RC 22085 (W0106-007) 5 June 2001 Electrical Engineering

# IBM Research Report

## An Almost Sure Convergence Proof of the Sliding Window Lempel Ziv Algorithm

**Luis A. Lastras**
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

**Research Division**
**Almaden - Austin - Beijing - Haifa - T. J. Watson - Tokyo - Zurich**

# An almost sure convergence proof of the sliding-window Lempel Ziv algorithm

Luis Lastras *

June 4, 2001

**Abstract**

An almost sure convergence proof of the sliding window Lempel Ziv algorithm (LZ77) is given. The proof is valid for those sources that in addition of being stationary and ergodic, have *exponential rates* for entropy.

## 1 Introduction

The sliding window Lempel-Ziv data compression algorithm was first proposed by Ziv and Lempel [1] in 1977. The optimality of the algorithm was established by Wyner and Ziv [2] who showed that as the window length approaches infinity, the *expected* compression ratio of 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 [3] include as a corollary an almost sure optimality proof of an infinite window version of the algorithm under the same source assumptions.

In this work, we show that if the source is stationary, ergodic, and possesses exponential rates for entropy (for a precise definition of these terms, see the Preliminaries), the compression ratio of the algorithm on each individual sequence approaches the entropy of the source as the window length approaches infinity, except for a set of source sequences with measure zero. This result contrasts with the infinite window assumption inherent in Ornstein and Weiss [3], and gives more precise information than the the results of Wyner and Ziv [2], since the statements are of a pointwise, rather than average nature. However, it must be noted that our proof is valid only for a more restricted class of sources, due to the assumption that the source has exponential rates for entropy (Markov chains are among the processes that have this property [4]).

The basic tool used in this work is Kac's formula for the expected recurrence time of a pattern, which first appeared in the context of Lempel-Ziv algorithms in the work of Willems [5].

## 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

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

*Luis Lastras is with the Multimedia Technologies Department, IBM TJ Watson Research Center, Yorktown Heights, NY 10598

Let $(\mathcal{A}_{-\infty}^{\infty}, \Sigma, \mu)$ be a probability space. The probability law $\mu$ 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 $\mu$ 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 $\mu$ is defined as

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

It is also assumed that $\mu$ has *exponential rates* for entropy. 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 $l$,

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

*Note:* Markov chains are among the class of processes that have this property (see [4] p. 165).

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

$$p_k = p_{k-1} + 1 + \max\{l \ge 0 : R(T^{p_{k-1}}\mathbf{x}, l) \le n_w\} \tag{1}$$

The $n_w$-memory Lempel-Ziv parsing the first $n$ symbols of $\mathbf{x}$ is obtained from the $n_w$-memory parsing of $\mathbf{x}$ in the following manner: Let

$$c_n = \min\{k : p_k \ge n\} \tag{2}$$

and define

$$p_{k,n} = \begin{cases} p_k & \text{if} \quad 0 \le k < c_n \\ n & \text{if} \quad k = c_n \end{cases} \tag{3}$$

The $k$th $(1 \le k \le c_n)$ *phrase* of the parsing is $\mathbf{x}_{p_{k-1,n}}^{p_{k,n}-1}$ and its associated length is

$$l_{k,n} = \begin{cases} 1 + \max\{l \ge 0 : R(T^{p_{k-1,n}}\mathbf{x}, l) \le n_w\} & \text{if} \quad 1 \le k < c_n \\ n - p_{k-1,n} & \text{if} \quad k = c_n \end{cases} \tag{4}$$

Note that for clarity purposes, the notation used for $c_n$ and $p_{k,n}$ did not explicitly state their dependence on $\mathbf{x}$ and $n_w$; the notation $c_n(\mathbf{x}, n_w)$ and $p_{k,n}(\mathbf{x}, n_w)$ will also be used when convenient.

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. Then, for each $k \in \{1, \cdots, c_n\}$, the source encoder transmits sequentially the following information

- The phrase length $l_{k,n}$ (using $\mathrm{LEN}(l_{k,n})$ bits)
- The pointer $R(T^{p_{k-1}}\mathbf{x}, l_{k,n}-1)$ (using $\lceil \log_2 n_w \rceil$ bits)
- The value of the symbol $x_{p_k-1}$ (using $\lceil \log_2 |\mathcal{A}| \rceil$ bits)

It will be assumed that the phrase length will be encoded using the technique described in the appendix of Wyner and Ziv [2] for comma-free encoding of positive integers. Alternative techniques can be found in Elias [6].

The cost of encoding the $k$th phrase is equal to

$$\mathrm{LEN}(l_{k,n}) + \lceil \log_2 n_w \rceil + \lceil \log_2 |\mathcal{A}| \rceil \qquad 1 \le k \le c_n \tag{5}$$

where $\mathrm{LEN} : Z \to Z$ is the length function associated with the code of [2]. This length function possesses the following (weak) upper bound:

$$\mathrm{LEN}(l) \le \gamma \log_2 l \tag{6}$$

where $\gamma$ is some constant. The bitstream produced by the encoder can be pictured as a sequence of concatenated packets $\Gamma\Theta_1\Theta_2\cdots\Theta_{c_n}$, where $\Gamma$ represents the bits associated with the uncoded transmission of $\mathbf{x}_{-n_w}^{-1}$ and $\Theta_i$ refers to the bits of phrase $i$ ($1 \le i \le c_n$). From the description of the operation of the encoder, it is clear that for any $i$, there exists a decoder $f_i : \{0,1\}^* \to \mathcal{A}^*$ such that

$$f_i(\Gamma\Theta_1\Theta_2\cdots\Theta_i) = \mathbf{x}_{-n_w}^{p_i-1}, \qquad 1 \le i \le c_n$$

The number of bits-per-symbol generated by the $n_w$-memory LZ algorithm when applied to the first $n$ symbols of the source sample $\mathbf{x}$ is defined as

$$b_n(\mathbf{x}, n_w) = \frac{n_w \lceil \log_2 |\mathcal{A}| \rceil}{n+n_w} + \frac{1}{n+n_w} \sum_{k=1}^{c_n(\mathbf{x},n_w)} \mathrm{LEN}(l_{k,n}(\mathbf{x}, n_w)) + \lceil \log_2 n_w \rceil + \lceil \log_2 |\mathcal{A}| \rceil$$

At the penalty of a small notation overload, also define $b_n(\mathbf{y}_{-n_w}^{n-1}, n_w) = b_n(\mathbf{x}, n_w)$ where $\mathbf{x}$ is any sequence for which $\mathbf{x}_{-n_w}^{n-1} = \mathbf{y}_{-n_w}^{n-1}$.

It is the purpose of this work to show the existence of a set $E$ with $\mu(E) = 1$ such that for all $\mathbf{x} \in E$,

$$\lim_{n_w\to\infty} \lim_{n\to\infty} b_n(\mathbf{x}, n_w) \quad = \quad h \tag{7}$$

# 3 Results

Our efforts will be directed towards establishing the following Theorems, which clearly can be combined in order to reach the desired conclusion (7).

**Theorem 1** *There exists a set $A$ with $\mu(A) = 1$ such that for all $\mathbf{x} \in A$,*

$$\limsup_{n_w\to\infty} \limsup_{n\to\infty} b_n(\mathbf{x}, n_w) \quad \le \quad h$$

**Theorem 2** *There exists a set $B$ with $\mu(B) = 1$ such that for all $\mathbf{x} \in B$*

$$h \leq \liminf_{n_w \to \infty} \liminf_{n \to \infty} b_n(\mathbf{x}, n_w)$$

*Note:* The proof of Theorem 2 is based on arguments published in [4], and will be relegated to the Appendix.

### 3.1 Proof of Theorem 1

The basic idea is to classify the phrases generated by the LZ algorithm in *short* and *long*. Define $\ell : Z_+ \times R_+ \longrightarrow Z$ as:

$$\ell(n_w, \xi) \stackrel{\Delta}{=} \left\lfloor \frac{\log_2 n_w}{h + \xi} \right\rfloor$$

and define $\mathcal{A}_{-\infty}^{\infty} \times Z_+ \times R \to Z$ functions $c_n^S$ and $c_n^L$ as

$$c_n^S(\mathbf{x}, n_w, \xi) = |1 \leq k \leq c_n(\mathbf{x}, n_w) : l_{k,n}(\mathbf{x}, n_w) < \ell(n_w, \xi) + 1\}| \quad (8)$$

$$c_n^L(\mathbf{x}, n_w, \xi) = |1 \leq k \leq c_n(\mathbf{x}, n_w) : l_{k,n}(\mathbf{x}, n_w) \geq \ell(n_w, \xi) + 1\}| \quad (9)$$

so that $c_n^S$ and $c_n^L$ denote the number of short and long phrases in the $n_w$-memory parsing of the first $n$ symbols of $\mathbf{x}$, respectively.

We will upper bound $b_n(\mathbf{x}, n_w)$ by splitting it in four pieces, each of which will be upper bounded separately. Define

$$\begin{aligned}
b_n^W(\mathbf{x}, n_w) &= \frac{n_w}{n + n_w} \lceil \log_2 |\mathcal{A}| \rceil \\
b_n^{L,P}(\mathbf{x}, n_w, \xi) &= \frac{\lceil \log_2 n_w \rceil}{n + n_w} c_n^L(\mathbf{x}, n_w, \xi) \\
b_n^{S,P}(\mathbf{x}, n_w, \xi) &= \frac{\lceil \log_2 n_w \rceil}{n + n_w} c_n^S(\mathbf{x}, n_w, \xi) \\
b_n^O(\mathbf{x}, n_w) &= \frac{1}{n + n_w} \sum_{k=1}^{c_n(\mathbf{x}, n_w)} \lceil \log_2 \mathcal{A} \rceil + \mathrm{LEN}(l_{k,n}(\mathbf{x}, n_w))
\end{aligned}$$

so that $b_n^W$ denotes the bits used in the uncoded transmission of the initial $n_w$-long window; $b_n^{S,P}$ and $b_n^{L,P}$ denote the amount of bits invested in the encoding of the *pointer* for the *short* and *long* phrases, respectively, and $b_n^O$ denotes the remaining bits.

Note that for any $\epsilon > 0$,

$$b_n(\mathbf{x}, n_w) = b_n^W(\mathbf{x}, n_w) + b_n^{S,P}(\mathbf{x}, n_w, \epsilon) + b_n^{L,P}(\mathbf{x}, n_w, \epsilon) + b_n^O(\mathbf{x}, n_w) \quad (10)$$

It is not difficult to see that for all $\mathbf{x}$,

$$\lim_{n_w \to \infty} \lim_{n \to \infty} b_n^W(\mathbf{x}, n_w) = 0 \quad (11)$$

and that for all $\epsilon > 0$ and $\mathbf{x}$,

$$\limsup_{n_w \to \infty} \limsup_{n \to \infty} b_n^{L,P}(\mathbf{x}, n_w, \epsilon) \leq h + \epsilon \quad (12)$$

It will also be shown that for all $\epsilon > 0$, there exists a set $A_\epsilon$ with $\mu(A_\epsilon) = 1$ such that for all $\mathbf{x} \in A_\epsilon$

$$\lim_{n_w \to \infty} \lim_{n \to \infty} b_n^{S,P}(\mathbf{x}, n_w, \epsilon) = 0 \quad (13)$$

$$\lim_{n \to \infty} b_n^{O}(\mathbf{x}, n_w) = 0 \quad (14)$$

and hence for all $\mathbf{x} \in A_\epsilon$,

$$\limsup_{n_w \to \infty} \limsup_{n \to \infty} b_n(\mathbf{x}, n_w) \leq h + \epsilon$$

From this, we obtain the conclusion of the Theorem as follows. Let $\{\epsilon_k\}_{k=1}^{\infty}$ be a sequence of positive numbers such that $\lim_{k \to \infty} \epsilon_k = 0$, and for each $k$, let $A_{\epsilon_k}$ be the associated set of full measure. Define $A = \bigcap_{k \geq 1} A_{\epsilon_k}$ and note that $\mu(A) = 1$ (this follows from the $\sigma$-subadditivity of $\mu$). Then, for all $\mathbf{x} \in A$ and for all $k$,

$$\limsup_{n_w \to \infty} \limsup_{n \to \infty} b_n(\mathbf{x}, n_w) \leq h + \epsilon_k$$

Taking the limit as $k \to \infty$ gives the statement of the Theorem.

### 3.1.1 Bounding $b_n^W$ and $b_n^{L,P}$

The fact that for all $\mathbf{x}$, $\lim_{n_w \to \infty} \lim_{n \to \infty} b_n^W(\mathbf{x}, n_w) = 0$ is a simple consequence of the definition of $b_n^{L,P}(\mathbf{x}, n_w)$. For any $\epsilon > 0$, $n$ ,$n_w$ and $\mathbf{x}$, the number of long phrases is upper bounded by

$$\begin{aligned} c_n^L(\mathbf{x}, n_w, \epsilon) &\leq \frac{n}{\ell(n_w, \epsilon) + 1} \\ &\leq \frac{n}{\frac{\log_2 n_w}{h+\epsilon} - 1 + 1} \\ &= \frac{n(h+\epsilon)}{\log_2 n_w} \end{aligned} \quad (15)$$

and hence,

$$\begin{aligned} b_n^{L,P}(\mathbf{x}, n_w, \epsilon) &\leq \frac{n}{n + n_w} \frac{1 + \log_2 n_w}{\log_2 n_w} (h + \epsilon) \\ &\leq \frac{1 + \log_2 n_w}{\log_2 n_w} (h + \epsilon) \end{aligned} \quad (16)$$

From (16) it is clear that for all $\epsilon > 0$ and $\mathbf{x}$,

$$\limsup_{n_w \to \infty} \limsup_{n \to \infty} b_n^{L,P}(\mathbf{x}, n_w, \epsilon) \leq h + \epsilon$$

## 3.2 Bounding $b_n^{S,P}$ and $b_n^O$

This subsection makes use of the following Theorem, which constitutes the central result of this paper:

**Theorem 3** *For all $\epsilon > 0$ there exists a $\psi > 0$, a $n_w^*$ and a set $A$ with $\mu(A) = 1$ such that for all $\mathbf{x} \in A$ and for all $n_w > n_w^*$*

$$\limsup_{n \to \infty} \frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} \leq n_w^{-\psi}$$

Let $\epsilon > 0$ be fixed. Let $\psi > 0$, $n_w^*$ and $A_\epsilon$ be the objects provided by the Theorem. Then for all $\mathbf{x} \in A_\epsilon$, and for all $n_w > n_w^*$,

$$\begin{aligned}\limsup_{n\to\infty} b_n^{S,P}(\mathbf{x}, n_w, \epsilon) &\leq (1 + \log_2 n_w) \limsup_{n\to\infty} \frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} \\ &\leq (1 + \log_2 n_w) n_w^{-\psi}\end{aligned}$$

Taking the limit as $n_w \to 0$ we conclude that for all $\mathbf{x} \in A_\epsilon$,

$$\lim_{n_w\to\infty} \lim_{n\to\infty} b_n^{S,P}(\mathbf{x}, n_w, \epsilon) = 0$$

Now note that for all $n$, $n_w$ and $\mathbf{x}$,

$$\begin{aligned}b_n^O(\mathbf{x}, n_w) &\leq \frac{1}{n} \sum_{k=1}^{c_n(\mathbf{x},n_w)} \lceil \log_2 |\mathcal{A}| \rceil + \mathrm{LEN}(l_{k,n}(\mathbf{x}, n_w)) \\ &\overset{(a)}{\leq} \frac{1}{n} \sum_{k=1}^{c_n(\mathbf{x},n_w)} \lceil \log_2 |\mathcal{A}| \rceil + \gamma \log_2(l_{k,n}(\mathbf{x}, n_w)) \\ &= \frac{c_n(\mathbf{x}, n_w)}{n} \left( \lceil \log_2 |\mathcal{A}| \rceil + \gamma \sum_{k=1}^{c_n(\mathbf{x},n_w)} \frac{\log_2 (l_{k,n}(\mathbf{x}, n_w))}{c_n(\mathbf{x}, n_w)} \right) \\ &\overset{(b)}{\leq} \frac{c_n(\mathbf{x}, n_w)}{n} \left( \lceil \log_2 |\mathcal{A}| \rceil + \gamma \log_2 \left( \frac{1}{c_n(\mathbf{x}, n_w)} \sum_{k=1}^{c_n(\mathbf{x},n_w)} l_{k,n}(\mathbf{x}, n_w) \right) \right) \\ &\overset{(c)}{=} \frac{c_n(\mathbf{x}, n_w)}{n} \left( \lceil \log_2 |\mathcal{A}| \rceil + \gamma \log_2 \left( \frac{n}{c_n(\mathbf{x}, n_w)} \right) \right) \qquad (17)\end{aligned}$$

where in (a) $\gamma$ is the constant defined in the Preliminaries in Equation (6), (b) follows from the Jensen inequality and the convexity $\bigcap$ of the logarithm and in (c), the summation has been simplified using Equations (1), (3) and (4).

The following Lemma is proved in the Appendix.

**Lemma 1** *There exists a constant $K$ such that for all $a$, $\log_2 a \leq K\sqrt{a}$*

With its aid, we can upper bound (17) as

$$\begin{aligned}b_n^O(\mathbf{x}, n_w) &\leq \frac{c_n(\mathbf{x}, n_w)}{n} \left( \lceil \log_2 |\mathcal{A}| \rceil + \gamma K \left( \frac{c_n(\mathbf{x}, n_w)}{n} \right)^{-1/2} \right) \\ &= \alpha_1 \frac{c_n(\mathbf{x}, n_w)}{n} + \alpha_2 \left( \frac{c_n(\mathbf{x}, n_w)}{n} \right)^{1/2} \\ &\leq \alpha_3 \left( \frac{c_n(\mathbf{x}, n_w)}{n} \right)^{1/2} \\ &= \alpha_3 \left( \frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} + \frac{c_n^L(\mathbf{x}, n_w, \epsilon)}{n} \right)^{1/2} \\ &\leq \alpha_3 \left( \frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} + \frac{h + \epsilon}{\log_2 n_w} \right)^{1/2}\end{aligned}$$

where the last inequality follows from (15). After taking the limit as $n \to \infty$, we obtain

$$\limsup_{n\to\infty} b_n^O(\mathbf{x}, n_w) \quad \leq \quad \alpha_3 \left( \limsup_{n\to\infty} \frac{c_n^S(\mathbf{x}, n_w)}{n} + \frac{h+\epsilon}{\log_2 n_w} \right)^{1/2}$$

and hence for all $\mathbf{x} \in A_\epsilon$, and for all $n_w > n_w^*$,

$$\limsup_{n\to\infty} b_n^O(\mathbf{x}, n_w) \quad \leq \quad \alpha_3 \left( n_w^{-\psi} + \frac{h+\epsilon}{\log_2 n_w} \right)^{1/2}$$

Finally, taking the limit as $n_w \to \infty$, we conclude that for all $\mathbf{x} \in A_\epsilon$,

$$\lim_{n_w\to\infty} \lim_{n\to\infty} b_n^O(\mathbf{x}, n_w) = 0$$

□

### 3.3 Proof of Theorem 3

In addition to Kac's Lemma (stated below as Lemma 2), the proof of Theorem 3 makes use of Lemmas 3, 4 and 5, each of which is stated and proved below.

**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 [7]. □

**Lemma 3** *For all $\epsilon > 0$, $\mathbf{x}$, $n$ and $n_w$,*

$$\frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} \leq \frac{1}{n} + \frac{1}{n} \sum_{r=0}^{n-1} a(T^r \mathbf{x}, n_w, \epsilon)$$

*where*

$$a(\mathbf{x}, n_w, \epsilon) \quad = \quad \begin{cases} 1 & \text{if } R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w \\ 0 & \text{otherwise.} \end{cases}$$

**Proof.** This Lemma is a consequence of the definitions made in Sections 2 and 3.

$$\begin{aligned} c_n^S(\mathbf{x}, n_w, \epsilon) &\overset{(a)}{=} |1 \leq k \leq c_n(\mathbf{x}, n_w) : l_{k,n}(\mathbf{x}, n_w) < \ell(n_w, \epsilon) + 1\}| \\ &\leq 1 + |\{1 \leq k \leq c_n(\mathbf{x}, n_w) - 1 : l_{k,n}(\mathbf{x}, n_w) < \ell(n_w, \epsilon) + 1\}| \\ &\overset{(b)}{\leq} 1 + |\{1 \leq k \leq c_n(\mathbf{x}, n_w) - 1 : 1 + \max\{l \geq 0 : R(T^{p_{k-1,n}}\mathbf{x}, l) \leq n_w\} < \ell(n_w, \epsilon) + 1\}| \\ &\leq 1 + |\{0 \leq r \leq n-1 : \max\{l \geq 0 : R(T^r \mathbf{x}, l) \leq n_w\} < \ell(n_w, \epsilon)\}| \\ &\overset{(c)}{=} 1 + |\{0 \leq r \leq n-1 : R(T^r \mathbf{x}, \ell(n_w, \epsilon)) > n_w\}| \end{aligned} \tag{18}$$

where (a) comes from the definition of $c_n^S(\mathbf{x}, n_w)$ (Equation (8)), (b) comes from the definition of $l_{k,n}$ (Equation (4)) and (c) comes from the fact that $\max\{l \geq 0 : R(T\mathbf{x}, l) \leq n_w\} < \ell(n_w, \epsilon)$ if and only if $R(T\mathbf{x}, \ell(n_w, \epsilon)) > n_w$. Finally, we rephrase (18) as

$$\frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} \leq \frac{1}{n} + \frac{1}{n}\sum_{r=0}^{n-1} a(T^r\mathbf{x}, n_w, \epsilon)$$

which is the statement of the Lemma. □

**Lemma 4** *For all $\epsilon > 0$ there exists a $\delta > 0$ and a $l^*$ such that for $l \geq l^*$,*

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

Let $\epsilon > 0$ be fixed. Define, for each $l$,

$$T^l \triangleq \left\{\mathbf{x}_0^{l-1} : \left|-\frac{1}{l}\log_2 \mu_l(\mathbf{x}_0^{l-1}) - h\right| < \frac{\epsilon}{2}\right\}$$

and note that for any $l$,

$$\begin{aligned}
&\mu\left(\left\{\mathbf{x} : R\left(\mathbf{x}, l\right) > 2^{l(h+\epsilon)}\right\}\right) \\
&\quad\leq \quad 1 - \mu_l\left(T^l\right) + \sum_{\mathbf{x}_0^{l-1} \in T^l} \mu\left(\left\{\mathbf{x} : R\left(\mathbf{x}, l\right) > 2^{l(h+\epsilon)}\right\} | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1}\right) \mu_l\left(\mathbf{x}_0^{l-1}\right) \qquad (19)
\end{aligned}$$

The summation is upper bounded as follows:

$$\begin{aligned}
&\sum_{\mathbf{x}_0^{l-1} \in T_l} \mu\left(\left\{\mathbf{x} : R(\mathbf{x}, l) > 2^{l(h+\epsilon)}\right\} | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1}\right) \mu_l\left(\mathbf{x}_0^{l-1}\right) \\
&\overset{(a)}{\leq} \quad \sum_{\mathbf{x}_0^{l-1} \in T_l} \frac{E\left[R(\mathbf{X}, l) | \mathbf{X}_0^{l-1} = \mathbf{x}_0^{l-1}\right]}{2^{l(h+\epsilon)}} \mu_l\left(\mathbf{x}_0^{l-1}\right) \\
&\overset{(b)}{=} \quad \sum_{\mathbf{x}_0^{l-1} \in T_l} \frac{\frac{1}{\mu_l(\mathbf{x}_0^{l-1})}}{2^{l(h+\epsilon)}} \mu_l\left(\mathbf{x}_0^{l-1}\right) \\
&\overset{(c)}{\leq} \quad \sum_{\mathbf{x}_0^{l-1} \in T_l} 2^{-l\epsilon/2} \mu_l\left(\mathbf{x}_0^{l-1}\right) \\
&\leq \quad 2^{-l\epsilon/2} \qquad (20)
\end{aligned}$$

where $(a)$ follows from the Markov inequality, $(b)$ follows from Kac's Lemma and $(c)$ follows from the definition of $T^l$.

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 $l$,

$$\mu_l\left(T^l\right) \geq 1 - 2^{-lK} \qquad (21)$$

Now define $\delta = \frac{1}{2}\min(\epsilon/2, K)$ and choose $l^* > 1/\delta$. Using (19), (20) and (21) 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 Lemma. □

**Lemma 5** *For all $\epsilon > 0$ there exists a $\psi > 0$ and a $n_w^*$ such that for all $n_w > n_w^*$*

$$\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) \leq n_w^{-\psi}$$

**Proof.** Let $\epsilon > 0$ be fixed and let $\delta > 0$ and $l^*$ be the constants provided by Lemma 4. Choose $n_w^*$ such that

$$\begin{aligned}\ell(n_w^*, \epsilon) &> l^* \\ \frac{1}{2}\log_2 n_w^* &> h + \epsilon\end{aligned}$$

and choose

$$\psi = \frac{\delta}{2(h+\epsilon)}$$

Provided that $n_w > n_w^*$,

$$\begin{aligned}\mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w\}\right) &\overset{(a)}{\leq} \mu\left(\left\{\mathbf{x} : R\left(\mathbf{x}, \ell(n_w, \epsilon)\right) > 2^{\ell(n_w,\epsilon)(h+\epsilon)}\right\}\right) \\ &\overset{(b)}{\leq} 2^{-\ell(n_w,\epsilon)\delta} \qquad (22)\end{aligned}$$

where $(a)$ follows from the definition of $\ell(n_w, \epsilon)$ and $(b)$ follows from the fact that $n_w > n_w^*$ implies that $\ell(n_w, \epsilon) > \ell(n_w^*, \epsilon) > l^*$ and therefore the conclusion of Lemma 4 applies. Using the fact that $\frac{1}{2}\log_2 n_w^* > h + \epsilon$ we can lower bound $\ell(n_w, \epsilon)$ as follows:

$$\ell(n_w, \epsilon) = \left\lfloor \frac{\log_2 n_w}{h+\epsilon} \right\rfloor \geq \frac{\log_2 n_w}{h+\epsilon} - 1 \geq \frac{\log_2 n_w}{2(h+\epsilon)}$$

Using this lower bound we conclude that for all $n_w > n_w^*$,

$$\begin{aligned}2^{-\delta\ell(n_w,\epsilon)} &\leq n_w^{-\frac{\delta}{2(h+\epsilon)}} \\ &= n_w^{-\psi}\end{aligned}$$

Combining this with (22) gives the desired result. □

We are now ready to prove Theorem 3:

**Theorem 3** *For all $\epsilon > 0$ there exists a $\psi > 0$, a $n_w^*$ and a set $A$ with $\mu(A) = 1$ such that for all $\mathbf{x} \in A$ and for all $n_w > n_w^*$*

$$\limsup_{n\to\infty} \frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} \leq n_w^{-\psi} \qquad (23)$$

**Proof.** Fix $\epsilon > 0$ and let $\psi$ and $n_w^*$ be the constants provided by Lemma 5. Let $a : \mathcal{A}_{-\infty}^{\infty} \times Z_+ \times R_+$ be the function defined in Lemma 3. For each $n_w > n_w^*$, the ergodic Theorem guarantees the existence of a set $A_{n_w}$ with $\mu(A_{n_w}) = 1$ such that for all $\mathbf{x} \in A_{n_w}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r=0}^{n-1} a(T^r \mathbf{x}, n_w, \epsilon) = \int a(\mathbf{x}, n_w, \epsilon) d\mu(\mathbf{x})$$

Next, define

$$A = \bigcap_{n_w > n_w^*} A_{n_w}$$

and note that

$$\begin{aligned} \mu(A) &= 1 - \mu\left( \bigcup_{n_w > n_w^*} A_{n_w}^c \right) \\ &\overset{(a)}{\geq} 1 - \sum_{n_w > n_w^*} \mu(A_{n_w}^c) \\ &= 1 \end{aligned}$$

where (a) follows from the $\sigma$-subadditivity of $\mu$. This allows us to conclude that $\mu(A) = 1$. Now, for $\mathbf{x} \in A$ and $n_w > n_w^*$,

$$\begin{aligned} \lim_{n\to\infty} \frac{c_n^S(\mathbf{x}, n_w, \epsilon)}{n} &\overset{(a)}{\leq} \lim_{n\to\infty} \frac{1}{n} + \frac{1}{n} \sum_{r=0}^{n-1} a(T^r \mathbf{x}, n_w, \epsilon) \\ &= \int a(\mathbf{x}, n_w, \epsilon) d\mu(\mathbf{x}) \\ &= \mu\left(\{\mathbf{x} : R(\mathbf{x}, \ell(n_w, \epsilon)) > n_w)\}\right) \\ &\overset{(b)}{\leq} n_w^{-\psi} \end{aligned}$$

where (a) follows from Lemma 3 and (b) follows from Lemma 5. □

# 4 Conclusions

We have established that the compression ratio of the sliding window Lempel Ziv algorithm converges almost surely to the entropy of the source as the window size grows to infinity, under the assumption that the source is stationary, ergodic, and has exponential rates for entropy. An interesting question is whether the same result holds without the last assumption. Also, it is of practical interest to obtain non-asymptotic almost sure statements on this same problem.

# 5 Acknowledgements

The author wishes to thank Vittorio Castelli for many interesting discussions.

# 6 Appendix

**Lemma 6** *There exists a constant $K$ such that for all $a > 0$, $\log_2 a \leq K\sqrt{a}$*

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

**Theorem 4** *There exists a set $B$ with $\mu(B) = 1$ such that for all $\mathbf{x} \in B$*

$$h \leq \liminf_{n_w \to \infty} \liminf_{n \to \infty} b_n(\mathbf{x}, n_w)$$

**Proof.** The proof is an adaptation of the argument of Lemma II.1.3 of Shields [4], which in turn is attributed to Barron [8]. Some definitions are needed before we enter the proof of the Theorem. A set $\Upsilon$ of elements of $\{0,1\}^*$ is called a *code*; the elements of $\Upsilon$ are called codewords. A code $\Upsilon$ is said to satisfy the *prefix* property if no codeword in $\Upsilon$ is a prefix of another codeword in $\Upsilon$. Let $l(c) : \Upsilon \to Z$ be the length (in bits) of codeword $c$. It is well known that the lengths of the codewords of a prefix code must satisfy the Kraft inequality:

$$\sum_{c \in \Upsilon} 2^{-l(c)} \leq 1$$

Let $\Phi_n(\mathbf{x}_{-n_w}^{n-1}, n_w)$ denote the binary codeword that the LZ $n_w$-memory algorithm assigns to the first $n$ symbols of $\mathbf{x}$ (and of course to the initial $n_w$-long window) and recall that $(n + n_w)b_n(\mathbf{x}_{-n_w}^{n-1}, n_w)$ is the corresponding code length. Define

$$\Upsilon_{n,n_w} = \{\Phi_n(\mathbf{x}_{-n_w}^{n-1}, n_w) : \mathbf{x}_{-n_w}^{n-1} \in \mathcal{A}^{n_w+n}\}$$

For all $n$ and $n_w$, $\Upsilon_{n,n_w}$ is a prefix code. To show this, suppose otherwise. Then there exist $\mathbf{v}_{-n_w}^{n-1} \in \mathcal{A}^{n+n_w}$ and $\mathbf{w}_{-n_w}^{n-1} \in \mathcal{A}^{n+n_w}$ such that $\mathbf{w}_{-n_w}^{n-1} \neq \mathbf{v}_{-n_w}^{n-1}$ and $\Phi_n(\mathbf{v}_{-n_w}^{n-1}, n_w)$ is a prefix of $\Phi_n(\mathbf{w}_{-n_w}^{n-1}, n_w)$. Let

$$\begin{aligned}
\Phi_n(\mathbf{v}_{-n_w}^{n-1}, n_w) &= \Gamma\Theta_1\Theta_2\cdots\Theta_{c_n} \\
\Phi_n(\mathbf{w}_{-n_w}^{n-1}, n_w) &= \Gamma\Theta_1\Theta_2\cdots\Theta_{c_n}\hat{\Theta}_{c_n+1}\hat{\Theta}_{c_n+2}\cdots\hat{\Theta}_{\hat{c}_n}
\end{aligned}$$

where we have used the notation introduced in the Preliminaries for $\Phi_n(\mathbf{v}_{-n_w}^{n-1}, n_w)$ and $\Phi_n(\mathbf{w}_{-n_w}^{n-1}, n_w)$, and potentially, the concatenated packets $\hat{\Theta}_{c_n+1}\hat{\Theta}_{c_n+2}\cdots\hat{\Theta}_{\hat{c}_n}$ are equal to the empty string. The contradiction comes from the fact that by applying the decoder $f_{c_n}$ (this decoder was introduced the Preliminaries) to all of $\Phi_n(\mathbf{v}_{-n_w}^{n-1}, n_w)$ and the corresponding left portion of $\Phi_n(\mathbf{w}_{-n_w}^{n-1}, n_w)$, we deduce that

$$\mathbf{v}_{-n_w}^{n-1} = \mathbf{w}_{-n_w}^{n-1}\mathbf{z}$$

for some (possibly empty) string $\mathbf{z} \in \mathcal{A}^*$, contradicting the assumption that $\mathbf{v}_{-n_w}^{n-1} \neq \mathbf{w}_{-n_w}^{n-1}$.

Next define

$$\begin{aligned}
V_{n,n_w} &= \left\{\mathbf{x}_{-n_w}^{n-1} : \log_2 \mu_{n+n_w}(\mathbf{x}_{-n_w}^{n-1}) + (n + n_w)b_n(\mathbf{x}_{-n_w}^{n-1}, n_w) < \log_2 \frac{1}{n^2}\right\} \\
W_{n,n_w} &= \left\{\mathbf{x} : \mathbf{x}_{-n_w}^{n-1} \in V_{n,n_w}\right\}
\end{aligned}$$

and note that

$$\begin{aligned}\mu(W_{n,n_w}) &= \sum_{\mathbf{x}_{-n_w}^{n-1} \in V_{n,n_w}} \mu_{n+n_w}(\mathbf{x}_{-n_w}^{n-1}) \\ &\leq \frac{1}{n^2} \sum_{\mathbf{x}_{-n_w}^{n-1} \in V_{n,n_w}} 2^{-(n+n_w)b_n(\mathbf{x}_{-n_w}^{n-1}, n_w)} \\ &\leq \frac{1}{n^2}\end{aligned}$$

where the last inequality follows from the prefix property of the code. Upon defining

$$B_{n_w}^1 = \bigcup_{k=1}^{\infty} \bigcap_{n \geq k} W_{n,n_w}^c$$

we obtain by the Borel-Cantelli Lemma,

$$\mu\left(B_{n_w}^1\right) = 1$$

and hence the measure of the set

$$B^1 \stackrel{\Delta}{=} \bigcap_{n_w \geq 1} B_{n_w}^1$$

is equal to one. By the entropy Theorem [4], there exists a set $B^2$ with $\mu(B^2) = 1$ such that for all $\mathbf{x} \in B^2$,

$$\begin{aligned}h &= \lim_{n\to\infty} -n^{-1} \log_2 \mu_n(\mathbf{x}_0^{n-1}) \\ &= \lim_{n\to\infty} -(n+n_w)^{-1} \log_2 \mu_n(\mathbf{x}_0^{n-1}) \\ &\leq \liminf_{n\to\infty} -(n+n_w)^{-1} \log_2 \mu_{n+n_w}(\mathbf{x}_{-n_w}^{n-1}) \qquad (24)\end{aligned}$$

where the last step follows from

$$\begin{aligned}\mu_{n+n_w}(\mathbf{x}_{-n_w}^{n-1}) &= \mu(\{\mathbf{y} : y_i = x_i, -n_w \leq i < n\}) \\ &\leq \mu(\{\mathbf{y} : y_i = x_i, 0 \leq i < n\}) \\ &= \mu_n(\mathbf{x}_0^{n-1})\end{aligned}$$

Define $B = B^1 \bigcap B^2$. Let $\mathbf{x} \in B$. Then for all $n_w$, there exists a $k$ such that for all $n \geq k$,

$$b_n(\mathbf{x}_{-n_w}^{n-1}, n_w) \geq (n+n_w)^{-1} \log_2 \frac{1}{n^2} - (n+n_w)^{-1} \log_2 \mu_{n+n_w}(\mathbf{x}_{-n_w}^{n-1})$$

Taking the limit as $n \to \infty$, we obtain that for all $n_w$,

$$\begin{aligned}\liminf_{n\to\infty} b_n(\mathbf{x}, n_w) &\geq \liminf_{n\to\infty} -(n+n_w)^{-1} \log_2 \mu_{n+n_w}(\mathbf{x}_{-n_w}^{n-1}) \\ &\geq h\end{aligned}$$

where the last inequality follows from (24). Taking the limit as $n_w \to \infty$, we obtain that for all $\mathbf{x} \in B$,

$$\liminf_{n_w\to\infty} \liminf_{n\to\infty} b_n(\mathbf{x}, n_w) \geq h$$

which is the statement of the Theorem.

## References


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

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

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

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

[5] Frans M.J. Willems. Universal Data Compression and Repetition Times. *IEEE Transactions on Information Theory*, 35:54–58, 1989.

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

[7] J. Ziv A. Wyner. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*,