Average profile and limiting distribution for a phrase size in thw Lempel-Ziv parsing algorithm

Guy Louchard, Wojciec Szpankowski

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm

Guy LOUCHARD
Wojciech SZPANKOWSKI

N° 1886
Avril 1993

PROGRAMME 2

Calcul Symbolique,
Programmation
et Génie logiciel

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

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 the average value 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.

ALGORITHME DE LEMPEL-ZIV: PROFIL MOYEN ET ISTRIBUTION LIMITE DE LA TAILLE D'UNE PHRASE

Résumé

Considérons l'algorithme d'analyse lexicale de Lempel-Ziv qui partitionne une suite de longueur n en phrases variables (blocks) de telle sorte qu'un nouveau block est le sous-mot le plus court non encore rencontré précédemment en tant que phrase. Les paramètres suivants sont d'un grand intérêt pratique: nombre de phrases, taille d'une phrase, nombre de phrases de taille donnée. Cet article est consacré à la taille d'une phrase *choisie au hasard* et au nombre moyen de phrases de taille donnée (*le profil moyen de la taille des phrases*). Ces paramètres peuvent être analysés efficacement à l'aide d'une représentation en arbre digital de recherche. Lorsque la source est sans mémoire, avec des probabilités inégales de génération de symboles (*modèle assymétrique de Bernoulli*), nous démontrons que la taille d'une phrase typique est asymptotiquement Gaussienne et nous donnons explicitement ses moyenne et variance. En terme d'arbre digital de recherche, nous prouvons le caractère asymptotiquement Gaussien de la profondeur d'un sommet (c'est-à-dire la longueur du chemin de la racine à un sommet choisi au hasard). Ce dernier résultat est démontré par une technique appartenant aux outils de "l'analyse analytique d'algorithmes", approche qui semble être nouvelle dans le contexte de la compression de données.

*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 several universal data compression schemes (cf. [32]) is the parsing algorithm due to Lempel and Ziv [17]. 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 [31], and explored by Grassberger [8] and Szpankowski [24], 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 a universal data compression scheme and its numerous applications such as efficient transmission of data (cf. [17], [31], and [28]), discriminating between information sources (cf. [7], [30]), test of randomness (cf. [30]), estimating the statistical model of individual sequences (cf. [29], [30]), 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 studying in the past as the first-order properties, that is, typical behaviors in the almost sure sense. Very few results – with a noble exception of the paper by Aldous and Shields [1] – are available up-to-date concerning the second order properties such as limiting distributions, large deviation results, concentration of mean, etc.

Recently, Gilbert and Kadota [7] argued for the necessity of such investigations. The authors of [7] 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 other results, 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], [15]).

We consider a special type of digital trees, namely a *digital search tree* (cf. [5], [6], [15], [16]). 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 left or the right child of the root depending whether its first symbol is "0" or "1". 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

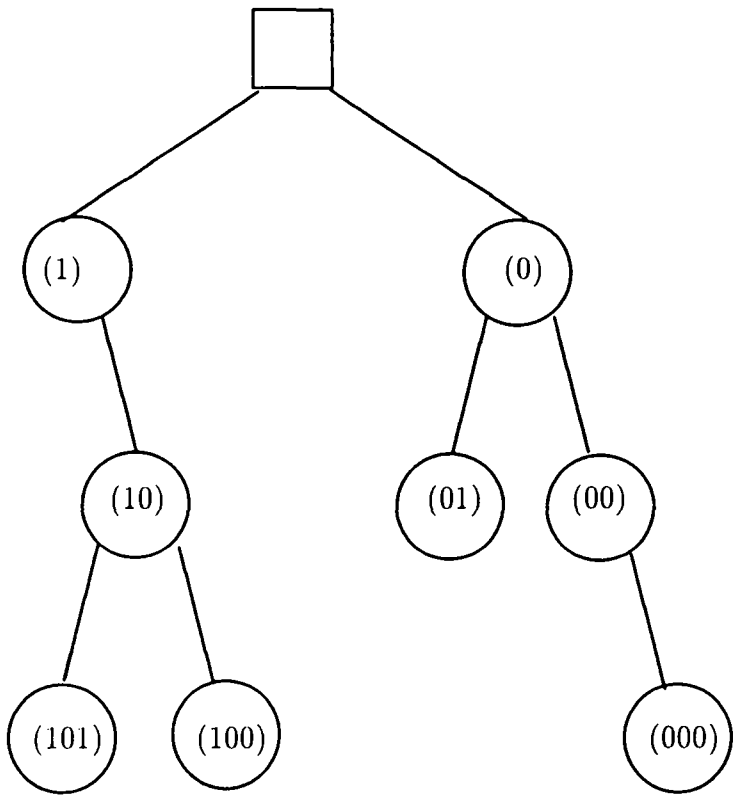

Figure 1: A digital tree representation of Ziv's parsing for the string 11001010001000100...

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 [15] and [20].

The Lempel-Ziv parsing algorithm can be efficiently implemented using the digital search tree structure. For simplicity, 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 tree construction. For example, for the binary alphabet, "0" in the input string means to move to the left and "1" means to 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, the edge and the new node are added to the tree. The phrases created in such a way are stored directly into the 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. [8], [24], [26]).

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 into two different frameworks. Either – as suggested by Gilbert of Kadota [7] – with *fixed number* m of parsed words, or – as originally discussed in Lempel and Ziv [17] – with *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. Therefore, any new result in this endouver will lead to a new finding in the area of digital trees, and reverse: we can apply many known results on such trees (for a survey see Mahmoud [20]) 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 are 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 the *Lempel-Ziv model*, as we shall call the latter model.

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, while m is to represent the number of *independent* strings used to built a digital tree 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) 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 $O(1)$.

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 *fixed* number of independent strings (cf. [5, 6, 10, 13, 14, 15, 16, 18, 22, 25]). 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 [16], Knuth [15], Flajolet and Sedgewick [5], and Flajolet and Richmond [6]. The average depth and the variance of the depth for the asymmetric Bernoulli model is given in Szpankowski [25] (the symmetric case was also analyzed in Kirschenhofer and Prodinger [12]), while the variance of the internal path length in the symmetric Bernoulli model was investigated in Kirschenhofer *et al.* [14]. Finally, Louchard [18], and Aldous and Shields [1] for the symmetric Bernoulli alphabet

obtained the limiting distribution of the depth. As mentioned above, in this paper – among other results – we directly extend Louchard's result to asymmetric Bernoulli model, while in a forthcoming paper we intend to generalize some of the Aldous and Shields [1] results concerning the limiting distribution of the internal path length.

On the other hand, 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 [11], [14]). The first-order property of the length of a phrase in the Lempel-Ziv parsing algorithm was recently reported by Ornstein and Weiss [21]. Finally, Gilbert and Kadota [7] 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.

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 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. Out of these words we build a sequence (of random length) according to the Lempel-Ziv scheme. Alternatively, we can construct a digital search trees from these m strings. Then, the length of a randomly selected phrase is the same as the length of a randomly selected depth (i.e., the path from the root to a node). Traditionally, in the area of digital trees this depth is denoted as D_m, and we shall adopt this notation. More precisely, let $D_n(i)$ be the length of the ith phrase (or the ith depth), where $1 \le i \le m$. Then, the typical depth D_m is defined 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 the sequence produced by the Lempel-Ziv parsing scheme from these m parsed words. 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.

The situation is similar, but *not* the same in the Lempel-Ziv model in which a sequence of 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 the lower and the upper bounds for the number of phrases. It is easy to see that there exist constants α_1 and α_2 such that

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

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

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 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...111 which is very skewed). Fortunately, we shall prove in sequel that $\Pr\{D_n^{LZ} = k | M_n = m\} = (1 + O(\sqrt{\log n / n}))\Pr\{D_m = k\}$.

In passing we note that the internal path length of the associated tree L_{M_n} in the Lempel-Ziv model satisfies the following relationship (cf. [11])

$$L_{M_n - 1} < n \leq L_{M_n} \ . \tag{4}$$

which is useful in estimating the limiting distribution of M_n.

We first consider the **digital 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 [15] (cf. [24]), we have the following relationship between the depth D_m and the average profile $\overline{B}_m(k)$

$$\Pr\{D_m = k\} = \frac{\overline{B}_m(k)}{m} \ . \tag{5}$$

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. [15], [24])

$$B_{m+1}(u) = 1 + u \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-1-j} (B_j(u) + B_{m-j}(u)) \tag{6}$$

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

A general recurrence of the above type was analyzed in Szpankowski [24] (cf. see also Flajolet and Richmond [6] for interesting extensions). A slight modification of Theorem 2.4 in [24] directly leads to the exact solution of (6), namely:

$$B_m(u) = m - (1-u) \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u) \tag{7}$$

where

$$Q_k(u) = \prod_{j=2}^{k+1} (1 - up^j - uq^j) \, . \tag{8}$$

We consider the symmetric and the asymmetric cases separately. For the symmetric model, we exactly compute the coefficients at u^k of $B_m(u)$ (i.e., $\overline{B}_m(k)$) directly from (7). For the asymmetric model, we use Goncharov's theorem (cf. [15]) applied to the probability generating function $D_m(u) = B_m(u)/m$ to establish the limiting normality of D_m (for details see Section 3.1). In the latter case we need one more result from [24] that is provided below for the reader convenience.

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

$$ED_m = \frac{1}{h} \left(\log m + \gamma - 1 + \frac{H}{2h} + \theta + \delta(m) \right) + O(\log m/m) \tag{9}$$

where h is the entropy, $H = p \log^2 p + q \log^2 q$, $\gamma = 0.577\ldots$ is the Euler constant, $\delta(m)$ is a fluctuating function with a small amplitude, and

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

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

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

where A is a constant and $\Delta(m)$ is a fluctuating function with a small amplitude. In the symmetric case, the coefficient at $\log m$ becomes zero, and then (cf. [14])

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

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 $\Delta(m)$ is a periodic function of $log_2 m$. ∎

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

Theorem 1. (i) *For the symmetric Bernoulli model the limiting distribution of D_m is*

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

for such real x that $x + \log_2 m$ is integer, with $Q_k = \prod_{j=1}^{k}(1 - 2^{-j})$.

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

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

where ED_m and $var\ D_m$ are given by (9) and (10), respectively. Moreover, the moments of D_m converges to the appropriate moments of the normal distribution. More precisely, for any complex ϑ

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

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

Remark 1. The limiting distribution for the symmetric case was obtained before by Louchard [18] by a different method that the one presented in Section 3.1. It should be noted that the limiting distribution of D_m for *every* real x does not exist. We can, however, write

$$\lim_{m\to\infty} \sup_x \left| \Pr\{D_m \le x\} - \frac{2^x}{m} \left(1 + \frac{1}{2Q_\infty} \sum_{i=0}^{\infty} (-1)^{i+1} \frac{2^{-i(i+3)/2}}{Q_i} \exp(-m2^{-(x-1-i)})\right)\right| = 0 \tag{15}$$

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

$$\Pr\{D_m \le j + 1\} = \frac{1}{m} \left(2^{j+1} - 1 + \sum_{k=1}^{j} 2^k \frac{(-1)^{j-k+1} 2^{(j-k)(j-k+1)/2}}{Q_{j-k} Q_{k-1}} (1 - 2^{-k})^{m-1}\right) \tag{16}$$

for all integers $j \geq 1$. □

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}. Aldous and Shields [1] proved the following deep result.

Fact 2. *In the symmetric Bernoulli model*

$$\frac{M_n - EM_n}{\sqrt{varM_n}} \to N(0,1) \tag{17}$$

where $N(0,1)$ denotes the standard normal distribution, with $EM_n \sim n/\log_2 n$ and $varM_n \sim \Theta(n/\log_2^3 n)$. ■

Remark 2. In fact, the authors of [1] established a stronger result, namely, the limiting distribution of $M_n(k)$. □

There is no corresponding limiting distribution for M_n for the asymmetric model. However, Jacquet and Szpankowski [11] conjectured the following:

Conjecture. *In the asymmetric Bernoulli model*

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

where $EM_n \sim nh/\log_2 n$ and $varM_n \sim c_2 h^3 n/\log^2 n$ with h being the entropy. Moreover, moments of M_n converge to the appropriate moments of the normal distribution. In other words, for some complex ϑ the following holds

$$e^{-\vartheta\sqrt{n/c_2 h}} E\left(e^{\vartheta M_n \log n/\sqrt{nh^3 c_2}}\right) \to e^{\vartheta^2/2} \tag{19}$$

where c_2 is defined above. ■

Here is the idea of the proof for the above conjecture. Consider the relationship (4) which can be rewritten as (cf. [11])

$$\Pr\{M_n > m\} = \Pr\{L_m \leq n\} \ . \tag{20}$$

Hence, knowing the limiting distribution of the internal path length in the digital tree model will suffice to compute the limiting distribution of M_n. We simply appeal to the renewal equation as in [11] (cf. Theorem 17.3 in Billingsley [2]). Using the idea of Jacquet and Régnier [9], we shall prove the limiting distribution of the internal path length L_m (cf. [11] for more details).

In passing, we note that the above approach turns out to be successful in estimating the coefficient at $n/\log_2^3 n$ in the variance of M_n in the symmetric case. Indeed, using the result of [14] we prove in [11] that

$$varM_n \sim (C + \delta(\log_2 n))\frac{n}{\log_2^3 n} \tag{21}$$

for the symmetric Bernoulli model where $\delta(\log_2 n)$ 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 [14], 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. We again point out that the difficulties of analyzing it arise from the fact that $\Pr\{D_m = k\} \neq \Pr\{D_{M_n} = k | M_n = m\}$, so we cannot directly applied Theorem 1. Nevertheless, the following is true. The proof is delayed till Section 3.2.

Theorem 2. (i) *The length of a randomly selected phrase for the symmetric Bernoulli model has the following limiting distribution*

$$\lim_{n\to\infty} \Pr\{D_n^{LZ} = x + \log_2(n/\log_2 n)\} = 2^{x-1}\left(1 + \frac{1}{Q_\infty}\sum_{i=0}^{\infty}(-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i}e^{-2^{-(x-1-i)}}\right) \tag{22}$$

for such real x that $x + \log_2(n/\log_2 n) = j$ is an integer. If x is any real number, then the limiting distribution does not exist, but (15) still holds with m replaced by $n/\log_2 n$.

(ii) *For the asymmetric Bernoulli model the typical depth D_n^{LZ} is normally distributed. More precisely,*

$$\frac{D_n^{LZ} - c_1\log(nh/\log n)}{\sqrt{c_2\log(nh/\log n)}} \to N(0,1) \tag{23}$$

provided our Conjecture is true. In fact, the rate of convergence is $1 + O(1/\sqrt{\log n})$. ■

Remark 3. (i) *Extensions.* It is plausible that our analysis can be extended to the 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 [10]).

(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

[22] 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,

$$\liminf_{m\to\infty} \frac{\ell_m}{\log m} = \frac{-1}{\log p_{\min}} \quad (a.s.) \qquad \limsup_{m\to\infty} \frac{\ell_m}{\log m} = \frac{-1}{\log p_{\max}} . \tag{24}$$

The same is true for D_m and D_n^{LZ} (cf. [11, 25, 26]).

(iii) *Average Profile.* The average profile $\overline{B}_m(k)$ directly follows from Theorem 1 and (5). The limiting distribution of the profile $B_m(k)$ is much 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. [11])

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

with $B_0^0(u) = 1$. This recurrence is much harder to solve that the one for the average profile since the above recurrence is a multiplicative one, while the recurrence (6) has an additive form. □

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, and are of their own interests.

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 required quite different techniques to establish the claimed results.

A. Symmetric Bernoulli Model

We pick our analysis where we left it in Section 2, that is, from recurrence (6) which – as we indicated – has solution (7). More precisely:

$$B_m(u) = m - (1-u) \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u) \tag{25}$$

where

$$Q_k(u) = \prod_{j=1}^{k} (1 - u2^{-j}) . \tag{26}$$

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, as in Louchard [18],

$$\frac{1}{Q_\infty(u)} = \sum_{i=0}^{\infty} \frac{u^i}{2^i Q_i} \quad , \quad Q_\infty(u) = -\sum_{i=0}^{\infty} u^i R_i \tag{27}$$

where

$$R_i = (-1)^{i+1} \frac{2^{-i(i+1)/2}}{Q_i} \tag{28}$$

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

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

Hence applying this to our basic solution (25) we obtain

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

Finally, after some tedious algebra one obtains (16) as in Louchard [18], and taking $m \to \infty$ we easily derive part (i) of Theorem 1 (see also Mahmoud [20], 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,

$$D_m(u) = 1 - \frac{1-u}{m} \sum_{k=2}^{m} (-1)^k \binom{m}{k} Q_{k-2}(u) \ . \tag{29}$$

Let $\mu_m = ED_m$ and $\sigma_m^2 = varD_m$. 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 $c_2 = (H - h^2)/h^3$. We use Goncharov's theorem to establish the normal distribution of D_m by showing that the following holds

$$\lim_{m \to \infty} e^{-\vartheta \mu_m/\sigma_m} D_m(e^{\vartheta/\sigma_m}) = e^{\vartheta^2/2} \tag{30}$$

where $\vartheta = ix$ for imaginary i. However, below we prove a stronger result, namely we show that (30) holds for *any* complex ϑ, and hence this will automatically establish convergence of moments (since every analytical function has its derivative).

We now derive an asymptotic expansion for the probability generating function $D(u)$ around $u = 1$. We assume $u = e^v$, and due to $\sigma_m \sim \sqrt{\log m}$, 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 (29) 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. [15], [27]). The Mellin-like approach is recalled below for the reader convenience.

Lemma 3. (Szpankowski [27]). *Let f_k be any sequence such that it has an analytical continuation $f(s)$ in the complex plane (i.e., $f(k) = f_k$) such that $f(s)$ does not grow faster than exponential for large s (for details see [27]). Then*

$$\sum_{k=2}^{m}(-1)^k\binom{m}{k}f_k = \frac{1}{2\pi i}\int_{-3/2-i\infty}^{-3/2+i\infty}\Gamma(s)f(s)m^{-s}ds + e_m \tag{31}$$

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

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 $varD_m$ 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 note that (cf. [5], [24])

$$Q(u,s) = \frac{P(u,0)}{P(u,s)} = \frac{Q_\infty(u)}{P(u,s)} \tag{32}$$

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

Using now Lemma 3 we obtain

$$1 - D_m(u) = \frac{1-u}{m}\int_{-3/2-i\infty}^{-3/2+i\infty}\Gamma(s)m^{-s}Q(u,-s-2)ds + \text{ smaller order terms }, \tag{33}$$

where the "smaller order terms" come from e_m and can be savely ignore in further computations (see for example [10] for more details).

We need to assess the integral in (33), but this is easy since we can apply the residue theorem. Note that 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 $Q(e^v, -s-2)$. That is, $s_k^j(v)$ are zeros of

$$p^{-s_k^j(v)-2+j} + q^{-s_k^j(v)-2+j} = e^{-v}\ . \tag{34}$$

It turns out (cf. [5], [10], [15], [20], [24]) that the dominating contribution to the asymptotics comes from $s_0^j(v)$. In this case, one can solve equation (34) (cf. [9] and [10])

to derive

$$s_0^j(v) = j - 3 - \frac{v}{h} - \frac{1}{2}\left(\frac{1}{h} - \frac{H}{h^3}\right)v^2 + O(v^3) \tag{35}$$

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 $(1 - e^v p^{-s_k^j(v)-2+j} + e^v q^{-s_k^j(v)-2+j})^{-1}$ at $s_k^j(v)$, and let $g(s) = \Gamma(s)Q(u,s)$. Then, by Cauchy's theorem we obtain

$$\begin{aligned} 1 - D_m(e^v) &= R_0^2(v)g(s_0^2(v))(1-e^v)m^{-1}m^{-s_0^2(v)} + \sum_{j=3}^{\infty} R_0^j(v)g(s_0^j(v))(1-e^v)m^{-1}m^{-s_0^j(v)} \\ &+ \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}\sum_{j=2}^{\infty} R_k^j(v)g(s_k^j(v))(1-e^v)m^{-1}m^{-s_k^j(v)} . \end{aligned} \tag{36}$$

We consider now the above three terms separately:

A. $j = 2$ and $k = 0$

Note that $v = \vartheta/\sigma_m = \vartheta/\sqrt{c_2 \log m}$. Hence by (35)

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

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 $1 - e^{-v} = v + O(1)$. Therefore, we obtain

$$e^{-\vartheta\mu_m/\sigma_m} R_0^2(v)g(s_0^2(v))(1-e^{-v})m^{-s_0^2(v)} \to e^{\vartheta^2/2} \tag{37}$$

B. $j \geq 3$ and $k = 0$

In this case we can repeat the analysis from case A to get

$$e^{-\vartheta\mu_m/\sigma_m} R_0^2(v)g(s_0^2(v))(1-e^{-v})m^{-s_0^2(v)} \to O(m^{2-j}e^{\vartheta^2/2}) , \tag{38}$$

so this term is of order magnitude smaller than the first term in (36).

C. $k \neq 0$

Fix $j = 2$. Then, as in Jacquet and Szpankowski [10] we can prove that

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

But, we also know ([9], [10]) that $\Re(s_k^2(v)) \geq s_0^2(\Re(v))$, so finally by (35) the above sum becomes

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

for some β. Finally, consider 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 for $v \to 0$ or $m \to \infty$

$$e^{-\vartheta\mu_m/\sigma_m} D_m(e^{\vartheta/\sigma_m}) = e^{\vartheta^2/2}(1 + O(vm^{-\beta v^2}) + O(1/m)) \to e^{\vartheta^2/2} \tag{39}$$

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

A. For large n

$$\Pr\{D_n^{LZ} = k | M_n = m\} = \left(1 + O(\sqrt{\log n / n})\right) \Pr\{D_m = k\} . \tag{40}$$

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

$$E e^{i\theta D_n^{LZ}} \sim E e^{i\theta D_{nh/\log n}} , \tag{41}$$

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 $m = nh/\log n$ 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

$$\overline{B}_m(k) \sim \frac{m}{\sqrt{2\pi c_2 \log m}} \exp\left(-\frac{(k - c_1 \log m)^2}{2c_2 \log m}\right) \tag{42}$$

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

Clearly, the number of nodes at level k, $B_m(k)$, is related to the internal path length L_m by $L_m = \sum_{k=1}^{m} kB_m(k)$. We assume that

$$\frac{L_m - c_1 m \log m}{\sqrt{c_2 m \log m}} \to N(0,1) , \tag{43}$$

which directly implies our Conjecture by the renewal theorem (cf. [2], [11]). In fact, we should have conjectured (43), and then (18) is a direct consequence.

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,

$$m = \frac{nht}{\log n}\left(1 + \frac{\log\log n}{\log n} + O(1/\log n)\right) \tag{44}$$

and by (43)

$$\frac{L_t - nt}{\sqrt{hc_2 n}} \to X_t , \tag{45}$$

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

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)

$$\tau = \min\{t : \ X_t \geq c\sqrt{n}(1-t)\} , \tag{46}$$

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

$$\Pr\{D_n^{LZ} = k | M_n = m\} = \frac{E\{B_\tau(k)|\tau = t_0\}}{m} . \tag{47}$$

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

$$y \ \sim \ c\sqrt{n}(1-\tau) \qquad n \to \infty , \tag{48}$$

$$x \ = \ c\sqrt{n}(1-\tau) . \tag{49}$$

Hence, by (48) τ is asymptotically normal with $E\tau = 1$ and $var\ \tau = hc_2/n$ (i.e, $\tau = 1$ (pr.), but in fact $\tau = 1$ (a.s.)). 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 $varM_n \sim (h^2n^2/\log^2 n) \cdot var\ \tau \sim nh^3c_2/\log^2 n$.

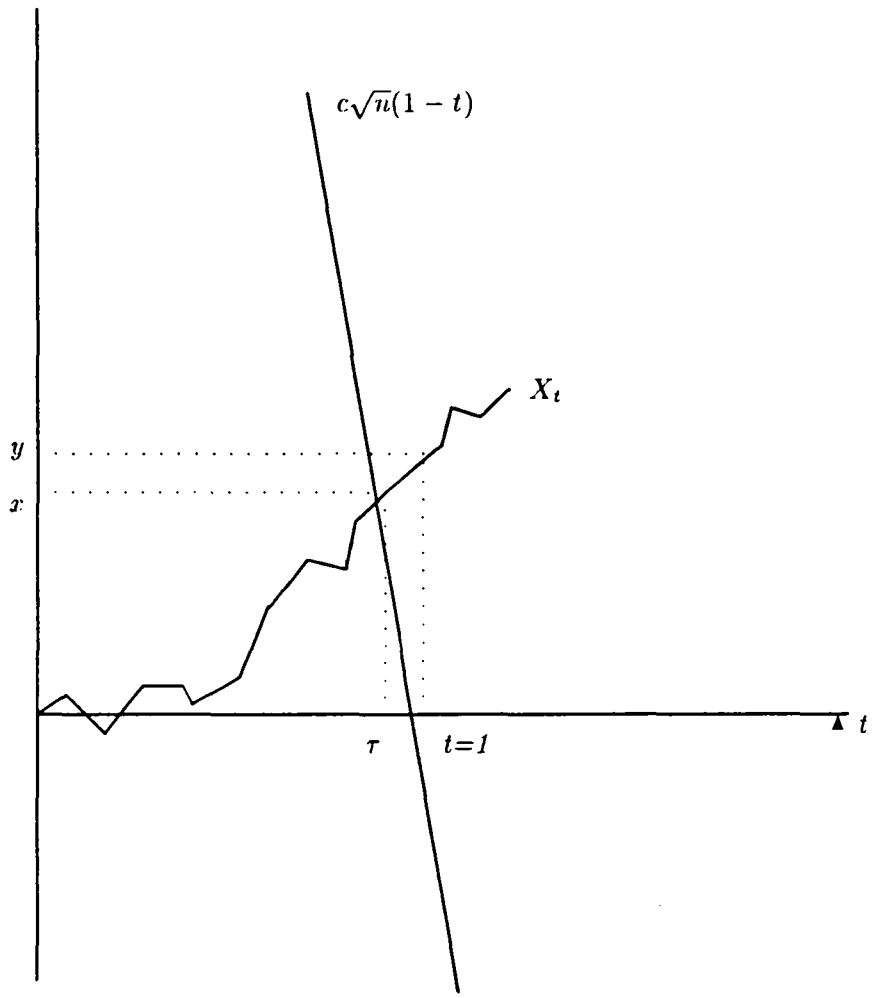

Figure 2: Illustration to the analysis.

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,

$$EB_{t_0}(k) = \overline{B}_{t_0}(k) + E\{Z_\tau(k)|X_\tau = x\} \ . \tag{50}$$

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

$$\overline{B}_{t_0}(k) \sim \overline{B}_1(k) = O(n/\log^{3/2} n) \tag{51}$$

where the right-hand side equation follows from (44).

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. [19]).

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

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

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

$$f(t) = \frac{\sqrt{n}\mu^{3/2}}{\sqrt{2\pi\alpha}\sigma} \ exp \ \left(\frac{-nT^2\mu^3}{2\alpha\sigma^2}\right)\left(1 + C_1T + O(T^2)\right) \tag{52}$$

$T = O(\frac{1}{\sqrt{n}})$ and 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 concluded that X_t is locally around τ like a Brownian motion. In our case, $\alpha/\mu = 1$, $\sigma = 1$ and the crossing time $X_\tau = x$ and T are related as $x = -c\sqrt{n}T$. Hence, by (52) of Lemma 4 we see that the density $f(x)$ of the crossing value $X_\tau = x$ is given by

$$f(x) \sim \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}\left(1 + C_2\frac{x}{\sqrt{n}}\right) \tag{53}$$

Our goal is to estimate $\int f(x)E\{Z_\tau(k)|X_\tau = x\}dx$. First of all, we observe that Theorem 1 and (40) imply for every k

$$\int_{-\infty}^{\infty} \frac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}} E\{Z_1(k)|X_1 = y\}dy = 0 \ . \tag{54}$$

Moreover, the following two estimates are easy to establish:

$$Z_\tau(k) = Z_1(k) + O(n|T|/\log n) = Z_1(k) + O(\sqrt{n}x/\log n) \ , \tag{55}$$

and $y = X_1$ becomes

$$y = X_\tau + \sqrt{|T|}\xi = x + C\frac{\sqrt{|x|}}{n^{1/4}}\xi \ , \tag{56}$$

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 (53), the density of $f(x)$ 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 (54) and the above

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)E\{Z_\tau(k)|X_\tau = x\} \ &\sim \ \int_{-\infty}^{\infty} \psi(y)\left(E\{Z_1(k)|X_1 = y\} + O(\sqrt{n}y/\log n)\right)dy \\ &= \ O(\sqrt{n}/\log n) \ .\end{aligned} \tag{57}$$

This completes the proof of (40) since $\overline{B}_{t_0} = O(n/\log^{3/2} n)$, as noticed in (51).

B. FINISHING THE PROOF

We first consider the asymmetric alphabet, and prove part (ii) of Theorem 2. From Theorem 1 (cf. (14) and (39)) we conclude that for some real θ

$$E\left(e^{i\theta D_m}\right) = \exp\left(i\theta c_1 \log m - (1/2)\theta^2 c_2 \log m\right)\left(1 + O\left(\theta/\sqrt{\log m}\right)\right) , \qquad (58)$$

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

$$F(\theta) = E\left\{\exp\left(i\theta c_1 \log M_n/f(n) - (1/2)\theta^2 c_2 \log M_n/f^2(n)\right)(1 + O(\theta/\sqrt{\log M_n})\right\} . \qquad (59)$$

Let $\xi_n := (M_n - EM_n)/\sqrt{varM_n}$, then by our Conjecture $Ee^{\vartheta\xi_n/g(n)} \to \exp(\vartheta^2/(2g^2(n)))$ for some complex ϑ and real-valued function $g(n)$. 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 $varM_n \sim cn/\log^2 n$. Therefore, from (58), (59) and the above one obtains

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

for some complex ϑ, where $g^{-1}(n) = f^{-1}(n)(1+O(1/f(n))$. But, according to (51) $O(\sqrt{n}) \le M_n \le O(n/\log_2 n)$, hence $\log(1 + \xi_n c/\sqrt{n}) = O(1)$. 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

$$e^{-i\theta c_1\sqrt{\log(nh/\log n)}}F(\theta) = e^{-c_2\theta^2/2}(1 + O(1/\sqrt{\log n})) , \qquad (60)$$

which completes the proof of part (ii).

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. (16)) 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 $M_n = m$. The only term needed to be investigated is $2^x e^{-2^{-x-1-i}}$ where $x = j - \log_2 M_n$. Note that for such x it 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]

$$M_n = \frac{n}{\log_2 n} + \xi_n O\left(\sqrt{n/\log_2^3 n}\right) = \frac{n}{\log_2 n}\left(1 + \xi_n O(1/\sqrt{n\log n})\right) \qquad (61)$$

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

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

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

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

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

[9] P. Jacquet and M. Régnier, Normal Limiting Distribution for the Size and the External Path Length of Tries, INRIA TR-827, (1988).

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

[11] P. Jacquet and W. Szpankowski, On the Lempel-Ziv Parsing Algorithm and Its Digital Tree Representation, INRIA Rapports de Recherche, 1833, (1992).

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

[13] 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).

[14] P. Kirschenhofer, H. Prodinger and W. Szpankowski, Digital Search Trees Again Revisited: The Internal Path Length Perspective, *SIAM J. Computing*, to appear.

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

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

[17] A. Lempel and J. Ziv, On the Complexity of Finite Sequences, *IEEE Information Theory* 22, 1, 75-81 (1976).

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

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

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

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

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

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

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

[25] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, to appear.

[26] W. Szpankowski, Asymptotic Properties of Data Compression and Suffix Trees, *IEEE Trans. Information Theory*, to appear.

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

[28] 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).

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

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

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

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

Unité de Recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399