A Functional Equation Arising in the Analysis of Algorithms

(Extended Abstract)

Philippe Jacquet*
jacquet@blagny.inria.fr

Wojciech Szpankowski†
spa@cs.purdue.edu

Abstract

We consider a functional-differential equation of the form $\alpha h'_z(z,u)+\beta h(z,u) = h(zup,u)\cdot h(zuq,u)+a(z,u)$ with $h(0,u)=1$ where $p+q=1$, α,β are nonnegative constants, $h'_z(\cdot,\cdot)$ denotes the partial derivative with respect to z, and u and z are complex numbers. This equation arises in numerous problems in combinatorics, computer science, data compression and molecular biology (e.g., complexity of the quicksort, number of phrases in the Lempel-Ziv parsing algorithm, enumeration of parsings of given length, external/internal path length in digital search trees, etc.). We obtain an asymptotic solution of this equation for large z, and apply it to settle some open problems in the area of analysis of algorithms. In particular, we solve a problem posed by Aldous and Shields concerning the limiting distribution of the number of phrases in the fundamental parsing scheme due to Lempel and Ziv. Actually, we reduce this problem to another one, namely, finding the limiting distribution of the internal path length in a digital search tree. In addition, we can enumerate parsings of given length, obtain limiting distribution of the external path lengths of other digital trees such as tries and Patricia tries, and so forth.

1. Introduction

The initial motivation for this work was the desire to understand the asymptotic behavior of the fundamental parsing algorithm on words due to Lempel and Ziv [17, 25, 26]. This scheme partitions a word into phrases (blocks) of variable sizes such that a new block is the shortest subword not seen in the past as a phrase. For example, the string 110010100010001000 is parsed into (1)(10)(0)(101)(00)(01)(000)(100).

This parsing algorithm plays a crucial rôle in universal data compression schemes and its numerous applications such as efficient transmission of data (cf. [17], [26]), discriminating between information sources (cf. [9], [25]), test of randomness (cf. [25]), estimating the statistical model of individual sequences (cf. [25]), and so forth. More importantly, such a partition of a word reflects its underlying periodicities, and therefore finds applications in several algorithms on words [2, 4, 8, 9, 23, 24]. 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.

By all means the most important parameter is the number of phrases which is studied by us in a probabilistic framework. We assume that a word of length n is generated by a probabilistic memoryless binary source. That is: *symbols are generated in an independent manner with "0" and "1" occurring respectively with probability p and $q=1-p$.* If $p=q=0.5$, then such a probabilistic model will be further called the *symmetric Bernoulli* model; otherwise we refer to the *asymmetric Bernoulli* model. If the length of a word is a random variable distributed as Poisson, then we refer to the *Poisson* model. Actually, our methodology should also handle Markovian models where some dependency among consecutive symbols is allowed.

In view of the above, the problem of finding the limiting distribution of the number of phrases in the Lempel-Ziv parsing algorithm is of a fundamental nature. (For example, if M_n represents the number of phrases constructed from a word of length n, then $(M_n+1)\log M_n$ is the length of compressed code in one of the Lempel-Ziv compression algorithm [5].) It is not an easy problem as attested by Aldous and Shields [1] who solve it only for the symmetric Bernoulli model. The authors of [1] write: "It is natural to conjecture that asymptotic normality holds for a larger class of processes ... . But in

*INRIA, Rocquencourt, 78153 Le Chesnay Cedex, France.
†Department of Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A. 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 Collaborative Grant 0057/89.

view of the difficulty of even the simplest case (i.e., the fair coin-tossing case we treat here) we are not optimistic about finding a general result. We believe the difficulty of our normality result is intrinsic ... ." In this paper, we extend the result of Aldous and Shields to the asymmetric Bernoulli model, and we also strengthen some of their results for the symmetric model. Actually, we do much more, and provide solution to some other problems that can be reduced to our basic functional equation; e.g., internal/external path lengths in digital trees (e.g., tries, Patricia and digital search trees), number of parsings of given length, etc.

The Lempel-Ziv parsing algorithm can be alternatively described on a *digital search tree* (cf. [1, 12, 16]). For the reader convenience, we show in Figure 1 the digital search tree associated with the word discussed at the beginning of this section. In particular, the root of the tree is empty (i.e., we start with an empty phrase). All other phrases of the Lempel-Ziv parsing algorithm 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, that is, symbol "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. In passing, we note that for a word of fixed length n the size of the associated digital search tree is *random* which gives a new twist to the analysis (cf. [18]).

Concerning our main results, we first present a simple combinatorial relationship (cf. (1)) that expresses the distribution of the number of phrases in the Lempel-Ziv algorithm in terms of the distribution of the internal path length in the associated digital search tree built from *fixed* number of independent words. We observe that the generating function of the internal path length in such a tree satisfies our functional equation (with $b = 0$ and $a(z, u) \equiv 0$). Hence, an asymptotic solution of such a equation leads to the limiting distribution of the internal path length, and in a consequence to the limiting distribution of the number of phrases.

There are other applications of our functional equation that are of interest to computer science and combinatorics. In particular, we shall obtain the limiting distributions of external path lengths for two other digital trees, namely tries and Patricia tries that are of prime interest to searching algorithms and algorithms on words (cf. [16, 19, 23, 24]). Interestingly enough, we can also enumerate the associated digital search trees with given internal path length. The last problem is motived by the question posed by Gilbert and Kadota [9] who asked for the number of parsings of given length in the Lempel-Ziv algorithm.

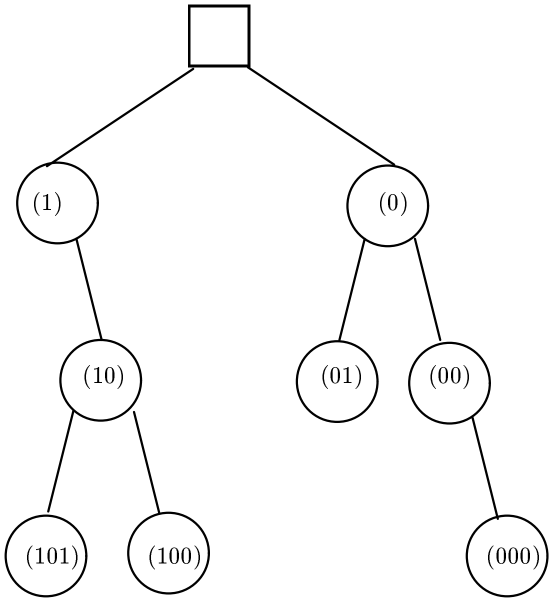

Figure 1: A digital tree representation of Ziv's parsing for the string 11001010001000100...

There is little we can say about literature on these problems. Second-order properties (e.g., limiting distributions) of the Lempel-Ziv algorithm were scarcely discussed with a noble exception of the work of Aldous and Shields [1] who studied only the symmetric model. External (internal) path lengths in digital trees are of interest to many algorithms on words but only recently variances of these quantities were obtained for symmetric alphabet by Kirschenhofer, Prodinger and Szpankowski [13, 14, 15], and the authors of [13, 14, 15] could not extend their methodologies to the asymmetric alphabet. We not only give the asymptotics of the variances for the asymmetric model, but we also complete the picture by deriving the limiting distributions.

2. Main Results

We first concentrate on the motivating example, namely the Lempel-Ziv parsing scheme. Let M_n be the number of phrases generated by such an algorithm from a word of length n. We also consider a digital search tree built from *fixed* number of, say m, statistically *independent* words (the so called independent tree), and we denote by $D_m(i)$ the length of the path from the root to the ith node, which is also the length of the ith phrase in the Lempel-Ziv scheme in the case when the parsing is built from m words. Then, the internal path length L_m becomes $L_m = \sum_{i=1}^m D_m(i)$. Below, we present an argument that allows to study properties of M_n through L_m.

Hereafter, we shall consistently use n as the length of a single word to be parsed, and m as the number of words used to construct an independent digital search tree.

In view of the above definitions and our discussion in the previous section (cf. Fig. 1), we note that M_n can be expressed in terms of L_m as

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

which directly implies that

$$\Pr\{M_n > m\} = \Pr\{L_m \le n\} \;. \qquad (2)$$

Indeed, consider building a dynamic digital search tree from phrases. Each time a phrase is created we add it as a new word to the digital tree. The tree grows, and we continue this process until for the first time the internal path length becomes n. Clearly, the number of inserted words at this time is M_n (cf. [9, 12, 18]).

Relationship (1) is known as the *renewal equation* (cf. Billingsley [3], Chap. 17). Theorem 17.3 of Billingsley [3] proves (with some trivial modifications) that if

$$\frac{L_m - \mu_m}{\sigma_m} \to N(0,1) \;, \qquad (3)$$

then

$$\frac{M_n - n/(\mu_n/n)}{\sigma_n(\mu_n/n)^{-3/2}} \to N(0,1) \qquad (4)$$

where $N(0,1)$ is the standard normal distribution, and μ_m and σ_m are positive constants that under mild standard uniform integrability arguments can be asymptotically interpreted as the mean and the variance of L_m.

In summary, we have reduced the evaluation of the limiting distribution of M_n to the estimation of the limiting distribution of the internal path length L_m in a digital search tree built from *fixed* number of *independent* words.

Hereafter, we study digital search trees and evaluate L_m. Define $L_m(u) = Eu^{L_m}$ and $L(z,u) = \sum_{m=0}^{\infty} L_m(u)\frac{z^m}{m!}$. Moreover, we set $\widetilde{L}(z,u) = L(z,u)e^{-z}$ which can be interpreted as the generating function of the internal path length in a family of digital search trees built from number of words that is distributed as Poisson with mean z.

There is a recurrence relationship on $L_m(u)$ and $L(z,u)$. Indeed, noting that left and right subtrees are characterized by $L_k(u)$ and $L_{m+1-k}(u)$, where k is the number of words with the first symbol being "0", we obtain

$$L_{m+1}(u) = u^m \sum_{k=0}^{m} \binom{m}{k} p^k q^{m-k} L_k(u) L_{m-k}(u) \;. \qquad (5)$$

with $L_0(u) = 1$. Hence, also

$$\frac{\partial L(z,u)}{\partial z} = L(pzu,u)L(qzu,u) \qquad (6)$$

with $L(z,0) = 1$. The above equation is the one mentioned in the abstract with $a(z) \equiv 0$ and $\alpha = 1,\ \beta = 0$.

Now, we can summarize our main results. All of them are based on an asymptotic solution of the functional equation (6). The first result deals with the limiting distribution of the internal path length L_m in a digital search tree built from m independent words. It proves an open problem of [1] and [15] (cf. [19]).

Theorem 1. *Consider a digital search tree under the* **asymmetric** *Bernoulli model.*

(i) *Asymptotically the average value EL_m and the variance* Var L_m *become*

$$\begin{aligned} EL_m &= \frac{m}{h}\left(\log m + \frac{h_2}{2h} + \gamma - 1 - \alpha + \delta_0(\log m)\right) \\ &+ \frac{1}{h}\left(\log m + \frac{h_2}{2h} - \gamma - \log p - \log q + \alpha\right) \\ &+ O(1) \qquad (7) \\ \mathrm{Var}\ L_m &= c_2 m\log m + O(m) \qquad (8) \end{aligned}$$

where $h = -p\log p - q\log q$ is the entropy of the alphabet, $\gamma = 0.571\ldots$ is the Euler constant, $h_2 = p\log^2 p + q\log^2 q$, and $c_2 = (h_2 - h^2)/h^3$,

$$\alpha = -\sum_{k=1}^{\infty} \frac{p^{k+1}\log p + q^{k+1}\log q}{1 - p^{k+1} - q^{k+1}} \;, \qquad (9)$$

and for $\log p/\log q$ rational $\delta_0(\log m)$ is fluctuating functions with small amplitude, while for $\log p/\log q$ irrational $\delta_0(x) \equiv 0$ (cf. [10, 11, 22]).

(ii) *Let $c_1 = 1/h$. Then*

$$\frac{L_m - EL_m}{\sqrt{\mathrm{Var}\ L_m}} \to N(0,1) \qquad (10)$$

where $N(0,1)$ is the standard normal distribution. In fact, a stronger result holds, namely for a complex ϑ in a neighbourhood of zero, and for any $\varepsilon > 0$

$$e^{-\vartheta c_1 m \log m} E e^{\vartheta L_m} = e^{c_2 \frac{\vartheta^2}{2} m \log m} \left(1 + O(1/m^{1/2+\varepsilon})\right) \tag{11}$$

(iii) *There exist positive constants A and $\mu < 1$ such that uniformly in k*

$$\Pr\left\{ \left| \frac{L_m - EL_m}{\sqrt{\mathrm{Var}\ L_m}} \right| > k \right\} < A\mu^k \tag{12}$$

for large m.

(iv) *For the* **symmetric** *Bernoulli model the following equivalence of (11) holds*

$$\begin{aligned} e^{-\vartheta m \log_2 m} E e^{\vartheta L_m^{sym}} &= e^{\frac{\vartheta^2}{2} m (C + \delta(\log_2 m))} \\ &\cdot \left(1 + O(1/m^{1/2+\varepsilon})\right) \end{aligned} \tag{13}$$

for any $\varepsilon > 0$, where $\delta(x)$ is a periodic continuous function of period 1, mean 0 and very small amplitude ($< 10^{-6}$). Similarly, uniformly in k

$$\Pr\left\{ \left| \frac{L_m^{sym} - EL_m^{sym}}{\sqrt{m(C + \delta(\log_2 m))}} \right| > k \right\} < A\mu^k \tag{14}$$

with $\mu < 1$. ■

A sketch of the proof for Theorem 1 is presented in Section 3. Here, we only outline its main ingredients for the reader convenience:

- We first analyze the Poisson model that is characterized by the exponential bivariate generating function $L(z,u)$ satisfying (6).
- In order to solve (6) we try to transform it into an additive functional equation by considering $\log L(z,u)$. This is only possible if one proves the existence of $\log L(z,u)$. We prove that there is a convex cone around the real axes such that for some $\kappa(u)$ we have $\log L(z,u) = \Theta(z^{\kappa(u)})$ (cf. Theorem 5).
- Next, we use Taylor expansion of $\log L(z,u)$ in the convex cone to show that for large z the generating function $L(z,u)$ appropriately normalized converges to the generating function of the normal distribution (cf. Theorem 6).
- To prove the above we must compute precise asymptotic expansion for the average and the variance of the internal path length in the Poisson model (Theorem 6).
- The final effort is to de-Poissonize the above result, that is, to transform the normal distribution of the Poisson model into the normal distribution of the Bernoulli model (cf. Theorem 7).

Our next finding concerns the number of phrases in the Lempel-Ziv parsing algorithm which was our main motivation for this work. Using our fundamental relationship (1) and the renewal theorem of Billingsley (cf. [3]) we directly obtain the following result that proves the problem posed by Aldous and Shields (cf. [12]).

Before we formulate it, we need some additional notation. Define $\Lambda(m) = EL_m$ for all integer m, and let for all $y \geq 0$, $\Lambda(y)$ be a linear interpolation of $\Lambda(m)$ between integer points. (Actually, we could define $\Lambda(m)$ as an approximation of EL_m only up to $O(m)$ term in (7).) Now, let $\Lambda^{-1}(x)$ be the inverse function of $\Lambda(y)$ defined for $x \geq 0$. Note that $\Lambda(x) \sim \frac{x \log x}{h}$ and $\Lambda^{-1}(x) \sim \frac{hx}{\log x}$ for $x \to \infty$. As easy to check, the function $\Lambda(\cdot)$ is convex, hence $\Lambda^{-1}(\cdot)$ is concave (this should be at least clear for $\Lambda(x)$ approximated by the first two terms in (7) which are convex functions).

Theorem 2. (i) *In the asymmetric Bernoulli model, the following weak convergence holds*

$$\frac{M_n - EM_n}{\sqrt{\mathrm{Var} M_n}} \to N(0,1) \tag{15}$$

with

$$EM_n \sim \frac{nh}{\log(n)} \tag{16}$$

$$\mathrm{Var}\ M_n \sim \frac{c_2 h^3 n}{\log^2 n} \tag{17}$$

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

(ii) *There exist positive real numbers A, α and an integer $n_0 \geq 1$ such that for all $n > n_0$, and for real number $x \geq 0$:*

$$\begin{aligned} \Pr\{M_n \geq (1 + \frac{x}{\sqrt{n}})\Lambda^{-1}(n)\} &\leq A \exp\left(-\alpha \frac{x}{\sqrt{1+x}}\right) \\ \Pr\{M_n < (1 - \frac{x}{\sqrt{n}})\Lambda^{-1}(n)\} &\leq A \exp(-\alpha x) \end{aligned}$$

Thus, for all $r \geq 0$ the random variable $Z_n = (M_n - \Lambda^{-1}(n))/\sqrt{c_2 h^3 n / \log^2 n}$ is r-uniform integrable, and therefore all moments of Z_n exist and converge to appropriate moments of the normal distribution. In particular, for any (fixed) $\varepsilon > 0$ and large n we have

$$\Pr\left\{ M_n > (1+\varepsilon)\frac{nh}{\log n}\left(1 + O\left(\frac{\log\log n}{\log n}\right)\right)\right\} \leq$$

$$\leq A \exp\left(-\frac{a\varepsilon}{c_2\sqrt{h(1-\varepsilon)}}\sqrt{n}\right)$$

for some constants $A, a > 0$.

(iii) *The above results are also true for the symmetric model if one replaces the variance by*

$$\text{Var } M_n^{sym} \sim \frac{n(C+\delta(\log_2 n))}{\log_2^3 n} \quad (18)$$

where $C = 0.26600\ldots$ and $\delta(x)$ is defined above. In (ii) one must replace $\sqrt{n}$ by $\sqrt{n/\log n}$. ■

Our functional equation has several other applications. We mention here only few. Consider first the combinatorial problem posed by Gilbert and Kadota [9]: How many parsings of total length n one can construct from m words? For example, for $m = 2$ we have four parsings of length three, namely: (0)(00), (0)(01), (1)(10) and (1)(11), and two parsings of length two, namely: (0)(1) and (1)(0).

Let $F_m(n)$ be the number of parsings built from m words of total length n, and let $F_m(x) = \sum_{n=0}^{\infty} F_m(n)x^n$ be its generating function. Note that (cf. [9])

$$F_{m+1}(x) = x^m \sum_{k=0}^{m} \binom{m}{k} F_k(x) F_{m-k}(x). \quad (19)$$

The result below provides an asymptotic solution to the problem of Gilbert and Kadota [9].

Theorem 3. *The number of parsings built from m words of total length n is*

$$F_m(n) = 2^n \Pr\{L_m^{sym} = n\} \ . \quad (20)$$

In particular, for $n = m\log_2 m + O(\sqrt{m})$ we obtain asymptotically

$$F_m(n) \sim \frac{2^n}{\sqrt{2\pi(C+\delta(\log_2 n))m}} \cdot \exp\left(-\frac{(n-m\log_2 m)^2}{2(C+\delta(\log_2 n))m}\right)$$

where C and $\delta(x)$ are defined in Theorem 1.

Proof. Consider the recurrence (19) and note that $F_m(x/2) = L_m^{sym}(x)$. The rest follows from Theorem 1 and the analysis of Section 3. ■

Finally, we consider the external path lengths in two other digital trees, namely tries and Patricia tries. As discussed in Knuth [16] and Mahmoud [19] these trees are widely applied in sorting and searching algorithms as well as in conflict resolution algorithms for broadcast communications. Let $L_m^T(u)$ and $L_m^P(u)$ be the probability generating functions for the internal path lengths in regular tries and Patricia tries. It is easy to check (cf. [13, 14, 16, 19]) that the bivariate generating functions $L^T(z,u)$ and $\bar{L}^P(z,u) = L^P(z,u) - 1$ satisfy the following function equations

$$\begin{aligned} L^T(z,u) &= L^T(zup,u)L^T(zuq,u) \\ &+ z(u-1) \end{aligned} \quad (21)$$

$$\begin{aligned} \bar{L}^P(z,u) &= \bar{L}^P(zup,u)\bar{L}^P(zuq,u) \\ &+ \bar{L}^P(zq,u) + \bar{L}^P(zp,u) \end{aligned} \quad (22)$$

that fall under our general functional equation defined in the abstract (see also (25)).

Theorem 4. *For the asymmetric alphabet we have*

$$\frac{L_m^T - EL_m^T}{\sqrt{c_2 m \log m}} \to N(0,1) \quad (23)$$

$$\frac{L_m^P - EL_m^P}{\sqrt{c_2 m \log m}} \to N(0,1) \ , \quad (24)$$

where

$$EL_m^T = \frac{m}{h}\log m + \frac{m}{h}\left(\gamma + \frac{h_2}{2h}\right) + O(1) \ ,$$

$$EL_m^P = \frac{m}{h}\log m + \frac{m}{h}\left(\gamma + \frac{h_2}{2h} - \rho\right) + O(1) \ ,$$

with $\rho = -p\log(1-p) - q\log(1-q)$. In the symmetric alphabet, the above is true if one replaces the variances (i.e., $c_2 m \log m$) by var $L_m^T \sim (C^T + \delta(\log_2 m))m$ and var $L_m^T \sim (C^T + \delta(\log_2 m))m$ where $C^T = 4.35\ldots$ (cf. [13]) and $C^P = 0.37\ldots$ (cf. [14]). As before, $\delta(x)$ denotes a fluctuating function with mean zero. ■

This completes the list of our main results. We shall delay the discussion of further consequences of our findings to the final version of the paper (cf. [12]).

3. Analysis of the Functional Equation

In this section we provide an asymptotic solution of the functional equation (6).

To make our discussion general, let L_m be a discrete random variable depending of a large parameter m (think of L_m as the internal path length in a digital tree built from m words). Define $h(u) = Eu^{L_m} = \sum_{k=0}^{\infty} \Pr\{L_m = k\}u^k$ and $h(z,u) = \sum_{m=0}^{\infty} h(u)\frac{z^m}{m!}$ as the probability generating function and the exponential generating function, respectively.[1] We also

[1] We switch to $h(u)$ and $h(u,z)$ to underline the fact that our analysis is not restricted to the internal path length in a digital tree.

work with the so called Poisson generating function defined as $\widetilde{h}(z,u) = h(z,u)e^{-z}$. The latter could be interpreted as the probability generating function of L_m when m is distributed according to Poisson with parameter z (i.e., the Poisson model).

We assume that $h(z,u)$ satisfies the following differential-functional equation

$$\alpha\frac{\partial h(z,u)}{\partial z} + \beta h(z,u) = h(zup,u)h(zuq,u) + a(z,u) \tag{25}$$

where α, β are constants, and $p+q = 1$, with $h(z,1) = e^z$, $h(0,u) = 1$, and $a(z,u) = e^{O(z)}$. It turns out that a methodology of solving the following differential-functional equation

$$\frac{\partial h(z,u)}{\partial z} = h(zup,u)h(zuq,u)\ . \tag{26}$$

does not differ significantly from what one needs to solve (25). In this conference version of the paper, we only investigate (26). Note that our basic functional equation (6) for the internal path length in a digital search tree satisfies exactly (26).

The idea of the proof is to take the logarithm of the multiplicative equation (26) and transform it into an additive equation, which is much easier to handle (cf. [6, 7, 16, 19, 22]). The problem is that one must find a neighborhood $\mathcal{U}(1)$ of $u = 1$ such that the logarithm of $h(z,u)$ exists for z near real positive axes. This turns out to be a non-trivial problem (cf. Theorem 5 below).

To formulate the next result we need additional notation. The domain of z we shall work with is a *convex cone* $C(D,\delta)$ defined as the set of complex numbers $z = x + iy$ such that $x \geq 0$ and $|y| \leq Dx^\delta$ where $\delta \geq 0$ and $D \geq 0$. In our analysis we will have $\delta < 1$ which implies a convex cone (cf. Figure 2).

Furthermore, we denote by κ a solution of the following equation

$$(pu)^\kappa + (qu)^\kappa = 1\ . \tag{27}$$

In fact, κ is a function of u, and if we need to emphasis this we shall write $\kappa(u)$. Note that $\kappa(u) = 1 + (u-1)h + O((u-1)^2)$. We have the following technical result that is crucial to our analysis.

Theorem 5. *There exists a convex cone $\mathcal{C}(D,\delta)$, a constant ξ, and a neighborhood $\mathcal{U}(1)$ of $u = 1$ such that for $|z| > \xi$ the logarithm $\log h(z,u)$ exists, and $\log h(z,u) = \Theta(z^\kappa)$ uniformly in $(z,u) \in \mathcal{C}(D,\delta)\times\mathcal{U}(1)$, where κ is defined in (27). Moreover, all derivatives with respect to u of $\log h(z,u)$ exist and are of order $\Theta(z^{\kappa+\beta})$ for any $\beta > 0$.*

Proof. The proof is quite involved and details can be found in [12]. Here we provide only a sketch for z real.

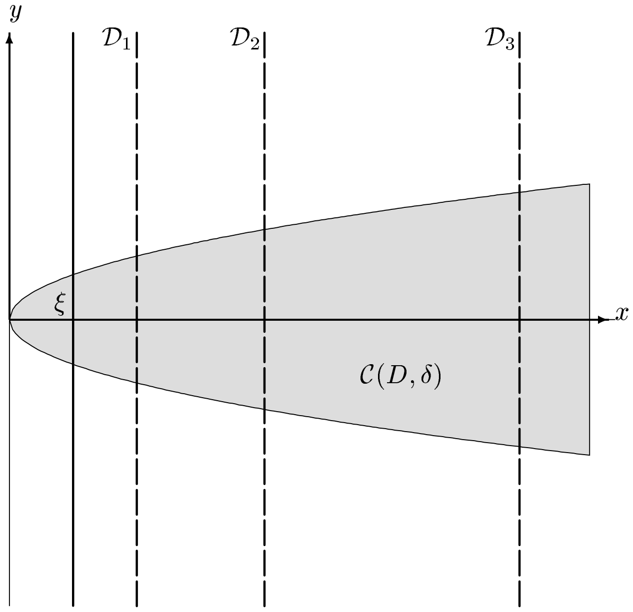

Figure 2: Illustration to the convex cone $\mathcal{C}(D,\delta)$ and the increasing domains $\mathcal{D}_m$.

It turns out that the original equation is too complicated, and therefore we introduce another one. Define $f(z) = h(z,u)/h'(z,u)$, then it fulfills the following differential-functional equation[2]

$$f'(z) = 1 - \left(\frac{pu}{f(puz)} + \frac{qu}{f(quz)}\right) f(z) \tag{28}$$

with $f(0,u) = 1$ and $f(z,1) = 1$. The function $f(z)$ has only polynomial growth, and is more easier to handle. Observe that $\log h(z,u) = \int_0^z f^{-1}(x)\mathrm{d}x$.

Thus, to prove a bound on $\log h(z,u)$ we need a lower and an upper bound on $f(z)$. In fact, it suffices to prove for large z that $a|z^{1-\kappa(u)}| \leq |f(z,u)| \leq A|z^{1-\kappa(u)}|$ for some a, A (which are in fact a function of u). The latter is proved in two steps. Namely, we first prove that for real z, say x, the above bounds hold, and in addition we show that for complex z there is $\delta < 1$ such that for all $\beta > 0$ the derivative of $f(z)$ with respect to z behaves like $|f'(z)| \leq B(u)|z|^{-\kappa(u)+\beta}$ for $|z| > \xi$. Observe that these two facts proves our theorem since $f(z) = f(\Re(z)) + \int_{\Re(z)}^z f'(x)\mathrm{d}x$, thus $f(z) - f(\Re(z)) = \Theta(z^{-\kappa(u)+\delta+\beta})$ which suffices for the proof.

In this conference version we only establish an upper bound on $f(z)$ for z real. That is, we prove that for large real x (e.g., $x > \xi$) the following holds

$$|f(x)| \leq Ax^{1-\kappa}\ .$$

[2] The function $f(z)$ clearly depends on u, but to avoid cumbersome notation, we write $f(z)$ whenever it is possible.

We carry out the proof by induction over the increasing domains $\mathcal{D}_m$ defined as $\mathcal{D}_m = [\xi, \nu^{-m}]$ for some $\xi > 0$, $\max\{p,q\} \le \nu < 1$, and a nonnegative integer m as shown in Figure 2.

Let A_m be an upper bound for $f(x)x^{\kappa-1}$ on the domain $\mathcal{D}_m$ (restricted to real line). If we prove that A_m is bounded (with respect to m), then we establish our result. Let now x denote a real number belonging to $\mathcal{D}_{m+1} - \mathcal{D}_m$. Note that then xp and xq belong to $\mathcal{D}_m$, hence by induction assumption, $f(pux) \le A_m(pux)^{1-\kappa}$ and $f(qux) \le A_m(qux)^{1-\kappa}$. By (28) we have $f'(x) \le 1 - x^{\kappa-1}A_m^{-1}f(x)$. It is easy to solve this equation (cf. [12]) and one obtains

$$\begin{aligned} f(x) &\le f(\nu x)\exp\left(-\frac{(1-\nu^\kappa)x^\kappa}{\kappa A_m}\right) \\ &+ \int_{\nu x}^{x} \exp\left(\frac{y^\kappa - x^\kappa}{\kappa A_m}\right) \mathrm{d}y\ . \end{aligned}$$

Multiplying the above by $x^{\kappa-1}$, we obtain $A_{m+1} \le \max\{A_m, A'_{m+1}\}$ where

$$\begin{aligned} A'_{m+1} &\le A_m \max_{x\in[\nu^{-m},\nu^{-m-1}]} \left\{ \int_{\nu x}^{x} \exp\left(\frac{y^\kappa - x^\kappa}{\kappa A_m}\right) \mathrm{d}y \right. \\ &+ \left. \nu^{-\kappa} \exp\left(-\frac{(1-\nu^\kappa)x^\kappa}{\kappa A_m}\right)\right\} \end{aligned}$$

We now proceed in two steps. First, we let $v_m = A_m\nu^{\kappa m}$, and prove that v_m exponentially decays to zero. Secondly, we prove that A_m is an increasing sequence uniformly bounded from the above.

It is easy to prove that (cf. [12, 21])

$$\begin{aligned} \int_{\nu x}^{x} \exp\left(\frac{y^\kappa - x^\kappa}{\kappa}\right) \mathrm{d}y &= x^{1-\kappa}\left(1 - e^{-x^\kappa(1-\nu^\kappa)}\right) \\ &\cdot\ (1 + O(1/x)) = x^{1-\kappa}\eta(x) \end{aligned}$$

where $\eta(x) = 1 + O(1/x)$. Thus

$$\begin{aligned} A_{m+1} &\le \max\left\{A_m, A_m\nu^{-\kappa}\right. \qquad (29) \\ &\cdot\ \left.\exp\left(-\frac{\nu^{-\kappa m}}{A_m\kappa}\right) + A_m\eta\left(\frac{\nu^{-m}}{A_m^{1/\kappa}}\right)\right\} \end{aligned}$$

which can be reduced to

$$v_{m+1} \le v_m F(u, v_m)$$

where we have $F(u,v_m) = \max\{\nu^\kappa, \exp(-(v_m\kappa)^{-1}) + \nu^\kappa\eta(v_m^{-1/\kappa})\}$. Clearly, there exists m_0 such that $F(1, v_{m_0}) < 1$ for $m > m_0$, which implies exponential decay of v_m. To see this it is enough to observe that $F(x,0) = \nu^\kappa < 1$ and $F(\cdot,\cdot)$ is a continuous function (for $u = 1$ one also has $A_m = 1$, as desired).

Now we return to (29) to get

$$\begin{aligned} A_{m+1} &\le A_m \max\left\{1, \left(\nu^{-\kappa}\exp(-(v_m\kappa)^{-1})\right.\right. \\ &+ \left.\left.\eta(v_m^{-1/\kappa})\right)\right\}\ , \end{aligned}$$

and then $A_{m+1} \le (1 + O(v_m^{1/\kappa}))A_m$. Thus

$$A_{m+1} \le A_0 \prod_{j=0}^{m}\left(1 + O(v_j^{1/\kappa})\right) < \infty\ , \tag{30}$$

which proves that A_m are uniformly bounded by A_∞. Clearly, as $\mathcal{U}(1)$ becomes smaller and smaller (i.e., $u \to 1$) the constant $A_\infty(u)$ tends to 1. This completes the proof. ■

To formulate our main results, we define the so called Poisson mean and variance respectively as $\widetilde{X}(z) = h'_u(z,1)$ and $\widetilde{V}(z) = h''_{uu}(z,1) + h'_u(z,1) - (h'_u(z,1))^2$ where $h'_u(z,1)$ and $h''_{uu}(z,1)$ denote the first and the second derivative of $h(z,u)$ with respect to u at $u = 1$. The reader can notice that $\widetilde{X}(z)$ and $\widetilde{V}(z)$ are the mean and the variance of our generic random variable L_m in the Poisson model.

Our main finding for the Poisson model can be formulated as follows.

Theorem 6. *In the Poisson model for $(z,u) \in \mathcal{C}[D,\delta]\times\mathcal{U}(1)$ the following holds for any $\beta > 0$ and $z \to \infty$ (with $u = e^t$ for some real t in the vicinity of zero)*

$$\log h(z,e^t) = z + \widetilde{X}(z)t + \widetilde{V}(z)\frac{t^2}{2} + O(t^3 z^{\kappa(u)+\beta})\ , \tag{31}$$

where the mean $\widetilde{X}(z)$ and the variance $\widetilde{V}(z)$ become asymptotically

$$\begin{aligned} \widetilde{X}(z) &= \frac{z}{h}\log z + \frac{z}{h}\left(\gamma - 1 + \frac{h_2}{2h} - \alpha - \delta_1(\log z)\right) \\ &+ O(1)\ , \end{aligned} \tag{32}$$

and

$$\begin{aligned} \widetilde{V}(z) &= \frac{z\log^2 z}{h^2} \\ &+ \frac{2z\log z}{h^3}\left(\gamma h + h_2 - \frac{h^2}{2} - \alpha h - h\delta_1(\log z)\right) \\ &+ O(z)\ , \end{aligned} \tag{33}$$

where

$$\alpha = -\sum_{k=1}^{\infty} \frac{p^{k+1}\log p + q^{k+1}\log q}{1 - p^{k+1} - q^{k+1}}\ . \tag{34}$$

The function $\delta_1(x)$ is identically equal to zero when $\frac{\log p}{\log q}$ is irrational, while for $\frac{\log p}{\log q} = \frac{r}{t}$ where r,t are integers such that $\gcd(r,t) = 1$, the function $\delta_1(\log z)$ is fluctuating with a small amplitude as given by the formula below

$$\delta_1(\log z) = \sum_{\substack{\ell=-\infty \\ \ell\ne 0}}^{\infty} \frac{\Gamma(s_0^\ell)Q(-2)}{Q(s_0^\ell - 1)} \exp\left(-\frac{2\pi i\ell r}{\log p}\log z\right)\ , \tag{35}$$

where

$$Q(s)=\prod_{k=0}^{\infty}(1-p^{-s+k}+q^{-s+k})\ , \qquad (36)$$

and $s_0^\ell=-1+2\pi i\ell r/\log p$ *for* $\ell=\mathcal{Z}=\pm1,\pm2,\ldots$ *is a solution of* $1=p^{-s}+q^{-s}$.

Proof. Using Taylor's expansion of $\log h(z,u)$ we obtain

$$\begin{aligned}\log h(z,u) &= \log h(z,1)+(u-1)\frac{\partial}{\partial u}\log h(z,1)\\ &+ \frac{(u-1)^2}{2}\frac{\partial^2}{\partial u^2}\log h(z,1)+R(z,u)\end{aligned}$$

with $R(z,u)$ being the remaining term of the form

$$R(z,u)=\int_0^u\frac{(v-1)^3}{3!}\frac{\partial^3}{\partial v^3}\log h(z,v)\mathrm{d}v\ .$$

Due to Theorem 5 the error term is $O(z^{\kappa+\beta}t^3)$ after substitution $u=e^t$. Now it suffices to note that $\log h(z,1)=z$ and $\log\widetilde{h}(z,u)=\log h(z,u)-z$.

We must now estimate asymptotically the mean and the variance. Since we need several terms of such asymptotic expansions (to prove our main result concerning the Bernoulli model) we use the Mellin transform. The reader may familiarize himself with the technique from [19].

Consider first the the mean $\widetilde{X}(z)$. Direct differentiation of our basic equation (26) leads to the following recurrence

$$\widetilde{X}(z)+\widetilde{X}'(z)=\widetilde{X}(zp)+\widetilde{X}(zq)+z\ . \qquad (37)$$

Let $X(s)$ denote the Mellin transform of $\widetilde{X}(z)$, that is, $X(s)=\int_0^\infty\widetilde{X}(z)z^{s-1}dz$. It can be proved that it exists in the strip $\Re(s)\in(-2,-1)$. Observe that the Mellin transform of $\widetilde{X}'(z)-z$ is also defined in $\Re(s)\in(-2,-1)$. Then, (37) translates into

$$X(s)-(s-1)X(s-1)=p^{-s}X(s)+q^{-s}X(s)\ . \quad (38)$$

To solve the functional equation (38) we make a substitution $X(s)=\gamma(s)\Gamma(s)$ where $\Gamma(s)$ is the gamma function, and $\gamma(s)$ satisfies the following recurrence

$$\gamma(s)-\gamma(s-1)=p^{-s}\gamma(s)+q^{-s}\gamma(s)\ ,$$

which can be seen have the following solution

$$\gamma(s)=\prod_{k=0}^{\infty}\frac{1-p^{k+2}-q^{k+2}}{1-p^{-s+k}+q^{-s+k}} \qquad (39)$$

for $\Re(s)\in(-2,-1)$. Applying the Cauchy residue theorem to the above (i.e., inverting the Mellin transform) one proves (39).

The variance is more intricate, as already seen in [13, 14, 15]. We observe (after quite tedious algebra) that the variance $\widetilde{V}(z)$ satisfies the following recurrence

$$\begin{aligned}\widetilde{V}(z)+\widetilde{V}'(z) &= \widetilde{V}(zp)+\widetilde{V}(zq)+2zp\widetilde{X}'(zp)\\ &+ 2zq\widetilde{X}'(zq)+(\widetilde{X}'(z))^2\ . \qquad (40)\end{aligned}$$

This functional equation is harder to solve due to the last term for which there is no close form expression for the Mellin transform. But, fortunately, we can prove that for the *asymmetric* alphabet the last term contributes only $O(z)$ and we need terms up to $O(z)$ (to recover the leading terms in the variance in the Bernoulli model). Leaving the proof of latter fact to the final version of the paper (it is rather simple and one can apply the same line of arguments as in the proof of Theorem 5), we now concentrate on estimating the remaining part of the variance which we denote as $\widetilde{V}_1(z)$. Its Mellin transform $V_1(s)$ becomes

$$\begin{aligned}V_1(s)-(s-1)V_1(s-1) &= (p^{-s}+q^{-s})V_1(s)\\ &- 2(p^{-s}+q^{-s})sX(s)\ .\end{aligned}$$

This functional equation has the following solution

$$V_1(s)=-2\Gamma(s)\sum_{k=0}^{\infty}\frac{(p^{-s}+q^{-s})(s-k)\gamma(s-k)}{\prod_{m=0}^{k}(1-p^{-s+m}-q^{-s+m})}\ , \qquad (41)$$

where $\gamma(s)$ is given in (39). Inverting the above one finally obtains (33). This completes the proof of Theorem 6. ∎

The major remaining problem is to de-Poissonize Theorem 6, that is, to translate the results of Theorem 6 into the original Bernoulli model. There are several technical problems that we shall not address in this conference version. In short, we use the Cauchy integral to find $h_m(u)$ from its Poisson generating function $h(z,u)$, that is, formally for $u\in\mathcal{U}(1)$

$$h_m(u)=\frac{m!}{2\pi i}\oint\frac{\widetilde{h}(z,u)e^z}{z^{m+1}}dz \qquad (42)$$

where the integration is around the origin. In a similar manner, the average $X_m=EL_m$ and the variance $V_m=\mathrm{Var}L_m$ in the Bernoulli model can be computed from the Poisson model. We can prove the following translation formulas (cf. [11])

$$\begin{aligned}X_m &= \widetilde{X}(m)-m\widetilde{X}''(m)/2+O(1)\ , \qquad (43)\\ V_m &+ \widetilde{V}(m)-m(\widetilde{X}'(m))^2+m^2(\widetilde{X}''(m))^2\\ &+ O(1) \qquad (44)\end{aligned}$$

With the above tools, we obtain our last result that proves Theorem 1 from Section 2.

Theorem 7. *In the asymmetric Bernoulli model, the normalized random variable* $(L_m - X_m)/\sqrt{V_m}$ *tends to the standard normal distribution. More precisely, for a complex* ϑ *and any* $\varepsilon > 0$

$$e^{-\vartheta c_1 m \log m} h_m(e^\vartheta) = e^{c_2 \frac{\vartheta^2}{2} m \log m} \left(1 + O(1/m^{1/2+\varepsilon})\right) \quad (45)$$

where c_1 *and* c_2 *are defined in Theorem 1.*

Proof. The proof is based on the Cauchy formula (42). But, Theorem 5 provides only an asymptotic expansion for $h(z,u)$, hence we need a special depoissonization result. We use the one discussed in [11] and [20]. The technicalities here are quite involved, and details can be found in [12]. ■

ACKNOWLEDGEMENT

It is a pleasure to acknowledge several enlighting discussions that the authors had with P. Flajolet, INRIA, on the topic of this paper. We also in debt to D. Aldous, Berkeley, for pointing out to us the renewal equation of Billingsley.

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] A. Amir and G. Benson, Two-Dimensional Periodicity and Its Application, *Proc. 3-rd Symposium on Discrete Algorithms*, Orlando, 1992.

[3] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York 1968.

[4] D. Breslauer and Z. Galil, An Optimal $O(\log \log n)$ Time Parallel String Matching Algorithm, *SIAM J. Computing*, 19, 1051-1058, (1990).

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley&Sons, New York (1991).

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

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

[8] Z. Galil and K. Park, Truly Alphabet Independent Two-Dimensional Pattern Matching, *Proc. 33-rd IEEE Conference on Foundation of Computer Science*, Pittsburgh, 1992.

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

[10] P. Jacquet, *Contribution de l'Analyse d'Algorithmes a l'Evaluation de Protocoles de Communication*, Thèse Université de Paris Sud-Orsay, 1989.

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

[12] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, Purdue University, CSD-TR-088, 1993.

[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, On the Balance Property of Patricia Tries: External Path Length Viewpoint , *Theoretical Computer Science*, 68, 1989, pp. 1–17.

[15] P. Kirschenhofer, H. Prodinger and W. Szpankowski, Digital Search Trees Again Revisited: The Internal Path Length Perspective, *SIAM J. Computing*, 24, (1994).

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

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

[18] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, INRIA Rapport de Recherche, 1993 (also *IEEE Information Theory*, to appear).

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

[20] B. Rais, P. Jacquet and W. Szpankowski, A Limiting Distribution for the Depth in PATRICIA Tries, *SIAM J. Discrete Mathematics*, 6, 197-213 (1993).

[21] L. Sirovich, *Techniques of Asymptotic Analysis*, Springer-Verlag, New York 1971.

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

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

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

[25] 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).

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