A Measure of Relative Entropy Between Individual Sequences with Application to Universal Classification

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

***Abstract*—A new notion of empirical informational divergence (relative entropy) between two individual sequences is introduced. If the two sequences are independent realizations of two finite-order, finite alphabet, stationary Markov processes, the empirical relative entropy converges to the relative entropy almost surely. This new empirical divergence is based on a version of the Lempel–Ziv data compression algorithm. A simple universal classification algorithm for individual sequences into a finite number of classes which is based on the empirical divergence, is introduced. It discriminates between the classes whenever they are distinguishable by some finite-memory classifier, for almost every given training sets and almost any test sequence from these classes. It is universal in the sense of being independent of the unknown sources.**

***Index Terms*—Lempel–Ziv algorithm, information divergence, finite-memory machines, finite-state machines, universal classification, universal hypothesis testing.**

I. Introduction

SUPPOSE one observes a sequence $\boldsymbol{x} = (x_1 \cdots x_n)$ emitted from an unknown lth-order stationary Markov process $p(\cdot)$ over a finite-alphabet A with $|\mathsf{A}|=A$ letters, and wishes to estimate the nth-order entropy, or equivalently $-n^{-1}\log p(\cdot)$. While the straightforward approach of calculating the lth-order conditional empirical distribution is computationally prohibitive for large l and is impossible if l is unknown, it has been shown in [1], [2] that the Lempel–Ziv (LZ) codeword length for $\boldsymbol{x}$ divided by the length n, is a computationally efficient, reliable estimate of the entropy, and hence also of $n^{-1}\log p(\cdot)$.

More precisely, let

$$p(x_1, x_2 \cdots x_n) = \prod_{i=1}^{n} p(x_i|s_{i-1}); x_i \in \mathsf{A}, \tag{1}$$

where $s_i = x_{i-\ell+1}^{i} = (x_{i-\ell+1}, x_{i-\ell+2} \cdots x_i)$ for $i \geq \ell$ and $s_i = (s_0, x_1, x_2 \cdots x_i)$ for $i < \ell$, s_0 being the initial state. Here s_i takes on values in the set A^ℓ of all length ℓ vectors with components in A.

Manuscript received January 10, 1992; revised December 20, 1992. This work was suported in part by the US–Israel Binational Science Foundation. This work was presented in part at the IEEE International Symposium on Information Theory, San Antonio, TX, January 17–22, 1993.

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

J. Ziv was on a Sabbatical leave at AT&T Bell Laboratories, Murray Hill, NJ 07974 while performing this research.

IEEE Log Number 9209850.

Let $c(\boldsymbol{x})$ denote the number of phrases in $\boldsymbol{x}$ resulting from the incremental parsing of $\boldsymbol{x}$ [1], i.e. sequential parsing of $\boldsymbol{x}$ into distinct phrases such that each phrase is the shortest string which is not a previously parsed phrase. Then, the LZ codeword length for $\boldsymbol{x}$ can be approximated by $c(\boldsymbol{x})\log c(\boldsymbol{x})$ and

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log p(\boldsymbol{x}) - \frac{1}{n}c(\boldsymbol{x})\log c(\boldsymbol{x})\right] = 0, \text{almost surely.}$$

In fact, this property continues to hold if $p(\cdot)$ is allowed to be any finite alphabet, stationary ergodic process, not necessarily a Markov process.

In this paper, we generalize this result to the case where there are two Markov probability measures, $p(\cdot)$ and $q(\cdot)$, each of order no larger than some positive integer l, in their steady -state modes. Let $\boldsymbol{x}$ be a realization of $p(\cdot)$ and let $\boldsymbol{z}$ be a realization of $q(\cdot)$. Given $\boldsymbol{x}$ and $\boldsymbol{z}$ we would like to get good estimators for $-n^{-1}\log p(\boldsymbol{z})$ and $-n^{-1}\log q(\boldsymbol{x})$. In particular, we seek an easily calculable function of both $\boldsymbol{x}$ and $\boldsymbol{z}$, that does not depend on ℓ, which will enable us to discriminate between two different unknown sources $p(\cdot)$ and $q(\cdot)$, based on their realizations $\boldsymbol{x}$ and $\boldsymbol{z}$.

The divergence $D(q\|p)$ is defined by:

$$D(q\|p) = \limsup_{n\to\infty}\frac{1}{n}\sum_{\mathsf{A}^n} q(\boldsymbol{x})\log\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}. \tag{2}$$

where logarithms are defined to be of base 2 unless specified otherwise. The divergence (or "relative entropy") is a measure of the statistical distance between two distributions [2].

Throughout this paper, we assume that $q(\boldsymbol{x}) > 0$ implies $p(\boldsymbol{x}) > 0$, i.e., an absolute continuity condition.

Now, since both $p(\cdot)$ and $q(\cdot)$ are assumed to be ℓth-order Markovian probability measures, we have by (1) that,

$$\sum_{\boldsymbol{x}\in\mathsf{A}^n} q(\boldsymbol{x})\log\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{a_1^\ell\in\mathsf{A}^\ell} q(a_1^\ell)\log\frac{q(a_1^\ell)}{p(a_1^\ell)} + (n-\ell)\sum_{a\in\mathsf{A};s\in\mathsf{A}^\ell} q(a,s)\log\frac{q(a|s)}{p(a|s)},$$

since the state s is the previous l letters. Thus,

$$D(q\|p) = D^\ell(q\|p) = \sum_{a\in\mathsf{A};s\in\mathsf{A}^\ell} q(a,s)\log\frac{q(a|s)}{p(a|s)}. \tag{3}$$

In particular, let $q_{\boldsymbol{z}}(a_1 \cdots a_{\ell+1}) = q_{\boldsymbol{z}}(a^{\ell+1})$ be the empirical measure of vectors of length $(\ell+1)$; $a^{\ell+1} \in \mathsf{A}^{\ell+1}$, where $q_{\boldsymbol{z}}(a^{\ell+1})$ is defined as

$$q_{\boldsymbol{z}}(a^{\ell+1}) = \frac{1}{n}\sum_{i=1}^{n} I\{z_i^{i+\ell} = a^{\ell+1}\} = q_{\boldsymbol{z}}(a,s), \qquad (4)$$

where $z_i^{i+\ell}$ denotes the string $(z_i, z_{i+1} \cdots z_{i+\ell})$ with the cyclic convention that $z_{-i} = z_{n-i}$, and $I\{D\}$ is the indicator function for the event D. Thus, by (1), (2), and (3)

$$-\log p(\boldsymbol{z}) = nH_{\boldsymbol{z}} + nD(q_{\boldsymbol{z}}\|p), \qquad (5)$$

where

$$\begin{aligned} H_{\boldsymbol{z}} &= -\sum_{\mathsf{A},\mathsf{A}^{\ell}} q_{\boldsymbol{z}}(a,s)\log\frac{q_{\boldsymbol{z}}(a,s)}{q_{\boldsymbol{z}}(s)} \\ &= -\sum_{\mathsf{A},\mathsf{A}^{\ell}} q_{\boldsymbol{z}}(a,s)\log q_{\boldsymbol{z}}(a|s); \quad a \in \mathsf{A}, \quad s \in \mathsf{A}^{\ell}. \end{aligned} \qquad (6)$$

In many cases of interest $p(\cdot)$ and $q(\cdot)$ or even their Markov orders are not known. However, we are given two n-sequences $\boldsymbol{x}$ and $\boldsymbol{z}$ and our objective is to decide whether or not they have emerged from the same source.

Analogously to the single source case, where $-\frac{1}{n}\log q(\boldsymbol{z})$ (and hence the entropy $H_{\boldsymbol{z}}$) is efficiently estimated by $\frac{1}{n}c(\boldsymbol{z})\log c(\boldsymbol{z})$, we introduce an empirical quantity $Q(\boldsymbol{z}\|\boldsymbol{x})$ which will be shown to have the property that,

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log p(\boldsymbol{z}) - Q(\boldsymbol{z}\|\boldsymbol{x})\right] = 0, \qquad (7a)$$

for almost every pair $(\boldsymbol{x},\boldsymbol{z})$ relative to the product probability measure $p(\boldsymbol{x})q(\boldsymbol{z})$, for every finite ℓ and every finite A. Following (5), the function $Q(\boldsymbol{z}\|\boldsymbol{x})$ can be decomposed into two terms, the first is an estimate of the empirical entropy associated with $\boldsymbol{z}$, i.e., $\frac{1}{n}c(\boldsymbol{z})\log c(\boldsymbol{z})$, and the second term, denoted by $\Delta(\boldsymbol{z}\|\boldsymbol{x})$ is an estimate of the divergence between $q_{\boldsymbol{z}}(\cdot)$ and $p(\cdot)$ with the property that,

$$\lim_{n\to\infty}[\Delta(\boldsymbol{z}\|\boldsymbol{x}) - D(q_{\boldsymbol{z}}\|p)] = 0, \qquad (7b)$$

almost surely, with respect to the product measure $p(\cdot)q(\cdot)$, for every finite ℓ and every finite A. Analogously to the fact that the entropy is estimated by *self* LZ incremental parsing of $\boldsymbol{z}$, here intuition suggests that $\Delta(\boldsymbol{z}\|\boldsymbol{x})$, which is an estimate of the *cross entropy* $D(q_{\boldsymbol{z}}\|p)$, will be associated with *cross parsing* of $\boldsymbol{z}$ with respect to $\boldsymbol{x}$.

These two related empirical quantities (which are functions of $\boldsymbol{x}$ and $\boldsymbol{z}$ but do not depend on ℓ or A) are based on a version of the Lempel–Ziv (LZ) data compression algorithm [1], [3]. These results are fully formulated, stated, and proved in Section II.

Intuitively, $\Delta(\boldsymbol{z}\|\boldsymbol{x})$ may serve as a reasonable discrimination function for universal classification of individual sequences. Indeed, we show in Section III that if there exists a finite-state (FS) classifier, in particular, a finite-memory classifier (with a rejection option), that is trained by a given set of training sequences from each class, which is both consistent (i.e., classifies correctly input sequences that are identical to one of the training sequences), and free from classification errors w.r.t two fixed sets of "source" sequences, then a classifier based on the comparison of $\Delta(\cdot\|\cdot)$ to a threshold, performs asymptotically as well for almost every data set.

By confining our interest to FS classifiers we limit the extraction process of the statistics from the data to be one which is realized by a finite-state machine (FSM). The classification rule is then assumed to be a function of the "type" that characterizes such a finite-state statistic. The number of resolution atoms is therefore upper bounded by the number of such possible "types"(which is, at most, polynomial in n). Thus, the number of training sequences may be kept relatively small. Observe that any family of classifiers that requires a relatively small number of equivalence classes and hence a small number of training sequences, can be a reasonable choice for a class of practical classification schemes. Thus, the family of FS classifiers is a good choice in that respect as well. These points will be elaborated on in Section III.

II. Formulation and Derivation of Main Results

The incremental LZ parsing algorithm [1] is a self parsing procedure of a sequence into $c(\boldsymbol{z})$ distinct phrases such that each phrase is the shortest string which is not a previously parsed phrase. For example, let $n = 11$ and $\boldsymbol{z} = (01111000110)$,then self incremental parsing yields $(0,1,11,10,00,110))$, namely , $c(\boldsymbol{z}) = 6$.

We now describe a variant of the LZ parsing algorithm [3] which is a sequential parsing of a vector $\boldsymbol{z}$ with respect to another vector $\boldsymbol{x}$. First, find the longest prefix of $\boldsymbol{z}$ that appears as a string in $\boldsymbol{x}$, i.e., the largest integer m such that $(z_1, z_2 \cdots z_m) = (x_i, x_{i+1} \cdots x_{i+m-1})$ for some i. The string $(z_1, z_2 \cdots z_m)$ is defined as the first phrase of $\boldsymbol{z}$ with respect to $\boldsymbol{x}$. If m=0 (i.e. z_1 does not appear in $\boldsymbol{x}$), the first phrase of $\boldsymbol{z}$ with respect to $\boldsymbol{x}$ is z_1. Thus, the case $m = 0$ is treated as though $m = 1$. Next, start from z_{m+1} and find, in a similar manner, the longest prefix of $z_{m+1}, z_{m+2} \cdots z_n$, which appears in $\boldsymbol{x}$, and so on. The procedure is terminated once the entire vector $\boldsymbol{z}$ has been parsed with respect to $\boldsymbol{x}$. Let $c(\boldsymbol{z}|\boldsymbol{x})$ denote the number of phrases in $\boldsymbol{z}$ with respect to $\boldsymbol{x}$. For example, let $\boldsymbol{z}$ be as before and $\boldsymbol{x} = (10010100110)$. Then, parsing $\boldsymbol{z}$ with respect to $\boldsymbol{x}$ yields: $(011, 110, 00110)$, that is $c(\boldsymbol{z}|\boldsymbol{x}) = 2$.

For two sequences $\boldsymbol{z}$ and $\boldsymbol{x}$ of length n, define the functions:

$$Q(\boldsymbol{z}\|\boldsymbol{x}) = \frac{1}{n}c(\boldsymbol{z}|\boldsymbol{x})\log n, \qquad (8)$$

$$\Delta(\boldsymbol{z}\|\boldsymbol{x}) = \frac{1}{n}[c(\boldsymbol{z}|\boldsymbol{x})\log n - c(\boldsymbol{z})\log c(\boldsymbol{z})]. \qquad (9)$$

The following two theorems describe the statistical behavior of these two quantities.

Theorem 1: Let $p(\cdot)$ be a stationary Markov source of order ℓ and let μ be an arbitrarily small number. Let $\boldsymbol{x}$ be a realization of $p(\cdot)$.

Then we have the following.

a) For every fixed $\boldsymbol{z} \in \mathsf{A}^n$ for which $p(\boldsymbol{z}) > 0$,

$$\begin{aligned} &p\left[\boldsymbol{x} : -\frac{1}{n}\log p(\boldsymbol{z}) - Q(\boldsymbol{z}\|\boldsymbol{x}) < -2\mu\log\frac{1}{\delta}\right] \\ &\le e^{-(n^{\mu}/\log n)K(\delta,\ell)+o(n^{\mu}/\log n)}, \end{aligned} \qquad (10)$$

where

1) $\delta = \min_{a^{\ell+1} \in A^{\ell+1}: p(a^{\ell+1})>0} p(a^{\ell+1})$

2) $K(\delta, \ell) = \delta^{\ell+1} \log \frac{1}{1-\delta}$,

b) For every $z \in A^n$ for which $p(z) > 0$,

$$p\left[x : -\frac{1}{n}\log p(z) - Q(z\|x) > 2\mu \log \frac{1}{\delta}\right] \leq n^{-\mu/2} + o(n^{-\mu/2}). \quad (11.a)$$

c) Let z be a realization of a stationary Markov source of order ℓ' with an underlying probability measure $q(\cdot)$. Then (11.a) can be replaced by a tighter bound for *almost* every $z \in A^n$ (relative to the probability measure $q(\cdot)$) as follows:

$$p\left[x : -\frac{1}{n}\log p(z) - Q(z\|x) > 2\mu \log \frac{1}{\delta}\right] \leq e^{-\frac{1}{2}\left(\frac{\mu^2}{\log^2 n}\right) n^{K'(\delta,\delta')}[1-o(n^{-\mu^2})]}, \quad (11.b)$$

for every $z \in A^n - B$ for which $p(z) > 0$, where B is a subset with the property:

$$q(z \in B) \leq e^{-K''(\delta,A,\mu)\frac{n^{\mu'}}{\log n} + o\left(\frac{n^{\mu'}}{\log n}\right)}, \quad (11.c)$$

and

1) $\delta' = \min_{a^{\ell+1} \in A^{\ell+1}: q(a^{\ell+1})>0} q(a^{\ell+1})$,

2) $K' = \frac{1}{4}\frac{\log\left(\frac{1}{1-\delta'}\right)}{\log \frac{1}{\delta}}$,

3) $K''(\delta, \delta', \mu) = \frac{\mu}{2(1+\mu)}\frac{\log \frac{1}{1-\delta}}{\log \frac{1}{\delta}} \log \frac{1}{1-\delta'}$,

4) $\mu' = 1 - \frac{1}{4}\log\left(\frac{1}{1-\delta'}\right) / \log \frac{1}{\delta}$.

Theorem 1 tells us that $Q(z\|x)$ exceeds $-\frac{1}{n}\log p(z)$ significantly with probability that decays "almost exponentially," but the probability of $Q(z\|x)$ being significantly smaller than $-\frac{1}{n}\log p(z)$ is upper bounded by a term that decays only *polynomially* with n. If, however, z belongs to some subset of high probability under $q(\cdot)$, the latter probability decays "almost exponentially" as well.

Corollary: For every $z \in Z^n$ for which $p(z) > 0$ and any arbitrarily small positive ϵ

$$\lim_{n\to\infty} p\left\{x : \left|-\frac{1}{n}\log p(z) - Q(z\|x)\right| > \epsilon\right\} = 0. \quad (12)$$

This follows directly from (9) and (11.a). Furthermore, it follows from (9), (11.b), (11.c), and the Borel–Cantelli Lemma that

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log p(z) - Q(z\|x)\right] = 0 \quad (13)$$

a.s. for every sequence of pairs (x, z), with respect to a probability measure for pairs given by

$$f(x, z) \triangleq p(x)q(z).$$

Theorem 2 follows from Theorem 1.

Theorem 2:

a) Let x be a realization of a stationary ℓth-order Markov source $p(\cdot)$.Then, for every $z \in A^n$ such that $p(z) > 0$

$$p\left(x : D(q_z\|p) - \Delta(z\|x) < -3\mu \log \frac{1}{\delta}\right) \leq e^{-(n^\mu/\log n)K(\delta,\ell) - o(n^\mu/\log n)} \quad (14)$$

where δ and $K(\delta, \ell)$ is defined in (9).

b) Let x be a realization of a stationary ℓth-order Markov source $p(\cdot)$ and let z be a realization of a stationary ℓ'th-order Markov source $q_z(\cdot)$.Then

$$p\left(x : D(q_z\|p) - \Delta(z\|x) > 3\mu \log \frac{1}{\delta}\right) \leq e^{-\frac{1}{2}\left(\frac{\mu^2}{\log^2 n}\right) n^{K'(\delta,\delta')}[1-o(n^{-\mu^2})]} \quad (15)$$

for every $z \in A^n - B$ such that $p(z) > 0$, where B is a subset with the property:

$$q(z \in B) \leq e^{-K''(\delta,A,\mu)\frac{n^{\mu'}}{\log n} + o\left(\frac{n^{\mu'}}{\log n}\right)} \quad (16)$$

and $K'(\delta, \delta'), \mu'$ and $K''(\delta, A, \mu)$ are given by (11.c).

The proofs of Theorem 1 and Theorem 2 that follow, are rather lengthy. It is, therefore, suggested that on first reading the reader should skip the proofs and move to Section III (application).

Proof of Theorem 1: a) Let

$$\min_{a\in A; s\in A^\ell: p(a,s)>0} p(a, s) \triangleq \delta \quad (17.a)$$

Then, ignoring zero probability events,

$$p(a|s) \geq \delta; p(s) \geq \delta. \quad (17.b)$$

Consider an auxiliary parsing of z into $\bar{c} = \bar{c}(n)$ phrases:

$$z = z^{L_1}, z^{L_2} \cdots z^{L_{\bar{c}}}, \quad (18)$$

where

$$z^{L_i} = z_{L_1+L_2+\cdots+L_{i-1}+1} \cdots z_{L_1+L_2+\cdots+L_i}; \quad z \in A,$$

and where $L_i, 1 \leq i \leq \bar{c}$, satisfies, for some arbitrarily small positive number μ,

$$\delta n^{-(1-\mu)} \leq p(z^{L_i}) < n^{-(1-\mu)}. \quad (19)$$

By (1) and (17), (19) always has a solution L_i and hence this parsing is well defined. Fix a positive integer L and let

$$\bar{\delta}(z^L) = \begin{cases} 1, & \text{if } z^L \neq x_i^{i+L-1} \text{for all } i; 1 \leq i \leq n-L, \\ 0, & \text{otherwise,} \end{cases} \quad (20)$$

where $x_i^j = x_i, x_{i+1} \cdots x_j$. Then, by (1),

$$E\bar{\delta}(z^L) \le \left[1 - \min_{s \in A^\ell} p(z^L|s)\right]^{\left(\frac{n}{L}-1\right)}, \tag{21}$$

where $E(\cdot)$ denotes expectation and s denotes the state that precedes z^L. (The factor $(\frac{n}{L}-1)$ should be replaced by $(\frac{n}{L})$ if L divides n). Now, let $y_1^L \in A^L$; then, by (1)

$$p(y_1^L|s) = p(y_1^\ell|s)p(y_{\ell+1}^L|y_1^\ell)$$

Therefore, by (1) and (17) for $L > l$,

$$\delta^\ell p(y_{\ell+1}^L|y_1^\ell) \le p(y_1^L|s) \le p(y_{\ell+1}^L|y_1^\ell)$$

Also, by (17)

$$p(y_1^L) = \sum_{s \in A^\ell} p(s, y_1^\ell)p(y_{\ell+1}^L|y_1^\ell) \ge \delta^\ell p(y_{\ell+1}^L|y_1^\ell)$$

and, since $p(s, y_1^\ell) \le p(s)$ we also have that

$$p(y_1^L) \le p(y_{\ell+1}^L|y_1^\ell)$$

Thus,

$$\delta^\ell p(y_1^L) \le p(y_1^L|s) \le \frac{1}{\delta^\ell} p(y_1^L) \tag{22}$$

Therefore, by (19), (21), and (22):

$$\begin{aligned} E\bar{\delta}(z^L) &\le \left[1 - \delta^\ell p(z^L)\right]^{\left(\frac{n}{L}-1\right)} \\ &\le \left[1 - \delta^{\ell+1} n^{-(1-\mu)}\right]^{\left(\frac{n}{L}-1\right)} \\ &\le e^{-n^{-(1-\mu)}\delta^{\ell+1}\left(\frac{n}{L}-1\right)}. \end{aligned} \tag{23}$$

Hence, by (23), the union bound and the fact that $\bar{c} \le n$,

$$p\left[\sum_{i=1}^{\bar{c}-1} \bar{\delta}(z^{L_i}) \ge 1\right] \le ne^{-n^{-(1-\mu)}\left(\frac{n}{L}-1\right)\delta^{\ell+1}}.$$

Also, by (17) and (19), the longest possible phrase must satisfy

$$(1-\delta)^{L_i} \ge n^{-(1-\mu)}\delta$$

or, equivalently

$$L_i \le \overline{L} \triangleq \frac{\log n + \log\frac{1}{\delta}}{\log\left(\frac{1}{1-\delta}\right)}. \tag{24}$$

Thus,

$$\begin{aligned} p\left[\sum_{i=1}^{\bar{c}-1} \bar{\delta}(z^{L_i}) \ge 1\right] &\le ne^{-n^{-(1-\mu)}\left(\frac{n}{L}-1\right)\delta^{\ell+1}} \\ &= e^{-\frac{n^\mu}{\log n}K(\delta,\ell)+o\left(\frac{n^\mu}{\log n}\right)}, \end{aligned} \tag{25}$$

where $K(\delta, \ell) = \delta^{\ell+1}\log\frac{1}{1-\delta}$,and where L is replaced by its upper bound $\overline{L}$ which is given by (24). Now, by (19) and (22)

$$\begin{aligned} -\log p(z) &\ge -\sum_{i=1}^{\bar{c}-1} \log p(z^{L_i}) - \bar{c}\ell\log\frac{1}{\delta} - \log p(z^{L_{\bar{c}}}) \\ &\ge (1-\mu)(\bar{c}-1)\log n - \bar{c}\ell\log\frac{1}{\delta}. \end{aligned} \tag{26}$$

But by (24), $\bar{c} \ge \frac{n}{\log n}\log\frac{1}{1-\delta} - o\left(\frac{n}{\log n}\right)$. Thus, by (24) and (26)

$$-\log p(z) \ge (1-\mu)\bar{c}\log n - o(\bar{c}\log n). \tag{27}$$

We are now going to relate $\bar{c}$ to $c(z|x)$.

By definition of $\bar{\delta}(z^\ell)$ (16), if $\bar{\delta}(z^L) = 0$ then z^L can not contain more than one comma of the LZ parsing of z with respect to x. Hence, by (25)

$$p[x : c(z|x) > \bar{c} - 1 + \overline{L}] \le e^{-\frac{n^\mu}{\log n}K(\delta,\ell)+o\left(\frac{n^\mu}{\log n}\right)}, \tag{28}$$

where the $\overline{L}$ term is an upper bound on the length of a phrase and therefore on the number of LZ commas that are contained in the last of the $\bar{c}$ phrases. Note that this last phrase does not necessarily satisfy $\bar{\delta}(z^{L_{\bar{c}}}) = 0$.

Combining (24), (26)–(28) yields (9), and hence, completes the proof of part a). □

b) Consider now the parsing of z into $\hat{a} = c(n)$ phrases

$$z = z^{L_1}, z^{L_2} \cdots z^{L_{\hat{a}}},$$

where, in contrast with (19) ,L_i now satisfies

$$\delta n^{-(1+\mu)} < p(z^{L_i}) \le n^{-(1+\mu)}. \tag{29}$$

Let

$$\hat{\delta}(z^L) \triangleq \begin{cases} 1, & \text{if } z^L = x_{i+1}^{i+L}, \text{for some } 0 \le i \le n-L, \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, by (29) and by the union bound, for every phrase except, perhaps, for the last one,

$$E\hat{\delta}(z^L) < np(z^L) \le n^{-\mu}.$$

Thus, by Chebychev's inequality:

$$p\left(\sum_{i=1}^{\hat{a}-1} \hat{\delta}(z^{L_i}) > (\hat{c}-1)n^{-\frac{\mu}{2}}\right) \le n^{-\mu/2}. \tag{30}$$

Now, by construction (29), within every one of the phrases for which $\delta(z^L) = 0$, at least one LZ phrase (associated with the LZ parsing of z with respect to x), is initiated. Hence, by (30),

$$p\left(c(z|x) < (\hat{c}-1) - (\hat{c}-1)n^{-\frac{\mu}{2}}\right) \le n^{-\mu/2} \tag{31}$$

Also, by (22)

$$-\log p(z) \le -\sum_{i=1}^{\hat{c}-1} \log p(z^{L_i}) + \hat{c}\ell\log\frac{1}{\delta} + \overline{L}\log\frac{1}{\delta}, \tag{32}$$

where the last term is due to the last phrase of the parsing according to (29), and where, by (29)

$$\frac{\log n}{\log\frac{1}{\delta}} \triangleq \underline{L} < L_i \le \overline{L} \triangleq \frac{(1+\mu)\log n + \log\frac{1}{\delta}}{\log\left(\frac{1}{1-\delta}\right)}. \tag{33}$$

But, by (33),

$$\hat{c} \le \frac{n}{\underline{L}} = \frac{n}{\log n}\log\frac{1}{\delta}. \tag{34}$$

Thus, by (32) and (34),

$$-\log p(z) \le (1+\mu)(\hat{c}-1)\log n + o((\hat{c}-1)\log n). \tag{35}$$

Combining (31) and (35) yields (11.a), completing the proof of part b). □

c) Consider again the parsing of $\boldsymbol{z}$ into $\hat{c}$ phrases according to (29). We assume, for the sake of simplicity, that D divides $\hat{c}$ and subdivide $\boldsymbol{z}$ into $\frac{\hat{c}}{D}$ blocks of D phrases each. For each such block of D phrases, the probability of the event F that all the phrases within the block are distinct given the preceding state s, is lower-bounded by

$$q(F|s) \geq [1 - D(1-\delta')^{\underline{L}}]^D, \tag{36}$$

where s is the initial state, where $\underline{L}$ is given by (29) and where

$$\delta' \triangleq \min_{q(a,s)>0} q(a,s) \quad .$$

Here, we took advantage of the fact that $[1 - D(1-\delta')^{\underline{L}}]$ is a lower bound on the probability that a phrase is not identical to any of the (D) preceding phrases regardless of the preceding state, thus enabling us to treat the phrases as being independent, in so far as this lower bound is concerned. Note that the lower bound on $q(F|s)$ given by (36) does not therefore depend on the state s.

Let

$$D = \left(\frac{1}{1-\delta'}\right)^{\frac{1}{4}\underline{L}} = \left(\frac{1}{1-\delta'}\right)^{\frac{1}{4}\frac{\log n}{\log\frac{1}{\delta}}} = n^{\frac{1}{4}\frac{\log\left(\frac{1}{1-\delta'}\right)}{\log\frac{1}{\delta}}}. \tag{37}$$

Then, by (36) and (37)

$$q(F|s) \geq \left[1 - \frac{1}{D^3}\right]^D \geq 1 - \frac{1}{D^2}.$$

Thus,

$$1 - q(F|s) \leq \frac{1}{D^2} = n^{-\frac{1}{2}\frac{\log\frac{1}{(1-\delta')}}{\log 1/\delta}}.$$

Now, let us denote the jth block of D phrases of $\boldsymbol{z}$ by $\boldsymbol{z}_j(D)$ and let

$$\delta(\boldsymbol{z}_j(D)) = \begin{cases} 1, & \text{if not all of the } D \text{ phrases in} \\ & \boldsymbol{z}_j(D) \text{ are distinct,} \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

Then, applying the Chernoff bound, and taking advantage of the fact that the moment generating function $E[e^{b\delta(\boldsymbol{z}_j(D))}|s]$ is upper bounded by $(e^b\frac{1}{D^2}+1)$ for all s and any positive b,

$$q\left[\sum_{j=1}^{\frac{\hat{c}}{D}}\delta(\boldsymbol{z}_j(D)) > \mu\frac{\hat{c}}{D}\right] \leq \left[e^b n^{-\frac{1}{2}\frac{\log\frac{1}{(1-\delta')}}{\log 1/\delta}} + 1\right]^{\frac{\hat{c}}{D}} e^{-b\mu\hat{c}/D}. \tag{39}$$

Let

$$e^b = n^{\frac{1}{2}\log\frac{1}{(1-\delta')}/\log\frac{1}{\delta}}.$$

Then,

$$q\left[\sum_{j=1}^{\hat{c}/D}\delta(\boldsymbol{z}_j(D)) > \mu\frac{\hat{c}}{D}\right] \leq e^{-\frac{1}{2}\left[\left[\log\left(\frac{1}{1-\delta'}\right)/\log\frac{1}{\delta}\right]\mu\log n - 2\ln 2\right]\frac{\hat{c}}{D}}. \tag{40}$$

Now, by (33) and (37),

$$\frac{\hat{c}}{D} > \frac{n}{\underline{L}D} = \frac{n\log\left(\frac{1}{1-\delta}\right)}{(1+\mu)\log n + \log\frac{1}{\delta}}\frac{1}{n^{\frac{1}{4}\left[\log\left(\frac{1}{1-\delta'}\right)/\log\frac{1}{\delta}\right]}}. \tag{41}$$

Thus,

$$q\left[\sum_{j=1}^{\frac{\hat{c}}{D}}\delta(\boldsymbol{z}_j(D)) > \mu\frac{\hat{c}}{D}\right] \leq e^{-K''(\delta,\delta',\mu)\frac{n^{\mu'}}{\log n}+o\left(\frac{n^{\mu'}}{\log n}\right)}, \tag{42}$$

where

$$K''(\delta,\delta',\mu) = \frac{\mu}{2(1+\mu)}\frac{\log\left(\frac{1}{1-\delta}\right)}{\log 1/\delta}\log\left(\frac{1}{1-\delta'}\right)$$

$$\mu' = 1 - \frac{1}{4}\log\left(\frac{1}{1-\delta'}\right)/\log\frac{1}{\delta}.$$

Consider now all blocks of D phrases for which $\delta(\boldsymbol{z}_j(D)) = 0$. Within each such block, consider the D_L phrases for which $L_j = L$ and denote these phrases by $\boldsymbol{z}_1^L, \boldsymbol{z}_2^L \cdots \boldsymbol{z}_{D_L}^L$. Consider the parsing of $\boldsymbol{x}$ into L-vectors

$$\boldsymbol{x} = x_1^L, x_{L+1}^{2L} \cdots x_{iL+1}^n; i = \frac{n}{L} - 1.$$

(For the sake of simplicity we have assumed that L divides n).

Now, let

$$\delta_1(\boldsymbol{z}^L) = \begin{cases} 1, & \text{if } \boldsymbol{z}^L = x_{jL+1}^{(j+1)L}, \quad \text{for some } 0 \leq j \leq \frac{n}{L} - 1, \\ 0, & \text{otherwise.} \end{cases} \tag{43}$$

Then, by the Chernoff bound,

$$p\left(\sum_{j=1}^{D_L}\delta_1(\boldsymbol{z}_j^L) > \frac{D\mu}{\log^3 n}\right) \leq \min_{b>0} e^{-\frac{bD\mu}{\log^3 n}} E\left(e^{b\sum_{j=1}^{D_L}\delta_1(\boldsymbol{z}_j^L)}\right). \tag{44}$$

We next overestimate the expectation on the right-hand side of (41), taking advantage of the fact that $\boldsymbol{z}_1^L, \boldsymbol{z}_2^L \cdots \boldsymbol{z}_{D_L}^L$ are all distinct.

By (29), the probability of the event that a single L-vector $\boldsymbol{z}_j^L$ is identical to one of the L-vectors in the parsed $\boldsymbol{x}$, is upper-bounded by

$$\frac{n}{L}\cdot n^{-(1+\mu)} \leq n^{-\mu}. \tag{45}$$

The probability of the event that K L-vectors $\boldsymbol{z}_{i_1}^L, \boldsymbol{z}_{i_2}^L \cdots \boldsymbol{z}_{i_K}^L$ are identical to K *distinct* L vectors among the $\frac{n}{L}$ L vectors in the parsed $\boldsymbol{x}$ is upper bounded by

$$\binom{D_L}{K} n^K \left[\max_{s\in\mathbf{A}^L} p(\boldsymbol{z}^L|s)\right]^K \tag{46}$$

But, by (17) and (29),

$$\max_s p(\boldsymbol{z}^L|s) \leq \frac{p(\boldsymbol{z}^L)}{\delta} \leq \frac{n^{-(1+\mu)}}{\delta} \tag{47}$$

Thus, by (46) and (47)

$$\begin{aligned} Ee^{b\sum_{i=1}^{D_L}\delta(\boldsymbol{z}_i^L)} &\le \sum_{K=0}^{D_L} n^{-K\mu}\left(\frac{1}{\delta}\right)^K e^{bK}\binom{D_L}{K} \\ &\le \sum_{K=0}^{D} n^{-K\mu}\left(\frac{1}{\delta}\right)^K e^{bK}\binom{D}{K} \\ &\le \left(1-e^b n^{-\mu}\frac{1}{\delta}\right)^{-D}\sum_{K=0}^{D}\binom{D}{K}\left[e^b\left(\frac{n^{-\mu}}{\delta}\right)\right]^K, \\ &\quad\cdot\left[1-e^b\frac{n^{-\mu}}{\delta}\right]^{D-K} \\ &= \left(1-e^b n^{-\mu}\frac{1}{\delta}\right)^{-D} \end{aligned}$$

provided that

$$e^b n^{-\mu}\frac{1}{\delta} < 1. \tag{48}$$

Now, set

$$b = \frac{1}{2}\mu\log n. \tag{49}$$

Then, $e^b n^{-\mu}\frac{1}{\delta} = n^{-\frac{1}{2}\mu}\frac{1}{\delta}$ and it tends to zero as $n\to\infty$. Thus, for large enough n,

$$Ee^{b\sum_{i=1}^{D_L}\delta_1(\boldsymbol{z}_i^L)} \le \left(1-\frac{n^{-\frac{1}{2}\mu}}{\delta}\right)^{-D} = e^{D\left[\frac{n^{-\frac{1}{2}\mu}}{\delta}+o\left(n^{-\frac{1}{2}\mu}\right)\right]}. \tag{50}$$

Therefore, by (44), (49), and (50),

$$\begin{aligned} &p\left(\sum_{j=1}^{D_L}\delta_1(\boldsymbol{z}_j^L) > \frac{D\mu}{\log^3 n}\right) \\ &\quad\le e^{-\left[\frac{1}{2}\mu^2/\log^2 n - \frac{n^{-\frac{1}{2}\mu}}{\delta} - o\left(n^{-\frac{1}{2}\mu}\right)\right]D} \\ &\quad\le e^{-\left[\frac{1}{2}\mu^2/\log^2 n - o\left(\frac{1}{\log^2 n}\right)\right]D}. \end{aligned} \tag{51}$$

In a similar fashion, we consider now the parsing of $\boldsymbol{x}$ into shifted L phrases

$$\boldsymbol{x} = x_1^K, x_{K+1}^{L+K}, x_{L+K+1}^{2L+K}\cdots x_{iL+K+1}^{(i+1)L},\cdots; 1\le K\le L-1$$

and let

$$\delta_K(\boldsymbol{z}^L) = \begin{cases} 1, & \text{if } \boldsymbol{z}^L = x_{iL+K+1}^{(i+1)L+K} \text{ for some } i=1,2,\cdots, \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

Then, similar to the derivation of (47), we get:

$$p\left[\sum_{j=1}^{D_L}\delta_K(\boldsymbol{z}_j^L) > \frac{D\mu}{\log^3 n}\right] \le e^{-\frac{1}{2}\left[\mu^2/\log^2 n - o\left(n^{-\frac{1}{2}\mu}\right)\right]D}. \tag{53}$$

Now, let $\boldsymbol{z}^L\in\boldsymbol{x}$ denote the event that $\boldsymbol{z}^L = x_i^{i+L}$ for some $1\le i\le n-L$ and let

$$\delta(\boldsymbol{z}^L) = \begin{cases} 1, & \text{if } \boldsymbol{z}^L\in\boldsymbol{x}, \\ 0, & \text{otherwise.} \end{cases} \tag{54}$$

Then, by (52)–(54) and the union bound

$$p\left(\sum_{j=1}^{D_L}\delta(\boldsymbol{z}_j^L) > \frac{D\mu}{\log^3 n}\right) \le Le^{-\left[\frac{1}{2}\mu^2/\log^2 n + o\left(n^{-\frac{1}{2}\mu}\right)\right]D} \tag{55}$$

Finally, as in (30), let

$$\hat{\delta}(\boldsymbol{z}^{L_i}) = \begin{cases} 1, & \text{if } \boldsymbol{z}^{L_i}\in\boldsymbol{x}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{z}^{L_i}$ is some phrase (of an arbitrary length L_i) in a block of D *distinct* phrases generated under (29).

Then, by (24) and (55),

$$p\left(\sum_{i=1}^{D}\hat{\delta}(\boldsymbol{z}_i^{L_i}) > \frac{D\mu(\overline{L})^2}{\log^3 n}\right) \le (\overline{L})^2 e^{-\left[\frac{1}{2}\mu^2/\log^2 n - o(n^{-\mu/2})\right]D} \tag{56}$$

where (56) follows from (55) and the union bound, since there are at most $\overline{L}$ different phrase lengths. Hence, by (33) and (56)

$$\begin{aligned} &p\left(\sum_{i=1}^{D}\hat{\delta}(\boldsymbol{z}_i^{L_i}) > \frac{D\mu(1+\mu)^2}{\log n\log^2\left(\frac{1}{1-\delta}\right)} + o\left(\frac{\mu(1+\mu)}{\log^2 n}\right)\right) \\ &\quad\le e^{-\left[\frac{1}{2}\mu^2/\log^2 n - o(n^{-\mu^2})\right]D}, \end{aligned} \tag{57}$$

where, by (37),

$$D = n^{1/4\left[\frac{\log\left(\frac{1}{1-\delta'}\right)}{\log 1/\delta}\right]}.$$

Equipped with (42) and (57) we can now claim that with high probability most phrases $\boldsymbol{z}^{L_i}$ in $\boldsymbol{z}$ are such that $\hat{\delta}(\boldsymbol{z}^{L_i}) = 0$and hence, at least one incremental LZ phrase starts within every such $\boldsymbol{z}^{L_i}$, thus relating $\hat{c}$ to $c(\boldsymbol{z}|\boldsymbol{x})$ as was done in the derivation of (31).

We first bound the contribution of "bad" D-blocks of phrases (namely blocks in which not all the D phrases are distinct) by (42) and then bound the contribution of "bad" phrases (i.e., phrases for which $\hat{\delta}(\boldsymbol{z}_i^{L_i}) = 1$) within a "good" D-block by (57).

Let

$$\mathbf{B} = \left\{\boldsymbol{z} : \sum_{j=1}^{\frac{\hat{c}}{D}}\delta(\boldsymbol{z}_j(D)) > \mu\frac{\hat{c}}{D}\right\} \tag{58}$$

Then, by (42)

$$q(\mathbf{B}) \le e^{-K''(\delta,\delta',\mu)\frac{n^{\mu'}}{\log n} + o\left(\frac{n^{\mu'}}{\log n}\right)}. \tag{59}$$

Now, for $\boldsymbol{z}\in A^\ell - \mathbf{B}$, we have by (30), (31) and (53) and for small enough μ that

$$\begin{aligned} &p\left[\boldsymbol{x} : \frac{1}{n}c(\boldsymbol{z}|\boldsymbol{x})\log n + \frac{1}{n}\log p(\boldsymbol{z}) < -2\mu\log\frac{1}{\delta}\right] \\ &\quad\le ne^{-\frac{1}{2}[\mu^2/\log^2 n - o(n^{-\mu^2})]n^{1/4}\frac{\log\frac{1}{1-\delta'}}{\log 1/\delta}} \end{aligned} \tag{60}$$

which proves part c) of Theorem 1. □

Remarks:

1) We have ignored the end-effects, for the sake of simplicity, by assuming, in the proof of part c) of Theorem 1, that D divides $\hat{a}$ and that L divides n. It is easy to show that (54) holds even without these assumptions.
2) Similar results can be derived for the more general case where n, the length of $\boldsymbol{x}$ is not the same as that of $\boldsymbol{z}$.

Now, by (9), we let

$$\Delta(\boldsymbol{z}\|\boldsymbol{x}) \stackrel{\Delta}{=} \frac{1}{n}[c(\boldsymbol{z}|\boldsymbol{x})\log n - c(\boldsymbol{z})\log c(\boldsymbol{z})]. \tag{61}$$

We next prove that $|\Delta(\boldsymbol{z}\|\boldsymbol{x}) - D(q_{\boldsymbol{z}}\|p)|$ tends to zero in probability.

It follows from [12] that for any finite-state, finite alphabet probability measure $q(\cdot)$ (e.g., a finite-order Markov source) and every $\boldsymbol{z} \in \mathcal{A}^{\ell}$,

$$\frac{1}{n}c(\boldsymbol{z})\log c(\boldsymbol{z}) < -\frac{1}{n}\log q(\boldsymbol{z}) + O\left(\frac{\log\log n}{\log n}\right) \tag{62}$$

and in particular, for $q(\boldsymbol{z}) = q_{\boldsymbol{z}}(\boldsymbol{z})$:

$$\frac{1}{n}c(\boldsymbol{z})\log c(\boldsymbol{z}) < H_{\boldsymbol{z}} + O\left(\frac{\log\log n}{\log n}\right).$$

Furthermore, since $c(\boldsymbol{z})\lceil \log c(\boldsymbol{z}) + \log A\rceil$ is a length-function of a uniquely decipherable code [1],then by the Kraft's inequality

$$\begin{aligned}
1 &\geq \sum_{\mathcal{A}^{\ell}} 2^{-c(\boldsymbol{z})[\log c(\boldsymbol{z})+\log A+1]} \\
&\geq \sum_{\boldsymbol{z}:c(\boldsymbol{z})\log c(\boldsymbol{z}) < -\log q(\boldsymbol{z})-\epsilon n} 2^{-c(\boldsymbol{z})[\log c(\boldsymbol{z})+\log A+1]} \\
&\geq \left[\sum_{\boldsymbol{z}:c(\boldsymbol{z})\log c(\boldsymbol{z}) < -\log q(\boldsymbol{z})-\epsilon n} 2^{-\log q(\boldsymbol{z})}\right] 2^{-2c(\boldsymbol{z})\log A+\epsilon n} = \\
&= 2^{-2c(\boldsymbol{z})\log A+\epsilon n} q(\boldsymbol{z} : c(\boldsymbol{z})\log c(\boldsymbol{z}) < -\log q(\boldsymbol{z}) - \epsilon n).
\end{aligned} \tag{63}$$

However, as was shown in [1],

$$c(\boldsymbol{z}) \leq O\left(\frac{n}{\log n}\right)$$

Hence

$$\begin{aligned}
q(\boldsymbol{z} : c(\boldsymbol{z})\log c(\boldsymbol{z}) < -\log q(\boldsymbol{z}) - \epsilon n) \\
\leq 2^{-\epsilon n + O\left(\frac{n}{\log n}\right)}.
\end{aligned} \tag{64}$$

We are now well equipped to prove Theorem 2 as follows.

Proof of Theorem 2: Part a) follows directly from (6), (10), (61), and (62).

As for part b), given $\boldsymbol{z}$, let $q_{\boldsymbol{z}}(a^{\ell'+1})$ denote the empirical probability of $(\ell'+1)$-vectors in $\boldsymbol{z}$.

Let $Q_{\boldsymbol{z}}^{\ell'+1} = \{q_{\boldsymbol{z}}(a^{\ell'+1}); a^{\ell'+1} \in \mathcal{A}^{\ell'+1}\}$ denote the empirical probability distribution of $(\ell'+1)$-vectors in $\boldsymbol{z}$.

The type $T^{\ell'+1}(\boldsymbol{z})$ is defined by

$$T^{\ell'+1}(\boldsymbol{z}) = \{\boldsymbol{y} \in \mathcal{A}^n : Q_{\boldsymbol{y}} = Q_{\boldsymbol{z}}\}$$

The number of types is no larger than $(n+1)^{A^{\ell'+1}}$ (see [13]). Thus, by (60),replacing $q(\cdot)$ by $q_{\boldsymbol{z}}(\cdot)$, it follows that the fraction of vectors in $T^{\ell'+1}(\boldsymbol{z})$ for which

$$c(\boldsymbol{z})\log(\boldsymbol{z}) < -\log q_{\boldsymbol{z}}(\boldsymbol{z}) - \epsilon n$$

is upper-bounded by

$$2^{-\epsilon(n)+O\left(\frac{n}{\log n}\right)}(n+1)^{A^{(\ell'+1)}} \tag{65}$$

Since $(n+1)^{-A^{(\ell'+1)}}$ is a lower bound on the probability $q_{\boldsymbol{z}}[T^{\ell'+1}(\boldsymbol{z})]$ [13] and since all the vectors in the type are equally probable.

Now, all n-vectors in each $T_{\boldsymbol{z}}^{\ell'+1}$ have the same $q_{\boldsymbol{z}}(\cdot)$ probability measure (since $q_{\boldsymbol{z}}(\cdot)$ is an ℓ'-order Markov measure). Hence, part b) of Theorem 2 follows from (5), (9), and (11) directly by (61). □

III. Universal Discrimination of Individual Sequences via Finite-State Classifiers: A Nonprobabilistic Approach

In this section, we demonstrate how the results of Section II can be applied to universal classification of individual sequences. The underlying idea is that the information divergence $D(\cdot\|\cdot)$, which serves as a useful discrimination function, is estimated by means of the cross-parsing procedure, that is by $\Delta(\cdot\|\cdot)$.

The problem of classifying information sources is traditionally posed in a probabilistic framework, where the goal is normally to minimize the probability of error or some other related performance criterion. In classical theory of hypothesis testing (see, e.g., [4], [5]), complete knowledge of the underlying probability measure is assumed. Since this assumption is rarely met in practice, a considerable effort has been made in recent years to relax its necessity and to develop for certain classes of sources, universal classification rules that are independent of the unknown underlying statistics and yet perform asymptotically as well as the optimal likelihood ratio test (see, e.g., [6]–[11]).

In this section, motivated by the results of Section II, we attempt to make an additional step toward universality. Similarly to the approach taken in [1], rather than modeling the data generation mechanism by an underlying probability measure, we allow the data to be arbitrary but limit the class of permissible classification rules to consist only of these which are implementable by a finite-state machines (FSM's). This set-up often reflects a realistic situation, where we have no faithful statistical model on one hand, and we are equipped with limited computational resources on the other.

We first formulate the classification problem under consideration. For the sake of simplicity and convenience, we shall consider a two-class problem and assume that all observation sequences are of the same length n. The results will extend straightforwardly to the more general situation.

The following generic notation will be used. Let $\phi_i = \{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i \cdots \boldsymbol{x}_k^i\}$, $i = 1, 2$, denote two disjoint given collections of k arbitrary vectors in $\mathcal{A}^n$. To avoid cumbersome notation, when we refer to an arbitrary member of either ϕ_1 or ϕ_2,

we shall sometimes omit the subscript j and the superscript i of $\boldsymbol{x}_j^i$. The sequences $\{\boldsymbol{x}_j^i\}$, $j = 1, 2 \cdots k$, will be referred to as *training sequences* and ϕ_i, $i = 1, 2$, are the *training sets* corresponding to "sources" σ_i, $i = 1, 2$. Here, a source represents an abstract entity which governs the respective training set in the sense of possessing certain characteristic features which are shared by vectors of one training set but not by vectors of the other. Therefore, it is convenient to think of a source as a large hidden collection of vectors which is partly seen via the training set, i.e., $\sigma_i \supseteq \phi_i$. Normally, each training set contains only a small fraction of vectors from the source. For example, the source may include an exponentially large number of sequences (as a function of n) while the training set size k is only polynomial or even fixed.

The classification problem is as follows. Upon observing a test vector $\boldsymbol{y} \in \mathsf{A}^n$ to be classified and given the training sets ϕ_i, $i = 1, 2$, decide whether $\boldsymbol{y} \in \sigma_i$, $i = 1, 2$ or reject $\boldsymbol{y}$, namely, decide that $\boldsymbol{y} \in \sigma_0 \stackrel{\Delta}{=} \sigma_1^c \cap \sigma_2^c$, where the superscript c denotes the complementary set. A classification rule M is a partition of A^n into three disjoint sets M_i, $i = 0, 1, 2$, where M_0 is the rejection region, and M_1 and M_2 are decision regions corresponding to the two sources σ_1 and σ_2, respectively. We seek a classification rule M that is both *consistent* with respect to the training sets, i.e., $M_i \supseteq \phi_i$, $i = 1, 2$, and *error free*, namely, $M_1 \cap \sigma_2 = M_2 \cap \sigma_1 = \emptyset$. These are two conflicting goals because the consistency requirement means that M_i should be "large enough" to include the corresponding training set, but on the other hand, it should be "reasonably small" to avoid confusion with the other source.

The permissible family M of classification rules M considered here consists of classifiers that can be realized by FSM's followed by modulo-n counters. Specifically, an S-state classifier is a triple $C = (\boldsymbol{S}, g, \Omega)$, where S is a finite set of states with S elements, $g : \mathsf{A} \times \boldsymbol{S} \rightarrow \mathrm{S}$ is a *next-state function*, and Ω is a partition of the space of empirical probability measures on $\mathsf{A} \times \boldsymbol{S}$, into three disjoint regions Ω_i, $i = 0, 1, 2$, depending on the training sets, where Ω_0 is the rejection region and Ω_1 and Ω_2 are acceptance regions of σ_1 and σ_2, respectively. When a test sequence $\boldsymbol{y} = y_1, y_2 \cdots y_n$ is fed into C, which in turn is initialized with $s_0 \in \mathrm{S}$, a state sequence $\boldsymbol{s} = s_1, s_2 \cdots s_n$, $s_i \in \boldsymbol{S}$, is generated by

$$s_t = g(y_t, s_{t-1}), \qquad t = 1, 2 \cdots n. \tag{66}$$

Let $n_{\boldsymbol{y}}^g(a, s)$, $a \in \mathsf{A}$, $s \in \boldsymbol{S}$, denote the joint count of $y_t = a$ and $s_{t-1} = s$ along the pair sequence $(\boldsymbol{y}, \boldsymbol{s})$, and let $q_{\boldsymbol{y}}^g(a, s) = n_{\boldsymbol{y}}^g(a, s)/n$ denote the empirical joint probability of a and s with respect to g. The empirical joint probability distribution $Q_{\boldsymbol{y}}^g = \{q_{\boldsymbol{y}}^g(a, s), a \in \mathsf{A}, s \in S\}$ serves as test statistics for classifying $\boldsymbol{y}$, that is, the classification rule $M = \{M_i\}_{i=0}^2$, associated with a FS classifier C is given by

$$M_i = \{\boldsymbol{y} \in \mathsf{A}^n : Q_{\boldsymbol{y}}^g \in \Omega_i\}, \qquad i = 0, 1, 2, \tag{67}$$

where the partition Ω (and hence also M) in turn depends on the training sequences via the empirical distributions $Q_{\boldsymbol{x}_j^i}^g$, $j = 1, 2 \cdots k$, $i = 1, 2$. These empirical distributions are precomputed in the training phase.

Two sequences $\boldsymbol{y}$ and $\boldsymbol{z}$ are said to be of the same *type with respect to* g, or, of the same *g-type* if $Q_{\boldsymbol{y}}^g = Q_{\boldsymbol{z}}^g$. The g-type of $\boldsymbol{z}$ is defined as

$$T_g(\boldsymbol{z}) \stackrel{\Delta}{=} \{\boldsymbol{y} \in \mathsf{A}^n : Q_{\boldsymbol{y}}^g = Q_{\boldsymbol{z}}^g\}. \tag{68}$$

It should be pointed out that by confining our interest to FS classifiers, we exclude uninteresting trivialities, e.g., the classification rule $M_i = \phi_i$, $i = 1, 2$, which requires an exponential number of states as a function of n. Furthermore, we avoid the need for an exponentially large number of training sequences in each set. Generally speaking, a good training set need not contain more than one representative from each g-type of every source sequence, because two training sequences of the same g-type carry exactly the same information accessible to a classifier C that employs g as a next-state function. Since there are less than $(n+1)^{AS}$ different types with respect to every g [13], there is no point in sampling more than $(n+1)^{AS}$ good training sequences for a given g from each source.

We first observe that for a given next-state function g, the smallest acceptance regions associated with a consistent FS classifier must include the entire g-type of each training sequence. Thus, the acceptance regions

$$M_i = \bigcup_{j=1}^{k} T_g(\boldsymbol{x}_j^i), \qquad i = 1, 2, \tag{69}$$

are the smallest possible for a given g in the previously defined sense. Thus, if for some g, the union of the g-types of training sequences from each class does not intersect with the other source set, then the above classification rule is also error free. Otherwise, the two sources are not separable (or distinguishable) by any S-state classifier. This observation, together with the fact that classification is performed w.r.t. the empirical distributions $\{Q_{\boldsymbol{x}_j^i}^g\}$, motivates the following definition.

Definition: A pair of sources (σ_1, σ_2) is said to be (ϵ, S)-*separable* relative to ϕ_1 and ϕ_2 if there exists a next-state function g of an S-state FSM such that for every $\boldsymbol{y} \in \sigma_1$, every $\boldsymbol{z} \in \sigma_2$, and some $\boldsymbol{x} \in \phi_1 \cup \phi_2$,

$$|D(Q_{\boldsymbol{z}}^g \| Q_{\boldsymbol{x}}^g) - D(Q_{\boldsymbol{y}}^g \| Q_{\boldsymbol{x}}^g)| > \epsilon, \tag{70}$$

where

$$D(Q_{\boldsymbol{z}}^g \| Q_{\boldsymbol{x}}^g) = \sum_{a \in \mathsf{A}, s \in \boldsymbol{S}} q_{\boldsymbol{z}}^g(a, s) \log \frac{q_{\boldsymbol{z}}^g(a|s)}{q_{\boldsymbol{x}}^g(a|s)}. \tag{71}$$

It should be kept in mind, on the other hand, that if every source sequence is represented by at least one training sequence of the same type, then M_i, as previously defined, covers σ_i. Another desirable property that we would like a classifier to have is robustness.

Definition: A classification rule M that employs a next-state function g is called *robust* w.r.t. a given training set (ϕ_1, ϕ_2), if for every $\nu > 0$, every $\boldsymbol{y} \in M_i$, $(i = 1, 2)$, and every n-sequence $\boldsymbol{y}'$ whose Hamming distance $d_H(\boldsymbol{y}, \boldsymbol{z})$ from $\boldsymbol{y}$ does not exceed $n \cdot \nu$, we have

$$\frac{1}{n} |\log Q_{\boldsymbol{x}}^g(\boldsymbol{y}) - \log Q_{\boldsymbol{x}}^g(\boldsymbol{y}')| \leq \delta_n(\nu) \tag{72}$$

where $\delta_n(\nu) \to 0$ as $\nu \to 0$ uniformly for every $\boldsymbol{x} \in \phi_1 \cup \phi_2$.

Note that $\boldsymbol{y}'$ may be considered a "noisy" version of $\boldsymbol{y}$. A classification rule is robust if it is not too sensitive to a small noise (or "dithering") level. This desirable insensitivity to noise allows us to assume that the test sequences to be classified all have strictly positive probabilities $Q_{\boldsymbol{x}}^g(\cdot)$. This in turn, will enable to invoke Theorem 2 later on. Intuitively speaking, robustness is associated with a *vanishing* memory of the classifier. Alternatively, if the memory is not vanishing, the classifier might not ever "recover" even from small perturbations in the remote past and would yield classification errors.

A natural choice of a FS machine g that has such a robustness, or vanishing memory property is that of an ℓth-order *Markovian* machine, or equivalently, a finite-memory machine. In this machine, the state s_t at time t, is the string of ℓ preceding letters, $x_{t-\ell}^{t-1}$. Hence, $q(a,s) = q(a^{\ell+1})$; $a^{\ell+1} \in \mathsf{A}^{\ell+1}$, $s \in \mathsf{A}^{\ell}$; $a \in \mathsf{A}^{\ell}$. Here the state might be affected by errors that occurred in no more than ℓ outcomes of the near past and therefore the memory is vanishing after ℓ steps. (An interesting problem in this respect is to characterize FSM's other than Markov with a vanishing memory property.) It is easy to show that a classifier based on a Markovian machine is robust if the training sequences induce conditional letter probabilities $\{q_{\boldsymbol{x}}^g(a|s)\}$ that are all bounded away from zero. This in turn can be accomplished by slightly dithering the training sequences as well, which leads to the following particular construction of a Markov classifier.

Denote by $T^{\ell}(\boldsymbol{z})$ the ℓth-order Markov type of a vector $\boldsymbol{z}$, that is, the same definition as in (68) but with g being the next-state function of a Markovian machine. Next, define the *ν-neighborhood* type $T_{\nu}^{\ell}(\boldsymbol{z})$ as the set of all sequences $\boldsymbol{y}$ such that $d_H(\boldsymbol{y}, \boldsymbol{z}') \le n \cdot \nu$ for some $\boldsymbol{z}' \in T^{\ell}(\boldsymbol{z})$. Let

$$U_i = \bigcup_{j=1}^{k} \bigcup_{\boldsymbol{x} \in T_{\nu}^{1}(\boldsymbol{x}_j^i)} \{\boldsymbol{z} : D(Q_{\boldsymbol{z}} \| Q_{\boldsymbol{x}}) \le \xi\}, \qquad i = 1, 2,$$

where $D(\cdot\|\cdot)$ is as previously defined, but w.r.t an ℓth-order Markovian machine g and $\xi > 0$ is arbitrarily small. Finally, let M^{ℓ} be defined by

$$M_1^{\ell} = U_1 - U_2$$

and

$$M_2^{\ell} = U_2 - U_1.$$

Note, that in order to avoid errors, the acceptance regions of σ_1 are subtracted from those of σ_2 and vice versa. Now, it follows from (65) and (66) that if the sources (σ_1, σ_2) are (ϵ, ℓ)-Markov separable, namely, $(\epsilon, A^{\ell})-$separable w.r.t to an ℓth-order Markovian machine, then there exists a sufficiently small positive number $\xi(\epsilon, \ell)$ such that for any $0 < \xi < \xi(\epsilon, \ell)$:

a) $M_1^{\ell} \supset \phi_1$ and $M_2^{\ell} \supset \phi_2$;
b) $M_1^{\ell} \cap M_2^{\ell} = \emptyset$.

Thus, if every ν-neighborhood Markov type of σ_i is represented in ϕ_i $(i = 1, 2)$, M^{ℓ} is essentially equivalent to M, in the sense of classifying sequences from σ_1 and σ_2 correctly. However, the sets M_i^{ℓ} might be much larger than σ_i $(i = 1, 2)$. Therefore, if we compare the performance of M^{ℓ} to that of any other robust classifier which utilizes a pair of training sets, we may allow the training sets to be drawn from M_1^{ℓ} and M_2^{ℓ}, respectively (rather than from σ_1 and σ_2 only). Therefore, hereafter the relevant reference sets for a given classifier M^{ℓ} will be M_1^{ℓ} and M_2^{ℓ} instead on the nonapparent sources σ_1 and σ_2.

We next show that a classification algorithm based on the empirical divergence given by (9) is as efficient as M^{ℓ} for almost any pair of training sets that are drawn from M_1^{ℓ} and M_2^{ℓ}, respectively, and for almost every test sequence that is sampled from either M_1^{ℓ}, M_2^{ℓ} or M_0^{ℓ}. Specifically, fix $\eta > 0$ and let us define $\hat{M}$ in the following manner.

$$\hat{M}_1 = \bigcup_{i=1}^{k} \{\boldsymbol{z} : \Delta(\boldsymbol{z}\|\boldsymbol{x}_i^1) \le \eta\} - \bigcup_{i=1}^{k} \left\{\boldsymbol{z} : -\frac{1}{n} c(\boldsymbol{z}) \log c(\boldsymbol{z}) < \Delta(\boldsymbol{z}\|\boldsymbol{x}_i^2) \le \eta\right\} \quad (73.a)$$

and

$$\hat{M}_2 = \bigcup_{i=1}^{k} \{\boldsymbol{z} : \Delta(\boldsymbol{z}\|\boldsymbol{x}_i^2) \le \eta\} - \bigcup_{i=1}^{k} \left\{\boldsymbol{z} : -\frac{1}{n} c(\boldsymbol{z}) \log c(\boldsymbol{z}) < \Delta(\boldsymbol{z}\|\boldsymbol{x}_i^1) \le \eta\right\}. \quad (73.b)$$

Observe first, that $\hat{M}$ is consistent for every sufficiently small $\eta > 0$, simply because $c(\boldsymbol{x}|\boldsymbol{x}) = 0$. A more interesting property of $\hat{M}$ is that it behaves "almost always" in a manner identical to M^{ℓ} provided that η is chosen sufficiently small (In particular, η should tend to ξ when $n \to \infty$ as will be seen later.) By "almost always" we mean the following. Let $\overline{\boldsymbol{z}}$ and $\{\overline{\boldsymbol{x}}_j^i\}$, $j = 1, \cdots k$, $i = 1, 2$, be arbitrary sequences drawn from M_1^{ℓ} and M_2^{ℓ}, respectively, and let $\boldsymbol{z}$ be an arbitrary sequence in $T^{\ell}(\overline{\boldsymbol{z}})$. Similarly, let $\phi_i = (x_1^i \cdots x_k^i)$ be an arbitrary training set in $\Phi_i = \times_{j=1}^{k} T_{\ell}(\overline{\boldsymbol{x}}_j^i)$, $i = 1, 2$. While M^{ℓ} provides the *same* classification of $\boldsymbol{z}$ for all triplets $(\boldsymbol{z}, \phi_1, \phi_2) \in G = T^{\ell}(\overline{\boldsymbol{z}}) \times \Phi_1 \times \Phi_2$, the theorem below states that the classification of $\hat{M}$ is identical to that of M^{ℓ} for every $(\boldsymbol{z}, \phi_1, \phi_2) \in G$ except for a vanishingly small fraction of triplets $(\boldsymbol{z}, \phi_1, \phi_2)$ in G. Before presenting the theorem, we observe that since most of the testing and training sequences that may be sampled from M_1^{ℓ} and M_2^{ℓ} induce strictly positive empirical transition probabilities from every any state $s \in A^{\ell}$ to any letter $a \in A$, we may assume that this is the case with the given training set and testing sequence $\boldsymbol{z}$, and thereby exclude merely a minority of possible situations for which $\hat{M}$ and M^{ℓ} might behave differently.

Theorem 3: Let the pair (σ_1, σ_2) be $(\epsilon, \ell)-$Markov separable relative to the training sets $\overline{\phi}_i = \{\overline{\boldsymbol{x}}_j^i, j = 1 \cdots k\}$, $i = 1, 2$, which are sampled from M_1^{ℓ} and M_2^{ℓ}, respectively. Assume further that every $\overline{\boldsymbol{x}}_j^i$ is associated with strictly positive empirical transition probabilities from every state $s \in A^{\ell}$ to any letter $a \in \mathsf{A}$. Then, there exists a subset $\mathbf{B}$ of test sequences

$\boldsymbol{z}$ and a subset $B_{\boldsymbol{z}}$ of training sequences, for every $\boldsymbol{z} \in \mathsf{A}^n - \mathbf{B}$, with the following properties.

a) If $\boldsymbol{z} \in \mathsf{A}^n - \mathbf{B}$ and $\boldsymbol{x}_j^i \in (\mathsf{A}^n - B_{\boldsymbol{z}}) \cap T^\ell(\overline{\boldsymbol{x}}_j^i)$ for every $j = 1 \cdots k$, $i = 1, 2$, then $\hat{M}$ classifies $\boldsymbol{z}$ in a manner identical to that of M^ℓ.

b) The set $\mathbf{B}$ satisfies

$$\lim_{n\to\infty} \frac{|T^\ell(\boldsymbol{z}) \cap \mathbf{B}|}{|T^\ell(\boldsymbol{z})|} = 0; \tag{74}$$

and for every $\boldsymbol{z} \in A^n - \mathbf{B}$,

$$\lim_{n\to\infty} \frac{|T^\ell(\overline{\boldsymbol{x}}_j^i) \cap B_{\boldsymbol{z}}|}{|T^\ell(\overline{\boldsymbol{x}}_j^i)|} = 0. \tag{75}$$

The theorem tells us that $\hat{M}$ is almost always a good approximation to M^ℓ as it accepts and rejects essentially the same vectors. Another implication is associated with the learning ability of $\hat{M}$ as compared to M^ℓ. Both rules require about the same number of training sequences in the sense explained above. It should be stressed, however, that unlike M^ℓ, the rule $\hat{M}$ does not depend on the unknown ℓ. This is an important point because the value ℓ for which the sources are (ϵ, ℓ)-Markov separable is normally unknown as it depends on the unknown σ_1 and σ_2. Furthermore, one *cannot* guarantee (ϵ, ℓ)-Markov separability by selecting "large enough" ℓ because (ϵ, ℓ)-Markov separability does not imply $(\epsilon, \ell + 1)$-Markov separability. Another advantage of $\hat{M}$ is that both the computational complexity and the memory-size associated with its implementation grow linearly with n [1], [14].

Proof of Theorem 3: Since $Q_{\boldsymbol{z}}$ and $\{Q_{\boldsymbol{x}_j^i}\}$ are assumed to induce strictly positive transition probabilities from all states to all letters, we may apply Theorem 2 to these empirical measures. In other words, let us invoke Theorem 2 with the *empirical* measures $Q_{\boldsymbol{x}}(\cdot)$ and $Q_{\boldsymbol{z}}(\cdot)$ playing the roles of the probabilistic measures $p(\cdot)$ and $q(\cdot)$ of Section II. It will now be shown that for almost all sequences in each such type, $D(Q_{\boldsymbol{z}}\|Q_{\boldsymbol{x}})$ may be efficiently replaced by $\Delta(\boldsymbol{z}\|\boldsymbol{x})$ and hence the classifier $\hat{M}$ may efficiently replace M^ℓ. This follows from a simple consideration. Since the number of ℓth-order Markov types is at most $(n+1)^{A^{\ell+1}}$, then $Q_{\boldsymbol{x}}(T^\ell(\boldsymbol{x})) \geq (n+1)^{-A^{\ell+1}}$ [13]. Thus, for every set F of n-vectors,

$$\frac{|T^\ell(\boldsymbol{x}) \cap F|}{|T^\ell(\boldsymbol{x})|} = \frac{Q_{\boldsymbol{x}}(\boldsymbol{x})|T^\ell(\boldsymbol{x}) \cap F|}{Q_{\boldsymbol{x}}(\boldsymbol{x})|T^\ell(\boldsymbol{x})|} = \frac{Q_x(T^\ell(\boldsymbol{x}) \cap F)}{Q_{\boldsymbol{x}}(T^\ell(\boldsymbol{x}))}$$

$$\leq (n+1)^{A^{\ell+1}} Q_{\boldsymbol{x}}(F),$$

which in turn, implies that

$$\lim_{n\to\infty} \frac{|T^\ell(\boldsymbol{x}) \cap F|}{|T^\ell(\boldsymbol{x})|} = 0,$$

for any set F such that $Q_{\boldsymbol{x}}(F)$ decreases with n faster than any polynomial . Similarly,

$$\lim_{n\to\infty} \frac{|T^\ell(\boldsymbol{z}) \cap \mathbf{B}|}{|T^\ell(\boldsymbol{z})|} = 0,$$

for any $\mathbf{B}$ such that $Q_{\boldsymbol{z}}(\mathbf{B})$ decreases with n in a rate faster than that of any polynomial. Now, let

$$F = B_{\boldsymbol{z}} = \left\{\boldsymbol{x} : D(Q_{\boldsymbol{z}}\|Q_{\boldsymbol{x}}) - \Delta(\boldsymbol{z}\|\boldsymbol{x}) < -3\mu_n \log\frac{1}{\delta}\right\},$$

where $\mu_n \to 0$ at a rate not faster than $\log\log n / \log n$, and let $\mathbf{B}$ be as in (54). Then, Theorem 3 now follows directly from Theorem 2. □

ACKNOWLEDGMENT

Helpful discussion with A. D. Wyner and useful comments made by the anonymous referees are greatly appreciated.

REFERENCES

[1] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*,vol. IT-24, no. 5, pp. 530–536, Sept. 1978.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[3] A. D. Wyner and J. Ziv. "Fixed-data-base version of the Lempel–Ziv data compression algorithm," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 878–880, May 1991.

[4] H. Van Trees, *Detection Estimation, and Modulation Theory*. New York: Wiley, 1968.

[5] R. E. Blahut, *Principles and Practice of Information Theory*. New York: Addison-Wesley, 1987.

[6] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 278–286, Mar. 1988.

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

[8] ——, "On tests for independence, tests for randomness and universal data compression," submitted for publication.

[9] 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, no. 5, pp. 1014–1019, Sept. 1989.

[10] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 285–290, Mar. 1991.

[11] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, Sept. 1992.

[12] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel–Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, no. 1, pp. 66–72, Jan. 1992.

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

[14] M. Rodeh, V. R. Pratt, and S. Even "Linear algorithm for data compression via string matching," *J. Assoc. Comput. Mach.*, vol. 28, sec. 3, pp. 16–24, Jan. 1981.

[15] 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, no. 4, pp. 431–438, July 1981.