol{s}^l) \left(\frac{j}{\delta^2}\right)^{n/l}, \end{aligned} \tag{A.7}$$

where $s_0^l = \hat{s}_0^l = s_0$. Since (A.7) holds for any $P \in \mathbb{P}_j^\delta$, it follows by (A.1) and (A.7) that

$$\begin{aligned} \max_{P \in \mathbb{P}_j} P(\boldsymbol{x}) &= \hat{P}(\boldsymbol{x}) \\ &\leq \left(1 - \delta j^2 X^2\right)^{-n} P'(\boldsymbol{x}) \\ &\leq \left(1 - \delta j^2 X^2\right)^{-n} \left(\frac{j}{\delta^2}\right)^{n/l} P'(\boldsymbol{x}, \boldsymbol{s}^l) \\ &\leq \left(1 - \delta j^2 X^2\right)^{-n} \left(\frac{j}{\delta^2}\right)^{n/l} \max_{P \in \mathbb{P}_j} P(\boldsymbol{x}, \boldsymbol{s}^l). \end{aligned} \tag{A.8}$$

Finally, by minimizing the factor $(1 - \delta j^2 X^2)^{-n} (j \delta^{-2})^{n/l}$ on the right-most side of (A.8) with respect to δ in the range $0 < \delta \leq (jX)^{-1}$, and using the assumption that $j \leq M$, the proof of Lemma 2 is complete.

References

[1] 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.

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

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.,* vol. 39, no. 1, pp. 1-38, 1977.

[4] R. G. Gallager, *Information Theory and Reliable Communications.* New York: J. Wiley, 1968.

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

[6] E. Plotnik, M. J. Wienberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and the redundancy of the Lempel-Ziv algorithm," submitted to *IEEE Trans. Inform. Theory.*

[7] J. Ziv, "Compression, tests for randomness, and estimating the statistical model of an individual sequence," in *Sequences,* R. M. Capocelli, Ed. New York: Springer-Verlag, 1990, pp. 366-373.

[8] N. Merhav, "Universal coding with minimum probability of codeword length overflow," *IEEE Trans. Inform. Theory,* vol. 37, pt. I, pp. 556-563, May 1991.

[9] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feed-back schemes," *IEEE Trans. Inform. Theory,* vol. IT-14, pp. 206-220, Mar. 1968.

[10] W. B. Davenport, Jr. and W. L. Root, *Random Signals and Noise.* New York: McGraw-Hill, 1958, pp. 322-324.

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

[12] M. J. Weinberger, A. Lempel, and J. Ziv, "A sequential algorithm for the universal coding of finite memory sources," submitted to *IEEE Trans. Inform. Theory.*

[13] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" submitted to *IEEE Trans. Inform. Theory.*