Analytical Information Theory: Average Profile of the Lempel-Ziv Scheme for a Markovian Source*

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

JING TANG
Applied Microsystems Corporation
Redmond, WA 98052 U.S.A.

June 3, 1997

*Collaboration with P. JACQUET, INRIA, France.

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

Outline of the Talk

1. Lempel-Ziv'78 Scheme and Its Digital Tree Construction
2. Digital Tree Model Versus Lempel-Ziv Model
3. Probabilistic Model
4. Main Results
   - Digital Tree Model
   - Lempel-Ziv Model
5. Sketch of Proof for Digital Tree Model
6. Sketch of Proof for Lempel-Ziv Model
7. Open Problems

Lempel-Ziv'78 Scheme

Lempel-Ziv'78 scheme partitions a sequence into phrases such that the next phrase is the shortest phrase not seen in the past.

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

What is the typical length of a Phrase Length?
How many phrases one can construct from a string of length n?
What is the distribution of the last symbol in a typical phrase?
What is the length of the longest phrase?
What is the redundancy of the LZ code?

Digital Tree Construction of LZ78

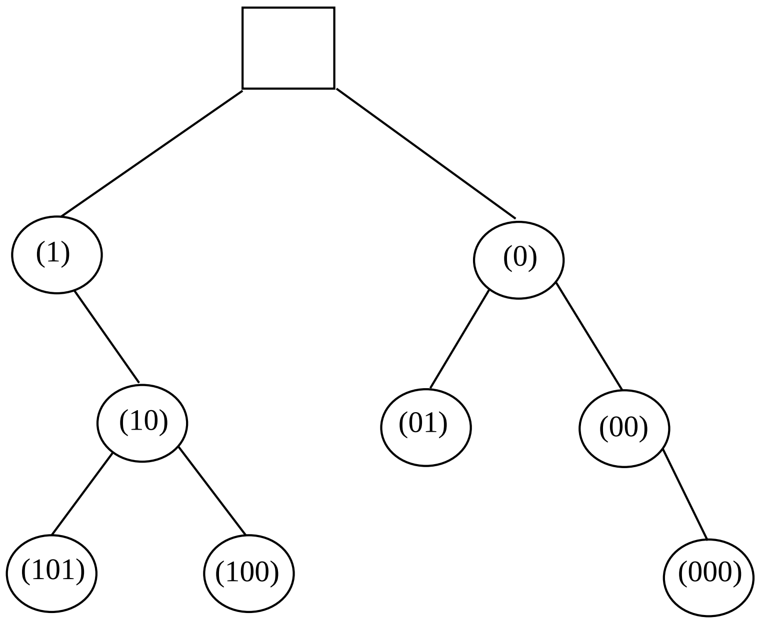

Figure 1: A digital tree representation of the Lempel-Ziv parsing for the string $11001010001000100\ldots$

Digital Tree Model Versus Lempel-Ziv model

In the **digital tree model** (DT model) or **Gilbert-Kadota** one constructs Lempel-Ziv phrases from m INDEPENDENT strings (possibly of infinite length).

The **Lempel-Ziv model** (LZ) assumes that the string is of fixed length, say n.

Parameters for Both Models

M_n – number of phrases in the LZ model
L_m – length of LZ sequence in the DT model
D_m – length of a randomly selected phrase in the digital tree model
D_n^{LZ} – length of a randomly selected phrase in the LZ model.

EXAMPLE: Let $m = 4$ sequences are given: $X_1 = 0000\ldots$, $X_2 = 1010\ldots$, $X_3 = 1111\ldots$ and $X_4 = 0101\ldots$. Then, the Lempel-Ziv sequence is

$$(0)(1)(11)(01)$$

of length $L_4 = 6$ and depth $D_4 = 1\frac{1}{2}$.

Probabilistic Model – Markov Source

Throughout this talk, we shall assume that sequences are generated by a MARKOV SOURCE over a finite alphabet $\mathcal{A} = \{1, 2, \ldots, V\}$, that is, the next symbol depends on a finite number of previous symbols.

We shall deal with Markov chains of ORDER ONE, and assume the transition probability matrix $P = \{p_{ij}\}_{i,j=1}^{V}$. Throughout, only STATIONARY Markov chains are considered (but our technique allows to extend our results to non-stationary case) with stationary vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_V)$ satisfying $\boldsymbol{\pi} P = \boldsymbol{\pi}$.

NOTATIONS

$$
\begin{aligned}
h &= -\sum_{i=1}^{V}\sum_{i=1}^{V} \pi_i p_{ij} \log p_{ij} \qquad \text{(entropy)} \\
h_\pi &= -\sum_{i=1}^{V} \pi_i \log \pi_i
\end{aligned}
$$

Markov Source – More Notations

To express our results, we need to introduce matrices over the complex plane. We denote by s a complex variable. Then:

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

Let now $Q^*(s) = \mathbf{adj}[Q(s)]$ be the adjoint matrix of $Q(s)$. Then

$$\beta = [\det Q(s)]''|_{s=-1},$$
$$\dot{Q}^* = \dot{Q}^*(s)|_{s=-1}.$$

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

where

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

and

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

Notation

Finally:

$$\boldsymbol{\pi}(s) = [\pi_1^{-s}, \pi_2^{-s}, \ldots, \pi_V^{-s}], \quad \boldsymbol{\pi} = \boldsymbol{\pi}(-1),$$

and

$$\omega = \det \begin{bmatrix} 1 & -p_{12} & \ldots & -p_{1V} \\ 1 & 1 - p_{22} & \ldots & -p_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & -p_{V2} & \ldots & 1 - p_{VV} \end{bmatrix} .$$

Main Results – Digital Tree Model

Theorem 1. *Consider a Markovian (stationary) source with transition probabilities* $\{p_{ij}\}_{i,j=1}^{V}$. *Then*

$$\begin{aligned}
\mathbf{E}[D_m] &= \frac{1}{h}\left(\log m + \gamma - 1 + h - h_\pi - \frac{\beta}{2\omega h}\right.\\
&- \left.\boldsymbol{\pi}\dot{X}(-2) + \delta_1(\log m)\right) + O\left(\frac{\log m}{m}\right)\\
\mathbf{Var}\ [D_m] &= \frac{1}{h^3}\left(-\frac{\beta}{\omega} - \frac{2}{\omega}\boldsymbol{\pi}\dot{Q}^*\Psi - h^2\right)\log m + O(1)\ ,
\end{aligned}$$

and

$$\frac{D_m - \mathbf{E}[D_m]}{\sqrt{\mathbf{Var}\ D_m}} \to N(0,1)$$

where $\gamma = 0.577\ldots$ *is the Euler constant,* $\delta_1(x)$ *is a fluctuating function with a small amplitude when*

$$\frac{\log p_{ij} + \log p_{1i} - \log p_{1j}}{\log p_{11}} = \text{rational}$$

for any $i,j = 1,2,\ldots,V$, *and* $\lim_{x\to\infty}\delta_1(x) = 0$ *otherwise.*

More precisely: Let $G_m(u)$ be the ordinary generating function of D_m (i.e., $G_m(u) = \mathbf{E}[u^{D_m}]$), $\mu_m = \mathbf{E}[D_m]$, and $\sigma_m = \sqrt{\mathbf{Var}\ D_m}$. Then, for a complex τ

$$e^{-\tau\mu_m/\sigma_m} G_m(e^{\tau/\sigma_m}) = e^{\frac{\tau^2}{2}} \left(1 + O\left(\frac{1}{\sqrt{\log m}}\right)\right)$$

as $m \to \infty$.

Remark. When the Markovian model is reduced to the Bernoulli model, that is, $p_{ji} = \pi_i$ for $i, j = 1, 2, ..., V$, then $\omega = 1$,

$$\beta = -h_2 = \sum_{i=1}^{V} \pi_i \log^2 \pi_i,$$

$$Q^{-1}(s) = \frac{1}{1 - \boldsymbol{\pi}(s)\Psi}[(1 - \boldsymbol{\pi}(s)\Psi)I + \Psi\boldsymbol{\pi}(s)],$$

and

$$\boldsymbol{\pi}\dot{X}(-2) = \sum_{k=1}^{\infty} \frac{\sum_{i=1}^{V} \pi_i^{k+1} \log \pi_i}{1 - \sum_{i=1}^{V} \pi_i^{k+1}}.$$

In summary, our findings for the Markovian model coincide with the Bernoulli model.

Main Results – Lempel-Ziv Model

Theorem 2. *For large n*

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

where $N(0,1)$ is the standard normal distribution, and

$$c_2 = \frac{1}{h^3}\left(-\frac{\beta}{\omega} - \frac{2}{\omega}\boldsymbol{\pi}\dot{Q}^*\Psi - h^2\right) \ .$$

Moreover, moments of D_n^{LZ} converge to the appropriate moments of the normal distribution.

Average Redundancy of LZ78

POINTWISE REDUNDANCY r_n of the LZ code is defined as

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

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

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

Theorem 1 indicates the following formula for $\bar{r}_n$ (binary alphabet)

$$
\begin{aligned}
\bar{r}_n &= h\frac{2 - \gamma + h_\pi - h + \frac{\beta}{2\omega h} + \boldsymbol{\pi}\dot{X}(-2)) - \delta_1(n)}{\log n} \\
&+ O\left(\frac{\log\log n}{\log^2 n}\right) \ .
\end{aligned}
$$

Previously Known Results for LZ78

- Ziv & Lempel (1978) – (a.s.) behavior of M_n (Wyner & Ziv)

- Aldous & Shields (1988) – limiting distribution for M_n for UNBIASED memoryless sources.
- E. Plotnik, M.J. Weinberger, and J. Ziv (1992) – upper bound on the LZ code redundancy.

- Louchard & Szpankowski (1995) – limiting distribution and moments of D_m and D_m^{LZ} for memoryless sources.

- Jacquet & Szpankowski (1995) – limiting distribution for M_n for memoryless sources.

- Louchard & Szpankowski (1995, 1997) – average redundancy of LZ78 for memoryless sources.

- Savari (1997) – average redundancy of LZ78 for memoryless and Markov sources.

Sketch of Proof for DTM – Main Ideas

1. Recurrence for the generating function of D_m. In fact, instead of D_m we shall deal with
   $B_{m,i}$ – number of nodes in the digital tree at level i (this is called **profile** of the DST or LZ scheme).
   Observe that

$$\Pr\{D_m = i\} = \frac{\mathbf{E}[B_{m,i}]}{m} = \frac{\bar{B}_{m,i}}{m}$$

2. Simplifying the recurrence on $B_m(u) = \mathbf{E}[u^{\bar{B}_{m,i}}]$ by considering the **Poisson model**, that is, replacing the fixed number of strings m by a Poisson stream of strings with mean m.

3. Solving asymptotically the Poisson model through the **Mellin transform** approach.

4. Depoissonization.

Recurrence

Let $B_m(u)$ denote the (ordinary) generating function of $\bar{B}_{m,k}$ (average # of nodes on DST at level k – so called PROFILE). In addition, we denote by $B_m^i(u)$ the generating function of $B_{m,k}^i$ – the profile for digital search trees built from strings starting with symbol $i \in \mathcal{A}$. Then for $i = 1, \ldots, V$

$$\begin{array}{ll} B_{m+1}^1(u) & = u \sum_{|\mathbf{j}|=m} \binom{m}{\mathbf{j}} p_{11}^{j_1} \ldots p_{1V}^{j_V} [B_{j_1}^1(u) + \ldots + B_{j_V}^V(u)] + 1 \\ \cdots & = \cdots \\ B_{m+1}^V(u) & = u \sum_{|\mathbf{j}|=m} \binom{m}{\mathbf{j}} p_{V1}^{j_1} \ldots p_{VV}^{j_V} [B_{j_1}^1(u) + \ldots + B_{j_V}^V(u)] + 1 \end{array}$$

where $\mathbf{j} = (j_1, j_2, \ldots, j_V)$, $|\mathbf{j}| = j_1 + j_2 + \ldots + j_V$ and $B_0^i(u) = 0$.

Observe that $B_m(u)$ then becomes

$$B_{m+1}(u) = u \sum_{|\mathbf{j}|=m} \binom{m}{\mathbf{j}} \pi_1^{j_1} \ldots \pi_V^{j_V} [B_{j_1}^1(u) + \ldots + B_{j_V}^V(u)] + 1$$

Poisson Model

Let

$$\tilde{B}^i(u,z) = \sum_{n=1}^{\infty} \tilde{B}^i_n(u)\frac{z^n}{n!}e^{-z}, i = 1, 2, ..., V,$$

$(\frac{\partial}{\partial z}\tilde{B}^i(u,z) = \tilde{B}^i_z(u,z))$ be the POISSON TRANSFORM of $B^i_m(u)$. Observe that $\tilde{B}^i(u,z) = \mathbf{E}[u^{B^i_{N(z)}}]$ where $N(z)$ is Poisson with mean z.

Then the above set of recurrences becomes:

$$\tilde{B}_z(u,z)+\tilde{B}(u,z) = u[\tilde{B}^1(u,\pi_1 z)+...+\tilde{B}^V(u,\pi_V z)]+1,$$

and (conditional Poisson transforms)

$$\tilde{B}^i_z(u,z)+\tilde{B}^i(u,z) = u[\tilde{B}^1(u,p_{i1}z)+...+\tilde{B}^V(u,p_{iV}z)]+1$$

for $i = 1, 2, ..., V$.

The above differential-functional equations are used to estimate the mean, the variance and the limiting distribution of B_m. Below, we only sketch our derivation for the mean EB_m.

Recurrence for the Mean

Taking the first derivative of $\tilde{B}(u,z)$ at $u = 1$ we can compute the mean EB_m^i (after DEPOISSONIZATION of $\tilde{B}_u(1,z)$)

Observe the following recurrence for $\tilde{B}_u(1,z)$ (the mean of the profile in the Poisson model):

$$\begin{array}{rcl} \tilde{B}^1_{zu}(1,z) + \tilde{B}^1_u(1,z) & = & z + [\tilde{B}^1_u(1,p_{11}z) + \ldots + \tilde{B}^V_u(1,p_{1V}z)] \\ \ldots & = & \ldots \\ \tilde{B}^V_{zu}(1,z) + \tilde{B}^V_u(1,z) & = & z + [\tilde{B}^1_u(1,p_{V1}z) + \ldots + \tilde{B}^V_u(1,p_{VV}z)], \end{array}$$

and

$$\tilde{B}_{zu}(1,z) + \tilde{B}_u(1,z) = z + [\tilde{B}^1_u(1,\pi_1 z) + \ldots + \tilde{B}^V_u(1,\pi_V z)].$$

Mellin Land

Mellin Transform $f^*(s)$ of a real-valued function $f(x)$ is defined as

$$f^*(s) = \mathcal{M}[f(x)] = \int_0^\infty f(x)x^{s-1}dx$$

provided it exists.

Let

$$\begin{aligned} B_i^*(s) = \mathcal{M}[B_u^i(1,z)] = \Gamma(s)x_i(s), \\ B^*(s) = \mathcal{M}[B_u(1,z)] = \Gamma(s)x(s), \end{aligned}$$

be Mellin transforms of $B_u^i(1,z)$ and $B_u(1,z)$, respectively, where $\Gamma(s)$ is the Euler gamma function. These Mellin transforms exist in $\Re(s) \in (-2,-1)$.

Observe

$$\begin{array}{rcl} -(s-1)B_1^*(s-1) + B_1^*(s) & = & B_1^*(s)p_{11}^{-s} + \ldots + B_V^*(s)p_{1V}^{-s}, \\ \ldots & = & \ldots \\ -(s-1)B_V^*(s-1) + B_V^*(s) & = & B_1^*(s)p_{V1}^{-s} + \ldots + B_V^*(s)p_{VV}^{-s}, \end{array}$$

and

$$-(s-1)B^*(s-1) + B^*(s) = B_1^*(s)\pi_1^{-s} + \ldots + B_V^*(s)\pi_V^{-s}.$$

Mellin Transform Solution

Using $\Gamma(s) = (s-1)\Gamma(s-1)$ we can reduce the above to the following matrix equation

$$X(s) - X(s-1) = P(s)X(s),$$

where $X(s) = [x_1(s), \ldots, x_V(s)]^T$. Thus:

$$X(s) = Q^{-1}(s)X(s-1) = [\prod_{i=0}^{\infty} Q^{-1}(s-i)]K,$$

where K is a constant that can be easily determined since we know that $X(-2) = [1, \ldots, 1]^T$.

Finally, observe that

$$B^*(s) = \boldsymbol{\pi}(s)\Xi(s) + \Gamma(s)x(s-1)$$

where $\Xi(s) = [B_1^*(s), \ldots, B_V^*(s)]^T$, $\boldsymbol{\pi}(s) = (\pi^{-s}, \ldots, \pi_V^{-s})$. Observe that $x(s-1) = 1 + O(s+1)$.

ASYMPTOTIC BEHAVIOR OF $B_u(1, z)$ FOR $z \to \infty$ DEPENDS ON THE SINGULARITIES OF THE MATRIX $Q^{-1}(s)$ THAT WE STUDY NEXT.

Singularities

Lemma 1. *Let $Q(s)$ be defined as above. Then the following is true:*

(i) $Q(s)$ is nonsingular for $\Re(s) < -1$, and $s_0 = -1$ is a simple pole.

(ii) If and only if

$$\frac{\log p_{ij} + \log p_{1i} - \log p_{1j}}{\log p_{11}} = \text{rational}$$

for any $i, j = 1, 2, ..., V$, then $Q(s)$ has simple poles on the line $\Re(s) = -1$, and for any integer l all such poles are

$$s_l = -1 + l\theta\jmath$$

with $\jmath = \sqrt{-1}$ and $\theta = \frac{n_1}{n_2}|\frac{2\pi}{\log p_{11}}|$, where integers n_1, n_2 are such that $\{|\frac{n_1}{n_2 \log p_{11}}(\log p_{ij} - \log p_{1i} + \log p_{1j})|\}_{ij=1}^{V}$ is a set of co-prime integers.

(iii) Finally, $Q(-1 + l\theta\jmath) = E^{-l}Q(-1)E^{l}$ where $E = \mathbf{diag}(1, e^{\theta_{12}\jmath}, \ldots, e^{\theta_{1V}\jmath})$ with $\theta_{ik} = -\theta \log p_{ik}$.

Asymptotics

The Laurent expansions of $X(s)$ and $V(s)$ at the dominant pole $s = -1$ give:

$$\begin{aligned} Q^{-1}(s) &= \frac{1}{s+1}Q_1 + Q_2 + O(s+1), \\ X(s-1) &= \Psi + \dot{X}(-2)(s+1) + O(s+1)^2, \\ \Gamma(s) = & \quad \frac{-1}{s+1} + \gamma - 1 + O(s+1). \end{aligned}$$

Thus:

$$B^*(s) = \frac{1}{(s+1)^2}\boldsymbol{\pi} A_1 + \frac{1}{s+1}[\boldsymbol{\pi} A_2 + \boldsymbol{\pi}'(-1)A_1 - 1] + O(1),$$

where (after tedious algebra)

$$\begin{aligned} A_1 &= -Q_1\Psi = \frac{1}{h}\Psi, \\ A_2 &= (\gamma-1)Q_1\Psi - Q_2\Psi - Q_1\dot{X}(-2), \\ &= -\frac{1}{h}(\gamma-1)\Psi + \frac{1}{\alpha h}\dot{Q}^*\Psi + \frac{\beta}{2\alpha h^2}\Psi + \frac{1}{h}\boldsymbol{\pi}\dot{X}(-2)\Psi. \end{aligned}$$

Finally ....

Finally, around the dominant pole $s = -1$ we obtain:

$$
\begin{aligned}
B^*(s) &= \frac{1}{(s+1)^2}\frac{1}{h} + \frac{1}{s+1}[-\frac{1}{h}(\gamma - 1) + \frac{1}{\omega h}\boldsymbol{\pi}\dot{Q}^*\Psi \\
&+ \frac{\beta}{2\omega h^2} + \frac{1}{h}\boldsymbol{\pi}\dot{X}(-2) + \frac{h_\pi}{h} - 1] + O(1) \ .
\end{aligned}
$$

The nondominant poles $s_l = -1 + l\theta j$ contribute:

$$
B^*(s) = -\frac{1}{h}\rho_l \boldsymbol{\pi}(-1 + l\theta\jmath)\Psi(l)\frac{1}{s + 1 - l\theta\jmath} + O(1)
$$

where $\Psi(l) = E^{-l}\Psi$.

Inverse Mellin

Taking the inverse Mellin (e.g., $f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f^*(s)x^{-s}ds$), leads to

$$
\begin{aligned}
\tilde{B}_u(1,z) &= \frac{1}{h}z\log z + \frac{1}{h}[\gamma - 1 - \frac{\beta}{2\omega h} - \frac{1}{\omega}\boldsymbol{\pi}\dot{Q}^*\Psi \\
&- \boldsymbol{\pi}\dot{X}(-2) + h - h_\pi]z + \delta_1(z) + O(\log z)
\end{aligned}
$$

where

$$
\delta_1(z) = -\frac{1}{h}\sum_{l=0} \rho_l \boldsymbol{\pi}\Psi(l) z^{1-l\theta_\jmath} \ .
$$

Observe that the other poles of $Q^{-1}(s-1)$, $Q^{-1}(s-2)$, ... at $s=0$, $s=1$, etc. contributed only $O(\log z)$, so they can be savely ignored for our analysis.

Analytical Depoissonization

So far, we have computed $\tilde{B}_u(1,z) = EB_N$ where N is Poisson with mean m, that is the Poisson mean. But, we need $ED_m = EB_m/m$, that is, we must depoissonize the above results.

DEPOISSONIZATION IDEA

In general, let g_n be a sequence, and

$$\widetilde{G}(z) = \sum_{n=0}^{\infty} g_n \frac{z^n}{n!} e^{-z} = \mathbf{E}[g_N]$$

its Poisson transform. One may ask:

How to infer asymptotics of g_n from an asymptotic expansion of $\widetilde{G}(z)$ for $z \to \infty$ in a cone?

EXAMPLE: Let $g_n = n(n-1)$, then $\widetilde{G}(z) = z^2$. Hence:

$$g_n \sim \widetilde{G}(n) \ .$$

Is it true in general?

Depoissonization

Lemma 1. (Depoissonization Lemma)
(i) *Let* $\widetilde{G}(z)$ *be the Poisson transform of a sequence* g_n *that is assumed to be an entire function of* z. *We postulate that for* $0 < |\theta| < \pi/2$ *the following two conditions simultaneously hold for some numbers* $A, B, \xi > 0$, β, *and* $\alpha < 1$:
(I) *For* $z \in S_\theta = \{z : \ |\arg(z)| < \theta\}$

$$|z| > \xi \quad \Rightarrow \quad |\widetilde{G}(z)| \leq B|z|^\beta \Psi(|z|) \ ,$$

where $\Psi(z)$ *is a slowly varying function (e.g.,* $\Psi(z) = \log^d z$ *for some* $d > 0$*),*

(O) *For* $z \notin S_\theta$

$$|z| > \xi \quad \Rightarrow \quad |\widetilde{G}(z)e^z| \leq A \exp(\alpha|z|) \ .$$

Then, for large n,

$$g_n = \widetilde{G}(n) + O(n^{\beta-1}\Psi(n)) \ ,$$

or more precisely:

$$g_n = \widetilde{G}(n) - \frac{1}{2}\widetilde{G}''(n) + O(n^{\beta-2}\Psi(n)) \ .$$

(ii) *If the above two conditions, namely (I) and (O), hold for* $\widetilde{G}(z,u) = \sum_{n=0}^{\infty} G_n(u) \frac{z^n}{n!} e^{-z}$ *for* u *belonging to a compact set* $\mathcal{U}$*, then*

$$G_n(u) = \widetilde{G}(n,u) + O(n^{\beta-1}\Psi(n))$$

for large n *and uniformly in* $u \in \mathcal{U}$*.*

Depo of Polynomially Growing Sequences

Lemma 2. *Let $g(z)$ be an analytical continuation of a sequence g_n whose Poisson transform is $\widetilde{G}(z)$, and such that $g(z) = O(z^\beta)$ in a linear cone. Then, for some θ_0 and for all linear cones $\mathcal{S}_\theta$ ($\theta < \theta_0$), there exists $\alpha < 1$ and $A > 0$ such that*

$$z \notin \mathcal{S}_\theta \qquad \Rightarrow \qquad |\widetilde{G}(z)e^z| \leq Ae^{\alpha|z|} ,$$

thus Lemma 1 can be applied automatically.

Finishing the proof ...

To complete the proof of our Theorem 1, it is enough to observe that

$$EB_m = O(m \log m) ,$$

and apply Lemmæ 1 and 2.

LZ Model – Upper Bound

Now, we use Theorem 1 to prove the limiting distribution of the typical phrase length D_n^{LZ} in the Lempel-Ziv model (i.e., when one partition a string of length n).

Upper Bound

We observe that the phrase length in the LZ'77 scheme (the longest prefix of X_{n+1}^{∞} that occurs in the "training sequence" X_1^n) cannot be shorter that the typical phrase length of LZ'78.

A.J. Wyner (1997) proved that the phrase length of LZ'77 in a Markovian model is normally distributed, thus we can infer a similar conclusion for the phrase length of LZ'78 (an upper bound).

LZ Model – Lower Bound

1. We first use the fact that

$$\Pr\{D_n^{LZ} = k\} \sim \Pr\{D_{\lfloor nh/\log n \rfloor}^{DST} = k\}$$

for $k = O(\log n)$ (for memoryless source see Louchard & Szpankowski (1995)).

2. **DST–Mixing Model**.
Fix $m = O(n/\log n)$. Select every Kth phrase from m phrases generated by LZ'78 algorithm. Built a new LZ sequence out of these m/K strings (in other words: build a DST from these m/K strings). A new typical phrase cannot be longer than in the original LZ'78 scheme. Denote this new typical phrase as: $\widetilde{D}_{m/K}^{DST}$. Observe that we are now almost in the digital tree model situation except that the strings are WEAKLY dependent.

Lower Bound

3. We shall prove that

$$\Pr\{\widetilde{D}_{m/K}^{DST} = k\} = (1 + o(1))\Pr\{D_m = k\} \qquad (*)$$

where D_m is the depth in a DST built from m Independent Markovian sequences. From Theorem 1 we know that

$$\frac{D_m - c_1 \log m}{\sqrt{c_2 \log m}} \to N(0, 1) \ .$$

(where $c_1 = 1/h$ and c_2 as in Theorem 1), which will imply

$$\frac{D_{m/K}^{DST} - c_1 \log m}{\sqrt{c_2 \log m}} \to N(0, 1) \ .$$

for $m = O(n/\log n)$.

Lower Bound

4. Mixing Model:
Let $\mathcal{F}_m^n$ be a σ-field generated by $X_{k=m}^n$ for $m \leq n$. There exists a function $\alpha(\cdot)$ of g such that: (i) $\lim_{g\to\infty} \alpha(g) = 0$, (ii) $\alpha(1) < 1$, and (iii) for any m, and two events $A \in \mathcal{F}_{-\infty}^m$ and $B \in \mathcal{F}_{m+g}^\infty$ the following holds

$$(1-\alpha(g))\Pr\{A\}\Pr\{B\} \leq \Pr\{AB\} \leq (1+\alpha(g))\Pr\{A\}\Pr\{B\}.$$

5. To prove (*), let $D_m^i(u)$ and $\widetilde{D}_m^i(u)$ be the probability generating functions of D_m and $D_{m/K}^{DST}$ for the Markov model when the first symbol of all strings is i. For simplicity consider only a binary alphabet $\{0,1\}$. Then:

$$\begin{aligned}
D_{m+1}^i(u) &= u \sum_{k=0}^m P_{m,k}(p_{i1} D_k^i(u) + p_{i2} D_{m-k}^i(u)), \\
\widetilde{D}_{m+1}^i(u) &= u \sum_{k=0}^m \widetilde{P}_{m,k}(p_{i1} \widetilde{D}_k^i(u) + p_{i2} \widetilde{D}_{m-k}^i(u)).
\end{aligned}$$

where $P_{m,k} = \binom{m}{k} p_{i1}^k p_{i2}^{m-k}$ and $\widetilde{P}_{m,k}$ satisfy the following mixing condition (with $m\alpha(m) \to 0$)

$$(1-\alpha(m))^m P_{m,k} \leq \widetilde{P}_{m,k} \leq (1+\alpha(m))^m P_{m,k}$$

Lower Bound

6. Using induction, and the above recurrence we can prove that for $m\alpha(m) \to 0$ we have

$$D_m^i(u) = (1 + O(m\alpha(m))\widetilde{D}_m^i(u).$$

Hence, the unconditional generating functions $D_m(u)$ and $\widetilde{D}_m(u)$ satisfy:

$$D_m(u) = (1 + O(m\alpha(m))\widetilde{D}_m(u).$$

which proves (*).

7. Now, it remains only to show that we can choose such an $\alpha(m)$ that $m\alpha(m) \to 0$ as $m \to \infty$. We observe that for any Markov chain over a finite alphabet there exists $\rho < 1$ such that the mixing condition $\alpha(g) = O(\rho^g)$ for some g. In our case, g is equal to K phrase lengths in the LZ model. With high probability every phrase is of length $O(\log m)$. More precisely, for almost all i except εm we have $D_m \sim 1/h \log m$ (for $m = O(n/\log n)$. Thus, $K \geq -h/\log \rho$ suffices.

Open Problem I: Limiting Distribution of M_n

The main open problem in lossless LZ78 is to prove the central limit theorem and large deviations for the number of phrases M_n. The problem is very difficult! We know now how to do the digital tree model, but must translate it to the LZ model (the above discussed trick does not work here).

Open Problem II: Lossy Extension of LZ'78

Consider a natural lossy extension of the LZ'78: that is, the next phrase the the longest phrase that is within distance D from a previous phrase (or more precisely: the shortest phrase that is within distance BIGGER than D from all previous phrases).

Let's introduce a D-ball $B_D(w_k)$ with center $w_k \in \mathcal{A}^k$ as $B_D(w_k) = \{x_1^k : \ d(w_k, x_1^k) \le D\}$. We shall write $P(B_D(X_1^k))$ for the probability of such a ball.

Definition. (Generalized Renyi Entropy) For a fixed $D > 0$ let

$$r_0(D) = \lim_{k\to\infty} \frac{-E \log P(B_D(X_1^k))}{k} ,$$

CONJECTURE. We conjecture the following is true with high probability (see Luczak&Szpankowski (1997) for LZ77 extension)

$$\begin{aligned} \frac{D_n(D)}{\log n} &\to \frac{1}{r_0(D)} \quad \text{(pr.)} \\ \frac{M_n(D) \log M_n(D)}{n} &\to r_0(D) \quad \text{(a.s.)} . \end{aligned}$$