Proceedings of the 29th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Uppsala, Poland, 25-29 June 2018. Submission

Towards LZ'78 Analysis for Markov Sources: Distribution of Tail Symbols in DST

Philippe Jacquet[1], Wojciech Szpankowski[2†]

[1] *Bell Labs- Nokia, France.*
[2] *Department of Computer Science, Purdue University, USA*

Lempel-Ziv'78 is one of the most popular data compression algorithm on words. Over the last decades we uncover its fascinating behavior and understand better many of its beautiful properties. Among others, in 1995 by settling the Ziv conjecture we proved that for *memoryless source* (i.e., when a sequence is generated by a source without memory) the number of LZ'78 phrases satisfies the Central Limit Theorem (CLT). Since then the quest commenced to extend it to Markov sources, however, despite several attempts this problem is still open.

In this conference paper, we revisit the issue and focus on a much simpler, but not trivial problem that may lead to the resolution of the LZ'78 dilemma. We consider the associated Digital Search Tree (DST) version of the problem in which the DST is built over a fixed number of Markov generated sequences. In such a model we shall count the number of of the so called "tail symbol", that is, the symbol that follows the last inserted symbol. Our goal here is to analyze this new quantity under Markovian assumption since it plays crucial role in the analysis of the original LZ'78 problem. We establish the mean, the variance, and the central limit theorem for the number of tail symbols. We accomplish it by applying techniques of analytic combinatorics on words also known as analytic pattern matching.

Keywords: Lempel-Ziv'78, Markov sources, digital trees, depoissonization, analytic combinatorics.

1 Introduction

The Lempel-Ziv compression algorithm [20] is a universal compression scheme. It partitions the text to be compressed into consecutive phrases such that the next phrase is the unique shortest prefix of the uncompressed text not seen before in the compressed portion of the text. For example, $11001010001000100\ldots$ is parsed into phrases $()(1)(10)(0)(101)\ldots$. The LZ'78 compression code for a word w over the alphabet $\mathcal{A}$ that we denote as $C(w)$ consists of a pointer to the previous phrase and the last symbol of the current phrase. It is well known that the average compression rate $|C(w)|/|w|$ tends to the source entropy rate h when $|w| \to \infty$, however, one is often more interested in the called called (normalized) redundancy rate $\frac{|C(w)|}{|w|} - h$, that is, the excess of the code length over the optimal code length represented by the entropy of the source. Its behavior is known now for memoryless sources [1, 6, 8] but the last fifty years failed to produce any significant progress for Markov sources.

†W. Szpankowski work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01, and and NCN grant 2013/09/B/ST6/02258.

It is convenient to organize the phrases (dictionary) of the Lempel-Ziv scheme in a *digital search tree* (DST) [11, 19] which represents a parsing tree. The root then contains an empty phrase. The first phrase is the first symbol, say "$a \in \mathcal{A}$" which is stored in a node appended to the root. The next phrase is either "$aa \in \mathcal{A}^2$" stored in another node that branches out from the node containing the first phrase "a" or a new symbol that is stored in a node attached to the root. This process repeats recursively until the text is parsed into full phrases. A detailed description can be found in [2, 6, 10, 19]; see also Fig. 1.

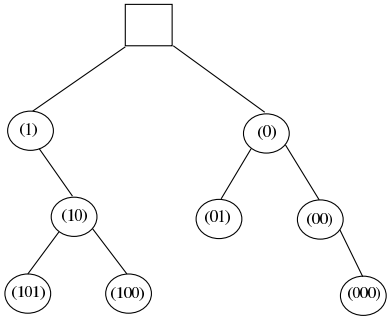

Fig. 1: A digital tree representation of the Lempel-Ziv parsing for the string 11001010001000100 ... into phrases ()(1)(10)(0)(101) ... where () is the empty phrase stored in the root

.

Let a text w be generated over an alphabet $\mathcal{A}$, and let $\mathcal{T}(w)$ be the associated digital search tree constructed by the algorithm. Each node in $\mathcal{T}(w)$ corresponds to a phrase in the parsing algorithm. Let $L(w)$ be the (total) path length in $\mathcal{T}(w)$, that is, the sum of all paths from the root to all nodes (i.e., the sum of phrases which is also the text length). We have $L(w) = |w|$ (if all phrases are full). We also note that the compression code $C(w)$ is a description of $\mathcal{T}(w)$, node by node in the order of creation. The compressed code length is then $|C(w)| = \sum_{k=1}^{M(w)} \lceil \log_2(k) \rceil + \lceil \log_2(|\mathcal{A}|) \rceil$ where $M(w)$ is the number of full phrases needed to phrases w, and the pointer to the kth node requires at most $\lceil \log_2 k \rceil$ bits, while and the next symbol costs $\lceil \log_2 |\mathcal{A}| \rceil$ bits. To simplify, we shall assume throughout that $|C(w)| = M(w)\,(\log(M(w)) + \log(|\mathcal{A}|))$.

To understand LZ'78 behavior one must analyze the limiting distribution of $M(w)$ and/or $L(w)$. Indeed, let $|w| = n$ and denote $L_n := L(w)$ and $M_n = M(w)$. It is is well known that $P(M_n > m) = P(L_m < n)$ linking the number of phrases M_n to the path length L_n in the associated DST. For memoryless sources, this relation and the so called renewal equation are sufficient to analyze the limiting distribution of M_n since one starts with a digital search tree built from m *independently* generated strings [6, 8]. However, this approach fail when the original sequence w is generated by a Markov source since the phrases carved by the LZ'78 algorithm are strongly correlated thought a forward and backward dependence. In this case we need to understand the behavior of the number of phrases that start with a symbol a and end with a symbol b. This is the main challenge one encounters when analyzing LZ'78 under Markovian assumption.

In this conference paper we propose a first step towards resolving the dependence problem between phrases. We consider *only* a digital search tree built from Markov sequences over a binary alphabet $\mathcal{A} = \{a, b\}$. We study here the asymptotic behavior of the following quantity. We consider n independent sequences generated by a Markov source with transition probability $\mathbf{P}$. When a sequence is inserted into the DST, we call the "tail symbol" the symbol that follows the last symbol inserted in the DST. For example, in Figure 1 the tail symbol after phrase (10) is "0".

Let now $c \in \mathcal{A}$ be an arbitrary symbol from the alphabet, that is, it is either an "a" or a "b". We denote $\mathbf{T}_n = (T_n^a, T_n^b)$ the random variable vector representing the number of times the symbol a appears as a tail symbol after the insertion of the n sequences *assuming* that all sequences start with symbol $a \in \mathcal{A}$ for T_n^a or symbol b representing T_n^b. For example, in Figure 1 we have $T_8^{(0)} = 5$. Notice that we have a

stochastic recursion

$$T^c_{n+1} = \delta_a + T^a_{n_a} + T^b_{n_b} \tag{1}$$

where δ_a is equal to 1 when the second symbol of the first sequence is a and is equal to 0 otherwise. The quantity n_c is the number of sequences inserted after the first sequence such that their second symbol is equal to c (thus they fork to the subtree corresponding to c in the DST). Our goal in this paper is to analyze probabilistic behavior of $\mathbf{T}_n$; in particular, its mean and variance, and its limiting distribution. In the Concluding remarks of this paper we show how analysis of T_n can lead to a characterization of LZ'78 algorithm.

Let us now briefly review literature on LZ'78 and DST analysis. Our ultimate goal is to prove the Central Limit Theorem (CLT) for the number of phrases and establish precise rate of decay of the LZ'78 code redundancy for Markov sources. For memoryless sources, this result was already proved in our 1995 paper [6] while the average redundancy was presented in [14, 16], It should be pointed out that since our 1995 paper [6] no simpler, in fact, no new proof of CLT was presented except the one by Neininger and Rüschendorf [15] but only for *unbiased* memoryless sources (as in [1]). The proof of [15] applies the so called *contraction method*. The only known to us analysis of LZ'78 for Markov sources is presented in [9], but the authors restricted their attention to a single phrase. An attempt to analyze of the LZ'78 for Markov sources was reported in [12].

Regarding analysis of digital search trees, and in general digital trees, more is known. The reader is refer to our book [10] for details. Digital trees for memoryless sources were analyzed in [1, 3, 14, 19]. Digital trees under Markovian models were analyzed in [5, 9, 13].

2 Main Results

In this section we present our main results delaying most of the proofs till the last section. We consider a stationary source generating a sequence of symbols drawn from a finite alphabet $\mathcal{A}$. We assume that the source is stationary and ergodic. We will consider a Markovian process of order 1 with the transition matrix $\mathbf{P} = [P(a|b)]_{a,b\in\mathcal{A}}$. For this conference paper we assume that $P(a|b) > 0$ for all $a, b\mathcal{A}$. Extensions to finite alphabet and higher order Markov is possible since a Markovian source of order r is simply a Markovian source of order 1 over the alphabet $\mathcal{A}^r$.

In this section we shall analyze T^c_n representing the number of times the symbol a appears as a tail symbol after the insertion of the n sequences *assuming* that all sequences are generated by a Markov source with transition probability $\mathbf{P}$ and all sequences start with symbol $c \in \mathcal{A}$. We have already observed that the vector T^c satisfies the stochastic equation (1). We will translate it now into the generating function world.

For $c \in \mathcal{A}$ let $D^c_{n,k} = P(T^c_n = k)$ and $D^c_n(u) = E[u^{T^c_n}]$ be the probability generating function of T^c_n defined for a complex variable u. We have the recursion:

$$D^c_{n+1}(u) = (P(a|c)u + 1 - P(a|c)) + \sum_k \binom{n}{k} P(a|c)^k P(b|c)^{n-k} D^a_k(u) D^b_{n-k}(u) \tag{2}$$

subject to (i) $D^c_0(u) = 1$ and (ii) $D^c_1(u) = P(a|c)u + 1 - P(a|c)$. Furthermore, define the bivariate Poisson transform

$$D_c(z,u) = \sum_{n\geq 0} \mathbf{E}[u^{T^c_n}] \frac{z^n}{n!} e^{-z}$$

be the Poisson transform of T_n^c. It satisfies the following differential-functional equation

$$\partial_z D_c(z,u) + D_c(z,u) = D_1^c(u) D_a(P(a|c)z,u)) \cdot D_b(P(b|c)z,u) \tag{3}$$

with $D_c(z,1) = 1$ where ∂_z is the partial derivative with respect to variable z. We also sometimes write $f_z(z,u) := \partial_z f(z,u)$.

We now focus on the first Poisson moment $X_c(z) = \partial_u D_c(z,1)$ where ∂_u is the derivative with respect to variable u. We also study the Poisson variance $V_c(z) = \partial_u^2 D_c(z,1) + X_c(z) - (X_c(z))^2$, and the limiting distribution of T_n^c. After finding asymptotic behavior of the Poisson mean $X_c(z)$ and variance $V_c(z)$ for large $z \to \infty$ we invoke the depoissonization lemma of [7] to extract the original mean and variance:

$$\mathbf{E}[T_n^c] = X^c(n) - \frac{1}{2} n X_z^c(n) + \cdots, \qquad \mathrm{Var}[T_n^c] = V^c(n) - n[X_z(n)]^2 z + \cdots .$$

Let us start with the Poisson mean $X^c(z)$. Taking the derivative of (3) with respect to u and setting $u = 1$ we find

$$\partial_z X_c(z) + X_c(z) = P(a|c) + X_a(P(a|c)z) + X_b(P(b|c)z). \tag{4}$$

To complete this equation we need to calculate the initial values of $\mathbf{E}[T_n^c]$. It is easy to see that

$$\mathbf{E}[T_0^c] = 0, \quad \mathbf{E}[T_1^c] = P(a|c), \quad \mathbf{E}[T_2^c] = P(a|c) + P(a|c)P(a|a) + P(b|c)P(a|b). \tag{5}$$

In a similar fashion we can derive the differential-functional equation for the Poisson variance. After some tedious algebra we arrive at

$$\partial_z V_c(z) + V_c(z) = P(a|c) - P^2(a|c) + [\partial_z X_c(z)]^2 + V_a(P(a|c)z) + V_b(P(b|c)z). \tag{6}$$

Both differential-functional system of equations (3) and (5) can be solved using complicated Mellin transform approach. We will provide details of our approach in the next section. For now we need to introduce some extra notation to present our main results.

For complex s define

$$\mathbf{P}(s) = \begin{bmatrix} P(a|a)^{-s} & P(b|a)^{-s} \\ P(a|b)^{-s} & P(b|b)^{-s} \end{bmatrix}. \tag{7}$$

For such $\mathbf{P}(s)$ we denote by $\lambda(s)$ the main eigenvalue. We also need another matrix

$$\mathbf{Q}(s) = \prod_{i \geq 1} (\mathbf{I} - \mathbf{P}(s-i))^{-1} \prod_{j=-\infty}^{j=-2} (\mathbf{I} - \mathbf{P}(j))$$

which is well defined for $\Re(s) \in (-2, 0)$. We also define $\langle \mathbf{x}, \mathbf{y} \rangle$ as the scalar product of vectors $\mathbf{x}$ and $\mathbf{y}$.

Now we are in the position to formulate our main result.

Theorem 1 *Consider a digital search tree (DST) built over n independent sequences generated by a Markov source. For $(a,b,c) \in \mathcal{A}^3$ define*

$$\alpha_{abc} = \log \left[\frac{P(a|b)P(c|a)}{P(c|b)} \right]. \tag{8}$$

(i) [Aperiodic case] *If not all* $\{\alpha_{abc}\}$ *are rational, then*

$$\mathbf{E}[T_n^a] = n\left(\pi_a + \frac{1}{\lambda'(-1)}\mathbf{1}\langle(\boldsymbol{\pi}'(-1) + \boldsymbol{\pi}\mathbf{Q}'(-1))\,(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{e}_a\rangle\right)\mathbf{1} + o(n). \tag{9}$$

[Periodic case] *If all* $\{\alpha_{abc}\}$ *are rationally related, then*

$$\mathbf{E}[T_n^a] = n\left(\pi_a + \frac{1}{\lambda'(-1)}\mathbf{1}\langle(\boldsymbol{\pi}'(-1) + \boldsymbol{\pi}\mathbf{Q}'(-1))\,(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{e}_a\rangle + O(n)\right)\mathbf{1} + O(n^{-1+\varepsilon}) \tag{10}$$

where $Q(n)$ *is a periodic function and some* $\varepsilon > 0$.
(ii) [Variance] *The variance* $\mathrm{Var}[T_n^a]$ *grows linearly, that is* $\mathrm{Var}[T_n^a] = O(n)$ *for* $a \in \mathcal{A}$.
(iii) [Central Limit Theorem] *For any* $c \in \mathcal{A}$ *we have*

$$\frac{T_n^c - \mathbf{E}[T_n^c]}{\mathrm{Var}[T_n^c]} \to N(0,1)$$

where $N(0,1)$ *denotes the standard normal distribution.*

We notice that, unexpectedly, the number of tail symbols equal to a *is not* converging to $n\pi_a$ as we should expect from a Markovian sequence converging to its stationary state when the length of the sequence increase. It departs from the stationary distribution by the term $\frac{1}{\lambda'(-1)}\langle(\boldsymbol{\pi}'(-1) + \boldsymbol{\pi}\mathbf{Q}'(-1))\,(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{e}_a\rangle$. The reason is that the tail symbol is picked up at random in the sequence but occurs when the sequence path leaves the tree and this introduce a non trivial bias.

3 Proofs

In this section we prove separately Theorem 1(i), then Theorem 1(ii), and finally Theorem 1(iii).

3.1 Proof of Theorem 1(i): Mean

We first analyze asymptotically $\mathbf{X}(z) = (X_a(z), X_b(z))$ that satisfies the system of differential-functional equations (4). We solve this system, and then apply Mellin transform and depoissonization to prove Theorem 1(i).

Since for all integer n, we have $T_n^c \le n$, we notice that the function $X_c(z)$ is $O(z)$ both when $z \to \infty$ and when $z \to 0$. Thus the function $\mathbf{X}(z)$ has no Mellin transform. To correct this we introduce $\tilde{X}_c(z) = X_c(z) - G_c(z)$ with $G_c(z) = (\mathbf{E}[T_1^c]z + \mathbf{E}[T_2^c]z^2/2)e^{-z}$ which is $O(z^3)$ when $z \to 0$, where $\mathbf{E}[T_1^c]$ and $\mathbf{E}[T_1^c]$ are defined in (5).

The Mellin transform $X_c^*(s)$ of $\tilde{X}_c(z)$ on the strip $\Re(s) \in]-3,-1[$ exists. The Mellin transform of $\partial_z\tilde{X}_c(z)$ exists too on the strip $\Re(s) \in]-2,0[$. Thus the two Mellin transforms coexist on the strip $\Re(s) \in]-2,-1[$ and satisfies [19]

$$-(s-1)(X_c^*(s-1) + G_c^*(s)) + X_c^*(s) + G_c^*(s) = P(a|c)^{-s}(X_a^*(s) + G_a^*(s)) + P(b|c)^{-s}(X_b^*(s) + G_b^*(s))$$

where $G_c^*(s)$ for $c \in \mathcal{A}$ is the Mellin transform of $G_c(z)$ and has the explicit expression $\mathbf{E}[T_1^c]\Gamma(1+s) + \mathbf{E}[T_2^c]\Gamma(s+2)/2$. This expression is here for completeness.

An alternative but convenient way to see this equations is to consider the vector $\mathbf{X}^*(s)$ made of the quantities $X_c^*(s)$, which is also the Mellin transform of the vector $\mathbf{X}(z)$ made of the coefficients $\tilde{X}_c(z)$. This yields the linear equation

$$-(s-1)(\mathbf{X}^*(s-1)+\mathbf{G}^*(s-1))+\mathbf{X}^*(s)+\mathbf{G}^*(s)=\mathbf{P}(s)(\mathbf{X}^*(s)+\mathbf{G}^*(s)) \tag{11}$$

where $\mathbf{G}^*(s)$ is the vector of the $G_c^*(s)$. It can be rewritten in

$$(s-1)(\mathbf{X}^*(s-1)+\mathbf{G}^*(s-1))=(\mathbf{I}-\mathbf{P}(s))(\mathbf{X}^*(s)+\mathbf{G}^*(s)).$$

This kind of equation has been studied in [9] where we introduce a new function $\mathbf{x}(s)$ defined as

$$\mathbf{X}^*(s)+\mathbf{G}^*(s)=\Gamma(s)\mathbf{x}(s)$$

for some function $\mathbf{x}(s)$. Thus the equation becomes $\mathbf{x}(s-1)=(\mathbf{I}-\mathbf{P}(s))\mathbf{x}(s)$, which leads to

$$\mathbf{x}(s)=\prod_{i\geq 0}(\mathbf{I}-\mathbf{P}(s-i))^{-1}\mathbf{K}$$

where $\mathbf{K}$ is a constant vector. Notice that the matrices very likely don't commute thus the product order is specified from the left to right. Indeed we have

$$\mathbf{K}=\left(\prod_{j\geq 2}(\mathbf{I}-\mathbf{P}(-j))^{-1}\right)^{-1}\mathbf{x}(-2)=\prod_{j=-\infty}^{j=2}(\mathbf{I}-\mathbf{P}(j))\mathbf{x}(-2). \tag{12}$$

To handle it we need an explicit formula for $\mathbf{x}(-2)$. The following lemma from [9] is useful in this regard.

Lemma 1 *Let $\{f_n\}_{n=0}^{\infty}$ be a sequence of real numbers having the Poisson transform $\widetilde{F}(z)=\sum_{n=0}^{\infty}f_n\frac{z^n}{n!}e^{-z}$, which is an entire function. Furthermore, let its Mellin transform $F(s)$ have the following factorization*

$$F(s)=\mathcal{M}[\widetilde{F}(z);s]=\Gamma(s)\gamma(s).$$

Assume that $F(s)$ exists for $\Re(s)\in(-2,-1)$, and that $\gamma(s)$ is analytic for $\Re(s)\in(-\infty,-1)$. Then

$$\gamma(-n)=\sum_{k=0}^{n}\binom{n}{k}(-1)^k f_k,\quad \text{for}\quad n\geq 2. \tag{13}$$

Now we can compute $\mathbf{x}(-2)$ using above and (5) leading to

$$\mathbf{x}(-2)=\begin{bmatrix} T_2^a-2P(a|a)\\ T_2^b-2P(a|b)\end{bmatrix}. \tag{14}$$

In another notation

$$\mathbf{x}(-2)=(\mathbf{P}^2-\mathbf{P})\mathbf{e}_a, \tag{15}$$

where $\mathbf{e}_a$ is the vector made of a single 1 at a position and zero otherwise.

Next, we notice that the vector

$$\Gamma(s)\prod_{i\geq 0}(\mathbf{I}-\mathbf{P}(s-i))^{-1}\prod_{j=-\infty}^{j=-2}(\mathbf{I}-\mathbf{P}(j))\mathbf{x}(-2)$$

may have a double pole on $s=-1$ since $\Gamma(s)$ has a pole and also $(\mathbf{I}-\mathbf{P}(s))^{-1}$ since $\mathbf{I}-\mathbf{P}(-1)=\mathbf{I}-\mathbf{P}$ is singular. But in fact the pole multiplicity is reduced by one, as prove below. Let us also define

$$\mathbf{Q}(s)=\prod_{i\geq 1}(\mathbf{I}-\mathbf{P}(s-i))^{-1}\prod_{j=-\infty}^{j=-2}(\mathbf{I}-\mathbf{P}(j)).$$

Then $\mathbf{x}(s)=(\mathbf{I}-\mathbf{P}(s))^{-1}\mathbf{Q}(s)\mathbf{x}(-2)$.

We notice that when $s\to -1$, then $\mathbf{Q}(s)=\mathbf{I}+(s+1)\mathbf{Q}'(-1)+O((s+1)^2)$. Furthermore let $\lambda(s)$ be the main eigenvalue of matrix $\mathbf{P}(s)$ and $\mathbf{1}(s)$ and $\boldsymbol{\pi}(s)$ be respectively the right and left main eigenvectors. We have $\lambda(-1)=1$, $\mathbf{1}(-1)$ is all made of one's, and $\boldsymbol{\pi}(-1)$ is the stationary distribution of the Markov source.

From the matrix spectral representation [19] we have

$$\mathbf{P}(s)=\lambda(s)\mathbf{1}(s)\otimes\boldsymbol{\pi}(s)+\mathbf{R}(s)=\lambda(s)\boldsymbol{\Pi}(s)+\mathbf{R}(s) \tag{16}$$

where $\mathbf{R}(s)$ is the automorphism of the eigenplan orthogonal to the main eigenvector and $\boldsymbol{\Pi}(s)=\mathbf{1}(s)\otimes\boldsymbol{\pi}(s)$. so skipping details finally we have

$$\lim_{s\to -1}\mathbf{x}(s)=\mathbf{P}\mathbf{e}_a-\pi_a\mathbf{1}-\frac{1}{\lambda'(-1)}\mathbf{1}\langle(\boldsymbol{\pi}'(-1)+\boldsymbol{\pi}\mathbf{Q}'(-1))\,(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{e}_a\rangle, \tag{17}$$

where π_a is the coefficient of the stationary distribution $\boldsymbol{\pi}$ on symbol a.

Now we are in position to establish asymptotics of $X_c(z)$ for large z and through depoissonization asymptotics of $\mathbf{E}[T_n^c]$. The inverse Mellin transform is

$$\tilde{X}_c(z)=\frac{1}{2i\pi}\int_{x-i\infty}^{x+i\infty}X_c^*(s)z^{-s}ds \tag{18}$$

valid for all $x\in]-2,-1[$. Remembering that $T_c(z)=\tilde{X}_c(z)+P(a|c)z$ we have indeed

$$\tilde{\mathbf{X}}(z)=\frac{1}{2i\pi}\int_{x-i\infty}^{x+i\infty}\Gamma(s)\mathbf{x}(s)z^{-s}ds-\frac{1}{2i\pi}\int_{x-i\infty}^{x+i\infty}\mathbf{G}^*(s)z^{-s}ds \tag{19}$$

We know that $\mathbf{T}(z)-\tilde{\mathbf{X}}(z)$ is decaying exponentially fast when $z\to\infty$.

Moving the line of integration toward the right, we meet a single pole at $s=-1$ of $\mathbf{G}^*(s)z^{-z}$ and its residues is $-z\mathbf{P}\mathbf{e}_a$. Then

$$\frac{1}{2i\pi}\int_{x-i\infty}^{x+i\inf ty}\mathbf{G}^*(s)z^{-s}ds=-\mathbf{P}\mathbf{e}_a+O(z^{-M})$$

for all $M>0$.

The value -1 is also a simple pole for $z^{-s}\Gamma(s)\mathbf{x}(s)$. We know that its residue has expression

$$-z\left(\mathbf{P}\mathbf{e}_a - \pi_a\mathbf{1} - \frac{1}{\lambda'(-1)}\mathbf{1}\langle(\boldsymbol{\pi}'(-1)+\boldsymbol{\pi}\mathbf{Q}'(-1))\,(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{e}_a\rangle\right). \tag{20}$$

Therefore we have

$$\mathbf{X}(z) = z\left(\pi_a + \frac{1}{\lambda'(-1)}\mathbf{1}\langle(\boldsymbol{\pi}'(-1)+\boldsymbol{\pi}\mathbf{Q}'(-1))\,(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{e}_a\rangle\right)\mathbf{1} + o(z). \tag{21}$$

For irrational case, we know that $s=-1$ is th only pole on the line $\Re(s)=-1$, leading to the error term $o(z)$ coming from other poles of $(\mathbf{I}-\mathbf{P}(s))^{-1}$ which may occur on the right half plan of $s=-1$.

But in the rational case, there is the possibility of other poles regularly spaced on the axis $\Re(s)=-1$ with some specific matrices $\mathbf{P}$ detailed in [9] where the coefficients α_{abc} are introduced. In these very specific cases (the uniform probability distribution on $\mathcal{A}$ is one of them) the $o(z)$ term should be replaced by a term $zQ_c(\log z)+O(z^{1-\epsilon})$, where Q_c is a periodic vector of very small amplitude and mean zero, and $\epsilon>0$ depends on the matrix $\mathbf{P}$. This proves Theorem 1(i).

3.2 Proof of Theorem 1(ii): Variance

We now analyze asymptotically $\mathbf{V}(z)=(V_a(z),V_b(z))$ that satisfies the system of differential-functional equations (6). In order to apply depoissonization, for $\theta\in[0,\pi/2]$ we define $\mathcal{C}(\theta)$ as the complex cone containing the complex number z such that $|\arg(z)|\le\theta$. on increasing domains [4, 19, 8]

$$\mathcal{C}_k(\theta)=\{z, z\in\mathcal{C}(\theta)\&|z|\le\rho^k\}$$

with $\rho=\min_c\{\frac{1}{P(a|c)},\frac{1}{P(b|c)}\}$.

Our first goal is to prove that $V_c(z)=O(z)$. We shall use use the increasing domain approach [19] applied to (6) following the footsteps of the proof of Lemma 7A of [6]. From Fact 1 of [6] we conclude that

$$V_c(z)=V_c(\rho z)e^{-z(1-\rho)}+e^{-z}\int_{\rho z}^{z}e^x\left(V_a(P(a|c)x)+V_b(P(b|c)x)+g(x)\right)dx \tag{22}$$

where $g(z)=P(a|c)-P^2(a|c)+[X_z^c(z)]^2=O(1)$. Indeed, it follows from Fact 1 of [6] that the differential equation like

$$f'(z)=b(z)-a(z)f(z) \tag{23}$$

satisfies

$$f(z)=f(z_0)e^{A(z_0)-A(z)}+\int_{z_0}^{z}b(x)e^{A(x)-A(z)}dx$$

where $A(z)=\int a(z)$ is the primitive function of $a(z)$. Setting in (23) $f(z)=V_c(z)$, $b(z)=V_a(P(a|c)z)+V_b(P(b|c)z)+g(z)$ and $a(z)=1$ we obtain (22).

Now we apply induction over the increasing domains. In short, we assume that for $z\in\mathcal{C}_k(\theta)$ we have $|V_c(z)|\le B_k|z|$ for some B_k. Using the induction of the increasing domains we prove, as in the Appendix of [6] that B_k are bounded. This completes the proof, after applying the depoissonization lemma of [7].

3.3 Proof of Theorem 1(iii): Central Limit Theorem

In this section we prove the CLT for T_n^c for any $c \in \mathcal{A}$. The proof is heavily based on our paper [8] to which we refer throughout this section. We should point out that in [8] we proved CLT for the path length in DST built over of memoryless sources. In this paper we extend the main technical part of [8] to Markov sources that hopefully will lead a the CLT for LZ'78.

Let us recall some notation. We define $D_n^c(u) = \mathbf{E}[u^{T_n^c}]$ as the probability generating function of T_n^c. To prove that T_n^c satisfies a central limit theorem, we use Levy's continuity theorem and show that for a complex τ we have

$$D_n^c\left(\exp\left(\frac{\tau}{\sqrt{\mathrm{Var}(T_n^c)}}\right)\right) e^{-\tau E(T_n^c)/\sqrt{\mathrm{Var}(T_n^c)}} \to e^{\tau^2/2}.$$

In order to accomplish it we first considering a Poisson version of $D_n^c(u)$, namely $D_c(z,u) = \sum_n D_n^c(u)\frac{z^n}{n!}e^{-z}$ that satisfies

$$\partial_z D_c(z,u) + D_c(z,u) = D_1^c(u) D_a(P(a|c)z,u) D_b(P(b|c)z,u) \tag{24}$$

where, for short, we write $D_1^c(u) = uP(a|c) + P(b|c)$.

To follow the footsteps of our proof from [8] we need to generalize deep technical result of [8], namely Theorem 10 (and its un-Poissonized version Theorem 6) to T_n^c with the Markov input. Once we prove below Theorem 2 we not only recover CLT but also large and moderate deviation for T_n^c. In this conference version we present our results for CLT.

For $\delta > 0$ we denote $|z^*|^{-\delta} = \max\{1, |z|^{-\delta}\}$ in order to avoid a singular behavior when $z \to 0$. Theorem below is equivalent to Theorem 10 of [8].

Theorem 2 *For all real number $\delta > 0$, there exists $0 < \theta < \pi/2$ and a complex neighborhood $\mathcal{U}(0)$ of 1 such that for $t \in \mathcal{U}(0)$ and $z \in \mathcal{C}(\theta)$ $\log(D_c(z, e^{t|z^*|^{-\delta}}))$ exists and $\log(D_c(z, e^{t|z^*|^{-\delta}}) = O(z)$ uniformly in $t \in \mathcal{U}(0)$.*

Proof: We notice that $D_c(z,1) = 1$ thus $\log D_c(z,1) = 0$ From now we will work with $\tilde{D}_c(z,u) = D_c(z,u)e^z$. As in [8] we define the kernel functions $f_c(z,u)$:

$$f_c(z,u) = \frac{\tilde{D}_c(z,u)}{\partial_z \tilde{D}_c(z,u)} = \frac{\tilde{D}_c(z,u)}{D_1^c(u)\tilde{D}_a\left(P(a|c)z,u\right)\tilde{D}_b\left(P(b|c)z,u\right)}.$$

Notice that $f_c(z,1) = 1$. We have $f_c(z,u) = \frac{1}{\partial_z \log \tilde{D}_c(z,u)}$. Clearly if $f_c(z,u)$ exists and has no roots in a given domain, then the function $\log \tilde{D}_c(z,u)$ exists in this domain. Again following [8] we are going to prove that for $u = e^{t|z|^{-\delta}}$ we have $f_c(z,u) = 1 + b(z,t)$ with $b(z,t) = O(|t|)$ when z is in a complex cone $\mathcal{C}(\theta)$. That way by selecting t sufficiently close to 0 (see (27) below) we will have $f_c(z,u) = \Theta(1)$. At the same time we prove that $\log \tilde{D}_c(z,u)$ exists and is $O(z)$ in $\mathcal{C}(\theta)$ since

$$\tilde{D}_c(z,u) = 1 + \int_0^z \frac{dx}{f_c(x,u)}$$

for $z \in \mathcal{C}(\theta)$ and the integration path in $\mathcal{C}(\theta)$.

In fact we will prove a slightly different result which indeed implies the main result. For an arbitrary number $\nu < 1$ and t complex we define the sequence $u_k = e^{\nu^k t}$ and the function $f_{c,k}(z) = f_c(z, u_k)$. To make the connection with the main result it suffice to set $\delta = -\frac{\log \nu}{\log \rho}$. We denote

$$f_{c,k}(z) = 1 + b_k(z,t), \qquad \frac{1}{f_{c,k}(z)} = 1 + a_k(z,t).$$

We will prove that for all integer k and for all $z \in \mathcal{C}_k(\theta)$ we uniformly have $a_k(z,t) = O(|t|)$ and $b_k(z,t) = O(|t|)$. We will prove this property by recursion on the increasing domains $\mathcal{C}_k(\theta)$. For $k = 0$ we already checked the proposition. In passing we notice that $a_{k+1}(z,t) = a_k(z, \nu t)$ and similarly $b_{k+1}(z,t) = b_k(z, \nu t)$. As in [8] we have the differential equation

$$\partial_z f_{c,k}(z) = 1 - f_{c,k}(z) g_{c,k}(z)$$

with

$$g_{c,k}(z) = \frac{P(a|c)}{f_{a,k}(P(a|c)z)} + \frac{P(b|c)}{f_{b,k}(P(b|c)z)}.$$

The resolution of the differential equation is (see also previous section)

$$f_{c,k}(z) = 1 + \int_0^z (1 - g_{c,k}(z)) \exp(G_{c,k}(z) - G_{c,k}(x)) dx$$

with the function $G_{c,k}(z)$ being a primitive of $g_{c,k}(z)$.

We notice now that when $z \in \mathcal{C}_k$, then $\forall (c,d)\ P(d|c)z \in \mathcal{C}_{k-1}(\theta)$, therefore we can apply the recurrence hypothesis. For this purpose we denote

$$a_k = \max_{z \in \mathcal{C}_k(\theta), c \in \mathcal{A}, t \in \mathcal{U}(0)} \left\{ \frac{|a_k(z,t)|}{|t|} \right\}.$$

We have

$$g_{c,k}(z) - 1 = P(a|c) a_{a,k-1}(P(a|c)z, \nu t) + P(b|c) a_{b,k-1}(P(b|c)z, \nu t)$$

and therefore $|g_{c,k}(z) - 1| \leq a_{k-1} \nu |t|$. Consequently

$$|f_{c,k}(z) - 1| \leq a_{k-1} \nu |t| \int_0^1 |z| \exp\left(\Re(G_{c,k}(z) - G_{c,k}(yz))\right) dy.$$

Now we observe that for any y

$$\begin{aligned} \Re\left(G_{c,k}(yz) - G_{c,k}(z)\right) &= -\Re(z)(1-y) + \int_y^1 \Re((g_{c,k}(xz) - 1)z) dx & (25) \\ &\leq (-\cos(\theta)|z| + \nu a_{k-1}|t|.|z|)(1-y). & (26) \end{aligned}$$

Consequently

$$b_k |t| \leq \nu a_{k-1} |t| \int_0^1 |z| e^{-(\cos\theta - \nu a_{k-1}|t|)|z|y} dy \leq \frac{\nu a_{k-1}|t|}{\cos\theta - \nu a_{k-1}|t|}.$$

As noticed in [8] the sequences b_k and a_k converges if $|t|$ is small enough and θ is selected such as $\nu/\cos\theta < 1$. □

The path from the previous result to the normal limiting distribution for the $D_c(z,u)$ and for the $D_n^c(u)$ is now via depoissonization as detailed in [8]. In short for all values of $\delta > 0$, in particular for the small values, we have $\log \tilde{D}_c(z, e^{|z|^{-\delta}t}) = O(z) = z + O(z|t|)$ for $z \to \infty$ which translates to

$$\log \tilde{D}_c(z, e^{|z|^{-\delta}t}) = z + T_c(z)|z|^{-\delta}t + Var_c(z)|z|^{-2\delta}t^2 + O(|z|^{1-3\delta}|t|^4).$$

The proof of Theorem 1(iii) is complete by setting as in [8]

$$t = \frac{\tau n^{\delta}}{\sqrt{\text{Var}T_n^c}} = O(n^{-1/2+\delta}) \to 0 \tag{27}$$

for $\delta < 1/2$.

4 Concluding Remarks

Our ultimate goal is to solve the fifty year open problem of a full characterization of the Lempel-Ziv'78 scheme under Markovian assumption. We claim that our Theorem 1 proved in this paper is a step in the direction of resolving this open problem. In this concluding remarks we provide some evidence justifying our confidence.

We recall that $D_n^c = P(T_n^c = k)$, and to make the notation under control we shall assume that the tail symbol involved in T_n^c could be an a or a b. It will be clear from the context. In Theorem 1 we presented our main results concerning D_n^c. Now, we show how it can be used to analyze LZ'78 scheme under Markovian assumption. It should be clear that to analyze LZ'78 with Markov input we need to study the number of phrases or blocks that start with a symbol $c \in \mathcal{A}$ and end with another symbol. However, it will be more convenient to consider the tail symbol, that is, the symbol that follows the last symbol of a phrase.

We shall prove the following claim.

Lemma 2 *Let $m_a + m_b = m$. Let $P^c_{m_a,m_b}$ the probability that m_a blocks/phrases among the first m start with symbol a while m_b start with symbol b assuming that the sequence starts with c. Then*

$$P^c_{m_a,m_b} \quad \leq \quad \sum_k D^a_{m_a,k} D^b_{m_b,m_a-1-k} + D^a_{m_a,k} D^b_{m_b,m_a-k},$$

or written differently $P^c_{m_a,m_b} \leq [u^{m_a}](1+u) D^a_{m_a}(u) D^b_{m_b}(u)$ where $D^c_n(u) = \mathbf{E}[u^{T^c_n}]$ is analyzed in Theorem 1.

Finally, we indicate a lower bound. Let $(c,d) \in \mathcal{A}^2$, we denote $D^{c,d}_{m,k}$ the probability that $m+1$ i.i.d Markovian sequences all starting with symbol c have k tail symbols equal to a and the last sequence has a tail symbol equal to d. The $D^{c,d}_{m,k}$ quantity have similar recursion as the $D^c_{m,k}$'s and we write $D^{c,d}_m(u) = \sum_k D^{c,d}_{m,k} u^k$.

Lemma 3 *We have $P^a_{m_a+1,m_b} \geq [u^{m_a}] D^{a,a}_{m_a}(u) D^b_{m_b}(u)$.*

This and Lemma 2 will allow us to bound P_{m_a,m_b} between two normal distributions of very similar mean and variance. Stay tuned!

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] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Second edition. John Wiley & Sons, New York, 2006.

[3] Fayolle, J., Ward, M. D. Analysis of the average depth in a suffix tree under a Markov model. In *International Conference on Analysis of Algorithms DMTCS*, proc. AD (Vol. 95, p. 104), 2005.

[4] P. Jacquet, P., M. Régnier, M. (1987). Normal limiting distribution of the size of tries. In Proceedings of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation (pp. 209-223). North-Holland Publishing Co..

[5] P. Jacquet, W. Szpankowski, Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*, 37(5), 1470-1475, 1991.

[6] P. Jacquet and W. Szpankowski, Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees, *Theoretical Computer Science*, 144, 161–197, 1995.

[7] P. Jacquet, W. Szpankowski, Analytical depoissonization and its applications. *Theoretical Computer Science*, 201(1), 1-62, 1998.

[8] P. Jacquet and W. Szpankowski, On the Limiting Distribution of Lempel Ziv'78 Redundancy for Memoryless Sources, *IEEE Trans. Information Theory*, 60, 6917-6930, 2014.

[9] P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, *Algorithmica*, 2002.

[10] P. Jacquet, W. Szpankowski, *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, Cambridge, 2015.

[11] D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

[12] K. Leckey, R. Neininger, and N. Wormald, Probabilistic Analysis of Lempel-Ziv Parsing for Markov Sources, preprint 2017.

[13] K. Leckey, R. Neininger and W. Szpankowski, Towards More Realistic Probabilistic Models for Data Structures: The External Path Length in Tries under the Markov Model, *SIAM-ACM Symposium on Discrete Algorithms* (SODA 2013), 877-886, New Orleans, 2013.

[14] G Louchard, W Szpankowski, On the average redundancy rate of the Lempel-Ziv code. *IEEE Transactions on Information Theory*, 43, 2–8, 1997.

[15] R. Neininger and L. Rüschendorf, A General Limit Theorem for Recursive Algorithms and Combinatorial Structures, *The Annals of Applied Probability*, 14, No. 1, 378-418, 2004.

[16] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. Information Theory*, 43, 9–21, 1997.

[17] R. Sedgewick, and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1995.

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

[19] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.

[20] J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Transactions on Information Theory*, 24, 530–536, 1978.