ON THE AVERAGE REDUNDANCY RATE OF THE LEMPEL-ZIV CODE*

May 22, 1996

Guy Louchard
Departement d'Informatique Théorique
Université Libre de Bruxelles
B-1050 Brussels
Belgium
louchard@ulb.ac.be

Wojciech Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

In this paper, we settle a long standing open problem concerning the average redundancy, r_n, of the Lempel-Ziv'78 (LZ78) code. We prove that for a memoryless source the average redundancy rate attains asymptotically $Er_n = (A + \delta(n))/\log n + O(\log\log n/\log^2 n)$, where A is an explicitly given constant that depends on source characteristics, and $\delta(x)$ is a fluctuating function with a small amplitude. We also derive the leading term for the kth moment of the number of phrases. We conclude by conjecturing a precise formula on the expected redundancy for a Markovian source. The main result of this paper is a consequence of recently obtained second-order properties of the Lempel-Ziv algorithm [P. Jacquet, W. Szpankowski, *Theoretical Computer Science*, 144, 1995]. These findings have been established by analytical techniques of the precise analysis of algorithms. We give a brief survey of these results since they are interesting in their own right, and shed some light on the probabilistic behavior of pattern matching based data compression.

Index Terms: Average redundancy rate, Lempel-Ziv parsing scheme, data compression, generalized Lempel-Ziv scheme, digital search trees, pattern matching, analytical techniques of the precise analysis of algorithms.

---

*Preliminary version of this paper was presented at *Data Compression Conference*, Snowbird 1996.

†This research was partially supported by NSF Grants NCR-9206315, NCR-9415491 and CCR-9201078, and NATO Collaborative Grant CRG.950060.

1. INTRODUCTION

The *redundancy* of a noiseless code measures how far the code is from being optimal for a given source of information. While asymptotically optimal codes require that the redundancy tends to zero (with the length of the code), sometimes a stronger requirement is necessary, namely: that the (average) redundancy per symbol goes to zero at some universal rate. For example, while there are several asymptotically optimal data compression codes (e.g., several versions of the Lempel-Ziv scheme [19]), one can further optimize the rate of convergence to the optimal compression ratio. This is of prime importance for some practical on-line and off-line data compression schemes.

It is known that some prefix codes exist for which the expected redundancy per symbol is $O(\log n/n)$ for a class of sources (e.g., Markov, finite-state sources, etc.). But, recently Shields [13] proved that such a redundancy rate cannot be achieved for general sources. Furthermore, it must be observed that for practical universal data compression codes the above redundancy rate is often not achievable.

In this paper, we investigate the redundancy rate of the Lempel-Ziv parsing scheme [19] – also known as LZ78 algorithm – which was proved to be universal and asymptotically optimal. A modification of this scheme (cf. [16]) is used in the UNIX `compress` command. To recall, the algorithm partitions a sequence of length n into variable phrases such that the next phrase is the shortest phrase not seen in the past. The code consists of pairs of numbers: each pair being a pointer to the previous occurrence of the prefix of the phrase and the last bit of the phrase. Hereafter, we consider only a binary alphabet, but an extension to any finite alphabet is straightforward. Thus, if M_n is the number of phrases constructed from a sequence of length n, then the code length ℓ_n is (cf. [3])

$$\ell_n = M_n(\log M_n + 1) \ . \tag{1}$$

We should point out that various modifications of the LZ78 code lead to slightly different formulæ for the code length. For example, a careful analysis of LZ'78 reveals that a more accurate formula for the code length is

$$\ell_n = \sum_{i=1}^{M_n} \lceil \log(i+1) \rceil \ .$$

But, using the following identity (cf. Knuth [8]) Ex. 1.2.4-42) which holds for integers $a \geq 2$ and $m \geq 1$

$$\sum_{i=1}^{m} \lceil \log_a(i+1) \rceil = (m+1)(\lfloor \log_a(m+1) \rfloor + 1) + \frac{(m+1)a^{1-\{\log_a(m+1)\}} - a}{a-1} \ ,$$

where $\{\log m\} = \log m - \lfloor \log m \rfloor$ (dense in $[0,1]$), we conclude that $\sum_{i=1}^{M_n} \lceil \log(i+1) \rceil = M_n \log M_n + O(M_n)$. Furthermore, the exact count of bits for a particular implementation of the LZ78 code may contribute $O(M_n)$ additional bits. The point to observe is that once we know how to analyze M_n and $M_n \log M_n$ we can obtain a precise count of the numbers of bits contributing to a particular code length. Now, it should be clear that from the asymptotic point of view, equation (1) is accurate enough, and is the most interesting. Therefore, throughout the paper we only consider the code length given by (1).

The pointwise redundancy r_n and its expected value $\overline{r}_n$ of the Lempel-Ziv code are respectively

$$r_n = \frac{M_n(\log M_n + 1) - nh}{n}, \tag{2}$$

$$\overline{r}_n := E r_n = \frac{E\{M_n(\log M_n + 1)\} - nh}{n}, \tag{3}$$

where h is the entropy rate of the source. Plotnik, Weinberger and Ziv [11] proved that the expected redundancy of the Lempel-Ziv code is $\overline{r}_n = O(\log\log n / \log n))$ for finite-state sources. But, the authors of [11] also noticed that there exist sequences for which the redundancy rate is bounded from below by $2/\log n$. In this paper, we prove that this lower bound is actually attainable for the expected redundancy rate $\overline{r}_n$, just closing the gap between the upper and the lower bounds. Moreover, we shall provide a precise asymptotic formula for $\overline{r}_n$ given by (3) which will indicate that the coefficient at $1/\log n$ contains a fluctuating function of a small amplitude. We should also point out that for practical implementations of the LZ'78 code one might need to add $O(EM_n)$ bits to the code length which contribute $O(1/\log n)$ to the expected redundancy, but we can precisely assess this additional contribution to $\overline{r}_n$. Finally, during revision of this paper we have learned that for Markovian sources Savari [12] claimed an upper bound $O(1/\log n)$ on the redundancy of the Lempel-Ziv code and its variations.

The paper is organized as follows. In the next section, we present our main results and their consequences. The proof is delayed until Section 3, where we also briefly survey some recent second-order results concerning the Lempel-Ziv'78 algorithm obtained by analytical techniques (cf. [7, 9]; these are instrumental in the proof of our delicate result. Finally, in our concluding remarks, we conjecture a precise formula on the expected redundancy for a Markovian source. We should also point out that the methods used in this paper can be used to optimize some parameters of various modifications and generalizations of the Lempel-Ziv scheme (cf. [10]), as well as to obtain some other second-order properties of the Lempel-Ziv scheme such as the variance of the number of phrases, and so forth.

2. MAIN RESULTS

We now present the main result of this paper concerning the average redundancy rate $\overline{r}_n$ of the Lempel-Ziv'78 code for a memoryless source. It will follow directly from a lemma discussed below whose proof is delayed till the next section.

Theorem. *Consider a memoryless binary source with symbol "0" occurring with probability* $p > 0$ *and symbol "1" appearing with probability* $q = 1 - p$. *The average redundancy of the Lempel-Ziv code is*

$$\overline{r}_n = h\frac{2 - \gamma - \frac{1}{2h}h_2 + \alpha - \delta_0(n)}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right) , \tag{4}$$

where $h = -p\log p - q\log q > 0$ *is the entropy,* $\gamma = 0.577\ldots$ *is the Euler constant,* $h_2 = p\log^2 p + q\log^2 q$, *and*

$$\alpha = -\sum_{k=1}^{\infty}\frac{p^{k+1}\log p + q^{k+1}\log q}{1 - p^{k+1} - q^{k+1}} . \tag{5}$$

The function $\delta_0(x)$ *is a fluctuating function with mean zero and a small amplitude for* $\log p/\log q$ *rational (e.g., the amplitude of* $\delta_0(x)$ *is smaller than* 10^{-6} *for the unbiased case, where* $p = q = 0.5$*), and* $\lim_{x\to\infty}\delta_0(x) = 0$ *otherwise.*

The above result will follow directly from the next lemma which is of interest in its own right. It describes asymptotic behavior of all moments of the number of phrases M_n created by the Lempel-Ziv scheme. To proceed, we need some notation. Define, for large (say, integer) x, a function $\mu(x)$ as follows:

$$\mu(x) = \frac{x}{h}\log x - \frac{A}{h}x + O\left(\frac{\log x}{h}\right) \tag{6}$$

where $A = O(1)$ and will be specified below. Let x_n be a positive solution of the following equation

$$\mu(x_n) = n . \tag{7}$$

Observe that the above equation has the following asymptotic solution for large n

$$x_n = \frac{nh}{\log n}\left(1 + \frac{\log\log n}{\log n} + \frac{A - \log h}{\log n} + O\left(\frac{(\log\log n)^2}{\log^2 n}\right)\right) . \tag{8}$$

We adopt the same assumptions as in Theorem. Then, our finding can be formulated as follows:

Lemma. *Let* M_n *be the number of phrases obtained after parsing a sequence of length* n *according to the Lempel-Ziv algorithm. Then, for any* $k \geq 1$

$$EM_n^k = x_n^k\left(1 + O\left(\sqrt{\frac{\log n}{n}}\right)\right) + O\left(\frac{n^{k-1}}{\log^{k-1} n}\right) , \tag{9}$$

and

$$E\{M_n(\log M_n + 1)\} = EM_n(\log EM_n + 1) + O(1/\log n) \tag{10}$$

for large n.

Our main result (4) of Theorem is an easy conclusion from (9) and (10). Indeed, since $EM_n = x_n(1 + O(\sqrt{\log n/n}))$, then by (10) we obtain

$$\begin{aligned}
\overline{r}_n &= \frac{EM_n(\log EM_n + 1) - nh}{n} + O\left(\frac{1}{n\log n}\right) = \frac{x_n(\log x_n + 1) - nh}{n} + O\left(\sqrt{\frac{\log n}{n}}\right) \\
&= h\frac{\log\log n}{\log n} - h\frac{\log\log n}{\log n} + h\frac{1+A}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right)
\end{aligned}$$

where A is the constant in (6), and its explicit formula is given in (19) below (cf. Fact A in Section 3). This proves (4) since the terms $\log\log n/\log n$ cancel out!

The above redundancy result should be compared to the average redundancy of another version of the Lempel-Ziv scheme [18] known as Lempel-Ziv'77 (in short: LZ77). In this scheme, one searches for the longest substring that is yet uncompressed which occurs in a "database sequence". Then, the code consists of a pointer to the database sequence and the length of the repeated substring. From recent results of Jacquet and Szpankowski [6] and Wyner and Wyner [17] we conclude that the average redundancy $\overline{r}_n$ for memoryless and Markovian sources of the LZ'77 code is

$$\overline{r}_n = h\frac{\log\log n}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right) .$$

It was conjectured that the performance of LZ77 and LZ78 are comparable to one another since the limitation on the number of phrases for LZ78 is balanced by the need to send the length function for LZ77. Our main finding answers this question in the negative, that is, LZ77 cost to send the length overhelms its gain.

3. ANALYSIS BY ANALYTICAL TECHNIQUES

In this section, we prove the Lemma. Actually, in the first part of this section (cf. Section 3.1) we survey some recent probabilistic results on second-order properties of the LZ'78 scheme established recently by analytical techniques (cf. Louchard and Szpankowski [9], and Jacquet and Szpankowski [7]). The results of [7] are instrumental in proving our Lemma (cf. Section 3.2). We should mention that these findings are interesting in their own right, and some of these results are even stronger than the ones presented in this paper. To the best of our knowledge, these are one of the first set of results on Lempel-Ziv'78 scheme – with the exception of [1] – that capture second-order probabilistic properties of the algorithm (e.g., variance, limiting distribution, large deviations, etc). For results that are true for all source sequences see for example [11].

3.1 Survey of Second-Order Properties for LZ78 Scheme

As introduced in Louchard and Szpankowski [9], it is convenient to consider two models for the LZ78 parsing scheme. The first one is called the *digital tree model* (or Gilbert-Kadota model) and assumes that the number of phrases is given, say, there are m statistically independent phrases, and one asks for the Lempel-Ziv sequence that can be built from these m phrases. Such a construction can be done efficiently using a data structure called the *digital search tree*, as described in [7, 9]. In another model, called the *Lempel-Ziv model*, the length of a sequence is given, and we assume it is equal to n. Hereafter, we shall consistently use n as the length of a single string to be parsed, and m as the number of words (phrases) used to construct the Lempel-Ziv sequence.

There are some parameters of interest in both models that we discuss next. Let $D_m(i)$ be the length of the ith phrase in the digital tree model. By D_m we denote the length of a randomly selected phrase, that is,

$$\Pr\{D_m = k\} = \frac{1}{m}\sum_{i=1}^{m}\Pr\{D_m(i) = k\} \ .$$

The length of the sequence built in the digital tree model is denoted by L_m, and it is clearly equal to $L_m = \sum_{i=1}^{m} D_m(i)$. In the Lempel-Ziv model, we write D_n^{LZ} for the length of a randomly selected phrase, and M_n for the number of phrases constructed from a string of length n.

There are relationships between the above parameters in both models. The most important relationship is the one between L_m and M_n, constituting the so called *renewal equation* studied by Billingsley [2]. It can be phrased as follows:

$$M_n = \max\{m : \ L_m = \sum_{k=1}^{m} D_m(i) \le n\} \ , \tag{11}$$

which directly implies that

$$\Pr\{M_n > m\} = \Pr\{L_m \le n\} \ . \tag{12}$$

The relationship (11) follows directly from the above definitions, and the interested reader may find more on it in [7, 9].

We next observe that the above relationship should lead to a complete characterization of M_n if one can analyze L_m. This follows from a result of Billingsley (cf. Theorem 17.3 in

[2]) which can be re-phrased for our needs as follows: If

$$\frac{L_m - \mu_m}{\sigma_m} \to N(0,1) \ , \tag{13}$$

then

$$\frac{M_n - n/(\mu_m/m)}{\sqrt{n(\sigma_m/m)/(\mu_m/m)^3}} \to N(0,1) \tag{14}$$

where $N(0,1)$ is the standard normal distribution, and μ_m and σ_m are positive constants that, under mild standard uniform integrability arguments, can be asymptotically interpreted as the mean and the variance of L_m.

Recently, Jacquet and Szpankowski [7] proved that for a memoryless source L_m is normally distributed, and so by Billingsley's result M_n is also normally distributed. But, to derive second order properties of the Lempel-Ziv algorithm (such as redundancy), we need the rate of convergence to a normal distribution. The authors of [7] obtained such a rate for L_m, but not for M_n (since (14) does not provide it). Therefore, we shall deal mostly with the digital tree model instead of the Lempel-Ziv model.

As mentioned above, the results of [7] and [9] establish some second-order properties of LZ78, and they are of interest on their own. We survey these below, in the form of two facts called Fact A (digital tree model) and Fact B (Lempel-Ziv model). We shall write

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{15}$$

for the distribution function of the standard normal distribution.

Now, we are ready to recall main results of Jacquet and Szpankowski [7] and Louchard and Szpankowski [9] for the digital tree model.

Fact A. (DIGITAL TREE MODEL) *Consider a digital tree model created by m independent words generated by a biased memoryless source.*

(i) *The phrase length D_m satisfies*

$$\begin{aligned} ED_m &= \frac{1}{h}\log m + \frac{1}{h}\left(\frac{h_2}{2h} + \gamma - 1 - H_{b-1} - \alpha + \delta_0(m)\right) + O\left(\frac{\log m}{m}\right) && (16) \\ \mathrm{Var} D_m &= c_2 \log m + O(1) \ , && (17) \end{aligned}$$

where $c_2 = (h_2 - h^2)/h^3$, $H_b = \sum_{i=1}^{b} \frac{1}{i}$ is the Harmonic number, and other quantities are defined in Theorem.
Furthermore, for large m

$$\Pr\{D_m \le ED_m + x\sqrt{\mathrm{Var} D_m}\} = \Phi(x)\left(1 + O(1/\sqrt{\log m})\right) \tag{18}$$

where $x = O(1)$.

(ii) *The length of the Lempel-Ziv sequence built from* m *independent phrases satisfies:*

$$\begin{aligned} EL_m &= \frac{m}{h}\left(\log m + \frac{h_2}{2h} + \gamma - 1 - \alpha + \delta_0(\log m)\right) \\ &+ \frac{1}{h}\left(\log m + \frac{h_2}{2h} - \gamma - \log p - \log q + h\alpha\right) + O(1) \qquad (19) \\ \text{Var } L_m &= c_2 m \log m + O(m) \ . \qquad (20) \end{aligned}$$

Furthermore, for large m *the following weak convergence takes place*

$$\Pr\{L_m \le EL_m + x\sqrt{\text{Var}L_m}\} = \Phi(x)\left(1 + O(1/\sqrt{m})\right) \qquad (21)$$

for $x = O(1)$. *Finally, there exist positive constants* A *and* $\mu < 1$ *such that uniformly in* k

$$\Pr\left\{\left|\frac{L_m - EL_m}{\sqrt{\text{Var } L_m}}\right| > k\right\} < A\mu^k \ . \qquad (22)$$

(iii) *The above is also true for an unbiased memoryless source (i.e., symbols occur with the same probability) if one replace the variance above by the following*

$$\text{Var } L_m^{sym} \sim m \cdot (C + \delta(\log_2 m)) \qquad (23)$$

where $C = 0.26600\ldots$ *and* $\delta(x)$ *is a fluctuating function with small amplitude.*

We next formulate recent findings of [7, 9] for the Lempel-Ziv model. Those results related to M_n are direct consequences of Fact A and Billingsley's theorem due to our basic relationships (11)-(12). For the phrase length distribution of D_n^{LZ} we must resort to Louchard and Szpankowski [9] who proved that

$$\Pr\{D_n^{LZ} = k | M_n = m\} = \left(1 + O(\sqrt{\log n / n})\right) \Pr\{D_m = k\} \ .$$

Then, we are able to report the following results of [9, 7]:

Fact B. (Lempel-Ziv Model) *Consider the Lempel-Ziv model in which a sequence of length* n *is generated by a biased memoryless source.*

(i) *The phrase length* D_n^{LZ} *satisfies:*

$$\Pr\{D_m \le \frac{1}{h}\log(nh/\log n) + x\sqrt{c_2 \log(nh/\log n)}\} = \Phi(x)\left(1 + O(1/\sqrt{\log m})\right) \qquad (24)$$

for $x = O(1)$, *and there exist two positive constants* A *and* $\alpha < 1$ *such that*

$$\Pr\left\{\left|\frac{D_m^{LZ} - c_1\log(nh/\log n)}{\sqrt{c_2 \log(nh/\log n)}}\right| > k\right\} \le A\alpha^k \qquad (25)$$

uniformly in k for large m.

(ii) *The sequence of random variables*

$$Z_n := \frac{M_n - EM_n}{\sqrt{\mathrm{Var} M_n}} \to N(0,1) \tag{26}$$

In addition, for all $r \geq 0$ the sequence $(Z_n)^r$ is uniformly integrable. Thus all moments of Z_n exist and converge to the appropriate moments of the normal distribution. In particular,

$$EM_n \sim \frac{nh}{\log(n)} \tag{27}$$

$$\mathrm{Var}\ M_n \sim \frac{c_2 h^3 n}{\log^2 n}\ . \tag{28}$$

Finally, for any $\varepsilon > 0$, there exist an integer $n_0 \geq 1$ such that for all $n > n_0$

$$\Pr\left\{|M_n - EM_n| > \varepsilon EM_n\right\} \leq A \exp\left(-a\varepsilon\sqrt{n}\right) \tag{29}$$

for some positive constants $A, a > 0$.

(iii) *The above results are also true for the symmetric (i.e., unbiased) memoryless source if one replaces the variance by*

$$\mathrm{Var}\ M_n^{sym} \sim \frac{n(C + \delta(\log_2 n))}{\log_2^3 n} \tag{30}$$

where the constant $C = 0.26600\ldots$. In (29) one must replace $\sqrt{n}$ by $\sqrt{n/\log n}$.

In passing, we should mention that our large deviations results (e.g., see (25) and (29)) are probably not the best possible. Actually, based on results of [7] and Gärtner-Ellis large deviations theorem, we conjecture that the following is true:

$$\lim_{m\to\infty} \frac{1}{m \log m} \log \Pr\{|EL_m - EL_m| > yEL_m\} = -I(y)$$

where $I(y) > 0$ is a function of $y > 0$, and

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\left\{M_n > \mu^{-1}\left(\frac{n}{1-y}\right)\right\} = -\frac{I(y)}{1-y}$$

for $0 < y < 1$, and $\mu^{-1}(x)$ is the inverse function to $\mu(m) = EL_m$ continued to real x.

3.2 Proof of Lemma

Now, we are in position to prove our Lemma from Section 2. From (12) we easily derive the kth moment of M_n. Indeed, a standard computation reveals that

$$EM_n^{k+1} = \sum_{m=0}^{\infty} m^{k+1} \Pr\{M_n = m\} = \sum_{m=0}^{\infty} m^{k+1} (\Pr\{M_n > m-1\} - \Pr\{M_n > m\})$$

$$
\begin{aligned}
&= \sum_{m=0}^{\infty}((m+1)^{k+1}-m^{k+1})\Pr\{M_n>m\} \\
&= (k+1)\sum_{m\geq 0} m^k \Pr\{M_n>m\} + O(EM_n^k) \ ,
\end{aligned}
$$

where the last equality follows from the identity: $a^k-b^k=(a-b)(a^{k-1}+a^{k-2}b+\cdots+b^{k-1})$. Then, the sum above becomes

$$
\begin{aligned}
\sum_{m\geq 0} m^k \Pr\{M_n>m\} &= \sum_{m\geq 0} m^k \Pr\{L_m \leq n\} \\
&= \int_0^{\infty} x^k \Pr\{L_x\leq n\}dx + O(EM_n^k)
\end{aligned}
$$

where, in the last line, we replace the sum by the integral, and use the Euler-Maclaurin formula [8] to estimate the error. To verify, it suffices to do some elementary algebra on the Euler-Maclaurin formula which is recalled below: *For any function $f(k)$, we have*

$$
\begin{aligned}
\sum_{k=a}^{b} f(k) &= \int_a^b f(x)dx - \frac{f(b)-f(a)}{2} + \sum_{k=1}^{j}\frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(b)-f^{(2k-1)}(a)\right) \\
&+ O\left((2\pi)^{-2j}\right)\int_a^b |f^{(2j)}(x)|dx
\end{aligned}
$$

where $f^{(j)}(x)$ is the jth derivative of $f(x)$, and B_k is the kth Bernoulli number. In the above, $f(x)=x^k\Pr\{M_n>x\}$, and the error term for $j=1$ becomes $\int_0^\infty f'(x)dx = O(EM_n^k)$. In summary, we obtain

$$
EM_n^{k+1} = (k+1)\int_0^{\infty} x^k \Pr\{L_x\leq n\}dx + O(EM_n^k) \tag{31}
$$

Thus by Fact A and the above we arrive at

$$
\begin{aligned}
EM_n^{k+1} &= (k+1)\int_0^{x_n} x^k(1+O(1/\sqrt{x}))\Phi\left(\frac{n-\mu(x)}{\sigma(x)}\right)dx \\
&+ (k+1)\int_{x_n}^{\infty} x^k(1+O(1/\sqrt{x}))\Phi\left(\frac{n-\mu(x)}{\sigma(x)}\right)dx + O(EM_n^k)
\end{aligned}
$$

where we use simplified notation $\mu(x)=EL_x$ and $\sigma(x)=\mathrm{Var}\ L_x$. The quantity x_n above is the same as in (7), that is, $n=\mu(x_n)$. Note that x_n is given asymptotically by (8).

In order to estimate the above integrals, we first recall definition of the error function $\mathrm{erf}(x)$:

$$
\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \qquad x\geq 0 \ .
$$

Observing that $\Phi(x) = 1/2 + 1/2\text{erf}(x/\sqrt{2})$ and $\text{erf}(-x) = -\text{erf}(x)$, we estimate the first integral above as (with the error term $1 + O(1/\sqrt{x})$ dropped for the simplicity of presentation)

$$\int_0^{x_n} x^k \Phi\left(\frac{n-\mu(x)}{\sigma(x)}\right) dx = \frac{x_n^{k+1}}{k+1} - \frac{1}{2}\int_0^{x_n} x^k \left(1 - \text{erf}\left(\frac{n-\mu(x)}{\sqrt{2}\sigma(x)}\right)\right) dx \ .$$

In a similar fashion, the second integral becomes

$$\int_{x_n}^{\infty} x^k \Phi\left(\frac{n-\mu(x)}{\sigma(x)}\right) dx = \frac{1}{2}\int_{x_n}^{\infty} x^k \left(1 - \text{erf}\left(\frac{\mu(x)-n}{\sqrt{2}\sigma(x)}\right)\right) dx \ .$$

In summary, we have

$$\begin{aligned} EM_n^{k+1} &= x_n^{k+1}(1+O(1/\sqrt{x_n})) - \frac{k+1}{2}\int_0^{x_n} x^k(1+O(1/\sqrt{x}))\left(1 - \text{erf}\left(\frac{n-\mu(x)}{\sqrt{2}\sigma(x)}\right)\right) dx \\ &+ \frac{k+1}{2}\int_{x_n}^{\infty} x^k(1+O(1/\sqrt{x}))\left(1 - \text{erf}\left(\frac{\mu(x)-n}{\sqrt{2}\sigma(x)}\right)\right) dx + O(x_n^k) \ . \end{aligned} \tag{32}$$

We compute the above two integrals separately. Let us denote them as follows

$$\begin{aligned} I_1 &= \int_0^{x_n} x^k \left(1 - \text{erf}\left(\frac{n-\mu(x)}{\sqrt{2}\sigma(x)}\right)\right) dx \ , \\ I_2 &= \int_{x_n}^{\infty} x^k \left(1 - \text{erf}\left(\frac{\mu(x)-n}{\sqrt{2}\sigma(x)}\right)\right) dx \ . \end{aligned}$$

We make the following change of variables

$$y = \frac{\mu(x)-n}{\sigma(x)}$$

and we simplify $\mu(x) \sim c_1 x \log x$ and $\sigma(x) \sim \sqrt{c_2 x \log x}$ where we write $c_1 = 1/h$. Observe that x is a function of y, so we often write $x(y)$. It satisfies the following equation $c_1 x \log x - y\sqrt{c_2 x \log x} - n = 0$, which implies

$$\begin{aligned} x(y)\log x(y) &= \left(\frac{y\sqrt{c_2} + \sqrt{c_2 y^2 + 4nc_1}}{2c_1}\right)^2 \\ &= nh\left(1 + \frac{c_2 h y^2}{2n} + \frac{y\sqrt{hc_2}}{\sqrt{n}}\sqrt{1 + \frac{hc_2 y^2}{4n}}\right) \\ &= nh\left(1 + \frac{y\sqrt{hc_2}}{\sqrt{n}} + \frac{y^2 c_2 h}{2n} + \frac{y^3 c_2 h\sqrt{hc_2}}{8n^{3/2}} + O(y^5/n^{5/2})\right) \ . \end{aligned}$$

From the above and after some further Taylor's expansion of $\log x(y)$ we derive the following (with the help of MAPLE)

$$\frac{dx(y)}{dy} = \sqrt{n}(1+G(n))\left(\frac{h^{3/2}\sqrt{c_2}}{\log n} + \frac{yh^2 c_2}{\sqrt{n}\log n} + O\left(\frac{1}{n}\right)\right)$$

and

$$x(y) = n(1+G(n))\left(\frac{h}{\log n} + \frac{yh^{3/2}\sqrt{c_2}}{\sqrt{n}\log n} + \frac{y^2h^2c_2}{2n\log n} + O\left(\frac{y^3}{n^{3/2}}\right)\right)$$

where

$$G(n) = \frac{\log\log n}{\log n} + \cdots .$$

The above substitutions are needed to compute the above integrals which become

$$\begin{aligned} I_1 &= \int_0^\infty x^k(-y)(1-\mathrm{erf}(y/\sqrt{2}))\frac{dx(-y)}{dy}dy ,\\ I_2 &= \int_0^\infty x^k(y)(1-\mathrm{erf}(y/\sqrt{2}))\frac{dx(y)}{dy}dy . \end{aligned}$$

Let us first estimate I_2 for $k=0$. We obtain

$$\begin{aligned} I_2 &= \int_0^\infty (1-\mathrm{erf}(y/\sqrt{2}))\frac{dx(y)}{dy}dy \\ &= (1+G(n))\int_0^\infty (1-\mathrm{erf}(y/\sqrt{2}))\left(\sqrt{\frac{n(h_2-h^2)}{\log^2(n)}} + \frac{yc_2h^2}{\log(n)} + O\left(\frac{y^2}{\sqrt{n}}\right)\right)dy \\ &= (1+G(n))\left(\sqrt{\frac{n(h_2-h^2)}{\pi\log^2(n)}} + \frac{c_2h^2}{2\log(n)} + O\left(\frac{1}{\sqrt{n}}\right)\right) . \end{aligned}$$

To compute the above integral we used the following well known identity (cf. [5] Eq. (6.281))

$$\int_0^\infty (1-\mathrm{erf}(px))x^{2q-1} = \frac{\Gamma(q+1/2)}{2\sqrt{\pi}qp^{2q}}$$

for $p>0$ and $q>0$. We also need $\Gamma(3/2)=\sqrt{\pi}/2$ (cf. [5]). In a similar manner we can compute I_1. Indeed, we have

$$I_1 = (1+G(n))\left(\sqrt{\frac{n(h_2-h^2)}{\pi\log^2(n)}} - \frac{c_2h^2}{2\log(n)} + O\left(\frac{1}{\sqrt{n}}\right)\right) .$$

Putting everything together we get

$$EM_n = x_n\left(1+O\left(\sqrt{\frac{\log n}{n}}\right)\right) + O(1)$$

This proves Lemma for $k=0$, and following the same lines of arguments we prove it for any k.

To prove our result on the average redundancy, we must evaluate $E\{M_n(\log M_n+1)\}$, that is, we now establish (10). Let N be a random variable distributed as the standard normal distribution. Then, by Fact B we can write $M_n = EM_n + N\sigma_n + X$ where $X=o(\sqrt{n})$ (pr.) (in fact, (29) suggests that $X=O(1)$ but we use only the above weaker condition to derive our results). We write: $\sigma_n = \sqrt{\mathrm{Var} M_n} \sim \sqrt{c_2 h^3 n / \log^2 n}$. Using (27)-(28), one easily proves the following

$$\begin{aligned} M_n \log M_n &= (EM_n + N\sigma_n + X) \log EM_n \left(1 + \frac{N\sqrt{c_2 h}}{\sqrt{n}} + \frac{X}{EM_n}\right) \\ &= (EM_n + N\sigma_n + X) \left(\log EM_n + \frac{N\sqrt{c_2 h}}{\sqrt{n}} + \frac{X}{EM_n} - \frac{N^2 c_2 h}{2n}\right) \end{aligned}$$

After some algebra and noting that $EN = EX = EN^3 = 0$ we obtain

$$E\{M_n(\log M_n + 1)\} = EM_n(\log EM_n + 1) + O(1/\log n) ,$$

which proves (10), and completes our derivation of the Lemma, and the proof of Theorem.

4. CONCLUDING REMARKS

In this final section, we consider an extension of our main result to Markovian sources. We provide a precise formula for the average redundancy $\overline{r}_n$, however, a rigorous proof of it requires an extension Fact A (cf. (21)) which is under our current investigations, and will be reported in a forthcoming paper.

As mentioned above, in order to extend our main results to Markovian sources we need to generalize Fact A and Fact B. At this point, we have an extension for the phrase length (part (i) of Fact A) that allows to predict a precise form of the average redundancy $\overline{r}_n$. We start with some notation. In this case, it is better to consider a general alphabet, say $\mathcal{A} = \{1, 2, \ldots, V\}$. We consider a stationary Markov chain over $\mathcal{A}$ such that $P = \{p_{ij}\}_{i,j=1}^{V}$ is its transition matrix, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_V)$ denotes the stationary probabilities. We define two matrices: $P(s) = \{p_{ij}^{-s}\}_{i,j=1}^{V}$ and $Q(s) = I - P(s)$ where s is a complex number. Furthermore: the entropy $h = -\sum_{i=1}^{V}\sum_{i=1}^{V} \pi_i p_{ij} \log p_{ij}$, and $h_\pi = -\sum_{i=1}^{V} \pi_i \log \pi_i$. Let also $\beta = [\det Q(s)]''|_{s=-1}$, and

$$\dot{\mathbf{X}}(-2) = \sum_{i=1}^{\infty} [Q^{-1}(-2) \ldots Q^{-1}(-i)(Q^{-1}(s))'|_{s=-i-1} Q^{-1}(-i-2) \ldots] K$$

where $K = \{\Pi_{i=0}^{\infty} Q^{-1}(-2-i)\}^{-1}\Psi$, and $\Psi = [1, 1, ..., 1]_{V\times 1}^{T}$. Finally, let ω be the determinant of a matrix formed from $Q(-1)$ by replacing its first column by the vector Ψ.

The following result was recently proved in [15]: Under the above assumptions, in the digital tree model the length of the generalized Lempel-Ziv sequence built from m phrases becomes

$$EL_m = \frac{m}{h}\left(\log m + \gamma - 1 + h - h_\pi - \frac{\beta}{2\omega h} - \boldsymbol{\pi}\dot{\mathbf{X}}(-2)) + \delta_1(\log m)\right) + O(\log m)$$

with notations defined as above. Here $\delta_1(x)$ is a fluctuating function with a small amplitude when $(\log p_{ij} + \log p_{1i} - \log p_{1j})/\log p_{11}$ is rational for any $i, j = 1, 2, ..., V$, and $\lim_{x\to\infty} \delta_1(x) = 0$ otherwise.

The above will yield the following formula on the average redundancy

$$\overline{r}_n = h\frac{2 - \gamma + h_\pi - h + \frac{\beta}{2\omega h} + \boldsymbol{\pi}\dot{\mathbf{X}}(-2)) - \delta_1(n)}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right) ,$$

provided we can prove the limiting distribution for L_m and M_n. That is, we must extend Fact A(ii) and Fact B(ii) to the Markovian case. Such an extension was outlined in Jacquet and Szpankowski [7], but details are still under development, and will be reported in a forthcoming paper.

ACKNOWLEDGEMENT

We would like to thank M. Feder, P. Jacquet, J. Kieffer and P. Shields for helpful comments and discussions regarding this work. We also thank the Associate Editor, M. Feder and anonymous referees for helping us to shape the final version of this paper.

References

[1] D. Aldous and P. Shields, A Diffusion Limit for a Class of Random-Growing Binary Trees, *Probab. Th. Rel. Fields*, 79, 509-542 (1988).

[2] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York 1968.

[3] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley&Sons, New York (1991).

[4] E. Gilbert and T. Kadota, The Lempel-Ziv Algorithm and Message Complexity, *IEEE Trans. Information Theory*, 38, 1839-1842, 1992.

[5] I. Gradshteyn and I. Ryznik, *Tables, Integrals, Series, and Products*, Academic press, New York 1980.

[6] P. Jacquet and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combin.Theory Ser. A*, 66, 237-269, 1994.

[7] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197, 1995.

[8] D. Knuth, *The Art of Computer Programming: Fundamental Algorithms,* Vol. 1. Addison-Wesley, 1973.

[9] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Trans. Information Theory*, 41, 478-488, 1995.

[10] G. Louchard, W. Szpankowski and J. Tang, Average Profile for Generalized Digital Search Trees and Lempel-Ziv Algorithm, Purdue University, CSD-TR-96-005, 1996 (preliminary version *Proc. 33-rd Annual Allerton Conference*, 891-900, 1995).

[11] E. Plotnik, M.J. Weinberger, and J. Ziv, Upper Bounds on the Probability of Sequences Emitted by Finite-State Sources and on the Redundancy of the Lempel-Ziv Algorithm, *IEEE Trans. Information Theory*, 38, 66-72, 1992.

[12] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, reprint.

[13] P. Shields, Universal Redundancy Rates Do Not Exist, *IEEE Information Theory*, 39, 520-524, 1993.

[14] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176-1198, 1993.

[15] J. Tang, *Probabilistic Analysis of Digital Search Trees*, Ph.D. Thesis. Purdue University, 1996.

[16] T. Welch, A Technique for High-Performance Data Compression, *Computer*, 17, 8-19, 1984.

[17] A.D. Wyner and A.J. Wyner, Improved Redundancy of a Version of the Lempel-Ziv Algorithm, *IEEE Trans. Information Theory*, 41, 723-732, 1995.

[18] J. Ziv and A. Lempel, A Universal Algorithm for Sequential Data Compression, *IEEE Trans. Information Theory*, 23, 337-343, 1977.

[19] J. Ziv and A. Lempel, Compression of Individual Sequences via Variable-rate Coding, *IEEE Trans. Information Theory*, 24, 530-536, 1978.