AVERAGE PROFILE AND LIMITING DISTRIBUTION FOR A PHRASE SIZE IN THE LEMPEL-ZIV PARSING ALGORITHM

October 13, 1994

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

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

Abstract

Consider the parsing algorithm due to Lempel and Ziv that partitions a sequence of length n into variable phrases (blocks) such that a new block is the shortest substring not seen in the past as a phrase. In practice the following parameters are of interest: number of phrases, the size of a phrase, the number of phrases of given size, and so forth. In this paper, we focus on the size of a *randomly* selected phrase, and the average number of phrases of a given size (the so called average *profile of phrase sizes*). These parameters can be efficiently analyzed through a digital search tree representation. For a memoryless source with *unequal* probabilities of symbols generation (the so called *asymmetric Bernoulli model*), we prove that the size of a typical phrase is asymptotically normally distributed with mean and the variance explicitly computed. In terms of digital search trees, we prove the normal limiting distribution of the typical depth (i.e., the length of a path from the root to a randomly selected node). The latter finding is proved by a technique that belongs to the toolkit of the "analytical analysis of algorithms", but which seems to be novel in the context of data compression.

Index Terms: Digital search trees, Lempel-Ziv parsing scheme, data compression, phrase length, typical depth in a digital tree, limiting distributions, average profile, Mellin transform, analytical analysis of algorithms.

*This research was primary done while the author was visiting INRIA in Rocquencourt, France. The author wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support. In addition, support was provided by NSF Grants NCR-9206315 and CCR-9201078 and INT-8912631, and from Grant AFOSR-90-0107, and in part by NATO Grant 0057/89.

1. INTRODUCTION

The heart of some universal data compression schemes is the parsing algorithm due to Lempel and Ziv [33]. It partitions a sequence into phrases (blocks) of variable sizes such that a new block is the shortest substring not seen in the past as a phrase. For example, the string 110010100010001000 is parsed into (1)(10)(0)(101)(00)(01)(000)(100). There is another possibility of parsing, as already noticed in [32], and explored by Grassberger [9] and Szpankowski [26], that allows overlapping in the course of creating the partition. For example, for the above sequence the latter parsing leads to (1)(10)(0)(101)(00)(01)(000100). In this paper, we only consider the former parsing algorithm.

These parsing algorithms play a crucial rôle in universal data compression schemes and theirs numerous applications such as efficient transmission of data (cf. [29, 32, 33]), discriminating between information sources (cf. [8], [31]), test of randomness (cf. [31]), estimating the statistical model of individual sequences (cf. [30], [31]), and so forth. The parameters of interest to these applications are: the number of phrases, the number of phrases of a given size, the size of a phrase, the length of a sequence built from a given number of phrases, etc. Some of these parameters have been studied in the past as the first-order properties, that is, typical behaviors in the almost sure sense. Very few results – with a notable exception of the paper by Aldous and Shields [1] – are available up-to-date concerning second order properties such as limiting distributions, large deviation results, concentration of mean, etc.

Recently, Gilbert and Kadota [8] have presented convincing arguments for the need of such investigations. The authors of [8] used numerical evaluations to obtain qualitative insights into some second-order behaviors of the Lempel-Ziv parsing algorithm. In particular, they studied the length of a sequence obtained from the first m phrases, and the length of the mth phrase. In this paper, among others, we provide for memoryless sources (the so called *Bernoulli model*) the limiting distribution for the latter quantity. We obtain these results by transforming the problem into another one on *digital trees* (cf. [1], [17]). In passing, we observe that digital trees have been studied in their own right for more than twenty years (cf. [17, 18]).

We consider a special type of digital trees, namely a *digital search tree* (cf. [5], [7], [17], [18]). This tree is constructed as follows (see also Figure 1). We consider m, possible infinite, strings of symbols from a finite alphabet Σ (however, for the simplicity of presentation we further work only with the binary alphabet $\Sigma = \{0, 1\}$). The first string is stored in the root, while the second string occupies the right or the left child of the root depending whether

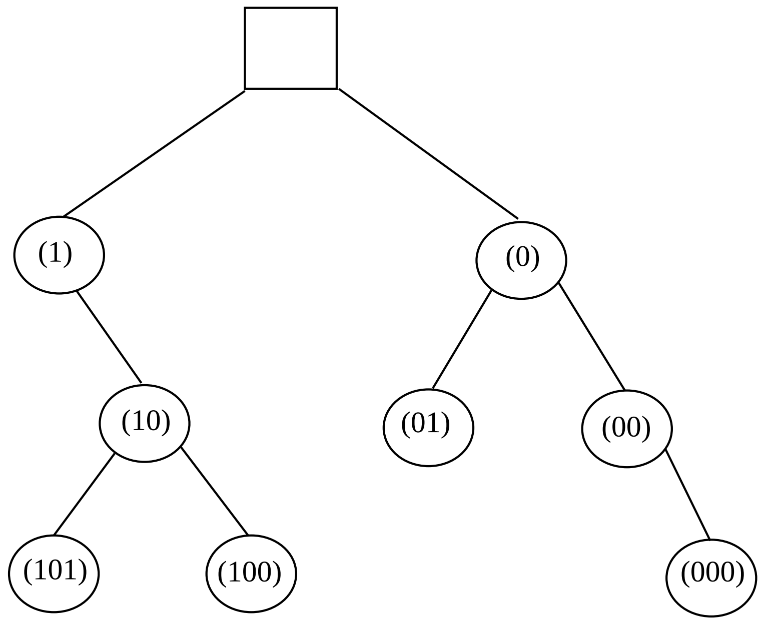

Figure 1: A digital tree representation of Ziv's parsing for the string 11001010001000100...

its first symbol is "1" or "0". The remaining strings are stored in available nodes (that are directly attached to nodes already existing in the tree). The search for an available node follows the prefix structure of a string. The rule is simple: if the next symbol in a string is "1" we move to the right, otherwise move to the left. The resulting tree has m internal nodes. The details can be found in [17] and [21].

The Lempel-Ziv parsing algorithm can be efficiently implemented using the digital search tree structure. We assume that the first phrase of the Lempel-Ziv scheme is an *empty* phrase. We store it in the root of a digital search tree, and all other phrases are stored in internal nodes. When a new phrase is created, the search starts at the root and proceeds down the tree as directed by the input symbols exactly in the same manner as in the digital search tree construction. For example, for the binary alphabet, "0" in the input string means move to the left and "1" means proceed to the right. The search is completed when a branch is taken from an existing tree node to a new node that has not been visited before. Then, an edge and a new node are added to the tree. Phrases created in such a way are stored directly in nodes of the tree (cf. Figure 1). In passing, we note that the second parsing algorithm discussed above (with overlapping between phrases) leads to another digital tree called the suffix tree (cf. [9], [26], [28]).

We consider the Lempel-Ziv algorithm in a probabilistic framework. We assume that a string of length n is generated according to the Bernoulli model. That is: *symbols are* *generated in an independent manner with "0" and "1" occurring respectively with probability p and $q = 1 - p$.* If $p = q = 0.5$, the the Bernoulli model is called *symmetric*, otherwise it is *asymmetric.*

The Lempel-Ziv algorithm can be analyzed in two different frameworks. Namely, with a *fixed number* m of parsed words or with a *fixed length* n of a sequence to be parsed. The former model falls exactly under the digital search tree framework with independent strings, and we further call it the *digital tree model*, as it is known for last twenty years [17, 18]. Therefore, any new result in this endouver will lead to a new finding in the area of digital search trees, and reverse: we can apply many known results of such trees (for a survey see Mahmoud [21]) to our problem. The latter problem is harder since by fixing the length of a string we introduce some dependency among phrases (even if they still do *not* overlap!). Nevertheless, this dependency is not strong enough to spoil the analysis, and we shall prove that the digital search tree results can be extended to this new situation. We coin a name *Lempel-Ziv model* for the latter framework.

Hereafter, we stick to some notation that we shall use throughout the paper. We always denote by n the length of a single string that is parsed (i.e., Lempel-Ziv model), while m is always the number of *independent* strings used to built a digital tree (i.e., digital tree model) or the number of parsed words used to construct a single string (of a random length!).

In this paper we report two main findings, namely: for both models we prove that the length of a randomly selected phrase (and the average number of phrases of a given size) in the asymmetric model is normally distributed around its mean with the variance of order $\Theta(\log n)$. We treat separately the symmetric Bernoulli model since the variance in this case is $O(1)$, and actually the limiting distribution does not exist in this case. However, we show that the limiting distribution centered around $\log_2 n$ resembles the double exponential distribution (i.e., $e^{-e^{-x}}$).

Digital trees, that is, *tries*, compact tries known also as *Patricia tries*, and *digital search trees* have been extensively analyzed in the past in the case of a *fixed* number of independent strings (cf. [5, 7, 12, 15, 16, 17, 18, 19, 23, 27]), and in some cases with dependent strings (cf. analysis of suffix tree in [27, 28]). In particular, the average length of the internal path length (i.e., the sum of all depths) and the average size of a digital search tree in the symmetric model was analyzed by Konheim and Newman [18], Knuth [17], Flajolet and Sedgewick [5], and Flajolet and Richmond [7]. The average depth and the variance of the depth for the asymmetric Bernoulli model is given in Szpankowski [27] (the symmetric case was also analyzed in Kirschenhofer and Prodinger [14]), while the variance of the internal path length in the symmetric Bernoulli model was investigated in Kirschenhofer *et al.* [16]. Finally, Louchard [19], and Aldous and Shields [1] for symmetric Bernoulli alphabet obtained the limiting distribution of the depth. As mentioned above, in this paper we directly extend Louchard's result to asymmetric Bernoulli model, while in another paper Jacquet and Szpankowski [13] generalize some of the Aldous and Shields [1] results concerning the limiting distribution of the internal path length.

As mentioned above, for the Lempel-Ziv parsing algorithm mostly only first-order properties have been investigated, with an exception of the work by Aldous and Shields [1]. It is well known that for a stationary and ergodic source the number of phrases is almost surely equal to $(nh/\log n)$ where h is the entropy of the alphabet. For the symmetric Bernoulli alphabet Aldous and Shields [1] proved that the number of phrases is normally distributed with mean $n/\log_2 n$ and variance $\Theta(n/\log_2^3 n)$ (for the coefficient at $n/\log_2^3 n$ in the variance see [13], [16]). The first-order property of the phase length in the Lempel-Ziv parsing algorithm was recently reported by Ornstein and Weiss [22]. Finally, Gilbert and Kadota [8] analyzed numerically the number of possible messages composed of m parsed phrases, as well as the length of a phrase in the digital tree model (see [13] for some theoretical solutions to these problems).

The paper is organized as follows. In the next section, we formulate our main results and present some consequences of them. The proof concerning the limiting distribution of the depth in a digital tree model is presented in Section 3.1, while the Lempel-Ziv model is analyzed in Section 3.2

2. MAIN RESULTS

Let us first consider the **digital tree model** in which the number of parsed words is fixed and equal to m. These words are statistically independent and satisfy the Bernoulli model. We construct a digital search tree from these m strings or alternatively we build a sequence (of random length) according to the Lempel-Ziv scheme. Then, the length of a randomly selected phrase in the Lempel-Ziv sequence composed of m phrases is the same as the length of a randomly selected depth (i.e., the path from the root to a node) in the associated digital tree. Traditionally, in the area of digital trees this depth is denoted as D_m, and we shall adopt this notation. Let also $D_m(i)$ be the depth of the ith node in the associated digital tree. Actually, observe that $D_m(i) = D_i(i)$ for $m \geq i$. Clearly, for various $i \leq m$ distributions of $D_m(i)$ are different, and therefore it makes sense to define a *typical depth* D_m as

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

Furthermore, we denote by L_m the internal path length of the digital search tree, that is, $L_m := \sum_{i=1}^{m} D_m(i)$. Note that L_m is the length of a sequence generated by the Lempel-Ziv parsing scheme from these m parsed words (i.e., in the digital tree model). Finally, we denote by $B_m(k)$ the number of nodes in the digital search tree at level k. Clearly, it is equal to the number of phrases of length k in the Lempel-Ziv scheme in the digital tree model.

The situation is similar, but *not* the same in the **Lempel-Ziv model** in which a sequence of a fixed length n is parsed into phrases. Let M_n and $M_n(k)$ denote the number of phrases and the number of phrases of size k, respectively, produced by the algorithm. Let also $D_n^{LZ}(i)$ be the length of the ith phrase in the Lempel-Ziv model, where $1 \leq i \leq M_n$. By the *typical phrase length* $D_{M_n}^{LZ}$ or shortly D_n^{LZ} we denote the length of a randomly selected phrase. The typical depth D_n^{LZ} in the Lempel-Ziv model can be estimated as follows

$$\Pr\{D_n^{LZ} = k\} = \sum_{m=m_L}^{m_U} \Pr\{D_n^{LZ} = k | M_n = m\} \Pr\{M_n = m\} \tag{2}$$

where m_L and m_U are lower and the upper bounds for the number of phrases M_n. One easily proves that for some constants α_1 and α_2

$$m_L := \alpha_1 \sqrt{n} \leq M_n \leq \alpha_2 n / \log_2 n =: m_U \ . \tag{3}$$

Indeed, the minimum number of phrases occurs only for two strings: either all zeros or all ones, and then $n = \sum_{i=1}^{M_n} D_n(i) = M_n(M_n+1)/2$, hence the lower bound $m_L = \Theta(\sqrt{n})$ follows. For the upper bound, we consider a complete binary tree with the internal path length equal to n. Thus, $n \geq \sum_{i=1}^{\log_2 M_n - 1} i2^i \geq (\log_2 M_n - 2)M_n$, and the upper bound $m_U = O(n/\log_2 n)$ follows.

According to (2), one needs to estimate the conditional probability $\Pr\{D_n^{LZ} = k | M_n = m\}$ in order to assess the distribution of D_n^{LZ}. It is tempting to assume that $\Pr\{D_n^{LZ} = k | M_n = m\} = \Pr\{D_m = k\}$ where the right-hand side of this equation refers to the depth in the digital tree model. But, this is *untrue* due to the fact that in the Lempel-Ziv model we consider *only* those digital search trees whose internal path length is fixed and equal to n. Clearly, this restriction affects the depth of a randomly selected node (think of a digital tree built from the string $11111\ldots111$ which is very skewed). Fortunately, we shall prove in Section 3.2 that $\Pr\{D_n^{LZ} = k | M_n = m\} = (1 + O(\sqrt{\log n / n}))\Pr\{D_m = k\}$.

We now present results for the **digital tree model**, and let $\overline{B}_m(k) := EB_m(k)$ be the average number of internal nodes at level k in a digital tree built over m independent strings. As in Knuth [17] (cf. also [26]), we have the following relationship between the depth D_m and the average profile $\overline{B}_m(k)$

$$\Pr\{D_m = k\} = \frac{\overline{B}_m(k)}{m} . \tag{4}$$

This follows from the definition (1) of D_m and the definition of $\overline{B}_m(k)$.

We shall work initially with the average profile, and we define the generating function $B_m(u) = \sum_{k=0}^{\infty} \overline{B}_m(k) u^k$ which satisfies the following recurrence (cf. [17], [26])

$$B_{m+1}(u) = 1 + u \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} (B_j(u) + B_{m-j}(u)) \tag{5}$$

with $B_0(u) = 0$. This recurrence arises naturally in our setting by considering the left and the right subtrees of the root.

A general recurrence of the above type was analyzed in Szpankowski [26] (cf. see also Flajolet and Richmond [7] for an interesting extension). A slight modification of Theorem 2.4 in [26] directly leads to the exact solution of (5), namely:

$$B_m(u) = m - (1-u) \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u) \tag{6}$$

where

$$Q_k(u) = \prod_{j=2}^{k+1} (1 - up^j - uq^j) \quad , \quad Q_0(u) = 1 . \tag{7}$$

Actually, the derivation of (6) is not too complicated, so we provide a sketch of the proof. Let us start with multiplying both sides of (6) by $z^m/m!$ to get $B'_z(z,u) = e^z + uB(pz,u)e^{qz} + uB(qz,u)e^{pz}$ where $B(z,u) = \sum_{m=0}^{\infty} B_m(u) \frac{z^m}{m!}$, and $B'_z(z,u)$ is the derivative of $B(z,u)$ with respect to z. We now multiply this functional equation by e^{-z} and introduce $\widetilde{B}(z,u) = B(z,u)e^{-z}$. This leads to a new functional equation, namely $\widetilde{B}'(z,u) + \widetilde{B}(z,u) = 1 + u(\widetilde{B}(zp,u) + \widetilde{B}(zq,u))$. Comparing now the coefficients at z^m one immediately obtains $\widetilde{B}_{m+1}(u) = \delta_{m,0} - \widetilde{B}_m(u)(1 - up^m - uq^m)$ where $\delta_{0,m}$ is the Kronecker symbol. To prove (6) it only suffices to note that $B_m(u) = \sum_{k=0}^{m} \binom{m}{k} \widetilde{B}_k(u)$.

We consider the symmetric and the asymmetric cases separately. For the symmetric model, we exactly compute the coefficients at u^k of $B_m(u)$ directly from (6). For the asymmetric model, we use Goncharov's theorem (cf. [17]) applied to the probability generating function $D_m(u) = B_m(u)/m$ to establish normal limiting distribution of D_m (for details see Section 3.1). In the latter case we need one more result from [26] that is provided below for the reader's convenience.

Fact 1. (i) *The average ED_m of the depth becomes as $m \to \infty$*

$$ED_m = \frac{1}{h}\left(\log m + \gamma - 1 + \frac{h_2}{2h} + \theta + \delta(m)\right) + O(\log m/m) \tag{8}$$

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

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

The function $\delta(x)$ is a fluctuating function with a small amplitude when $\log p/\log q$ is rational, and $\delta(x) \equiv 0$ for $\log p/\log q$ irrational. More precisely, for $\log p/\log q = r/t$ where r, t are integers,

$$\delta_1(x) = \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \frac{\Gamma(s_0^\ell)Q(-2)}{Q(s_0^\ell - 1)} \exp\left(-\frac{2\pi i\ell r}{\log p}\log x\right) , \tag{9}$$

where $s_0^\ell = -1 + 2\pi i\ell r/\log p$.

(ii) *The variance of D_m for large m satisfies*

$$var\ D_m = \frac{h_2 - h^2}{h^3}\log m + A + \Delta(m) + O(\log^2 m/m) \tag{10}$$

where A is a constant and $\Delta(x)$ is a fluctuating function with a small amplitude. In the symmetric case, the coefficient at $\log m$ becomes zero, and then (cf. [16])

$$var\ D_m = \frac{1}{12} + \frac{1}{\log^2 2}\cdot\frac{\pi^2}{6} - \alpha - \beta + \Delta(m) + O(\log^2 m/m) \tag{11}$$

where

$$\alpha = \sum_{j=1}^{\infty}\frac{1}{2^j - 1} \quad , \quad \beta = \sum_{j=1}^{\infty}\frac{1}{(2^j-1)^2}$$

and the function $\Delta(x)$ is continuous with period 1 and mean zero. ■

In Section 3.1 we prove our first main result concerning the limiting distribution of D_m (hence, also for the average profile $\overline{B}_m(k)$) in the digital tree model.

Theorem 1. (i) Symmetric Case. *Let $Q_k = \prod_{j=1}^{k}(1-2^{-j})$, and define $\psi(m) = \log_2 m - \lfloor\log_2 m\rfloor$. Then, for the symmetric Bernoulli we obtain for any integer K*

$$\lim_{m\to\infty} \Big|\Pr\{D_m \le \log_2 m + K\} - 2^{K-\psi(m)}\left(1 + \frac{1}{2Q_\infty}\sum_{i=0}^{\infty}(-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i}e^{-2^{-(K-\psi(m)-1-i)}}\right)\Big| = 0 . \tag{12}$$

The function $\psi(m)$ is dense in $[0,1]$ but not uniformly dense, thus the limiting distribution of D_m does not exist (see Remark 1(i) below).

(ii) Asymmetric Case. *In the asymmetric case, the limiting distribution of D_m is normal, that is,*

$$\frac{D_m - ED_m}{\sqrt{\mathrm{Var}\ D_m}} \to N(0,1) \tag{13}$$

where ED_m and Var D_m *are given by (8) and (10), respectively, and the moments of D_m converge to the appropriate moments of the normal distribution. More generally, for any complex ϑ such that $\Re(\vartheta) > 0$*

$$e^{-\vartheta c_1 \log m} E(e^{\vartheta D_m}) = e^{c_2 \frac{\vartheta^2}{2} \log m} \left(1 + O(\frac{1}{\sqrt{\log m}})\right) \tag{14}$$

where $c_1 = 1/h$ and $c_2 = (h_2 - h^2)/h^3$.

(iii) *In the asymmetric case, there exist positive constants A and $\alpha < 1$ such that*

$$\Pr\left\{\left|\frac{D_m - c_1 \log m}{\sqrt{c_2 \log m}}\right| > k\right\} \le A\alpha^k \tag{15}$$

uniformly in k for large m. ■

Remark 1. (i) The limiting distribution for the symmetric case was obtained before by Louchard [19] by a different method than the one presented in Section 3.1. Actually, we can write (12) alternatively as

$$\lim_{m\to\infty} \sup_x \left| \Pr\{D_m \le x\} - \frac{2^x}{m}\left(1 + \frac{1}{2Q_\infty}\sum_{i=0}^{\infty}(-1)^{i+1}\frac{2^{-i(i+3)/2}}{Q_i}\exp(-m2^{-(x-1-i)})\right)\right| = 0 \tag{16}$$

where x is any real number. Moreover, in the symmetric model we can, following Louchard, also give exact distribution of the depth, that is,

$$\Pr\{D_m \le j+1\} = \frac{1}{m}\left(2^{j+1} - 1 + \sum_{k=1}^{j} 2^k \frac{(-1)^{j-k+1} 2^{(j-k)(j-k+1)/2}}{Q_{j-k}Q_{k-1}}(1-2^{-k})^{m-1}\right) \tag{17}$$

for all integers $j \ge 1$.

(ii) One may wonder why the limiting distribution in the symmetric case is not normal, and actually what kind of "known" distribution it resembles. First of all, the central limit theorem holds in the asymmetric case since by definition (1) the indicator function of the typical depth could be viewed as the average sum of indicator functions of all depths. Thus, after proper normalization (i.e., $\sqrt{\mathrm{Var} D_m}$) one may expect normal distribution of D_m *provided* $\mathrm{Var} D_m \to \infty$ as $m \to \infty$. This holds in the asymmetric case (indeed, $\mathrm{Var} D_m = O(\log m)$) but not in the symmetric model where $\mathrm{Var} D_m = O(1)$. To predict the behavior of the limiting distribution for the symmetric model, one should have a closer look at the definition of $D_m(i)$. Let C_{ij} be the length of the longest common prefix of the ith and jth strings. Then,

$$D_m(i) = \max\{\min\{C_{i1}, D_m(1)\}, \ldots, \min\{C_{i,i-1}, D_m(i-1)\}\} .$$

It suggests that the depth $D_m(m)$ is a maximum over $(m-1)$ *dependent* random variables. If those random variables would be independent, one should expect "double exponential" (i.e., $e^{-e^{-x}}$) limiting distribution for $D_m(m)$. Actually, the limiting distribution of D_m is a combination of double-exponential distributions as can be verified by inspecting carefully formula (12).

(iii) We observe that the large deviation result (iii) of Theorem 1 is a direct consequence of (14) from Theorem 1(ii). Clearly, it follows from the Markov inequality along the same lines as in Flajolet and Soria [6]. □

Now, we turn our attention to the **Lempel-Ziv model**. Before we present our main finding, we review some known results for the number of phrases M_n, which we further need to analyze the depth D_n^{LZ}.

Fact 2. (i) (Aldous and Shields [1]) *In the* **symmetric** *Bernoulli model*

$$\frac{M_n - EM_n}{\sqrt{\mathrm{Var} M_n}} \to N(0,1) \tag{18}$$

where $N(0,1)$ *denotes the standard normal distribution, with* $EM_n \sim n/\log_2 n$ *and* $\mathrm{Var} M_n \sim \Theta(n/\log_2^3 n)$.

(ii) (Jacquet and Szpankowski [13]) *In the* **asymmetric** *Bernoulli model*

$$\frac{M_n - EM_n}{\sqrt{\mathrm{Var} M_n}} \to N(0,1) \tag{19}$$

where $EM_n \sim nh/\log n$ *and* $\mathrm{Var} M_n \sim c_2 h^3 n/\log^2 n$. *Moreover, all moments of* M_n *converge to the appropriate moments of the normal distribution.* ■

Remark 2. Actually, using Kirschenhofer *et al.* [16] and the approach from Jacquet and Szpankowski [13] one can estimate the coefficient at $n/\log_2^3 n$ in the variance of M_n in the symmetric case. After some algebra, we derives (cf. [16])

$$\mathrm{Var} M_n \sim (C + \delta(\log_2 n))\frac{n}{\log_2^3 n} \tag{20}$$

where $\delta(x)$ is a fluctuating continuous function with period 1, mean zero, and amplitude smaller than 10^{-6}. The constant C has an explicit, but complicated formula as derived in [16], and its numerical value is $C = 0.26600...$ with all five digits significant. □

We are now ready to present our result concerning the Lempel-Ziv model. The proof can be found in Section 3.2.

Theorem 2. Symmetric Case. (i) *Let* $\psi_1(n) = \log_2(n/\log_2 n) - \lfloor \log_2(n/\log_2 n) \rfloor$, *and let* K *be an integer. Then, the asymptotic distribution for the length of a randomly selected phrase for the symmetric Bernoulli model becomes*

$$\lim_{m\to\infty} \left| \Pr\{D_n^{LZ} \le \log_2(n/\log_2 n) + K\} - 2^{K-\psi_1(n)} \left(1 + \frac{1}{2Q_\infty} \sum_{i=0}^{\infty} (-1)^{i+1} \frac{2^{-i(i+1)/2}}{Q_i} e^{-2^{-(K-\psi_1(n)-1-i)}}\right) \right| = 0 \ . \tag{21}$$

The function $\psi_1(n)$ *is dense in* $[0,1]$ *but not uniform dense, hence the limiting distribution of* D_n^{LZ} *does not exist.*

(ii) Asymmetric Model. *For the asymmetric Bernoulli model the typical depth* D_n^{LZ} *is normally distributed, i.e.,*

$$\frac{D_n^{LZ} - c_1 \log(nh/\log n)}{\sqrt{c_2 \log(nh/\log n)}} \to N(0,1) \ . \tag{22}$$

More precisely, for some complex ϑ *with* $\Re(v) > 0$

$$e^{-\vartheta c_1 \sqrt{\log(nh/\log n)}} E\left(e^{\vartheta D_n^{LZ}/\sqrt{\log(nh/\log n)}}\right) = e^{c_2 \vartheta^2/2}(1 + O(1/\sqrt{\log n})) \ . \tag{23}$$

Furthermore, the above implies the existence of two positive constants A *and* $\alpha < 1$ *such that*

$$\Pr\left\{ \left| \frac{D_m^{LZ} - c_1 \log(nh/\log n)}{\sqrt{c_2 \log(nh/\log n)}} \right| > k \right\} \le A\alpha^k \tag{24}$$

uniformly in k *for large* m. ■

Remark 3. (i) *Markovian Model.* It is plausible that our analysis can be extended to Markovian model in which the next symbol in a sequence depends on a finite number of previous ones. Such an extension was already obtained for the depth D_m in another digital tree, namely trie (cf. Jacquet and Szpankowski [12]).

(ii) *Almost Sure Behaviors.* Surprisingly enough, the almost sure behavior of D_m and D_n^{LZ} are not implied by Theorems 1 and 2. In fact, $D_m/\log m$ and $D_n^{LZ}/\log n$ do *not* converge almost surely. The same applies to the length of the last phrase, or the depth of insertion, which we denote as ℓ_m. Indeed, this is a consequence of the profound results of Pittel [23] concerning digital trees. He proved, among other things, that $\ell_m/\log m$ converges *in probability* to $1/h$, but does *not* converge almost surely. Let $p_{\min} = \min\{p,q\}$ and $p_{\max} = \max\{p,q\}$. Then,

$$\liminf_{m\to\infty} \frac{\ell_m}{\log m} = \frac{-1}{\log p_{\min}} \qquad (a.s.) \qquad \limsup_{m\to\infty} \frac{\ell_m}{\log m} = \frac{-1}{\log p_{\max}} \ . \tag{25}$$

The same is true for D_m and D_n^{LZ} (cf. [13, 27, 28]).

(iii) *Average Profile.* The average profile $\overline{B}_m(k)$ directly follows from Theorem 1 and (4). The limiting distribution of the profile $B_m(k)$ is harder to obtain. Aldous and Shields [1] established it for the symmetric case. In the asymmetric case the limiting distribution is unknown. For the digital tree model it is easy to establish a recurrence for $B_m(k)$. Define $B_m^k(u) = Eu^{B_m(k)}$. Then (cf. [13])

$$B_{m+1}^k(u) = \sum_{l=0}^{m} \binom{m}{l} p^l q^{m-l} B_l^{k-1}(u) B_{m-l}^{k-1}(u) \ ,$$

with $B_0^0(u) = 1$. Let now $B^k(z,u) = \sum_{m=0}^{\infty} B_m^k(u) \frac{z^m}{m!}$. Then, the above becomes

$$\frac{\partial B^k(z,u)}{\partial z} = B^{k-1}(pz,u) B^{k-1}(qz,u)$$

with $B^0(z,u) = u(e^z - 1) + 1$. We conjecture that for $k = O(\log m)$ the limiting distribution of B_m^k is normal with mean $\overline{B}_m(k)$ established in Theorem 1.

(iv) *Extensions and Open Problems.* One may consider an extension of digital search trees called b-digital search trees, and its corresponding Lempel-Ziv parsing scheme. In a b-digital search tree every node can store up to b strings [7, 21] (with possible exception of the root). Based on this generalization, one can extend the Lempel-Ziv parsing scheme as follows: We postulate that the next phrase in the generalized Lempel-Ziv algorithm is the longest phrase seen in the past by *at most* $b-1$ phrases. For example, the sequence from Figure 1 is parsed as follows: (1)(1)(0)(0)(10)(10)(00)(100)(01)(00). Note that the number of *distinct* phrases (the ones that count in the possible extension of the data compression scheme) is equal to six compared to eight in the original Lempel-Ziv parsing scheme. What is the length of a randomly selected phrase in such a generalization? What is the distribution of the number of phrases? Etc. Those and other questions possibly can be answered if one solves the b-digital search tree model as we did in this paper for $b = 1$. As pointed out in Flajolet and Richmond [7] (cf. also [17]) the analysis of b-digital search trees is not that simple. Indeed, our basic recurrence equation (5) becomes now for $m \geq 0$

$$B_{m+b}(u) = b + u \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} (B_j(u) + B_{m-j}(u)) \tag{26}$$

with $B_i(u) = i$ for $i \leq b$. As in the case of $b = 1$, to solve the above we introduce the exponential generating function $B(z,u) = \sum_{m=0}^{\infty} B_m(u) \frac{z^m}{m!}$ that satisfies the following differential-functional equation

$$\frac{\partial^b B(z,u)}{\partial z^b} = a(z) + uB(pz,u)e^{qz} + uB(zq,u)e^{pz}$$

where $a(z)$ is a poly-exponential function. This functional equation and the recurrence (26) do not have closed-form solutions as in the case of $b = 1$. Thus, the general solution from [26] cannot be applied. Another approach is needed, and possibly the one suggested by Flajolet and Richmond [7] should lead to a solution. We address this problem in a forthcoming paper. □

3. ANALYSIS

In this section we prove Theorem 1 (cf. Section 3.1) and Theorem 2 (cf. Section 3.2). Those proofs, as it turns out, require quite different approaches, and they might be useful in the analysis of other problems on data compression.

3.1 Digital Search Tree Model

We study a digital search tree built from m independent strings generated according to the Bernoulli model. We consider separately the symmetric model and the asymmetric one since they require quite different techniques.

A. Symmetric Bernoulli Model

We pick our analysis where we left it in Section 2, that is, from recurrence (5) which has solution (6), that is,

$$B_m(u) = m - (1-u)\sum_{k=2}^{m}(-1)^k\binom{m}{k}Q_{k-2}(u) \tag{27}$$

where

$$Q_k(u) = \prod_{j=1}^{k}(1-u2^{-j}) \tag{28}$$

with $Q_0(u) = 1$. Since the formula for $Q_k(u)$ is relatively simple, we can extract coefficients of $B_m(u)$ "by hand".

Note that $Q_k(u) = Q_\infty(u)/Q_\infty(u2^{-k})$, and by Euler's identities (cf. [17]) as in Louchard [19],

$$\frac{1}{Q_\infty(u)} = \sum_{i=0}^{\infty}\frac{u^i}{2^iQ_i} \quad , \quad Q_\infty(u) = -\sum_{i=0}^{\infty}u^iR_i \tag{29}$$

where

$$R_i = (-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i} \tag{30}$$

with $Q_i = Q_i(1)$. Let now $[u^k]f(u)$ denote the coefficient at u^k of $f(u)$. Then,

$$[u^n]Q_{k-2}(u) = -\sum_{l=0}^{n}\frac{R_{n-l}}{Q_l2^{l(k-1)}} \ .$$

Hence, applying this to our basic solution (27) we obtain

$$
\begin{aligned}
[u^{j+1}]B_m(u) &= \sum_{l=0}^{j+1} \frac{2^l R_{j-l+1}}{Q_l} \Big((1-2^{-l})^m - 1 - m/2\Big) \\
&- \sum_{l=0}^{j} \frac{2^l R_{j-l}}{Q_l} \Big((1-2^{-l})^m - 1 - m/2\Big) \ .
\end{aligned}
$$

Finally, after some tedious algebra one obtains (17) as in Louchard [19], and taking $m \to \infty$ we easily derive part (i) of Theorem 1 (see also Mahmoud [21], Ex. 6.12).

B. Asymmetric Bernoulli Model

In this case, we rather work with the probability generating function $D_m(u)$ for the depth which is equal to $B_m(u)/m$, that is,

$$
D_m(u) = 1 - \frac{1-u}{m} \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u) \ . \tag{31}
$$

Let $\mu_m = ED_m$ and $\sigma_m^2 = \mathrm{Var}D_m$. Fact 1 implies $\mu_m \sim c_1 \log m$ and $\sigma_m^2 \sim c_2 \log m$ where $c_1 = 1/h$ and $c_2 = (h_2 - h^2)/h^3$. We use Goncharov's theorem to establish the normal distribution of D_m by showing that

$$
\lim_{m\to\infty} e^{-\vartheta\mu_m/\sigma_m} D_m(e^{\vartheta/\sigma_m}) = e^{\vartheta^2/2} \tag{32}
$$

where $\vartheta = ix$ for imaginary i. However, below we prove a stronger result, namely we show that (32) holds for *any* complex ϑ (with $\Re(\vartheta) > 0$), and hence this will automatically establish convergence of moments (since every analytical function has its derivatives).

We now derive an asymptotic expansion for the probability generating function $D(u)$ around $u = 1$. We assume $u = e^v$, and due to $\sigma_m = O(\sqrt{\log m})$, we define $v = \vartheta/\sigma_m \to 0$. Hereafter, we use the complex variable v that tends to zero as $m \to \infty$.

Note that $1 - D_m(u)$ given in (31) has the form of an alternating sum. Such a sum can be handled either by Rice's method (cf. [5]) or by the Mellin-like approach (cf. [17], [25]). The Mellin-like approach is recalled below for the reader convenience.

Lemma 3. (Szpankowski [25]). *Let f_k be any sequence such that it has an analytical continuation $f(s)$ (i.e., $f(k) = f_k$) in the complex plane right to the line $(-\frac{3}{2} - i\infty, -\frac{3}{2} + i\infty)$ such that $f(s)$ does not grow faster than exponential for large s (for details see [25]). Then*

$$
\sum_{k=2}^{m} (-1)^k \binom{m}{k} f_k = \frac{1}{2\pi i} \int_{-3/2-i\infty}^{-3/2+i\infty} \Gamma(s) f(-s) m^{-s} ds + e_m \tag{33}
$$

where $\Gamma(s)$ is the gamma function, and the error term e_m is of order magnitude smaller than the leading term as shown in (35).

Proof. For completeness, we present a sketch of the proof. Details can be found in [25]. Let

$$S_m = \sum_{k=2}^{m} (-1)^k \binom{m}{k} f_k \ ,$$

and define $[m;s] = \Gamma(m+1)/\Gamma(m+1+s)$. We shall first prove that

$$S_m = \frac{1}{2\pi i} \int_{-3/2-i\infty}^{-3/2+i\infty} \Gamma(s) f(-s)[m;s] ds \ . \tag{34}$$

To evaluate the above integral, we use the Cauchy residue theorem [10]. Consider a large rectangle $R_{\alpha,\beta}$ with corners at $(\frac{1}{2} - \beta \pm i\alpha)$ and $(-\frac{3}{2} \pm i\alpha)$ left to the line of integration $(-\frac{3}{2} - i\infty, -\frac{3}{2} + i\infty)$. Then, the integral in (34) is the sum of of residues in $R_{\alpha,\beta}$ minus the integrals on the bottom, top and left lines of $R_{\alpha,\beta}$. Since $f(\cdot)$ cannot grow faster than any exponential function, we can prove that the integrals on the bottom, top and left lines of the rectangle vanish as $\alpha, \beta \to \infty$. Thus, we must estimate the residues to the left of the line $(-\frac{3}{2} - i\infty, -\frac{3}{2} + i\infty)$. But, $\Gamma(s)$ has singularities with residue of value $(-1)^k/k!$ at $s = -k$, where k is an integer ($k \geq 2$), and $[m;-k] = m!/(m-k)!$, thus by the residue theorem we immediately prove (34). To establish (33) we observe that by Stirling's formula $[m;s] = m^{-s}(1 + sO(1/m))$ (cf. [10]). Then,

$$e_m = O(1/m) \int_{-3/2-i\infty}^{-3/2+i\infty} s\Gamma(s) m^{-s} f(-s) ds \ , \tag{35}$$

as desired. ■

We use Lemma 3 to obtain precise asymptotics of $D(u)$. (In fact, we can use it to re-derive the average ED_m and the variance $\mathrm{Var}D_m$ of D_m given in Fact 1.) To do so, however, we need an analytical continuation of $Q_k(u)$. Denote it as $Q(u,s)$, and observe that (cf. [5], [26])

$$Q(u,s) = \frac{P(u,0)}{P(u,s)} = \frac{Q_\infty(u)}{P(u,s)} \tag{36}$$

where $P(u,s) = \prod_{j=2}^{\infty}(1 - up^{s+j} - uq^{s+j})$.

Using now Lemma 3 we obtain

$$1 - D_m(u) = \frac{1-u}{m2\pi i} \int_{-3/2-i\infty}^{-3/2+i\infty} \Gamma(s) m^{-s} Q(u, -s-2) ds + e_m \ , \tag{37}$$

where $e_m = O(1/m^2) \int_{-3/2-i\infty}^{-3/2+i\infty} \Gamma(s) m^{-s} s Q(u, -s-2) ds$, and as we shall see $e_m = O(1/m)$, so we can safely ignore it in further computations (see for example [12] for more details).

We now evaluate the integral in (37) by the residue theorem. However, this time we compute residues *right* to the line of integration in (37). More precisely, as in the proof of Lemma 3, we consider a large rectangle right to the line of integration, and after observing that the integral over bottom, right and top lines are small, we are left with residues right to the line of integration.

The gamma function has its singularities at $s_{-1} = -1$ and $s_0 = 0$, and in addition we have infinite number of zero $s_k^j(v)$ ($j = 2, 3, \ldots$, $k = 0 \pm 1, \pm 2, \ldots$) of $P(e^v, -s-2)$ of the denominator of $Q(e^v, -s-2)$ where we substituted $u = e^v$ with $\Re(v) > 0$. More precisely, $s_k^j(v)$ are zeros of

$$p^{-s-2+j} + q^{-s-2+j} = e^{-v} \ . \tag{38}$$

It turns out (cf. [5], [12], [17], [21], [26]) that the dominating contribution to the asymptotics comes from $s_0^j(v)$. Indeed, the contributions of the first two singularities at s_{-1} and s_0 are respectively $-(1-u)Q(u,-1)$ and $(1-u)Q(u,-2)/m$. They can be safely ignored after we multiply everything by $e^{-\vartheta\mu_m/\sigma_m} = e^{-O(\sqrt{\log m})}$. Thus, now we concentrate on the contribution coming from $s_0^j(v)$. In this case, one can solve equation (38) (cf. [11] and [12]) to derive

$$s_0^j(v) = j - 3 - \frac{v}{h} - \frac{1}{2}\left(\frac{1}{h} - \frac{h_2}{h^3}\right) v^2 + O(v^3) \tag{39}$$

for integer $j \geq 2$ and $v \to 0$. We also note that $\Im(s_k^j(v)) \neq 0$ for $k \neq 0$.

Let now $R_k^j(v)$ denote the residue of $(1 - e^v p^{-s_k^j-2+j} + e^v q^{-s_k^j-2+j})^{-1}$ at $s_k^j(v)$, and let $g(s) = \Gamma(s)Q(u, -s-1)$. In the sequel, we use the following expansion which derivation can be found in [26]

$$\begin{aligned} Q(u,-s-2) &= \frac{1}{1-u(p^{-s}+q^{-s})} \cdot \frac{Q_\infty(u)}{P(u,-s-1)} \\ &= -\frac{w^{-1}}{h} - \frac{\theta}{h} + \frac{h_2}{2h^2} + w\frac{\theta h_2}{2h^2} + O(w^2) \end{aligned}$$

where $w = s - s_0^j(v)$. Then, after some straightforward calculations as in [12, 26], by Cauchy's theorem we obtain

$$\begin{aligned} 1 - D_m(e^v) &= R_0^2(v) g(s_0^2(v))(1-e^v) m^{-1} m^{-s_0^2(v)} + \sum_{j=3}^{\infty} R_0^j(v) g(s_0^j(v))(1-e^v) m^{-1} m^{-s_0^j(v)} \\ &+ \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{j=2}^{\infty} R_k^j(v) g(s_k^j(v))(1-e^v) m^{-1} m^{-s_k^j(v)} + O(1) \ . \end{aligned} \tag{40}$$

We consider now the above three terms separately:

(a) $j = 2$ and $k = 0$

Set $v = \vartheta/\sigma_m = \vartheta/\sqrt{c_2 \log m}$ with $\Re(\vartheta) > 0$. Then by (39)

$$m^{-s_0^2(v)} = m \exp\left(\frac{\vartheta}{h}\sqrt{\frac{\log m}{c_2}} + \frac{\vartheta^2}{2}\right) .$$

In addition, the following holds: $R_0^2(v) = -1/h + O(v)$, and $g(s_0^2(v)) = -h/v + O(1)$, and finally $1 - e^{-v} = v + O(1)$ (cf. [12]). Therefore, we obtain

$$e^{-\vartheta\mu_m/\sigma_m} R_0^2(v) g(s_0^2(v))(1 - e^{-v}) m^{-s_0^2(v)} \to -e^{\vartheta^2/2} \tag{41}$$

(b) $j \geq 3$ and $k = 0$

In this case we can repeat the analysis from case (a) to get

$$e^{-\vartheta\mu_m/\sigma_m} R_0^2(v) g(s_0^2(v))(1 - e^{-v}) m^{-s_0^2(v)} \to O(m^{2-j} e^{\vartheta^2/2}) , \tag{42}$$

so this term is of order magnitude smaller than the first term in (40).

(c) $k \neq 0$

Fix $j = 2$. Then, as in Jacquet and Szpankowski [12] we can prove that

$$\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} R_k^2(v) g(s_k^2(v))(1 - e^{v}) m^{-1} m^{-s_k^2(v)} = O(v m^{-\Re(s_k^2(v))}) .$$

But, we also know ([11], [12]) that $\Re(s_k^2(v)) \geq s_0^2(\Re(v))$, so finally by (39) the above sum becomes

$$\begin{aligned} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} R_k^2(v) g(s_k^2(v))(1 - e^{v}) m^{-1} m^{-s_k^2(v)} &= m^{-\Re(s_0^2(v))} O(v m^{\Re(s_0^2(v)) - s_0^2(\Re(v))}) \\ &= m^{-\Re(s_0^2(v))} O(v m^{-\beta v^2}) \end{aligned}$$

for some β. Finally, consider general $j \geq 3$. As in the case (b), we note that $m^{-s_k^j(v)}$ contributes $O(m^{2-j})$, so this term is negligible.

Putting everything together: we note that as $v \to 0$ or $m \to \infty$ we have $e^{-\vartheta\mu_m/\sigma_m}(1 - D_m(e^{\vartheta/\sigma_m})) \sim -e^{-\vartheta\mu_m/\sigma_m} D_m(e^{\vartheta/\sigma_m})$ for $\Re(v) > 0$, and finally

$$e^{-\vartheta\mu_m/\sigma_m} D_m(e^{\vartheta/\sigma_m}) = e^{\vartheta^2/2}(1 + O(v m^{-\beta v^2}) + O(1/m)) \to e^{\vartheta^2/2} \tag{43}$$

which proves part (ii) of Theorem 1.

3.2 Lempel-Ziv Model

We now prove Theorem 2. To assess the distribution of D_n^{LZ} we need to estimate the conditional probability $\Pr\{D_n^{LZ} = k | M_n = m\}$ (cf. (2)). We have pointed out before that $\Pr\{D_n^{LZ} = k|M_n = m\} \neq \Pr\{D_m = k\}$ where D_m is the depth in the digital tree model already estimated in Theorem 1. Nevertheless, we show that these probabilities are not far away.

In the sequel we prove the following two facts that suffice to establish Theorem 2:

A. For large n

$$\Pr\{D_n^{LZ} = k|M_n = m\} = \left(1 + O(\sqrt{\log n/n})\right) \Pr\{D_m = k\} \ . \tag{44}$$

B. For the asymmetric alphabet when $n \to \infty$

$$Ee^{i\theta D_n^{LZ}} \sim Ee^{i\theta D_{\lfloor nh/\log n \rfloor}} \ , \tag{45}$$

that is, the limiting distribution of D_n^{LZ} is asymptotically the same as the limiting distribution of the depth in the digital tree model with $m = \lfloor nh/\log n \rfloor$ nodes. A similar statement is also true for the symmetric model.

A. Reducing to the Digital Tree Model

Let us first fix the number of phrases m. Then, as in the digital tree model $\Pr\{D_m = k\} = \overline{B}_m(k)/m$, and by Theorem 1

$$\overline{B}_m(k) \sim \frac{m}{\sqrt{2\pi c_2 \log m}} \exp\left(-\frac{(k - c_1 \log m)^2}{2c_2 \log m}\right) \tag{46}$$

for $k = O(\log m)$, where as before $c_1 = 1/h$ and $c_2 = (h_2 - h^2)/h^3$. Furthermore, we define $Z_m(k) := B_m(k) - \overline{B}_m(k)$ that represents a deviation of $B_m(k)$ around its mean.

Clearly, the number of nodes at level k is related to the internal path length L_m by $L_m = \sum_{k=1}^m kB_m(k)$. From Jacquet and Szpankowski [13] we know that

$$\frac{L_m - c_1 m \log m}{\sqrt{c_2 m \log m}} \to N(0,1) \ . \tag{47}$$

Actually, the above is an archi-fact used to prove Fact 2(ii) through the renewal theorem (cf. Theorem 17.3 in [2]).

Consider now the Lempel-Ziv model. We must estimate the average number of internal nodes at level k *under the condition that* $L_m = n$. As the first step, let us vary the number of nodes m by introducing a parameter t such that $m \log m/h = nt$. Clearly,

$$m = \frac{nht}{\log n}\left(1 + \frac{\log\log n}{\log n} + O(1/\log n)\right) \tag{48}$$

and by (47)

$$\frac{L_t - nt}{\sqrt{hc_2 n}} \to X_t \ , \tag{49}$$

where X_t is a Gaussian non-Markovian process with $EX_t = 0$ and $\mathrm{Var}X_t = t$. In passing, we note that $X_t \sim \sum_{k=0}^{m} kZ_t(k)/\sqrt{hc_2 n}$.

In order to capture properties of the Lempel-Ziv model, we introduce a random variable τ which represents the first time L_t attains level n. More precisely, $\tau = \min\{t : \ L_t \geq n\}$. Equivalently, τ can be defined as (cf. Fig. 2)

$$\tau = \min\{t : \ X_t \geq c\sqrt{n}(1-t)\} \ , \tag{50}$$

where $c = 1/\sqrt{hc_2}$. Then, for $\tau = t_0$

$$\Pr\{D_n^{LZ} = k | M_n = m\} = \frac{E\{B_\tau(k) | \tau = t_0\}}{m} \ . \tag{51}$$

We need to estimate $EB_{t_0}(k) := E\{B_\tau(k)|\tau = t_0\}$. Let $X_1 = y$ and $X_\tau = x$. From Figure 2 we see that

$$\begin{aligned} y &\sim c\sqrt{n}(1-\tau) \qquad n \to \infty \ , & (52) \\ x &= c\sqrt{n}(1-\tau) \ . & (53) \end{aligned}$$

Hence, by (52) τ is asymptotically normal with $E\tau = 1$ and Var $\tau = hc_2/n$ (i.e, $\tau = 1$ (pr.)). As a direct consequence of the above, we also re-discover that $EM_n \sim nh/\log n E\tau \sim nh/\log n$ and $\mathrm{Var}M_n \sim (h^2n^2/\log^2 n) \cdot \mathrm{Var}\ \tau \sim nh^3c_2/\log^2 n$.

Now, we wrestle with the computation of $EB_{t_0}(k)$. Note that conditioning on $\tau = t_0$ is equivalent to conditioning on $X_\tau = x$. Hence,

$$E\{B_{t_0}(k)|X_\tau = x\} = \overline{B}_{t_0}(k) + E\{Z_\tau(k)|X_\tau = x\} \ . \tag{54}$$

Moreover, since $t_0 = 1 + O(1/\sqrt{n})$

$$\overline{B}_{t_0}(k) \sim \overline{B}_1(k) = O(n/\log^{3/2} n) \tag{55}$$

where the right-hand side of the above follows from (46).

To assess the error we need to estimate $E\{Z_\tau(k)|X_\tau = x\}$. For this we need a more precise estimate of τ. The following lemma is well known (cf. [20]).

Lemma 4. *Consider an ordinary Brownian motion $B(t)$. Define*

$$\tau = \inf\{t : \ \mu t + B(t)\sigma/\sqrt{n} = \alpha\} \ .$$

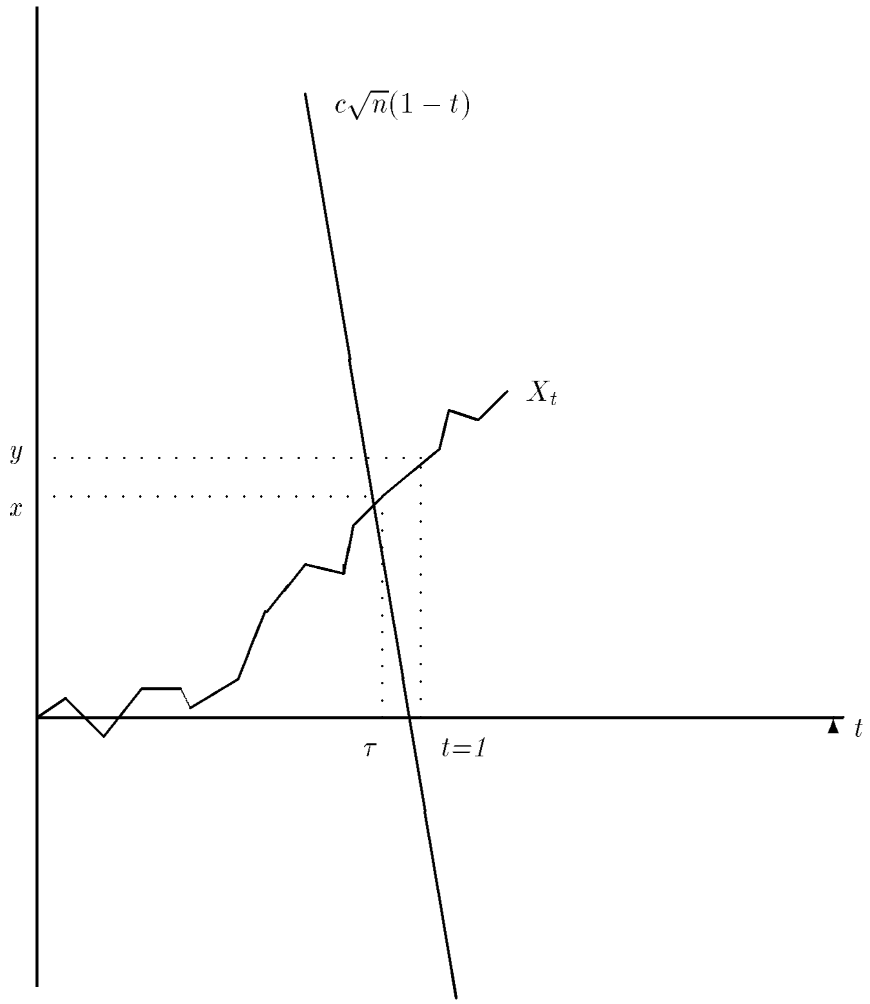

Figure 2: Illustration to the analysis.

Let $T = \tau - \alpha/\mu$. Then, the asymptotic density $f(t)$ for T becomes

$$f(t) = \frac{\sqrt{n}\mu^{3/2}}{\sqrt{2\pi\alpha}\sigma} \exp\left(\frac{-nt^2\mu^3}{2\alpha\sigma^2}\right)\left(1 + C_1 t + O(t^2)\right) \tag{56}$$

Furthermore, $T = O(\frac{1}{\sqrt{n}})$ (pr.) and the relative error in the density with respect to the Gaussian density is also $O(\frac{1}{\sqrt{n}})$.

Proof. By a classical result, the density $f(u)$ for τ is given by

$$f(u) = \frac{\alpha\sqrt{n}}{\sigma\sqrt{2\pi}u^{3/2}} \exp\left(\frac{-n(\alpha-\mu u)^2}{2u\sigma^2}\right) \tag{57}$$

Let $T := \tau - \alpha/\mu$. Setting $t := u - \alpha/\mu$, and expanding (57) around $t = 0$, we derive (56). Obviously, $T = O(\frac{1}{\sqrt{n}})$ (pr.) since the density of T is of order $O(e^{-Ant^2})$. In addition, the relative error in the density is also $O(\frac{1}{\sqrt{n}})$. ∎

To apply Lemma 4, we refer to Durbin [3] from whom we conclude that around τ the process X_t behaves locally like a Brownian motion. In our case, $\alpha/\mu = 1$, $\sigma = \alpha/c$ and the crossing time $X_\tau = x$ and T are related by $x = -c\sqrt{n}T$ $c = 1/\sqrt{hc_2}$. Hence, by (56) of Lemma 4 we see that the density $f(x)$ of the crossing value $X_\tau = x$ is given by

$$f(x) \sim \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}\left(1 + C_2\frac{x}{\sqrt{n}}\right) \tag{58}$$

In order to assess $EB_{t_0}(k)$ we need to estimate $\int_{-\infty}^{\infty} f(x)E\{Z_\tau(k)|X_\tau = x\}dx$ as suggested by (54). First of all, we observe that Theorem 1 and (54) imply for every k

$$\int_{-\infty}^{\infty} \frac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}} E\{Z_1(k)|X_1 = y\}dy = 0 \ . \tag{59}$$

Moreover, since during T at most $O(n|T|/\log n)$ phrases can be generated (cf. (8)), the following two estimates are easy to establish:

$$Z_\tau(k) = Z_1(k) + O(n|T|/\log n) = Z_1(k) + O(\sqrt{n}x/\log n) \ , \tag{60}$$

and $y = X_1$ becomes

$$y = X_\tau + \sqrt{|T|}\xi = x + C\frac{\sqrt{|x|}}{n^{1/4}}\xi \ , \tag{61}$$

where C is a constant and ξ a random variable distributed according to the standard normal distribution. Note that $T = O(1/\sqrt{n})$, hence $x = y - \sqrt{y}\xi/n^{1/4} + O(1/\sqrt{n})$.

Putting everything together. From (58), the density $f(x)$ in terms of y becomes

$$\psi(y) \sim \frac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}}\left(1 + C_4\frac{y^{3/2}\xi}{n^{1/4}} + C_2\frac{y}{\sqrt{n}}\right) \ .$$

Then, by (59) and the above

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)E\{Z_\tau(k)|X_\tau = x\} &\sim \int_{-\infty}^{\infty} \psi(y)\left(E\{Z_1(k)|X_1 = y\} + O(\sqrt{n}y/\log n)\right)dy \\ &= O(\sqrt{n}/\log n) \ . \end{aligned} \tag{62}$$

This completes the proof of (44) since $\overline{B}_{t_0} = O(n/\log^{3/2} n)$, as noticed in (55).

B. Finishing the Proof

We first consider the asymmetric alphabet, and prove part (ii) of Theorem 2. From Theorem 1 (cf. (14) and (43)) we conclude that for some real θ (for simplicity we consider $\vartheta = i\theta$ but as easy to see our proof works for any complex ϑ such that $\Re(\vartheta) > 0$)

$$E\left(e^{i\theta D_m}\right) = \exp\left(i\theta c_1 \log m - (1/2)\theta^2 c_2 \log m\right)\left(1 + O\left(\theta/\sqrt{\log m}\right)\right) \ , \qquad (63)$$

where D_m is the depth in the digital tree model. Let now $f(n) = \sqrt{\log(nh/\log n)}$. Define $F(\theta) = Ee^{i\theta D_n^{LZ}/f(n)}$. Then, from the above

$$F(\theta) = E\left\{\exp\left(i\theta c_1 \log M_n/f(n) - (1/2)\theta^2 c_2 \log M_n/f^2(n)\right)(1 + O(\theta/\sqrt{\log M_n})\right\} \ . \quad (64)$$

Let $\xi_n := (M_n - EM_n)/\sqrt{\mathrm{Var} M_n}$, and observe that $Ee^{i\theta\xi_n/g(n)} \to \exp(-\theta^2/(2g^2(n)))$ for some real θ and real-valued function $g(n)$ (cf. Fact 2). Note that $\log M_n = \log(nh/\log n) + \log(1 + \xi_n c/\sqrt{n})$ for some constant c, due to $EM_n \sim nh/\log n$ and $\mathrm{Var}M_n \sim cn/\log^2 n$. Therefore, from (63), (64) and the above one obtains

$$\begin{aligned} F(\theta) &= \exp(i\theta c_1 \sqrt{\log(nh/\log n)} - (1/2)\theta^2 c_2) \\ &\cdot \ E\left(e^{\frac{\vartheta}{g(n)}\log(1+\xi_n c/\sqrt{n})}\left(1 + O(1/\sqrt{\log M_n})\right)\right) \end{aligned}$$

where $\frac{1}{g(n)} = \frac{1}{f(n)}(1 + O(1/f(n)))$. But, according to (3) $O(\sqrt{n}) \le M_n \le O(n/\log_2 n)$, hence $\log(1 + \xi_n c/\sqrt{n}) = \Theta(1)$. Therefore, by the *bounded convergence theorem* (cf. [4]) we immediately obtain $Ee^{(\vartheta/g(n))\log(1+\xi_n c/\sqrt{n})} \to 1$, and finally

$$e^{-i\theta c_1\sqrt{\log(nh/\log n)}} F(\theta) = e^{-c_2\theta^2/2}(1 + O(1/\sqrt{\log n})) \ , \qquad (65)$$

which completes the proof of part (ii). Clearly, the above is true if $i\theta$ is replaced by a complex ϑ such that $\Re(\vartheta) > 0$.

Now, we turn our attention to the symmetric alphabet, and establish part (i) of Theorem 2. Since in this case we have exact distribution for D_m (cf. (17)) we can easily by-pass most of analytical difficulties. Therefore, we rather present a sketch of the proof leaving most of the details to the interested reader. We consider the limiting distribution (12) as a conditional distribution with $M_n = m$. Ignoring for a moment $\psi_1(n)$, we need only to investigate $2^x e^{-2^{-x-1-i}}$ where $x = j - \log_2 M_n$ for some integer j. Note that for such x the above expression becomes $2^x e^{-2^{-x-1-i}} = (2^j/M_n)e^{-\alpha M_n}$ where $\alpha = 2^{-j+1+i}$. By the result of Aldous and Shields [1] (cf. Fact 2(i))

$$M_n = \frac{n}{\log_2 n} + \xi_n O\left(\sqrt{n/\log_2^3 n}\right) = \frac{n}{\log_2 n}\left(1 + \xi_n O(1/\sqrt{n \log n})\right) \qquad (66)$$

where $\xi_n \to N(0,1)$. To complete the proof it suffices to estimate the following

$$\frac{e^{-\alpha n/\log_2 n}}{n/\log_2 n}\int_{-\infty}^{\infty}\frac{e^{-\alpha x O(\sqrt{n/\log_2^3 n})}}{1+xO(1/\sqrt{n\log n})}dF_\xi(x)=\frac{e^{-\alpha n/\log_2 n(1+O(1/\log^2 n))}}{n/\log_2 n}(1+O(1/\log^2 n))$$

where $F_\xi(x)$ is the standard normal distribution function. Clearly, the above proves part (i), and this completes the proof of Theorem 2.

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] J. Durbin, The first-Passage Density of Continuous Gaussian Process to a General Boundary, *J. Appl. Probab.*, 22, 99-122 (1985).

[4] Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley & Sons, New York (1971).

[5] P. Flajolet and R. Sedgewick, Digital Search Trees Revisited, *SIAM J. Computing*, 15, 748-767 (1986).

[6] P. Flajolet and M. Soria, General Combinatorial Schemas: Gaussian Limit Distributions and Exponential Tails, *Discrete Mathematics*, 114, 159-180 (1993).

[7] P. Flajolet and B. Richmond, Generalized Digital Trees and Their Difference-Differential Equations, *Random Structures & Algorithms*, 3, 305-320 (1992).

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

[9] P. Grassberger, Estimating the Information Content of Symbol Sequences and Efficient Codes, *IEEE Trans. Information Theory*, 35, 669-675 (1991).

[10] P. Henrici, *Applied and Computational Complex Analysis*, John Wiley&Sons, New York 1977.

[11] P. Jacquet and M. Régnier, Trie Partitioning Process: Limiting Distributions, *Lecture Notes in Computer Science*, vol. 214, 196-210, Springer-Verlag New York (1986).

[12] P. Jacquet and W. Szpankowski, Analysis of Digital Tries with Markovian Dependency, *IEEE Trans. Information Theory*, 37, 1470-1475 (1991).

[13] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, to appear (also Purdue University, CSD-TR-93-088, 1992).

[14] P. Kirschenhofer and H. Prodinger, Further Results on Digital Search Trees, *Theoretical Computer Science*, 58, 143-154 (1988).

[15] P. Kirschenhofer, H. Prodinger and W. Szpankowski, On the Variance of the External Path in a Symmetric Digital Trie, *Discrete Applied Mathematics*, 25, 129–143 (1989).

[16] P. Kirschenhofer, H. Prodinger and W. Szpankowski, Digital Search Trees Again Revisited: The Internal Path Length Perspective, *SIAM J. Computing*, 23, 598-616 (1994).

[17] D. Knuth, *The Art of Computer Programming. Sorting and Searching,* Addison-Wesley (1973).

[18] A. Konheim and D.J. Newman, A Note on Growing Binary Trees, *Discrete Mathematics*, 4, 57-63 (1973).

[19] G. Louchard, Exact and Asymptotic Distributions in Digital and Binary Search Trees, *RAIRO Theoretical Inform. Applications*, 21, 479-495 (1987).

[20] G. Louchard and R. Schott, Probabilistic Analysis of Some Distributed Algorithms, *Random Structures & Algorithms*, 2, 151-185 (1991).

[21] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York (1992).

[22] D. Ornstein and B. Weiss, Entropy and Data Compression Schemes, *IEEE Information Theory*, 39, 78-83 (1993).

[23] B. Pittel, Asymptotic Growth of a Class of random Trees, *Annals of Probability*, 13, 414 - 427 (1985).

[24] J. Rissanen, A Universal Data Compression System, *IEEE Trans. Information Theory*, 29, 656-664 (1983).

[25] W. Szpankowski, The Evaluation of an Alternating Sum with Applications to the Analysis of Some Data Structures, *Information Processing Letters*, 28, 13-19 (1988).

[26] W. Szpankowski, A Characterization of Digital search Trees From the Successful Search Viewpoint, *Theoretical Computer Science*, 85, 117-134 (1991).

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

[28] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, 39, 1647-1659 (1993).

[29] A. Wyner and J. Ziv, Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression, *IEEE Trans. Information Theory*, 35, 1250-1258 (1989).

[30] J. Ziv, On Classification with Empirically Observed Statistics and Universal Data Compression, *IEEE Trans. Information Theory*, 34, 278-286 (1988).

[31] J. Ziv, Compression, Test of Randomness, and Estimating the Statistical Model of Individual Sequences, *SEQUENCES*, R. Capocelli, Ed. New York: Springer-Verlag, 366-373 (1990).

[32] J. Ziv and A. Lempel, A Universal Algorithm for Sequential Data Compression, *IEEE Trans. Information Theory*, 23, 3, 337-343 (1977).

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