b) Assume (9) and (11) and define Q as in (A.3). Then Q satisfies (8). By (8) and (9) Q is an optimum test channel. On the other hand, (A3) and (11) yield (10), which shows the optimality of p.

Discussion: By (8), the transition probabilities $Q_{\min}(k \mid j)$ of an optimum test channel for a source given by $p_{\max}$ are not defined for $j \in \Omega$ with $p_{\max}(j) = 0$. The reason is that for such $j \in \Omega$ the transition probabilities do not affect $J_s(p_{\max}, Q_{\min})$. On the other hand, the transition probabilities of a saddle point are not freely selectable for $j \in \Omega$ with $p_{\max}(j) = 0$ due to their optimum property (7a). If the saddle point is uniquely determined, these probabilities are specified by (A.3).

Acknowledgment

The late Prof. H. Brehm of the Technical University of Erlangen-Nürnberg, who died in 1990, was the author's mentor at the time the work on this correspondence was carried out. This correspondence is dedicated to his memory.

References

[1] T. Berger, "The source coding game," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 71–76, Jan. 1971.
[2] D. J. Sakrison, "The rate distortion function for a class of sources," *Inform. Contr.*, vol. 15, pp. 165-195, 1969.
[3] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[4] W. G. Bath and V. D. Vandelinde, "Robust memoryless quantization for minimum signal distortion," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 296–306, Mar. 1982.
[5] D. Kazakos, "Robust noiseless source coding through a game theoretic approach," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 576–583, July 1983.
[6] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[7] G. Owen, *Spieltheorie*. New York: Springer Verlag, 1971.
[8] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.
[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. Singapore: McGraw-Hill, 1987.
[10] K. Trottler, "Informationstheoretische Untersuchungen zur Vektorquantisierung von sphärisch invarianten Sprachmodellprozessen," Dr.Ing. Dissert., Univ. Erlangen-Nürnberg, 1987.

Variable-to-Fixed Length Codes Provide Better Large Deviations Performance than Fixed-to-Variable Length Codes

Neri Merhav, *Member, IEEE,* and David L. Neuhoff, *Senior Member, IEEE*

Abstract—**It is proved that for finite-alphabet, finite-state unifilar sources, variable-to-fixed length codes provide better large deviations performance of the empirical compression ratio, than fixed-to-variable length codes. It is shown how to construct a universal variable-to-fixed length code that achieves the optimal performance.**

Index Terms–**Universal data compression, variable-rate coding, variable-to-fixed length codes, fixed-to-variable length codes, large deviations, unifilar sources, finite-state sources.**

Manuscript received March 21, 1990; revised March 6, 1991.

N. Merhav was with AT&T Bell Laboratories, Murray Hill, NJ. He is now with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

D. L. Neuhoff was with AT&T Bell Laboratories, Murray Hill, NJ. He is now with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109.

IEEE Log Number 9102762.

I. Introduction

Lossless fixed-to-variable (F-V) length codes have been compared to variable-to-fixed (V-F) length codes under various performance criteria. Krichevsky and Trofimov [1] showed that optimal F-V and V-F length codes are equivalent in terms of tradeoff between the redundancy r and the delay d. Specifically, it was shown in [1] that for both types of coding schemes the minimum redundancy decays as fast as $1/d$ in the case of a known source with limited memory, and at the rate of $d^{-1}\log d$ when the source statistics are unknown. In [2] and [3] Jelinek *et al.* demonstrated that the best F-V and V-F length codes for memoryless sources provide the same exponential decay rate of the buffer overflow probability as the buffer size grows indefinitely. In [4] this result was generalized to unifilar (state calculable) Markov sources. Ziv [5] has shown that for Markovian sources with *long* memory there exists a V-F length code that provides a better compression ratio than any F-V length code with the same number of codewords. Recently, a result in the same spirit has been proved [6] for universal coding of binary memoryless sources. V-F length coding is also advantageous in the sense of avoiding error propagation and hence, easy to integrate with error correcting codes [1].

In this correspondence, we show that the best V-F length code provides a better large deviations performance than any F-V length encoder with the same number of codewords. Specifically, we introduce a random variable, referred to as the empirical compression ratio (ECR), which is defined as the length (in bits) of the encoder output word divided by the length (in bits) of the input word. As a measure of performance, we are interested in the exponential decay rate of the probability that the ECR exceeds a given threshold R in the range $H < R < 1$, where H is the entropy of the (binary) source. This is different from the commonly used performance measure of compression ratio, defined as the ratio between the *expected* output word length and the *expected* input word length (see e.g., [5]–[8]), as it quantifies the rate of convergence of the ECR and provides insight on its tail behavior. It is shown that for *any* unifilar finite-state (FS) source, the exponential decay rate of the probability that the ECR exceeds R, for the best V-F length code, is $1/R$ times faster than that of the best F-V length code with the same number of codewords, i.e., essentially the same amount of storage. Thus, the results here are more general than in [5] and [6] in the sense that memoryless sources as well as Markov sources are special cases of unifilar FS sources. Furthermore, the best performance in both F-V and V-F length code classes are attained by universal codes that depend neither on the source nor on the value of R.

II. Preliminaries and Problem Formulation

A unifilar finite-alphabet, FS source is characterized by an alphabet X, a state set S, a conditional probability matrix $p(\cdot \mid \cdot)$, and a next-state function $f: X \times S \to S$. At each time instant t, the source emits a letter $X_t \in X$ and moves to a new state $S_{t+1} \in S$. For a fixed initial state S_0, the probability of a given source sequence $X_1^n = (X_1, X_2, \cdots, X_n)$ is given by

$$P(X_1^n) = \prod_{t=1}^{n} p(X_t \mid S_t), \tag{2.1}$$

0018-9448/92$03.00

where S_t is uniquely determined by the previous source letter X_{t-1} and the previous state S_{t-1} using the recursion, $S_t = f(X_{t-1}, S_{t-1})$. This property enables one to reconstruct the state sequence $S_1^n = (S_1, S_2, \cdots, S_n)$ from X_1^n and the initial state S_0.

For the sake of simplicity, we shall assume throughout that the source is binary ($X = \{0, 1\}$). Clearly, this does not affect the generality, as every $|X|$-ary unifilar FS source with $|S|$ states can be easily transformed into a binary unifilar FS source having the same entropy, at the expense of adding states. This can be done by representing each source letter by $\log_2 |X|$ bits and adding substates associated with transitions between bits within each original source letter. It will be assumed also that the source is ergodic and that the initial state S_0 is fixed and known to both the encoder and decoder.

A code is characterized by a super-alphabet C_n of $N = 2^n$ binary strings (words) $X_1, \cdots, X_N$, where X_i is of length $l(X_i)$ bits, and by a one-to-one mapping from vectors $X_i \in C_n$ into binary strings (codewords) of length $L(X_i)$ bits. It is assumed that the code is proper and complete [3], i.e., every infinite binary string $X = (X_1, X_2, \cdots)$ has one and only one prefix $X_i \in C_n$. For the sake of simplicity, we shall adopt throughout the sequel the abbreviated notations $l(X)$ and $L(X)$ for infinitely long sequences X, where $l(X) \triangleq l(X_i)$, and $L(X) \triangleq L(X_i)$, X_i being the prefix of X in C_n.

The compression ratio of a code is defined as

$$\rho \triangleq \frac{EL(X)}{El(X)}, \tag{2.2}$$

where $E(\cdot)$ denotes expectation with respect to the source P. The *empirical compression ratio* (ECR), associated with a code C_n and an infinite source string X is defined as

$$\rho(X) \triangleq \frac{L(X)}{l(X)}. \tag{2.3}$$

A F-V length code is a code for which $C_n = \{0, 1\}^n$ and $l(X_i) = \log N = n$, $1 \le i \le N$. A V-F length code is one for which $L(X_i) = n$, $1 \le i \le N$. Hence, the ECR of a V-F and F-V length codes with 2^n codewords are given, respectively, by

$$\rho_{FV}(X) = \frac{L(X)}{n}, \tag{2.4a}$$

$$\rho_{VF}(X) = \frac{n}{l(X)}. \tag{2.4b}$$

It is well known [9] that the compression ratio ρ of any lossless code is lower bounded by the source entropy H, which for a stationary source P is given by

$$H \triangleq -\lim_{n\to\infty} \frac{1}{n} \sum_{X_1^n \in X^n} P(X_1^n) \log P(X_1^n). \tag{2.5}$$

Our objective is to compare the best achievable performance of sequences of V-F length codes to that of sequences of F-V length codes, in the sense of maximizing

$$\theta(R) \triangleq \liminf_{n\to\infty} \left[-\frac{1}{n} \log \Pr\{\rho(X) > R\} \right], \tag{2.6}$$

for every $H < R < 1$.

Note that for the commonly used compression ratio (2.2) the expectations in the numerator and the denominator are taken *independently*, hence, it does not necessarily provide a faithful statistical characterization for the behavior of the random variable $\rho(X)$. Unfortunately, investigating $E[\rho(X)]$ as an alternative to (2.2) seems to be much more difficult in the general case. Instead, the proposed performance measure (2.6) is associated with the cumulative probability distribution function $F_\rho(R)$ of $\rho(X)$, which for the range $R \in (H, 1)$ (corresponding to the distribution tail), can be evaluated by large deviations techniques or combinatorial techniques [10], [11].

III. Main Results

Define the empirical distribution,

$$q_X(x, s) = \frac{1}{n} \sum_{t=1}^{n} \delta(X_t = x, S_t = s), \quad x \in X, \quad s \in S, \tag{3.1}$$

where $\delta(X_t = x, S_t = s)$ is the indicator function for $X_t = x$ jointly with $S_t = s$. Also, let $q_X(s) = \sum_{x\in X} q_X(x, s)$ and

$$q_X(x \mid s) = \begin{cases} q_X(x, s)/q_X(s), & q_X(s) > 0, \\ 0, & q_X(s) = 0. \end{cases} \tag{3.2}$$

Let $Q_X \triangleq \{q(x, s), x \in X, s \in S\}$ and define the empirical entropy as

$$H(Q_X) \triangleq -\sum_{s\in S} \sum_{x\in X} q_X(x, s) \log q_X(x \mid s). \tag{3.3}$$

The Kullback–Leibler divergence is defined as

$$D(Q_X \| P) \triangleq \sum_{s\in S} \sum_{x\in X} q_X(x, s) \log \frac{q_X(x \mid s)}{p(x \mid s)}. \tag{3.4}$$

where $P = \{p(x, s), x \in X, s \in S\}$, $p(x, s) \triangleq \Pr\{X_t = x, S_t = s\}$, and $p(x \mid s) \triangleq \Pr\{X_t = x \mid S_t = s\}$. It is easy to show by (2.1), (3.3) and (3.4) that for unifilar sources considered here,

$$P(X_1^n) = \exp_2\{-n[H(Q_X) + D(Q_X \| P)]\}. \tag{3.5}$$

A. F-V Length Codes

We first present an upper bound on $\theta(R)$, as defined in (2.6), for F-V length codes with 2^n codewords. Then, a simple universal F-V length code that attains this bound is demonstrated. The results stated in this subsection summarize Section III of [12].

Theorem 1: For every sequence $\{C_n\}_{n\ge 1}$ of uniquely decipherable F-V length codes, C_n with 2^n codewords, any finite-state unifilar binary source P, and every $H < R < 1$,

$$\limsup_{n\to\infty} \left[-\frac{1}{n} \log \Pr\{\rho_{FV}(X) > R\} \right] \le \theta_{FV}^*(R), \tag{3.6}$$

where

$$\theta_{FV}^*(R) \triangleq \min_{\{Q:\, H(Q)\ge R\}} D(Q \| P). \tag{3.7}$$

The proof appears in [12, (18)–(20)].

Note that the event $\rho_{FV}(X) > R$, or equivalently, $L(X) > nR$ can be interpreted as an error event associated with a rate R fixed-to-fixed (F-F) encoder, which consists of a F-V encoder

followed by truncation to nR bits. Theorem 1 tells us that the error exponent is overbounded by $\theta^*_{\mathrm{FV}}(R)$. It should be pointed out that for a general rate R, F-F length encoder, which assigns nR bits to each one the 2^{nR} most likely source n-tuples and makes an error for all the rest, the error exponent is again tightly overbounded by $\theta^*_{\mathrm{FV}}(R)$ [10], [13], [14]. This means that there is no loss of optimality, in that sense, when using a F-V length encoder followed by truncation.

To achieve the asymptotic exponential rate $\theta^*_{\mathrm{FV}}(R)$, assume first that the next-state function f is known and consider the well-known simple universal code that consists of $|S|\log(n+1)$ bits (neglecting roundoff terms) allocated to encode the estimated source parameters Q_X, followed by $-\log Q_X(X_1^n)$ bits assigned to Huffman coding of X_1^n with respect to Q_X. This results in a length function,

$$\begin{aligned} L(X) &= -\log Q_X(X_1^n) + |S|\log(n+1). \\ &= nH(Q_X) + |S|\log(n+1). \end{aligned} \tag{3.8}$$

It is also shown in [12, Theorem 1] that the Lempel–Ziv algorithm [15] attains $\theta^*_{\mathrm{FV}}(R)$ when used as a F-V length code for n-tuples. If f is unknown, then $2\cdot|S|\log|S|$ extra bits are needed to encode the index of the best f in the sense of minimizing $H(Q_X)$, among all $|S|^{2|S|}$ possible functions.

B. V-F Length Codes

For a V-F length code with 2^n codewords, the event $\rho_{\mathrm{VF}}(X) > R$ is equivalent to the event $l(X) < n/R$. The next theorem establishes an upper bound on $\theta(R)$, as defined in (2.6), for V-F length codes.

Theorem 2: For every sequence $\{C_n\}_{n\ge 1}$ of complete and proper V-F length codes, C_n with 2^n codewords, any finite-state unifilar binary source P, and every $H < R < 1$,

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log\Pr\{\rho_{\mathrm{VF}}(X) > R\}\right] \le \theta^*_{\mathrm{VF}}(R) \triangleq \frac{1}{R}\theta^*_{\mathrm{FV}}(R). \tag{3.9}$$

Proof: Let C_n be a given V-F length code with 2^n codewords and a length function $l(X)$. As C_n is assumed proper and complete, it can be represented by a complete binary tree with 2^n leaves, where each internal node has two children. All words of length $l(X)$ greater than n/R can be shortened to n/R bits, with no loss in performance, by pruning all subtrees with roots at depth n/R, at the benefit of growing words which were originally shorter than n/R. This is possible because there are as many as $2^{n/R}$ possible words of length n/R, while only 2^n words are needed. We now have a modified tree C'_n with all codewords no longer than n/R, and with probability $\Pr\{l(X) < n/R\}$ not greater than that of the original code C_n. Consider next a transformation of C'_n into an F-V length encoder with $2^{n/R}$ codewords in the following manner: Every word with length $l(X) < n/R$, is extended to n/R bits by all $\exp_2[n/R - l(X)]$ possible suffixes, and accordingly, the n-bit codeword for this word is also extended by all possible $(n/R - l(X))$-bit suffixes. We now have a F-V length code C''_n with $2^{n/R}$ codewords and with length function denoted by $L''(X)$. Note that the event $l(X) < n/R$ for C_n is equivalent to the event $L''(X) > n$ for C''_n. Hence, by applying Theorem 1 to F-V length codes for blocks of size n/R, we arrive at

$$\begin{aligned} &\limsup_{n\to\infty}\left[-\frac{1}{n}\log\Pr\{\rho_{\mathrm{VF}}(X) > R\}\right] \\ &\quad= \limsup_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left\{l(X) < \frac{n}{R}\right\}\right] \\ &\quad\le \limsup_{n\to\infty}\left[-\frac{1}{n}\log\Pr\{L''(X) > n\}\right] \\ &\quad= \limsup_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left\{\frac{L''(X)}{n/R} > R\right\}\right] \\ &\quad\le \frac{1}{R}\theta^*_{\mathrm{FV}}(R) = \theta^*_{\mathrm{VF}}(R). \end{aligned} \tag{3.10}$$

This completes the proof of Theorem 2. □

We now demonstrate a simple universal V-F length code with no more than 2^n codewords, which asymptotically attains $\theta^*_{\mathrm{VF}}(R)$ uniformly for every P and every R.

Let $H(Q_X^l) \triangleq H(Q_{X_1^l})$ be the empirical entropy associated with X_1^l for some positive integer l. We would like each word in the code to be essentially $n/H(Q_X^l)$ bits long. Let C_n^* be the set of all source strings X_1^l, where l is the smallest integer such that

$$l\cdot\left[H(Q_X^l) + \epsilon_n\right] > n, \tag{3.11}$$

and ϵ_n is chosen such that the total number of codewords produced does not exceed 2^n.

We first show that $\epsilon_n = O(n^{-1}\log n)$ is large enough for this purpose. To do this, we need to overbound the number N_{l+1} of words in C_n^* whose length is $l+1$. Let $T_X^l \triangleq T_{X_1^l} \subset X^l$ denote the type of an l-vector X_1^l. By construction, every $X_1^{l+1} \in C_n^*$ satisfies $l[H(Q_X^l) + \epsilon_n] \le n$. Hence, for every $Y_1^l \in T_X^l$, at least one of the one-bit extensions is a member of C_n^*. In other words, either (Y_1^l, "0") or (Y_1^l, "1") or both belong to C_n^*. Thus, since $|T_X^l| \le 2^{nH(Q_X^l)}$ (which can be easily seen from (3.5) and the fact that $\Pr\{T_X^l\} \le 1$), we find that

$$\begin{aligned} N_{l+1} &\le 2\cdot\sum_{T_X^l:\, lH(Q_X^l)\le n - l\epsilon_n} |T_X^l| \\ &\le 2\cdot\sum_{T_X^l:\, lH(Q_X^l)\le n - l\epsilon_n} 2^{lH(Q_X^l)} \\ &\le 2\cdot\sum_{T_X^l:\, lH(Q_X^l)\le n - l\epsilon_n} 2^{n - l\epsilon_n} \\ &\le 2(l+1)^{|S|}2^{n-l\epsilon_n}, \end{aligned} \tag{3.12}$$

where the last inequality follows from the fact that the number of types of l-vectors does not exceed $(l+1)^{|S|}$ in the binary case. Now, by (3.11), the shortest word in C_n is essentially $n/(1+\epsilon_n)$ bits long. Hence, for any $\delta > 0$ and n sufficiently large, the cardinality of C_n^* is overbounded as follows.

$$\begin{aligned} |C_n^*| &\le \sum_{l:\, l+1 \ge n/(1+\epsilon_n)} N_{l+1} \le 2^n \\ &\quad\cdot\left[2\cdot\sum_{l\ge n(1-\delta)}(l+1)^{|S|}2^{-l\epsilon_n}\right]. \end{aligned} \tag{3.13}$$

It is now easy to verify that in order to keep the right-hand side of (3.13) less than 2^n, ϵ_n should be $O(n^{-1}\log n)$.

Before we show that the V-F length code defined in (3.11) attains $\theta^*_{\mathrm{VF}}(R)$, we first establish the fact that $\{lH(Q_X^l)\}_{l\ge 1}$ is a monotonically increasing sequence for any string X. Let $X_1^l \in X^l$ be the

prefix of $X_1^{l+1} \in X^{l+1}$. Clearly, for any source P in the class $\boldsymbol{P}_S$ of binary unifilar Markov sources with $|S|$ states, $P(X_1^l) \geq P(X_1^{l+1})$ and hence, by (3.5) and by the fact that $D(Q_X \| P) \geq 0$,

$$2^{-lH(Q_X^l)} = \max_{P \in \boldsymbol{P}_S} P(X_1^l) \geq \max_{P \in \boldsymbol{P}_S} P(X_1^{l+1}) = 2^{-(l+1)H(Q_X^{l+1})}, \quad (3.14)$$

or equivalently, $lH(Q_X^l) \leq (l+1)H(Q_X^{l+1})$.

To prove the asymptotic optimality of C_n^*, we note that the event $l(X) \leq n/R$ is equivalent to the event

$$\frac{n}{R}\left[H(Q_X^{n/R}) + \epsilon_n\right] > n, \quad (3.15)$$

by the monotonicity of the left hand side of (3.11) as a function of l. Finally,

$$\begin{aligned}\Pr\{\rho(X) > R\} &= \Pr\left\{l(X) < \frac{n}{R}\right\} \\ &\leq \Pr\left\{l(X) \leq \frac{n}{R}\right\} \\ &= \Pr\left\{\frac{n}{R}\left[H(Q_X^{n/R}) + \epsilon_n\right] > n\right\} \\ &= \Pr\{H(Q_X^{n/R}) > R - \epsilon_n\}. \end{aligned} \quad (3.16)$$

By a technique similar to [12, (19), (20)] but with n replaced by n/R, it is easy to verify that (3.16) decays with the exponential rate

$$\theta_{\mathrm{VF}}^*(R) = \frac{1}{R}\theta_{\mathrm{FV}}^*(R). \quad (3.17)$$

IV. Concatenating Short Codewords

The exponential growth of the storage needed for both V-F and F-V length codes is certainly the main practical difficulty in their implementation. Consider a codebook of 2^n codewords (henceforth, a 2^n-codebook) defined by n/k successive usages of a subcodebook of 2^k codewords, where k divides n and assumed *fixed*. It is interesting to find the best achievable exponent (in n) for sequences of F-V and V-F length codes of this structure. It turns out, as we show in this section, that for memoryless sources it is possible to attain an exponential rate $\theta(R)$, which is smaller than the optimal rate for 2^n-codebooks, but only by a term that decays as fast as $k^{-1}\log k$. In other words, for k reasonably large but fixed, it is possible to attain performance arbitrarily close to the optimum, just by repeatedly using a 2^k-subcodebook. We demonstrate this fact for V-F length codes, however, the same technique can be used for the F-V case as well.

Consider the 2^n-codebook of a V-F length code generated by all possible combinations of n/k words from the 2^k-codebook of a V-F length subcode. Since each word from the subcodebook is assigned to a k-bit subcodeword, a total of $k \cdot n/k = n$ bits are needed for the full codeword. To encode a given string X, we first parse it into n/k phrases $X_1, X_2, \cdots, X_{n/k}$ with respect to the subcodebook, and then encode each one of them separately using the subcode. As the V-F length subcode, we use the one described in Section III-B, (3.11), with n replaced by k. Let $l(X_i)$ be the length of the ith phrase. We are interested in the asymptotic behavior of

$$\Pr\left\{\sum_{i=1}^{n/k} l(X_i) < n/R\right\} \quad (4.1)$$

in the memoryless case. Using the Chernoff bound we obtain,

$$\begin{aligned}\Pr\left\{\sum_{i=1}^{n/k} l(X_i) < n/R\right\} &\leq \min_{\alpha \geq 0} E \exp_2\left\{\alpha\left[\frac{n}{R} - \sum_{i=1}^{n/k} l(X_i)\right]\right\} \\ &= \min_{\alpha \geq 0} 2^{\alpha n/R}\left[E2^{-\alpha l(X)}\right]^{n/k}. \end{aligned} \quad (4.2)$$

We now focus on the term $E2^{-\alpha l(X)}$. To overbound this term, we use the facts that $l(X) \geq k/[H(Q_X^l) + \epsilon_k]$, where $\epsilon_k = O(k^{-1}\log k)$, $|T_X^l| \leq \exp_2[lH(Q_X^l)]$, and that the number of types of binary l-vectors is $(l+1)$.

$$\begin{aligned} E2^{-\alpha l(X)} &\leq \sum_{l=k(1+\epsilon_k)}^{k/\epsilon_k+1} 2^{-\alpha l} \sum_{X_1^l:\, H(Q_X^l) \geq k/l-\epsilon_k} P(X_1^l) \\ &\leq \sum_{l=k/(1+\epsilon_k)}^{k/\epsilon_k+1} 2^{-\alpha l} \sum_{T_X^l:\, H(Q_X^l) \geq k/l-\epsilon_k} |T_X^l| \exp_2 \\ &\quad \cdot\left\{-l\left[H(Q_X^l) + D(Q_X^l \| P)\right]\right\} \\ &\leq \left(\frac{k}{\epsilon_k} + 1\right) \max_{0 \leq k/l \leq 1+\epsilon_k} 2^{-\alpha l}(l+1) \\ &\quad \cdot \max_{Q_X^l:\, H(Q_X^l) \geq k/l-\epsilon_k} 2^{-lD(Q_X^l \| P)} \\ &\leq \left(\frac{k}{\epsilon_k} + 2\right)^2 \max_{0 \leq k/l \leq 1+\epsilon_k} 2^{-\alpha k \cdot \frac{l}{k}} \\ &\quad \cdot \max_{Q_X^l:\, H(Q_X^l) \geq k/l-\epsilon_k} 2^{-k \cdot \frac{l}{k} D(Q_X^l \| P)} \\ &\leq \left(\frac{k}{\epsilon_k} + 2\right)^2 \max_{0 \leq r \leq 1+\epsilon_k} 2^{-\frac{\alpha}{r}k} \\ &\quad \cdot \max_{Q:\, H(Q) > r-\epsilon_k} 2^{-\frac{k}{r}D(Q \| P)} \\ &= \exp_2\left\{-k\left[\min_{0 \leq r \leq 1+\epsilon_k} \frac{1}{r}\right.\right. \\ &\quad \left.\left. \cdot\left(\alpha + \min_{Q:\, H(Q) > r-\epsilon_k} D(Q \| P)\right) - \delta_k\right]\right\} \\ &= \exp_2\left\{-k\left[\min_{0 \leq r \leq 1+\epsilon_k} \frac{1}{r}\left(\alpha + \theta_{\mathrm{FV}}^*(r - \epsilon_k)\right)\right.\right. \\ &\quad \left.\left. - \delta_k\right]\right\}, \end{aligned} \quad (4.3)$$

where $\delta_k = k^{-1}\log(k/\epsilon_k + 2)^2 = O(k^{-1}\log k)$. Combining

(4.2) and (4.3), we arrive at

$$\begin{aligned}\Pr\left\{\sum_{i=1}^{n/k} l(X_i) < n/R\right\} &\le \min_{\alpha\ge 0} \exp_2\left\{\alpha\frac{n}{R} - n\left[\min_{0\le r\le 1+\epsilon_k}\frac{1}{r}\cdot\left(\alpha+\theta^*_{\mathrm{FV}}(r-\epsilon_k)\right) - \frac{\delta_k}{k}\right]\right\}\\ &= \exp_2\left\{-n\left[\max_{\alpha\ge 0}\left(\min_{0\le r\le 1+\epsilon_k}\frac{1}{r}\cdot\left(\alpha+\theta^*_{\mathrm{FV}}(r-\epsilon_k)\right) - \frac{\alpha}{R}\right) - \frac{\delta_k}{k}\right]\right\}\\ &= \exp_2\left\{-n\left[\beta(R) - \frac{\delta_k}{k}\right]\right\}, \qquad (4.4)\end{aligned}$$

where

$$\beta(R) \triangleq \max_{\alpha\ge 0}\min_{r\in[0,1+\epsilon_k]}\frac{1}{r}\left[\alpha+\theta^*_{\mathrm{FV}}(r-\epsilon_k)\right] - \frac{\alpha}{R}. \qquad (4.5)$$

We now show that $\beta(R) \approx \theta^*_{\mathrm{VF}}(R)$. It was shown in [16] (see also [3], [12]) that for memoryless sources considered here,

$$\theta^*_{\mathrm{FV}}(r) = \max_{\lambda\ge 0}\lambda(r - H_\lambda), \qquad (4.6)$$

where H_λ is the Rényi entropy of the source,

$$H_\lambda = \frac{\lambda+1}{\lambda}\log\sum_{x\in X} p(x)^{\frac{1}{1+\lambda}}. \qquad (4.7)$$

Hence, by (4.5) and (4.6),

$$\begin{aligned}\beta(R) &= \max_{\alpha\ge 0}\min_{r\in[0,1+\epsilon_k]}\max_{\lambda\ge 0}\left[\lambda\left(1-\frac{H_\lambda+\epsilon_k}{r}\right)+\frac{\alpha}{r}-\frac{\alpha}{R}\right]\\ &\ge \max_{\alpha\ge 0}\max_{\lambda\ge 0}\min_{r\in[0,1+\epsilon_k]}\cdot\left\{\frac{1}{r}\left[\alpha-\lambda(H_\lambda+\epsilon_k)\right]+\lambda-\frac{\alpha}{R}\right\}\\ &\triangleq \max_{\alpha\ge 0}\max_{\lambda\ge 0} h(\alpha,\lambda), \qquad (4.8)\end{aligned}$$

where

$$h(\alpha,\lambda) = \begin{cases}\alpha\left(\dfrac{1}{1+\epsilon_k}-\dfrac{1}{R}\right)+\lambda\dfrac{1-H_\lambda}{1+\epsilon_k}, & \alpha\ge\lambda(H_\lambda+\epsilon_k),\\ -\infty, & \alpha<\lambda(H_\lambda+\epsilon_k)\end{cases}. \qquad (4.9)$$

Since $R < 1 < 1+\epsilon_k$, we have

$$\frac{1}{1+\epsilon_k}-\frac{1}{R} < 0, \qquad (4.10)$$

and thus, the value of α that maximizes $h(\alpha,\lambda)$ is $\alpha_\lambda = \lambda(H_\lambda + \epsilon_k)$. Hence,

$$\begin{aligned}\beta(R) &\ge \max_{\lambda\ge 0} h(\alpha_\lambda,\lambda)\\ &= \max_{\lambda\ge 0}\left[\lambda(H_\lambda+\epsilon_k)\left(\frac{1}{1+\epsilon_k}-\frac{1}{R}\right)+\lambda\frac{1-H_\lambda}{1+\epsilon_k}\right]\\ &= \frac{1}{R}\max_{\lambda\ge 0}\lambda(R-\epsilon_k-H_\lambda)\\ &= \frac{1}{R}\theta^*_{\mathrm{FV}}(R-\epsilon_k)\\ &= \theta^*_{\mathrm{VF}}(R) - O(k^{-1}\log k), \qquad (4.11)\end{aligned}$$

where we used the differentiability of $\theta^*_{\mathrm{VF}}(R)$ and the fact that $\epsilon_k = O(k^{-1}\log k)$. Clearly, the term $O(k^{-1}\log k)$ dominates $k^{-1}\delta_k$ in (4.4), and we conclude that

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left\{\sum_{i=1}^{n/k} l(X_i) < n/R\right\}\right] \ge \theta^*_{\mathrm{VF}}(R) - O\left(\frac{\log k}{k}\right). \qquad (4.12)$$

Finally, it should be pointed out that the redundancy term $O(k^{-1}\log k)$ in (4.9) can be decreased to $O(1/k)$ if one uses a subcode matched to the source rather than a universal subcode.

V. Conclusion

We have seen that the best V-F length code provides an exponential rate of $\Pr\{\rho(X) > R\}$, which is $1/R$ times faster than the best F-V length code with the same number of codewords 2^n. An interesting question for future research is whether or not $\theta^*_{\mathrm{VF}}(R)$ is the best achievable exponential rate for $\Pr\{\rho(X) > R\}$ among all *variable-to-variable* codes with 2^n codewords, as well.

References

[1] R. E. Krichevsky and V. E. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory,* vol. IT-27, pp. 199–207, Mar. 1981.

[2] F. Jelinek, "Buffer overflow in variable-length coding of fixed rate sources," *IEEE Trans. Inform. Theory,* vol. IT-14, pp. 409–501, May 1968.

[3] F. Jelinek and K. S. Schneider, "On variable-length-to-block coding," *IEEE Trans. Inform. Theory,* vol. IT-18, pp. 765–774, Nov. 1972.

[4] —, "Variable-length encoding of fixed-rate Markov sources for fixed-rate channels," *IEEE Trans. Inform. Theory,* vol. IT-20, pp. 750–755, Nov. 1974.

[5] J. Ziv, "Variable-to-fixed length codes are better than fixed-to-variable length codes for Markov sources," *IEEE Trans. Inform. Theory,* vol. 36, pp. 861–863, July 1990.

[6] T. J. Tjalkens and F. M. J. Willems, "A universal variable-to-fixed source code based on Lawrence's algorithm," preprint.

[7] —, "Variable to fixed-length codes for Markov sources," *IEEE Trans. Inform. Theory,* vol. IT-33, pp. 246–257, Mar. 1987.

[8] B. Fitingof and Z. Waksman, "Fused trees and some new approaches to source coding," *IEEE Trans. Inform. Theory,* vol. 34, pp. 417–424, May 1988.

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

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

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

[12] N. Merhav, "Universal coding with minimum probability of code-

word length overflow," *IEEE Trans. Inform. Theory,* vol. 37, pt. I, pp. 556-563, May 1991.
[13] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory,* vol. IT-20, pp. 197-199, Mar. 1974.
[14] R. E. Krichevskii and V. K. Trofimov, "Optimal coding of unknown and inaccurately known sources," *Topics Inform. Theory, Proc. Coll. Math. Soc. I,* Bolyai Keszthely, Hungary, 1975, pp. 425-430.
[15] 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.
[16] F. Jelinek, *Probablistic Information Theory.* New York: McGraw-Hill, 1968.

A Universal Model Based on Minimax Average Divergence

Cheng-Chang Lu and James George Dunham, *Member, IEEE*

***Abstract*—Given a set of training samples, the commonly used approach to determine a universal model is accomplished by averaging the statistics over all training samples. It is suggested to use average divergence as a measurement for the effectiveness of a universal model and propose a minimax universal model that minimizes the maximum average divergence among all training samples. Efficient searching algorithms are developed and experimental results are also presented.**

***Index terms*—Source coding, minimax methods.**

I. Introduction

For a data compression system, it is reasonable to split the task into two considerations: source modeling and string encoding [1]. Several efficient source coding techniques have been proposed and widely applied [2]. Given the availability of these source coding techniques, finding the best way to determine the source model becomes a crucial problem in data compression.

Source modeling is intended to capture the structure and statistics of the entire information string. For a nonadaptive data compression system, the source model required by the encoding unit has to be determined from the input sequence before the corresponding codes can be generated. This requires each individual input sequence to pass the source modeling part of the system first and then go through the coding part. Such implementation may not be acceptable in terms of extra memory and time required to describe and calculate the source model for each individual input sequence. One alternative to eliminate this overhead is to seek a universal model for a class of sequences that can be utilized to model similar data sets. Given a set of training samples, the commonly used approach to determine this universal model is accomplished by averaging the statistics over all training samples. In this correspondence, we suggest to use average divergence as a measurement for the effectiveness of a universal model and also propose a minimax universal model that minimizes the maximum average divergence among all training samples.

Manuscript received April 23, 1990; revised March 11, 1991. This work was presented in part at the IEEE International Symposium on Information Theory. San Diego, CA, January 14-19, 1990.

C.-C. Lu is with the Department of Mathematical Sciences, Kent State University, Kent, OH 44242.

J. D. Dunham is with the Department of Electrical Engineering, Southern Methodist University, Dallas, TX 75275.

IEEE Log Number 9102760.

This correspondence is organized as follows. In Section II, we first analyze the conditioning-tree source model and its performance bounds. In Section III, we present a universal model based on minimax average divergence. In Section IV, efficient searching algorithms are developed and experimental results are also presented.

II. Conditioning-Tree Source Model

The conditioning-tree source model was proposed in [2], [3] and has been utilized as a source structure to determine the context for each input symbol. Consider the binary conditioning tree with 3 terminal nodes as shown in Fig. 1. This tree can be used to parse sequence U as follow. When the input u_n is received, if the previous input u_{n-1} is 0, then the terminal node t_1 is reached and the context of U is set to be t_1; otherwise, the context of U is set to be t_2 if u_{n-2} is equal to 0 or the context of U is set to be t_3 if u_{n-2} is equal to 1.

The T-conditioning entropy of the conditioning tree T for a given input sequence i is defined by

$$H(Q_i \mid T) = \sum_{t \in R} H_i(a \mid t) \cdot q_i(t), \tag{1}$$

where R is a set of terminal nodes of the conditioning tree T, $q_i(t)$ is the probability of the terminal node t appeared in the sequence i, and

$$H_i(a \mid t) = \sum_{a \in S} - q_i(a \mid t) \cdot \log q_i(a \mid t), \tag{2}$$

where $q_i(a \mid t)$ is the conditional probability for symbol "a" given terminal node t in the input sequence i, S is a set of source alphabets and Q_i is the stochastic matrix based on $q_i(a \mid t)$. It is known that $H(Q_i \mid T)$ is the lower bound for the code rate that can be achieved in compressing the given input sequence i if the conditioning tree T is employed as the source structure and the empirical distribution Q_i estimated from the input sequence i is used for probability distributions. As mentioned previously, we need to seek a universal model for a class of sequences to eliminate storage and computation overhead. Assume that the same conditioning tree T is used and the stochastic matrix P is chosen as the universal probability distribution, then T-conditioning inaccuracy for a given input sequence i with stochastic matrix Q_i

$$E(Q_i \| P \mid T) = \sum_{t \in R} \left\{ \left[- \sum_{a \in S} q_i(a \mid t) \cdot \log p(a \mid t) \right] \cdot q_i(t) \right\} \tag{3}$$

is the lower bound for the code rate [3]. The average divergence [4] between the stochastic matrix Q_i of the ith training sequence and the universal model P is defined by

$$\begin{aligned} D(Q_i \| P \mid T) &= E(Q_i \| P \mid T) - H(Q_i T) \\ &= \sum_{t \in R} q_i(t) \cdot \left\{ \sum_{a \in S} q_i(a \mid t) \cdot \log \frac{q_i(a \mid t)}{p(a \mid t)} \right\}. \end{aligned} \tag{4}$$

The average divergence is the redundancy introduced by using P to design a code for the input sequence i.

III. Minimax Universal Model

Given a set of training samples, a universal model for a class of sequences needs to be found to eliminate storage and computation