Universal Codes for Finite Sequences of Integers Drawn from a Monotone Distribution

Dean P. Foster, Robert A. Stine & Abraham J. Wyner

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6302

November 29, 2001

Abstract

We offer two noiseless codes for blocks of integers $X^n = (X_1, \ldots, X_n)$. We provide explicit bounds on the relative redundancy that are valid for any distribution F in the class of memoryless sources with a possibly infinite alphabet whose marginal distribution is monotone. Specifically we show that the expected code length $L(X^n)$ of our first universal code is dominated by a linear function of the entropy of X^n. Further, we present a second universal code that is efficient in that its length is bounded by $n\,H_F + o(n\,H_F)$, where H_F is the entropy of F which is allowed to vary with n. Since these bounds hold for any n and any monotone F we are able to show that our codes are strongly minimax with respect to relative redundancy (as defined by Elias).

Key Phrases: Universal noiseless coding of integers, Elias codes, Wyner's inequality, relative redundancy, strongly minimax.

1 Introduction

Consider the problem of lossless compression of a finite collection of n positive integers $X^n = (X_1, \ldots, X_n)$ into a prefix code of shortest expected length. The $X_i \geq 1$ are independent, integer-valued random variables that share a common, unknown probability distribution F which is assumed monotone, meaning that $P_F(X = i) = p_i \geq p_{i+1}$, $i = 1, 2, \ldots$, and $P_F(X > 0) = 1$. We will denote the set of all monotone distributions with finite entropy as $\mathcal{M}$. For $F \in \mathcal{M}$, all $X_i = 1$ when the entropy

$$H_F = -\sum_i p_i \log p_i$$

is 0.

We wish to encode X^n as efficiently as possible, regardless of the entropy H_F. Given F, one can construct an arithmetic coder whose code length $L_F(X^n)$ is on average within one bit of the minimum attainable length,

$$n\,H_F \leq E\,L_F(X^n) \leq 1 + n\,H_F\ .$$

If F is unknown, we seek a universal code whose loss relative to this utopian performance is bounded. The usual way (see definition 3 of [1]) to evaluate such codes is to show that the average redundancy goes to zero, namely to show that

$$\lim_{n\to\infty} \frac{E\,L(X^n) - n\,H_F}{n} = 0 \tag{1}$$

for every distribution F in a class $\mathcal{C}$. If the members of $\mathcal{C}$ satisfy (1) for some code, then $\mathcal{C}$ is said to be *weakly* universally encodable in the sense of Davisson. Györfi, Páli, and van der Meulen [5] show that, without restrictions on the class C of memoryless sources, such a code does not exist. They prove that universal codes are impossible even in this weak sense for infinite alphabet memoryless sources, even those restricted to finite entropy. However, under the additional assumption of monotonicity Elias [2] shows that universal codes do exist. Györfi *et al* [4] extend this result to an even larger class of memoryless distributions with finite entropy. Such results, however, do not imply how well the code performs for finite values of n. Unless we know the distribution F, we do not know if the particular n we are considering is large enough for the limit to be useful. What is hoped for is that the class $\mathcal{M}$ is *strongly* universally encodable in the sense of Davisson:

$$\lim_{n\to\infty} \sup_{F\in\mathcal{M}} \frac{E\,L(X^n) - n\,H_F}{n} = 0 \; . \tag{2}$$

Strong universality implies a uniform bound on the redundancy. The following lemma (proved in the appendix) shows that, for monotonic distributions, this is impossible:

Lemma 1 *For any prefix code with length L,*

$$(\forall n) \quad \sup_{F\in\mathcal{M}} \frac{E\,L(X^n) - n\,H_F}{n} = \infty.$$

Because one cannot uniformly bound the per-symbol redundancy, we consider a different measure of performance for which one *can* obtain a uniform bound, even for distributions on infinite alphabets. Some notation is needed to describe our results. Elias [2] measures the performance of a prefix code as the ratio of its expected code length to the minimum attainable length,

$$R_n(L,F) = \frac{E\,L(X^n)}{\max(1, n\,H_F)} \; . \tag{3}$$

Since R_n quantifies the performance of a code in relative terms with respect to the entropy we will sometimes call R_n the *relative redundancy.* A code with length function $L(X^n)$ is *universal* in the spirit of Elias [2, page 201] for some class of distributions $\mathcal{F}$ if the relative redundancy is bounded,

$$L \text{ is universal for } \mathcal{F} \iff \forall n \sup_{F\in\mathcal{F}} R_n(L,F) \le c_n < c < \infty \; . \tag{4}$$

A universal code guarantees a level of performance, but that performance may not be very good. To address the performance of a code, we say that the code with length $L(X^n)$ is *efficient* if its expected code length grows only slightly faster than the best possible,

$$L \text{ is efficient } \iff E\,L(X^n) \le nH_F + g(n\,H_F) \; , \tag{5}$$

where the remainder g is a given sublinear function not depending on n or F,

$$\lim_{x\to\infty} \frac{g(x)}{x} = 0 \; .$$

Combining these aspects, an efficient universal code guarantees not only uniform performance but also short codes. In Theorem 3, we prove the existence of an efficient universal code for the class $\mathcal{M}$.

The ratio criterion R_n is more restrictive than the usual redundancy for sources of low entropy. If $H_F > 1$, the ratio criterion R_n is weaker than the usual redundancy. When the entropy is small, however, the ratio criterion is more strict. As an example, consider a Lempel-Ziv coder; its relative redundancy converges to zero for a wide class of sources, including finite-alphabet, memoryless sources. Suppose the data X^n is generated i.i.d. from distribution F on $(1, 2)$ with $p_1 = 1 - 1/n$ and $p_2 = 1/n$. When encoding X^n, the LZ coder will create a collection of code words for strings of zeros, with about $\sqrt{n}$ code words each representing a longer string of zeros than the previous. So, to code the entire sequence, the LZ coder will take about $(\frac{1}{2} \log n) \sqrt{n}$ bits. The LZ coder does well in the usual sense of redundancy in this case since $n\, H_F \approx \log n$ and its code length is $o(n)$. In terms of the relative criterion R_n, however, the LZ coder does poorly (Kosaraju and Manzini [6] discuss the fact that it performs poorly for sources of lower entropy). A good code would use about $\log n$ bits whereas the LZ code is longer by a factor of $\sqrt{n}$. If sources from a wide range of entropies are considered, simply achieving a bounded ratio can be a challenge.

The rest of this paper develops as follows. For coding a monotone source, we show in Section 2 that that it is possible to encode X^n so that the length of the prefix code is bounded in expectation by a linear function of the entropy of the source,

$$(\forall n > 0,\, \forall F \in \mathcal{M}) \qquad E\, L(X^n) \leq c_0 + c_1\, n\, H_F \ , \tag{6}$$

where the constants c_0 and $c_1 > 1$ are invariant of n and F. The lower bound for c_1 is a consequence of Lemma 1. From (6) it follows that such a code is universal since

$$R_n(L, F) \leq c_0 + c_1 \ .$$

This code is a simple modification of the concatenation of scalar universal codes. It produces a universal code with $c_0 = 3$ and $c_1 = \frac{9}{2}$. Surprisingly, the only modification is the optional compression of the leading bits of each universal code so that the code is competitive when the source entropy is near 0. Following in Section 3, we extend this approach significantly to produce an efficient universal code for monotone sources with arbitrary entropy. Our main result (Theorem 2) finds an efficient code whose expected length satisfies the following bound:

$$(\forall n > 0, \forall F \in \mathcal{M}) \qquad EL(X^n) \leq n\, H_F + \left(20 + 10 \frac{n\, H_F \log\log(n\, H_F)}{\log(n\, H_F)}\right) \ .$$

A slightly modified version of the efficient code is strongly minimax in the sense that (Corollary 2)

$$\lim_{n\to\infty} \sup_{F\in\mathcal{M}} \frac{EL(X^n) - (nH_F + 1)}{nH_F} = 0 \ .$$

2 A Universal Code for Blocks

Universal codes for integers are well-known. A particularly good example of these for sources of large entropy is the "penultimate" code, denoted by ω in Elias [2]. (See [4] and [5] for extensions of Elias.) This procedure encodes the positive integer $x \geq 1$ with an idealized length

$$L_\omega(x) = c_\omega + \log x + \log\log x + \cdots \ ,$$

where the log (base 2) terms are accumulated while positive and $c_\omega \approx 2.865$ (see [8]). The penultimate code is universal in the sense of (6) because its expected length is bounded by a linear function of the entropy: for $X \sim F$ [2],

$$(\forall F \in \mathcal{M}) \qquad E\,L_\omega(X) \leq 1 + \tfrac{5}{2}\,H_F \ . \tag{7}$$

The penultimate code is also asymptotically optimal as defined in [2]. When coding one integer from a monotone distribution of large entropy, the length of the penultimate code approaches the entropy in the sense of the limit

$$\lim_{H_F\to\infty} \frac{E\,L_\omega(X)}{H_F} = 1 \ . \tag{8}$$

The concatenation of scalar universal codes does not automatically produce a universal code for a block of several integers. The weakness occurs when coding sources of small entropy. In the extreme case when $H_F = 0$ and all $X_i = 1$, the accumulation of symbols for each X_i implies that the total length is proportional to n. For example, the penultimate code for $X_i = 1$ is a single 0 bit, so the code for X^n is a sequence of n consecutive zeros. For coding such sequences of low entropy, Elias [2] proposes a recursive procedure that is optimal in the limit as n and $n\,H_F$ increase. His procedure, however, requires that one assume a positive lower bound on the source entropy. The code described below avoids this condition. It is both universal and asymptotically optimal as the entropy grows. The coding procedure makes use of two codes for integers,

Table 1: Each member of X^n is represented as a column of bits in the ragged array U of unary codes for the input sequence X^n.

X^n	5	1	3	1	2
u_1	0	1	0	1	0
	0		0		1
	0		1		
	0				
	1				

the penultimate code and the unary code. The unary code for the integer $x \geq 1$ is a sequence of $x-1$ zeros followed by a single 1:

$$U(x) = \begin{cases} 1\ , & x = 1\ , \\ \underbrace{0 \cdots 0}_{x-1}\ 1\ , & \text{otherwise.} \end{cases}$$

To construct a universal code for X^n, we begin by forming a ragged array U of n columns with a varying number of rows. The columns of U are the unary codes for X_i stacked side-by-side, as in Table 1 (which mimics that of Elias [2]). As shown in the table, the first row of U holds the leading bits of the sequence of unary codes, which we denote $U_1 = (U_{11}, \ldots, U_{1n})$ with U_{ri} denoting the r^{th} element in the unary code for X_i. The sum

$$N_1 = \sum_i U_{1i}$$

of the elements in this first row counts the number of $X_i = 1$. When subtracted from n,

$$Y = n - N_1 \tag{9}$$

counts the number of "large" $X_i > 1$. Because of the monotonicity of F, if the source entropy is small, then Y will also be small.

We obtain a universal code from this array by optionally compressing the leading bits U_1 from the first row of U, with the remaining bits encoded as a sequence of universal codes. It was surprising to us that such a simple adjustment produces a code that bounds the ratio (6) for a monotone source of arbitrary entropy. To be more specific about the code, we encode the leading bits U_1 in two parts. First, we encode the count of large values as $\omega(Y)$. The positions of the zeros in U_1 identify those $X_i > 1$ requiring further encoding. We represent these locations using $1 + \lfloor \log \binom{n}{y} \rfloor$ bits and encode the remaining bits as the concatenation of the universal codes $\omega(X_i - 1)$ for those $X_i > 1$. The length obtained in this fashion is then compared to the length of a direct concatenation of universal codes, with the shorter format adopted. A leading bit denotes the adopted format. The total length of the code for X^n is thus

$$L(X^n) = 1 + \min(1 + \log \binom{n}{Y} + L_\omega(Y) + \sum_{i:X_i>1} L_\omega(X_i - 1), \sum_{i=1}^{n} L_\omega(X_i)) \; . \tag{10}$$

The efficient universal code described in the following section extends this scheme by varying the number of rows of U that are encoded in this fashion.

Our proof that $E\,L(X^n)$ satisfies (6) concentrates on the low entropy setting. Cases of relatively high entropy can be handled easily using the results of Elias [2]. We state our claim along with the qualifying assumptions as

Theorem 1. *Let $X^n = (X_1, \ldots, X_n)$ denote a sequence of iid, integer-valued random variables $X_i \geq 1$ following a monotone probability distribution F having entropy H_F. Then the expected length of the proposed coding procedure is linearly bounded by the entropy with $c_0 = 3$ and $c_1 = \frac{9}{2}$. In other words,*

$$(\forall n > 0, \forall F \in \mathcal{M}) \qquad E\; L(X^n) \leq 3 + \tfrac{9}{2}\, n\, H_F \; . \tag{11}$$

Proof. First consider cases with $n = 1$, 2, or 3. There is little value in such cases for considering the compression of U_1, so encode the data using a universal code for each X_i. From (7), it follows that for $n \leq 3$ that

$$E\; L(X^n) \leq n\left(1 + \tfrac{5}{2} H_F\right) \leq 3 + \tfrac{5}{2}\, n\, H_F \; .$$

Now consider larger blocks with $n \geq 4$. We split these into two categories, those with "high" entropy $H_F \geq 1/2$ and the rest with "low" entropy. For $H_F \geq 1/2$, we can ignore the option to compress U_1 and bound the length as

$$E\; L(X^n) \leq 1 + n\, E\, L(X_1) \leq 1 + n\,(1 + \tfrac{5}{2} H_F) \leq 1 + \tfrac{9}{2}\, n\, H_F \; ,$$

so that (11) holds.

For cases with "low" entropy $H_F < 1/2$, the proof of (11) works by decomposing the total entropy $H(X^n)$ into terms that can be matched to summands in the expression (10) for the code length. For this, we require two lemmas that are proved in the appendix. For the rest of this section we overload our notation for the entropy and let $H(X)$ denote the entropy associated with the random variable X. The first lemma shows that a term in the entropy of a Bernoulli sequence dominates the mean of the associated binomial count.

Lemma 2 *Let $Y \sim Bi(n,p)$ denote a binomial random variable with parameters $n \geq 4$ and $p < 1/3$. Then*

$$E\,Y \leq E\,\log \binom{n}{Y} .$$

The second lemma establishes a property analogous to (7) for a binomial random variable that does not satisfy the monotonicity condition.

Lemma 3 *If $Y \sim Bi(n,p)$ with $p < 1/3$ and $n \geq 4$, then the expected length of the penultimate code for Y is dominated by a linear function of its entropy,*

$$EL_\omega(Y) \leq 2 + 2\,H(Y) . \tag{12}$$

For Y defined as in (9), the assumption that $H_F < \frac{1}{2}$ implies both Lemma 2 and Lemma 3 hold. To show that the conditions of the lemmas are satisfied, we decompose $H(X^n)$ into the entropy of a Bernoulli sequence and remaining terms. We first define the indicators $b^n = (b_1, \ldots, b_n)$ where

$$b_i = \begin{cases} 0, & X_i = 1 \\ 1, & X_i > 1 \end{cases} .$$

Since $b^n = 1 - U_1$, we note that $Y = \sum_i b_i$. The b_i are independent Bernoulli trials so that $Y \sim \mathrm{Bi}(n, P_1)$ with

$$P_1 = P(X > 1)$$

for $X \sim F$. The assumed monotonicity of F implies that its entropy is bounded below by

$$H_F = E\,\log \frac{1}{p_i} \geq \log \frac{1}{\max p_i} = \log \frac{1}{p_1} ,$$

where $p_i = P(X = i)$. Consequently, because $H_F < 1/2$, p_1 is at least

$$p_1 \geq 2^{-H_F} > \frac{1}{\sqrt{2}} > .7 \ , \tag{13}$$

and $P_1 < 0.3$ so that Lemma 2 and Lemma 3 hold.

We now match and compare the terms of the entropy (14) to those in the expected code length (10) when the leading bits are compressed. The joint probability for X^n can be written as

$$P(X^n) = P(X^n, b^n, Y) = P(X^n | b^n, Y) P(b^n | Y) P(Y) \ ,$$

so that the total entropy is

$$n\, H_F = H(X^n) = E Y\ H(X | X > 1) + E\ \log \binom{n}{Y} + H(Y) \ , \tag{14}$$

where $X \sim F$, and $H(X|X > 1)$ denotes the entropy of the conditional distribution of X given $X > 1$. Using first Lemma 2 and then Lemma 3, we can bound the expected length of the compressed version of the proposed code by $2 + \frac{5}{2} n\, H_F$,

$$\begin{aligned}
2 + \tfrac{5}{2} n\, H_F &= \tfrac{5}{2} E Y\ H(X|X > 1) + \tfrac{5}{2} E \log \binom{n}{Y} + (2 + \tfrac{5}{2} H(Y)) \\
&\geq \tfrac{5}{2} E Y\ H(X|X > 1) + E Y + \tfrac{3}{2} E \log \binom{n}{Y} + (2 + \tfrac{5}{2} H(Y)) \\
&\geq E Y (1 + \tfrac{5}{2} H(X|X > 1)) + E\ \log \binom{n}{Y} + E\, L_\omega(Y) \\
&\geq E Y\ L_\omega(X|X > 1) + E\ \log \binom{n}{Y} + E\, L_\omega(Y) \ .
\end{aligned}$$

With one bit added to allow for the option of compressing U_1, we see that the expected length satisfies (11). $\square$

It is easy to see that this coding procedure is asymptotically optimal, coding efficiently for sources of large entropy. The argument exploits the asymptotic optimality of the underlying penultimate code (8). In particular, holding n fixed and letting $H_F \to \infty$,

$$\begin{aligned}
\lim_{H_F \to \infty} \frac{E\, L(X^n)}{n H_F} &\leq \lim_{H_F \to \infty} \frac{1 + n\, E L_\omega(X_1)}{n\, H_F} \\
&= \lim_{H_F \to \infty} \frac{E\, L_\omega(X_1)}{H_F} = 1 \ .
\end{aligned}$$

3 An Efficient Universal Code

One can improve upon the previous universal code in the presence of greater entropy. That code handles the case when $n\,H_F$ is small quite well. It leaves room for improvement, however, as the entropy of the source grows. With more entropy present it becomes useful to not only compress the leading bits of the unary codes, but to compress subsequent "layers" as well. For example, suppose that $P(X_i = 1) = P(X_i = 2) = \frac{1}{2}$ and $n = 100$ so that the entropy lower bound for a code is $n\,H_F = 100$ bits. When represented as the simple concatenation of penultimate codes, the expected code length is 200 bits. Were it coded instead using the previous universal code with the one level of compression, the expected length falls to 157 bits. Two levels of compression, as shown next, reduce the expected length down to 108 bits. We show in this section that a code that adaptively selects the degree of compression is both efficient as well as universal. We summarize this result as

Theorem 2. *There exists a uniquely decodable prefix code for X^n whose length function $L(X^n)$ satisfies*

$$(\forall n > 0, \forall F \in \mathcal{M}) \qquad EL(X^n) \leq 20 + nH_F\left(1 + 10\frac{\log\log(n\,H_F)}{\log(n\,H_F)}\right)\ ,$$

where we take $\log(\log(x))/\log(x)$ to be zero for $x \leq 2$.

Note: The pair of constants (20,10) in this theorem can be improved to (30,5) but we do not provide a proof of this. The proof does show that for any positive ϵ, a pair can be found of the form $(c_0(\epsilon), 2+\epsilon)$.

Before proceeding to the proof, it is instructive to consider the application of a universal source coding algorithm (such as LZ) after the application of the Elias penultimate code, for example. The application of the Elias code results in a binary sequence for X^n with entropy nH_F. Application of LZ results in a code with an expected length of nH_F bits and a redundancy that is $o(n)$. This redundancy rate is not adequate. No improvements can be made in this approach; while the stationary encoding of X^n is still ergodic, it is not i.i.d. (or even finite memory). Thus there can be no further specification of the rate at which LZ (any version) will converge to H_F. When H_F may be arbitrarily small, the redundancy dominates and we must seek a different approach.

Proof (Theorem 2). Let m denote a fixed constant. As in the derivations of the previous section, our proposed code separates the coding task into two parts: "small" values for which $X_i \le m$ and "large" values for which $X_i > m$. Expanding the notation of the previous section, we let

$$P_m = P(X_i > m)$$

denote the probability of a "large value". This decomposition follows the division of the entropy as

$$n\,H_F = H(X^n) = H(X^n \wedge (m+1)) + n\,P_m\,H(X_1|X_1 > m)\;. \tag{15}$$

The collection of n "small" values $X^n \wedge (m+1)$ is handled using a multinomial coder. The penultimate code is competitively optimal, for m sufficiently large, when applied to the subset of X_i which are "large". Values coded as $m+1$ in the multinomial code identify the positions of the "large" X_i.

Here are the details. Suppose, first, that a fixed integer m is chosen (independently of X^n) and, along with n, is known to both the encoder and the decoder. (For the universal code of Section 2, $m = 0$ or 1 as indicated by a leading flag bit.) Define the counts

$$N_k = \sum_{i=1}^{n} 1\{X_i = k\}\;, \quad S_m = \sum_{k=1}^{m} N_k\;, \quad \text{and } B_m = N - S_m.$$

Thus, N_k is the number of times that X_i equals k and S_m is the number of times that $X_i \le m$, which is the size of the "small" set. Our encoder first specifies the multinomial counts N_k for $k = 2,\ldots,m$ and B_m. (For notational simplicity, let $m \ge 2$.) Given these counts, $X^n \wedge (m+1)$ is uniformly distributed,

$$P(X^n \wedge (m+1) = x^n | N_2, \ldots, N_m, B_m) = \binom{n}{N_1\;N_2\;\cdots N_m\;B_m}^{-1}.$$

Thus we can encode $X^n \wedge (m+1)$ using on average at most

$$\begin{aligned} 1 + E\,\log\binom{n}{N_1\;N_2\;\cdots N_m\;B_m} &\le 1 + H(X^n \wedge (m+1)|N_1,\ldots,N_m,B_m) \\ &\le 1 + H(X^n \wedge (m+1)) \text{ bits.} \end{aligned}$$

To complete the multinomial portion of the code, we need to represent N_k and B_m; a crude upper bound gives $L(N_k) \le 2 + 2\log N_k$. Taking this approach and using Jensen's inequality and the monotonicity of the source distribution F, the number of bits to code the marginal counts $N_2, \ldots, N_m$ and B_m is bounded by

$$\sum_{k=2}^{m} E\,L(N_k) + E\,L(B_m) \;\le\; 2(m + \sum_{k=2}^{m} E\log N_k + E\,\log B_m)$$

$$
\begin{aligned}
&\le 2(m + \sum_{k=2}^{m} \log EN_k + \log EB_m) \\
&\le 2(m + \sum_{k=1}^{m} \log(n\,P_k)) \\
&\le 2m(1 + \log(nP_1)) \ .
\end{aligned}
$$

Thus, if $L(X^n \wedge (m+1))$ is the total number of bits needed to encode the truncated sequence $X^n \wedge (m+1)$, we have shown that

$$E\,L(X^n \wedge (m+1)) \le 1 + H(X^n \wedge (m+1)) + 2m(1 + \log(nP_1)) \ .$$

Note that we omit a code for N_1 because it can be recovered from $N_1 = n - \sum_2^m N_k - B_m$.

The encoder now faces the task of specifying those $X_i > m$. Recall that the position of these values are denoted by those terms in the multinomial code with value $m+1$. Let $Z_k = X_{i_k}$, $k = 1, \ldots, B_m$ denote this subsequence where $X_i > m$. We encode Z_k using a code whose expected code length $L(Z_1)$ satisfies [2]

$$E\,L(Z_1) \le 1 + H(Z_1) + \log(1 + H(Z_1)) \ .$$

Now $nP_m = E\,B_m$ is the expected number of X_i to be coded using the Elias code. Thus, the *total* number of bits for parts one and two is bounded above by

$$
\begin{aligned}
E\,L(X^n) \ \le \ & 1 + H(X^n \wedge (m+1)) + 2m(1 + \log(nP_1)) \\
& + n\,P_m\,[1 + \ H(X_1|X_1 > m) + \log(1 + H(X_1|X_1 > m))\,] \ .
\end{aligned}
$$

The decomposition (15) implies that

$$E\,L(X^n) \le 1 + H(X^n) + 2m(1 + \log(nP_1)) + n\,P_m(1 + \log(1 + H(X_1|X_1 > m)) \ .$$

We now consider the asymptotic behavior of P_m. First, let $h(p)$ for $0 \le p \le 1$ denote the Boolean entropy function. It is easy to prove the following decomposition of the entropy which we state as

Lemma 4 *The entropy $H(X)$ of a random variable X may be partitioned as*

$$H(X) = P_m H(X|X > m) + (1 - P_m) H(X|X \le m) + h(P_m) \ .$$

The monotonicity condition on the probabilities p_i implies Wyner's inequality [10]:

Lemma 5 *If $X \sim F$, then*

$$E \log(X) \leq H_F \ .$$

A direct application of Markov's inequality to Wyner's inequality proves

Lemma 6 (Wyner-Markov) *If $X \sim F$ and m is any positive integer, then*

$$P_m = P\{X \geq m+1\} \leq \frac{H_F}{\log(m+1)} \ .$$

The importance of Wyner-Markov (Lemma 6) is that it provides a simple bound on the tail probability P_m in terms of the entropy.

To bound the remaining code length, we observe that Lemma 4 implies

$$H(X_1 | X_1 > m) \leq \frac{H_F}{P_m} \ .$$

It follows that

$$\begin{aligned} n\, P_m \log\left(1 + H(X_1|X_1 > m)\right) &\leq n\, P_m \log(1 + H_F/P_m) \\ &\leq \frac{n\, H_F}{\log(m+1)} \log(1 + \log(m+1)) \ , \end{aligned}$$

where the second inequality holds because the function $f(x) = x \log(1 + c/x)$ is monotone increasing in x for $0 < x \leq 1$.

Collecting together the terms and using Wyner-Markov again, we have shown the bound

$$E\, L(X^n) \leq 1 + n\, H_F + 2m(1 + \log n\, H_F) + \frac{n\, H_F}{\log(m+1)} \left(1 + \log(1 + \log(m+1))\right) \ . \quad (16)$$

Let $m = \max(\sqrt{nH_F} - 1, 1)$ and let $x = nH_F$. If $x \leq 4$, then $m = 1$ and we are using the same code as in Theorem 1 so the bound $EL(X^n) \leq 3 + (9/2)x$ applies. For $x > 4$ we have the inequalities

$$\begin{aligned} \log x &> 2, \\ \log \log x &> 1, \text{ and} \\ \frac{\log x}{\sqrt{x}} - \frac{\log x}{x} &\leq \frac{\log \log x}{\log x} \ . \end{aligned}$$

From these, we have

$$EL(X^n) \le 1 + x(1 + 8\frac{\log\log x}{\log x}) \ ,$$

and thus

$$EL(X^n) \le \begin{cases} 1 + x(1 + 8\frac{\log\log x}{\log x}) & x > 4 \ , \\ 3 + 9/2x & x \le 4 \ . \end{cases}$$

Both bounds equal 21 bits at $x = 4$. Coding m takes at most $1 + \log(1 + n\,H_F)$ bits. So we have the universally true bound that

$$E\,L(X^n) \le 20 + n\,H_F \left(1 + 10\frac{\log\log nH_F}{\log nH_F}\right)$$

where $\frac{\log\log nH_F}{\log nH_F}$ is taken to be zero if $nH_F < 2$. □

Corollary 1 *The relative redundancy converges to zero for any sequence of n_i and F_i such that $n_i H_i \to \infty$, where H_i is the entropy of the distribution F_i,*

$$E\,L(X^{n_i}) = n_i\,H_i + o(n_i H_i) \ .$$

Thus, the relative redundancy goes to 0, asymptotically, as the minimum *expected number of bits goes to infinity. More precisely the rate can be expressed as*

$$\frac{E\,L(X^{n_i})}{n_i H_i} \le 1 + \frac{2\log\log(n_i H_i) + O(1)}{\log(n_i H_i)} \ .$$

Proof. Dividing equation (16) by $n\,H_F$ to compute a relative redundancy, we find

$$\frac{E\,L(X^n)}{n\,H_F} \le 1 + \frac{1}{n\,H_F} + \frac{2m(1 + \log(n\,H_F))}{n\,H_F} + \frac{1 + \log(1 + \log(m+1))}{\log(m+1)} \ .$$

With $m = \sqrt{n\,H_F} - 1$, this reduces to

$$\frac{EL(X^n)}{n\,H_F} \le 1 + \frac{1}{n\,H_F} + \frac{2(1 + \log(n\,H_F))}{\sqrt{n\,H_F}} + \frac{1 + \log(\log(n\,H_F))}{(1/2)\log(n\,H_F)} \ .$$

Noting $(\log x)/\sqrt{x} = o(1/\log x)$ we get the result. □

If we take the distribution to be constant, namely $F_i = F$, then the corollary shows that our code is weakly minimax in the usual sense. Namely,

$$\sup_{F \in \mathcal{M}} \lim_{n\to\infty} \frac{EL(X^n) - nH_F}{n} = 0.$$

We also obtain a rate which is a direct extension the results of [2] and [5].

Our goal now is to provide a firm upper bound on the code length for all sequences. To obtain such a bound, we modify our algorithm slightly. We first send a flag bit indicating if the sequence is all 1. If the sequence is non-trivial we compress using the previous two-part code. In the appendix we prove that this modified code achieves the following performance:

Theorem 3. *There exists a constant c and a uniquely decodable prefix code for X^n whose length function $L(X^n)$ satisfies*

$$(\forall n \geq 16, \forall F \in \mathcal{M}) \qquad E\, L(X^n) \leq 1 + nH_F \left(1 + \frac{c \log\log\log n}{\log\log n}\right) .$$

Corollary 2. *The modified code is strongly minimax in the sense that*

$$\lim_{n\to\infty} \sup_{F\in\mathcal{M}} \frac{EL(X^n) - (nH_F + 1)}{nH_F} = 0 ,$$

where we take $0/0 = 0$ if $H_F = 0$.

Appendix: Supplemental Proofs

Proof of Lemma 1: We prove Lemma 1 by establishing that for finite alphabets, even if one knew more about the distribution than that it was monotone, one still cannot create a uniform bound for the per-symbol redundancy. Define the class $\mathcal{S}_m$ as consisting of the uniform distributions U_k on the integers $1, ..k$, for all $m \geq k \geq 1$. Let $\mathcal{S} = \bigcup S_m$. Since $\mathcal{S} \subset \mathcal{M}$, the following implies Lemma 1: For any prefix code with length L,

$$(\forall n) \quad \sup_{F\in\mathcal{S}} \frac{E\, L(X^n) - n\, H_F}{n} = \infty.$$

To prove this, first note that for all $n \geq 1$,

$$\sup_{F\in\mathcal{S}} EL(X^n) - nH_F \geq \sup_{F\in\mathcal{S}} EL(X^1) - H_F.$$

this converts the problem to the single dimensional case. We can now apply the redundancy-capacity theorem (see for example [7]) to the family $\mathcal{S}_m$ by considering a discrete memoryless channel C_m whose input and output are the integers $1, ..m$. The k^{th} row of the channel matrix is the probability vector associated with distribution U_k, namely

$$(1/k, 1/k, ..., 1/k, 0, 0, ..., 0) \ .$$

That is, we input the parameter k and output a random integer from U_k. The capacity of C_m is well-known (see [9]) to be

$$C_m = \log \sum_{k=1}^{m} \frac{(k-1)^{k-1}}{k^k} \ .$$

Hence,

$$\sup_{F \in S} EL(X) - H_F \geq \sup_m C_m = \infty.$$

□

Lemma 2 *Let $Y \sim Bi(n,p)$ denote a binomial random variable with parameters $n \geq 4$ and $p < 1/3$. Then*

$$E \ \log \binom{n}{Y} \geq EY \ .$$

Proof. Let $p_k = \binom{n}{k} p^k (1-p)^{n-k}$ and $q_k = 2^k p^k (1-p)^{n-k}$. So long as the q_k define a sub-probability function with $S_n = \sum_{k=0}^{n} q_k \leq 1$, the relative entropy $D(p\|q) \geq 0$ and thus

$$\sum_{k=0}^{n} p_k \log \binom{n}{k} \geq \sum_{k=0}^{n} p_k \log 2^k = EY \ . \tag{17}$$

To show that $S_n \leq 1$, write the sum of the q_k as

$$\begin{aligned} S_n &= (1-p)^n \sum_{k=0}^{n} \left(\frac{2p}{1-p} \right)^k \\ &= (1-p)^n \left(\frac{1 - \left(\frac{2p}{1-p}\right)^{n+1}}{1 - \frac{2p}{1-p}} \right) \\ &= \frac{(1-p)^{n+1} - (2p)^{n+1}}{1-3p} \ . \end{aligned} \tag{18}$$

We first show that $S_4 \leq 1$ and continue by induction. The lemma also holds for $n = 3$, but that situation requires a different proof. (It fails for $n = 1, 2$.)

For $n = 4$, notice that the numerator in the fraction of (18) clearly has a root at $1/3$, so that the denominator can be canceled. This gives

$$S_4 = 1 - 2p + 4p^2 + 2p^3 + 11p^4 \ .$$

With $w = 3p$, the ratio

$$\frac{S_4 - 1}{p} = \frac{11w^3 + 6w^2 + 36w - 54}{27}$$

is negative for $0 \le w \le 1$ because $53 < 54$. Thus $S_4 \le 1$ for $0 \le p \le 1/3$. For the induction, it follows that for $n \ge 4$,

$$\begin{aligned} S_{n+1} &= (1-p)S_n + (2p)^{n+1} \\ &\le (1-p) + (2p)^{n+1} \\ &\le 1 \,. \end{aligned}$$

The final inequality holds because $2(2p)^2 \le 1$ implies that $(2p)^{n+1} \le p$, again for $n \ge 4$.

Lemma 3. *If $Y \sim Bi(n,p)$, then for $p < 1/3$ and $n \ge 4$, the expected length of the penultimate code for Y is dominated by a linear function of the entropy $H(Y)$,*

$$2 + 2\,H(Y) \ge EL_\omega(Y) \,. \tag{19}$$

Proof. Define $p_i = P(Y=i) = \binom{n}{i} p^i (1-p)^{n-i}$. Because $p_{i+1}/p_i = p(n-i)/((1-p)(i+1))$, the maximum of the binomial density occurs at $x = \lfloor np \rfloor$ or $x = 1 + \lfloor np \rfloor$. The entropy is then at least

$$\begin{aligned} -E \log P_Y(Y) &\ge -E \log \max_i p_i \\ &= -\max(\log p_x, \log p_{x+1}) \\ &\ge -\log \max \left(\binom{n}{x} (\hat{p}_0)^x (1-\hat{p}_0)^{n-x}, \right. \\ &\qquad \left. \binom{n}{x+1} (\hat{p}_1)^{x+1} (1-\hat{p}_1)^{n-x-1} \right) , \end{aligned} \tag{20}$$

where $\hat{p}_0 = x/n$ and $\hat{p}_1 = (x+1)/n$ are the possible maximum likelihood estimates for p. This expression simplifies because the maximum of these two functions is always the first term, for these values of n and p. To show this, we need the following upper bound for the natural log function,

$$\log_e(1+\delta) \le \delta - c\,\delta^2, \quad -1 \le \delta \le 1 \,, \tag{21}$$

with the constant $c = 1 - \log_e 2$ chosen to achieve equality for $x = 1$. This expression then gives a bound on the ratio of the functions in (20). After canceling numerous terms, we see that the log of the ratio of density functions is negative,

$$\begin{aligned} \log_e \frac{\binom{n}{x+1}\hat{p}_1^{x+1}(1-\hat{p}_1)^{n-x-1}}{\binom{n}{x}\hat{p}_0^x(1-\hat{p}_0)^{n-x}} &= x \log_e\left(1 + \frac{1}{x}\right) + (n-x-1)\log_e\left(1 - \frac{1}{n-x}\right) \\ &\le x\left(\frac{1}{x} - \frac{c}{x^2}\right) + (n-x-1)\left(\frac{-1}{n-x} - \frac{c}{(n-x)^2}\right) \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{1}{n-x} - \frac{c}{x}\right) + \frac{c}{n-x}\left(\frac{1}{n-x} - 1\right) \\
&\leq 0 \; ,
\end{aligned}
$$

because both summands are negative. The first summand is negative for small values of x in the range,

$$
x < \frac{cn}{1+c} \approx .4n \; ,
$$

containing the values covered by this lemma. To complete the proof of the lemma, we need to cover the case $x = 0$ separately. In this case, we obtain the trivial lower bound $H(Y) \geq 0$. For $x > 0$, we use Stirling's approximation,

$$
\log k! = k \log k - k \log e + \tfrac{1}{2} \log(2\pi k) + \epsilon_k \; ,
$$

with an error term of the form $\frac{\log e}{12k+1} < \epsilon_k < \frac{\log e}{12k}$ [3]. Substituting this expression for the log factorials and canceling, we find that the entropy of Y is bounded below by

$$
\begin{aligned}
H(Y) &\geq -\log \binom{n}{x} \hat{p}_0^x (1-\hat{p}_0)^{n-x} \\
&= \tfrac{1}{2} \log n\hat{p}_0(1-\hat{p}_0) + \tfrac{1}{2} \log 2\pi + (\epsilon_x + \epsilon_{n-x} - \epsilon_n) \\
&> \tfrac{1}{2} \log np + \tfrac{1}{2} \log \frac{\hat{p}_0}{p} + \tfrac{1}{2} \log \pi \\
&> \tfrac{1}{2} \log np = \tfrac{1}{2} \log EY \; , \qquad (22)
\end{aligned}
$$

because the sum of the error terms $\epsilon_x + \epsilon_{n-x} - \epsilon_n > 0$, and $x = \lfloor np \rfloor \geq 1$ implies $\pi\hat{p}_0 > p$. Finally, using Jensen's inequality, the average code length is bounded above by

$$
E\, L_\omega(Y) \leq E\; 2(1 + \lfloor \log Y \rfloor) \leq 2 + 2 \log EY \; ,
$$

and (19) follows when this inequality is combined with (22).

Our proof of Theorem 3 makes use of the following lemma. We consider the conditional entropy of the data given that the sequence is non-trivial, that is, it is not all 1. We state and prove

Lemma 7 *Let X_i be i.i.d. with a common monotonic distribution on the positive integers. Let $A = \{X_i > 1$ for at least one $i = 1, \ldots n\}$. Then*

$$
H(X_1, \ldots, X_n | A) \geq \log n \left(1 - \frac{q_1^n}{1-p_1^n}\right) \; ,
$$

where $p_1 = P\{X_1 = 1\}$ and $q_1 = 1 - p_1$.

Proof. To prove the lemma we consider three simple sets that partition the set of all possible outcomes: (i) the all one sequence (ii) the set of sequences that have no ones and (iii) the set of sequences with at least one 1, but not all. Formally, we let

$$A^c = \{X_i = 1 \text{ for all } i = 1, \dots, n\}$$

and we introduce the set

$$B = \{X_i = 1 \text{ for some } i \text{ and } X_j \neq 1 \text{ for at some } j \neq i\}$$

as well as the set

$$C = \{X_i \neq 1 \text{ for all } i\}.$$

Note that $A = B \cup C$ and that $\Omega = B \cup C \cup A^c$.

Using these sets we partition the entropy of X^n given A:

$$\begin{aligned} H(X^n|A) &= P(B|A)H(X^n|A,B) + P(C|A)H(X^n|A,C) + h(P(B|A)) \\ &\geq P(B|A)H(X^n|A,B) \\ &= P(B|A)H(X^n|B). \end{aligned}$$

Now it is easy to see that $P(A) = P(B) + P(C)$ and that $P(A) = 1 - P(A^c) = 1 - p_1^n$. It thus follows that

$$\begin{aligned} P(B|A) &= \frac{P(B)}{P(A)} = \frac{P(A) - P(C)}{P(A)} \\ &= 1 - \frac{P(C)}{P(A)} = 1 - \frac{q_1^n}{1 - p_1^n}. \end{aligned}$$

Thus we need only find a lower bound for $H(X^n|B)$. To this end, define $Z_j = 1\{X_j \neq 1\}$. From the data processing inequality, it follows that

$$H(X^n|B) \geq H(Z_1^n | 0 < \sum_{j=1}^{n} Z_j < n).$$

For any binary n-vector z_1^n, its *type* is defined in the usual way:

$$T(z_1^n) = \sum_{j=1}^{n} z_j.$$

That is the type of z_1^n is the number of 1's. Because Z_j is i.i.d. it follows that the probability of z_1^n depends only on its type. Now let

$$T(k) = \{z_1^n \in \{0,1\}^n : \text{ such that } T(z_1^n) = k\}.$$

The cardinality of $T(k)$ is easily seen to be $\binom{n}{k}$. Thus

$$H(Z_1^n|T(Z_1^n)=k) = \log\binom{n}{k} \geq \log n.$$

Furthermore we have the simple decomposition:

$$H(Z_1^n|0<T(Z_1^n)<n) = \sum_{k=1}^{n-1} P(T(Z_1^n)=k|0<T(Z_1^n)<n)H(Z_1^n|T(Z_1^n)=k).$$

Thus it follows that

$$H(Z_1^n|0<\sum_{i=1}^{n} Z_i<n) \geq \log n \ ,$$

which in turn implies that

$$H(X^n|B) \geq \log n.$$

This completes the proof of the lemma. □

We can now prove

Theorem 3. *There exists a uniquely decodable prefix code for X^n whose length function $L(X^n)$ satisfies*

$$E\,L(X^n) \leq 1 + n\,H_F\,\left[1 + O\left(\frac{\log\log\log n}{\log\log n}\right)\right] \ .$$

Proof. First observe that if $nH_F > \log n$, then the theorem follows immediately from Theorem 2. Our proof thus focuses on the low-entropy case $n\,H_F \leq \log n$. As before, let $p_i = P\{X_i = i\}$. If the input data is constant, $X^n = 1^n$, then we simply code a flag bit. If $X^n \neq 1^n$, we apply the code used in Theorem 2. With the flag bit added, the expected length of the code is

$$E\,L(X^n) = 1 + (1-p_1^n)\ E[L(X^n)|X^n \neq 0] \ .$$

Observe that

$$\begin{aligned} n\,H_F &= H(X^n) = P(X^n=1^n)H(X^n|X^n=1^n) + P(X^n\neq 1^n)H(X^n|X^n\neq 1^n) \\ &= (1-p_1^n)H(X^n|X^n\neq 1) \ , \end{aligned}$$

and so

$$\frac{E\,L(X^n)}{H(X^n)} = \frac{1}{H(X^n)} + \frac{(1-p_1^n)E[L(X^n)|X^n\neq 0]}{H(X^n)}$$

$$= \frac{1}{H(X^n)} + \frac{E\left[L(X^n)|X^n \neq 0\right]}{H(X^n|X^n \neq 0)} .$$

Before we can apply Theorem 2, we need to show that the distribution of X_i is monotonic conditional upon $X^n \neq 1^n$ for $n > 2$. To see that this is the case, suppose first that $p_1 \leq 1 - 1/n$. In this case, $H(X_i)$ is minimized when $p_2 = 1 - p_1$. However, by definition, the boolean entropy $h(1 - p_1) > (\log n)/n$, implying the contradiction $H(X^n) > \log n$. Thus, $p_1 > 1 - 1/n$. Now, it is simple to check that

$$P(X_i = 1|X^n \neq 1) = \frac{p_1(1 - p_1^{n-1})}{1 - p_1^n} ,$$

and for $k > 1$

$$P(X_i = k|X^n \neq 1) = \frac{p_k}{1 - p_1^n} .$$

Thus, the distribution of X_i given that the sequence is non-trivial ($X^n \neq 1^n$) is monotonic iff

$$p_1(1 - p_1^{n-1}) > p_2 ,$$

or

$$p_1 - p_2 > p_1^n . \tag{23}$$

Since $p_1 > 1 - 1/n$, (23) holds for all $n > 2$. Thus we can apply Theorem 2. Since Lemma 7 implies that $H(X^n|X^n \neq 1) > \log n$, our proof is complete. □

References

[1] L.D. Davisson. Universal noiseless coding. *IEEE Trans. on Info. Theory*, 19:783–795, 1973.

[2] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. on Info. Theory*, 21:194–203, 1975.

[3] W. Feller. *An Introduction to Probability Theory and Its Applications.* Wiley, New York, 1968.

[4] L. Györfi, I. Páli, and E. C. van der Meulen. On universal noiseless source coding for infinite source alphabets. *European Transactions on Telecommunications and Related Technologies*, 4:125–132, 1993.

[5] L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for an infinite source alphabet. *IEEE Trans. on Info. Theory*, 40:267, 1994.

[6] S. Kosaraju and G. Manzini. Compression of low entropy string with lempel-ziv algorithms. In *Sequences 1997 Workshop on Compression and Complexity of Sequences*, pages 107–121, Positano, Italy, 1997.

[7] N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. on Info. Theory*, 41:714–722, 1995.

[8] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

[9] J. Sayir. On coding by probability transformation. In J. Massey, editor, *ETH Series in Information Processing*, volume 12, page 80. Hartung-Gorre Verlag Konstanz, 1999.

[10] A. D. Wyner. An upper bound on the entropy series. *Inform. Contr.*, 20:176–181, 1972.