Redundancy of the Lempel–Ziv Incremental Parsing Rule

Serap A. Savari, *Member, IEEE*

***Abstract*— The Lempel–Ziv codes are universal variable-to-fixed length codes that have become virtually standard in practical lossless data compression. For any given source output string from a Markov or unifilar source, we upper-bound the difference between the number of binary digits needed to encode the string and the self-information of the string. We use this result to demonstrate that for unifilar or Markov sources, the redundancy of encoding the first n letters of the source output with the Lempel–Ziv incremental parsing rule (LZ'78), the Welch modification (LZW), or a new variant is $O((\ln n)^{-1})$, and we upper-bound the exact form of convergence. We conclude by considering the relationship between the code length and the empirical entropy associated with a string.**

***Index Terms*— Lempel–Ziv codes, Markov sources, unifilar sources, renewal theory.**

I. Introduction

An important and challenging problem in data compression is trying to better understand the Lempel–Ziv incremental parsing rule [1], which has motivated many practical lossless data compression schemes. We assume that we are encoding the output of a Markov source or a unifilar source; the terms are often used interchangeably (see, e.g., [2, Sec 3.6], [3, Sec 6.4], and [4]). We define a Markov source, i.e., a unifilar source, with finite alphabet $\{0, 1, \cdots, K-1\}$ and set of states $\{0, 1, \cdots, R-1\}$ by specifying, for each state s and letter j,

1) the probability $p_{s,j}$ that the source emits j from state s
2) the unique next state $S[s, j]$ after j is issued from state s.

For any source string σ with an initial state s_0, these rules inductively specify its final state $S[s_0, \sigma]$, its probability $P(\sigma \mid s_0)$, and its self-information in natural units, $I(\sigma \mid s_0) = -\ln P(\sigma \mid s_0)$. In the underlying Markov chain, let $f_{s,r}$ denote the transition probability from state s to state r; then

$$f_{s,r} = \sum_{j:S[s,j]=r} p_{s,j}.$$

We assume the source has a single recurrent class of states; i.e., the underlying Markov chain has a single recurrent class of states or, equivalently, for each pair of states s and r, there is a string σ with positive probability that drives the source to state r from state s. Let π_s denote the steady-state probability that the source is in state s and let $\mathcal{H}$ represent the entropy of the source in natural units. If $\boldsymbol{F} = [f_{s,r}]$, $\boldsymbol{\pi} = (\pi_0 \cdots \pi_{R-1})$, and $\boldsymbol{e}$ denotes the column vector of length R consisting of all ones, then $\boldsymbol{\pi}$ and $\mathcal{H}$ are given by

$$\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \boldsymbol{F}$$

$$\boldsymbol{\pi} \cdot \boldsymbol{e} = 1$$

$$\mathcal{H} = -\sum_{s=0}^{R-1} \sum_{j=0}^{K-1} \pi_s p_{s,j} \ln p_{s,j}.$$

The class of sources that can be modeled by a Markov source is fairly general and includes, for each $l \geq 1$, the family of sources for which each output depends statistically only on the l previous output symbols.

We also assume that the output of the source is encoded into a uniquely decodable sequence of letters from a binary channel alphabet. In [5], Shannon established that the average number of binary digits per source symbol that can be achieved by any such source coding technique is lower-bounded by $\mathcal{H} \log_2 e$. The *redundancy* of a source code is the amount by which the average number of binary digits per source symbol associated with that code exceeds Shannon's entropy bound. One of the goals in developing source coding algorithms is to minimize redundancy.

There are many well-known source codes such as the Huffman code [6], the Tunstall code (see [7] and [8]) and arithmetic codes (see [9]) for which the average number of binary digits per source symbol comes arbitrarily close to Shannon's entropy bound. A practical disadvantage of each of these algorithms is that they require an *a priori* knowledge of the source model. The alternative is to use an adaptive or universal source code, i.e., a code which needs no *a priori* assumptions about the statistical dependencies of the data to be encoded. There is an extensive literature on universal coding, and there are many types of universal codes. For example, the dynamic Huffman code (see, e.g., [10]) is an adaptive version of the Huffman code, and there is a way to use arithmetic coding in an adaptive way (see [11]). However, among the existing universal codes, the encoding techniques motivated by the 1977 Lempel–Ziv algorithm (see [12]) and the 1978 Lempel–Ziv algorithm (see [1]) are virtually standard in practical lossless data compression because they empirically achieve good compression, and they are computationally efficient. However, the Lempel–Ziv codes are not as well

Manuscript received February 26, 1996; revised June 1, 1996. This work was supported by an AT&T Bell Laboratories GRPW Fellowship and a Vinton Hayes Fellowship.

The author was with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA. She is now at the Computing Science Research Center, Lucent Technologies, Murray Hill, NJ 07974 USA.

Publisher Item Identifier S 0018-9448(97)00178-8.

understood as many other well-known codes. We will focus on gaining insight into the 1978 Lempel–Ziv code, often called the Lempel–Ziv incremental parsing rule or LZ'78, and two of its variants.

The Lempel–Ziv incremental parsing rule starts off with a dictionary consisting of the K source symbols. At any parsing point, the next parsed phrase σ is the unique dictionary entry which is a prefix of the unparsed source output. For all three of the encoding procedures we will consider, if the dictionary contains M entries, then $\lceil \log_2 M \rceil$ bits are used to encode the next parsed phrase. Once this phrase has been selected, the dictionary for the Lempel–Ziv incremental parsing rule is enlarged by replacing σ with its K single-letter extensions. As an example, suppose that we have a ternary source, and the source output is the string $0\ 0\ 0\ 0\ 2\ \cdots$.

- Initially, the dictionary is $\{0,\ 1,\ 2\}$.
- The first parsed string is 0, and the dictionary is updated to $\{00,\ 01,\ 02,\ 1,\ 2\}$. At this point, the unparsed source output is $0\ 0\ 0\ 2\ \cdots$.
- The second parsed string is 00, and the revised dictionary is $\{000,\ 001,\ 002,\ 01,\ 02,\ 1,\ 2\}$. Now, the unparsed source sequence is $0\ 2\ \cdots$.
- The third parsed string is 02, resulting in the dictionary $\{000,\ 001,\ 002,\ 01,\ 020,\ 021,\ 022, 1,\ 2\}$.

Practical implementations of the Lempel–Ziv incremental parsing rule often differ somewhat from the original LZ'78 algorithm. We will focus on the LZW algorithm introduced by Welch in [13]. Initially, the dictionary entries are the K source symbols. At any parsing point, the next parsed phrase is the longest dictionary entry which is a prefix of the unparsed source output. Thus far, the parsing rule is identical to the one used by LZ'78. The difference is in the way the dictionaries are updated in the two procedures. In the Lempel–Ziv incremental parsing rule, the last parsed phrase is replaced by its K single-letter extensions. For LZW, the dictionary is enlarged by adding the last parsed phrase concatenated with the first symbol of the unparsed source output. According to [13], LZW achieves very similar compression to LZ'78, but is easier to implement. Miller and Wegman discussed a "character extension improvement" algorithm in [14] that is identical to LZW. They claimed that the algorithm empirically achieves better compression than LZ'78 on English text, especially for small dictionary sizes. Miller and Wegman attributed the empirical success of LZW to the addition of one new dictionary string per parsed string versus the net gain of $K-1$ dictionary strings per parsed string created by LZ'78; i.e., each parsed string is represented by approximately $\log_2(K-1)$ fewer binary digits. Note that any string can appear as a parsed phrase at most once for the original Lempel–Ziv incremental parsing rule, while it can occur as a parsed phrase up to K times for LZW. Let us continue the previous example by examining how the LZW parser would segment the source output sequence $0\ 0\ 0\ 0\ 2\ \cdots$.

- Initially, the dictionary is $\{0,\ 1,\ 2\}$.
- The first parsed string is 0, and the remaining source output is $0\ 0\ 0\ 2\ \cdots$. Hence, the dictionary is enlarged to $\{0,\ 00,\ 1,\ 2\}$.
- The next parsed string is 00, and the unparsed source sequence is now $0\ 2\ \cdots$. The dictionary is expanded to $\{0,\ 00,\ 000,\ 1,\ 2\}$.
- The third parsed string is 0, and the rest of the source output is $2\ \cdots$. The new dictionary is $\{0,\ 00,\ 02,\ 000,\ 1,\ 2\}$ and the fourth parsed phrase is 2.

For LZ'78, it is clear that the decoder can use the sequence of code symbols to simulate the evolution of the parser's dictionary and subsequently reconstruct the source output; it is less obvious that the LZW decoder has this property. For any string σ and letter j, define the string $\sigma \circ j$ as the string formed by appending j to the string σ. The LZW decoder can easily determine the first source output symbol u_1. The new dictionary entry is of the form $u_1 \circ j$ for some source symbol j. To find j, the decoder looks at the code letters corresponding to the second phrase. If these code letters indicate that the second parsed phrase is u_1 or $u_1 \circ j$, then j and u_1 are the same symbol. Otherwise, the second parsed phrase is some u_2 which is distinct from u_1 and, therefore, j is the same as u_2. This argument can be extended to show that it is possible to accurately decode any source string from its corresponding string of code letters.

There are many small modifications that can be made to LZ'78 or LZW in order to create new encoding rules. For example, Gallager [15] proposed a variant G of LZW. Suppose that a string σ has occurred $K-2$ times as a parsed string for LZW. Then it has two single-letter extensions, say $\sigma \circ j_1$ and $\sigma \circ j_2$, which are not dictionary entries. Without loss of generality, assume that σ is next used as a parsed string when $\sigma \circ j_1$ is a prefix of the unparsed source output starting from a parsing point. Then $\sigma \circ j_1$ will be the new dictionary entry and σ will be used as a parsed string for the Kth time if and only if there is a parsing point at which $\sigma \circ j_2$ is a prefix of the unparsed source output. In G, when a string is used as a parsed string for the $K-1$st time, the dictionary is updated by replacing the string with its two single-letter extensions which are not already in the dictionary. Note that the size of the dictionary for G grows by one each time a string is parsed, and a string can be used as a parsed phrase up to $K-1$ times. For $K=2$, the rule G is the same as LZ'78. Let us continue our example and see how G would segment the source output sequence $0\ 0\ 0\ 0\ 2\ \cdots$.

- Initially, the dictionary is $\{0,\ 1,\ 2\}$.
- The first parsed string is 0, and the remaining source output is $0\ 0\ 0\ 2\ \cdots$. The new dictionary is $\{0,\ 00,\ 1,\ 2\}$.
- The second parsed string is 00, and the dictionary is enlarged to $\{0,\ 00,\ 000,\ 1,\ 2\}$. Now, the unparsed source sequence is $0\ 2\ \cdots$.
- The next parsed string is 0 and the remainder of the source output is $2\ \cdots$. Since this is the second time that 0 is a parsed string, it will be removed from the dictionary and the strings 01 and 02 will be added. The new dictionary is $\{00,\ 01,\ 02,\ 000,\ 1,\ 2\}$ and the fourth parsed phrase is 2.

Let u_1^n symbolize the string $u_1, \cdots, u_n$. It is assumed that the decoder knows the length n of u_1^n in advance. If the last parsed phrase σ is a partial phrase, the encoder will transmit the codeword corresponding to any dictionary entry which has σ as a prefix. Let $\mathcal{L}^{\mathrm{LZ}}(u_1^n)$, $\mathcal{L}^{\mathrm{W}}(u_1^n)$, and $\mathcal{L}^{\mathrm{G}}(u_1^n)$ denote the length of the encoding of the string u_1^n in bits for LZ'78, LZW, and G, respectively. Let U_1^n be the random string corresponding to the first n letters emitted from the source. The redundancies $\mathcal{R}^{\mathrm{LZ}}$, $\mathcal{R}^{\mathrm{W}}$, and $\mathcal{R}^{\mathrm{G}}$ of the codes in bits are

$$\mathcal{R}^{\mathrm{LZ}} = E\left(\frac{1}{n}\mathcal{L}^{\mathrm{LZ}}(U_1^n)\right) - \mathcal{H}\log_2 e \qquad (1)$$

$$\mathcal{R}^{\mathrm{W}} = E\left(\frac{1}{n}\mathcal{L}^{\mathrm{W}}(U_1^n)\right) - \mathcal{H}\log_2 e \qquad (2)$$

and

$$\mathcal{R}^{\mathrm{G}} = E\left(\frac{1}{n}\mathcal{L}^{\mathrm{G}}(U_1^n)\right) - \mathcal{H}\log_2 e \qquad (3)$$

where the expectations are taken over all n-tuples. It was demonstrated in [1] that

$$\lim_{n\to\infty} \mathcal{R}^{\mathrm{LZ}} = 0.$$

In [4], it was established that for a binary source, every source output string u_1^n satisfies

$$\frac{\mathcal{L}^{\mathrm{LZ}}(u_1^n) - I(u_1^n \mid s_0)\log_2 e}{n} \le \frac{\ln\ln n}{\ln n} + O\left(\frac{1}{\ln n}\right)$$

and hence,

$$\mathcal{R}^{\mathrm{LZ}} \le \frac{\ln\ln n}{\ln n} + O\left(\frac{1}{\ln n}\right).$$

In [4], it was conjectured that $\mathcal{R}^{\mathrm{LZ}} = \Theta((\ln\ln n)/(\ln n))$. However, in recent years, much of the data compression community believed that $\mathcal{R}^{\mathrm{LZ}} = \Theta((\ln n)^{-1})$. This question was next addressed in [16]; it was claimed there that for a binary, memoryless source, there exists a constant $\mathcal{C}$ which is a function of source parameters and a fluctuating function $\delta(n)$ with small amplitude that satisfies

$$\mathcal{R}^{\mathrm{LZ}} = \frac{\mathcal{C} + \delta(n)}{\ln n} + O\left(\frac{\ln\ln n}{(\ln n)^2}\right)$$

and that an extension of this result exists for those Markov sources for which each output symbol depends statistically only on the previous output symbol. The redundancy bound of [4] was obtained by studying the number of parsed phrases of a given length in order to bound the total number of parsed phrases. The results of [16] were derived using [17], which states that as n increases, the number of phrases in the parsing of U_1^n approaches a normal distribution with a mean and variance which are functions of n and the source parameters.

Our approach to analyzing the performance of an encoding rule is new. Instead of focusing on the number of parsed phrases of a given length, we show how to use renewal theory to bound the number of phrases corresponding to the parsing of a string in terms of the self-information of the string, and this leads to an upper bound on the length of the encoding of the string. Our main result is to demonstrate that for each of the three encoding rules that we are investigating

$$\mathcal{R} \le O\left(\frac{1}{\ln n}\right)$$

furthermore, the number of binary digits used to represent any string u_1^n satisfies

$$\frac{\mathcal{L}(u_1^n) - I(u_1^n \mid s_0)\log_2 e}{n} \le O\left(\frac{1}{\ln n}\right). \qquad (4)$$

In every case, we upper-bound the exact rate of convergence.

II. New Redundancy Bound

In evaluating $\mathcal{L}^{\mathrm{LZ}}(u_1^n)$, $\mathcal{L}^{\mathrm{W}}(u_1^n)$, and $\mathcal{L}^{\mathrm{G}}(u_1^n)$, we will use the following elementary result.

Lemma 1: For any integer $k \ge 2$, and real number $x \ge 0$,

$$\begin{aligned}\sum_{i=1}^{k} \lceil x+\log_2 i\rceil &\le k \cdot \lfloor \log_2(2^x k)\rfloor \\ &\quad + k - 2\cdot 2^{-x+\lfloor \log_2(2^x k)\rfloor} + O(\ln k) \\ &\le k\log_2 k + kx + k\log_2\left(\frac{\log_2 e}{e}\right) + O(\ln k).\end{aligned}$$

The proof of Lemma 1 can be found in Appendix I. A related result is presented in [18, Example 1.2.4.42].

Let c^{LZ}, c^{W}, and c^{G} represent the number of complete phrases obtained by parsing u_1^n according to LZ'78, LZW, and G, respectively. For LZ'78, the dictionary starts with K entries and has a net gain of $K-1$ strings per parse; hence, the size of the dictionary used to select the ψth parsed string is $\psi(K-1)+1$. Since $\lceil \log_2 M\rceil$ bits are used to encode any entry of a dictionary of size M for each parsing rule, $\mathcal{L}^{\mathrm{LZ}}(u_1^n)$ satisfies

$$\begin{aligned}\mathcal{L}^{\mathrm{LZ}}(u_1^n) &< \sum_{j=1}^{c^{\mathrm{LZ}}+1} \lceil \log_2(j(K-1)+1)\rceil \\ &< \sum_{j=1}^{c^{\mathrm{LZ}}+1} \lceil \log_2(K-1) + \log_2(j+1)\rceil.\end{aligned}$$

Therefore, it follows from Lemma 1 that

$$\begin{aligned}\mathcal{L}^{\mathrm{LZ}}(u_1^n) &\le c^{\mathrm{LZ}}\log_2 c^{\mathrm{LZ}} \\ &\quad + c^{\mathrm{LZ}}\log_2\left(\frac{(K-1)\log_2 e}{e}\right) + O(\ln c^{\mathrm{LZ}}).\end{aligned} \qquad (5)$$

For LZW and G, the number of possibilities for the ψth parsed string is $\psi + K - 1$ and thus Lemma 1 implies that

$$\mathcal{L}^{\mathrm{W}}(u_1^n) \le c^{\mathrm{W}}\log_2 c^{W} + c^{\mathrm{W}}\log_2\left(\frac{\log_2 e}{e}\right) + O(\ln c^{\mathrm{W}}) \qquad (6)$$

and

$$\mathcal{L}^{\mathrm{G}}(u_1^n) \le c^{\mathrm{G}}\log_2 c^{\mathrm{G}} + c^{\mathrm{G}}\log_2\left(\frac{\log_2 e}{e}\right) + O(\ln c^{\mathrm{G}}). \qquad (7)$$

From (5) to (7), we see that upper bounds on c^{LZ}, c^{W}, and c^{G} lead to upper bounds on $\mathcal{L}^{\mathrm{LZ}}$, $\mathcal{L}^{\mathrm{W}}$, and $\mathcal{L}^{\mathrm{G}}$, respectively. We have the following results.

Theorem 1: Assume that the source has positive entropy $\mathcal{H}$ and that the self-information corresponding to the symbols emitted by the source has a *nonarithmetic* distribution; i.e., there is no constant Λ such that $-\ln p_{s,j}$ is an integer multiple of Λ for all pairs of states s and symbols j satisfying $p_{s,j} > 0$. Abbreviate $I(u_1^n \,|\, s_0) < \infty$ by I. For the three encoding rules we are studying, we have the following asymptotic relationships between the number of phrases associated with the parsing of u_1^n and the self-information of u_1^n:

$$c^{\mathrm{LZ}} \cdot \frac{\ln I}{I} \le 1 + o(1) \tag{8}$$

$$(c^{\mathrm{LZ}} \log_2 c^{\mathrm{LZ}} - I \log_2 e) \cdot \frac{\ln I}{I} \le \log_2\left(\frac{Re}{\mathcal{H}}\right) + o(1) \tag{9}$$

$$c^{\mathrm{W}} \cdot \frac{\ln I}{I} \le 1 + o(1) \tag{10}$$

$$(c^{\mathrm{W}} \log_2 c^{\mathrm{W}} - I \log_2 e) \cdot \frac{\ln I}{I} \le \log_2\left(\frac{RKe}{\mathcal{H}}\right) + o(1) \tag{11}$$

$$c^{\mathrm{G}} \cdot \frac{\ln I}{I} \le 1 + o(1) \tag{12}$$

$$(c^{\mathrm{G}} \log_2 c^{\mathrm{G}} - I \log_2 e) \cdot \frac{\ln I}{I} \le \log_2\left(\frac{R(K-1)e}{\mathcal{H}}\right) + o(1). \tag{13}$$

The asymptotic relationship between the length of the encoding of u_1^n and the self-information of u_1^n satisfies

$$\ln n \cdot \left(\frac{\mathcal{L}^{\mathrm{LZ}}(u_1^n) - I \log_2 e}{n}\right) \le \frac{I}{n} \log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right) + o(1) \tag{14}$$

$$\ln n \cdot \left(\frac{\mathcal{L}^{\mathrm{W}}(u_1^n) - I \log_2 e}{n}\right) \le \frac{I}{n} \log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right) + o(1) \tag{15}$$

$$\ln n \cdot \left(\frac{\mathcal{L}^{\mathrm{G}}(u_1^n) - I \log_2 e}{n}\right) \le \frac{I}{n} \log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right) + o(1). \tag{16}$$

Proof: We introduce the following notation for the Lempel–Ziv incremental parsing rule. Let

- σ_i denote the ith phrase of the source output
- ψ_i denote the source state just before phrase σ_i
- $\Omega = \{\tau : I(\sigma \,|\, s) = \tau \text{ for some string } \sigma \text{ and state } s\}$
- $c(\tau) = |\{\sigma_i : 1 \le i \le c^{LZ},\ I(\sigma_i \,|\, \psi_i) = \tau\}|$
- $\mu(\tau) = |\{(s, \sigma) :\ I(\sigma \,|\, s) = \tau\}|$
- $\gamma(\tau) = |\{(s, \sigma) :\ I(\sigma \,|\, s) \le \tau\}|$
- $\tilde{I} = \sum_{i=1}^{c^{\mathrm{LZ}}} I(\sigma_i \mid \psi_i); \tilde{I} \le I$ because of the possible final partial phrase.

Note that

$$c^{\mathrm{LZ}} = \sum_{\tau \in \Omega} c(\tau) \tag{17}$$

and

$$\tilde{I} = \sum_{\tau \in \Omega} \tau \cdot c(\tau). \tag{18}$$

To upper-bound c^{LZ} for a given $I(u_1^n)$, we maximize

$$\sum_{\tau \in \Omega} c(\tau)$$

subject to the constraints

$$\sum_{\tau \in \Omega} \tau \cdot c(\tau) \le I$$

and $0 \le c(\tau) \le \mu(\tau)$ for all $\tau \in \Omega$. We will show that the number of phrases is maximized by selecting as many phrases with small self-information as possible. In particular, we are going to pick a "threshhold" self-information $\bar{\tau}$ and consider the set $\mathcal{S}$ of strings with self-information upper-bounded by $\bar{\tau}$. Our choice of $\bar{\tau}$ is determined by the criterion that the cumulative self-information of the strings in $\mathcal{S}$ is approximately I. We will upper-bound c^{LZ} by the size of $\mathcal{S}$. More precisely, we have the following result.

Lemma 2: An upper bound on c^{LZ} is given by

$$c^{\mathrm{LZ}} \le \gamma(\bar{\tau}) \tag{19}$$

where $\bar{\tau}$ is chosen so that

$$\sum_{\tau \in \Omega : \tau < \bar{\tau}} \tau \cdot \mu(\tau) < I \le \sum_{\tau \in \Omega : \tau \le \bar{\tau}} \tau \cdot \mu(\tau). \tag{20}$$

Note that $\bar{\tau}$ is a nondecreasing function of I and that as n approaches infinity, both I and $\bar{\tau}$ approach infinity.

Proof of Lemma 2: To arrive at a contradiction, suppose that (19) is false. Then $c^{\mathrm{LZ}} \ge \gamma(\bar{\tau}) + 1$. The encoding rule ensures that every complete phrase in the parsing is distinct. Let h_i denote the self-information of phrase i; i.e., $h_i = I(\sigma_i \,|\, \psi_i)$. Without loss of generality, reorder the self-informations so that $h_1 \le h_2 \le \cdots \le h_{c^{\mathrm{LZ}}}$. Now consider the $\gamma(\bar{\tau})$ distinct pairs (s, ϕ) with $I(\phi \,|\, s) \le \bar{\tau}$ and order the pairs so that

$$I(\phi_1 \,|\, s_1) \le I(\phi_2 \,|\, s_2) \le \cdots \le I(\phi_{\gamma(\bar{\tau})} \mid s_{\gamma(\bar{\tau})}).$$

Observe that for each $i \in \{1, 2, \ldots, \gamma(\bar{\tau})\}$, $h_i \ge I(\phi_i \,|\, s_i)$, and for each $i > \gamma(\bar{\tau})$, $h_i > \bar{\tau}$. Hence

$$\begin{aligned} I \ge \tilde{I} = \sum_{i=1}^{c^{\mathrm{LZ}}} h_i \ge \sum_{i=1}^{\gamma(\bar{\tau})+1} h_i > \sum_{i=1}^{\gamma(\bar{\tau})} I(\phi_i \mid s_i) + \bar{\tau} \\ = \sum_{\tau \in \Omega : \tau \le \bar{\tau}} \tau \cdot \mu(\tau) + \bar{\tau} \ge I + \bar{\tau} \end{aligned}$$

which is a contradiction. □

To understand the relationships among $\bar{\tau}, \gamma(\bar{\tau})$, and

$$\sum_{\tau \in \Omega : \tau \le \bar{\tau}} \tau \cdot \mu(\tau)$$

we investigate how the source generates self-information. We can model the generation of self-information as a *renewal process* (see [19, Sec 3] and [20, Sec 3]). In a renewal process, renewals occur at randomly chosen epochs in time and successive interrenewal periods, i.e., the intervals between renewals, are independent and identically distributed nonnegative random variables; we assume that the process starts evolving at time zero. A related type of stochastic process is a *delayed renewal process.* A delayed renewal process is almost identical to an ordinary renewal process; the only difference is that the first interrenewal period of a delayed renewal process may have a different probability distribution from the remaining ones. For our purposes, we choose the interrenewal periods to represent the self-information generated by the source between successive returns to some given state ψ; an epoch can then be interpreted as the self-information of the source string upon an entrance of the source into state ψ. Because of the Markovian nature of the source, these interrenewal periods are independent random variables and all but the first interrenewal period are also identically distributed. That is, if the source is initially in state ψ, the stochastic process defined above is a renewal process; otherwise, it is a delayed renewal process. For each state ψ and integer $k \geq 2$, we let $J_k^{(\psi)}$ symbolize the self-information, in natural units, generated by the source between the $k-1$th and the kth occurrences of state ψ; $J_1^{(\psi)}$ denotes the self-information produced until the source reaches state ψ for the first time. Let $T_k^{(\psi)} = J_1^{(\psi)} + \cdots + J_k^{(\psi)}$, and let $\{N^{(\psi,s)}(t);\, t \geq 0\}$ be the renewal or delayed renewal process defined, for each state ψ, and starting state s, by specifying the random variable $N^{(\psi,s)}(t)$ as the number of renewals until the self-information reaches t, i.e., the largest integer k for which $T_k^{(\psi)} \leq t < T_{k+1}^{(\psi)}$. Let

$$m^{(\psi,s)}(t) = E[N^{(\psi,s)}(t)].$$

Given the starting state s and the source output $v_1, v_2, \cdots$, each prefix v_1^i with $S[s, v_1^i] = \psi$ corresponds to a renewal in the corresponding sample function; the expected number of renewals in the interval $(t, t+dt]$ is given by

$$m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t) = \sum_{\sigma: I(\sigma \mid s) \in (t, t+dt], \psi = S[s,\sigma]} P(\sigma \mid s). \tag{21}$$

For each string σ with $I(\sigma \mid s) \in (t, t+dt]$, we have that $e^{-t-dt} \leq P(\sigma \mid s) < e^{-t}$. Hence, it follows from (21) that

$$\begin{aligned} e^{-t-dt} \cdot |\sigma : I(\sigma \mid s) \in (t, t+dt], \psi = S[s,\sigma]| &\\ \leq m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t) &\\ < e^{-t} \cdot |\sigma : I(\sigma \mid s) \in (t, t+dt], \psi = S[s,\sigma]| & \end{aligned}$$

and thus the number of strings σ with $I(\sigma \mid s) \in (t, t+dt]$ and $\psi = S[s, \sigma]$ is in the interval

$$\begin{aligned} (e^t[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)], &\\ e^{t+dt}[m^{(\psi,s)}(t+dt) - m^{(\psi,s)}(t)]]. & \end{aligned}$$

We have that

$$\gamma(\bar{\tau}) = \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^+} e^x \, dm^{(\psi,s)}(x) \tag{22}$$

and

$$\sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^-} x e^x dm^{(\psi,s)}(x) < I \leq \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^+} x e^x \, dm^{(\psi,s)}(x). \tag{23}$$

Hence, by multiplying both sides of (22) and (23) by $e^{-\bar{\tau}}$, we find that

$$\gamma(\bar{\tau}) \cdot e^{-\bar{\tau}} = \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^+} e^{x-\bar{\tau}} dm^{(\psi,s)}(x) \tag{24}$$

and

$$\begin{aligned} &\sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^-} x e^{x-\bar{\tau}} dm^{(\psi,s)}(x) \\ &< I \cdot e^{-\bar{\tau}} \leq \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^+} x e^{x-\bar{\tau}} dm^{(\psi,s)}(x) \\ &= \sum_{s=0}^{R-1} \sum_{\psi=0}^{R-1} \int_0^{\bar{\tau}^+} (x - \bar{\tau}) e^{x-\bar{\tau}} dm^{(\psi,s)}(x) + \bar{\tau}\gamma(\bar{\tau}) \cdot e^{-\bar{\tau}}. \end{aligned} \tag{25}$$

In Appendix II, we establish the following result for sources in which the self-information per symbol has a nonarithmetic distribution.

Lemma 3: For all states s, as $\bar{\tau}$ increases

$$\int_0^{\bar{\tau}^+} e^{x-\bar{\tau}} dm^{(\psi,s)}(x) \longrightarrow \frac{\pi_\psi}{\mathcal{H}} \tag{26}$$

$$\int_0^{\bar{\tau}^-} (x - \bar{\tau}) e^{x-\bar{\tau}} dm^{(\psi,s)}(x) \longrightarrow -\frac{\pi_\psi}{\mathcal{H}} \tag{27}$$

and

$$\int_0^{\bar{\tau}^+} (x - \bar{\tau}) e^{x-\bar{\tau}} dm^{(\psi,s)}(x) \longrightarrow -\frac{\pi_\psi}{\mathcal{H}}. \tag{28}$$

It follows from (24)–(28) that as $\bar{\tau}$ increases,

$$\gamma(\bar{\tau}) \cdot e^{-\bar{\tau}} = \frac{R}{\mathcal{H}} + o(1) \tag{29}$$

and

$$(\bar{\tau}\gamma(\bar{\tau}) - I) \cdot e^{-\bar{\tau}} = \frac{R}{\mathcal{H}} + o(1). \tag{30}$$

Taking the logarithm of both sides of (29), we find that as $\bar{\tau}$ increases,

$$\log_2 \gamma(\bar{\tau}) - \bar{\tau} \log_2 e = \log_2 \left(\frac{R}{\mathcal{H}} \right) + o(1). \tag{31}$$

Multiplying (29) by (31), we see that

$$(\gamma(\bar{\tau}) \log_2 \gamma(\bar{\tau}) - \bar{\tau}\gamma(\bar{\tau}) \log_2 e) \cdot e^{-\bar{\tau}} = \frac{R}{\mathcal{H}} \log_2 \left(\frac{R}{\mathcal{H}} \right) + o(1). \tag{32}$$

Multiplying both sides of (30) by $\log_2 e$ and adding the resulting expression to (32), we find that as $\bar{\tau}$ increases,

$$(\gamma(\bar{\tau})\log_2\gamma(\bar{\tau}) - I\log_2 e)\cdot e^{-\bar{\tau}} = \frac{R}{\mathcal{H}}\log_2\left(\frac{Re}{\mathcal{H}}\right) + o(1). \tag{33}$$

Next, we would like to determine the asymptotic relationship between $e^{-\bar{\tau}}$ and I. Substituting (29) into (30), we see that

$$\bar{\tau}\frac{R}{\mathcal{H}} + o(\bar{\tau}) - I\cdot e^{-\bar{\tau}} = 0. \tag{34}$$

Define δ to satisfy

$$\bar{\tau} = \ln\left(\frac{\frac{\mathcal{H}I}{R}}{\ln\left(\frac{\mathcal{H}I}{R}\right)}(1+\delta)\right). \tag{35}$$

Substituting (35) into (34), we find that

$$\frac{R}{\mathcal{H}}\ln\left(\frac{\mathcal{H}I}{R}\right) + \frac{R}{\mathcal{H}}\ln(1+\delta) - \frac{R}{\mathcal{H}}\ln\ln\left(\frac{\mathcal{H}I}{R}\right) + o\left(\ln\left(\frac{\mathcal{H}I}{R}(1+\delta)\right)\right) - \frac{R}{\mathcal{H}(1+\delta)}\ln\left(\frac{\mathcal{H}I}{R}\right) = 0$$

and dividing both sides of this equation by $R\ln(\mathcal{H}I/R)/\mathcal{H}$, we see that

$$1 + \frac{\ln(1+\delta)}{\ln\left(\frac{\mathcal{H}I}{R}\right)} - o(1) + o\left(1 + \frac{\ln(1+\delta)}{\ln\left(\frac{\mathcal{H}I}{R}\right)}\right) - \frac{1}{1+\delta} = 0.$$

Thus

$$\delta = o(1)$$

and hence, as $\bar{\tau}$ and I increase,

$$e^{-\bar{\tau}} = \left(\frac{R}{\mathcal{H}I}\ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1+o(1)). \tag{36}$$

Substituting (36) into (29) and (33), we see that as $\bar{\tau}$ and I increase

$$\gamma(\bar{\tau})\cdot\left(\frac{R}{\mathcal{H}I}\ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1+o(1)) = \frac{R}{\mathcal{H}} + o(1) \tag{37}$$

and

$$(\gamma(\bar{\tau})\log_2\gamma(\bar{\tau}) - I\log_2 e)\cdot\left(\frac{R}{\mathcal{H}I}\ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1+o(1)) = \frac{R}{\mathcal{H}}\log_2\left(\frac{Re}{\mathcal{H}}\right) + o(1). \tag{38}$$

By Lemma 2, $c^{\mathrm{LZ}} \le \gamma(\bar{\tau})$. Hence, by (37) and (38), as I increases

$$c^{\mathrm{LZ}}\cdot\left(\frac{R}{\mathcal{H}I}\ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1+o(1)) \le \frac{R}{\mathcal{H}} + o(1) \tag{39}$$

and

$$(c^{\mathrm{LZ}}\log_2 c^{\mathrm{LZ}} - I\log_2 e)\cdot\left(\frac{R}{\mathcal{H}I}\ln\left(\frac{\mathcal{H}I}{R}\right)\right)(1+o(1)) \le \frac{R}{\mathcal{H}}\log_2\left(\frac{Re}{\mathcal{H}}\right) + o(1). \tag{40}$$

Formulas (39) and (40) are equivalent to (8) and (9), respectively. From (5), (8), and (9), we have

$$(\mathcal{L}^{\mathrm{LZ}}(u_1^n) - I\log_2 e)\cdot\frac{\ln I}{I} \le \log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right) + o(1). \tag{41}$$

Since the source has positive entropy, $I = \Theta(n)$ and, therefore, (14) is equivalent to (41).

We use the same ideas to prove the rest of Theorem 1. For LZW and G, the lemma corresponding to Lemma 2 is

Lemma 4:

$$c^{\mathrm{W}} \le K\gamma(\tau^{\mathrm{W}})$$

and

$$c^{\mathrm{G}} \le (K-1)\gamma(\tau^{\mathrm{G}})$$

where τ^{W} and τ^{G} are chosen so that

$$K\sum_{\tau\in\Omega:\tau<\tau^{\mathrm{W}}}\tau\cdot\mu(\tau) < I \le K\sum_{\tau\in\Omega:\tau\le\tau^{\mathrm{W}}}\tau\cdot\mu(\tau)$$

$$(K-1)\sum_{\tau\in\Omega:\tau<\tau^{\mathrm{G}}}\tau\cdot\mu(\tau) < I \le (K-1)\sum_{\tau\in\Omega:\tau\le\tau^{\mathrm{G}}}\tau\cdot\mu(\tau).$$

Proof of Lemma 4: The difference between Lemma 2 and Lemma 4 lies in the number of times a given string can appear as a parsed phrase for each encoding rule. In LZ'78, a string can occur at most once. Any string can occur up to K times as a parsed phrase for LZW and up to $K-1$ times as a parsed phrase for G. □

For LZW, the counterparts to (29) and (30) are that as τ^{W} increases,

$$\gamma(\tau^{\mathrm{W}})\cdot\exp(-\tau^{\mathrm{W}}) = \frac{R}{\mathcal{H}} + o(1) \tag{42}$$

and

$$(K\tau^{\mathrm{W}}\gamma(\tau^{\mathrm{W}}) - I)\cdot\exp(-\tau^{\mathrm{W}}) = \frac{RK}{\mathcal{H}} + o(1) \tag{43}$$

and for G, as τ^{G} increases,

$$\gamma(\tau^{\mathrm{G}})\cdot\exp(-\tau^{\mathrm{G}}) = \frac{R}{\mathcal{H}} + o(1) \tag{44}$$

and

$$((K-1)\tau^{\mathrm{G}}\gamma(\tau^{\mathrm{G}}) - I)\cdot\exp(-\tau^{\mathrm{G}}) = \frac{R(K-1)}{\mathcal{H}} + o(1). \tag{45}$$

With these modifications, the proofs of (10)–(13), (15), and (16) are identical to the proofs of (8), (9), and (14). □

Next, we will briefly consider the situation in which the self-information corresponding to the source symbols issued has an *arithmetic* distribution with period Λ. Lemma 3 no longer holds, and Theorem 1 is not true in general for arithmetic distributions. However, there are some analogous results that we present in Appendix III. For the remainder of the paper, we consider only nonarithmetic distributions.

An immediate consequence of Theorem 1 is

Corollary 1: Assume that the source has positive entropy. Let $h_{\max}$ denote the maximum self-information generated by the source upon emitting a symbol. Note that $h_{\max}$ is finite. Then for all source output strings u_1^n with nonzero probability

$$\begin{aligned}
&\frac{\mathcal{L}^{\mathrm{LZ}}(u_1^n) - I\log_2 e}{n} \\
&\quad\le \frac{h_{\max}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
&\frac{\mathcal{L}^{\mathrm{W}}(u_1^n) - I\log_2 e}{n} \\
&\quad\le \frac{h_{\max}}{\ln n}\left(\log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
&\frac{\mathcal{L}^{\mathrm{G}}(u_1^n) - I\log_2 e}{n} \\
&\quad\le \frac{h_{\max}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right).
\end{aligned}$$

Proof: We have that $I \le n \cdot h_{\max}$. This fact and (14)–(16) imply the result. □

For a very probable collection of strings, we can further tighten the bound presented in Corollary 1. In particular, we have the following result.

Corollary 2: Assume that the source has positive entropy. With probability $1 - O(\frac{1}{\sqrt{n}})$,

$$\begin{aligned}
&\frac{\mathcal{L}^{\mathrm{LZ}}(U_1^n) - I(U_1^n \mid s_0)\log_2 e}{n} \\
&\quad\le \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
&\frac{\mathcal{L}^{\mathrm{W}}(U_1^n) - I(U_1^n \mid s_0)\log_2 e}{n} \\
&\quad\le \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
&\frac{\mathcal{L}^{\mathrm{G}}(U_1^n) - I(U_1^n \mid s_0)\log_2 e}{n} \\
&\quad\le \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right).
\end{aligned}$$

Proof: Since

$$\lim_{n\to\infty} E(I(U_1^n \mid s_0)/n) = \mathcal{H}$$

and the variance and third moment of the self-information of every source symbol emitted is finite, it follows from [21] that

$$\mathcal{H} - o\left(\frac{1}{\ln n}\right) \le \frac{I(U_1^n \mid s_0)}{n} \le \mathcal{H} + o\left(\frac{1}{\ln n}\right) \quad \text{with probability } 1 - O\left(\frac{1}{\sqrt{n}}\right). \tag{46}$$

The corollary follows from (46) and (14)–(16). □

It is also easy to derive upper bounds on the redundancy of the codes using Theorem 1 and (46). We have the following result.

Theorem 2: Assume the source has positive entropy. Then

$$\begin{aligned}
\mathcal{R}^{\mathrm{LZ}} &\le \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right) \\
\mathcal{R}^{\mathrm{W}} &\le \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right)
\end{aligned}$$

and

$$\mathcal{R}^{\mathrm{G}} \le \frac{\mathcal{H}}{\ln n}\left(\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right)\right) + o\left(\frac{1}{\ln n}\right).$$

Proof: Taking the expected value of both sides of (14)–(16) with respect to U_1^n, we see that

$$\begin{aligned}
&\ln n \cdot E\left(\frac{\mathcal{L}^{\mathrm{LZ}}(U_1^n) - I(U_1^n \mid s_0)\log_2 e}{n}\right) \\
&\quad\le E\left(\frac{I(U_1^n \mid s_0)}{n}\right)\cdot\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right) + o(1)
\end{aligned} \tag{47}$$

$$\begin{aligned}
&\ln n \cdot E\left(\frac{\mathcal{L}^{\mathrm{W}}(U_1^n) - I(U_1^n \mid s_0)\log_2 e}{n}\right) \\
&\quad\le E\left(\frac{I(U_1^n \mid s_0)}{n}\right)\cdot\log_2\left(\frac{RK\log_2 e}{\mathcal{H}}\right) + o(1)
\end{aligned} \tag{48}$$

$$\begin{aligned}
&\ln n \cdot E\left(\frac{\mathcal{L}^{\mathrm{G}}(U_1^n) - I(U_1^n \mid s_0)\log_2 e}{n}\right) \\
&\quad\le E\left(\frac{I(U_1^n \mid s_0)}{n}\right)\cdot\log_2\left(\frac{R(K-1)\log_2 e}{\mathcal{H}}\right) + o(1).
\end{aligned} \tag{49}$$

We have that for all strings u_1^n,

$$0 \le \frac{I(u_1^n \mid s_0)}{n} \le h_{\max}. \tag{50}$$

Since (46) holds with probability $1 - O(1/\sqrt{n})$ and (50) applies for the remaining cases, we have that

$$\mathcal{H} - o\left(\frac{1}{\ln n}\right) \le E\left(\frac{I(U_1^n \mid s_0)}{n}\right) \le \mathcal{H} + o\left(\frac{1}{\ln n}\right). \tag{51}$$

(47)–(49) and (5) imply the theorem. □

Thus far, we have been assuming that the source has positive entropy. To finish our analysis of these three encoding rules, we will investigate the performance of the algorithms on the output of a source with *zero* entropy. In this case, for every state s, there is a letter j_s such that

$$p_{s,j_s} = 1. \tag{52}$$

This implies that from any starting state s_0, there is only one possibility for the output from the source. We have the following result.

Lemma 5: Assume that the source has zero entropy. Let m be any positive integer. For each of the three encoding rules we are studying, the parsing of the source output u_1^n can result in at most R complete parsed phrases of length m.

Proof: For each algorithm, (52) implies that the m letters following a parsing point are uniquely determined by the state of the source at that parsing point. Therefore, since the source has R states, there are at most R distinct strings of length m that can appear as a parsed phrase. In the case of LZ'78, the lemma holds because all complete parsed phrases must be distinct. For LZW and G, let $v_1 \circ \cdots \circ v_{m+1}$ be the $m+1$ letters following a parsing point. If the parsed string following this parsing point is $v_1 \circ \cdots \circ v_m$, then $v_1 \circ \cdots \circ v_{m+1}$ is not in the current dictionary. However, if at any future point $v_1 \circ \cdots \circ v_{m+1}$ is a prefix of the unparsed source output, then the encoding rules for LZW and G guarantee that $v_1 \circ \cdots \circ v_{m+1}$ will be a prefix of the corresponding parsed string. Our earlier argument implies that there are at most R distinct possibilities for the string $v_1 \circ \cdots \circ v_{m+1}$. Therefore, there can be at most R complete parsed phrases of length m. □

From Lemma 5, we can obtain the following bound on the number of phrases associated with the parsing of u_1^n.

Lemma 6: Assume that the source has zero entropy. Let λ denote the smallest integer for which

$$\frac{R\lambda(\lambda+1)}{2} + \lambda + 1 > n.$$

Then for each of the encoding rules, an upper bound on the number of complete phrases c in the parsing of u_1^n is given by

$$c \le R\lambda. \tag{53}$$

Proof: We will establish (53) by contradiction. Suppose that (53) is false for one of the encoding rules. Then for that encoding rule, $c \ge R\lambda + 1$. Let l_i represent the length of phrase i. Without loss of generality, reorder the lengths so that $l_1 \le l_2 \le \cdots \le l_c$. From Lemma 5, we know that for every positive integer m, there can be at most R complete phrases of length m. Hence, for $R(m-1)+1 \le j \le Rm$, we have that $l_j \ge m$, and $l_{R\lambda+1} \ge \lambda + 1$. Thus

$$\begin{aligned} n \ge \sum_{i=1}^{c} l_i \ge \sum_{i=1}^{R\lambda+1} l_i &= l_{R\lambda+1} + \sum_{m=1}^{\lambda} \sum_{j=R(m-1)+1}^{Rm} l_j \\ &\ge (\lambda+1) + \sum_{m=1}^{\lambda} Rm \\ &= (\lambda+1) + \frac{R\lambda(\lambda+1)}{2} > n \end{aligned}$$

which is a contradiction. □

Using Lemma 6, we obtain the following result.

Theorem 3: Assume the source has zero entropy. For the three encoding rules we are considering, we have the following asymptotic bound on the length of the encoding of the source output string u_1^n:

$$\begin{aligned} \frac{\mathcal{L}^{\mathrm{LZ}}(u_1^n)}{n} \le \sqrt{\frac{R}{2n}} \log_2 n &+ \sqrt{\frac{2R}{n}} \\ &\times \log_2\left(\frac{\sqrt{2R}(K-1)\log_2 e}{e}\right) + O\left(\frac{\ln n}{n}\right) \end{aligned} \tag{54}$$

$$\begin{aligned} \frac{\mathcal{L}^{\mathrm{W}}(u_1^n)}{n} \le \sqrt{\frac{R}{2n}} \log_2 n &+ \sqrt{\frac{2R}{n}} \\ &\times \log_2\left(\frac{\sqrt{2R}\log_2 e}{e}\right) + O\left(\frac{\ln n}{n}\right) \end{aligned} \tag{55}$$

$$\begin{aligned} \frac{\mathcal{L}^{\mathrm{G}}(u_1^n)}{n} \le \sqrt{\frac{R}{2n}} \log_2 n &+ \sqrt{\frac{2R}{n}} \\ &\times \log_2\left(\frac{\sqrt{2R}\log_2 e}{e}\right) + O\left(\frac{\ln n}{n}\right). \end{aligned} \tag{56}$$

Proof: By Lemma 6, we have that for each of the encoding rules, the number of complete parsed phrases c satisfies

$$c \le \sqrt{2Rn}. \tag{57}$$

By substituting (57) into (5)–(7), we obtain (54)–(56), respectively. □

It is interesting to compare the redundancy of the code associated with the Lempel–Ziv incremental parsing rule with that of a different type of Lempel–Ziv code. LZ'77, i.e., the encoding rule developed by Ziv and Lempel in 1977 (see [12]), uses a greedy parsing scheme. Essentially, if the first n symbols u_1^n of the source output have been parsed, the next parsed string is the longest prefix σ of the unparsed source output that is of the form u_m^{m+l-1} for some $m \le n$. The string is encoded by using $\lceil \log_2 n \rceil$ bits to represent m and $\Theta(\log l)$ bits to represent l. For a fixed-database implementation of LZ'77, Wyner demonstrated in [22] that as n increases, the length l of the next parsed phrase is asymptotically normally distributed with mean $(\ln n)/\mathcal{H} + O(1)$ and variance $O(\ln n)$. This suggests that the redundancy of encoding the first n letters of the source output using LZ'77 is $\Theta((\ln \ln n)/(\ln n))$. Since LZ'78 and its variants can also be viewed as greedy procedures, it was conjectured that these schemes would also have redundancy $\Theta((\ln \ln n)/(\ln n))$. For LZ'78, LZW, and G, we upper-bounded the redundancies of the algorithms by minimizing the self-information per parsed phrase. The goal of LZ'77 is to maximize the self-information per phrase, but the resulting decrease in the number of phrases is more than offset by the size of the "dictionary" corresponding to the parsed string u_1^n, and this is why LZ'78 and its variants asymptotically outperform LZ'77. However, recent trends in practical lossless data compression suggest that LZ'77 and its variants perform as well as, if not better than, LZ'78 and its variants. There are a few explanations for this. For small to moderate values of n, the difference between $\Theta((\ln \ln n)/(\ln n))$ and $\Theta((\ln n)^{-1})$ is not significant and an understanding of the lower order terms is very important in order to make a fair comparison. The other limitations to our analysis are the assumptions that the source statistics do not change over time and that the dictionary of source strings can grow arbitrarily large. If either of these assumptions are violated, then the situation may be very different.

III. Bound on Pointwise Code Length

The results of the last section and earlier analyses of the compression achieved by the Lempel–Ziv codes assume a model for the source and bound the average number of code symbols per source symbol used by the code for a random output from this source. In practical situations, we would like to be able to bound the number of code symbols per source symbol needed for the encoding of a particular string u_1^n. The appropriate way to modify the analysis carried out in the last section is to select a source model and then choose the parameters of this model to maximize the likelihood that u_1^n is emitted. For example, given a positive integer l, one standard source model assumes that each output depends statistically only on the l previous output symbols. As usual, let us suppose the source letters come from a finite alphabet $\{0, 1, \cdots, K-1\}$. We will apply a model for the source in which there are a set of states $\{0, 1, \cdots, R-1\}$ with an initial state s_0 and $S[s, j]$ defines the next state if symbol j is emitted from state s. In order to complete the definition of the model, we need to specify $\theta_{s,j}$, the probability that the source emits symbol j from state s. Let $\hat{p}_{s,j}$ and $\hat{\pi}_s$ be the empirical probability that j is emitted from state s and the empirical probability that the source is in state s, respectively. That is,

$$\hat{p}_{s,j} = \frac{\text{number of times } j \text{ is emitted from state } s \text{ in } u_1^n}{\text{number of times the source is in state } s \text{ in } u_1^n}$$

and

$$\hat{\pi}_s = \frac{\text{number of times the source is in state } s \text{ in } u_1^n}{n}.$$

The empirical entropy $\hat{\mathcal{H}}_n$ for this model is given by

$$\hat{\mathcal{H}}_n = -\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \ln \hat{p}_{s,j}$$

and the self-information $I(u_1^n \mid s_0)$ of u_1^n assuming a model with the probabilities $\{\theta_{s,j}\}$ is

$$I(u_1^n \mid s_0) = n\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \ln\left(\frac{1}{\theta_{s,j}}\right).$$

Lemma 7: The choice of the probabilities $\theta_{s,j}$ to minimize the self-information $I(u_1^n \mid s_0)$ of u_1^n is $\theta_{s,j} = \hat{p}_{s,j}$ for all states s and symbols j.

Proof: The problem of minimizing $I(u_1^n \mid s_0)$ is equivalent to selecting the $\theta_{s,j}$ to minimize the *empirical divergence* $\hat{D}_n$ defined by

$$\hat{D}_n \doteq -\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \ln\left(\frac{\theta_{s,j}}{\hat{p}_{s,j}}\right).$$

Using the inequality that for all $x > 0, \ln x \le x - 1$ with equality if and only if $x = 1$, we have that

$$\hat{D}_n \ge -\sum_{s=0}^{R-1}\sum_{j=0}^{K-1} \hat{\pi}_s \hat{p}_{s,j} \left(\frac{\theta_{s,j}}{\hat{p}_{s,j}} - 1\right)$$
$$= 0$$

and $\hat{D}_n = 0$ if and only if $\theta_{s,j} = \hat{p}_{s,j}$ for all states s and symbols j. □

Let $\hat{I}(u_1^n \mid s_0)$ be the self-information of the string assuming the empirical model; i.e., when $\theta_{s,j} = \hat{p}_{s,j}$ for all states s and symbols j. Note that Lemma 7 implies that for this model of the source, we have that

$$\hat{I}(u_1^n \mid s_0) = n\hat{\mathcal{H}}_n. \tag{58}$$

The analysis on individual sequences carried out in the last section applies for u_1^n assuming the empirical model of the source. We have the following results.

Theorem 4: If $\hat{\mathcal{H}}_n$ is positive, then

$$\left(\frac{1}{n}\mathcal{L}^{\mathrm{LZ}}(u_1^n) - \hat{\mathcal{H}}_n \log_2 e\right) \cdot \ln\left(\frac{\hat{\mathcal{H}}_n^2 n}{R}\right) \le \hat{\mathcal{H}}_n\left(\log_2\left(\frac{R(K-1)\log_2 e}{\hat{\mathcal{H}}_n}\right)\right) + o(1) \tag{59}$$

$$\left(\frac{1}{n}\mathcal{L}^{\mathrm{W}}(u_1^n) - \hat{\mathcal{H}}_n \log_2 e\right) \cdot \ln\left(\frac{\hat{\mathcal{H}}_n^2 n}{RK}\right) \le \hat{\mathcal{H}}_n\left(\log_2\left(\frac{RK\log_2 e}{\hat{\mathcal{H}}_n}\right)\right) + o(1) \tag{60}$$

$$\left(\frac{1}{n}\mathcal{L}^{\mathrm{G}}(u_1^n) - \hat{\mathcal{H}}_n \log_2 e\right) \cdot \ln\left(\frac{\hat{\mathcal{H}}_n^2 n}{R(K-1)}\right) \le \hat{\mathcal{H}}_n\left(\log_2\left(\frac{R(K-1)\log_2 e}{\hat{\mathcal{H}}_n}\right)\right) + o(1). \tag{61}$$

If $\hat{\mathcal{H}}_n$ is zero, then

$$\frac{\mathcal{L}^{\mathrm{LZ}}(u_1^n)}{n} \le \sqrt{\frac{R}{2n}}\log_2 n + \sqrt{\frac{2R}{n}} \times \log_2\left(\frac{\sqrt{2R}(K-1)\log_2 e}{e}\right) + O\left(\frac{\ln n}{n}\right) \tag{62}$$

$$\frac{\mathcal{L}^{\mathrm{W}}(u_1^n)}{n} \le \sqrt{\frac{R}{2n}}\log_2 n + \sqrt{\frac{2R}{n}} \times \log_2\left(\frac{\sqrt{2R}\log_2 e}{e}\right) + O\left(\frac{\ln n}{n}\right) \tag{63}$$

$$\frac{\mathcal{L}^{\mathrm{G}}(u_1^n)}{n} \le \sqrt{\frac{R}{2n}}\log_2 n + \sqrt{\frac{2R}{n}} \times \log_2\left(\frac{\sqrt{2R}\log_2 e}{e}\right) + O\left(\frac{\ln n}{n}\right). \tag{64}$$

Proof: We will demonstrate (59) and (62). Expression (59) follows from (39), (40), (5), and (58). Expressions (57), (5), and the fact that $R = O(n)$ imply (62). The proofs of (60), (61), (63), and (64) are similar to the proofs of (59) and (62). □

IV. Conclusion

We have investigated the redundancy of LZ'78, LZW, and G as the length n of the source string encoded tends toward infinity and we established that for each algorithm, the redundancy decreases toward zero as $\Theta((\ln n)^{-1})$. Moreover, we upper-bounded the number of code symbols per source symbol needed for the encoding of a particular string u_1^n. The main idea was to upper-bound the number of parsed phrases used to encode u_1^n by the maximum number of phrases with cumulative self-information less than $I(u_1^n \mid s_0)$ and to use renewal theory to understand the asymptotic behavior of the second quantity.

Appendix I

Proof of Lemma 1: Let $\epsilon = x - \lfloor x \rfloor$. Then $0 \le \epsilon < 1$ and

$$\sum_{i=1}^{k} \lceil x + \log_2 i \rceil = k\lfloor x \rfloor + \sum_{i=1}^{k} \lceil \epsilon + \log_2 i \rceil$$

so it is sufficient to prove the lemma assuming $0 \le x < 1$. Let $z = \lfloor \log_2(2^x k) \rfloor$ and suppose that l is the largest integer for which $z = \lceil \log_2(2^x l) \rceil$; i.e., $2^x l \le 2^z < 2^x(l+1)$, and so $l \le 2^{z-x} < l+1$. We have that

$$\begin{aligned}\sum_{i=1}^{k} \lceil x + \log_2 i \rceil &= \sum_{i=1}^{k} \lceil \log_2 (2^x i) \rceil \\ &= \sum_{j=1}^{z} j \cdot |\text{integers } i\colon 2^{j-1} < 2^x i \le 2^j| \\ &\quad + (k-l)(z+1) \\ &= \sum_{j=1}^{z} j \cdot |\text{integers } i\colon 2^{j-x-1} < i \le 2^{j-x}| \\ &\quad + (k-l)(z+1) \\ &\le \sum_{j=1}^{z} j \cdot 2^{j-x-1} + (k - 2^{z-x} + 1)(z+1) \\ &= 2^{z-x}(z-1) + 2^{-x} \\ &\quad + (k - 2^{z-x})(z+1) + z + 1 \\ &= kz + k - 2 \cdot 2^{z-x} + O(z) \qquad (65)\end{aligned}$$

which is equivalent to the first inequality. Let $y = 2^{z-x}$. Since $\log_2 y = \lfloor \log_2(2^x k) \rfloor - x$ we have that $k/2 < y \le k$. The expression $k \log_2 y - 2y$ is maximized at $y = k(\log_2 e)/2$, and for this value of y, we find that

$$z = \log_2 k + x + \log_2(\log_2 e) - 1.$$

Substituting this value of z into (65), we obtain the second inequality. □

Appendix II

In order to prove Lemma 3, we employ the following well-known renewal theorem (see [19, Secs. 3.4 and 3.5]).

Theorem 5: For all states s and ψ, if $J_k^{(\psi)}$, $k > 1$ has a nonarithmetic distribution and if $h(t)$ is directly Riemann integrable, then

$$\lim_{t \to \infty} \int_0^t h(t-x)\, dm^{(\psi,s)}(x) = \frac{1}{E[J_2^{(\psi)}]} \int_0^\infty h(t)\, dt.$$

We have the following relationship among $E[J_2^{(\psi)}]$, $\mathcal{H}$, and π_ψ.

Lemma 8: $E[J_2^{(\psi)}] = \mathcal{H}/\pi_\psi$, $\psi \in \{0, \cdots, R-1\}$.

Proof of Lemma 8: Let $\mathcal{H}(\psi)$ represent the entropy in natural units of the next source symbol, given that the source is in state ψ; i.e.,

$$\mathcal{H}(\psi) = -\sum_{j=0}^{K-1} p_{\psi,j} \ln p_{\psi,j}, \psi \in \{0, \cdots, R-1\}.$$

Just as we can define a renewal process where the inter-renewal variable is the self-information generated by the source, we can view the process by which the source generates self-information as a semi-Markov process (see, e.g., [19, Sec 4.8]) with the properties that whenever the source enters state ψ:

1) The next state it will enter is state r with probability $f_{\psi,r}$.
2) Given that the next symbol to be emitted is letter j, the amount of self-information generated until the transition from ψ to $S[\psi, j]$ is $-\ln p_{\psi,j}$.

Note that $\mathcal{H}(\psi)$ is the mean information growth produced by this semi-Markov process from its entrance into state ψ until it makes a transition. $E[J_2^{(\psi)}]$ can now be interpreted as the information growth between successive transitions into state ψ. To complete the proof, we note that the long run proportion of self-information generated while the process is in state ψ can be shown (see [19, Sec 4.8]) to be equal to both $\mathcal{H}(\psi)/E[J_2^{(\psi)}]$ and

$$\pi_\psi \mathcal{H}(\psi) / \left(\sum_{r=0}^{R-1} \pi_r \mathcal{H}(r) \right).$$

To finish the proof, note that

$$\mathcal{H} = \sum_{r=0}^{R-1} \pi_r \mathcal{H}(r). \qquad \square$$

Proof of Lemma 3: For (26), we let $h(t) = e^{-t}$ and for the other equations, we let $h(t) = -te^{-t}$. The lemma follows from Theorem 5 and Lemma 8. □

Appendix III

Throughout this appendix, we assume the self-information associated with the source symbols emitted has an arithmetic distribution with period Λ. Furthermore, we presume that for all pairs of states ψ and symbols j, $p_{\psi,j} < 1$. If necessary, it is possible to change the alphabet and set of states in order to satisfy this assumption.

In the results that follow, we use a stronger version of Lemmas 2 and 4 that more accurately relates the self-information of a string to the number of phrases associated with the parsing of that string by more carefully counting the number of strings at the threshhold self-information. Let

$$\hat{\gamma}(\tau) \doteq |\{(s, \sigma)\colon I(\sigma \mid s) < \tau\}|.$$

We have the following result.

Lemma 9: The number of phrases in the parsing of u_1^n satisfies

$$c^{\mathrm{LZ}} \le \rho^{\mathrm{LZ}} \doteq \hat{\gamma}(\bar{\tau}) + \alpha_{\mathrm{LZ}} \tag{66}$$

$$c^{\mathrm{W}} \le \rho^{\mathrm{W}} \doteq K\hat{\gamma}(\tau^{\mathrm{W}}) + \alpha_{\mathrm{W}} \tag{67}$$

and

$$c^{\mathrm{G}} \le \rho^{\mathrm{G}} \doteq (K-1)\hat{\gamma}(\tau^{\mathrm{G}}) + \alpha_{\mathrm{G}} \tag{68}$$

where $\bar{\tau}$, τ^{W}, τ^{G}, α_{LZ}, α_{W}, and α_{G} are chosen so that

$$J_{\mathrm{LZ}} \doteq \sum_{\tau\in\Omega:\tau<\bar{\tau}} \tau\cdot\mu(\tau) + \bar{\tau}\cdot\alpha_{\mathrm{LZ}} \ge I > J_{\mathrm{LZ}} - \bar{\tau} \tag{69}$$

$$J_{\mathrm{W}} \doteq K \sum_{\tau\in\Omega:\tau<\tau^{\mathrm{W}}} \tau\cdot\mu(\tau) + \tau^{\mathrm{W}}\cdot\alpha_{\mathrm{W}} \ge I > J_{\mathrm{W}} - \tau^{\mathrm{W}} \tag{70}$$

$$J_{\mathrm{G}} \doteq (K-1) \sum_{\tau\in\Omega:\tau<\tau^{\mathrm{G}}} \tau\cdot\mu(\tau) + \tau^{\mathrm{G}}\cdot\alpha_{\mathrm{G}} \ge I > J_{\mathrm{W}} - \tau^{\mathrm{G}} \tag{71}$$

$$1 \le \alpha_{\mathrm{LZ}} \le \mu(\bar{\tau}) \tag{72}$$

$$1 \le \alpha_{\mathrm{W}} \le K\mu(\tau^{W}) \tag{73}$$

$$1 \le \alpha_{\mathrm{G}} \le (K-1)\mu\gamma_s(\tau^{G}). \tag{74}$$

Proof: The proof of Lemma 9 is identical to the proofs of Lemmas 2 and 4. □

We first consider sources that are *acyclic* in the sense that there is no integer D greater than one for which the self-information generated by the source between successive occurrences of state ψ is an integer multiple of $D\Lambda$ for all ψ. We use Blackwell's theorem (see [19, Secs 3.4 and 3.5]).

Theorem 6: If $J_k^{(\psi)}, k \ge 1$ has an arithmetic distribution with period Λ, then

$$\lim_{m\to\infty} E[\text{number of renewals at } m\Lambda] = \frac{\Lambda}{E[J_2^{(\psi)}]}.$$

Since $p_{\psi,j} < 1$ for all pairs of states ψ and symbols j, at time $m\Lambda$, there will either be no renewal or one renewal. From Lemma 8, $E[J_2^{(\psi)}] = \mathcal{H}/\pi_\psi$ (see [20, Sec 5.6]). Hence, it follows from Theorem 6 and Lemma 8 that as m increases, the probability of a renewal at $m\Lambda$ (for the process associated with returns to state ψ) is $\pi_\psi\Lambda/\mathcal{H}$. For all states ψ and starting states s

$$\begin{aligned}\frac{\pi_\psi\Lambda}{\mathcal{H}} &= \lim_{m\to\infty} \mathrm{Prob}\{\sigma\colon \mathrm{I}(\sigma \mid \mathrm{s}) = \mathrm{m}\Lambda \text{ and } \mathrm{S}[\mathrm{s},\sigma] = \mathrm{i}\}\\ &= \lim_{m\to\infty} \sum_{\sigma\colon I(\sigma\mid s)=m\Lambda, S[s,\sigma]=\psi} P(\sigma\mid s)\\ &= \lim_{m\to\infty} e^{-m\Lambda}|\{\sigma\colon I(\sigma\mid s) = m\Lambda \text{ and } S[s,\sigma]=\psi\}|.\end{aligned} \tag{75}$$

Consequently, we have the following counterpart to (26)–(28) for acyclic, arithmetic sources.

Lemma 10: As m increases, the number of ordered pairs of states s and strings σ with self-information $m\Lambda$ starting from state s satisfies

$$\lim_{m\to\infty} e^{-m\Lambda}|\{(s,\sigma)\colon I(\sigma\mid s) = m\Lambda\}| = \frac{\Lambda R}{\mathcal{H}}. \tag{76}$$

Hence

$$\lim_{m\to\infty} e^{-m\Lambda}|\{(s,\sigma)\colon I(\sigma\mid s) < m\Lambda\}| = \frac{\Lambda R}{\mathcal{H}(e^\Lambda - 1)} \tag{77}$$

$$\lim_{m\to\infty} e^{-m\Lambda}|\{(s,\sigma)\colon I(\sigma\mid s) \le m\Lambda\}| = \frac{\Lambda R e^\Lambda}{\mathcal{H}(e^\Lambda - 1)} \tag{78}$$

$$\lim_{m\to\infty} e^{-m\Lambda}\sum_{j=1}^{m} j\Lambda|\{(s,\sigma)\colon I(\sigma\mid s) = (m-j)\Lambda\}| = \frac{\Lambda^2 R e^\Lambda}{\mathcal{H}(e^\Lambda - 1)^2} \tag{79}$$

$$\lim_{m\to\infty} e^{-m\Lambda}\sum_{j=0}^{m} j\Lambda|\{(s,\sigma)\colon I(\sigma\mid s) = (m-j)\Lambda\}| = \frac{\Lambda^2 R e^\Lambda}{\mathcal{H}(e^\Lambda - 1)^2}. \tag{80}$$

Proof: Equation (76) follows by summing both sides of (75) over s and ψ. To demonstrate (77), note that

$$\begin{aligned}&e^{-m\Lambda}|\{(s,\sigma)\colon I(\sigma\mid s) < m\Lambda\}|\\ &\quad= \sum_{j=1}^{m} e^{-j\Lambda}\left(e^{-(m-j)\Lambda}|\{(s,\sigma)\colon I(\sigma\mid s) = (m-j)\Lambda\}|\right)\end{aligned} \tag{81}$$

and (76) implies that

$$\lim_{m\to\infty} e^{-(m-j)\Lambda}|\{(s,\sigma)\colon I(\sigma\mid s) = (m-j)\Lambda\}| = \frac{\Lambda R}{\mathcal{H}}. \tag{82}$$

Equation (77) follows from (81) and (82). We obtain (78) by adding together (76) and (77). Equations (79) and (80) follow from (82) and the fact that

$$\lim_{m\to\infty}\sum_{j=0}^{m} j\Lambda e^{-j\Lambda}\cdot\frac{\Lambda R}{\mathcal{H}} = \lim_{m\to\infty}\sum_{j=1}^{m} j\Lambda e^{-j\Lambda}\cdot\frac{\Lambda R}{\mathcal{H}} = \frac{\Lambda^2 R e^\Lambda}{\mathcal{H}(e^\Lambda - 1)^2}.$$

□

Using Lemmas 9 and 10, we have the following analog to (14)–(16).

Theorem 7: Assume that the source is arithmetic with period Λ and acyclic. As usual, we abbreviate $I(u_1^n\mid s_0)$ by I. Let

$$\mathcal{C} \doteq (\max\{0, \Lambda - 1\} + (e^\Lambda - 1)^{-1})\cdot\log_2 e.$$

For the three encoding rules we are considering, we have the following asymptotic relationships between the number of binary digits used to represent u_1^n and the self-information of u_1^n:

$$\begin{aligned}&\ln n\cdot\left(\frac{\mathcal{L}^{\mathrm{LZ}}(u_1^n) - I\log_2 e}{n}\right)\\ &\quad\le \frac{I}{n}\left(\log_2\left(\frac{\Lambda R(K-1)\log_2 e}{\mathcal{H}(e^\Lambda - 1)}\right) + \mathcal{C}\right) + o(1)\end{aligned} \tag{83}$$

$$\ln n \cdot \left(\frac{\mathcal{L}^{\mathrm{W}}(u_1^n) - I\log_2 e}{n}\right) \le \frac{I}{n}\left(\log_2\left(\frac{\Lambda RK\log_2 e}{\mathcal{H}(e^{\Lambda}-1)}\right) + \mathcal{C}\right) + o(1) \tag{84}$$

$$\ln n \cdot \left(\frac{\mathcal{L}^{\mathrm{G}}(u_1^n) - I\log_2 e}{n}\right) \le \frac{I}{n}\left(\log_2\left(\frac{\Lambda R(K-1)\log_2 e}{\mathcal{H}(e^{\Lambda}-1)}\right) + \mathcal{C}\right) + o(1). \tag{85}$$

Proof: As with the proof of Theorem 1, we will focus on obtaining the result for the Lempel–Ziv incremental parsing rule, and then slightly modify the analysis for LZW and G. Define β_{LZ} by

$$\beta_{\mathrm{LZ}} \doteq \rho^{\mathrm{LZ}} e^{-\bar{\tau}}. \tag{86}$$

Expressions (66), (72), (77), and (78) imply that as $\bar{\tau}$ increases,

$$\frac{\Lambda R}{\mathcal{H}(e^{\Lambda}-1)} + o(1) \le \beta_{\mathrm{LZ}} \le \frac{\Lambda R e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)} + o(1). \tag{87}$$

From (69), (79), and (80), we find that

$$\left(\bar{\tau}\rho^{\mathrm{LZ}} - J_{\mathrm{LZ}}\right) \cdot e^{-\bar{\tau}} = \frac{\Lambda^2 R e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)^2} + o(1)$$

and since $J_{\mathrm{LZ}} - \bar{\tau} < I \le J_{LZ}$, we have that

$$\left(\bar{\tau}\rho^{\mathrm{LZ}} - I\right) \cdot e^{-\bar{\tau}} = \frac{\Lambda^2 R e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)^2} + o(1). \tag{88}$$

Taking the logarithm of both sides of (86), we see that as $\bar{\tau}$ increases,

$$\log_2 \rho^{\mathrm{LZ}} - \bar{\tau}\log_2 e = \log_2 \beta_{\mathrm{LZ}}. \tag{89}$$

Multiplying (86) by (89), we obtain

$$\left(\rho^{\mathrm{LZ}}\log_2\rho^{\mathrm{LZ}} - \bar{\tau}\rho^{\mathrm{LZ}}\log_2 e\right) \cdot e^{-\bar{\tau}} = \beta_{\mathrm{LZ}}\log_2\beta_{\mathrm{LZ}}. \tag{90}$$

Multiplying both sides of (88) by $\log_2 e$ and adding the resulting expression to (90), we find that as $\bar{\tau}$ increases

$$\left(\rho^{\mathrm{LZ}}\log_2\rho^{\mathrm{LZ}} - I\log_2 e\right) \cdot e^{-\bar{\tau}} = \beta_{\mathrm{LZ}}\log_2\beta_{\mathrm{LZ}} + \frac{\Lambda^2 R e^{\Lambda}\log_2 e}{\mathcal{H}(e^{\Lambda}-1)^2} + o(1). \tag{91}$$

To determine the asymptotic relationship among $e^{-\bar{\tau}}$, I, and β_{LZ}, we substitute (86) into (88) and deduce

$$\bar{\tau} \cdot \beta_{\mathrm{LZ}} + O(1) - I \cdot e^{-\bar{\tau}} = 0. \tag{92}$$

Hence, as $\bar{\tau}$ and I increase,

$$\bar{\tau} = \left(\ln\left(\frac{I}{\beta_{\mathrm{LZ}}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)}\right)\right)(1+o(1)) \tag{93}$$

and

$$e^{-\bar{\tau}} = \left(\frac{\beta_{\mathrm{LZ}}}{I}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)\right)(1+o(1)). \tag{94}$$

Substituting (94) into (86) and (88), we find that

$$\rho^{\mathrm{LZ}} \cdot \left(\frac{\beta_{\mathrm{LZ}}}{I}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)\right)(1+o(1)) = \beta_{\mathrm{LZ}} \tag{95}$$

and

$$\left(\rho^{\mathrm{LZ}}\log_2\rho^{\mathrm{LZ}} - I\log_2 e\right) \cdot \left(\frac{\beta_{\mathrm{LZ}}}{I}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)\right)(1+o(1)) = \beta_{\mathrm{LZ}}\log_2\beta_{\mathrm{LZ}} + \frac{\Lambda^2 R e^{\Lambda}\log_2 e}{\mathcal{H}(e^{\Lambda}-1)^2} + o(1). \tag{96}$$

By Lemma 9, $c^{\mathrm{LZ}} \le \rho^{\mathrm{LZ}}$. Therefore, (95) and (96) imply

$$c^{\mathrm{LZ}} \cdot \left(\frac{1}{I}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)\right)(1+o(1)) \le 1 \tag{97}$$

and

$$\left(c^{\mathrm{LZ}}\log_2 c^{\mathrm{LZ}} - I\log_2 e\right) \cdot \left(\frac{1}{I}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)\right)(1+o(1)) \le \log_2\beta_{\mathrm{LZ}} + \frac{\Lambda^2 R e^{\Lambda}\log_2 e}{\beta_{\mathrm{LZ}}\mathcal{H}(e^{\Lambda}-1)^2} + o(1). \tag{98}$$

From (5), (97), and (98), it follows that

$$\left(\mathcal{L}^{\mathrm{LZ}}(u_1^N) - I\log_2 e\right) \cdot \left(\frac{1}{I}\ln\left(\frac{I}{\beta_{\mathrm{LZ}}}\right)\right) \le \log_2\left(\frac{\beta_{\mathrm{LZ}}(K-1)\log_2 e}{e}\right) + \frac{\Lambda^2 R e^{\Lambda}\log_2 e}{\beta_{\mathrm{LZ}}\mathcal{H}(e^{\Lambda}-1)^2} + o(1). \tag{99}$$

By maximizing the right-hand side of (99) over the range of β_{LZ} given in (87), we find that the maximum occurs at one of the endpoints of the interval. We have that

$$\left(\mathcal{L}^{\mathrm{LZ}}(u_1^N) - I\log_2 e\right) \cdot \left(\frac{1}{I}\ln\left(\frac{I}{\beta_{LZ}}\right)\right) \le \log_2\left(\frac{\Lambda R(K-1)\log_2 e}{\mathcal{H}(e^{\Lambda}-1)}\right) + \mathcal{C} + o(1)$$

which is equivalent to (83) because $I = \Theta(n)$.

For LZW, the analogs to (86)–(88) are

$$\beta_{\mathrm{W}} = \rho^{\mathrm{W}}\exp(-\tau^{\mathrm{W}})$$

$$\frac{\Lambda RK}{\mathcal{H}(e^{\Lambda}-1)} + o(1) \le \beta_{\mathrm{W}} \le \frac{\Lambda RK e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)} + o(1)$$

$$\left(\tau^{\mathrm{W}}\rho^{\mathrm{W}} - I\right) \cdot e^{-\tau^{\mathrm{W}}} = \frac{\Lambda^2 RK e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)^2} + o(1)$$

respectively, and for G, we have that

$$\beta_{\mathrm{G}} = \rho^{\mathrm{G}}\exp(-\tau^{\mathrm{G}})$$

$$\frac{\Lambda R(K-1)}{\mathcal{H}(e^{\Lambda}-1)} + o(1) \le \beta_{\mathrm{G}} \le \frac{\Lambda R(K-1)e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)} + o(1)$$

$$\left(\tau^{\mathrm{G}}\rho^{\mathrm{G}} - I\right) \cdot e^{-\tau^{\mathrm{G}}} = \frac{\Lambda^2 R(K-1)e^{\Lambda}}{\mathcal{H}(e^{\Lambda}-1)^2} + o(1).$$

With these modifications, we can use the steps in the derivation of (83) to prove (84) and (85). □

Given Theorem 7, it is straightforward to find counterparts for the remaining results in Sections II and III. We omit the details here.

Finally, we briefly consider the situation in which the set of source states is *cyclic*. Unfortunately, the results are substantially more complicated and less insightful than they are for the acyclic, arithmetic case. Since (75) is the foundation for Theorem 7, we will limit our discussion to finding an analog for (75) when the set of source states is cyclic. To further specify our cyclic source, let D be the maximum integer for which the self-information generated by the source between consecutive occurrences of any given state is an integer multiple of $D\Lambda$. As an example, we consider the following:

Example: Suppose the source has states $\{0, 1, 2\}$ and always emits a symbol corresponding to the next state. Let

$$\boldsymbol{F} = \begin{pmatrix} 0.25 & 0.25 & 0.5 \\ 0.25 & 0.25 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}. \tag{100}$$

It is straightforward to demonstrate that this source is arithmetic with period $\ln 2$ and the self-information generated between consecutive occurrences of any given state is an integer multiple of $2 \ln 2$.

For each starting state s, the set of states can be partitioned into $1 < d \leq D$ categories with the property that state ψ is in category c relative to s, denoted $\mathcal{C}(c \,|\, s)$, where $c \in \{0, \cdots, D-1\}$, if and only if the (possibly delayed) renewal process for state ψ has renewals at epochs of the form $mD\Lambda + c\Lambda$. The counterpart to Theorem 6 is

Theorem 8: If state ψ is in category c relative to the starting state, and if $J_k^{(\psi)}$, $k > 1$ has an arithmetic distribution with period $D\Lambda$, then for $c' \in \{0, \cdots, D-1\}$

$$\lim_{m\to\infty} E[\text{number of renewals at } mD\Lambda + c'\Lambda] = \begin{cases} \frac{D\Lambda}{E\left[J_2^{(\psi)}\right]}, & c' = c \\ 0, & c' \neq c. \end{cases} \tag{101}$$

Since Lemma 8 continues to be valid, the analog to (75) is

$$\begin{aligned} \frac{\pi_\psi D\Lambda}{\mathcal{H}} &= \lim_{m\to\infty} \mathrm{Prob}\{\sigma \colon \mathrm{I}(\sigma \mid \mathrm{s}) = \mathrm{mD}\Lambda + \mathrm{c}\Lambda \text{ and} \\ &\qquad S[s,\sigma] = \psi \in \mathcal{C}(c \mid s)\} \\ &= \lim_{m\to\infty} \sum_{\sigma: I(\sigma|s)=mD\Lambda+c\Lambda,\ S[s,\sigma]=\psi\in\mathcal{C}(c|s)} P(\sigma \mid s) \\ &= \lim_{m\to\infty} e^{-mD\Lambda - c\Lambda} |\{\sigma \colon I(\sigma \,|\, s) = mD\Lambda + c\Lambda \text{ and} \\ &\qquad S[s,\sigma] = \psi \in \mathcal{C}(c \mid s)\}|. \end{aligned} \tag{102}$$

As we indicated earlier, it is possible to use (102) find the counterpart of Theorem 7.

ACKNOWLEDGMENT

This work was derived from a chapter of [8], which was supervised by Robert G. Gallager, who suggested using renewal theory in analyzing source codes. The author is grateful to Prof. Gallager and M. Feder for many helpful discussions and comments.

REFERENCES

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

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

[3] R. Ash, *Information Theory* New York: Wiley, 1965.

[4] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, pp. 66–72, 1992.

[5] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[6] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. IRE*, vol. 40, pp. 1098–1101, 1952.

[7] B. P. Tunstall, "Synthesis of noiseless compression codes," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1967.

[8] S. A. Savari, "Variable-to-fixed length codes for sources with known and unknown memory," Ph.D. dissertation, MIT, Cambridge, MA, 1996.

[9] S. A. Savari and R. G. Gallager, "Arithmetic coding for finite-state noiseless channels," *IEEE Trans. Inform. Theory*, vol. 40, pp. 100–107, 1994.

[10] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 668–674, 1978.

[11] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, 1986.

[12] J. Ziv and A. Lempel, "A universal algorithm for data compression," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 337–343, 1977.

[13] T. A. Welch, "A technique for high-performance data compression," *IEEE Computer*, vol. 17, no. 6, pp. 8–19, 1984.

[14] V. S. Miller and M. N. Wegman, "Variations on a theme by Ziv and Lempel," in *Combinatorial Algorithms on Words*, vol. F12, (NATO ASI Series), A. Apostolico and Z. Galil, Eds. Berlin, Germany: Springer-Verlag, 1985, pp. 131–140.

[15] R. G. Gallager, personal communication.

[16] G. Louchard and W. Szpankowski, "On the average redundancy rate of the Lempel-Ziv code," in *Proc. 1996 IEEE Data Compression Conf.* (Snowbird, UT, 1996).

[17] P. Jacquet and W. Szpankowski, "Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees," *Theor. Comput. Sci.*, vol. 144, pp. 161–197, 1995.

[18] D. E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, vol. 1. Reading, MA: Addison-Wesley, 1973.

[19] S. M. Ross, *Stochastic Processes*. New York: Wiley, 1983.

[20] R. G. Gallager, *Discrete Stochastic Processes*. Boston, MA: Kluwer, 1996.

[21] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*. Cambridge, MA: Addison-Wesley, 1968.

[22] A. J. Wyner, "String matching theorems and applications to data compression and statistics," Ph.D. dissertation, Stanford University, Stanford, CA, 1993.