ON THE OPTIMAL ASYMPTOTIC PERFORMANCE
OF UNIVERSAL ORDERING AND DISCRIMINATION
OF INDIVIDUAL SEQUENCES†

Marcelo J. Weinberger, Jacob Ziv, and Abraham Lempel

Technion - Israel Institute of Technology
Haifa 32000, Israel

Abstract. We consider the problem of ordering strings of a fixed length over a
discrete alphabet, according to decreasing probabilities of having been emitted by an
unknown finite-state source. We apply data compression to derive a universal algorithm
that solves this problem with an optimal asymptotic performance. We employ the above
result in the solution of the following problem: discriminate an individual sequence as
emitted by an i.i.d. random source or as a signal corrupted by noise. We give tight lower
and upper bounds on the asymptotic performance of finite-state discriminators.

† This research was partially supported by the US-ISRAEL Binational Science Foundation.
SUMMARY

A unifilar, ergodic, finite-state-machine (FSM) probabilistic source, over a discrete alphabet $A$ of $k$ letters, is defined by a finite set of states $S$ of cardinality $k$, probability measures $p(a | z)$, $a \in A$, $z \in S$, and a "next-state" transition function $f$ that maps $S \times A$ into $S$. The probability that an $n$-tuple $x^n_1 = x_1 x_2 \cdots x_n$, $x_i \in A$, $1 \leq i \leq n$, be emitted by the source when started at a given state $z_0$, is given by

$$P(x^n_1) = \prod_{i=1}^{n} p(x_i | z_{i-1}) \cdot \Delta f(z_{i-1}, x_i), \quad 1 \leq i \leq n.$$  

In this paper we consider the problem of ordering the $n$-tuples over $A$, according to decreasing probabilities of having been emitted by an unknown FSM source. For an $n$-tuple $x^n_1$ emitted by an FSM source with an $(a-1)k$-vector of parameters $\Theta$, let us denote by $M_\Theta(x^n_1)$ the cardinality of the set

$$\left\{ y^n_1 \in A^n : \left[ P(x^n_1) < P(y^n_1) \right] \land (y^n_1 \leq x^n_1) \right\},$$

where $y^n_1 \leq x^n_1$ means that $x^n_1$ does not precede $y^n_1$ in the lexicographic order on $A^n$. That is, $M_\Theta(x^n_1)$ is the ranking of $x^n_1$ according to its actual (unknown) probability, with $M_\Theta(u^n_1) = 1$ for the most likely sequence $u^n_1$, and $M_\Theta(v^n_1) = \alpha^n$ for the least likely sequence $v^n_1$, with ties being treated lexicographically. Note that $M_\Theta( \cdot )$ is a one-to-one mapping from $A^n$ to the subset of the integers $I_n = \{ 1, 2, \cdots, \alpha^n \}$.

Let $M( \cdot )$ be any one-to-one mapping from $A^n$ to $I_n$, independent of $\Theta$, that is, $M(x^n_1)$ denotes the ranking of $x^n_1$ in some arbitrary ordering of $n$-tuples. Finally, let $E^n_\Theta[ \cdot ]$ denote expectation relative to the actual probability distribution of $n$-tuples.
emitted by the source.

We are interested in a universal mapping $M_u(\cdot)$, in the sense that

$$E_\theta\left[\frac{M_u(x^n)}{M_\theta(x^n)}\right]$$

is sufficiently small for any $(\alpha-1)k$-vector of parameters $\theta$. This problem was first addressed in [1], where a mapping $M_\theta(\cdot)$ is proposed satisfying

$$\lim_{n \to \infty} \frac{1}{n} \log E_\theta\left[\frac{M_\theta(x^n)}{M_\theta(x^n)}\right] = 0$$

for any $(\alpha-1)k$-vector of parameters $\theta$. This universal mapping was obtained by ordering the $n$-tuples over $A$ according to increasing values of the code length obtained for $x^n$ by the LZ data compression algorithm [2]. Here, we prove the following asymptotic lower bound on the exponent

$$\frac{1}{n} \log E_\theta\left[\frac{M_u(x^n)}{M_\theta(x^n)}\right]$$

attained by any universal mapping $M_u(\cdot)$.

**Theorem 1:** For every $k$ such that $(\alpha-1)k > 2$, and for sufficiently large $n$, we have

$$\frac{1}{n} \log E_\theta\left[\frac{M_u(x^n)}{M_\theta(x^n)}\right] \geq O\left(\frac{\log n}{n}\right).$$

for almost every $\theta \in \Gamma^{(\alpha-1)k}$, where $\Gamma^{(\alpha, \beta)} = \{a, b\}$, $0 < a \leq b < 1$. (I.e., for all $\theta$ except a subset whose volume tends to zero as $n$ tends to infinity).

Theorem 1 is closely related to previous work dealing with noiseless coding of FSM sources ([3],[4]), where the rate of convergence of the compression ratio to its asymptotic value has a similar lower bound. Unlike the case of noiseless coding, it can be shown that the claim of Theorem 1 does not hold for $k=1$ and $\alpha=2$. (There exists a
universal mapping \( \overline{M}_a(\cdot) \) such that \( \frac{\overline{M}_a(x^*_T)}{M_0(x^*_T)} \leq 2 \) for every \( x^*_T \in \{0,1\}^n \) and every \( \theta \in \Gamma^{-\{0,5\}} \). The status of the case where either \( k=\alpha=2 \) or \( k=\alpha=3 \), remains open.

Next, we present a universal mapping that attains the asymptotic lower bound of Theorem 1. To this end, we need a preliminary mapping, in conjunction with Theorem 2 below, where we assume complete knowledge of the set \( S \) of states and of the next-state function \( f \), but no knowledge of the vector of parameters \( \theta \). In Theorem 2, we use the conditional, empirical "entropy" of \( x^*_T \) with respect to \( S \), defined by

\[
\hat{H}(x^*_T \mid S) \Delta \frac{-1}{\log \alpha} \sum_{a \in A} \sum_{t \in S} P_{\{1,\alpha\}}(za) \log P_{\{1,\alpha\}}(a \mid z),
\]

where

\[
P_{\{1,\alpha\}}(za) \Delta \frac{1}{n} \sum_{i=1}^n \delta(z_{i-1}, z; x_i, a),
\]

\[
\delta(z_{i-1}, z; x_i, a) \Delta \begin{cases} 
1 & \text{if } z_{i-1} = z \text{ and } x_i = a \\
0 & \text{otherwise}
\end{cases},
\]

and

\[
P_{\{1,\alpha\}}(a \mid z) \Delta \begin{cases} 
0 & \text{if } \sum_{a \in A} P_{\{1,\alpha\}}(za) = 0 \\
\frac{P_{\{1,\alpha\}}(za)}{\sum_{a \in A} P_{\{1,\alpha\}}(za)} & \text{otherwise}.
\end{cases}
\]
Theorem 2: Define $M_a^{(0)}(x^n_1)$ as the cardinality of the set

$$\left\{ y^n_1 \in A^n : \left( H(y^n_1 | S) < \hat{H}(x^n_1 | S) \right) \lor \left( H(y^n_1 | S) = \hat{H}(x^n_1 | S) \land (y^n_1 \leq x^n_1) \right) \right\}.$$ 

Then, for every $\theta$,

$$\frac{1}{n} \log E_\theta^A \left[ \frac{M_a^{(0)}(x^n_1)}{M_a(x^n_1)} \right] \leq O \left( \frac{\log n}{n} \right).$$

If the set $S$ of states is unknown, but an upper bound $K$ on its cardinality is given, we can still define a mapping $M_a^{(1)}(\cdot)$ that satisfies Theorem 2, by ordering the $n$-tuples over $A$ according to increasing values of the function

$$d_1(x^n_1) = \min_{f, R} \hat{H}(x^n_1 | R),$$

where the minimum is taken over all the sets $R$ of cardinality $K$ and all functions $f$. Ties are treated lexicographically. If $K$ is also unknown, Theorem 2 is satisfied using a technique similar to the MDL criterion [3], and a mapping $M_a^{(2)}(\cdot)$ is defined by the function

$$d_2(x^n_1) = \min_{f, R} \left[ \hat{H}(x^n_1 | R) + \frac{\alpha |R|}{n \log \alpha} \log |R| (n+1) + \frac{2 \log |R| l}{n \log \alpha} \right],$$

where the minimum is taken over all possible FSM’s and $|R|$ denotes the cardinality of the set $R$. Equivalently, we can order the $n$-tuples according to increasing values of the codelength obtained for $x^n_1$ by the coding scheme of [3]. Unfortunately, we cannot reduce the number of machines to be searched (which, as in [3], is finite) by using the state estimators presented in [5,6]. While the coding schemes presented in [2] and [3] have the property of “pointwise asymptotic optimality” (i.e., the difference between the
code-length obtained for $x^n_i$ and $\hat{H}(x^n_i | S)$ tends to zero as $n$ tends to infinity, for every $x^n_i$, the asymptotic optimality of the coding scheme of [5] was proved for the average redundancy only. Average asymptotic optimality suffices if we address the (easier) question of upper-bounding $E_0 \left[ \log \frac{M_u(x^n_i)}{M_0(x^n_i)} \right]$, which is closer to data compression.

Now, we turn to the question of universal discrimination between i.i.d. random vectors (hypothesis $H_0$) and signals corrupted by noise (hypothesis $H_1$), using a finite-state discriminator. This problem was introduced in [7], where it was shown to be equivalent to the following. Given an FSM with $k$ states and an input $x^n_i \in A^n$, assume that the output for every $x_i$, $1 \leq i \leq n$, is fed into an accumulator, and the input is "accepted" (decision $H_1$) if the accumulator exceeds a given threshold, or "rejected" ($H_0$) otherwise. We also assume that a set $S_1$ of exactly $\alpha^{(1-\lambda)n}$ input $n$-tuples is accepted. Thus, a finite-state discriminator is characterized by a set $S$ of $k$ states, a next-state function $f$, a (real) output function $g$ defined over $A^n$, and a real parameter $\lambda \in (0,1)$. We look for a universal machine $U$ such that for any given $\lambda$ and $\varepsilon > 0$, it accepts a set $S_U(\lambda)$ satisfying:

$$S_1 \subseteq S_U(\lambda),$$

and

$$|S_U(\lambda)| \leq |S_1| \alpha^{n\varepsilon},$$

for every finite-state discriminator that accepts exactly $S_1$. Note that if, in addition to the i.i.d. probability $P_0(x^n_i)$ of the input $x^n_i$ given $H_0$, we assume the existence of the probability $P_1(x^n_i)$ given $H_1$, then (1) and (2) imply
In the sequel, we restrict our discussion to universal machines $U$ satisfying the following additional condition

$$\lambda_1 \geq \lambda_2 \implies S_U(\lambda_1) \subseteq S_U(\lambda_2) \quad (3)$$

Clearly, (3) is satisfied by any finite-state discriminator.

Using our results on universal ordering we can now extend the results of [7] as follows.

**Theorem 3:**

a) If there exists a universal machine $U$ satisfying (1), (2), and (3), for every finite-state discriminator having a set $S$ of $k$ states, $k(\alpha-1) > 2$, and a next-state function $f$, then $\epsilon \geq \epsilon(n)$, where $\epsilon(n) = O\left(\frac{\log n}{n}\right)$.

b) There exists a universal machine (not modeled as an FSM), satisfying (1), (2), and (3), for every finite-state discriminator, every $\epsilon > 0$, and every $n$ such that $\epsilon \geq \epsilon(n)$, where $\epsilon(n) = O\left(\frac{\log n}{n}\right)$.

A universal discriminator satisfying Theorem 3 (b) is defined by the "accepted" set

$$\overline{S}_U(\lambda) \triangleq \left\{ x^*_1 \in A^n : 1 - d_2(x^*_1) - \frac{\log \pi^2/6}{n \log \alpha} > \lambda - \epsilon \right\}.$$
References


