Proceedings of 33-rd Allerton Conference, pp. 891-900, 1995

ON GENERALIZED DIGITAL SEARCH TREES WITH APPLICATIONS TO A GENERALIZED LEMPEL-ZIV ALGORITHM

Wojciech Szpankowski*
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Jing Tang
Department of Mathematics
Purdue University
W. Lafayette, IN 47907
U.S.A.
tangj@math.purdue.edu

Abstract

The goal of this research is twofold: (i) to analyze generalized digital search trees, and (ii) to derive the average profile (i.e., phrase length) of a generalization of the well known parsing algorithm due to Lempel and Ziv. In the generalized Lempel-Ziv parsing scheme, one partitions a sequence of symbols from a finite alphabet into phrases such that the new phrase is the longest substring seen in the past by at most $b-1$ phrases. The case $b=1$ corresponds to the original Lempel-Ziv scheme. As expected, such a scheme can be analyzed by considering a generalized digital search tree in which every node is capable of storing up to b strings. Such a digital tree is known as b-digital search tree. In this paper, we concentrate on investigating the depth of a randomly selected node in such a tree. We apply analytical techniques of analysis of algorithms to establish the mean, the variance and the limiting distribution of the depth. Based on these results, we infer the probabilistic behavior of a randomly selected phrase in the generalized Lempel-Ziv algorithm.

1. INTRODUCTION

The heart of several universal data compression schemes is the parsing algorithm due to Lempel and Ziv [16]. 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, 11001010001000100011 is parsed into (1)(10)(0)(101)(00)(01)(000)(100)(11). This scheme is used in the UNIX `compress` command and in a CCITT standard for data compression for modems. But, the original Lempel-Ziv scheme does not work very well for sequences with long repeated symbols (i.e., in such a case a run-length code is more useful). For example, the sequence 0000000000 has the Lempel-Ziv code of length ten (and thus it requires the same number of bits as the original string!). To remedy this situation, Louchard and Szpankowski [10] introduced a generalization of the Lempel-Ziv parsing scheme. It parses a sequence into phrases such that the next phrase is the longest phrase seen in the past by *at most* $b-1$ phrases ($b=1$ corresponds to the original Lempel-Ziv algorithm). A data compression code for this new algorithm may consists of pairs of numbers: one being a pointer to the previous occurrence of the prefix of the phrase, and the second number is either empty or the last bit of the phrase in the case it is the bth phrase. For example, the previous string 0000000000 is parsed into (0)(0)(00)(00)(000)(0) for $b=2$ and requires only eight bits to code it.

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

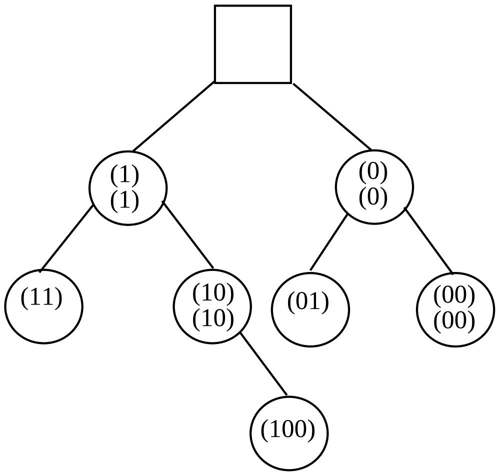

Figure 1: A 2-digital search tree representation of the generalized Lempel-Ziv parsing for the string 1100101000100010011

Our goal is to investigate the probabilistic behavior of a typical phrase length. As already observed in Louchard and Szpankowski [9] (cf. [5]), one can analyze the Lempel-Ziv algorithm by exploring a digital search tree (cf. [1, 7, 11]). In the case of the generalized Lempel-Ziv, we must consider a generalization of digital trees called *b-digital search tree*, or shortly b-DST (cf. [2, 11]) which is built from a fixed number, say m, of strings over a binary alphabet $\Sigma = \{0, 1\}$ (extension to a finite alphabet is easy). This tree is constructed as follows: The first b strings are stored in the root. The remaining $m - b$ strings are stored in an available space in a node which is not full, i.e., containing less than b strings. The search for an available space 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 until either we find a node with less than b strings or, if all nodes are full on this path, we create a new node. The details of such a construction can be found in [2, 7, 11]. To represent the generalized Lempel-Ziv scheme, we only must append the string to be parsed with b empty strings that are stored in the root. 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 of the string exactly in the same manner as in the digital tree construction untill either we find a node with less than b phrases or we need to create a new node. In Figure 1 we show a 2-DST constructed from the sequence 1100101000100010011. Observe that for a fixed length string, the number of nodes in the associated digital tree is random.

Digital trees appear in a variety of computer and communication applications including searching, sorting, dynamic hashing, codes stack communication protocols, and data compression as mentioned above (cf. [1, 2, 7, 8, 11, 9, 10, 5]). Although there is some unpleasant dependency between consecutive phrases in the Lempel-Ziv scheme, Louchard and Szpankowski [9] proved that this dependency is not too strong and one can infer the probabilistic behavior of the length of a randomly selected phrase from the depth of a randomly selected node in a b-DST built from *fixed* number of nodes. Therefore, in this conference version of the paper, we concentrate on studying b-digital search trees with fixed number, say m, strings.

We consider a digital tree in a probabilistic framework in which all m strings are statistically independent, and every string 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*.

Our results can be summarized as follows: For the b-digital search tree we establish the mean, the variance and the limiting distribution of the depth in the asymmetric Bernoulli model. In particularly, we prove that the depth appropriately normalized is normally distributed. We also show that the same is true for a randomly selected phrase in the generalized Lempel-Ziv scheme. We establish these results in a uniform manner by a novel technique that belongs to the toolkit of the analytical analysis of algorithms. More precisely, it was already observed by Flajolet and Richmond [2] that b-digital trees are harder to analyze than the regular ($b = 1$) digital search tree. We further observe that the asymmetric model considered in this paper adds a new hardship compared to the symmetric model analyzed by Flajolet and Richmond [1]. In particular, in the asymmetric model one cannot solve exactly functional equations involved (since the solution is no any longer a harmonic sum), and one must restore to asymptotic solutions through poissonization, singularity analysis, Mellin transform, and depoissonization. In addition, we had to develop a technique to estimate some constants (of Mellin transform) occuring in the final solution. These constants are somewhat important since they carry the information about the dependence of b on the depth and phrase length.

Digital search trees for $b = 1$ have been analyzed in the past in the case of *fixed* number of independent strings (cf. [1, 7, 8, 11, 14]). The situation is quite different when general b-digital search trees are considered. The first non-trivial analysis of the size of such trees in the symmetric Bernoulli model was proposed by Flajolet and Richmond [2]. The variance of the size and internal path length – still for the symmetric model – was discussed by Hubalek [4]. To the best of our knowledge, b-DST have not yet been analyzed for the asymmetric Bernoulli model. On the other hand, for the original Lempel-Ziv parsing algorithm ($b = 1$) mostly first-order properties have been investigated, with an exception of some recent works of Louchard and Szpankowski [9], and Jacquet and Szpankowski [6], and there is no analysis of the generalized Lempel-Ziv scheme except preliminary discussion of [10].

2. MAIN RESULTS

We consider a b-digital tree built over m statistically independent words. Let D_m^i be the depth of the i-th string in a b-DST, that is, the length of a path from the root to a node containing the ith string. In a variety of applications, one is interested in the *typical* depth D_m that is defined as the depth of a randomly selected string, that is,

$$P\{D_m = k\} = \sum_{i=1}^{m} \frac{P\{D_m^i = k\}}{m} \qquad (1)$$

As argued in Louchard and Szpankowski [9], the depth D_m is closely related to the length of a randomly selected phrase in the generalized Lempel-Ziv parsing scheme. We denote it as D_n^{LZ} where n is the length of the string to be parsed. Our goal is to study moments and distribution of D_m and D_n^{LZ}, and their dependence upon the parameter b.

As mentioned in the introduction, in this conference version of the paper we mostly concentrate on the depth D_m in a b-DST built over a fixed number of independent strings generated according to an asymmetric Bernoulli model (with "0" and "1" occurring respectively with probability p and $q = 1 - p$,). Let B_m^k be the average number of items (strings) on level k of a randomly built b-digital search tree. It is to observe that $\Pr\{D_m = k\} = B_m^k/m$ (cf. [7, 14]), thus one can alternatively study the average B_m^k

which is further called *the average profile*. Let $B_m(u) = \sum_{k\geq 0} B_m^k u^k$ be its generating function. A simple analysis of the digital search tree reveals (cf. [9]) that $B_m(u)$ satisfies the following recurrence:

$$B_{m+b}(u) = b + u\sum_{i=0}^{m}\binom{m}{i}p^i q^{m-i}\left(B_i(u) + B_{m-i}(u)\right) \tag{2}$$

with initial conditions

$$B_i(u) = i \qquad \text{for} \qquad i = 0, 1, \ldots, b \tag{3}$$

For $b = 1$ the above recurrence can be solved as shown in [13] (cf. [9]). Unfortunately, for $b > 1$ the recurrence becomes too complicated to be solved exactly. This was already noted by Flajolet and Richmond [2] who developed a special technique to deal with such recurrences for $b > 1$. Unfortunately again, the technique of [2] was designed for the symmetric Bernoulli model, and becomes very intricate for the asymmetric Bernoulli models (due to the fact that some sums occurring in the solution of (2) cease to become harmonic sums in the asymmetric case).

In view of this, we approach the general recurrence (2) from a new "angle". First of all, we "poissonize" the model, that is, we introduce the Poisson transform (or Poisson generating function) as

$$\widetilde{B}(u,z) = \sum_{i=0}^{\infty} B_i(u)\frac{z^i}{i!}e^{-z} \ .$$

Then, the recurrence becomes a slightly more manageable functional equation, namely:

$$(1 + \frac{\partial}{\partial z})^b \widetilde{B}(u,z) = b + u\left(\widetilde{B}(u,pz) + \widetilde{B}(u,qz)\right) \tag{4}$$

where $(1 + \frac{\partial}{\partial z})^b f(z) =^{\text{def}} \sum_{i=0}^{b}\binom{b}{i}\frac{\partial^i f(z)}{\partial z^i}$.

In the next section, we use the Mellin transform [3] and the depoissonization lemma [12, 6] to solve the above equation. We prove the following main result.

Theorem 1 (i) *Let D_m be the typical depth in a b-digital tree built over m statistically independent strings under the* **asymmetric** *Bernoulli model. Then*

$$\begin{aligned} ED_m &= \frac{1}{h_1}\log m + \frac{1}{h_1}\left(\frac{h_2}{2h_1} + \gamma - 1 - H_{b-1} - \Delta(b,p) + \delta(\log_2 m)\right) + O\left(\frac{\log m}{\sqrt{m}}\right) \\ \text{Var } D_m &= \frac{h_2 - h_1^2}{h_1^3}\log m + O(1) \end{aligned}$$

where $h_1 = -p\log p - q\log q$ is the entropy, $h_2 = p\log^2 p + q\log^2 q$, and $\gamma = 0.577...$ is Euler constant, while $H_{b-1} = \sum_{i=1}^{b-1}\frac{1}{i}$, $H_0 = 0$ is the harmonic sum. The constant $\Delta(b,p)$ can be computed as follows

$$\Delta(b,p) = \sum_{n=2b+1}^{\infty}\bar{f}_n\sum_{i=1}^{b}\frac{(i+1)b!}{(b-i)!n(n-1)...(n-i-1)} \tag{5}$$

where $\bar{f}_n$ is given recursively by

$$\begin{cases} f_{m+b} = m + \sum_{i=0}^{m}\binom{m}{i}p^i q^{m-i}(f_i + f_{m-i}), \ for \ m > 0, \\ f_0 = f_1 = ... = f_b = 0, \\ \bar{f}_{m+b} = f_{m+b} - m > 0, \ m \geq b+1. \end{cases}$$

Finally, $\delta(\log_2 m)$ *is a fluctuating function with a small amplitude when* $\log p/\log q$ *is rational, and* $\delta(\log_2 m) \equiv 0$ *otherwise.*

(ii) *Let* $G_m(u)$ *be the ordinary generating function of* D_m *(i.e.,* $G_m(u) = Eu^{D_m}$*),* $\mu_m = ED_m$*, and* $\sigma_m = \sqrt{\mathrm{Var}\ D_m}$*. Then, for a complex* τ *such that* $\Re(\tau) > 0$

$$e^{-\tau\mu_m/\sigma_m} G_m(e^{\tau/\sigma_m}) = e^{\frac{\tau^2}{2}}\left(1 + O\left(\frac{1}{\sqrt{\log m}}\right)\right) \tag{6}$$

Thus, the limiting distribution of $\frac{D_m - \mu_m}{\sigma_m}$ *is normal, and it converges in moments to the appropriate moments of the standard normal distribution. Also, there exist positive constants* A *and* $\alpha < 1$ *such that*

$$\Pr\left\{\left|\frac{D_m - c_1 \log m}{\sqrt{c_2 \log m}}\right| > k\right\} \le A\alpha^k \tag{7}$$

uniformly in k *for large* m*.*

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 denote the number of phrases produced by the algorithm. Let also $D_n^{LZ}(i)$ be the length of the ith phrase in the Lempel-Ziv model, where $1 \le i \le M_n$. By the *typical phrase length* 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

$$\Pr\{D_n^{LZ} = k\} = \sum_{m=m_L}^{m_U} \Pr\{D_n^{LZ} = k | M_n = m\} \Pr\{M_n = m\} \tag{8}$$

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 $m_L = \alpha_1\sqrt{n/b} \le M_n \le \alpha_2 (n/b)/\log_2(n/b) = m_U$. Indeed, the minimum number of phrases occurs only for two strings: either all zeros or all ones, and then $n = \sum_{i=1}^{M_n} D_n(i) \le b \sum_{i=1}^{M'n} i$, hence the lower bound $m_L = \Omega(\sqrt{n/b})$. 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 $O((n/b)/\log_2(n/b))$.

According to (8), 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 phrase (think of a digital tree built from the string $11111\ldots111$ which is very skewed). Fortunately, Louchard and Szpankowski [9] proved for $b = 1$ (and one can adopt without any significant changes the proof to the case $b > 1$) that $\Pr\{D_n^{LZ} = k | M_n = m\} = (1 + O(\sqrt{\log n/n}))\Pr\{D_m = k\}$.

Thus, the following second main result follows:

Theorem 2 *Let* D_n^{LZ} *be the length of a randomly selected phrase in the generalized Lempel-Ziv parsing of a string of length* n *under the* **asymmetric** *Bernoulli model. Then*

$$ED_n^{LZ} \sim \frac{1}{h_1}\log(nh_1/\log n) \tag{9}$$

$$\mathrm{Var}\ D_n^{LZ} \sim \frac{h_2 - h_1^2}{h_1^3}\log(nh_1/\log n)\ . \tag{10}$$

Furthermore, the limiting distribution of $\frac{D_n^{LZ}-ED_n^{LZ}}{\sqrt{\mathrm{Var}D_n^{LZ}}}$ *is normal.*

In passing, we should point out that the symmetric Bernoulli model differs significantly from the asymmetric model due to the fact (which we prove in a journal version of our paper) that $\mathrm{Var}D_m = O(1)$. In particular, the limiting distribution does not exist in this case, but it can be approximated by an extreme distribution.

3. ANALYSIS

In this section, we only prove Theorem 1 part (i) delaying the remaining proofs to the journal version.

As defined in the previous section, $B_m(u)$ is the generating function of the average profile B_m^k. Observe that $B_m(1) = m$, and $ED_m = B_m'(1)/m$, and finally $B_m''(1)/m = E\{D_m(D_m - 1)\} = \mathrm{Var}D_m - ED_m + (ED_m)^2$. Thus, Var $D_m = B_m''(1)/m + B_m'(1)/m - (B_m'(1)/m)^2$. We will use the above formula to derive asymptotics of ED_m and Var D_m as $m \to \infty$.

Our approach is analytical, and as mentioned in the previous section, we first derive the mean and the second factorial moment in the Poisson model. For this we need the first and the second derivative with respect to u at $u = 1$ of $B(u, z)$ derived in (4). One obtains:

$$
\begin{aligned}
(1+\frac{\partial}{\partial z})^b \widetilde{B}_u(u,z) &= \left(\widetilde{B}(u,pz)+\widetilde{B}(u,qz)\right) + u\left(\widetilde{B}_u(u,pz)+\widetilde{B}_u(u,qz)\right) \ , \\
(1+\frac{\partial}{\partial z})^b \widetilde{B}_{uu}(u,z) &= 2\left(\widetilde{B}_u(u,pz)+\widetilde{B}_u(u,qz)\right) + u\left(\widetilde{B}_{uu}(u,pz)+\widetilde{B}_{uu}(u,qz)\right) \ .
\end{aligned}
$$

Let $\widetilde{B}_u(1,z) = \widetilde{X}(z)$, $\widetilde{B}_{uu}(1,z) = \widetilde{W}(z)$ are the Poisson mean and the Poisson second factorial moment. Then, from the above we find

$$
(1+\frac{\partial}{\partial z})^b \widetilde{X}(z) = z + \widetilde{X}(pz) + \widetilde{X}(qz) \ , \tag{11}
$$

$$
(1+\frac{\partial}{\partial z})^b \widetilde{W}(z) = 2\left(f(pz) + \widetilde{X}(qz)\right) + \left(\widetilde{W}(pz)+\widetilde{W}(qz)\right) \ . \tag{12}
$$

Our goal is now to solve asymptotically (as $z \to \infty$ in a cone around $\Re(z) > 0$) the above two functional equations. It is well know that equations like these are amiable to the attack of Mellin transform. To recall, for a function $f(x)$ or real valued x, we define its Mellin transform $F^*(s)$ as

$$
F^*(s) = \mathcal{M}[f] = \int_0^\infty f(t)t^{s-1}dt \ .
$$

The interested reader can find more on Mellin transform in an excellent survey by Flajolet, Gourdon and Dumas [3].

Let now

$$
X(s) = \mathcal{M}[\widetilde{X}(t)] = \Gamma(s)\gamma(s) \ , \tag{13}
$$

$$
Y(s) = \mathcal{M}[\widetilde{W}(t)] = \Gamma(s)\beta(s) \tag{14}
$$

be Mellin transforms of the Poisson mean and the second factorial moment. They exist in proper strips as proved in the lemma below:

Lemma 1 . *The following is true:* (i) *$X(s)$ exists for $s \in (-b-1,-1)$, and $Y(s)$ exists for $s \in (-2b,-1)$.*

(ii) *Furthermore, $\gamma(-1-i) = 0$ for $i = 1,...,b-1$, $\gamma(-1-b) = (-1)^{b+1}$, and $\beta(-1-i) = 0$ for $i = 1,...,b$;*

Proof: Observe that $B_i(u) = i$ for $i = 0,1,\dots,b$ and thus $B_i(u) = b+(i-b)u$ for $i = 1+b,\dots,2b$ by recurrence (2). Taking derivatives, we find $\frac{\partial B_i(u)}{\partial u} = 0$ for $i = 0,1,\dots,b$ and $\frac{\partial B_i(u)}{\partial u} = i-b$ for $i = b,1+b,\dots,2b$. Furthermore, the second derivative is $\frac{\partial^2 B_i(u)}{\partial u^2} = 0$ for $i = 0,1,\dots,2b$. Thus, as $z \to 0$

$$\begin{aligned}
\widetilde{X}(z) &= \sum_{i=b+1}^{\infty} B_i'(1)\frac{z^i}{i!}e^{-z} = z^{(b+1)}/(b+1)! + O(z^{b+2}) \quad \text{as} \quad z \to 0 \\
\widetilde{W}(z) &= O(z^{2b+1}) \quad \text{as} \quad z \to 0
\end{aligned}$$

This proves part (i), while part (ii) follows directly from the lemma below. ■

Lemma 2 *Let $\{x_n\}_{n=0}^{\infty}$ be a sequence of real numbers, and let its Poisson generating function $\widetilde{F}(z) = \sum_{i=0}^{\infty} x_n \frac{z^n}{n!}e^{-z}$ be an entire function. Furthermore, let its Mellin transform $F^*(s)$ has the following factorization $F^*(s) = \mathcal{M}[f] = \Gamma(s)\gamma(s) = \mathcal{M}[f]$, and it exists in the strip $(-2,-1)$ with $\gamma(s)$ being analytical at $s = -2,-3,\dots$. Then,*

$$\gamma(-n) = \sum_{k=0}^{n}\binom{n}{k}(-1)^k x_k, \ \ for \ n \geq 2. \tag{15}$$

Proof: Let

$$F(z) = \sum_{n=1}^{\infty} g_n\frac{z^n}{n!}, \qquad \text{and} \qquad g_n = \sum_{k=0}^{n}\binom{n}{k}(-1)^{n-k}x_k \ ,$$

and let $F_N(z) = \sum_{n=1}^{N-1} g_n \frac{z^n}{n!}$, $N > 2$. Due to our assumptions, we can analytically continue $F^*(s)$ to the whole complex plane, and then for $s \in (-N,-N+1)$

$$F^*(s) = \mathcal{M}[F(z) - F_N(z)]$$

Since at $s = -N$,

$$F^*(s) = \frac{1}{s+N}\frac{(-1)^N}{N!}\gamma(-N) + O(1) \ ,$$

then by the inverse Mellin transform, we have

$$F(z) - F_N(z) = \frac{(-1)^N}{N!}\gamma(-N)z^N + O(z^{N+1}), \ \ as \ z \to 0 \ . \tag{16}$$

But,

$$F(z) - F_N(z) = \sum_{i=N}^{\infty} g_n\frac{z^n}{n!} = g_N\frac{z^N}{N!} + O(z^{N+1}) \ . \tag{17}$$

Thus, by comparing (16) and (17), we prove that $\gamma(-n) = (-1)^n g_n = \sum_{k=0}^{n}\binom{n}{k}(-1)^k x_k$ for $n > 2$. ■

Now, we are in a position to compute the Mellin transforms of $\widetilde{X}(z)$ and $\widetilde{W}(z)$. From (11) and (12), after taking Mellin transforms and using (13)-(14), we obtain

$$\begin{aligned}
\sum_{i=0}^{b}\binom{b}{i}(-1)^i\gamma(s-i) &= (p^{-s}+q^{-s})\gamma(s)\\
\sum_{i=0}^{b}\binom{b}{i}(-1)^i\beta(s-i) &= 2(p^{-s}+q^{-s})\gamma(s)+(p^{-s}+q^{-s})\beta(s)
\end{aligned}$$

and according to Lemma 1 $\gamma(s)$ exists in $\Re(s)\in(-b-1,-1)$ while $\beta(s)$ is well defined in the strip $\Re(s)\in(-2b,-1)$. To simplify it, we introduce a new notation, namely for any function $g(s)$ we define

$$\widehat{g}(s)=\sum_{i=1}^{b}\binom{b}{i}(-1)^{i+1}g(s-i) \tag{18}$$

provided $g(s-1),\ldots,g(s-b)$ exist. Then, we finally obtain

$$\gamma(s) = \frac{1}{1-p^{-s}-q^{-s}}\widehat{\gamma}(s) \tag{19}$$

$$\beta(s) = \frac{1}{1-p^{-s}-q^{-s}}\widehat{\beta}(s)+\frac{2(p^{-s}+q^{-s})}{(1-p^{-s}-q^{-s})^2}\widehat{\gamma}(s) \tag{20}$$

Let now s_k, $k=0,\pm1,\pm2,...$ be roots of $1-p^{-s}-q^{-s}=0$. Observe that $s_0=-1$. At $s=s_k$ we have

$$\frac{1}{1-p^{-s}-q^{-s}}=-\frac{1}{h(s_k)}\frac{1}{s-s_k}+\frac{h_2(s_k)}{2h^2(s_k)}+O(s-s_k) \tag{21}$$

where $h(t)=-p^{-t}\log p-q^{-t}\log q$. Expending $\Gamma(s)\widehat{\gamma}(s)$ around $s=s_k$ we find

$$\Gamma(s)\widehat{\gamma}(s)=\Gamma(s_k)\widehat{\gamma}(s_k)+\left(\Gamma(s_k)\widehat{\gamma}'(s_k)+\Gamma'(s_k)\widehat{\gamma}(s_k)\right)(s-s_k)+O((s-s_k)^2)\ .$$

Observe that the main contribution to the asymptotics comes from the residue at $s=s_0=-1$. After some algebra, we find

$$\gamma(s) = -\frac{1}{h_1}\frac{1}{s+1}+\frac{h_2}{2h_1^2}-\frac{\widehat{\gamma}'(-1)}{h_1}+O(s+1) \tag{22}$$

$$\beta(s) = \frac{2}{h_1^2}\frac{1}{(s+1)^2}+\left(-2\frac{h_2-h^2}{h_1^3}+2\widehat{\gamma}'(-1)\frac{1}{h_1^2}\right)\frac{1}{s+1}+O(1)\ . \tag{23}$$

In the above, we used the fact that $\widehat{\gamma}(-1)=1$ and $\widehat{\beta}(-1)=0$, which follow directly from Lemma 2. Observe now that $\Gamma(s)=-\frac{1}{s+1}+(\gamma-1)+O(s+1)$, hence the Laurent expansion of $X(s)$ at $s=-1$ is

$$X(s)=\Gamma(s)\gamma(s)=\frac{1}{h_1}\frac{1}{(s+1)^2}-\left(\frac{h_2}{2h_1^2}-\frac{1}{h_1}\widehat{\gamma}'(-1)+\frac{\gamma-1}{h_1}\right)\frac{1}{s+1} \tag{24}$$

The above formula for $X(s)$ can be now used to obtain asymptotics of $\widetilde{X}(z)$ using standard arguments (cf. [1, 3, 9]) of the inverse Mellin transform, that is,

$$\widetilde{X}(z)=\frac{1}{2\pi i}\int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty}X(s)z^{-s}ds\ .$$

After some algebra, we finally arrive at

$$\widetilde{X}(z) = \frac{1}{h_1} z \log z + \left(\frac{h_2}{2h_1^2} - \frac{1}{h_1}\widehat{\gamma}'(-1) + \frac{\gamma - 1}{h_1} \right) z + \sum_{k \neq 0} \frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h(s_k)} z^{-s_k} + O(1) \quad (25)$$

The above asymptotic formula concerns the behavior of the Poisson mean as $z \to \infty$. Our original goal was to derive asymptotics of the mean ED_m in the Bernoulli model. To infer Bernoulli model behavior from its Poisson model asymptotics, we must apply the so called *depoissonization lemma*. This lemma basically says that $mED_m \sim \widetilde{X}(m)$ under some weak conditions that are easy to verify in our case. The reader is referred to [6, 12] for more details about the depoissonization lemma. Thus,

$$ED_m = \frac{1}{h_1} \log m + \frac{h_2}{2h_1^2} - \frac{1}{h_1}\widehat{\gamma}'(-1) + \frac{\gamma - 1}{h_1} + \sum_{k \neq 0} \frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h(s_k)} m^{-1-s_k} + O\left(\frac{\log m}{\sqrt{m}} \right) .$$

Similarly, at $s = -1$,

$$Y(s) = -\frac{2}{h_1^2} \frac{1}{(s+1)^3} + \frac{2}{h_1} \left(\frac{h_2 - h_1^2}{h_1^2} - \frac{1}{h_1}\widehat{\gamma}'(-1) + \frac{\gamma - 1}{h_1} \right) \frac{1}{(s+1)^2} + O\left(\frac{1}{s+1} \right)$$

hence by the inverse Mellin transform, from the above equation we obtain

$$\widetilde{W}(z) = \frac{1}{h_1^2} z \log^2 z + \frac{2}{h_1} \left(\frac{h_2 - h_1^2}{h_1^2} - \frac{\widehat{\gamma}'(-1) + 1 - \gamma}{h_1} \right) z \log z + 2 \sum_{k \neq 0} \frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h^2(s_k)} z^{-s_k} \log z + O(1) .$$

Another application of the depoissonization lemma gives us

$$\begin{aligned} ED_m(D_m - 1) = \frac{\widetilde{W}(m)}{m} &= \frac{1}{h_1^2} \log^2 m + 2\frac{1}{h_1} \left(\frac{h_2 - h_1^2}{h_1^2} - \frac{1}{h_1}\widehat{\gamma}'(-1) + \frac{\gamma - 1}{h_1} \right) \log m \\ &\quad + 2 \sum_{k \neq 0} \frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h^2(s_k)} m^{-1-s_k} \log m + O\left(\frac{\log m}{\sqrt{m}} \right) . \end{aligned}$$

After some manipulation, we finally arrive at

$$\begin{aligned} \mathrm{Var}\ D_m &= \widetilde{W}(m)/m + ED_m - (ED_m)^2 \\ &= \frac{h_2 - h_1^2}{h_1^3} \log m + 2 \sum_{k \neq 0} \frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h(s_k)} \left(\frac{1}{h(s_k)} - \frac{1}{h_1} \right) m^{-1-s_k} \log m + C \quad (26) \end{aligned}$$

where C is a constant. If $\Re(s_k) = -1$, then $h(s_k) = h_1$, and

$$\mathrm{Var}\ D_m = \frac{h_2 - h_1^2}{h_1^3} \log m + O(1) .$$

If $\Re(s_k) > -1$, then $m^{-1-s_k} \log m = o(1)$, thus $\mathrm{Var}\ D_m = \frac{h_2 - h_1^2}{h_1^3} \log m + C$. We know that $\Re(s_k) = -1$ whenever $\log p / \log q$ is rational (cf. [5]), otherwise $\Re(s_k) > -1$.

Finally, we must show how to evaluate the constant $\widehat{\gamma}'(-1)$. Observe that $\gamma(s) = \mathcal{M}[\widetilde{X}(z)]$ and $\widetilde{X}(z) = \sum_{n=1}^{\infty} f_n \frac{z^n}{n!} e^{-z}$ where the sequence $f_n = ED_n$ is defined in Theorem 1(i). Thus, since $\mathcal{M}[z^n e^{-z}] = \Gamma(s+n)$, we obtain

$$\gamma(s) = \sum_{n=b+1}^{\infty} \frac{f_n}{n!} s(s+1)...(s+n-1)$$

for $\Re(s) \in (-b-2,-1)$. Hence,

$$\gamma'(s) = \sum_{n=b+1}^{\infty} \frac{f_n}{n!} s(s+1)...(s+n-1) \sum_{i=0}^{n-1} \frac{1}{s+i} .$$

After some tedious manipulation, one can evaluate $\widehat{\gamma}'(-1)$ as in (5). Thus, Theorem 1(i) is proved.

References

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

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

[3] P. Flajolet, X. Gourdon, P. Dumas, Mellin transforms and asymptotics: harmonic sums, *Theoretical Computer Science*, 144, 3-58, 1995.

[4] F. Hubalek, *Beiträge zur Analyse Verallgemeinerter Digitaler Suchbäume*, Ph. D. Technische Universität Wien, 1994.

[5] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, June 1995.

[6] P. Jacquet and W. Szpankowski, Analytical Depoissonization Lemma and Its Applications, preprint.

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

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

[9] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Information Theory*, 41, 478-488, 1995.

[10] G. Louchard and W. Szpankowski, Generalized Lempel-Ziv Parsing Scheme and its Preliminary Analysis of the Average Profile, *Proc. Data Compression Conference*, 262-271, Snowbird, 1995.

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

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

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

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

[15] 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).

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