RC23813 (W0511-159) November 22, 2005
Mathematics

IBM Research Report

On Certain Pathwise Properties of the Sliding Window Lempel-Ziv Algorithm

Luis Alfonso Lastras-Montaño
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

On certain pathwise properties of the sliding-window Lempel Ziv algorithm

Luis Alfonso Lastras-Montaño *

July 8, 2004

Abstract

We derive a number of almost sure results related to the sliding window Lempel-Ziv (SWLZ) algorithm. A principal result is a pathwise lower bound to the redundancy which is off by a factor of two from the main term in the lower bound of Wyner [1] and Yang and Kieffer [2], which hold in the expected sense for the Fixed Database Lempel Ziv algorithm, for different definitions of the compression ratio. Other results include a relatively weak upper bound on the redundancy and reasonably tight upper and lower bounds for the number of bits spent on the encoding of the phrase lengths. The main theme in deriving these results is a simple phrase length thresholding technique that traces its roots back to the work of Wyner and Ziv [3]; a significant aspect the present work can be viewed as exploring the extent to which said thresholding idea can give more detailed information about the properties of SWLZ. Another aspect of the present work studies the asymptotic behavior of the ratio of the number of phrases to the length of the parsed string for any finite sliding window size; in here we exploit results of Kieffer and Rahe [4]. In all cases it is assumed that the source is stationary and that in the most restrictive case it is an irreducible and aperiodic Markov chain; some of the results hold for sources that have exponential rates for entropy and more generally for the ergodic setting.

1 Introduction

The sliding window Lempel-Ziv (SWLZ) data compression algorithm was first proposed by Ziv and Lempel [5] in 1977. The optimality of the algorithm was established by Wyner and Ziv [3] who showed that as the window length approaches infinity, the *expected* compression ratio of

*Luis Alfonso Lastras-Montaño is with IBM TJ Watson Research Center, Yorktown Heights, NY 10598

the algorithm approaches the entropy of the source, under the assumption that this source is a stationary and ergodic random process. The results of Ornstein and Weiss [6] include as a corollary an almost sure optimality proof of a version of the algorithm whose database length grows as more data is parsed, under the same source assumptions. Savari [7] gives pathwise upper bounds on the redundancy of the same algorithm. In the work of Shields [8] we can find an almost sure convergence proof that holds under the general stationary ergodic setting; in [9] the same result was derived using different techniques for the more restricted class of sources that have exponential rates for entropy. A closely related algorithm is the Fixed Database Lempel Ziv (FDLZ) algorithm, which assumes that the database consists of a block of symbols that are statistically independent from the data being encoded but nevertheless share the same statistics with said data. This variant is attractive from the theoretical perspective since the independence assumption is often perceived to facilitate the analysis of the algorithm. Redundancy results for FDLZ can be found in the works of A.J. Wyner [1] and Yang and Kieffer [2]; principal ideas in the former precede the latter since they have appeared in [11] previously. These two works are of special interest since among other results, they contain lower bounds on redundancy of FDLZ that are essentially equal to the one offered here, although our bound is weaker in that it has an additional numerical factor of 1/2, and stronger in that it holds for every sample path whereas the earlier results hold in the expected sense. We also would like to point out that the works of A.J. Wyner and Yang et. al. differ in that the performance metric being studied is a ratio of expectations in the former whereas it is the expectation of a ratio in the latter. Note that in [1], Wyner argues that the two definitions in fact coincide for FDLZ, that they could differ for other algorithms and that furthermore, they both are of independent interest. Our work is in the almost sure domain and we simply define the bit rate as the ratio of the number of bits produced divided by the number of symbols parsed.

As previewed, a principal result in this work is a pathwise lower bound to the redundancy, which is given by

$$(h - o(1)) \left\{ \frac{1}{2} \frac{\log_2 \log_2 n_w}{\log_w n_w} + O\left(\frac{\log_2 \log_2 \log_2 n_w}{\log_2 \log_2 n_w} \right) \right\}$$

as $n_w \to \infty$, and holds for stationary aperiodic and irreducible Markov chains. Another significant result in this paper gives matching upper and lower bounds on the number of bits used in the encoding of the phrase length bits (Theorem 2) for sources that have exponential rates for entropy. We also explore some of the general properties of the various limits involved in this study. For example, we succeed in demonstrating that for sources that have exponential rates for entropy, the average number of phrases per symbol exists and is constant with probability one, a result that proved to be suprisingly elusive to obtain in spite of its apparent simplicity. In here, we make use of the results of Kieffer and Rahe [4] and an observation due to Vittorio Castelli [10] concerning the relation between Lempel-Ziv parsings of shifted versions of the same sequence. Possibly the weakest result in this paper is Theorem 1 which although to the knowledge of the author gives in conjunction with Theorem 2 the first upper bound on redundancy of this *particular* algorithm (SWLZ), fails to meet the strength of the bound of Savari [7] for an unbounded memory version of a similar algorithm or the upper bounds of Wyner [1] and Yang et.al. for FDLZ [2]. A.J. Wyner has informed the author that both our lower bound and upper bound have equivalent counterparts in his PhD thesis [11] for FDLZ, again the the sense of ratio of expectations.

To the knowledge of the author the results presented here are novel at least in that they are proved for a variant of LZ77 (with a sliding window) for which redundancy results had not been provided before. This variant is very natural for practical implementations, since it gives protection against changes in the statistics of the data (but note that the possibility of this phenomenon is not considered in the present work). To deal effectively with the sliding window in the almost sure domain, we found the need to introduce some interesting new ideas such as the ones used to prove Theorem 4. Additionally, we believe that there is value in some of the (apparently novel) approximations that we allow ourselves to make since the end results (e.g. the lower bound on redundancy or the computation of the phrase length bits) are competitive in quality with known results for other variants of LZ77.

This paper is organized as follows: in Section II we introduce the source models, the algorithm to be analyzed and the useful notion of the Lempel-Ziv shift operator T_{LZ}. In Section III we state our results. The proofs for the bounds are in Section IV; the fact that the pathwise average number of phrases per symbol exists and is constant almost surely is proved in Section V along with other more general results. Concluding remarks are offered in Section VI; all support material is relegated to the Appendix.

2 Preliminaries

Let $\mathcal{A}$ be a finite set, let $\mathcal{A}^l$ denote the l-th order cartesian product and let $\mathcal{A}_{-\infty}^{\infty}$ denote the doubly-infinite cartesian product. Let $\mathcal{A}^*$ be the set of all finite sequences of elements of $\mathcal{A}$, and let $\{0,1\}^*$ be the set of all finite binary sequences (we shall include the empty string in both definitions). Let

T (resp. T^{-1}) denote the left-shift (resp. right-shift) operator, i.e. if $\mathbf{x} \in \mathcal{A}_{-\infty}^{\infty}$ then

$$(T\mathbf{x})_k = x_{k+1}$$

Let $(\mathcal{A}_{-\infty}^{\infty}, \Sigma, \mu)$ be a probability space. In all cases, the probability law μ is assumed to be stationary, i.e. for all $E \in \Sigma$

$$\mu(T^{-1}E) = \mu(E)$$

where $T^{-1}E = \{\mathbf{x} : T\mathbf{x} \in E\}$. Also, μ is assumed to be ergodic, i.e. for any $E \in \Sigma$

$$T^{-1}E = E \implies \mu(E) = 0 \text{ or } \mu(E) = 1$$

For any positive integer l, define $\mu_l : \mathcal{A}^l \to [0,1]$ as

$$\mu_l(\mathbf{y}_0^{l-1}) = \mu(\{\mathbf{x} : x_i = y_i, 0 \le i < l\})$$

The entropy of μ is defined as

$$h = \lim_{l \to \infty} \frac{1}{l} \sum_{\mathbf{x}_0^{l-1} \in \mathcal{A}^l} -\mu_l(\mathbf{x}_0^{l-1}) \log_2 \mu_l(\mathbf{x}_0^{l-1})$$

In this work we provide a number of results each of which holds under a specific class of models for the information source. At all times we assume that the source has a finite alphabet, positive entropy and is stationary and ergodic. The memory models considered here are (in increasing order of generality) i) aperiodic and irreducible Markov of finite order, ii) the class of ergodic processes that possess *exponential rates* for entropy [12] and iii) general ergodic processes. An ergodic process is said to have exponential rates for entropy if for each $\epsilon > 0$, there exists a constant K such that for all l,

$$\mu_l(\{\mathbf{x}_0^{l-1} : 2^{-l(h+\epsilon)} \le \mu_l(\mathbf{x}_0^{l-1}) \le 2^{-l(h-\epsilon)}\}) \quad \ge \quad 1 - 2^{-Kl}$$

Given a source sample $\mathbf{x}$, define the recurrence time

$$R(\mathbf{x}, l) \quad = \quad \text{smallest } k > 0 \text{ such that } \mathbf{x}_{-k}^{-k+l-1} = \mathbf{x}_0^{l-1} \qquad l > 0$$

and $R(\mathbf{x}, 0) = 0$. Also define

$$L(\mathbf{x}, n_w) \quad = \quad 1 + \max\{l \ge 0 : R(\mathbf{x}, l) \le n_w\} \tag{1}$$

$$T_{LZ}\mathbf{x} \quad = \quad T^{L(\mathbf{x}, n_w)}\mathbf{x} \quad \forall\, \mathbf{x} \text{ s.t. } L(\mathbf{x}, n_w) < +\infty \tag{2}$$

$$T_{LZ}^k\mathbf{x} \quad = \quad T_{LZ}T_{LZ}^{k-1}\mathbf{x}$$

where the last definition is valid only if $T_{LZ}^{k-1}\mathbf{x}$ is defined and $L(T_{LZ}^{k-1}\mathbf{x}, n_w) < +\infty$. We refer to T_{LZ} as the Lempel-Ziv shift operator; note that its dependence on n_w has been suppressed in benefit of clarity. The set of sequences for which every phrase has finite length is defined as:

$$\mathcal{F}_{n_w} = \left\{ \mathbf{x} : \forall\ k \geq 0, \quad L(T_{LZ}^{k}\mathbf{x}, n_w) < +\infty \right\}$$

It will be proved that $\mu(\mathcal{F}_{n_w}) = 1$ if the information source is stationary ergodic and has positive entropy and therefore no loss in generality is suffered if attention is restricted to those sequences in $\mathcal{F}_{n_w}$. We now introduce a useful notation: if $t : \mathcal{A}_{-\infty}^{\infty} \rightarrow \mathcal{A}_{-\infty}^{\infty}$ is an operator that accepts as argument a sequence $\mathbf{y} \in \mathcal{A}_{-\infty}^{\infty}$ and returns $T^{s(\mathbf{y})}\mathbf{y}$ for some integer $s(\mathbf{y}) \geq 0$, then we say that $|t|\,\mathbf{y} = s(\mathbf{y})$. Thus for example the following is true:

$$\begin{aligned}
|T|\,\mathbf{x} &= 1 \\
|T_{LZ}|\,\mathbf{x} &= L(\mathbf{x}, n_w) \\
\left|T^{d}\right|\mathbf{x} &= d \qquad d \geq 0 \\
\left|T_{LZ}^{k}\right|\mathbf{x} &= \begin{cases} \sum_{j=0}^{k-1} L(T_{LZ}^{j}\mathbf{x}, n_w) & k \geq 1 \\ 0 & k = 0 \end{cases} \\
\left|T_{LZ}^{k}T^{d}\right|\mathbf{x} &= d + \begin{cases} \sum_{j=0}^{k-1} L(T_{LZ}^{j}T^{d}\mathbf{x}, n_w) & k \geq 1 \\ 0 & k = 0 \end{cases}
\end{aligned}$$

By a parsing of $\mathbf{x}$ we mean a strictly increasing sequence of integers $\{p_i\}_{i=0}^{\infty}$ with $p_0 = 0$. The n_w-memory Lempel-Ziv parsing of a source string $\mathbf{x}$ selects the $\{p_i\}$ sequentially according to

$$p_k = p_{k-1} + |T_{LZ}^{k-1}|\mathbf{x} \tag{3}$$

and thus by definition $p_k = \left|T_{LZ}^{k}\right|\mathbf{x}$. When addressing questions regarding the compression efficiency of Lempel-Ziv algorithms, it is useful to examine the behavior of the ratio

$$r_k(\mathbf{x}, n_w) \quad = \quad \frac{k}{\left|T_{LZ}^{k}\right|\mathbf{x}}$$

which is a number in the interval $[0, 1]$. We now describe the sliding-window Lempel Ziv algorithm with window size n_w. It will be assumed that n_w is a positive power of two in order to avoid unnecessary complications. Initially, the source encoder transmits $\mathbf{x}_{-n_w}^{-1}$ in an uncompressed form to the receiver, thus employing $n_w \lceil \log_2 |\mathcal{A}| \rceil$ bits. The source encoder transmits the the kth phrase by sending the following information

- The phrase length $L_k = L(T_{LZ}^{k-1}\mathbf{x}, n_w)$ (using $\mathrm{LEN}(L_k)$ bits)
- The pointer $R(T_{LZ}^{k-1}\mathbf{x}, L_k - 1)$ (using $\log_2 n_w$ bits)
- The value of the symbol $\mathbf{x}_{|T_{LZ}^k|\mathbf{x}-1}$ (using $\lceil \log_2 |\mathcal{A}| \rceil$ bits)

It will be assumed that the phrase length will be encoded a technique for the comma-free encoding of positive integers found in Elias [13]. The cost of encoding the kth phrase is equal to

$$\mathrm{LEN}(L_k) + \log_2 n_w + \lceil \log_2 |\mathcal{A}| \rceil \tag{4}$$

where $\mathrm{LEN} : Z \to Z$ is the length function associated with the code of [3]. This length function possesses the following bounds:

$$\log_2 l + 2\log_2 \log_2 l \le \mathrm{LEN}(l) \le \gamma + \log_2 l + 2\log_2 \log_2 l \tag{5}$$

where γ is a positive constant.

The number of bits-per-symbol generated by the n_w-memory LZ algorithm after the first k phrases of $\mathbf{x}$ have been parsed is then

$$b_k(\mathbf{x}, n_w) = b_k^P(\mathbf{x}, n_w) + b_k^L(\mathbf{x}, n_w)$$

where we have completely disregarded the bits used for the transmission of the initial uncoded window contents and

$$b_k^P(\mathbf{x}, n_w) = \left(\lceil \log_2 |\mathcal{A}| \rceil + \log_2 n_w\right) r_k(\mathbf{x}, n_w) \tag{6}$$

$$b_k^L(\mathbf{x}, n_w) = \left(\frac{1}{k}\sum_{i=1}^{k} \mathrm{LEN}(L_i(\mathbf{x}, n_w))\right) r_k(\mathbf{x}, n_w) \tag{7}$$

3 Results

As discussed in the Preliminaries, we assume that the source has positive entropy rate, a finite alphabet, is stationary and is also ergodic, although these assumptions are not explicitly written in the theorem statements. Our main results are as follows:

Theorem 1 (pointer bits) *Assume that the source is an irreducible, aperiodic finite alphabet Markov chain. Then with probability 1,*

$$\lim_{k\to\infty} b_k^P(\mathbf{X}, n_w) \le h + O\left(\sqrt{\frac{\log_2 \log_2 n_w}{\log_2 n_w}}\right)$$

as $n_w \to \infty$. The $O(\cdot)$ term does not depend on the particular path $\mathbf{X}$.

Theorem 2 (phrase length bits) *Assume that the source has exponential rates for entropy. Then with probability one*

$$\lim_{k\to\infty} b_k^L(\mathbf{X}, n_w) \leq h(1+o(1))\left(\frac{\gamma + \log_2\log_2 n_w^{1/h}}{\log_2 n_w} + 2\frac{\log_2\log_2\log_2 n_w^{1/h}}{\log_2 n_w}\right)$$
$$\lim_{k\to\infty} b_k^L(\mathbf{X}, n_w) \geq h(1-o(1))\left(\frac{\log_2\log_2 n_w^{1/(h+o(1))}}{\log_2 n_w} + 2\frac{\log_2\log_2\log_2 n_w^{1/(h+o(1))}}{\log_2 n_w}\right)$$

as $n_w \to \infty$. The $o(1)$ functions do not depend on the particular path $\mathbf{X}$.

Theorem 3 (average number of phrases per symbol) *There exists a positive constant r_{n_w} such that with probability 1, $\liminf_{k\to\infty} r_k(\mathbf{X}, n_w) = r_{n_w}$.*

Theorem 4 (limit existence) *Suppose that $EL(\mathbf{X}, n_w) < +\infty$. If f is measurable, nonnegative, integer valued and $Ef(X) < +\infty$, then*

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=0}^{k-1} f(T_{LZ}^i \mathbf{x})$$

exists and is finite with probability one.

Corollary 1 *If the source possesses exponential rates for entropy, there exists a positive constant r_{n_w} such that with probability 1, $\lim_{k\to\infty} r_k(\mathbf{X}, n_w) = r_{n_w}$.*

Theorem 5 (lower bound on redundancy) *For stationary Markov chains that are irreducible and aperiodic and with probability 1,*

$$\lim_{k\to\infty} b_k^L(\mathbf{x}, n_w) + \lim_{k\to\infty} b_k^P(\mathbf{x}, n_w) - h \geq (h - o(1))\left\{\frac{1}{2}\frac{\log_2\log_2 n_w}{\log_w n_w} + O\left(\frac{\log_2\log_2\log_2 n_w}{\log_2\log_2 n_w}\right)\right\}$$

as $n_w \to \infty$

Remarks:

- Implicit in the statements of Theorems 1, 2 and 5 is the fact that the limits addressed therein exist. This is by no means evident and will be established for sources that posses exponential rates for entropy through the use of Theorem 4.

- The fact that the limits exist with probability one does not imply that the limit is the same for every sample path. In this regard, note that Corollary 1 together with Definition (6) imply that with probability one $\lim_{k\to\infty} b_k^P(\mathbf{x}, n_w)$ does not depend on the choice of $\mathbf{x}$; however we do *not* make the same assertion for $\lim_{k\to\infty} b_k^L(\mathbf{x}, n_w)$. Rather than embracing this possibility we currently regard this as a weakness of the theory.

- Theorems 3 and 4 may not seem difficult to establish due to the fact that the source is stationary and ergodic; however we encounter the interesting question of whether the assumption that T is stationary and ergodic implies that the T_{LZ} operator is also stationary and ergodic. The author was unable to resolve this matter in the time he allotted to this question (in fact, the author believes neither is true) and pursued a path that did not require such a result. Theorem 4 is not a restatement of the results of [4] because the function is allowed to be unbounded.

4 Proofs of Theorem 1 and Theorem 2

4.1 On the existence of the limits

Before entering the main arguments, we dispose of the issue raised by the first remark in the previous section. The limits addressed by Theorems 1 and 2 exist if the limits

$$\lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1} L(T_{LZ}^{i}\mathbf{x}, n_w) \qquad \lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1} \mathrm{LEN}(L(T_{LZ}^{i}\mathbf{x}, n_w)) \tag{8}$$

exist. By Theorem 4 we only have to show that $EL(\mathbf{X}, n_w) < +\infty$ and $E\mathrm{LEN}(L(\mathbf{X}, n_w)) < +\infty$. In the Appendix we show that

Lemma 1 (finite expectation) *If the source has exponential rates for entropy, then for any* $n_w \geq 1$, $EL(\mathbf{X}, n_w) < +\infty$

It is easy to see that if M is large enough then for any positive i, the crude bound $\mathrm{LEN}(i) < M + i$ holds. Thus $E\mathrm{LEN}\,(L(\mathbf{X}, n_w)) < +\infty$ and both limits exist with probability one for either model for the source (irreducible, aperiodic Markov sources have exponential rates for entropy).

Remark: From Theorems 3 and 4 and Lemma 1 it is easy to see that Corollary 1 is true.

4.2 Preliminaries

The principal theme in proving both Theorem 1 and 2 is a classification of the phrases generated by the LZ algorithm in *short* and *long*; this idea is traces its roots back to [3] and [6]. Define

$$\ell(n_w, \xi) \quad \triangleq \quad \left\lfloor \frac{\log_2 n_w}{h + \xi} \right\rfloor$$

and let

$$c_k^S(\mathbf{x}, n_w, \xi) \quad = \quad |0 \leq i < k : L(T_{LZ}^i \mathbf{x}, n_w) < \ell(n_w, \xi) + 1| \tag{9}$$

$$c_k^L(\mathbf{x}, n_w, \xi) \quad = \quad |0 \leq i < k : L(T_{LZ}^i \mathbf{x}, n_w) \geq \ell(n_w, \xi) + 1| \tag{10}$$

so that c_k^S and c_k^L denote the number of short and long phrases in the n_w-memory parsing of the first k phrases of $\mathbf{x}$, respectively and $c_k^S + c_k^L = k$. For any $\epsilon > 0$, k ,n_w and $\mathbf{x}$,

$$\begin{aligned} \frac{c_k^L(\mathbf{x}, n_w, \epsilon)}{|T_{LZ}^k|\mathbf{x}} &\leq \frac{1}{\ell(n_w, \epsilon) + 1} \\ &\leq \frac{1}{\frac{\log_2 n_w}{h+\epsilon} - 1 + 1} \\ &= \frac{h + \epsilon}{\log_2 n_w} \end{aligned} \tag{11}$$

Therefore, we have that the average number of phrases per symbol when k phrases have been parsed can be bounded as

$$r_k(\mathbf{x}, n_w) \leq \frac{h + \epsilon}{\log_2 n_w} + \frac{c_k^S(\mathbf{x}, n_w, \epsilon)}{|T_{LZ}^k|\mathbf{x}} \tag{12}$$

The average number of short phrases is monotonically decreasing with ϵ whereas the function ϵ itself is monotonically increasing; therefore a good bound would have to select ϵ judiciously as a function of n_w. This selection will necessarily be heavily influenced with the quality of the estimates that we may be able to obtain for the second term in the right hand side. Intuitively, one would like to equate said term to the probability of observing a short phrase divided by the average phrase length. However, it is not easy to reason about the manner in which the T_{LZ} operator samples the process and it is not even clear that the sampled process is stationary. The careful reader will notice that these difficulties permeate virtually every aspect of this work and motivate us to try the following idea: overestimate the number of short phrases by examining every position in the string for an occurrence of a short phrase, in spite of the fact that this will likely give a number that is about $h^{-1} \log_2 n_w$ times larger than the true number. Because we are assuming that the process is stationary and ergodic with respect to the single shift operator T, this overestimate for the empirical density of short phrases is readily related to a probability through the application of the ergodic theorem: for a given $\epsilon > 0$ and n_w,

$$\limsup_{k \to \infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon)}{|T_{LZ}^k|\mathbf{x}} \leq \mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) \tag{13}$$

almost surely. Since as mentioned in the Introduction, we do not believe we have a tight upper bound on the redundancy since Theorem 1 is likely loose, it is appropriate to ask whether this is due to the thinking that led to (13). The answer is that although a fraction of the inefficiency of our analysis can be attributed to (13), the major source of slackness comes from underestimating the length of every long phrase as one plus the length of the longest short phrase. A sharp result should probably not make use of this idea as its main line of attack, but this was not clear when the research started. We point out that a similar trick is used in the proof of Theorem 3.

Now let $\{\epsilon_{n_w}\}$ be any sequence. Since there are countably many tuples (n_w, ϵ_{n_w}), intersecting the countably many sets of full measure that the ergodic theorem will give for every tuple (n_w, ϵ_{n_w}) in (13) we obtain a set of full measure again. Therefore, the following statement holds: there exists a set of full measure such that for every n_w,

$$\limsup_{k\to\infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \leq \mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon_{n_w})) > n_w\}\right) \tag{14}$$

is true. Thus the natural direction to follow is that of establishing upper bounds on the term in the right of (14). In this regard we offer

Theorem 6 *For irreducible and aperiodic Markov chains there exist constants* $\lambda_1, \lambda_2, \lambda_3, \lambda_4 > 0$ *such that if* $\log_2 n_w \geq \lambda_2/\epsilon$ *and* $\epsilon < 1/\lambda_3$

$$\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) \quad \leq \quad \lambda_4 \frac{(\log_2 n_w)^{1+|\mathcal{A}|^2} / 2^{\lambda_1 \epsilon^2 \log_2 n_w}}{\log_2 n_w}$$

or the weaker

Theorem 7 *For stationary, ergodic sources that possess exponential rates for entropy, there exists a sequence* $\epsilon_{n_w} \to 0$ *such that*

$$\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon_{n_w})) > n_w\}\right) \quad \leq \quad \frac{o(1)}{\log_2 n_w}$$

as $n_w \to \infty$.

which is valid under less restrictive assumptions on the source's memory structure.

Remark: A very useful property of Theorem 6 is that it holds for all values for the pair (n_w, ϵ) that satisfy $\log_2 n_w \geq \lambda_2/\epsilon$ and $\epsilon < 1/\lambda_3$. This contrasts with Theorem 7, which is valid only for a given sequence ϵ_{n_w} which decays to zero in an unspecified manner.

4.3 Upper bound in Theorem 1

Let m be an integer equal or greater than 2. Taking advantage of the flexibility that Theorem 6 affords us we choose

$$\epsilon_{n_w} = \sqrt{\frac{m + |\mathcal{A}|^2}{\lambda_1} \frac{\log_2 \log_2 n_w}{\log_2 n_w}} \tag{15}$$

Then with probability one, if n_w is large enough so that $\log_2 n_w \geq \lambda_2/\epsilon_{n_w}$ and $\epsilon_{n_w} < 1/\lambda_3$, we know that

$$\limsup_{k\to\infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \leq \lambda_4 \left(\frac{1}{\log_2 n_w}\right)^m \tag{16}$$

Then

$$\begin{aligned} \limsup_{k\to\infty} r_k(\mathbf{x}, n_w) &\leq \frac{h + \sqrt{\frac{m+|\mathcal{A}|^2}{\lambda_1} \frac{\log_2 \log_2 n_w}{\log_2 n_w}}}{\log_2 n_w} + \lambda_4 \left(\frac{1}{\log_2 n_w}\right)^m \\ &\leq \frac{h + O\left(\sqrt{\frac{\log_2 \log_2 n_w}{\log_2 n_w}}\right)}{\log_2 n_w} \end{aligned} \tag{17}$$

Note that the choice $m = 2$ is sufficient to obtain (17). The upper bound for b_k^P is obtained by combining its Definition (6) with (17). If on the other hand we wished to make use of Theorem 7 instead, we would have

$$\limsup_{k\to\infty} r_k(\mathbf{x}, n_w) \leq \frac{h + o(1)}{\log_2 n_w} \tag{18}$$

as $n_w \to \infty$ which is insufficient for redundancy calculation purposes.

4.4 Upper bound in Theorem 2

For all k, n_w and $\mathbf{x}$,

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} \mathrm{LEN}(L_i(\mathbf{x}, n_w)) &\overset{(a)}{\leq} \sum_{i=1}^{k} \frac{1}{k} (\gamma + \log_2 L_i(\mathbf{x}, n_w) + 2\log_2 \log_2 L_i(\mathbf{x}, n_w)) \\ &\overset{(b)}{\leq} \gamma + \log_2 \left(\frac{\sum_{i=1}^{k} L_i(\mathbf{x}, n_w)}{k}\right) + 2\log_2 \log_2 \left(\frac{\sum_{i=1}^{k} L_i(\mathbf{x}, n_w)}{k}\right) \\ &= \gamma + \log_2 \frac{1}{r_k(\mathbf{x}, n_w)} + 2\log_2 \log_2 \frac{1}{r_k(\mathbf{x}, n_w)} \end{aligned} \tag{19}$$

where (a) is due to (5) and (b) follows from Jensen's inequality and the convexity $\bigcap$ of the $\log_2$ and $\log_2 \log_2$ functions. Therefore

$$\begin{aligned} \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w) \leq\ & \limsup_{k\to\infty} \gamma r_k(\mathbf{x}, n_w) + \limsup_{k\to\infty} \left(r_k(\mathbf{x}, n_w) \log_2 \frac{1}{r_k(\mathbf{x}, n_w)}\right) \\ & + 2\limsup_{k\to\infty} \left(r_k(\mathbf{x}, n_w) \log_2 \log_2 \frac{1}{r_k(\mathbf{x}, n_w)}\right) \end{aligned} \tag{20}$$

Theorem 7 implies that with probability one,

$$\limsup_{k\to\infty} r_k(\mathbf{x}, n_w) \le \frac{h}{\log_2 n_w} + \frac{o(1)}{\log_2 n_w}$$

as $n_w \to \infty$ where the $o(1)$ term is a nonnegative function of n_w that is equal for all $\mathbf{x}$.

The functions $t \log_2 \frac{1}{t}$ and $t \log_2 \log_2 \frac{1}{t}$ are strictly increasing in a domain $t \in (0, t^*)$ where t^* is a positive constant given by Lemma 11 (see the Appendix). Therefore, for n_w sufficiently large (so that the right hand side of (21) is less than say, $t^*/2$)

$$\limsup_{k\to\infty} \left(r_k(\mathbf{x}, n_w) \log_2 \frac{1}{r_k(\mathbf{x}, n_w)} \right) \quad < \quad \frac{h(1+o(1))}{\log_2 n_w} \log_2 \log_2 n_w^{1/h}$$

A bound for $\limsup_{k\to\infty} r_k(\mathbf{x}, n_w) \log_2 \log_2 1/r_k(\mathbf{x}, n_w)$ may be obtained similarly. We conclude that

$$\limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w) \quad \le \quad h(1+o(1)) \left(\frac{\gamma + \log_2 \log_2 n_w^{1/h}}{\log_2 n_w} + 2 \frac{\log_2 \log_2 \log_2 n_w^{1/h}}{\log_2 n_w} \right) \tag{21}$$

as $n_w \to \infty$.

Remark: The upper bounds (18) and (21) give an almost sure convergence convergence proof alternative to the one given by Shields in [8] but valid only for a more restricted class of sources. This result was published in [9] without knowledge of [8].

4.5 Lower bound in Theorem 2

In light of Theorem 7 most phrases are expected to be long. However the Elias encoding of the length of the long phrases will require a certain minimum amount of bits. This observation gives the lower bound:

$$\begin{aligned} \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{i=1}^{k} \mathrm{LEN}(L_i(\mathbf{x}, n_w)) \quad &\ge \quad \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{i=1}^{k} \log_2 L_i(\mathbf{x}, n_w) + 2 \log_2 \log_2 L_i(\mathbf{x}, n_w) \\ &\ge \quad \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \left(\log_2 \frac{\log_2 n_w}{h + \epsilon_{n_w}} + 2 \log_2 \log_2 \frac{\log_2 n_w}{h + \epsilon_{n_w}} \right) \end{aligned} \tag{22}$$

By definition,

$$\frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} + \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} = \frac{k}{|T_{LZ}^k|\mathbf{x}} = r_k(\mathbf{x}, n_w)$$

and in particular,

$$\liminf_{k\to\infty} \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} + \limsup_{k\to\infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} = \liminf_{k\to\infty} r_k(\mathbf{x}, n_w)$$

By Corollary 1 with probability 1, the lim inf in the right hand side can be replaced with a lim to yield

$$\liminf_{k\to\infty} \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \quad \geq \quad \lim_{k\to\infty} r_k(\mathbf{x}, n_w) - \limsup_{k\to\infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \tag{23}$$

The almost sure pointwise converse for all codes that succeed in transmitting the source losslessly guarantees that with probability one (Theorem II.1.2 [12], p. 122)

$$(\lceil \log_2 |\mathcal{A}| \rceil + \log_2 n_w) \lim_{k\to\infty} r_k(\mathbf{x}, n_w) \quad \geq \quad h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w)$$

This observation together with (23) leads to

$$\liminf_{k\to\infty} \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \geq \frac{h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w)}{\lceil \log_2 A \rceil + \log_2 n_w} - \limsup_{k\to\infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \tag{24}$$

It will be convenient to simplify the shape of the first term:

$$\begin{aligned} \frac{h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w, \epsilon)}{\lceil \log_2 A \rceil + \log_2 n_w} &= \frac{\log_2 n_w}{\lceil \log_2 A \rceil + \log_2 n_w} \frac{h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w, \epsilon)}{\log_2 n_w} \\ &= \left(1 - O\left(\frac{1}{\log_2 n_w}\right)\right) \frac{h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w, \epsilon)}{\log_2 n_w} \end{aligned} \tag{25}$$

To give a lower bound for the above one can use Theorem 7 along with a weakened form of Equation (21) to obtain

$$\liminf_{k\to\infty} \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \geq \frac{h(1 - o(1))}{\log_2 n_w}$$

almost surely. Using this result in conjunction with (22) implies

$$\liminf_{k\to\infty} b_k^L(\mathbf{x}, n_w) \quad \geq \quad h(1 - o(1)) \left(\frac{\log_2 \log_2 n_w^{1/(h+o(1))}}{\log_2 n_w} + 2 \frac{\log_2 \log_2 \log_2 n_w^{1/(h+o(1))}}{\log_2 n_w} \right) \tag{26}$$

as desired.

4.6 Proof of the Lower Bound on Redundancy

We have established in a relatively satisfactory manner the number of bits that are devoted to the encoding of the phrase length bits, but the resolution of the simple phrase length thresholding techniques that we have used is apparently unable to yield good upper estimates for the pointer bits. It thus may come as a surprise that we are able to give a lower bound for the total redundancy that is off by a factor of two from the best known lower bound on redundancy for LZ-type algorithms that use a parsing rule originating from LZ77 [2]. The expression that we want to lower bound is

$$\lim_{k\to\infty} b_k^L(\mathbf{x}, n_w) + \lim_{k\to\infty} b_k^P(\mathbf{x}, n_w) - h \tag{27}$$

Now suppose that we had an alternative method which spends $\hat{b}_k^L(\mathbf{x}, n_w)$ bits for encoding the phrase lengths. Further assume that with probability one,

$$\limsup_{k\to\infty} \hat{b}_k^L(\mathbf{x}, n_w) \leq \xi_1 \tag{28}$$

$$\lim_{k\to\infty} b_k^L(\mathbf{x}, n_w) \geq \xi_2 \tag{29}$$

for some ξ_1, ξ_2 satisfying

$$\xi_2 - \xi_1 > 0 \tag{30}$$

We have chosen to use a lim sup instead of a lim in the above because we want to avoid having to settle the issue of the existence of the lim for the alternative method.

The almost sure converse for lossless source coding gives (Theorem II.1.2 [12], p. 122) implies that

$$\xi_1 + \lim_{k\to\infty} b_k^P(\mathbf{x}, n_w) \geq h$$

Using this to lower bound $\lim_{k\to\infty} b_k^P(\mathbf{x}, n_w)$ in (27), we obtain

$$\lim_{k\to\infty} b_k^L(\mathbf{x}, n_w) + \lim_{k\to\infty} b_k^P(\mathbf{x}, n_w) - h \geq \xi_2 - \xi_1$$

which is non trivial if (30) holds. This method for lower bounding the redundancy of an algorithm is generally applicable and particularly attractive in the context of this paper because it does not require direct lower bounds for the pointer bits. The method requires conceiving an algorithm that achieves better performance than the algorithm under consideration and thus it may appear discouraging from a practical point of view; after all it is quite reasonable to attempt to make statements about the redundancy of the best possible algorithm that one can design. Nevertheless, note that the alternative algorithm need not be universal, or even computationally attractive; it must simply achieve the transmission of any data string without any loss. The author has been informed by A.J. Wyner that in his PhD thesis [11] a similar idea is used to give a lower bound to FDLZ. Furthermore, better (practical) techniques for encoding the phrase lengths have been devised by Bender et.al. [14] and further analyzed by Wyner et.al. [15]. A discussion on the similarities and differences of A.J. Wyner's work with the present can be found in the Introduction.

In the spirit of this discussion, let us now consider a different method for encoding the phrase lengths. Said method uses an additional flag bit which indicates whether the phrase length is short or long, and encodes the length of the phrase in a manner appropriate to the case. Discriminating phrases in the basis of their length requires the specification of a sequence ϵ_{n_w}; for this we will choose the sequence in (15) with $m = 2$. As mentioned in the Theorem statement, this discussion will be valid for Markov chains. Without factoring in the expense of the flag bit, it can be easily seen that the average number of bits that are needed to encode the phrase lengths of the short phrases is no greater than

$$\limsup_{k\to\infty} \frac{c_k^S(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \left[\log_2\left(\frac{\log_2 n_w}{h+\epsilon_{n_w}}\right)\right]$$

bits. Using Equation (16), the above is seen to have an upper bound of

$$o(1)\frac{\log_2\log_2 n_w}{\log_2 n_w} \tag{31}$$

as $n_w \to \infty$. The length L of a long phrase is transmitted by encoding the difference between L and $\left\lfloor \frac{\log_2 n_w}{h+\epsilon_{n_w}} \right\rfloor$ and will employ at most

$$\mathrm{LEN}\left(L - \left\lfloor\frac{\log_2 n_w}{h+\epsilon_{n_w}}\right\rfloor\right) \le \gamma + \log_2\left(L - \frac{\log_2 n_w}{h+\epsilon_{n_w}} + 1\right) + 2\log_2\log_2 L$$

bits. Note that in the second term the difference $L - \lfloor \log_2 n_w/(h+\epsilon_{n_w})\rfloor$ has been upper bounded by L; this is done to avoid complications that are unnecessary since all that is is currently of concern is the first order properties in the redundancy lower bound. The total number of bits spent for the encoding of the long phrase lengths after k phrases have been parsed is at most

$$\sum_{0\le j<k: L(T_{LZ}^j\mathbf{x}, n_w)\ge \left\lfloor\frac{\log_2 n_w}{h+\epsilon_{n_w}}\right\rfloor+1} \gamma + \log_2\left(L(T_{LZ}^j\mathbf{x}, n_w) - \frac{\log_2 n_w}{h+\epsilon_{n_w}} + 1\right) + 2\log_2\log_2\left(L(T_{LZ}^j\mathbf{x}, n_w))\right)$$

There are exactly $c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})$ indexes j in the summation above. Notice also that the sum of the lengths of the long phrases found after k phrases have been parsed is always upper bounded by $|T_{LZ}^k|\mathbf{x}$. By multiplying and dividing the summation by $c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})$, noting the convexity $\bigcap$ of the $\log_2$ and $\log_2\log_2$ functions (for the latter note that $\lambda\log_2\log_2 a + (1-\lambda)\log_2\log_2 b \le \log_2(\lambda\log_2 a + (1-\lambda)\log_2 b) \le \log_2\log_2(\lambda a + (1-\lambda)b))$ and using Jensen's inequality, it follows that an upper bound to the average number bits used for the lengths of the long phrases is

$$\limsup_{k\to\infty}\left\{\frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}}\left[\gamma + \log_2\left(\frac{|T_{LZ}^k|\mathbf{x}}{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})} - \frac{\log_2 n_w}{h+\epsilon_{n_w}} + 1\right) + 2\log_2\log_2\left(\frac{|T_{LZ}^k|\mathbf{x}}{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}\right)\right]\right\}$$

Note from Inequality (11) that

$$\limsup_{k\to\infty}\frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \le \frac{h+\epsilon_{n_w}}{\log_2 n_w} \tag{32}$$

Furthermore, from Inequalities (16), (24) and Equality (25) we can deduce that

$$\limsup_{k\to\infty} \frac{|T_{LZ}^k|\mathbf{x}}{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})} \leq \frac{\log_2 n_w}{h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w) - O(1/\log_2 n_w)} \tag{33}$$

Note that the value of $\limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w)$ is independent of the partition threshold that classifies phrases as long or short. Therefore, the bounds in Theorem 2 are valid in this situation, in spite of the fact that in order to derive them, we used a partition threshold that is unrelated to the one being presently used. A weak form of (21) states that

$$\limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w) = o(1)$$

as $n_w \to \infty$. Combining the above with (32) and (33) we get

$$\limsup_{k\to\infty} \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \log_2 \log_2 \frac{|T_{LZ}^k|\mathbf{x}}{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})} = O\left(\frac{\log_2 \log_2 \log_2 n_w}{\log_2 n_w}\right) \tag{34}$$

Subtracting $\log_2 n_w/(h + \epsilon_{n_w})$ from the right hand side in (33), we obtain

$$\frac{\epsilon_{n_w} + \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w) + O(1/\log_2 n_w)}{\left(h - \limsup_{k\to\infty} b_k^L(\mathbf{x}, n_w) - O(1/\log_2 n_w)\right)(h + \epsilon_{n_w})} \log_2 n_w$$

Using the bound in (21) and the definition of ϵ_{n_w} in (15) it can be seen that the above is equal to

$$O(\epsilon_{n_w}) \log_2 n_w \tag{35}$$

as $n_w \to \infty$. Thus,

$$\begin{aligned}
&\limsup_{k\to\infty} \frac{c_k^L(\mathbf{x}, n_w, \epsilon_{n_w})}{|T_{LZ}^k|\mathbf{x}} \log_2 \left(\frac{|T_{LZ}^k|\mathbf{x}}{c_k^L(\mathbf{x}, n_w)} - \frac{\log_2 n_w}{h + \epsilon_{n_w}} + 1\right) \\
\leq\ & (h + \epsilon_{n_w}) \frac{\log_2 \left((1/\log_2 n_w + O(\epsilon_{n_w})) \log_2 n_w\right)}{\log_2 n_w} \\
=\ & (h + \epsilon_{n_w}) \left\{\frac{\log_2 O\left(\sqrt{\frac{\log_2 \log_2 n_w}{\log_2 n_w}}\right)}{\log_2 n_w} + \frac{\log_2 \log_2 n_w}{\log_2 n_w}\right\} \\
=\ & (h + \epsilon_{n_w}) \left\{\frac{1}{2} \frac{\log_2 \frac{\log_2 \log_2 n_w}{\log_2 n_w}}{\log_2 n_w} + \frac{\log_2 \log_2 n_w}{\log_2 n_w} + O\left(\frac{1}{\log_2 n_w}\right)\right\} \\
=\ & (h + \epsilon_{n_w}) \left\{\frac{1}{2} \frac{\log_2 \log_2 n_w}{\log_2 n_w} + O\left(\frac{\log_2 \log_2 \log_2 n_w}{\log_2 n_w}\right)\right\}
\end{aligned} \tag{36}$$

On the other hand from (26) we have

$$\liminf_{k\to\infty} b_k^L(\mathbf{x}, n_w) \geq (h - o(1)) \left\{\frac{\log_2 \log_2 n_w}{\log_2 n_w} + O\left(\frac{\log_2 \log_2 \log_2 n_w}{\log_2 n_w}\right)\right\} \tag{37}$$

for the original algorithm; we reiterate that the fact that this result was derived using a partition of the phrase lengths different from the one considered here does not preclude its applicability in this situation because it is a statement about the total number of phrase length bits. Finally, from (31), (34), (36) and (37) we obtain the desired result.

4.7 Proofs of Theorem 6 and Theorem 7

We now proceed to demonstrate how to obtain the estimates in Theorem 6 and Theorem 7. For the latter, we employ the results of Lastras [16] which describe extensions to the theory of Markov types that enable us to compute more flexible upper bounds to the probability of the ϵ-typical set for n-sequences; note nevertheless that although sufficient, the techniques in [16] are not perceived by the author as necessary in any strong sense. The material is organized in a way such that for the most part, first the arguments relevant to Theorem 6 are presented and then the simplifications needed for Theorem 7 are pointed out.

4.7.1 Preliminary material

Recall that the ϵ-atypical set is defined as

$$A_{l,\epsilon} \triangleq \left\{ \mathbf{x}_0^{l-1} : \left| -\frac{1}{l} \log_2 \mu_l(\mathbf{x}_0^{l-1}) - h \right| > \epsilon \right\}$$

We will also need Kac's Lemma:

Lemma 2 *(Kac's Lemma)*

$$E\left[R(\mathbf{X}, l) | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1} \right] = \frac{1}{\mu_l(\mathbf{x}_0^{l-1})}$$

Proof. See the Appendix in [17]. □

Lemma 3 *If the source is a finite alphabet Markov chain that is irreducible and aperiodic, there exist constants $\kappa_1(P) > 0$ and $\kappa_2(P) > 0$ such that for every pair (ϵ, l) that satisfies $l-1 > \kappa_2(P)/\epsilon$ the probability of the ϵ-atypical set satisfies*

$$\mu\left(A_{l,\epsilon}\right) \leq |\mathcal{A}| \left(1 + \frac{l-1}{|\mathcal{A}|}\right)^{|\mathcal{A}|^2} 2^{-(l-1)\kappa_1(P)\epsilon^2}$$

Proof. See Lastras [16].

4.7.2 The proofs' skeleton

Proof of Theorem 6. The proof is obtained by making the substitution $l \to \ell(n_w, \epsilon)$ in

Lemma 4 *If the source is an irreducible and aperiodic Markov chain then there exist constants $\kappa_1(P)$ and $\kappa_2(P)$ (that coincide with those of Lemma 3) such that if $l-1 > \kappa_2(P)/\epsilon$ and $\epsilon < 2/\kappa_1(P)$,*

$$\mu\left(\left\{\mathbf{x} : R(\mathbf{x}, l) > 2^{l(h+\epsilon)}\right\}\right) \leq 2|\mathcal{A}|\left(1 + \frac{l-1}{|\mathcal{A}|}\right)^{|\mathcal{A}|^2} 2^{-(l-1)\kappa_1(P)(\epsilon/2)^2}$$

The only step that may not be immediately obvious makes use of the observation that

$$\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) \quad \leq \quad \mu\left(\left\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > 2^{\ell(n_w,\epsilon)(h+\epsilon)}\right\}\right)$$

□

Proof of Theorem 7. We make the same substitution in the following variant of Lemma 4:

Lemma 5 *If the source has exponential rates for entropy, then for every $\epsilon > 0$ there exists a $\delta_\epsilon > 0$ and a constant l_ϵ such that if $l \geq l_\epsilon$*

$$\mu\left(\left\{\mathbf{x} : R(\mathbf{x}, l) > 2^{l(h+\epsilon)}\right\}\right) \leq 2^{-l\delta_\epsilon}$$

For the substitution to be possible, for every ϵ n_w has to be large enough so that $\ell(n_w, \epsilon) \geq l_\epsilon$. Upon making said substitution, and considering the cases $\delta_\epsilon \geq 1$ and $\delta_\epsilon < 1$ separately, the lemma becomes

$$\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) \leq 2^{-\left\lfloor \frac{\log_2 n_w}{h+\epsilon} \right\rfloor \delta_\epsilon} \leq 2^{-\left(\frac{\log_2 n_w}{h+\epsilon} - 1\right)\delta_\epsilon} \leq 2\max\left\{n_w^{-\frac{1}{h+\epsilon}}, n_w^{-\frac{\delta_\epsilon}{h+\epsilon}}\right\}$$

for n_w large enough, and therefore for a fixed $\epsilon > 0$,

$$\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) \leq \frac{o(1)}{\log_2 n_w}$$

as $n_w \to \infty$. The remaining arguments are analytic in nature. Let $\mu_{n_w,\epsilon} = \mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right)$. Thus for every ϵ, $\{\mu_{n_w,\epsilon}\}_{n_w}$ is a sequence of real numbers with the property that

$$\mu_{n_w,\epsilon} \log_2 n_w = o(1) \tag{38}$$

as $n_w \to \infty$. The goal is to show that there exists a sequence $\{\epsilon_{n_w}\}$ with the properties that

$$\epsilon_{n_w} = o(1) \qquad \mu_{n_w,\epsilon_{n_w}} = o(1)$$

as $n_w \to \infty$. As a useful intermediate construction, for every m, let $s_m > s_{m-1}$ be large enough so that

$$\mu_{s_m,1/m} \log_2 s_m \quad \leq \quad 1/m$$

Now define

$$\epsilon_{n_w} = \begin{cases} 1/(m-1) & s_{m-1} \leq n_w < s_m \\ 1/m & n_w = s_m \end{cases}$$

It is clear that $\epsilon_{n_w} \to 0$ and little effort is required to see that $\mu_{n_w,\epsilon_{n_w}} \log_2 n_w = o(1)$. The proof is complete. The reader may appreciate that we have no knowledge regarding how fast $\mu_{n_w,\epsilon_{n_w}} \log_2 n_w$ or ϵ_{n_w} decay to zero, which renders it difficult to obtain redundancy results starting from this Theorem; these arguments motivate weakening the source's memory assumptions to obtain the stronger Theorem 6, although even in this case the resulting redundancy bound may be judged to be too large.

4.7.3 The proofs of Lemmas 4 and 5

Proof of Lemma 4. For any l,

$$\begin{aligned} &\mu\left(\left\{\mathbf{x} : R\left(\mathbf{x}, l\right) > 2^{l(h+\epsilon)}\right\}\right) \\ &\leq \mu_l\left(A_{l,\epsilon/2}\right) + \sum_{\mathbf{x}_0^{l-1} \in A^c_{l,\epsilon/2}} \mu\left(\left\{\mathbf{x} : R\left(\mathbf{x}, l\right) > 2^{l(h+\epsilon)}\right\} | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1}\right) \mu_l\left(\mathbf{x}_0^{l-1}\right) \end{aligned} \tag{39}$$

The summation is upper bounded as follows:

$$\begin{aligned} &\sum_{\mathbf{x}_0^{l-1} \in A^c_{l,\epsilon/2}} \mu\left(\left\{\mathbf{x} : R(\mathbf{x}, l) > 2^{l(h+\epsilon)}\right\} | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1}\right) \mu_l\left(\mathbf{x}_0^{l-1}\right) \\ &\overset{(a)}{\leq} \sum_{\mathbf{x}_0^{l-1} \in A^c_{l,\epsilon/2}} \frac{E\left[R(\mathbf{X}, l) | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1}\right]}{2^{l(h+\epsilon)}} \mu_l\left(\mathbf{x}_0^{l-1}\right) \\ &\overset{(b)}{=} \sum_{\mathbf{x}_0^{l-1} \in A^c_{l,\epsilon/2}} \frac{\frac{1}{\mu_l(\mathbf{x}_0^{l-1})}}{2^{l(h+\epsilon)}} \mu_l\left(\mathbf{x}_0^{l-1}\right) \\ &\overset{(c)}{\leq} \sum_{\mathbf{x}_0^{l-1} \in A^c_{l,\epsilon/2}} 2^{-l\epsilon/2} \mu_l\left(\mathbf{x}_0^{l-1}\right) \\ &\leq 2^{-l\epsilon/2} \end{aligned} \tag{40}$$

where (a) follows from the Markov inequality, (b) follows from Kac's Lemma and (c) follows from the definition of $A^c_{l,\epsilon/2}$. By Lemma 3 we know that there exist constants $\kappa_1(P)$ and $\kappa_2(P)$ such that as long as $l-1 > \kappa_2(P)/\epsilon$,

$$\mu_l\left(A^c_{l,\epsilon/2}\right) \leq |\mathcal{A}|\left(1+\frac{1}{|\mathcal{A}|}\right)^{|\mathcal{A}|^2} 2^{-(l-1)\kappa_1(P)(\epsilon/2)^2}$$

Combining both bounds we obtain

$$\begin{aligned}\mu\left(\left\{\mathbf{x} : R(\mathbf{x},l) > 2^{l(h+\epsilon)}\right\}\right) &\leq 2^{-l\epsilon/2} + |\mathcal{A}|\left(1+\frac{l-1}{|\mathcal{A}|}\right)^{|\mathcal{A}|} 2^{-l\kappa_1(P)(\epsilon/2)^2} \\ &\leq 2|\mathcal{A}|\left(1+\frac{l-1}{|\mathcal{A}|}\right)^{|\mathcal{A}|^2} 2^{-(l-1)\kappa_1(P)(\epsilon/2)^2}\end{aligned}$$

where the latter is a gross bound that is true under the sufficient condition $\epsilon < 2/\kappa_1(P)$. □

Proof of Lemma 5. Since the process is assumed to have exponential rates for entropy, there exists a constant K (which depends on $\epsilon/2$) such that for all l,

$$\mu_l\left(A_{l,\epsilon/2}\right) \leq 1 - 2^{-lK} \tag{41}$$

Now define $\delta = \frac{1}{2}\min(\epsilon/2, K)$ and choose $l^* > 1/\delta$. Using (39), (40) and (41) it follows that for all $l \geq l^*$,

$$\begin{aligned}\mu\left(\left\{\mathbf{x} : R(\mathbf{x},l) > 2^{l(h+\epsilon)}\right\}\right) &\leq 2^{-l\epsilon/2} + 2^{-lK} \\ &\leq 2^{-l(2\delta - 1/l)} \\ &\leq 2^{-l\delta}\end{aligned}$$

thus ending the proof of the corollary. □

5 Proof of Theorems 3 and 4

Theorem 3 (restated) *Assume that the source has positive entropy, a finite alphabet, is stationary and ergodic. Then there exists a positive constant* r_{n_w} *such that with probability 1,* $\liminf_{k\to\infty} r_k(\mathbf{X}, n_w) = r_{n_w}$.

The gist of the proof consists on showing the following Lemma:

Lemma 6 *If* $\mathbf{x} \in \mathcal{F}_{n_w}$ *then* $\liminf_{k\to\infty} r_k(\mathbf{x}, n_w)$ *is invariant under the* T *operator.*

It then follows that for every integer $q > 0$ and $\Delta > 0$ the set

$$\mathcal{I}_{q,\Delta} = \mathcal{F}_{n_w} \bigcap \left\{ \mathbf{x} : (q-1)\Delta \leq \liminf_{k\to\infty} r_k(\mathbf{x}, n_w) < q\Delta \right\}$$

is invariant under T (note that by Lemma 14, $T\mathcal{F}_{n_w} = \mathcal{F}_{n_w}$) and therefore invoking the ergodicity of the source with respect to the T operator, $\mu(\mathcal{I}_{q,\Delta}) \in \{0,1\}$. Recall that $\mu(\mathcal{F}_{n_w}) = 1$. Since $r_k(\mathbf{x}, n_w) \in [0,1]$, from $1 = \mu(\Omega) = \mu(\bigcup_q \mathcal{I}_{q,\Delta}) = \sum_q \mu(\mathcal{I}_{q,\Delta})$ we conclude that there is exactly one q which depends on Δ for which $\mu(\mathcal{I}_{q,\Delta}) = 1$. Since Δ can be chosen to be as small as needed, we conclude that $\liminf_{k\to\infty} r_k(\mathbf{X}, n_w)$ is almost surely constant, said constant computable from (almost) any realization of the process. It only remains to show that this constant is not equal to zero. The source coding converse result gives (Theorem II.1.2 [12], p. 122)

$$\liminf_k b_k^P(\mathbf{X}, n_w) + \liminf_k b_k^L(\mathbf{X}, n_w) \geq h$$

with probability one. Using Lemma 12 we can obtain the following upper bound for the left hand side:

$$(\log_2 n_w + \lceil \log_2 |\mathcal{A}| \rceil) \liminf_k r_k(\mathbf{X}, n_w) + c\sqrt{\liminf_k r_k(\mathbf{X}, n_w)}$$

for some positive constant c. From the assumption that $h > 0$ it then follows that $\liminf_k r_k(\mathbf{X}, n_w) > 0$ with probability one. $\square$

We point out that Lemma 6 would not be needed if we could assume that the source is ergodic with respect to T_{LZ} for $\liminf_{k\to\infty} r_k(T_{LZ}^k \mathbf{x}, n_w)$ is trivially invariant under said operator. We were unable to show the ergodicity of the source with respect to T_{LZ} for ergodic sources.

5.1 Proof of Lemma 6

The proof of Lemma 6 is based on an observation due to Vittorio Castelli which was discussed in the context of a different problem. Said observation is summarized in:

Lemma 7 (Vittorio's Lemma) *Let $\mathbf{z} \in \mathcal{F}_{n_w}$. Then for any $0 \leq i \leq L(\mathbf{z}, n_w)$,*

$$L(\mathbf{z}, n_w) \leq i + L(T^i \mathbf{z}, n_w)$$

Vittorio's Lemma can then be used to prove the more general result:

Lemma 8 *Let $\mathbf{x} \in \mathcal{F}_{n_w}$. Then for any $0 \leq d \leq L(\mathbf{x}, n_w)$ and for any integer $k \geq 0$*

$$\left|T_{LZ}^k\right| \mathbf{x} \leq d + \left|T_{LZ}^k\right| T^d \mathbf{x} \leq \left|T_{LZ}^{k+1}\right| \mathbf{x}$$

Finally, we will also use the following simple analysis result:

Lemma 9 *Let d be any number and let $\{a_i\}_{i=0}^{\infty}$ and $\{b_i\}_{i=0}^{\infty}$ be sequences of positive integers such that $a_0 \leq b_0 + d \leq a_1 \leq b_1 + d \leq a_2 \leq b_2 + d \cdots$ and further assume that for all k, $a_k < a_{k+1}$ and $b_k < b_{k+1}$. Then*

$$\liminf_{k\to\infty} \frac{k}{a_k} = \liminf_{k\to\infty} \frac{k}{b_k}$$

We are ready to prove Lemma 6. Let $\mathbf{x} \in \mathcal{F}_{n_w}$. Recall the definitions

$$\begin{aligned} r_k(\mathbf{x}, n_w) &= \frac{k}{\left|T_{LZ}^k\right| \mathbf{x}} \\ r_k(T\mathbf{x}, n_w) &= \frac{k}{\left|T_{LZ}^k\right| T\mathbf{x}} \end{aligned}$$

and let $a_k = \left|T_{LZ}^k\right| \mathbf{x}$ and $b_k = \left|T_{LZ}^k\right| T\mathbf{x}$. Applying Lemma 8 we can see that

$$a_0 \leq d + b_0 \leq a_1 \leq d + b_1 \leq a_2 \leq d + b_2 \cdots$$

and also note that for all k, $a_k < a_{k+1}$ and $b_k < b_{k+1}$. Thus the conditions of Lemma 9 are satisfied and therefore

$$\liminf_{k\to\infty} \frac{k}{a_k} = \liminf_{k\to\infty} \frac{k}{b_k}$$

which concludes the proof of the lemma. □

We point out that the proof of Theorem 4 relies on Theorem 3 and thus the assumptions of the latter are part of those of the former.

Theorem 4 (restated) *Suppose that $EL(\mathbf{X}, n_w) < +\infty$. If f is measurable, nonnegative, integer valued and $Ef(X) < +\infty$, then*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f(T_{LZ}^i \mathbf{x})$$

exists and is finite with probability one.

Proof. We cannot assume that the measure is invariant with respect to the T_{LZ} operator and therefore a straightforward application of the ergodic theorem is not possible. Fortunately, we have at our disposal a result of Kieffer and Rahe [4] which when specialized to the setting of interest assures us that if

$$EL(\mathbf{X}, n_w) < +\infty \tag{42}$$

and if a function g is bounded and measurable then the limit

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} g(T_{LZ}^i \mathbf{x}) \tag{43}$$

exists with probability one. We are assuming that (42) is true, however [4] does not apply directly because the function being averaged is unbounded in general. Nevertheless, there is a way to circumvent this difficulty. The first item in the agenda is to show that with probability one,

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{X}\right) < +\infty$$

By Theorem 3 the limsup

$$\kappa \;\stackrel{\Delta}{=}\; \limsup_{k\to\infty} \frac{1}{r_k(\mathbf{x}, n_w)} \tag{44}$$

is almost surely finite and constant with probability one. Now note that

$$\begin{aligned} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{x}\right) &= \frac{1}{r_k(\mathbf{x}, n_w)} \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{x}\right) \\ &\leq \frac{1}{r_k(\mathbf{x}, n_w)} \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{j=0}^{|T_{LZ}^k|\mathbf{x}-1} f\left(T^j \mathbf{x}\right) \end{aligned}$$

Taking the lim sup as $k \to \infty$, and using the ergodic theorem and (44) we obtain that with probability one,

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{x}\right) \leq \kappa E f(\mathbf{X}) < +\infty$$

as desired. We now proceed to prove the lemma. For any positive k and t

$$\frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{x}\right) = \frac{1}{k} \sum_{0\leq i<k: f(T_{LZ}^i \mathbf{x})\leq t} f\left(T_{LZ}^i \mathbf{x}\right) + \frac{1}{k} \sum_{0\leq i<k: f(T_{LZ}^i \mathbf{x})>t} f\left(T_{LZ}^i \mathbf{x}\right) \tag{45}$$

The results of Kieffer and Rahe [4] assure us that the limit

$$F_t(\mathbf{x}) \triangleq \lim_{k\to\infty} \frac{1}{k} \sum_{0\le i<k: f(T_{LZ}^i \mathbf{x})\le t} f\left(T_{LZ}^i \mathbf{x}\right)$$

exists with probability one. The nonnegativity of f implies that the sequence $\{F_t(\mathbf{x})\}_t$ is monotonically increasing. Also

$$F_t(\mathbf{x}) \le \limsup_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{x}\right) < +\infty$$

with probability one. It follows that the $\{F_t(\mathbf{x})\}_t$ is almost surely a bounded sequence of real numbers and therefore the limit

$$F(\mathbf{x}) \triangleq \lim_{t\to\infty} F_t(\mathbf{x})$$

exists with probability one. As for the second term in (45), one can write

$$\begin{aligned}
\frac{1}{k} \sum_{0\le i<k: f(T_{LZ}^i \mathbf{x})>t} f\left(T_{LZ}^i \mathbf{x}\right) &= \frac{1}{r_k(\mathbf{x}, n_w)} \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{0\le i<k: f(T_{LZ}^i \mathbf{x})>t} f\left(T_{LZ}^i \mathbf{x}\right) \\
&\le \frac{1}{r_k(\mathbf{x}, n_w)} \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{0\le j<|T_{LZ}^k|\mathbf{x}: f(T^j \mathbf{x})>t} f\left(T^j \mathbf{x}\right)
\end{aligned}$$

An application of the ergodic theorem demonstrates that with probability one and for any t

$$\lim_{k\to\infty} \frac{1}{|T_{LZ}^k|\mathbf{x}} \sum_{0\le j<|T_{LZ}^k|\mathbf{x}: f(T^j \mathbf{x})>t} f\left(T^j \mathbf{x}\right) \quad \le \quad \sum_{i=t+1}^{\infty} i\mu\left(f(\mathbf{X}) = i\right)$$

From the assumption that $Ef(\mathbf{X}) < +\infty$, we can deduce that

$$\sum_{i=t+1}^{\infty} i\mu\left(f(\mathbf{X}) = i\right) = o(1)$$

as $t \to \infty$. Thus for every fixed t and taking the lim sup as $k \to \infty$ we obtain

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{0\le i<k} f(T_{LZ}^i \mathbf{x}) \le \lim_{k\to\infty} \frac{1}{k} \sum_{0\le i<k: f(T_{LZ}^i \mathbf{x})\le t} f\left(T_{LZ}^i \mathbf{x}\right) + o(1)\kappa$$

with probability one. Taking the limsup as $t \to \infty$,

$$\begin{aligned}
\limsup_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^i \mathbf{x}\right) &\le \lim_{t\to\infty} F_t(\mathbf{x}) + \limsup_{t\to\infty} o(1)\kappa \\
&= F(\mathbf{x})
\end{aligned}$$

On the other hand, taking the lim inf as $k \to \infty$ and then the limit as $t \to \infty$ we obtain

$$\liminf_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^{i}\mathbf{x}\right) \geq F(\mathbf{x})$$

The conclusion is that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} f\left(T_{LZ}^{i}\mathbf{x}\right) = F(\mathbf{x})$$

and the theorem is proved. □

6 Conclusions

The main motivation behind the present work was to understand the extent to which a simple short/long phrase length partitioning idea motivated by earlier works of Wyner and Ziv [3] and Ornstein and Weiss [6] could be used to give insight into the more detailed properties of SWLZ. Difficulties in using this idea to provide good upper estimates for the pointer bits arose quite rapidly, and it was somewhat suprising to see that we were capable of giving good upper and lower bounds for the phrase length bits. Further progress revealed that we could additionally give a lower bound on the redundancy which is roughly half of that of A.J. Wyner's [1] and Yang and Kieffer's [2], which hold for a closely related algorithm (see the Introduction). We insisted on obtaining results that would hold in the almost sure domain and in developing this theory we encountered the need to establish some basic asymptotic properties on the average number of phrases per symbol of SWLZ, which we addressed by a judicious application of the theory of Kieffer and Rahe [4] and the use of some interesting ideas concerning Lempel Ziv parsings of shifted versions of the same sequence. The problem of obtaining good upper estimates for the number of bits used to encode the pointers is the most important issue to be resolved in this direction.

7 Acknowledgements

The author wishes to thank Vittorio Castelli for many insightful discussions and specially the conversation that led to Lemma 7. The author also acknowledges feedback given by Ioannis Kontoyiannis and A.J. Wyner.

8 Appendix

Lemma 1 *If the source has exponential rates for entropy, then for any* $n_w \geq 1$, $EL(\mathbf{X}, n_w) < +\infty$

Proof. Let $\mathcal{L}_{n_w,i} = \left\{\mathbf{x}^i_{-n_w} : L(\mathbf{x}, n_w) = i\right\}$. From the Definition of the sliding window Lempel-Ziv algorithm it is not difficult to see that for any positive integer i, $|\mathcal{L}_{n_w,i}| \leq |\mathcal{A}|^{n_w}$. Let $\epsilon \in (0, h/2)$ be fixed. Since the source is assumed to have exponential rates for entropy, there exists i_ϵ and K_ϵ such that if $i > i_\epsilon$ then

$$\begin{aligned}
\mu_{n_w+i}\left(\mathcal{L}^i_{-n_w}\right) &\leq \mu_{n_w+i}\left(\mathcal{L}^i_{-n_w} \bigcap A^c_{n_w+i,\epsilon}\right) + \mu\left(A_{n_w+i,\epsilon}\right) \\
&\leq |\mathcal{A}|^{n_w} 2^{-(i+n_w)(h-\epsilon)} + 2^{-(i+n_w)K_\epsilon}
\end{aligned}$$

Forming the appropriate infinite summation, it is easy to see that there exists a constant M_ϵ such that

$$\begin{aligned}
EL(\mathbf{X}, n_w) &= \sum_{i=1}^{\infty} i\mu\left(L(\mathbf{X}, n_w) = i\right) \\
&\leq i_\epsilon + M_\epsilon
\end{aligned}$$

The Theorem is proved. □

Lemma 11 *The function* $t \log_2 \frac{1}{t}$ *is strictly increasing in the range* $(0, e^{-1})$ *and the function* $t \log_2 \log_2 \frac{1}{t}$ *is strictly increasing in the range* $(0, e^{-2/(1-\ln\ln 2)})$.

Proof. The first statement follows directly from the derivative of $t \log_2 \frac{1}{t}$. The derivative of the second function is $(-\ln\ln 2 + \ln\ln\frac{1}{t} - 1/\ln\frac{1}{t})/\ln 2$; using $\ln x \geq 1 - x^{-1}$ to lower bound said derivative gives the desired result.

8.1 All of the phrases of almost every process realization have finite length

Recall that

$$\mathcal{F}_{n_w} = \left\{\mathbf{x} : \forall\ k \geq 0,\quad L(T^k_{LZ}\mathbf{x}, n_w) < +\infty\right\}$$

In the Preliminaries we stated that $\mu(\mathcal{F}_{n_w}) = 1$, a useful fact that simplifies most of the discussions in this work by guaranteeing that the T_{LZ} operator can be safely iterated as many times as needed. This result follows from basic information theoretic considerations; the following simple results will be required in the proof:

Lemma 12 *There exists a constant K such that for all $a > 0$, $\log_2 a \le K\sqrt{a}$.*

Proof. From the widely used inequality $\log a \le a - 1$ we know that for all $a > 0$, $\log a^{1/2} \le \sqrt{a}$ and hence $\log_2 a \le K\sqrt{a}$ where $K = 2/\log 2$. $\Box$

Lemma 13 *There exists a $\hat{K}$ such that if $\eta \in [0, 1]$ then $H_b(\eta) \le \hat{K}\sqrt{\eta}$*

Proof. Using Lemma 12 and $\log a \le a - 1$

$$\begin{aligned} H_b(\eta) &= \eta \log_2 \frac{1}{\eta} + (1-\eta)\log_2 \frac{1}{1-\eta} \\ &\le \frac{2}{\log 2}\eta\frac{1}{\sqrt{\eta}} + \frac{(1-\eta)}{\log 2}\left(\frac{1}{1-\eta} - 1\right) \\ &\le \frac{2}{\log 2}\sqrt{\eta} + \frac{2}{\log 2}\eta \\ &\le \frac{4}{\log 2}\sqrt{\eta} \end{aligned}$$

where the last inequality follows from the fact that if $\eta \in [0, 1]$ then $\eta \le \sqrt{\eta}$. Defining $\hat{K} = 4/\log 2$ concludes the proof of the lemma. $\Box$

Lemma 14 *Let $\mathbf{x} \in \mathcal{F}^c_{n_w}$ and let r be any integer. Then $T^r\mathbf{x} \in \mathcal{F}^c_{n_w}$.*

Proof. Define $\mathbf{y} = T^r\mathbf{x}$ and assume that $\mathbf{y} \in \mathcal{F}_{n_w}$. Then by definition, for all $j \ge 0$, $L(T^j_{LZ}\mathbf{y}, n_w) < +\infty$ and therefore, for every $j \ge 0$ the quantity $|T^j_{LZ}|\mathbf{y}$ is well defined and the infinite chain of strict inequalities

$$|T^0_{LZ}|\mathbf{y} < |T^1_{LZ}|\mathbf{y} < |T^2_{LZ}|\mathbf{y} < \cdots \tag{46}$$

is true. Since it has been assumed that $\mathbf{x} \in \mathcal{F}^c_{n_w}$, there exists an integer $k \ge 0$ and $\delta \in \{1, 2, \cdots, n_w\}$ such that

$$\mathbf{x}_i = \mathbf{x}_{i-\delta} \qquad \forall\ i \ge |T^k_{LZ}|\mathbf{x} \tag{47}$$

Next let $\hat{k}$ be such that

$$r + |T^{\hat{k}}_{LZ}|\mathbf{y} \ge |T^k_{LZ}|\mathbf{x} \tag{48}$$

Note that such $\hat{k}$ is guaranteed to exist since due to the infinite chain of inequalities (46). For every $\hat{i} \ge |T^{\hat{k}}_{LZ}|\mathbf{y}$, we have that

$$\mathbf{y}_{\hat{i}} = \mathbf{x}_{\hat{i}+r} = \mathbf{x}_{\hat{i}+r-\delta} = \mathbf{y}_{\hat{i}-\delta}$$

where the second equality is obtained by combining Equations (47) and (48) with the assumption that $\hat{i} \geq |T_{LZ}^{\hat{k}}|\mathbf{y}$. This allows us to conclude that

$$L(T^{\hat{k}}\mathbf{y}, n_w) = +\infty$$

in clear violation of our assumption that $\mathbf{y} \in \mathcal{F}_{n_w}$. Therefore, $\mathbf{y} \in \mathcal{F}_{n_w}^c$ and the lemma is proved □We can now proceed to demonstrate that

Lemma 15 *If the pair (T, μ) is ergodic and $h > 0$, then for any database length $n_w \geq 1$, $\mu(\mathcal{F}_{n_w}) = 1$*

Proof. Using Lemma 14 with $r = 1$, we can easily demonstrate that $T\mathcal{F}_{n_w}^c \subset \mathcal{F}_{n_w}^c$. Using the lemma again with $r = -1$, we can see that $\mathcal{F}_{n_w}^c \subset T\mathcal{F}_{n_w}^c$, which implies that $\mathcal{F}_{n_w}^c = T\mathcal{F}_{n_w}^c$, and by ergodicity of the source, $\mu(\mathcal{F}_{n_w}^c) \in \{0, 1\}$. If $\mu(\mathcal{F}_{n_w}^c) = 0$ there is nothing to be proved, so we assume the only other possibility. The (necessarily finite) number of finite length phrases in every $\mathbf{x} \in \mathcal{F}_{n_w}^c$ is denoted by $\mathcal{K}(\mathbf{x}, n_w)$:

$$\mathcal{K}(\mathbf{x}, n_w) \triangleq |\{k : L(T_{LZ}^k\mathbf{x}, n_w) < +\infty\}| \qquad \mathbf{x} \in \mathcal{F}_{n_w}^c$$

Now define for every $j \geq 0$

$$\mathcal{F}_{n_w,j}^c \triangleq \left\{\mathbf{x} \in \mathcal{F}_{n_w}^c : \left|T_{LZ}^{\mathcal{K}(\mathbf{x},n_w)}\right|\mathbf{x} = j\right\}$$

Then

$$\begin{aligned} 1 &= \mu(\mathcal{F}_{n_w}^c) \\ &\overset{(a)}{=} \mu\left(\bigcup_{j\geq 0} \mathcal{F}_{n_w,j}^c\right) \\ &\overset{(b)}{=} \sum_{j=0}^{\infty} \mu(\mathcal{F}_{n_w,j}^c) \end{aligned}$$

where (a) follows from $\mathcal{F}_{n_w}^c = \bigcup_{j\geq 0} \mathcal{F}_{n_w,j}^c$, and (b) follows from the fact that the sets $\{\mathcal{F}_{n_w,j}^c\}_{j=0}^{\infty}$ are mutually disjoint and by σ-additivity of the measure.

Now let $\epsilon \in (0, 1)$, let j_ϵ^* be large enough so that

$$\sum_{j=0}^{j_\epsilon^*} \mu(\mathcal{F}_{n_w,j}^c) > 1 - \epsilon, \tag{49}$$

define

$$\mathcal{C}_{n_w,\epsilon} \triangleq \bigcup_{j=0}^{j_\epsilon^*} \mathcal{F}^c_{n_w,j} \tag{50}$$

and for every $n > 0$, let

$$\mathcal{D}_{n_w,\epsilon,n} \quad \triangleq \quad \left\{\mathbf{x}_0^{n-1} \in \mathcal{A}^n : \exists\ \mathbf{y} \in \mathcal{C}_{n_w,\epsilon} \text{ such that } \mathbf{x}_0^{n-1} = \mathbf{y}_0^{n-1}\right\}$$

Note that while every element of $\mathcal{C}_{n_w,\epsilon}$ is a sequence of infinite length, $\mathcal{D}_{n_w,\epsilon,n}$ contains vectors from $\mathcal{A}^n$. From the definition of $\mathcal{F}^c_{n_w,j}$, we know that if $\mathbf{x} \in \mathcal{C}_{n_w,\epsilon}$ then there exists a $\delta_{\mathbf{x}} \in \{1, \cdots, n_w\}$ such that

$$\mathbf{x}_i = \mathbf{x}_{i-\delta_{\mathbf{x}}} \qquad \forall\ i \geq j_\epsilon^*$$

In particular, $n_w + j_\epsilon^*$ symbols from $\mathcal{A}$ and the pointer $\delta_{\mathbf{x}} \in \{1, \cdots, n_w\}$ is all that suffices in order to specify, for any value of $n > 0$, the symbols that comprise $\mathbf{x}_0^{n-1}$. This observation enables us to deduce a bound on the cardinality of $\mathcal{D}_{n_w,\epsilon,n}$:

$$|\mathcal{D}_{n_w,\epsilon,n}| \leq n_w\, |\mathcal{A}|^{n_w + j_\epsilon^*} \qquad \forall\ n > 0 \tag{51}$$

Moreover, the following is true:

$$\begin{aligned} \mu_n(\mathcal{D}_{n_w,\epsilon,n}) &= \mu(\{\mathbf{x} \in \mathcal{A}_{-\infty}^{\infty} : \mathbf{x}_0^{n-1} \in \mathcal{D}_{n_w,\epsilon,n}\}) \\ &\overset{(a)}{\geq} \mu(\mathcal{C}_{n_w,\epsilon}) \\ &\overset{(b)}{>} 1 - \epsilon \end{aligned} \tag{52}$$

where (a) is due to the fact that

$$\mathcal{C}_{n_w,\epsilon} \subset \{\mathbf{x} \in \mathcal{A}_{-\infty}^{\infty} : \mathbf{x}_0^{n-1} \in \mathcal{D}_{n_w,\epsilon,n}\}$$

and (b) is due to Equation (49) and Definition (50). From the definition of the nth order approximation to the entropy rate we know that

$$
\begin{aligned}
nh_n &= \sum_{\mathbf{x}_0^{n-1}\in\mathcal{A}_0^{n-1}} \mu_n(\mathbf{x}_0^{n-1})\log_2\frac{1}{\mu_n(\mathbf{x}_0^{n-1})} \\
&= \sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}_{n_w,\epsilon,n}} \mu_n(\mathbf{x}_0^{n-1})\log_2\frac{1}{\mu_n(\mathbf{x}_0^{n-1})} + \sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}^c_{n_w,\epsilon,n}} \mu_n(\mathbf{x}_0^{n-1})\log_2\frac{1}{\mu_n(\mathbf{x}_0^{n-1})} \\
&= H_b(\mu(\mathcal{D}_{n_w,\epsilon,n})) \\
&\quad +\mu_n(\mathcal{D}_{n_w,\epsilon,n}) \sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}_{n_w,\epsilon,n}} \frac{\mu_n(\mathbf{x}_0^{n-1})}{\mu_n(\mathcal{D}_{n_w,\epsilon,n})}\log_2\frac{\mu_n(\mathcal{D}_{n_w,\epsilon,n})}{\mu_n(\mathbf{x}_0^{n-1})} \\
&\quad +\mu_n(\mathcal{D}^c_{n_w,\epsilon,n}) \sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}^c_{n_w,\epsilon,n}} \frac{\mu_n(\mathbf{x}_0^{n-1})}{\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})}\log_2\frac{\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})}{\mu_n(\mathbf{x}_0^{n-1})} \\
&\le \hat{K}\sqrt{\epsilon} + \sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}_{n_w,\epsilon,n}} \frac{\mu_n(\mathbf{x}_0^{n-1})}{\mu_n(\mathcal{D}_{n_w,\epsilon,n})}\log_2\frac{\mu_n(\mathcal{D}_{n_w,\epsilon,n})}{\mu_n(\mathbf{x}_0^{n-1})} \\
&\quad +\epsilon \sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}^c_{n_w,\epsilon,n}} \frac{\mu_n(\mathbf{x}_0^{n-1})}{\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})}\log_2\frac{\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})}{\mu_n(\mathbf{x}_0^{n-1})} \qquad (53)
\end{aligned}
$$

where (53) follows from Lemma 13, the use of $\mu_n(\mathcal{D}_{n_w,\epsilon,n}) \le 1$ and Equation (52).

Next note that $\mu_n(\cdot)/\mu_n(\mathcal{D}_{n_w,\epsilon,n})$ and $\mu_n(\cdot)/\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})$ are probability measures for sources with alphabet $\mathcal{D}_{n_w,\epsilon,n}$ and $\mathcal{D}^c_{n_w,\epsilon,n}$, respectively. It is a basic result of information theory that the entropy of a finite-alphabet source is upper bounded by the logarithm of the cardinality of its alphabet. Therefore, for all $n > 0$,

$$
\begin{aligned}
\sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}_{n_w,\epsilon,n}} \frac{\mu_n(\mathbf{x}_0^{n-1})}{\mu_n(\mathcal{D}_{n_w,\epsilon,n})}\log_2\frac{\mu_n(\mathcal{D}_{n_w,\epsilon,n})}{\mu_n(\mathbf{x}_0^{n-1})} &\le \log_2|\mathcal{D}_{n_w,\epsilon,n}| \\
&\overset{(a)}{\le} \log_2 n_w|\mathcal{A}|^{n_w+j^*_\epsilon} \qquad (54)
\end{aligned}
$$

where (a) follows from (51) and

$$
\begin{aligned}
\sum_{\mathbf{x}_0^{n-1}\in\mathcal{D}^c_{n_w,\epsilon,n}} \frac{\mu_n(\mathbf{x}_0^{n-1})}{\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})}\log_2\frac{\mu_n(\mathcal{D}^c_{n_w,\epsilon,n})}{\mu_n(\mathbf{x}_0^{n-1})} &\le \log_2|\mathcal{D}^c_{n_w,\epsilon,n}| \\
&\overset{(b)}{\le} \log_2|\mathcal{A}|^n \\
&= n\log_2|\mathcal{A}| \qquad (55)
\end{aligned}
$$

where (b) is due to the fact that $\mathcal{D}^c_{n_w} \subset \mathcal{A}^n$. Combining Equations (53), (54) and (55) we obtain that for every $\epsilon \in (0,1)$ and for every $n > j^*_\epsilon$

$$
nh_n \le \frac{\hat{K}\sqrt{\epsilon}}{n} + \frac{\log_2 n_w|\mathcal{A}|^{n_w+j^*_\epsilon}}{n} + \epsilon\log_2|\mathcal{A}|
$$

Taking the limit as $n \to \infty$, we obtain

$$h = \lim_{n \to \infty} h_n \leq \epsilon \log_2 |\mathcal{A}|$$

Since this is true for every $\epsilon \in (0,1)$, we have arrived at the conclusion that $h = 0$, in contradiction with the assumption that $h > 0$. □

Lemma 7 (restated) *Let* $\mathbf{z} \in \mathcal{F}_{n_w}$. *Then for any* $0 \leq i \leq L(\mathbf{z}, n_w)$,

$$L(\mathbf{z}, n_w) \leq i + L(T^i \mathbf{z}, n_w)$$

Proof. First note that for any $\mathbf{z} \in \mathcal{F}_{n_w}$, $L(\mathbf{z}, n_w) < +\infty$ and by Lemma 14, $T^i \mathbf{z} \in \mathcal{F}_{n_w}$ and therefore $L(T^i \mathbf{z}, n_w) < +\infty$ as well.

Case A $L(\mathbf{z}, n_w) = 1$. For this case it is sufficient to observe that $i \geq 0$ and $L(\cdot, \cdot) \geq 1$, see Definition (1).

Case B $L(\mathbf{z}, n_w) > 1$. We split this in two cases:

Case B1 $i \geq L(\mathbf{z}, n_w) - 1$. Then

$$\begin{aligned} L(\mathbf{z}, n_w) &\leq i + 1 \\ &\leq i + L(T^i \mathbf{z}, n_w) \end{aligned}$$

Case B2 $i < L(\mathbf{z}, n_w) - 1$. By Definition (1), there exists an integer $\delta_{\mathbf{z}} \in \{1, \cdots, n_w\}$ such that

$$\mathbf{z}_{j-\delta_{\mathbf{z}}} = \mathbf{z}_j \quad 0 \leq j < L(\mathbf{z}, n_w) - 1$$

Since $i \geq 0$, we may weaken this statement to

$$\mathbf{z}_{j-\delta_{\mathbf{z}}} = \mathbf{z}_j \quad i \leq j < L(\mathbf{z}, n_w) - 1 \tag{56}$$

Note that the interval specified for j is not empty as we are assuming that the strict inequality $i < L(\mathbf{z}, n_w) - 1$ holds. Applying the fact that for any integer k, $\mathbf{z}_k = \{T^i \mathbf{z}\}_{k-i}$ to Equation (56), we obtain

$$\{T^i \mathbf{z}\}_{j-\delta_{\mathbf{z}}-i} = \{T^i \mathbf{z}\}_{j-i} \quad 0 \leq j - i < L(\mathbf{z}, n_w) - 1 - i$$

Setting $\hat{j} = j - i$,

$$\{T^i \mathbf{z}\}_{\hat{j}-\delta_{\mathbf{z}}} = \{T^i \mathbf{z}\}_{\hat{j}} \quad 0 \leq \hat{j} < L(\mathbf{z}, n_w) - 1 - i$$

Combining this finding with the definition of $L(\cdot,\cdot)$ in (1), we obtain

$$L(T^i\mathbf{z},n_w) \geq L(\mathbf{z},n_w) - i$$

which is the statement of the lemma. □

Lemma 8 *Let* $\mathbf{x} \in \mathcal{F}_{n_w}$*. Then for any* $0 \leq d \leq L(\mathbf{x},n_w)$ *and for any integer* $k \geq 0$

$$\left|T_{LZ}^k\right|\mathbf{x} \leq d + \left|T_{LZ}^k\right|T^d\mathbf{x} \leq \left|T_{LZ}^{k+1}\right|\mathbf{x}$$

Proof. The proof is by induction. In the case $k = 0$ we know that

$$\begin{aligned} \left|T_{LZ}^0\right|\mathbf{x} &= 0 \\ \left|T_{LZ}^0 T^d\right|\mathbf{x} &= d \\ \left|T_{LZ}^1\mathbf{x}\right| &= L(\mathbf{x},n_w) \end{aligned}$$

Combining this observation with the assumption that $0 \leq d \leq L(\mathbf{x},n_w)$ shows that the theorem is true for $k = 0$. We now assume that

$$\left|T_{LZ}^{k-1}\right|\mathbf{x} \leq d + \left|T_{LZ}^{k-1}\right|T^d\mathbf{x} \leq \left|T_{LZ}^k\right|\mathbf{x} \tag{57}$$

Subtracting $\left|T_{LZ}^{k-1}\right|\mathbf{x}$ from all terms we obtain

$$0 \leq r_1 \leq L(T_{LZ}^{k-1}\mathbf{x})$$

where

$$r_1 = d + \left|T_{LZ}^{k-1}\right|T^d\mathbf{x} - \left|T_{LZ}^{k-1}\right|\mathbf{x}$$

and therefore by Lemma 7

$$L(T_{LZ}^{k-1}\mathbf{x},n_w) \quad \leq \quad r_1 + L(T^{r_1}T_{LZ}^{k-1}\mathbf{x},n_w)$$

Moreover, since

$$\begin{aligned} T^{r_1}T_{LZ}^{k-1}\mathbf{x} &= T^{d+\left|T_{LZ}^{k-1}\right|T^d\mathbf{x}}\mathbf{x} \\ &= T^{\left|T_{LZ}^{k-1}\right|T^d\mathbf{x}}T^d\mathbf{x} \\ &= T_{LZ}^{k-1}T^d\mathbf{x} \end{aligned}$$

we have that

$$L(T_{LZ}^{k-1}\mathbf{x}, n_w) \quad \leq d + \left|T_{LZ}^{k-1}\right| T^d\mathbf{x} - \left|T_{LZ}^{k-1}\right| \mathbf{x} + L(T_{LZ}^{k-1}T^d\mathbf{x}, n_w)$$

which implies

$$\left|T_{LZ}^{k}\right| \mathbf{x} \leq d + \left|T_{LZ}^{k}\right| T^d\mathbf{x} \tag{58}$$

thus establishing the first inequality of the theorem. To establish the second inequality we will make use of the fact that we just proved. Combining (57) and (58) we see that

$$d + \left|T_{LZ}^{k-1}\right| T^d\mathbf{x} \leq \left|T_{LZ}^{k}\right| \mathbf{x} \leq d + \left|T_{LZ}^{k}\right| T^d\mathbf{x}$$

Subtracting $d + \left|T_{LZ}^{k-1}\right| T^d\mathbf{x}$ from all terms, we obtain

$$0 \leq r_2 \leq L(T_{LZ}^{k-1}T^d\mathbf{x}, n_w)$$

where

$$r_2 = \left|T_{LZ}^{k}\right| \mathbf{x} - \left|T_{LZ}^{k-1}\right| T^d\mathbf{x} - d$$

Employing Lemma 7 again we have that

$$L(T_{LZ}^{k-1}T^d\mathbf{x}, n_w) \quad \leq \quad r_2 + L(T^{r_2}T_{LZ}^{k-1}T^d\mathbf{x}, n_w)$$

Moreover since

$$\begin{aligned} T^{r_2}T_{LZ}^{k-1}T^d\mathbf{x} &= T^{\left|T_{LZ}^{k}\right|\mathbf{x} - \left|T_{LZ}^{k-1}\right|T^d\mathbf{x} - d}T_{LZ}^{k-1}T^d\mathbf{x} \\ &= T^{\left|T_{LZ}^{k}\right|\mathbf{x}}\mathbf{x} \\ &= T_{LZ}^{k}\mathbf{x} \end{aligned}$$

we have that

$$L(T_{LZ}^{k-1}T^d\mathbf{x}, n_w) \leq \left|T_{LZ}^{k}\right| \mathbf{x} - \left|T_{LZ}^{k-1}\right| T^d\mathbf{x} - d + L(T_{LZ}^{k}\mathbf{x}, n_w)$$

or equivalently

$$d + \left|T_{LZ}^{k}\right| T^d\mathbf{x} \leq \left|T_{LZ}^{k+1}\right| \mathbf{x}$$

which concludes the proof of the theorem. □

Lemma 9 *Let d be any number and let $\{a_i\}_{i=0}^{\infty}$ and $\{b_i\}_{i=0}^{\infty}$ be sequences of positive integers such that $a_0 \leq b_0 + d \leq a_1 \leq b_1 + d \leq a_2 \leq b_2 + d \cdots$ and further assume that for all k, $a_k < a_{k+1}$ and $b_k < b_{k+1}$. Then*

$$\liminf_{k\to\infty} \frac{k}{a_k} \quad = \quad \liminf_{k\to\infty} \frac{k}{b_k}$$

and the lim sup *is finite.*

Proof. Since $a_k \leq b_k + d \leq a_{k+1}$ for all k, it follows that

$$\frac{k}{a_k} \geq \frac{k}{b_k + d} \geq \frac{k}{a_{k+1}}$$

or equivalently

$$\frac{k}{a_k} \geq \frac{k}{b_k + d} \geq \frac{k+1}{a_{k+1}} - \frac{1}{a_{k+1}}$$

Taking the lim inf as $k \to \infty$ we obtain

$$\liminf_{k\to\infty} \frac{k}{a_k} \geq \liminf_{k\to\infty} \frac{k}{b_k + d} \geq \liminf_{k\to\infty} \frac{k}{a_k}$$

where $\lim_k \frac{1}{a_{k+1}} = \limsup_k \frac{1}{a_{k+1}} = \liminf_k \frac{1}{a_{k+1}} = 0$ due to the fact that the sequence $\{a_k\}$ is assumed to be increasing. Finally,

$$\begin{aligned}
\liminf_{k\to\infty} \frac{k}{b_k + d} &= \frac{1}{\limsup_{k\to\infty} \frac{b_k+d}{k}} \\
&= \frac{1}{\limsup_{k\to\infty} \frac{b_k}{k}} \\
&= \liminf_{k\to\infty} \frac{k}{b_k}
\end{aligned}$$

This concludes the proof of the lemma. □

References

[1] A.J. Wyner. The Redundancy and Distribution of the Phrase Lengths of the Fixed-Database Lempel-Ziv Algorithm. *IEEE Transactions on Information Theory*, pages 1452–1464, 1997.

[2] En-Hui Yang, John C. Kieffer. On the redundancy of the fixed-database Lempel-Ziv algorithm for phi-mixing sources. *IEEE Transactions on Information Theory*, pages 47–65, 1997.

[3] A. Wyner, J. Ziv. The sliding-window Lempel-Ziv algorithm is asymptotically optimal. *Proceedings of the IEEE*, 82:872–877, 1994.

[4] John C. Kieffer, Maurice Rahe. The Pointwise Ergodic Theorem For Transformations Whose Orbits Contain or are Contained in the Orbits of a measure-preserving Transformation. *Can. J. Math*, XXXIV(6):1303–1318, 1982.

[5] J. Ziv, A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.

[6] D. Ornstein, B. Weiss. Entropy and Data Compression. *IEEE Transactions on Information Theory*, 1993.

[7] S. A. Savari. Redundancy of the Lempel-Ziv String Matching Code. *IEEE Transactions on Information Theory*, pages 787–791, 1998.

[8] Paul C. Shields. Performance of LZ Algorithms on Individual Sequences. *IEEE Transactions on Information Theory*, IT-45:1283–1288, 1999.

[9] Luis A. Lastras-Montaño. An Almost Sure Convergence Proof of the Sliding Window Lempel Ziv Algorithm. *Proceedings of the 2002 IEEE International Symposium on Information Theory*, page 121, 2002.

[10] Vittorio Castelli. Personal Communication.

[11] A.J. Wyner. *String Matching Theorems and Applications to Data Compression and Statistics.* PhD thesis, 1993.

[12] P. C. Shields. *The Ergodic Theory of Discrete Sample Paths.* American Mathematical Society, 1996.

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

[14] P.E. Bender, J.K. Wolf. New Asymptotic Bounds and Improvements on the Lempel-Ziv Data Compression Algorithm. *IEEE Transactions on Information Theory*, 37:721–729, 1991.

[15] A.D Wyner, A.J. Wyner. Improved redundancy of a Version of the Lempel-Ziv Algorithm. *IEEE Transactions on Information Theory*, pages 723–731, 1995.

[16] Luis A. Lastras-Montaño. Non-Asymptotic Upper Bounds on the Probability of the ϵ-atypical set for Markov chains. *Submitted to the IEEE Transactions on Information Theory.*

[17] A. Wyner, J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35:125–1258, 1989.