[17] L. Cheded and P. A. Payne, "The exact impact of amplitude quantization on multi-dimensional, high-order moments estimation," *Signal Processing*, vol. 39, no. 3, pp. 293–315, Sept. 1994.

[18] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.

[19] R. M. Gray and T. G. Stockham, Jr., "Dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 805–812, May 1993.

[20] D. W. E. Schobben, R. A. Beuker, and W. Oomen, "Dither and data compression," *IEEE Trans. Signal Processing*, vol. 45, pp. 2097–2101, Aug. 1997.

[21] V. K Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis, and algorithms," *IEEE Trans. Inform. Theory*, vol. 44, pp. 16–31, Jan. 1998.

[22] V. K Goyal, J. Zhuang, and M. Vetterli, "Universal transform coding based on backward adaptation," in *Proc. IEEE Data Compression Conf.*, J. A. Storer and M. Cohn, Eds., Snowbird, UT, Mar. 1997, pp. 231–240.

[23] V. K Goyal and M. Vetterli, "Block transform adaptation by stochastic transform coding of Gaussian vectors," in *Proc. IEEE Digital Signal Processiong Workshop*, Bryce Canyon, UT, Aug. 9–12, 1998.

[24] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344–347, May 1985.

[25] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorizations and Its Applications*. San Diego, CA: Academic, 1979.

[26] I. E. Telatar, private communication, July 1999.

[27] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1989.

[28] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Reading, MA: Addison-Wesley, 1974.

[29] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.

Universal Coding of Nonstationary Sources

Karthik Visweswariah, Sanjeev R. Kulkarni, *Senior Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

Abstract—In this correspondence we investigate the performance of the Lempel–Ziv incremental parsing scheme on nonstationary sources. We show that it achieves the best rate achievable by a finite-state block coder for the nonstationary source. We also show a similar result for a lossy coding scheme given by Yang and Kieffer which uses a Lempel–Ziv scheme to perform lossy coding.

Index Terms—Data compression, entropy, Lempel–Ziv algorithm, nonstationary sources, universal source coding.

I. INTRODUCTION

We investigate the use of universal coding methods for coding nonstationary sources. It is widely known that Lempel–Ziv coding methods are asymptotically optimal for the coding of stationary ergodic sources. We will show that for lossless coding of finite possibly nonstationary sources Lempel–Ziv coding methods perform as well as any finite-state

block coding scheme. We will also consider as an example of a nonstationary source the Arbitrarily Varying Source and investigate the performance of universal noiseless coding schemes for this source. For lossy coding of finite sources we show that the Yang–Kieffer coding scheme asymptotically performs better than any block code even when applied to nonstationary sources.

Lempel–Ziv coding techniques are known to be asymptotically optimal for individual sequences in the sense that they perform as well as any finite-state coding scheme [11]. There are also *uniform* bounds on the performance of the incremental-parsing techniques in terms of the coding performance achievable by a finite-state coder [5]. Kieffer [3], [4] gave the optimal rate for coding nonstationary sources using finite-state coders. Given the performance of Lempel–Ziv techniques on individual sequences it is natural to investigate whether they achieve the optimal rates given by Kieffer for nonstationary sources.

II. LOSSLESS CODING

Let us consider the lossless coding of a source $\boldsymbol{X} = (X_1, X_2, \cdots)$ with distribution P. Let $X^n = (X_1, X_2, \cdots, X_n)$ take values in A^n governed by the marginal distribution P_n, where A is a finite set. The probability measure P is possibly nonstationary. We code the source using the Lempel–Ziv incremental parsing scheme given in [11]. We will define the stationary hull of P as defined in [4] and denote it by $S(P)$. For the sake of completeness we repeat the definition here.

Definition 1: Consider a (possibly nonstationary) random process $\boldsymbol{X}$ taking values in a finite set A. We say that a process $\boldsymbol{Z}$ belongs to the stationary hull of $\boldsymbol{X}$ if there exists a sequence of natural numbers $n_0, n_1, \cdots$ such that

$$\lim_{j \to \infty} \frac{1}{n_j} \sum_{i=1}^{n_j} E(f(X_i, X_{i+1}, \cdots, X_{i+m-1}))$$
$$= E(f(Z_1, Z_2, \cdots, Z_m))$$

for all real-valued functions f that depend only on finitely many coordinates.

Processes in the stationary hull capture properties of finite-dimensional distributions along various convergent subsequences. In particular, if for a given size m the best m-block code has bad performance on the nonstationary source along a particular subsequence then there is a source in the stationary hull which reflects this. It is shown in [3] that the best possible average rate at which the source can be coded using a finite-state adaptive block to variable-length code is given by

$$R(P) = \sup_{Z \in S(P)} H(Z)$$

where $H(Z)$ is the entropy rate of the stationary source Z. We will show that this rate is achieved asymptotically by the Lempel–Ziv coding scheme. Let $LZ(x^n)$ denote the length of x^n when coded by the Lempel–Ziv algorithm. We will denote by x_i^j the string $(x_i, x_{i+1}, \cdots, x_j)$. We will see that the key property we will require is a uniform bound on $LZ(x^n)$ in terms of a certain finite-state code. One such bound is given by Lemma 1 in [10, Appendix], but we could have also used the result in [5].

Theorem 1: Suppose $\boldsymbol{X}$ is a source with distribution P that takes values in a finite set. then

$$\limsup_{n \to \infty} \frac{E(LZ(X^n))}{n} \leq R(P).$$

Proof: Fix $\epsilon > 0$. From Lemma A2 in [3, Appendix] there exists a block length t and a prefix-free code $\phi : A^t \to \{0, 1\}^\infty$ such that

$$E\left(\frac{l(\phi(Z_1^t))}{t}\right) \leq R(P) + \frac{\epsilon}{2} \qquad (1)$$

for $Z \in S(P)$. Using Lemma 1 in [10, Appendix] with block length t and code ϕ we have a sequence δ_n of positive numbers tending to zero as $n \to \infty$ such that

$$LZ(x^n) \leq \min_{1 \leq j \leq t} \sum_{\substack{i=j \bmod t \\ 1 \leq i \leq n-t+1}} l\left(\phi\left(x_i^{i+t-1}\right)\right) + n\delta_n.$$

This implies that

$$LZ(x^n) \leq \frac{1}{t} \sum_{1 \leq j \leq t} \sum_{\substack{i=j \bmod t \\ 1 \leq i \leq n-t+1}} l\left(\phi\left(x_i^{i+t-1}\right)\right) + n\delta_n.$$

Dividing both sides by n and taking expectations and limits we have

$$\limsup_{n \to \infty} \frac{1}{n} E(LZ(X^n))$$

$$\leq \limsup_{n \to \infty} \left(\frac{1}{nt} \sum_{1 \leq j \leq t} \sum_{\substack{i=j \bmod t \\ 1 \leq i \leq n-t+1}} E\left(l\left(\phi\left(X_i^{i+t-1}\right)\right)\right) + \delta_n\right)$$

$$\leq \limsup_{n \to \infty} \left(\frac{1}{nt} \sum_{1 \leq i \leq n-t+1} E\left(l\left(\phi\left(X_i^{i+t-1}\right)\right)\right)\right) + \limsup_{n \to \infty} \delta_n$$

$$= \frac{1}{t} E\left(l\left(\phi\left(Z_i^{i+t-1}\right)\right)\right)$$

$$\leq R(P) + \frac{\epsilon}{2}$$

where the equality is for some $Z \in S(P)$ and follows from Lemma 1 in the Appendix, and the last inequality follows from (1). Now $\epsilon > 0$ was arbitrary so we have that

$$\limsup_{n \to \infty} \frac{E(LZ(X^n))}{n} \leq R(P). \qquad \square$$

III. EXAMPLE: CODING AN ARBITRARILY VARYING SOURCE

In this section we will use bounds on the performance of universal codes on individual sequences to investigate their performance on the arbitrarily varying source, studied as an example in [2]. The arbitrarily varying source can be used to model, for example, the piecewise-stationary Bernoulli source studied in [9].

Let S be a finite set and A be the finite alphabet on which the source takes values. Let $p(\cdot|s)$ be a probability distribution on A for each $s \in S$. An Arbitrarily Varying Source (AVS) is a nonstationary source defined by an infinite sequence $s \in S^\infty$. The probability of a string $x_1^n \in A^n$ occurring is given by

$$P(x_1^n|s_1^n) = \prod_{i=1}^n p(x_i|s_i).$$

If the underlying state sequence is known then the optimal fixed-variable coding rate is clearly

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^n H(P(\cdot|s_i)).$$

This is also the block-to-block coding rate (see [7]).

Consider a string $x_1^n \in A^n$ and a string $y_1^m \in A^m$ with $m < n$. Let us define the empirical probability distribution $\hat{P}$ by

$$\hat{P}_m^n(y_1^m) \triangleq \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} \delta(x_i^{i+m-1}, y_1^m)$$

where $\delta(x, y) = 1$ if $x = y$ and is zero otherwise. Let $\hat{H}_m$ denote the entropy of the empirical distribution on m blocks defined above. We can also define the conditional empirical probability distribution and entropy by

$$\hat{P}^n(a|y_1^m) \triangleq \frac{\hat{P}(y_1^m a)}{\hat{P}(y_1^m)}$$

and

$$\hat{H}_{1|m} = \sum_{y_1^m \in A^m} \hat{P}(y_1^m) H(\hat{P}(\cdot|y_1^m)).$$

Similarly, we can define the empirical distribution of m blocks on the state sequence s_1^n, we denote this distribution by $\hat{Q}_m^n$. Note that when we write xy where x and y are two strings we mean y concatenated with x. We also drop the m and n in the notation for the empirical distribution, the dimension of the distribution, and the length of the underlying string when these are clear from the context.

Suppose we have a code ϕ such that for any fixed m

$$l(\phi(x_1^n)) \leq n\hat{H}_{1|m} + o(n). \qquad (2)$$

Note that the $o(n)$ term does not depend on the sequence x_1^n. Codes based on the Lempel–Ziv incremental parsing scheme and the infinite depth context tree weighting method can be shown to have such a property [6], [5], [8]. We can bound the performance of such codes on arbitrarily varying sources and the result is given in Theorem 2. The theorem implies that if the underlying state sequence has low complexity then codes can learn the source and perform as well asymptotically as if the underlying state sequence were known.

Theorem 2: If a code ϕ satisfies (2) then for an arbitrarily varying source X with an underlying state sequence s and for any integer $m > 0$ we have

$$\limsup_{n \to \infty} E_s\left(\frac{l(\phi(X^n))}{n}\right) \leq \limsup_{n \to \infty} \sum_{z \in S} \hat{Q}^n(z) H(P(\cdot|z))$$

$$+ \limsup_{n \to \infty} \frac{1}{m+1} H(\hat{Q}_{m+1}^n)$$

where $\hat{Q}_m^n$ is the empirical m-dimensional distribution of s_1^n.

Proof: To prove the theorem we will require a bound on $\hat{H}_{1|m}$ in terms of $\hat{H}_{m+1}$. This would be easy if the empirical distributions were stationary. We note that the empirical distributions defined above are not stationary, that is,

$$\sum_{y_2^m \in A^{m-1}} \hat{P}(ay_2^m) \neq \sum_{y_1 y_3^m \in A^{m-1}} \hat{P}(y_1 a y_3^m).$$

To make this distribution stationary we can define the empirical distribution by

$$\hat{P}'(y_1^m) \triangleq \frac{1}{n} \sum_{i=1}^n \delta(x_i^{i+m-1}, y_1^m)$$

where all indices are considered $\bmod n$. We then have for the corresponding entropies

$$\hat{H}'_{1|m} \leq \frac{1}{m+1} \hat{H}'_{m+1}.$$

Now

$$|\hat{H}_{1|m} - \hat{H}'_{1|m}| \leq \sum_{y_1^m \in A^m} \hat{P}(y_1^m)|H(\hat{P}(\cdot|y_1^m)) - H(\hat{P}'(\cdot|y_1^m))|$$

$$+ \sum_{y_1^m \in A^m} |\hat{P}(y_1^m) - \hat{P}'(y_1^m)|H(\hat{P}'(\cdot|y_1^m))$$

$$\leq \sum_{y_1^m \in A^m} \hat{P}(y_1^m)l_1(\hat{P}(\cdot|y_1^m),$$

$$\hat{P}'(\cdot|y_1^m))\log \frac{|A|}{l_1(\hat{P}(\cdot|y_1^m), \hat{P}'(\cdot|y_1^m))}$$

$$+ \sum_{y_1^m \in A^m} l_1(\hat{P}(\cdot y_1^m), \hat{P}'(\cdot|y_1^m))H(\hat{P}'(\cdot|y_1^m)) \quad (3)$$

where $l_1(\cdot, \cdot)$ denotes the l_1 distance between two distributions on the same alphabet. The last inequality follows for sufficiently large n from [1, Theorem 16.3.2] since we can show that

$$l_1(\hat{P}(\cdot|y_1^m), \hat{P}'(\cdot|y_1^m)) \leq \frac{2m|A|}{n-m} \quad (4)$$

and hence satisfies the condition required in the theorem for sufficiently large n. Using (3) and (4) we have

$$\left|\hat{H}_{1|m} - \hat{H}'_{1|m}\right| \leq \frac{2m|A|}{n-m}\log\frac{n-m}{2m} + \frac{2m|A|2^m}{n-m}\log|A|.$$

Similarly we can show that

$$\left|\hat{H}_{m+1} - \hat{H}'_{m+1}\right| \leq \frac{2m2^m}{n-m}\log\frac{n-m}{2m}.$$

Now we can bound the performance of the code ϕ as follows:

$$l(\phi(x^n))$$

$$\leq n\hat{H}_{1|m} + o(n)$$

$$\leq n\hat{H}'_{1|m}$$

$$+ n\left(\frac{2m|A|}{n-m}\log\frac{n-m}{2m} + \frac{2m|A|2^m}{n-m}, \log|A|\right) + o(n)$$

$$\leq \frac{n}{m+1}\hat{H}'_{m+1}$$

$$+ n\left(\frac{2m|A|}{n-m}\log\frac{n-m}{2m} + \frac{2m|A|2^m}{n-m}\log|A|\right) + o(n)$$

$$\leq \frac{n}{m+1}\hat{H}_{m+1}$$

$$+ n\left(\frac{2m(|A|+2^m)}{n-m}\log\frac{n-m}{2m} + \frac{2m|A|2^m}{n-m}\log|A|\right) + o(n).$$

Let us consider the underlying state sequence $\boldsymbol{s}$. We can define, as before, a conditional empirical distribution $\hat{W}(y^{m+1}|z^{m+1})$ where $y^{m+1} \in A^{m+1}$ and $z^{m+1} \in S^{m+1}$. We will denote the empirical distribution of the underlying state sequence by $\hat{Q}$. Now we have

$$\sum_{z^{m+1} \in S^{m+1}} \hat{Q}(z^{m+1})H\left(\hat{W}(\cdot|z^{m+1})\right) + I\left(\hat{P}_{m+1}, \hat{W}_{m+1}\right)$$

$$= H\left(\hat{P}_{m+1}\right).$$

Thus we have

$$l(\phi(x^n)) \leq \frac{n}{m+1}\left(\sum_{z^{m+1} \in S^{m+1}} \hat{Q}(z^{m+1})H\left(\hat{W}(\cdot|z^{m+1})\right)\right.$$

$$\left. + I\left(\hat{Q}_{m+1}, \hat{W}_{m+1}\right)\right)$$

$$+ n\left(\frac{2m(|A|+2^m)}{n-m}\log\frac{n-m}{2m} + \frac{2m|A|2^m}{n-m}\log|A|\right)$$

$$+ o(n).$$

Note that, although it is not explicit in the notation, all the empirical distributions depend on the sequence x_1^n and/or the underlying state sequence. We can take expectations on both sides of the previous inequality assuming a fixed underlying state sequence. Also since

$$I\left(\hat{Q}_{m+1}, \hat{W}_{m+1}\right) \leq H\left(\hat{Q}_{m+1}\right)$$

we have

$$E_s\left(\frac{l(\phi(X^n))}{n}\right) \leq \frac{1}{m+1}\left(\sum_{z^{m+1} \in S^{m+1}} \hat{Q}(z^{m+1})\right.$$

$$\left. E_s\left(H\left(\hat{W} \cdot (\cdot|z^{m+1})\right)\right) + H\left(\hat{Q}_{m+1}\right)\right)$$

$$+ \frac{2m(|A|+2^m)}{n-m}\log\frac{n-m}{2m}$$

$$+ \frac{2m|A|2^m}{n-m}\log|A| + \frac{o(n)}{n}.$$

Since $H(\cdot)$ is concave we have

$$\limsup_{n \to \infty} E_s\left(\frac{l(\phi(X^n))}{n}\right)$$

$$\leq \limsup_{n \to \infty} \frac{1}{m+1}\sum_{z^{m+1} \in S^{m+1}} \hat{Q}^n(z^{m+1})$$

$$\cdot H\left(E_s\hat{W}^n(\cdot|z^{m+1})\right) + H\left(\hat{Q}_{m+1}^n\right)$$

$$= \limsup_{n \to \infty} \frac{1}{m+1}\sum_{z^{m+1} \in S^{m+1}} \hat{Q}^n(z^{m+1})$$

$$\cdot H\left(P_{m+1}(\cdot|z^{m+1})\right) + H\left(\hat{Q}_{m+1}^n\right)$$

$$\leq \limsup_{n \to \infty}\sum_{z \in S}\hat{Q}^n(z)H(P(\cdot|z))$$

$$+ \limsup_{n \to \infty}\frac{1}{m+1}H\left(\hat{Q}_{m+1}^n\right). \quad (5)$$

The last equality follows because $P_{m+1}(\cdot|z^{m+1})$ has a product form so that the corresponding entropy is just the sum of the entropies of the one dimensional distributions. $\qquad\square$

If we assume that the state sequence has zero empirical entropy rate (as defined in [11]) then as $m \to \infty$ the second term in (5) goes to zero. Assuming that the underlying state has zero empirical entropy rate means that the state sequence has patterns that can be learned by the coding algorithm. The first term in (5) is independent of m and is equal to the coding performance that would be achieved if the underlying state sequence were known.

IV. LOSSY CODING

We consider a source $\boldsymbol{X} = (X_1, X_2, \cdots)$ with distribution P. As before, let the source $X^n = (X_1, X_2, \cdots, X_n)$ take values in A^n. Let P_n govern the probabilities of n strings. The probability measure P is possibly nonstationary. We assume that A is a finite set. We assume that the reproduction alphabet is also A. Let $\rho : A \times A \to [0, \infty)$ be the distortion measure. Using ρ we can define distortion for n strings as

$$\rho_n((x_1, \cdots, x_n), (y_1, \cdots, y_n)) = \frac{1}{n}\sum_{i=1}^n \rho(x_i, y_i).$$

A block code of block size N is defined by a map $\phi : A^N \to A^N$. The average distortion for a code ϕ is defined as

$$\overline{\rho}(P, \phi) = \limsup_{n \to \infty}\sum_{x^n \in A^n} P(x^n)\rho_n(x^n, \phi(x^n)).$$

Clearly, if ϕ is a block code of size N and Q is a stationary source then

$$\overline{\rho}(Q, \phi) = \sum_{x^N \in A^N} Q\left(x^N\right) \rho_N\left(x^N, \phi\left(x^N\right)\right).$$

It is shown in [3] that the best possible distortion achievable for a source with measure P using block codes of rate R is given by

$$\mathcal{D}_b^*(R, P) = \sup_{Q \in S(P)} D_b(R, Q).$$

Since Q is stationary $D_b(R, Q)$ is known.

We will use the coding method given in [10]. As before, let $LZ(x^n)$ denote the Lempel–Ziv coding length of a string x^n. By Lempel–Ziv coding we mean the incremental parsing scheme given in [11]. Let

$$B_n(R) = \{x^n \in A^n \; : \; LZ(x^n) \le nR\}.$$

Now the size of $B_n(R)$ is no more than $2^{\lfloor nR \rfloor}$. To code a string x^n we choose that string in $B_n(R)$ that has minimum distortion with x^n. Thus we code the source using $B_n(R)$ as our code book. We can code each n block with no more than $\lfloor nR \rfloor$ bits.

Given a set C_n which is a subset of A^n let $\rho(C_n)(x^n)$ denote the minimum distortion incurred when coding the string x^n using C_n as a code book.

Theorem 3: Suppose $\boldsymbol{X}$ is a source with distribution P that takes values in a finite set then

$$\limsup_{n \to \infty} E(\rho(B_n(R))(X^n)) \le \mathcal{D}_b^*(R, P).$$

Proof: We have from the discussion after [3, eq. (3.8)], that for any $\epsilon > 0$ there exists a block code $\phi \; : \; A^N \to A^N$ of rate less than $R - \epsilon$ and block length $N > 4/\epsilon$ such that

$$\overline{\rho}(Q, \phi) \le \mathcal{D}_b^*(R - 2\epsilon, P) + \epsilon/4.$$

Since the rate of the code is at most $R - \epsilon$ there exists a length function σ which satisfies Kraft's inequality and such that $\sigma(y^N) \le N(R - \epsilon) + 2$ for any string y^N in $\phi(A^N)$.

As in [10], from the block code ϕ we can define a code ϕ_n^j as follows: $\phi_n^j(x^n)_i^{i+N-1} \triangleq \phi(x_i^{i+N-1})$ if $1 \le i \le n - N + 1$ and $i = j \mod N$. At coordinates not defined by the above equation let $\phi_n^j(x^n)_k \triangleq a_0$ for some a_0 in A. Now using Lemma 1 in [10, Appendix] we have

$$LZ(\phi_n^j(x^n)) \le n(R - \epsilon + 2/N + \delta_n)$$

where $\delta_n \to 0$ as $n \to \infty$. Once again we point out that we could use the bound in [5] instead of Lemma 1 in [10, Appendix] and obtain essentially the same result. Since $2/N < \epsilon/2$ we have that for sufficiently large n, ϕ_n^j maps A^n into $B_n(R)$. Thus we have for $1 \le j \le N$

$$n\rho(B_n(R))(X^n) \le \sum_{i=j \bmod N, \, 1 \le i \le n-N+1} N\rho_N$$
$$\cdot \left(X_i^{i+N-1}, \phi_n^j\left(X_i^{i+N-1}\right)\right) + N\rho_{\max}$$

where $\rho_{\max} = \max_{x \in A} \rho(x, a_0)$. Averaging over j and dividing by n we get

$$\rho\left(B_n(R)\right)(X^n)$$
$$\le \frac{1}{n} \sum_{i=1}^{n-N+1} \rho_N\left(X_i^{i+N-1}, \phi\left(X_i^{i+N-1}\right)\right) + \frac{N\rho_{\max}}{n}.$$

Thus we have from Lemma 1 in the Appendix that

$$\limsup_{n \to \infty} E\left(\rho\left(B_n(R)\right)(X^n)\right) \le E\left(\rho_N\left(Z^N, \phi\left(Z^N\right)\right)\right)$$

for some source $\boldsymbol{Z}$ with measure Q in the stationary hull of P.

But $E(\rho_N(Z^N, \phi(Z^N))) = \overline{\rho}(Q, \phi)$ and so

$$E\left(\rho_N\left(Z^N, \phi\left(Z^N\right)\right)\right) \le \mathcal{D}_b^*(R - 2\epsilon, P) + \epsilon/4.$$

Since $\epsilon > 0$ was arbitrary and since $\mathcal{D}_b^*$ is a convex (and hence continuous) function of R (from [3]) we have

$$\limsup_{n \to \infty} E\left(\rho\left(B_n(R)\right)(X^n)\right) \le \mathcal{D}_b^*(R, P). \qquad \square$$

APPENDIX

We now state and prove a lemma which is useful in proving Theorems 1 and 3. We are given a random process $\boldsymbol{X}$ with distribution P which takes values in A.

Lemma 1: For any real function ϕ defined on A^t, there exists $\boldsymbol{Z} \in S(P)$ such that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^n E\left(\phi\left(X_i^{i+t-1}\right)\right) = E\left(\phi\left(Z_1^t\right)\right).$$

Proof: It is clear that we can find a subset of the natural numbers N^0 such that

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^n E\left(\phi\left(X_i^{i+t-1}\right)\right)$$
$$= \lim_{n \to \infty, \, n \in N^0} \frac{1}{n} \sum_{i=1}^n E\left(\phi\left(X_i^{i+t-1}\right)\right). \quad (6)$$

The limit on the right-hand side denotes a limit along integers in N^0. Since the alphabet A is finite we can find $N^1 \subseteq N^0$ such that

$$\lim_{n \to \infty, \, n \in N^1} \frac{1}{n} \sum_{i=1}^n P\left(X_i = a\right)$$

exists for each $a \in A$. Let this limiting one-dimensional distribution be P_1'. Continuing this argument we can find $N^1 \subseteq N^2 \subseteq N^3 \subseteq \cdots$ such that the k-dimensional distributions converge along integers in N^k to a distribution P_k'. Let $\boldsymbol{Z}$ be a process defined by these stationary distributions. Now from the sets $N^1 \subseteq N^2 \subseteq N^3 \subseteq \cdots$ we can form a set N^* in the following way: Pick the smallest element n_0 from N^0. For each k pick $n_k \in N^k$ such that $n_k > n_{k-1}$. By the construction of N^* and $\boldsymbol{Z}$ it is easy– to see that

$$\lim_{n \to \infty, \, n \in N^*} \sum_{i=1}^n E\left(f\left(X_i, X_{i+1}, \cdots\right)\right) = E(f(\boldsymbol{Z}))$$

holds for all real-valued functions f that depend only on finitely many coordinates. Thus $\boldsymbol{Z}$ is in the stationary hull of $\boldsymbol{X}$. Also since $N^* \subseteq N^0$ we have using (6)

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^n E\left(\phi\left(X_i^{i+t-1}\right)\right)$$
$$= \lim_{n \to \infty, \, n \in N^*} \frac{1}{n} \sum_{i=1}^n E\left(\phi\left(X_i^{i+t-1}\right)\right)$$
$$= E\left(\phi\left(Z_1^t\right)\right). \qquad \square$$

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications)*.　New York: Wiley, 1991.

[2] M. Feder and N. Merhav, "Hierarchical universal coding," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1354–1364, Sept. 1996.

[3] J. C. Kieffer, "Fixed rate encoding of nonstationary sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 651–655, Sept. 1987.

[4] ——, "Finite-state adaptive block to variable-length noiseless coding of a nonstationary information source," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1259–1263, Nov. 1989.

[5] E. Plotnik, M. Weinberger, and J. Ziv, "Upper bounds on the probability of a sequence emitted from a finite-state source and on the redundancy of the Lempel–Ziv data compression algorithm," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1259–1263, Nov. 1989.

[6] S. Savari, "Redundancy of the Lempel–Ziv incremental parsing rule," *IEEE Trans. Inform. Theory*, vol. 43, pp. 9–21, Jan. 1997.

[7] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1158, July 1994.

[8] F. M. J. Willems, "The context-tree weighting method: Extensions," in *IEEE Int. Symp. Information Theory*, Trondheim, Norway, 1994.

[9] ——, "Coding for a binary independent piecewise-identically-distributed-source," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2210–2217, Nov. 1996.

[10] E. Yang and J. C. Kieffer, "Simple universal lossy data compression schemes derived from the Lempel–Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 42, pp. 239–245, Jan. 1996.

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

A Dynamic Programming Algorithm for Constructing Optimal "1"-Ended Binary Prefix-Free Codes

Sze-Lok Chan and Mordecai J. Golin, *Member, IEEE*

Abstract—The generic *Huffman-Encoding Problem* of finding a minimum cost prefix-free code is almost completely understood. There still exist many variants of this problem which are not as well understood, though. One such variant, requiring that each of the codewords ends with a "1," has recently been introduced in the literature with the best algorithms known for finding such codes running in exponential time. In this correspondence we develop a simple $O(n^3)$ time algorithm for solving the problem.

Index Terms—Dynamic programming, one-ended codes, prefix-free codes.

I. INTRODUCTION

In this correspondence we discuss the problem of efficiently constructing minimum-cost binary prefix-free codes having the property that each codeword ends with a "1."

We start with a quick review of basic definitions. A *code* is a set of binary words $C = \{w_1, w_2, \cdots, w_n\} \subset \{0, 1\}^*$. A word $w = \sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_l}$ is a *prefix* of another word $w' = \sigma'_{i_1}\sigma'_{i_2}\cdots\sigma'_{i_{l'}}$ if w is the start of w'. Formally, this occurs if $l \leq l'$ and, for all $j \leq l$, $\sigma_{i_j} = \sigma'_{i_j}$. For example, 00 is a prefix of 00011. Finally, a code is said to be *prefix-free* if for all pairs $w, w' \in C$, w is not a prefix of w'.

Let $P = \{p_1, p_2, p_3, \cdots, p_n\}$ be a discrete probability distribution, that is, $\forall i,\ 0 \leq p_i \leq 1$ and $\sum_i p_i = 1$. The cost of code C with distribution P is

$$\text{Cost}(C, P) = \sum_i |w_i| \cdot p_i$$

where $|w|$ is the length of word w; $\text{Cost}(C, P)$ is, therefore, the average length of a word under probability distribution P. The *prefix-coding problem* is, given P, to find a prefix-free code C that minimizes $\text{Cost}(C, P)$. It is well known that such a code can be found in

$\mathbf{C_1}$:	10	11	000	001	010	011
$\mathbf{C_2}$:	01	11	001	101	0001	1001
$\mathbf{C_3}$:	1	01	001	0001	00001	000001

Fig. 1. Code C_1 is an optimal prefix-free code for the distribution $(1/6), (1/6), (1/6), (1/6), (1/6), (1/6)$. C_2 is an optimal *One-ended* prefix-free code for the same distribution. C_3 is an optimal one-ended code for the distribution $0.9, 0.09, 0.009, 0.0009, 0.00009, 0.00001$.

$O(n \log n)$ time using the greedy *Huffman-Encoding* algorithm, see, e.g., [5] or even $O(n)$ time if the p_i are already sorted [6].

In 1990, Berger and Yeung [1] introduced a new variant of this problem. They defined a *feasible* or 1-*ended* code to be a prefix-free code in which every word is restricted to end with a "1." Such codes are used, for example, in the design of self-synchronizing codes [3] and testing. Given P, the problem is to find the minimum-cost 1-ended code. Fig. 1 gives some examples.

In their paper, Berger and Yeung derived properties of such codes, such as the relationship of a min-cost feasible code to the entropy of P, and then described an algorithm to construct them. Their algorithm works by examining all codes of a particular type, returning the minimum one. They noted that experimental evidence seemed to indicate that their algorithm runs in time exponential in n. A few years later, Capocelli, De Santis, and Persiano [4] noted that the min-cost code can be shown to belong to a *proper* subset of the code-set examined by Berger and Yeung. They, therefore, proposed a more efficient algorithm that examines only the codes in their subset. Unfortunately, even their restricted subset contains an exponential number of codes[1] so their algorithm also runs in exponential time.

In this correspondence we describe another approach to solving the problem. Instead of enumerating all of the codes of a particular type it uses dynamic programming to find an optimum one in $O(n^3)$ time.

II. TREES AND CODES

There is a very well-known standard correspondence between prefix-free codes and binary[2] trees. In this section we quickly discuss its restriction to the 1-ended code problem. This will permit us to reformulate the min-cost feasible code problem as one that finds a min-cost tree. In this new formulation we will require that $p_1 \geq p_2 \geq \cdots \geq p_n \geq 0$ but will no longer require that $\sum_i p_i = 1$.

Definition 1: Let T be a binary tree. A leaf $u \in T$ is a *left leaf* if it is a left child of its parent; it is a *right leaf* if it is a right child of its parent.

The *depth* of a node $v \in T$, denoted by $\text{depth}(v)$, is the number of edges on the path connecting the root to v.

We build the correspondence between trees and codes as follows. First let T be a tree. Label every left edge in T with a 0 and every right edge with a 1. Associate with a leaf v the word $w(v)$ read off by following the path from the root of T down to v. Now let $v_1, v_2, \cdots, v_n$ be the set of right leaves of T. Then $C(T) = \{w(v_1), w(v_2), \cdots, w(v_n)\}$ is the code associated with T. Note that this code is feasible since all of its words end with a 1. Note also that there can be many trees corresponding to the same feasible code. See Fig. 2 for an example.

---

[1]The proof of this fact is a straightforward argument that recursively builds an exponentially sized set of codes that belong to the restricted subset. Because of space considerations we do not include it here but the interested reader can find the details in [2].

[2]In this correspondence we use the slightly nonstandard convention that a binary tree is a tree in which every internal node has *one or two* children.