Universal Codeword Sets and Representations of the Integers

PETER ELIAS, FELLOW, IEEE

Abstract—**Countable prefix codeword sets are constructed with the universal property that assigning messages in order of decreasing probability to codewords in order of increasing length gives an average codeword length, for any message set with positive entropy, less than a constant times the optimal average codeword length for that source. Some of the sets also have the asymptotically optimal property that the ratio of average codeword length to entropy approaches one uniformly as entropy increases. An application is the construction of a uniformly universal sequence of codes for countable memoryless sources, in which the nth code has a ratio of average codeword length to source rate bounded by a function of n for all sources with positive rate; the bound is less than two for $n = 0$ and approaches one as n increases.**

I. Introduction

THERE ARE problems of interest to a theory of concrete computational complexity dealing with the flow, storage, and manipulation of information in which information-theoretic considerations are dominant. The number of binary comparisons required to sort a list into order is an example in which the obvious informational lower bound of $\log n!$ is closely approached by a number of schemes [12]. The rate at which strictly equiprobable independent random bits can be derived from a biased and possibly a Markov binary sequence is a more far-fetched case, of possible interest for simulation, in which the obvious informational lower bound is approached [3]. Floyd's problem of rotating a binary matrix available row by row has an informational flavor, and he uses an entropy of mixing to find a lower bound that his algorithms essentially attain [7]. Minsky and Papert's discussion of exact and approximate match uses information-theoretic bounds [14], as do other analyses of related problems of information storage and retrieval [16], [3], [4], [6]. It seems less likely that informational arguments will be as important in the analysis of computation *per se*. Nonetheless it seemed worthwhile to explore the strictly informational restrictions on the performance of computers, allowing the freedom of choice of representations for the input and output that is characteristic of information theory but is avoided by Turing theorists. Such an investigation is carried out in [5]. This paper presents some byproducts that have communications applications.

Manuscript received February 12, 1974; revised August 16, 1974. This work was supported in part by the U.S. Army Research Office, Durham, under Contract DAH-C04-71-C-0039, and in part by the National Science Foundation under Grant GK-37582. A report based in part on this paper was presented at the 1973 International Symposium on Information Theory, Ashkelon, Israel.

The author is with the Department of Electrical Engineering and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

Turing theorists are usually interested in functions the domain of which is the positive integers $N^+ = \{1,2,3,\cdots\}$ or the nonnegative integers $N = \{0,1,2,3,\cdots\}$. To represent the integers as strings of symbols on the tape of a Turing machine requires the variable-length encoding of an infinite message set, called an encoding or a representation of the integers. Many such representations are possible but few have been discussed concretely.

In the usual noiseless source coding approach, one chooses a particular representation of the integers so as to minimize the average codeword length for some specified probability distribution on the integers. There is no obvious distribution to assume and no convenient source of experimental data. It was, therefore, a pleasant surprise when it turned out that certain countably infinite prefix sets of codewords have the *universal* property that given any countable set M of messages and any probability distribution P defined on M, assigning messages in order of decreasing probability to codewords in order of increasing length gives an average codeword length that is bounded above by $K_1 + K_2H(P)$, where K_1 and K_2 are constants ≥ 1 and $H(P)$ is the entropy of the distribution P.

The first such universal codeword sets were discovered in [5]. This paper gives a more systematic treatment and introduces the new class of asymptotically optimal universal codeword sets, which can be used to encode all discrete memoryless sources with an efficiency that approaches one uniformly as source entropy increases. Section II gives definitions and variable-length noiseless coding results for the countable case. The results are standard (see Gallager [10, ch. 3 and problem 3.7] for the extensions to the countable case) except for some obvious facts about effective encoding and decoding, which are not relevant to the finite case. Section III defines and constructs universal and asymptotically optimal sets for all alphabet sizes. Theorem 3 is essentially in [5]; the rest is new, and strengthens a result due to Wyner [17] on the entropy of decreasing distributions. Section IV reviews the standard representations of the integers and shows that some are universal but that there are no universal binary representations among them. Section V constructs three new simple universal binary representations of the integers, two of which are asymptotically optimal. Section VI gives an application to a communications problem, using universal sets to construct a code that is quite simple to implement and which gives a ratio of codeword length to entropy of ~ 1.6 for any discrete memoryless source with finite nonvanishing entropy. Using an asymptotically optimal (a.o.) set and adding another step to the coding procedure gives a sequence

of universal codes for all memoryless sources in which the ratio of codeword length to entropy is uniformly bounded for each member of the sequence and the bounds approach one.

The fact that regular sets have finite-state acceptors is used in discussing the complexity of the various representations. Kohavi [13] gives a brief presentation of the relevant notation and results. Golomb [9] discussed a class of probability distributions on the integers in which the integer j has probability j^{-s}, $s > 1$, properly normalized. The optimal encodings of such distributions are universal sets but are not asymptotically optimal. Universal coding schemes for sources with unknown parameters are reviewed by Davisson [1].

II. Countable Sources and Minimal Codes

A countable source (M,P) is a countable set M of messages and a probability distribution function $P: M \to (0,1]$ that assigns a positive probability $P(m) > 0$ to each message $m \in M$.

Let B be a finite set of symbols of size $|B|$ (for example, $B = \{0,1\}$, $|B| = 2$). Let B^* denote the set of all finite sequences of symbols, each symbol selected from the fixed set B. A (variable-length) $|B|$-ary *code* for the source (M,P) is a codeword set $C \subseteq B^*$ of size $|C| = |M|$ and a one-to-one function $\gamma: M \to C$, which assigns a distinct codeword $\gamma(m) \in C$ to each message $m \in M$. The set C and the code γ are said to be *uniquely decipherable* iff every concatenation of a finite number of codewords from C is a distinct sequence of symbols in B^* so that the extension $\hat{\gamma}: M^* \to C^*$ of γ, which maps a sequence of messages into the concatenation of the corresponding codewords, is also a one-to-one function from M^* to B^*. A uniquely decipherable set $C \subseteq B^*$ is said to be *complete* iff adding any new sequence $c' \in B^*$, $c' \notin C$, to C gives a set $C' = C \cup \{c'\}$ that is not uniquely decipherable.

A codeword set $C \subseteq B^*$ is said to be a *prefix* set, and a one-to-one code $\gamma: M \to C$ is said to be a prefix code, iff no codeword in C is the beginning of another. A prefix set is uniquely decipherable since any concatenation of codewords in C^* has only one prefix that is a codeword in C.

A codeword set $C \subseteq B^*$ is said to be *effective* iff there is an effective procedure (e.g. a Turing machine) that can decide whether a given sequence in B^* is a codeword in C or not. If C is uniquely decipherable, there is an effective procedure that distinguishes concatenations of codewords in C^* from other sequences in B^* iff C is also effective. For if C is not effective no algorithm can distinguish C itself from B^*, while if C is effective an algorithm can partition a given sequence in B^* into segments in all possible ways and test the segments for membership in C.

Index the numbers of $C = \{c_1,c_2,c_3,\cdots\}$ in order of increasing length, and define the length function L of the set C by setting

$$L(j) = |c_j| \leq L(j+1) = |c_{j+1}|, \qquad 1 \leq j < |C|$$

where $|c_j|$ denotes the length of the sequence c_j. Theorem 1 relates properties of the length function L to properties of the set C.

Theorem 1:

I. Let $C \subseteq B^*$ be a uniquely decipherable set with length function L. Then

i) L is an increasing function

$$L(j+1) \geq L(j), \qquad 1 \leq j < |C|;$$

ii) L satisfies the Kraft inequality

$$\sum_{c \in C} |B|^{-|c|} = \sum_{j=1}^{|C|} |B|^{-L(j)} \leq 1;$$

iii) if there is equality in ii), C is complete;

iv) if C is effective, L is an effectively computable function.

II. Let L satisfy I i) and I ii) for some integer $|C| \in N^+$ or $|C| = |N^+| = \infty$. Then

i) there is a prefix set $C \subseteq B^*$ with length function L;

ii) if the prefix set C in II i) is complete, then every sequence in B^* is either the prefix of some sequence in C or has some sequence in C as prefix. If in addition $|C|$ is finite, there is equality in I ii);

iii) if L is effectively computable, the prefix set C in II i) is effective.

Proof: I i) is a matter of the definition of L. I ii) is standard (see [10]). I iii) follows since if there is equality in ii), adding c' to C adds a term to the sum and violates the inequality. I iv) follows since generating B^* in order of increasing length and testing for membership in C gives an algorithm that can compute L. II i) is in Gallager, as is an algorithm that generates all members of a prefix set having length k when all shorter codewords and the number of codewords of length k are known. Thus if L is computable, C is effective, proving II iii). If C is effective, testing each prefix of c for membership gives an algorithm that halts on the rightmost symbol of any $c \in C$. II ii) follows since if some $c' \in B^*$ is neither the prefix of any member of C nor has any member of C as a prefix, it can be added to C to give a prefix set C', which is, therefore, uniquely decipherable. The necessity of equality in I ii) for a complete finite prefix C is in Gallager, but there are infinite complete prefix C with arbitrarily small Kraft sums.

Let (M,P) be a countable source and let $C \subseteq B^*$ be a uniquely decipherable set of size $|C| \geq |M|$. Index $M = \{m_1,m_2,m_3,\cdots\}$ in order of decreasing probability

$$P(m_1) \geq P(m_2) \geq P(m_3) \geq \cdots$$

and recall that $C = \{c_1,c_2,c_3,\cdots\}$ is already indexed in order of increasing length

$$|c_1| \leq |c_2| \leq |c_3| \leq \cdots.$$

Then no one-to-one code from M into C has a smaller average length than does the encoding $m_j \to c_j$, which assigns longer codewords to less probable messages and

has the minimal average codeword length

$$\sum_{j=1}^{|M|} P(m_j)|c_j| \tag{1}$$

determined by C and P alone. Such a code from M to C is called *minimal*. The average codeword length (1) of a minimal code does not depend on the details of the set C, but only on its length function L and is just the average value of the function L averaged with respect to the given distribution P, denoted by $E_P(L)$

$$E_P(L) = \sum_{j=1}^{|M|} P(m_j)L(j) = \sum_{j=1}^{|M|} P(m_j)|c_j|.$$

By Theorem 1, if any uniquely decipherable code for a source (M,P) is minimal, there is a prefix code with the same average codeword length, and the prefix set is effective if the uniquely decipherable set is also. There is, therefore, no advantage in either average codeword length or effective decipherability to be gained by using a uniquely decipherable set that is not a prefix set. Theorem 2 gives well-known lower and upper bounds to $E_P(L)$ for uniquely decipherable and prefix sets.

Theorem 2: Let B be a finite set and let (M,P) be a countable source with entropy $H(P) = H$, where the entropy

$$H(P) = \sum_{j=1}^{|M|} P(m_j) \log \frac{1}{P(m_j)}$$

is computed using logarithmic base $|B|$.

i) If $C \subseteq B^*$ is uniquely decipherable with length function L and size $|C| \geq |M|$,

$$E_P(L) \geq \begin{cases} 0, & H = 0 \\ \max\{1,H\}, & 0 < H \leq \infty. \end{cases}$$

ii) There is a prefix set $C(P) \subseteq B^*$ with length function L_P given by

$$L_P(j) = \left\lceil \frac{\log 1}{P(j)} \right\rceil, \qquad 1 \leq j \leq |C|$$

(where $\lceil x \rceil$ is the least integer not less than x) and minimal average codeword length

$$E_P(L_P) \leq \begin{cases} 0, & H = 0 \\ 1 + H, & 0 < H < \infty. \end{cases}$$

The proof is standard (see [10]). Note that $H(P) = 0$ iff M has only one member $M = \{m\}$, in which case $P(m) = 1$. Then the null string λ of no symbols is in B^* and will do for the single codeword, with length $|\lambda| = 0$. The set $C = \{\lambda\}$ satisfies the prefix condition by default. For $H > 0$ there must be more than one message. Then λ is not a codeword since for any $c \in C$, $\lambda c = c$ so $\{c,\lambda\}$ is not a uniquely decipherable set. Thus all codeword lengths (and their average) must be ≥ 1. The upper bound $1 + H$, for $H > 0$, in ii) is the best possible in the sense that for every $H > 0$ and $\varepsilon > 0$ there is a distribution P with $H(P) = H$ and with $E_P(L) > 1 + H - \varepsilon$, for every L that satisfies the Kraft inequality.

For a source (M,P) with entropy H, $0 < H < \infty$, the average codeword length bound

$$E_P(L_P) \leq 1 + H$$

satisfied by the set $C(P)$ gives a bound $R_P(H)$ on the ratio of that average codeword length to its minimal possible value

$$\frac{E_P(L_P)}{\max\{1,H\}} \leq R_P(H) = \begin{cases} 1 + H, & 0 < H \leq 1 \\ 1 + \dfrac{1}{H}, & 1 \leq H < \infty \end{cases}$$

where the function $R_P(H)$ is itself bounded by a constant

$$R_P(H) \leq K_P = 2$$

and

$$\lim_{H\to\infty} R_P(H) = \lim_{H\to 0} R_P(H) = 1.$$

The limit $R_P(H) \to 1$, for large H, gives the (block to variable-length) *noiseless source coding theorem* for a stationary memoryless source, which generates an infinite sequence of messages selected from the set M, selecting successive messages with statistical independence from the fixed probability distribution P. Let P_n: $M^n \to (0,1)$ be the probability distribution on n-tuples of messages. Then $H(P) > 0$ and

$$H(P_n) = nH(P)$$

and encoding P_n into the set $C(P_n)$ of Theorem 2 ii) gives

$$\lim_{n\to\infty} R_{P_n}(H(P_n)) = \lim_{n\to\infty} R_{P_n}(nH(P)) = \lim_{n\to\infty} 1 + \frac{1}{nH(P)} = 1$$

so that for any $\varepsilon > 0$ choosing $n > 1/\varepsilon H(P)$ gives an average codeword length $< (1 + \varepsilon)H(P_n)$.

III. Universal Codeword Sets

The upper bounds $1 + H$, $R_P(H)$, and K_P in the last section need not be satisfied by a minimal code for the source (M,P) unless the codeword set chosen is the set $C(P)$ of which the length function L_P is specifically designed to match the distribution function P of that particular source. We next consider a different situation in which a single set $C_\rho \subseteq B^*$ is used for the minimal encoding of any countable source with entropy H, for $0 < H < \infty$.

Let C_ρ be a countably infinite uniquely decipherable set in B^* and let $\rho: N^+ \to C_\rho$ map the positive integers $N^+ = \{1,2,3,\cdots\}$ into the members of C_ρ in order of increasing length, so that setting $\rho(j) = c_j$ gives the kind of indexing with $|c_j| \leq |c_{j+1}|$, $j \in N^+$, assumed before. Then the function ρ is called a $|B|$-ary *representation of the integers.*

Let L_ρ be the length function of the set C_ρ so that

$$L_\rho(j) = |\rho(j)| = |c_j|.$$

Coding a source (M,P) into the set C_ρ using a minimal code gives the minimal average codeword length $E_P(L_\rho)$. Coding the same source into the set $C(P)$ of Theorem 1 or any other set could not give an average codeword length

less than max $\{1,H(P)\}$, for $0 < H < \infty$. We say that the representation ρ and the set C_ρ are *universal*, iff the ratio

$$\frac{E_P(L_\rho)}{\max\{1,H(P)\}}$$

is bounded above by a constant K_ρ independent of P, for all P with $0 < H(P) < \infty$. We say that a universal set C_ρ is *asymptotically optimal* iff the ratio is bounded above by a function of $H(P)$,

$$\frac{E_P(L_\rho)}{\max\{1,H(P)\}} \le R_\rho(H(P)) \le K_\rho$$

with

$$\lim_{H\to\infty} R_\rho(H) = 1.$$

To construct a universal representation $\rho_1: N^+ \to B^*$ of the integers, let $B = \{0, 1, \cdots, |B| - 1\}$ be the first $|B|$ nonnegative integers, and construct the codeword $c_j = \rho_1(j)$ by first writing the integer $j \in N^+$ in standard $|B|$-ary notation as a sequence of

$$1 + \lfloor \log j \rfloor$$

symbols in B^*. (All logarithms are taken to base $|B|$ unless otherwise specified, and $\lfloor x \rfloor$ means the greatest integer not greater than x.) Then insert zero between each pair of symbols and one after the last symbol in that sequence. The resulting codeword has length

$$|\rho_1(j)| = 2(1 + \lfloor \log j \rfloor)$$

and the codeword set $C_1 = \{\rho_1(j) \mid j \in N^+\}$ is the regular set

$$C_1 = (B - \{0\})(0B)^*1.$$

(See, e.g., Kohavi [13] for a definition of *regular set*.) Sequences in C_1 are accepted by a simple device that receives $\rho_1(j)$ one symbol at a time, reads and prints the symbols that arrive at odd times, continues without printing when zero arrives at an even time, and halts without printing when one arrives at an even time.

The fact that C_1 is universal follows from an inequality due to Wyner [17]. Let (M,P) be any countable source with M indexed in the standard way so that $P(m_1) \ge P(m_2) \ge \cdots$. Then

$$1 \ge \sum_{i=1}^{j} P(m_i) \ge jP(m_j)$$

so

$$j \le \frac{1}{P(m_j)}$$

$$\log j \le \frac{\log 1}{P(m_j)}.$$

Averaging with respect to P gives Wyner's inequality

$$E_P(\log) \triangleq \sum_{j=1}^{|M|} P(j) \log j \le \sum_{j=1}^{|M|} P(j) \log \frac{1}{P(j)} = H(P).$$

It follows that for C_1, with length function L_1,

$$\begin{aligned} E_P(L_1) &= \sum_{j=1}^{|M|} P(j)|\rho_1(j)| = 2 \sum_{j=1}^{\infty} P(j)(1 + \lfloor \log j \rfloor) \\ &\le 2 + 2E_P(\log) \\ &\le 2 + 2H(P) \end{aligned}$$

so that C_1 is universal with bounds

$$E_P(L_1) \le 2(1 + H(P))$$

$$\frac{E_P(L_1)}{\max\{1,H(P)\}} \le 2R_P(H(P)) \le K_1 = 4.$$

However, C_1 is not asymptotically optimal, as will be seen later.

For each $k \in N^+$, another set $C_k \subseteq B^*$ can be constructed by first writing $j \in N^+$ in base $|B|^k$ notation, which takes

$$1 + \lfloor \log_{|B|^k} j \rfloor$$

base $|B|^k$ symbols. Using the sequences in B^k as base $|B|^k$ symbols, separating them with $0 \in B$, and terminating the sequence with $1 \in B$ gives a representation $\rho_k(j)$ of length

$$\begin{aligned} |\rho_k(j)| &= (1 + \lfloor \log_{|B|^k} j \rfloor)(k + 1) \\ &\le \left(1 + \frac{1}{k} \log_{|B|} j\right)(k + 1) \\ &\le 1 + k + \left(1 + \frac{1}{k}\right) \log j \end{aligned}$$

which can be decoded by a $(k + 2)$-state acceptor. Averaging with respect to P and using Wyner's inequality proves the following.

Theorem 3: For each $|B| \ge 2$ and each $k \in N^+$ there is a universal representation

$$\rho_k: N^+ \to C_k$$

with length function L_k, bounds

$$E_P(L_k) \le U_k(H(P)) = 1 + k + \left(1 + \frac{1}{k}\right) H(P)$$

$$\frac{E_P(L_k)}{\max\{1,H(P)\}} \le R_k(H(P)) = \frac{U_k(H(P))}{\max\{1,H(P)\}}$$

$$R_k(H\ P)) \le K_k = 2 + k + \frac{1}{k},$$

and limit

$$\lim_{H\to\infty} R_k(H) = 1 + \frac{1}{k}.$$

The set C_k is a regular prefix set in B^* and has a $(k + 2)$-state halting acceptor.

Theorem 3 can be used to strengthen another result of Wyner's [17] that $E_P(\log)$ and $H(P)$ converge or diverge together. Using Theorem 2 i), Wyner's inequality, and the bound on $|\rho_k(j)|$ averaged with respect to P gives

$$E_P(\log) \le H(P) \le E_P(L_k) \le 1 + k + \left(1 + \frac{1}{k}\right) E_P(\log).$$

Choosing

$$k = \lceil \sqrt{E_P(\log)} \rceil$$

gives Theorem 4.

Theorem 4: For any decreasing probability distribution $P(1) \geq P(2) \geq \cdots$,

$$0 \leq H(P) - E_P(\log) \leq 2(1 + \sqrt{E_P(\log)})$$

so for any sequence $P_1, P_2, \cdots, P_n, \cdots$ of distributions with $H(P_n) < \infty$ and $\lim_{n\to\infty} H(P_n) = \infty$,

$$\lim_{n\to\infty} \frac{E_{P_n}(\log)}{H(P_n)} = 1.$$

Choosing increasing values of k in Theorem 3 gives sets C_k, which are asymptotically more nearly optimal, but no single C_k is itself an a.o. set. To construct an a.o. representation $\rho: N^+ \to C_\rho$, for $C_\rho \subseteq B^*$, given $j \in N^+$, find the $n = n(j) \in N^+$ such that

$$\frac{n(n+1)}{2} \geq 1 + \lfloor \log j \rfloor \geq \frac{n(n-1)}{2} + 1 \qquad (2)$$

and let

$$m = \frac{n(n+1)}{2} - 1 - \lfloor \log j \rfloor \leq n - 1.$$

Then construct $\rho(j)$ by padding the $|B|$-ary representation of j on the left with m initial zeros and partitioning the resulting sequence of $n(n+1)/2$ symbols from B into n segments, the kth segment of length k. Insert zero between each pair of segments and one after the last segment to form $\rho(j)$. Then

$$|\rho(j)| = \frac{n(n+1)}{2} + n = n + m + 1 + \lfloor \log j \rfloor$$
$$\leq 2n + \lfloor \log j \rfloor.$$

From the right side of (2),

$$8(1 + \lfloor \log j \rfloor) \geq 4n(n-1) + 8 = (2n-1)^2 + 7,$$

so

$$1 + \sqrt{1 + 8\lfloor \log j \rfloor} \geq 2n$$

and using $\lfloor \log j \leq \log j \rfloor$ gives

$$|\rho(j)| \leq \log j + 1 + \sqrt{1 + 8\lfloor \log j \rfloor}.$$

Using the Wyner inequality and the convexity $\cap$ of the square root in averaging this expression gives

$$E_P(L_\rho) \leq 1 + H(P) + \sqrt{1 + 8H(P)}$$

which proves Theorem 5.

Theorem 5: The universal representation $\rho: N^+ \to C_\rho$ is asymptotically optimal for any $|B| \geq 2$ with length function L_ρ, bounds

$$E_P(L_\rho) \leq U_\rho(H(P)) = 1 + H(P) + \sqrt{1 + 8H(P)}$$

$$\frac{E_P(L)}{\max\{1, H(P)\}} \leq R_\rho(H(P)) = \frac{U_\rho(H(P))}{\max\{1, H(P)\}}$$

$$R_\rho(H(P)) \leq K_\rho = 5$$

and limit

$$\lim_{H\to\infty} R_\rho(H) = 1.$$

The set C_ρ is not regular but is accepted by a machine or algorithm with a counter, which sets $k = 1$, starts, reads and prints the next k symbols, halts if the $(k+1)$st symbol is one, and sets $k \leftarrow k + 1$ and returns to "start" if the $(k+1)$st symbol is zero.

IV. Standard Representations of Integers

Standard representations are widely used for doing arithmetic and counting operations. Decoding an arbitrary universal set into a standard representation is possible by Theorem 1 if L is computable but need not be easy. Fortunately some standard representations are universal. Unfortunately none of them are universal binary representations or are asymptotically optimal. We will discuss several.

Unary Encoding

The simplest binary representation of $N^+ = \{1,2,3,\cdots\}$ is known in Turing machine theory as *unary encoding*. The unary codeword set

$$C_\alpha = 0^*(B - \{0\})$$

is complete and has a two-state acceptor that halts on the first nonzero symbol. The unary encoding $\alpha: N^+ \to C_\alpha$ is onto, and for $|B| > 2$ its inverse

$$\alpha^{-1}(0^k b) = k(|B| - 1) + b, \qquad k \in N, b \in B - \{0\}$$

is easier to define. The magnitude

$$|\alpha(j)| = \left\lfloor \frac{(j-1)}{(|B|-1)} \right\rfloor + 1, \qquad j \in N^+$$
$$= j \text{ when } |B| = 2$$

is essentially linear in j. Unary codes are essentially optimal for some exponential distributions (see Golomb [8]) but are not universal since

$$P(j) = \frac{6}{\pi^2 j^2}, \qquad j \in N^+$$

has finite entropy but $E_P(L_\alpha) = \infty$.

$|B|$-ary Encoding

What is usually called "the standard $|B|$-ary representation" is a code $\beta: N^+ \to C_\beta$, where $C_\beta \subseteq B^*$ is the regular set

$$C_\beta = (B - \{0\})B^*$$

and β is defined inductively by

$$\beta(j) = j, \qquad j \in B - \{0\}$$
$$\beta(k|B| + j) = \beta(k)j, \qquad \text{for } k \in N^+, j \in B.$$

By the induction

$$|\beta(j)| = k, \qquad \text{iff } |B|^{k-1} \leq j < |B|^k$$
$$|\beta(j)| = 1 + \lfloor \log_{|B|} j \rfloor = \lfloor \log_{|B|} (j+1) \rfloor.$$

Unfortunately β is not a representation since it is not uniquely decipherable. For $|B| = 2$, $\beta(5) = 101 = \beta(2)\beta(1)$ illustrates the problem.

$|B|$-ary Representation

What is usually meant by "the standard $|B|$-ary representation" has an alphabet $\{0, 1, \cdots, |B| - 1, \square\}$, where $\square$ is an end-of-word symbol. The representation $\tau: N^+ \to C_\tau$, $C_\tau \subset (B \cup \{\square\})^*$, is defined by

$$C_\tau = (B - \{0\})B^*\square$$

$$\tau(j) = \beta(j)\square$$

$$|\tau(j)| = 1 + |\beta(j)| = 2 + \lfloor \log_{|B|} j \rfloor, \qquad j \in N^+.$$

There is widespread confusion between τ and β since the square frame around $\square$ is usually omitted, leaving only the space it encloses! Yet τ is a prefix, and so is a representation. However, for $|B| = 2$, τ is a *ternary*, not a binary representation.

The Wyner inequality (using logarithmic base $|B \cup \{\square\}| = |B| + 1$) shows that τ is universal

$$E_P(L_\tau) \le 2 + H(P) \log_{|B|} (|B| + 1) = U_\tau(H(P))$$

$$R_\tau(H) = \frac{U_\tau(H)}{\max\{1,H\}} \le K = 2 + \log_{|B|} (|B| + 1).$$

The set C_τ is not complete. A representation $\hat{\tau}: N^+ \to C_{\hat{\tau}}$ with

$$C_{\hat{\tau}} = B^*\square$$

is defined by enumerating $C_{\hat{\tau}}$ ordered by length and lexicographically within each length. Then

$$|\hat{\tau}(j)| = 1 + \lfloor \log_{|B|} j(|B| - 1) \rfloor$$

$$E_P(L_{\hat{\tau}}) \le U_{\hat{\tau}}(H(P)) = 1 + \log_{|B|} (|B| - 1) + H(P) \log_{|B|} (|B| + 1)$$

$$R_{\hat{\tau}}(H) = \frac{U_{\hat{\tau}}(H)}{\max\{1,H\}} \le K_{\hat{\tau}} = 1 + \log_{|B|} (|B|^2 - 1) < 3.$$

For $|B| = 2$, $\hat{\tau}$ has a simple decoding. Since the binary encoding $\beta(j)$ always starts with one, that one can be deleted to give $\hat{\beta}: N^+ \to C_{\hat{\beta}}$ with

$$\hat{\beta}(1) = \lambda \qquad \hat{\beta}(2j) = \hat{\beta}(j)0 \qquad \hat{\beta}(2j + 1) = \hat{\beta}(j)1$$

where λ represents the null sequence of no binary symbols.

Then $\hat{\tau}(j) = \hat{\beta}(j)\square$. Representations essentially equivalent to $\hat{\tau}$ were suggested by Shannon [15] and used by Elias [2] in run-length coding. Their universal character was not realized at that time, although the universal character of the resulting run-length codes was, and these codes are evaluated in [2].

V. Universal Binary Representations

For $|B| = 2$ the unary encoding α is not universal, the binary encodings β and $\hat{\beta}$ are not representations, and the "binary" representations τ and $\hat{\tau}$ are ternary, so there is no standard universal binary representation. The universal representations ρ_k in Theorem 3 and ρ in Theorem 5 are binary for $|B| = 2$ but not complete. The representations $\hat{\tau}$ for $|B| = 2^k - 1$, $k \ge 2$, can be mapped into $\{0,1\}^*$ by mapping the alphabet $B \cup \{\square\}$ one-to-one onto $\{0,1\}^k$ (see [2] and [15]) and are then complete and binary, but their decoding into a standard representation is nontrivial, and their shortest codeword has length $k \ge 2$. It turns out that the two limiting properties of the set $C(P)$ in Theorem 2,

$$\lim_{H\to\infty} R_P(H) = \lim_{H\to\infty} R_P(H) = 1$$

are both useful in applications. We construct three complete universal binary representations γ,δ,ω, which all share the first limit and have simple acceptors and decodings into $\beta(j)$ or $\tau(j)$. Two (δ and ω) share the second limit and are, therefore, asymptotically optimal.

(*Note added in proof:* Richard Karp [18] has done some closely related work on universal binary representations.)

Compound Representation γ

The representation ρ_1 in Theorem 3 is obtained by following each of the bits of the binary representation $\beta(j)$ by a bit of the unary representation $\alpha(|\beta(j)|)$ of the length $|\beta(j)|$ of $\beta(j)$. Since in the binary case the first symbol of $\beta(j)$ is always one, it can be dropped. This leaves a representation $\gamma: N^+ \to C_\gamma$ in which each bit of $\hat{\beta}(j)$ is inserted between a pair of bits in $\alpha(|\beta(j)|)$ to give

$\gamma(1) = 1$ $\quad$ $\gamma(3) = 0\,\bar{1}\,1$ $\quad$ $\gamma(5) = 0\,\bar{0}\,0\,\bar{1}\,1$

$\gamma(2) = 0\,\bar{0}\,1$ $\quad$ $\gamma(4) = 0\,\bar{0}\,0\,\bar{0}\,1$ $\quad$ $\gamma(6) = 0\,\bar{1}\,0\,\bar{0}\,1\cdots$.

The overlined bits are $\hat{\beta}(j)$, and the remainder are $\alpha(|\beta(j)|)$. The length of $\gamma(j)$ is

$$|\gamma(j)| = |\rho_1(j)| - 1 = |\beta(j)| + |\hat{\beta}(j)| = 1 + 2\lfloor \log j \rfloor.$$

The set

$$C_\gamma = (0\{0,1\})^*1$$

is regular and like C_1 is accepted by a three-state halting acceptor. Like C_α and C_τ, C_γ is complete.

A permutation of the bits in $\gamma(j)$ gives a representation $\gamma'(j) = \alpha(|\beta(j)|)\hat{\beta}(j)$, which has the same length function and is easier for people to read

$\gamma'(1) = \bar{1}$ $\quad$ $\gamma'(3) = 0\,\overline{1\,1}$ $\quad$ $\gamma'(5) = 0\,0\,\overline{1\,0\,1}$

$\gamma'(2) = 0\,\overline{1\,0}$ $\quad$ $\gamma'(4) = 0\,0\,\overline{1\,0\,0}$ $\quad$ $\gamma'(6) = 0\,0\,\overline{1\,1\,0}\cdots$

since the last bit of $\alpha(|\beta(j)|)$ is always one, and it can do double duty as the first bit of the overlined sequence $\beta(j) = 1\hat{\beta}(j)$. However, the set

$$C_{\gamma'} = \bigcup_{k\ge 0} 0^k 1\{0,1\}^k$$

is not regular, and an acceptor for C_γ needs a counter.

Both γ and γ' are universal since Wyner's inequality gives the bounds

$$E_P(L_\gamma) = E_P(L_{\gamma'}) \le 1 + 2H(P) = U_\gamma(H(P))$$

$$R_\gamma(H) = \frac{U_\gamma(H)}{\max\{1,H\}} \le K_\gamma = 3.$$

Doubly Compound Representation δ

While γ is universal it is not asymptotically optimal since

$$|\gamma(j)| = 1 + 2\lfloor \log j \rfloor > 2 \log j - 1$$

$$E_P(L_\gamma) > 2E_P(\log) - 1$$

so by Theorem 4

$$\frac{E_P(L_\gamma)}{H(P)} \to 2$$

as $H(P) \to \infty$. To work better for large $H(P)$, the length of $\beta(j)$ can be represented by $\gamma(|\beta(j)|)$ rather than by $\alpha(|\beta(j)|)$, which is more compact for large j.

Let

$$\delta(j) = \gamma(|\beta(j)|)\hat{\beta}(j).$$

Then

$$\delta(1) = \gamma(1) = \overline{1} \qquad \delta(3) = \gamma(2)1 = 0\,0\,\overline{1\,1}$$

$$\delta(2) = \gamma(2)0 = 0\,0\,\overline{1\,0} \qquad \delta(4) = \gamma(3)0 = 0\,1\,\overline{1\,0\,0}$$

where the overlined symbols are $\beta(j)$, i.e., the last one in $\gamma(|\beta(j)|)$ followed by $\hat{\beta}(j)$.

A decoding algorithm for $\delta(j)$ first uses a decoder for $\gamma(|\beta(j)|)$ to find $|\beta(j)|$, and then prints one followed by the last $|\beta(j)| - 1$ symbols of $\delta(j)$ and halts. It needs a counter: the set C_δ is not regular, nor is any other set with the same length function L_δ since

$$\begin{aligned} L_\delta(j) &= |\delta(j)| = |\gamma(|\beta(j)|) + |\hat{\beta}(j)| \\ &= 1 + 2\lfloor \log |\beta(j)| \rfloor + \lfloor \log j \rfloor \\ &= 1 + \lfloor \log j \rfloor + 2\lfloor \log (1 + \lfloor \log j \rfloor) \rfloor \end{aligned}$$

increases by two units whenever j increases by one from

$$2^{2^k - 1} - 1 \text{ to } 2^{2^k - 1}$$

while the set of lengths of the members of any regular set is ultimately periodic. The Wyner inequality and the convexity $\cap$ of the logarithm prove asymptotic optimality

$$E_P(L_\delta) \le U_\delta(H) = 1 + H + 2 \log (1 + H)$$

$$R_\delta(H) \le K_\delta = 4$$

$$\lim_{H \to \infty} R_\delta(H) = \lim_{H \to \infty} \frac{U_\delta(H)}{H} = 1.$$

Penultimate Representation ω

The representation $\omega: N^+ \to C_\omega$ represents some early integers by

$$\omega(1) = 0 \qquad \omega(4) = \overline{1\,0}\,\overline{1\,0\,0}\,0 \qquad \omega(15) = \overline{1\,1}\,\overline{1\,1\,1}\,1\,0$$

$$\omega(2) = \overline{1\,0}\,0 \qquad \omega(7) = \overline{1\,0}\,\overline{1\,1\,1}\,0 \qquad \omega(16) = \overline{1\,0}\,\overline{1\,0\,0}\,\overline{1\,0\,0\,0\,0}\,0$$

$$\omega(3) = \overline{1\,1}\,0 \qquad \omega(8) = \overline{1\,1}\,\overline{1\,0\,0\,0}\,0 \qquad \omega(32) = \overline{1\,0}\,\overline{1\,0\,1}\,\overline{1\,0\,0\,0\,0\,0}\,0.$$

The rightmost overlined group is $\beta(j)$, except for $j = 1$. Each earlier group is the binary encoding of the length less one of the following group, and the process halts on the left with a group of length 2. To encode,

i) write zero
ii) start
iii) if $\lfloor \log j \rfloor = 0$, halt
iv) write $\beta(j)$ to the left of previous writing
v) $j \leftarrow \lfloor \log j \rfloor$
vi) return to start.

There is an equally simple left-to-right decoding algorithm.

Define $l^k(j)$ by the induction

$$l^1(j) = \lfloor \log j \rfloor \qquad l^{k+1}(j) = l^1(l^k(j)), \qquad k \in N^+.$$

Then

$$|\beta(j)| = l^1(j) + 1$$

and

$$\begin{aligned} |\omega(j)| &= \sum_{m=1}^{k} \beta(l^{k-m}(j)) + 1 \\ &= 1 + \sum_{m=1}^{k} (l^m(j) + 1) \end{aligned}$$

where the summation stops with that integer k such that $l^k(j) = 1$.

An examination of $|\omega(j)|$ shows that

$$|\omega(j)| \le 1 + \tfrac{5}{2}\lfloor \log j \rfloor$$

$$E_P(L_\omega) \le 1 + \tfrac{5}{2}H(P), \qquad K_\omega \le \tfrac{7}{2}.$$

Asymptotic optimality follows from Theorem 4 and the limits

$$\lim_{j \to \infty} \frac{l^{m+1}(j)}{l^m(j)} = 0, \qquad \text{so } \lim_{j \to \infty} \frac{|\omega(j)|}{l^1(j)} = 1.$$

The representation ω is not quite ultimate. For $k \ge 5$, deleting the k initial ones from each block and the terminal zero, and prefixing the result with $\gamma(k + 1)$ (or with $\delta(k + 1)$ or $\omega(k + 1)$, for larger k) works better, but only for j much larger than Eddington's estimate of the number of protons and electrons in the universe!

VI. Universal Codes

A universal code is a code that works for all sources in some class. To explore several such codes requires definition of the corresponding classes.

The class $\mathcal{A}$ of stationary countable finite-entropy sources without memory includes a source (M,P), iff:

i) $M = N^+ = \{1,2,3,\cdots\}$;
ii) the distribution $P: M \to [0,1)$ has $0 < H(P) < \infty$;
iii) the source output sequence of length n takes the value $\boldsymbol{m} = m(1),m(2),\cdots,m(n)$ with probability

$$P_n(\boldsymbol{m}) = \prod_{j=1}^{n} P(m_j).$$

The class $\mathcal{M} \subset \mathcal{A}$ of *monotonic* sources satisfies the additional constraint

iv) $1 > P(j) \geq P(j+1), j \in N^+$.

The subset $\mathcal{A}_k \subset \mathcal{A}$ of k-ary (or k-letter) sources has, in addition,

v) $P(j) > 0$, iff $1 \leq j \leq k$;

and the monotonic k-ary sources $\mathcal{M}_k$ are the set

vi) $\mathcal{M}_k = \mathcal{A}_k \cap \mathcal{M}$.

The source coding theorem in Section II shows that for each $P \in \mathcal{A}$ there is a sequence of sets C_n and of codes $\mu_n: M^n \to C_n$ such that the entropy performance measure

$$\sum_{\boldsymbol{m} \in M^n} \frac{P_n(\boldsymbol{m})|\mu_n(\boldsymbol{m})|}{H(P_n)} \tag{3}$$

approaches one as n increases. Theorem 5 shows that there is a single a.o. set C_ρ such that for each $P \in \mathcal{A}$ there is a sequence $\mu_n': M^n \to C_\rho$ that has the same limiting performance, where the codes depend on P (since μ_n maps M^n onto C_ρ in order of decreasing probability) but C_ρ does not. Other work on sequences of codes reviewed by Davisson [1] shows that there is a single sequence $\mu_n'': M^n \to C_n$ that has the same limiting performance for all $P \in \mathcal{A}$. Davisson calls this sequence a universal code for $\mathcal{A}$. We prefer to reserve that name for a single code rather than a sequence of codes.

Define a code to be *universal* for a class of sources with respect to a performance measure if there is a uniform bound to the measure for all sources in the class. Then a universal representation $\rho: N^+ \to C_\rho$ is a universal code for the monotonic class $\mathcal{M}$, with respect to the average codeword length performance measure

$$\sum_{j \in N^+} \frac{P(j)|\rho(j)|}{\max\{1, H(P)\}}$$

by the definition of a universal representation.

No single code in any one of the three sequences of codes described has an *entropy* performance measure (4) that is bounded uniformly on $\mathcal{A}$ or $\mathcal{M}$, however. For $0 < H(P) < \varepsilon$ the average codeword length is ≥ 1 by Theorem 2 i) so the entropy measure is $> 1/n\varepsilon$ for any uniquely decipherable μ_n, and $\to \infty$ for fixed n as $\varepsilon \to 0$. Thus the sequence μ_n'' may be described as an asymptotically optimal universal *sequence* of codes for $\mathcal{A}$ but not as an a.o. sequence of *universal* codes for $\mathcal{A}$. The coder will need to use different n for different customers if he is to guarantee a uniform level of entropy performance. He need not know P, but he must know (a lower bound >0 to) $H(P)$. We construct universal codes for $\mathcal{M}_2$, $\mathcal{A}_2$, $\mathcal{M}$, $\mathcal{A}$, and a.o. sequences of such codes. In Davisson's terminology, such an a.o. sequence could be called a uniformly universal code.

Coding Runs of Zeros

For the class $\mathcal{M}_2$, let $M = \{0,1\}$ and $P(0) = q$, $P(1) = p$. Then the monotonic constraint is just $p \in (0,\frac{1}{2}]$. Universal run-length codes for $\mathcal{M}_2$ with entropy performance bounded by ~ 1.6 for $p \in (0,\frac{1}{2}]$ are analyzed in [2].

Universal run-length coding treats the infinite source output sequence as a concatenation $\alpha(j_1)\alpha(j_2)\alpha(j_3)\cdots$ of unary encodings of a sequence $j_1, j_2, j_3, \cdots$ of integers (the lengths of runs of zeros followed by one) and decodes and re-encodes the integers into the concatenation $\rho(j_1)\rho(j_2)\rho(j_3)\cdots$ of their universal representations.

The probability of the integer j is

$$Q(j) = q^{j-1}p$$

decreasing in j, and the entropy per run is

$$H(Q) = \frac{H(P)}{p} \geq 2, \qquad p \in (0,\tfrac{1}{2}].$$

It follows that $\max\{1, H(Q)\} = H(Q)$ so that

$$R_\rho(H(Q)) = \frac{E_Q(L_\rho)}{\max\{1, H(Q)\}} = \frac{E_Q(L_\rho)}{H(Q)} \leq K_\rho$$

gives a uniform bound of K_ρ to the entropy measure (4). Since Q is known as a function of p, the ratio R can be computed for $p \in (0,\frac{1}{2}]$ for any particular universal ρ and gives much tighter bounds than K_ρ for the representations $\hat{\tau}$ used in [2] and γ, δ, and ω developed in Section V.

Coding Runs of Zeros and Ones

The infinite source output sequence $\boldsymbol{m} = m(1)m(2)m(3)\cdots$ is also a concatenation

$$\begin{aligned}\boldsymbol{m} &= m(1)^{j_1}\overline{m}(1)^{k_1}m(1)^{j_2}\overline{m}(1)^{k_2}\cdots \\ &= m(1)\alpha_{m(1)}(j_1)\alpha_{\overline{m}(1)}(k_1)\alpha_{m(1)}(j_2)\alpha_{\overline{m}(1)}(k_2)\cdots\end{aligned}$$

of $m(1)$ and a sequence of unary encodings, where $\overline{m}(1) = 1 - m(1)$ is the complement of $m(1)$ and

$$\alpha(j) = \alpha_0(j) = 0^{j-1}1 \qquad \alpha_1(j) = 1^{j-1}0 = \bar{\alpha}(j).$$

A universal run-length encoding for $\mathcal{A}_2$ decodes into the sequence $m(1), j_1, k_1, j_2, k_2, \cdots$, and reencodes into the concatenation $m(1)\rho(j_1)\rho(k_1)\rho(j_2)\rho(k_2)\cdots$. The distributions of the j and k are

$$Q_0(j) = q^{j-1}p \qquad Q_1(j) = p^{j-1}q$$

both decreasing functions and a pair (j,k) always contains one integer from each distribution, so the concatenation $\rho(j)\rho(k)$ has entropy performance measure

$$\frac{(E_{Q_0}(L_\rho) + E_{Q_1}(L_\rho))}{(H(Q_0) + H(Q_1))} \tag{4}$$

which is the same as the performance of the original scheme $\alpha(j) \to j \to \rho(j)$ at $p = q = \frac{1}{2}$ (when $Q_0 = Q_1$) and again in the limit $p \to 0$ (when $H(Q_0) \to \infty$ and $E_{Q_0}(L_\rho)/H(Q_0) \to 1$ for a.o. ρ) but is a little worse in between. Computing the ratio as a function of $p \in (0,1)$ gives a uniform bound ~ 1.6 for several choices of ρ.

This scheme has practical interest since it works well for a memoryless source and works even better for a Markov source whose state is the last output symbol and whose conditional probabilities $P_0(0), P_1(1)$ of staying in the same state are greater than the corresponding steady-state

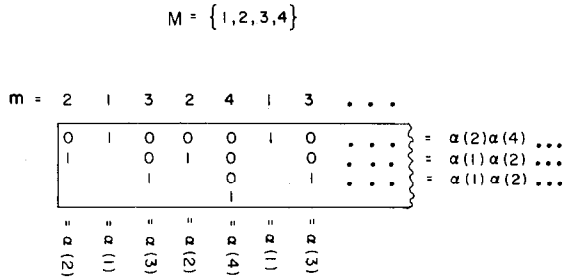

Fig. 1.

probabilities $P(0),P(1)$. It should, therefore, do well for both line drawings and large objects in facsimile encoding.

Universal Codes for $\mathscr{A}_k$

A source in $\mathscr{A}_k$, with $k > 2$, can be universally encoded by factoring it into $k-1$ independent sources in $\mathscr{A}_2$ and using a universal run-length encoding of their outputs. Let $M = \{1,2,\cdots,k\}$. Fig. 1 illustrates the procedure.

Let $\boldsymbol{m} = j_1,j_2,\cdots,j_n$ be the first n message. Represent each occurrence of the integer $j \in M$ in $\boldsymbol{m}$ by the unary codeword $\alpha(j) = 0^{j-1}1$. Write $\alpha(j_s)$ as the sth column of an array of k rows by n columns, the first symbol of each codeword occupying the first row. Then the entry in column s of row j is one, if $j_s = j$; is blank, if $j_s < j$; and is zero, if $j_s > j$; so the first row has no blanks and the kth row has no zeros.

The locations of the blanks in row j are determined by the locations of the ones in earlier rows. The values of the nonblank symbols in row j are independent of one another and of the values in earlier rows. Thus the nonblank symbols in each of the first $k-1$ rows are output sequences of $k-1$ independent sources in $\mathscr{A}_2$.

Encoding the nonblank symbols in each row by the run-length scheme $\alpha(j)\alpha(k) \to j$, $k \to \rho(j)\rho(k)$ described for $\mathscr{A}_2$, the ratio of the average codeword length to entropy still satisfies the uniform bound K_ρ and the tighter uniform bounds in [2]. An algorithm suggested by Gallager [11] sequences the transmission of the representations of the run-length integers and gives a universal code for $\mathscr{A}_k$.

For $1 \le j \le k-1$, place a marker j in row j, initially to the left of column 1. Start and choose the leftmost column containing any marker and the smallest (highest) marker in that column, say marker j. Move marker j to the right in row j, passing $r-1$ zeros, and halting on the first one to the right of the initial position of marker j in row j, and send the codeword $\rho(r)$. Then return to start. The received sequence always determines the source sequence up to the column from which the last marker moved, and determines at most $k-2$ message values lying to the right of that point.

Universal Codes for $\mathscr{M}$ and $\mathscr{A}$

When M is infinite the algorithm used for $\mathscr{A}_k$ will never send the second pair of runs for any symbol, and a less symmetric algorithm is needed. To start the new marker-moving algorithm, start a cycle. Start a cycle by moving marker 1 to the right side of its first pair of runs and send the encoding of that pair of runs. Return to start a cycle when every column to the left of marker 1 has appeared as the end of a run, so that decoding is complete up to the column occupied by marker 1. If a new cycle is not started, the next move is made by the smallest (highest) marker that lies to the left of marker 1.

The new algorithm works for $\mathscr{M}$ but not for $\mathscr{A}$, since $P(1) = 0$ is possible for $\mathscr{A}$, in which case marker 1 never moves. To encode $\mathscr{A}$, add labels to the rows of the array in Fig. 1 and send the labels to the receiver. The label l_1 on row 1 is that message $l_1 \in N^+$ which first completes a pair

$$\{l_1\}^j(N^+ - \{l_1\})^k \qquad \text{or } (N - \{l_1\})\{l_1\}^k$$

of runs, and the label on row $j+1$ is the message which first completes such a pair in that subsequence of the message sequence that remains when occurrences of $l_1,l_2,\cdots,l_j$ have been deleted. Then $\rho(l_j)$ is sent as a prefix to the code $m(s)\rho(j)\rho(k)$ for the first pair of runs that occur in row j.

Optimal Sequences of Universal Codes

Universal codes for $\mathscr{M}_2$ and $\mathscr{A}_2$ (and thus $\mathscr{A}_k$ and $\mathscr{A}$) that have better entropy performance than the bound of ~ 1.6 in [2] are constructed by using an a.o. representation ρ and an intermediate mapping between the two steps $\alpha(j) \to j$ and $j \to \rho(j)$ or $\alpha(j)\bar{\alpha}(k) \to j$, $k \to \rho(j)\rho(k)$ of a run-length encoding.

Let $h: N^+ \times N^+ \to N^+$ be the usual one-to-one mapping of pairs of integers onto integers

$$h(j_1,j_2) = \frac{(j_1 + j_2 - 1)(j_1 + j_2 - 2)}{2} + j_1.$$

For $\mathscr{M}_2$ the probability of a successive pair j_1,j_2 of runs is

$$Q_0(j_1)Q_0(j_2) = q^{j_1+j_2}\left(\frac{p}{q}\right)^2$$

a function of $j_1 + j_2$ alone, so the distribution

$$Q_2(h) = q^j\left(\frac{p}{q}\right)^2 \qquad \frac{(j-1)(j-2)}{2} + 1 \le h \le \frac{j(j-1)}{2}$$

of h is a nonincreasing function of h. Since h is one-to-one,

$$H(Q_2) = 2H(Q) \ge 4, \qquad p \in (0,\tfrac{1}{2}].$$

Therefore, using the encoding $\alpha(j_1)\alpha(j_2) \to j_1$, $j_2 \to h(j_1,j_2) \to \rho(h(j_1,j_2))$ gives an entropy performance of

$$\frac{E_{Q_2}(L_\rho)}{H(Q_2)} = R_\rho(H(Q_2)) = R_\rho(2H(Q))$$

and an n-fold iteration of the h-mapping stage in the coding process maps each 2^n-tuple $j_1,j_2,\cdots,j_{2^n}$ into a single integer with decreasing distribution Q_n and then into a single codeword in C_ρ with entropy performance

$$\frac{E_Q(L_\rho)}{H(Q_n)} = R_\rho(H(Q_n)) = R_\rho(2^nH(Q))$$

which approaches one uniformly on $\mathcal{M}_2$ as n increases for any a.o. ρ, since $H(Q) \geq 2$.

The same procedure is used for $\mathcal{A}_2$, but the sequence is

$$\alpha(j_1)\bar{\alpha}(k_1)\alpha(j_2)\bar{\alpha}(k_2)\cdots \rightarrow j_1,k_1,j_2,k_2\cdots$$
$$\rightarrow h(j_1,j_2),h(k_1,k_2)\cdots \rightarrow \rho(h(j_1,j_2))\rho(h(k_1,k_2))\cdots$$

so that the 2^n integers mapped by h^n are always drawn from the same distribution (Q_0 or Q_1). Using this encoding on the rows of the array in Fig. 1, sending representations of 2^n runs of zeros and 2^n runs of ones from each row, and using the appropriate marker-moving algorithm gives an a.o. sequence of universal codes for $\mathcal{A}_k$, $\mathcal{M}$, and $\mathcal{A}$.

References

[1] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[2] P. Elias, "Predictive coding," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, esp. pp. 30–33, Mar. 1955.
[3] ——, "The efficient construction of an unbiased random sequence," *Ann. Math. Statist.*, vol. 43, pp. 865–870, 1972.
[4] ——, "Efficient storage and retrieval by content and address of static files," *J. Ass. Comput. Mach.*, vol. 21, pp. 246–260, 1974.
[5] ——, "Minimum times and memories needed to compute the values of a function," *J. Comput. Syst. Sci.*, Oct. 1974.
[6] R. A. Flower, "Computer updating of a data structure," Research Lab. Electron., M.I.T., Cambridge, Mass., Quart. Progress Rep. 110, pp. 147–154., July 15, 1973.
[7] R. W. Floyd, "Permuting information in idealized two-level storage," in *Complexity of Computer Computations*, Miller, Thatcher and Bohlinger, Eds. New York: Plenum, 1972, pp. 105–109.
[8] S. W. Golomb, "Run-length encodings," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-12, pp. 399–401, July 1966.
[9] ——, "A class of probability distributions on the integers," *J. Number Theory*, vol. 2, pp. 189–192, 1970.
[10] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[11] ——, personal communication.
[12] D. E. Knuth, *The Art of Computer Programming*, vol. 3. Reading, Mass.: Addison-Wesley, 1973, esp. pp. 181–218.
[13] A. Kohavi, *Switching and Finite Automata Theory*. New York: McGraw-Hill, 1970, esp. ch. 16.
[14] M. Minsky and S. Papert, *Perceptrons*. Cambridge, Mass.: M.I.T. Press, 1969, esp. pp. 215–226.
[15] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, Ill.: University of Illinois Press, 1949, esp. p. 64.
[16] T. Welch, "Bounds on information retrieval efficiency in static file structures," M.I.T., Cambridge, Mass., MAC TR-88, Project MAC, 1971.
[17] A. D. Wyner, "An upper bound on the entropy series," *Inform. Contr.*, vol. 20, pp. 176–181, 1972.
[18] R. Karp, personal communication.

The Algebraic Decoding of Goppa Codes

N. J. PATTERSON

***Abstract*—An interesting class of linear error-correcting codes has been found by Goppa [3], [4]. This paper presents algebraic decoding algorithms for the Goppa codes. These algorithms are only a little more complex than Berlekamp's well-known algorithm for BCH codes and, in fact, make essential use of his procedure. Hence the cost of decoding a Goppa code is similar to the cost of decoding a BCH code of comparable block length.**

I. Introduction

LET K be the finite field $GF(q^m)$. Let J be the finite field $GF(q)$. Let $g(x)$ be a polynomial of degree $n \geq 1$ with coefficients in K, and let L be a subset of K with the property that no element of L is a root of g. We define a *Goppa code* $\mathcal{G}$ with *Goppa polynomial* g and symbol field J as follows. It is convenient to index the coordinates of $\mathcal{G}$ by L. Then $\boldsymbol{C}$ is a codeword of $\mathcal{G}$, if and only if

$$\sum_{\gamma \in L} \frac{C_\gamma}{x - \gamma} \equiv 0 \bmod g(x). \tag{1}$$

Manuscript received January 22, 1974; revised October 20, 1974.
The author is with the Government Communications Headquarters, Cheltenham, England.

Let $\boldsymbol{C}$ be a codeword and $\boldsymbol{R}$ the received word, so that the error vector $\boldsymbol{E}$ is given by

$$\boldsymbol{R} = \boldsymbol{C} + \boldsymbol{E}$$

so that

$$\sum_{\gamma \in L} \frac{R_\gamma}{x - \gamma} \equiv \sum_{\gamma \in L} \left(\frac{C_\gamma}{x - \gamma} + \frac{E_\gamma}{x - \gamma} \right) \bmod g(x)$$
$$\equiv \sum_{\gamma \in L} \frac{E_\gamma}{x - \gamma} \bmod g(x).$$

It is natural then to define the syndrome $S(x)$ as the polynomial of degree less than n such that

$$S(x) \equiv \sum_{\gamma \in L} \frac{R_\gamma}{x - \gamma} \bmod g(x). \tag{2}$$

We define

$$\sigma(x) = \prod_{\substack{\gamma \in L \\ E_\gamma \neq 0}} (x - \gamma)$$

(thus deg σ = number of errors), and we define $\eta(x)$ of degree less than n by

$$\eta(x) \equiv \sigma(x)S(x) \bmod g(x). \tag{3}$$