On the Average Redundancy Rate of the Lempel-Ziv Code*

Guy Louchard
Laboratoire d'Informatique Théorique
Université Libre de Bruxelles
B-1050 Brussels
Belgium

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

March 29, 1996

*To appear in *IEEE Trans. Information Theory.*

Outline of the Talk

1. Lempel-Ziv'78 Parsing Scheme and Its Code
2. Redundancy of a Code
3. Main Result
4. Extension (Markov Source, Generalized LZ'78)
5. Sketch of the Proof

Lempel-Ziv'78 Scheme

Lempel-Ziv'78 parsing scheme is used in UNIX `compress` command and ina CCITT standard for data compression for modems.

Example:
Consider a string $ababbbabbaaaba$ which is parsed and coded as follows:

Phrase No:	1	2	3	4	5	6	7
Sequence:	(a)	(b)	(ab)	(bb)	(abb)	(aa)	(aba)
Code:	-a	-b	1b	2b	3b	1a	3a

If M_n denote the number of phrases that a sequence of length n is parsed, then the length ℓ_n of the LZ'78 code is

$$\ell_n = M_n(\log M_n + 1) \ .$$

Code Redundancy

POINTWISE REDUNDANCY r_n is defined as

$$\begin{aligned} r_n &= \frac{\ell_n - nh}{n} \\ &= \frac{M_n(\log M_n + 1) - nh}{n} \end{aligned}$$

where h is the entropy rate of the source.

The AVERAGE redundancy $\overline{r}_n$ of the Lempel-Ziv code is

$$\overline{r}_n := Er_n = \frac{E\{M_n(\log M_n + 1)\} - nh}{n} \ .$$

Previous Results and Conjectures

Plotnik, Weinberger and Ziv
IEEE Trans. Information Theory, 38, 1992
proved that

$$\overline{r}_n = O\left(\frac{\log\log n}{\log n}\right)$$

but they also noticed that

$$\overline{r}_n = \Omega\left(\frac{1}{\log n}\right) \quad .$$

Plotnik, Weinberger and Ziv conjectured that

$$\overline{r}_n = \Theta\left(\frac{\log\log n}{\log n}\right) \quad ,$$

as Lempel-Ziv'77 scheme.

Main Result

THEOREM:
Let us consider a memoryless source generated a binary sequence with "1" occurring with probability p and "0" appearing with probability $q = 1 - p$.
Then, the average redundancy of the Lempel-Ziv code is:

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

where

$$\begin{aligned} h &= -p\log p - q\log q \\ h_2 &= p\log^2 p + q\log^2 q \ , \end{aligned}$$

$\gamma = 0.577\ldots$ is the Euler constant, and $\delta_0(x)$ is a fluctuating functions with small amplitude for $\log p/\log q$ rational, and $\lim_{x\to\infty}\delta_0(x) = 0$ otherwise. Finally, the constant α is defined as:

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

Some Additional Definitions

We also proved some other results that will directly implied our main finding.

Definitions:
Let for large (say, integer) x a function $\mu(x)$ is defined as:

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

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

$$\mu(x_n) = n\ .$$

Observe that for large n

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

Other Results That Implied Main Theorem

THEOREM:
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 $k \geq 1$

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

In addition:

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

The main results on redundancy $\overline{r}_n$ follow since

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

Comment on Lempel-Ziv'77 Redundancy

One can prove that Lempel-Ziv 77 data compression scheme (i.e., finding the longest prefix in a "training sequence") has the following redundancy

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

This follows for *memoryless* source from:

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.

For *Markov* source it was recently proved by:

A.J. Wyner,
The Redundancy and Distribution of the Phrase Lengths of the Fixed-Database Lempel-Ziv Algorithm, preprint.

Extension: Markovian Source

A much more elaborate analysis is need for Markovian sources. However, one can again prove that:

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

but meanings of the constants α, h_2 are different.

For example: we define α in the case of a stationary Markov chain over a V-ary alphabet with transition probabilities $P = \{p_{ij}\}_{i,j=1}^{V}$ and stationary probabilities $\{\pi_i\}_{i=1}^{V}$, as follows:
Let

$$Q(s) = I - P(s) \quad , \qquad P(s) = \{p_{ij}^{-s}\}_{i,j=1}^{V} \ .$$

Then, $\alpha = \boldsymbol{\pi} \cdot \dot{\mathbf{X}}(-2)$ where

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

where $K = \{\Pi_{i=0}^{\infty}Q^{-1}(-2-i)\}^{-1}\Psi$, and

$$\Psi = [1, 1, ..., 1]_{V\times 1}^{T}.$$

Extension: Generalized Lempel-Ziv'78

Louchard and Szpankowski in a paper presented in DCC'95 introduced a generalization of LZ'78 that works as follows: Choose an integer $b \geq 1$. A sequence is partitions in phrases such that the next phrase is the SHORTEST phrase seen in the past by *at most* $b-1$ phrases. The case $b=1$ corresponds to the original Lempel-Ziv algorithm.

Example:
Let $b=2$, and consider the string $ababbbabbaaaba$ which is parsed and coded as follows:

Phrase:	1	2	-	-	3	4	-	5
Sequence:	(a)	(b)	(a)	(b)	(bb)	(ab)	(bb)	(aa)
Code:	a	b	1	2	2b	1b	3	1a

DOES AN OPTIMAL VALUE OF b EXIST?

Generalized Digital Search Tree

HOW TO CONSTRUCT EFFICIENTLY PHRASES?

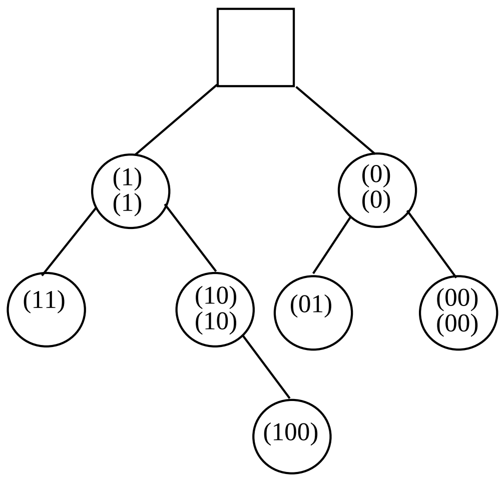

Figure 1: A 2-digital search tree representation of the generalized Lempel-Ziv parsing for the string 11001010001000010011

Performance of the Generalized LZ'78

Let
$M_n(b)$ – number of *all* phrases
$M_n'(b)$ – number of *distinct* phrases.

Length of the code:

$$\ell_n(X_1^n) = M_n(b)(\log M_n'(b) + I(b))$$

where where $I(b)$ is equal to one if the phrase consists of a previous prefix and one more bit (i.e., already $b-1$ phrases have occurred), and zero otherwise.

Redundancy of the Generalized LZ

The following can be proved:

$$\overline{r}_n(b) = h\frac{1-\gamma-\frac{h_2}{2h}-\Delta(b)+H_{b-1}+q_0(b)+\log q_0(b)-\delta(n)}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right) ,$$

$\delta(n)$ – a fluctuating function with a small amplitude,
$q_0(b)$ – the proportional constant in the number of nodes in the associated b-digital tree (i.e., the number of nodes is approximately $mq_0(b)$),
H_b – Harmonic numbers
$\Delta(b)$ – a constant that can be computed recursively (e.g., $\Delta(1,03)=1.25$, $\Delta(5,0.3)=0.83$, $\Delta(50,0.3)=0.36$).

For example, for unbiased memoryless source we have:

$$\overline{r}_n(1) = \frac{2.68+\delta(n)}{\log n}+O\left(\frac{\log\log n}{\log^2 n}\right) ,$$

$$\overline{r}_n(\infty) = \frac{1.87+\delta(n)}{\log n}+O\left(\frac{\log\log n}{\log^2 n}\right) .$$

Sketch of the Proof

The proof is based a result proved recently by
P. Jacquet and W. Szpankowski,
Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees,
Theoretical Computer Science, 144, 161-197, 1995.

We summarize this result as well as some others on LZ'78 that have been recently proved by us (Louchard, Jacquet and Szpankowski) by ANALYTICAL METHOD.

DIGITAL TREE MODEL VERSUS LEMPEL-ZIV MODEL
In the **digital tree model** one asks for the length L_m of a string composed of m phrases.
The **Lempel-Ziv** model assumes that the string is of fixed length, say n. and investigates the number of phrases M_n generated by the scheme.

Some parameters of interest: M_n – number of phrases constructed from a fixed length string
D_m – randomly selected phrase in the digital tree model
D_n^{LZ} – randomly selected phrase in the LZ model.

Second-Order Properties in Digital Tree Model

THEOREM: In **digital tree model** the following holds:

$$
\begin{aligned}
ED_m &= \frac{1}{h}\log m + \frac{1}{h}\left(\frac{h_2}{2h} + \gamma - 1 - H_{b-1} - \Delta(b,p)\right. \\
&+ \left.\delta(m)\right) + O\left(\frac{\log m}{m}\right) \qquad (1) \\
\mathrm{Var}D_m &= c_2 \log m + O(1)\ ,
\end{aligned}
$$

$$\frac{D_m - ED_m}{\sqrt{\mathrm{Var}D_m}} \to N(0,1)$$

where $h = -p\log p - q\log q$ is the entropy, $h_2 = p\log^2 p + q\log^2 q$,

$$c_2 = \frac{h_2 - h^2}{h^3}\ ,$$

and $\gamma = 0.577\ldots$ is the Euler constant, while $H_{b-1} = \sum_{i=1}^{b-1}\frac{1}{i}$, $H_0 = 0$ are harmonic numbers. The constant $\Delta(b,p)$ can be computed recursively (e.g., $\Delta(1,03) = 1.25$, $\Delta(5,0.3) = 0.83$, $\Delta(50,0.3) = 0.36$).

Finally, $\delta(x)$ is a fluctuating function with a small amplitude when $\log p/\log q$ is rational, and $\lim_{x\to\infty}\delta(x) = 0$ otherwise.

Continuation

Furthermore:

$$EL_m = mED_m \quad , \quad \mathrm{Var}L_m \sim m\mathrm{Var}D_m \ ,$$

$$\frac{L_m - EL_m}{\sqrt{\mathrm{Var}L_m}} \to N(0,1) \ .$$

More precisely, for a complex ϑ in a neighbourhood of zero, and for any $\varepsilon > 0$

$$e^{-\vartheta c_1 m \log m} E e^{\vartheta L_m} = e^{c_2 \frac{\vartheta^2}{2} m \log m} \left(1 + O(1/m^{1/2+\varepsilon})\right) \ .$$

where $c_1 = 1/h$.

We also conjectue that:

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

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

Second-Order Properties of LZ Model

THEOREM: Let $\Lambda(x)$ be a continuous approximation of EL_m, thus $\Lambda(x) = \frac{x}{h}\log x - \frac{A}{h}x + O(\log x)$ where A represents the constant term of (1). Define x_n as a solution of $\Lambda(x_n) = n$, so

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

Then:

$$\frac{D_n^{LZ} - h^{-1}\log(nh/\log n)}{\sqrt{c_2\log(nh/\log n)}} \to N(0,1),$$

$$\begin{aligned} EM_n^k &= x_n^k\left(1 + O\left(\sqrt{\frac{\log n}{n}}\right)\right) + O\left(\frac{n^{k-1}}{\log^{k-1} n}\right) \ , \\ \mathrm{Var}M_n &\sim \frac{c_2 h^3 n}{\log^2 n} \ , \end{aligned}$$

$$\frac{M_n - EM_n}{\sqrt{\mathrm{Var}M_n}} \to N(0,1) \ ,$$

Moreover, moments of M_n converge to the appropriate moments of the normal distribution.

Continuation

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

$$\Pr\{|M_n - EM_n| > \varepsilon EM_n\} \leq A \exp\left(-a\varepsilon\sqrt{n}\right) \qquad (2)$$

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

We conjecture that the following is true:

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

where $0 < y < 1$, and $I(y)$ is a positive function of y.

Relationships Between DT and LZ Models

Let $D_m(i)$ be the the length of the ith phrase in the digital tree model.

Renewal Equation:

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

which implies:

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

Also, by Billingsley's result: If

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

then

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

where $N(0,1)$ is the standard normal distribution, and μ_m and σ_m are positive constants.

Main Idea of the Proof

Observe that:

$$\begin{aligned} EM_n^{k+1} &= (k+1)\sum_{m\geq 0} m^k \Pr\{M_n > m\} \\ &= (k+1)\sum_{m\geq 0} m^k \Pr\{L_m \leq n\} \\ &= (k+1)\int_0^\infty x^k \Pr\{L_x \leq n\} dx + O(EM_n^k)\ . \end{aligned}$$

But:

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

where $\Phi(x) = 1/2 + 1/2\mathrm{erf}(x/\sqrt{2})$ with

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

Continuation

But

$$\int_0^{x_n} x^k \Phi\left(\frac{n-\mu(x)}{\sigma(x)}\right) dx = \frac{x_n^{k+1}}{k+1} - \underbrace{\frac{1}{2}\int_0^{x_n} x^k \left(1 - \operatorname{erf}\left(\frac{n-\mu(x)}{\sqrt{2}\sigma(x)}\right)\right) dx}_{I_1} ,$$

and

$$I_2 = \int_{x_n}^{\infty} x^k \Phi\left(\frac{n-\mu(x)}{\sigma(x)}\right) dx = \frac{1}{2}\int_{x_n}^{\infty} x^k \left(1 - \operatorname{erf}\left(\frac{\mu(x)-n}{\sqrt{2}\sigma(x)}\right)\right) dx .$$

We can prove that

$$I_2 - I_1 = O(1/\log n) .$$