1996

Average Profile of Generalized Digital Search Trees and the Generalized Lempel-Ziv Algorithm

Guy Louchard

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Jing Tang

Report Number:
96-005

AVERAGE PROFILE OF THE GENERALIZED DIGITAL SEARCH TREE AND THE GENERALIZED LEMPEL-ZIV ALGORITHM

Guy Louchard
Wojciech Szpankowski
Jing Tang

CSD-TR 96-005
Janurary 1996
(Revised November 1996)

AVERAGE PROFILE OF THE GENERALIZED DIGITAL SEARCH TREE AND THE GENERALIZED LEMPEL-ZIV ALGORITHM*

November 14, 1996

Guy Louchard	Wojciech Szpankowski†	Jing Tang
Dept. d'Informatique	Dept. of Computer Science	Applied Microsystems Co.
Université Libre de Bruxelles	Purdue University	5020 NE 148th Ave.
B-1050 Brussels	W. Lafayette, IN 47907	Redmond, WA 98052
Belgium	U.S.A.	U.S.A.
louchard@ulb.ac.be	spa@cs.purdue.edu	jtang@amc.com

Abstract

The goal of this research is threefold: (i) to analyze generalized digital search trees, (ii) to derive the average profile (i.e., phrase length) of a generalization of the well known parsing algorithm due to Lempel and Ziv, and (iii) to provide analytical tools to analyze asymptotically certain partial differential functional equations often arising in the analysis of digital trees. 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 shortest substring seen in the past by at most $b-1$ phrases ($b = 1$ corresponds to the original Lempel-Ziv scheme). Such a scheme can be analyzed through a generalized digital search tree in which every node is capable to store up to b strings. In this paper, we investigate the depth of a randomly selected node in such a tree and the length of a randomly selected phrase in the generalized Lempel-Ziv scheme. These findings and some of our recent results allow to compute the average redundancy of the generalized Lempel-Ziv code and compare it to the ordinary Lempel-Ziv code leading to an optimal value of b. Analytical techniques of (precise) analysis of algorithms are used to establish most of these conclusions.

Index Terms: Generalized Lempel-Ziv parsing scheme, generalized digital search trees, average redundancy, partial differential functional equations, singularity analysis, asymptotic expansions, depoissonization, Mellin transform.

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

†This author was additionally supported by NSF Grants NCR-9415491 and CCR-9201078.

1. INTRODUCTION

The heart of several universal data compression schemes is the parsing algorithm due to Lempel and Ziv [39] (e.g., it is used in the UNIX `compress` command and in a CCITT standard for data compression for modems). It is a dictionary based algorithm that 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. The Lempel-Ziv code consists of pairs of numbers: each pair being a pointer to the previous occurrence of the prefix of the phrase and the last symbol of the phrase (cf. [3]). For example, `ababcbababaaaaaaaaaca` is parsed into `(a)(b)(bb)(c)(ba)(bab)(aa)(aaa)(aaaa)(ca)` and its code becomes: `0a0b1b0c2a3b1a7a8a4a` (observe that there is no need for a separator between phrases). Let us count the length of this code in bits assuming a ternary alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. There are ten phrases, thus we need four bits to code each phrase. Every symbol requires two bits and there are ten symbols in the code, hence the total length of the code is 60 bits.

It is known that the original Lempel-Ziv scheme does not cope very well with sequences containing a long string of repeated symbols. To somewhat remedy this situation, Louchard and Szpankowski [24] introduced a generalization of the Lempel-Ziv parsing scheme: It parses a sequence into phrases such that the next phrase is the shortest 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 consists of pairs of number: one being a pointer to the previous *first* occurrence of the prefix of the phrase, and the second number either contains the last symbol of a new phrase in the case it is the first phrase among b identical phrases or otherwise it is empty. For example, the sequence above is parsed with $b = 2$ as follows: `(a)(b)(a)(c)(ba)(ba)(baa)(aa)(aa)(aaa)(ca)` which has eight *distinct* phrases (and twelve phrases). Its code is as follows: `0a0b10c2a44a1a66a3a` which is of length 40 bits (i.e., eight phrases each needed three bits and eight symbols every one requires two bits). We saved 20 bits!

In Section 2 we describe more precisely the code associated with the generalized Lempel-Ziv code. We also prove theoretically that this generalization improves slightly the redundancy of the code where by redundancy we mean a measure of how far the code is from being optimal for a given source of information. The computation of the average redundancy for Lempel-Ziv like codes was an open problem for awhile, but recently Louchard and Szpankowski [25] (cf. also [32]) proposed a method to evaluate it provided one computes *accurately* the average phrase length (see Section 2 for a more precise statement).

Our goal is to investigate the probabilistic behavior of a typical phrase length. As already

observed in Louchard and Szpankowski [23] (cf. [15]), the Lempel-Ziv algorithm can be modeled in two ways, namely as a *digital tree model* or a *Lempel-Ziv model.* In the former, one constructs the Lempel-Ziv sequence from m independent strings (of possibly infinite lengths). For example, let $m = 4$ sequences are given: $X_1 = 0000\ldots$, $X_2 = 1010\ldots$, $X_3 = 1111\ldots$ and $X_4 = 0101\ldots$. Then, for $b = 1$ the Lempel-Ziv sequence (0)(10)(11)(01) is of length $L_4 = 7$ and a typical (i.e., randomly selected) phrase is of length $1\frac{3}{4}$. In the Lempel-Ziv model there is only one sequence of fixed length, say n, and one partitions the sequences according to the Lempel-Ziv algorithm as described above. Clearly, these models are related as already observed in [23, 15]. We shall study both models in this paper.

Let us have a closer look at the **digital tree model**. To justify its name we shall show that it can be represented by a digital search tree (cf. [6, 19, 26]). In this case, we consider an extension of digital search trees called *b-digital search tree*, or (for short) b-DST (cf. [8, 26]) which is built from a fixed number, say $m + b$, of strings. Hereafter, we consider only the binary alphabet $\Sigma = \{0, 1\}$, but an extension to any finite alphabet is straightforward. This tree is constructed as follows: The first b strings are stored in the root. The remaining m 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 left , otherwise move to the right 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 [8, 19, 26].

Let us now discuss a digital tree representation for the **Lempel-Ziv model**. In this case, we assume that the first b phrases are empty and we store them in the root of a b-DST. All other phrases are stored in internal nodes and they are constructed on-line in the course of building the associated b-DST. 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 b-digital tree construction until either we find a node with less than b phrases or we create a new node. In Figure 1 we show the 2-DST constructed from the sequence 1100101000100010011. Observe that for a fixed length string, the number of nodes in the associated digital tree is a random variable that is equal to the number of distinct phrases of the generalized Lempel-Ziv scheme.

In this paper, we study both models in a probabilistic framework in which 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 Bernoulli model is called *symmetric*, otherwise it is *asymmetric.*

Digital trees appear in a variety of computer and communications applications includ-

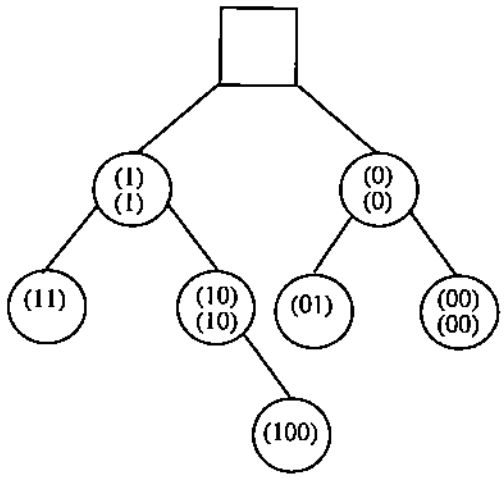

Figure 1: A 2-digital search tree representation of the generalized Lempel-Ziv parsing for the string 1100101000100010011

ing searching, sorting, dynamic hashing, codes, conflict resolution protocols for multiaccess communications, and data compression (cf. [6, 8, 19, 26, 23, 24, 15, 35]). Thus, better understanding of their behavior is desirable and could lead to some algorithmic improvements. One parameter that is of interest to these applications is the depth of a randomly (uniformly) selected node (i.e., the length of the path from the root to the chosen node). It can represent the search time for a key word or the length of a phrase in the generalized Lempel-Ziv algorithm (cf. Figure 1).

In this paper, we fully characterize the probabilistic behavior of the depth in a b-digital search tree under the digital tree model. We derive asymptotic expansions for the mean and the variance, as well as for large deviations and the limiting distribution of the depth. In particularly, we prove that the depth appropriately normalized is asymptotically normally distributed in the asymmetric Bernoulli model.

The *Lempel-Ziv model* is somewhat more intricate since there is some unpleasant dependency between consecutive phrases. Fortunately, Louchard and Szpankowski [23] proved that this dependency is not too strong (cf. equation (24) in Section 2), 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 a fixed number of nodes (i.e., in the digital tree model). In addition, using another recent finding of Louchard and Szpankowski [25] (concerning redundancy of the ordinary Lempel-Ziv code, i.e., for $b = 1$) we are able to compute the average redundancy of the generalized Lempel-Ziv code. The average redundancy measures how far the code is from being optimal for a given source of information (thus it

requires quite precise asymptotics of the average length of the Lempel-Ziv sequence in the digital tree model). This allows us to justify theoretically why the generalized Lempel-Ziv scheme is slightly better than the ordinary Lempel-Ziv scheme from the point of view of the average redundancy (cf. Theorem 3 in Section 2 and the discussion after).

We believe our contribution is also of a methodological nature: We establish our results in a consistent manner by a technique that belongs to the toolkit of the analytical analysis of algorithms. More precisely, it was already observed by Flajolet and Richmond [8] that b-digital trees are harder to analyze than the ordinary ($b = 1$) digital search trees. The difficulty boils down ultimately to a solution of the following general recurrence in x_n: Let $x_1, \ldots, x_b$ be given. Then, for a given sequence a_n and a constant u

$$x_{n+b} = a_n + u \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (x_k + x_{n-k}) \quad n \geq 0 \tag{1}$$

(cf. recurrence (4) in Section 2), which can be reduced to the following partial differential functional equation in terms of the Poisson generating function of x_n defined as $\tilde{X}(z) = \sum_{n \geq 0} x_n \frac{z^n}{n!} e^{-z}$:

$$\sum_{i=0}^{b} \binom{b}{i} \frac{\partial^i \tilde{X}(z)}{\partial z^i} = \tilde{A}(z) + u(\tilde{X}(pz) + \tilde{X}(qz)) \tag{2}$$

where $q = 1 - p$ (cf. (6) in Section 2 and (31)-(32) in Section 3).

The above recurrence can be solved exactly for $b = 1$ (cf. e.g., [34]), but attempts at extensions to $b > 1$ have partially failed. Flajolet and Richmond [8] (cf. also [12]) used a new technique to solve this recurrence for $p = 1/2$ (i.e., symmetric Bernoulli model). Unfortunately, this technique seems to be restricted to the symmetric Bernoulli model since some sums involved in the asymmetric Bernoulli model (i.e., $p \neq 1/2$) cease to be harmonic sums. To circumvent this difficulty we devise another approach that is asymptotic in its nature. In order to accomplish this, we use some other techniques such as analytical poissonization and depoissonization, singularity analysis, and Mellin transforms.

In passing, we should mention that differential functional equations such as (2) were already discussed in 1924 by Flamant [10] who provided an iterative solution. Our approach is completely different, and we present an asymptotic solution as $z \to \infty$ (which suffices to obtain an asymptotic solution of the original recurrence). Finally, during the course of the analysis we face a problem of numerical evaluation of some constants involving Mellin transforms. These constants are somewhat important since they carry the information about the dependence of b on the final solution. We propose here a method to evaluate numerically such constants (cf. Section 3.3) that is of its own interest and can be applied to other

problems. We should mention that similar numerical problems can be encountered in other analyses (cf. [17]).

Digital search trees for $b = 1$ have been analyzed in the past in the case of a *fixed* number of independent strings (cf. [6, 14, 18, 19, 20, 21, 28, 29, 30, 34, 35]). Much less is known about b-digital search trees. As mentioned above, the first non-trivial analysis of the size of such trees for the symmetric Bernoulli model was proposed by Flajolet and Richmond [8]. The variance of the size and the internal path length — still for the symmetric model — was discussed by Hubalek [12]. To the best of our knowledge, b-DST have not yet been analyzed for the asymmetric Bernoulli model. In the companion paper, one of us (cf. Louchard [22]) presents an alternative probabilistic approach to obtain some of our results, namely, the limiting distribution (without the rate of convergence) but not the large deviation results and precise evaluation of the moments (see Section 4.2 for the derivation of the asymptotic distribution in the symmetric case using this approach). In [22] Louchard also evaluates the average number of nodes in a b-digital search tree, thus directly extending the Flajolet and Richmond result [8] to the asymmetric Bernoulli model.

For the original Lempel-Ziv parsing algorithm ($b = 1$) mostly only first-order properties (e.g., leading terms in almost sure convergence) have been analyzed, with an exception of the work by Aldous and Shields [1], and recent works of Louchard and Szpankowski [23], and Jacquet and Szpankowski [14]. The first-order property of the length of a phrase in another Lempel-Ziv parsing algorithm (known as Lempel-Ziv'77 scheme [38]) was recently reported by Ornstein and Weiss [27] and Szpankowski [35]. Finally, Gilbert and Kadota [11] 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 present our main results concerning the digital tree model and the generalized Lempel-Ziv scheme. Proofs are deferred to Sections 3 and 4, where in the former we treat the asymmetric Bernoulli model, while in the latter the symmetric case. The proofs are analytical with the exception of the distribution in the symmetric Bernoulli model discussed in Section 4.2.

2. MAIN RESULTS

We consider a b-digital tree built over m statistically independent words. Let $D_m(i) = D_i(i)$ be the depth of the i-th string (of possibly infinite length) in such a tree, 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 defined as the depth of a randomly selected string,

that is,

$$\Pr\{D_m = k\} = \frac{1}{m}\sum_{i=1}^{m} P\{D_m(i) = k\} \ . \tag{3}$$

As argued in Louchard and Szpankowski [23], 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 parameter b.

2.1 Digital Tree Model

We now concentrate on the depth D_m in a b-DST built over a *fixed* number, say m, 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 *expected* number of strings on level k of a randomly built b-digital search tree. From the above we immediately obtain: $\Pr\{D_m = k\} = B_m^k/m$, 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 the generating function of the average profile.

A digital tree is a recursive structure. Suppose that there are $m + b$ strings to store. The root of such a tree contains b strings, and the remaining m strings are split between the left subtree and the right subtree. If i strings go to the left subtree, then its average profile is characterized by $uB_i(u)$ while $uB_{m-i}(u)$ is the generating function for the right subtree. Finally, the probability that i strings end up in the left subtree is equal to the probability that i out of m strings start with "0", and this is equal to $\binom{m}{i}p^i q^{m-i}$. Thus we have the following recurrence for $m \geq 0$:

$$B_{m+b}(u) = b + u\sum_{i=0}^{m}\binom{m}{i}p^i q^{m-i}\left(B_i(u) + B_{m-i}(u)\right) \tag{4}$$

with initial conditions

$$B_i(u) = i \qquad \text{for} \qquad i = 0, 1, \ldots, b-1 \ . \tag{5}$$

For $b = 1$ the above recurrence can be solved exactly as discussed in [33] (cf. [23]). Unfortunately, for $b > 1$ the recurrence becomes too complicated and no exact solution is known. This was already noted by Flajolet and Richmond [8] who developed a special technique to deal with such recurrences for $b > 1$. Unfortunately again, the technique of [8] 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 (4) cease to become harmonic sums in the asymmetric case).

In view of this, we approach the general recurrence (4) from a different "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 differential functional equation, namely:

$$\left(1 + \frac{\partial}{\partial z}\right)^b \widetilde{B}(u,z) = b + u\left(\widetilde{B}(u,pz) + \widetilde{B}(u,qz)\right) \tag{6}$$

where $(1+\frac{\partial}{\partial z})^b f(z) =^{\text{def}} \sum_{i=0}^{b} \binom{b}{i} \frac{\partial^i f(z)}{\partial z^i}$. We shall study $\widetilde{B}(u,z)$ for $z \to \infty$ in a cone around the real axis and u in a compact set around $u = 1$. This will suffice to recover asymptotics of $B_m(u)$, as discuss in Section 3.2.

In passing, we should point out that $\widetilde{B}(u,z)$ represents the average profile in the so called Poisson model in which the fixed number of strings is replaced by a random number of strings distributed according to a Poisson with mean z. To take the full advantage of this new model, however, we shall postulate that z is a complex variable. After "depoissonization" (cf. Section 3.2) we expect that $B_m(u) \sim \widetilde{B}(u,m)$.

In the next section, we use the Mellin transform [9], singularity analysis [7], and the depoissonization lemma [16, 31] to solve the above equation, and to prove the following main result.

Theorem 1 (ASYMMETRIC BERNOULLI MODEL)
(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*

$$ED_m = \frac{1}{h_1}\log m + \frac{1}{h_1}\left(\frac{h_2}{2h_1} + \gamma - 1 - H_{b-1} - \Delta(b,p) + \delta_1(m)\right) + O\left(\frac{\log m}{m}\right) \tag{7}$$

$$\text{Var } D_m = \frac{h_2 - h_1^2}{h_1^3}\log m + O(1) \tag{8}$$

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 the Euler constant, while $H_{b-1} = \sum_{i=1}^{b-1} \frac{1}{i}$, $H_0 = 0$ are harmonic numbers. The constant $\Delta(b,p)$ can be computed as follows (see Table 1 in Section 3.3 for numerical values):

$$\Delta(b,p) = \sum_{n=2b+1}^{\infty} \tilde{f}_n \sum_{i=1}^{b} \frac{(i+1)b!}{(b-i)!n(n-1)...(n-i-1)} < \infty \tag{9}$$

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

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

Finally, $\delta_1(x)$ is a fluctuating function with a small amplitude (see (49)) when $(\log p)/(\log q)$ is rational, and $\lim_{x\to\infty} \delta_1(x) = 0$ otherwise.

(ii) *Let $G_m(u)$ be the probability generating function of D_m (i.e., $G_m(u) = Eu^{D_m}$), $\mu_m = ED_m$, and $\sigma_m = \sqrt{\text{Var } D_m}$. Then, for complex τ*

$$e^{-\tau\mu_m/\sigma_m} G_m(e^{\tau/\sigma_m}) = e^{\frac{\tau^2}{2}} \left(1 + O\left(\frac{1}{\sqrt{\log m}}\right)\right) \tag{10}$$

Thus, the limiting distribution of $\frac{D_m - \mu_m}{\sigma_m}$ is normal, and it converges in moments (i.e., in mean of any order) to the appropriate moments of the standard normal distribution. Also, there exist positive constants A and $\alpha < 1$ such that uniformly in k for large m

$$\Pr\left\{\left|\frac{D_m - c_1 \log m}{\sqrt{c_2 \log m}}\right| > k\right\} \leq A\alpha^k \tag{11}$$

where $c_1 = 1/h_1$ and $c_2 = (h_2 - h_1^2)/h_1^3$.

The **symmetric** Bernoulli model must be treated differently since we shall prove below that Var $D_m = O(1)$, and hence a central limit theorem may not hold, which is actually the case. We use the Flajolet and Richmond [8] technique to prove this fact (cf. Section 4.1). Using a probabilistic approach we also establish the exact distribution of D_m (cf. Section 4.2). Both results are summarized in Theorem 2 below.

Before we present our findings, we must introduce some additional notation. Let

$$Q(t) = \prod_{k=0}^{\infty} (1 + t2^{-k}) \ , \tag{12}$$

and for integer s and complex z we define

$$H(s) = \frac{\partial^s}{\partial z^s} \left(\frac{1}{Q^b(-z)}\right)\Big|_{z=1} \ , \tag{13}$$

$$R_i(s) = -\frac{\partial^s}{\partial z^s} \left(\prod_{k=1}^{i} (1 - z2^k)^{-b}\right)\Big|_{z=1} \ , \quad R_0(s) = -1 \ . \tag{14}$$

Our second main result can be summarized as follows:

Theorem 2 (SYMMETRIC BERNOULLI MODEL)

(i) *Let us consider the symmetric Bernoulli model (with $p = q = 1/2$). The mean value ED_m is given by (7), while the variance becomes*

$$\mathrm{Var}D_m = \frac{1}{12} + \frac{1}{L^2}\left(1 + \frac{\pi^2}{6}\right) + \frac{1}{L^2}\left(J''(0) - (J'(0))^2\right) + \frac{1}{L}\delta_2(\log_2 m) - [\delta_1^2]_0 + O\left(\frac{\log^2 m}{m}\right) \tag{15}$$

where $L = h_1 = \log 2$ and

$$J'(0) = \int_0^1 \left(\frac{1}{Q(t)^b} - 1\right)\frac{dt}{t} + \int_0^\infty \frac{1}{Q(t)^b}\frac{dt}{t}\ , \tag{16}$$

$$J''(0) = -\frac{\pi^2}{3} + 2\int_0^1 \left(\frac{1}{Q(t)^b} - 1\right)\frac{\log t}{t}dt + 2\int_0^\infty \frac{1}{Q(t)^b}\frac{\log t}{t}dt\ , \tag{17}$$

and $\delta_2(\cdot)$ is a periodic function with mean zero and period 1, and $[\delta_1^2]_0$ is a very small constant (e.g., $[\delta_1^2]_0 \le 10^{-10}$ for $b = 1$). More precisely: as in Hubalek [12] with $\chi_k = 2k\pi i/L$ for $k = \pm 1, \pm 2, \ldots$

$$[\delta_1^2]_0 = \frac{1}{L^2}\sum_{k\neq 0}\frac{I(\chi_k)I(-\chi_k)}{\Gamma(2+\chi_k)\Gamma(2-\chi_k)}\ ,$$

where $\Gamma(s)$ is Euler's Gamma function, and

$$I(\chi_k) = \frac{1}{\chi_k} + \int_0^1 \left(Q^{-b}(t) - 1\right)t^{\chi_k - 1}dt + \int_1^\infty Q^{-b}(t)t^{\chi_k}dt\ .$$

(ii) *The exact distribution of D_m is given by*

$$m\Pr\{D_m \le j\} = b - \frac{1}{(b-1)!}\sum_{k=1}^{j}(1-2^k)^{-b} \times \tag{18}$$
$$\times \frac{\partial^{b-1}}{\partial z^{b-1}}\left(\frac{z^{2b}}{(z-1)^2}(z^{-b} - z^{-m})\prod_{1\le \ell\le j;\ell\neq k}\left(\frac{2^{-\ell}z}{1-(1-2^{-\ell})z}\right)^b\right)\Bigg|_{z=(1-2^{-k})^{-1}}$$

for any positive integer j.

(iii) *Let now $j = \lfloor \log_2 m\rfloor + \kappa$ for an integer κ, and define $\{\log_2 m\} = \log_2 m - \lfloor \log_2 m\rfloor$. Then, the "asymptotic distribution" of D_m can be expressed as*

$$\lim_{m\to\infty}\Big|\Pr\{D_m \le \lfloor \log_2 m\rfloor + \kappa\} - \sum_{\substack{\ell+s+t=b-1\\ \ell,s,t\ge 0}}\frac{(s+1)}{\ell!}\sum_{i=0}^{\infty}\frac{K_i(t)e^{-2^{-(\kappa-\{\log_2 m\}-i-1)}}}{2^{-(\ell-1)(i-(\kappa-\{\log_2 m\}))}}$$
$$+ \sum_{s+t=b-1}(s+1)\sum_{i=0}^{\infty}K_i(t)\left(e^{-2^{-(\kappa-\{\log_2 m\}-i-1)}} - 1\right)2^{\kappa-\{\log_2 m\}-i}\Big| = 0$$

where $K_i(t) = \sum_{s_1+s_2=t}\frac{(-1)^t}{s_1!s_2!}R_i(s_2)H(s_1)$ exponentially decreases with t. Observe that the limiting distribution of D_m does not exist in the symmetric case due to the term $\{\log_2 n\}$ which is dense in $[0,1]$ but not uniformly dense.

In passing it should be noted that the "asymptotic distribution" established in part (iii) above resembles a "mixture" of double exponential distributions (i.e., $e^{-2^{-x}}$), as in the case $b = 1$. An intuitive explanation for different behavior in the symmetric case is given in [23], but it follows basically from the fact that Var $D_m = O(1)$. We should also point out that numerical values of $J'(0)$ and $J''(0)$ can be found in Hubalek [12].

2.2 Lempel-Ziv Model

The situation is similar, but *not* the same, in the **Lempel-Ziv model** in which a sequence of *fixed length* n is parsed into random number of phrases. Let M_n denote the number of *full* phrases produced by the algorithm (the last incomplete phrase is ignored). We should mention that for $b > 1$ the number of full *distinct* phrases, M_n', is not equal to the total number of phrases M_n. 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 mean the length of a randomly selected phrase, that is, conditioned on M_n each phrase has probability $1/M_n$ of being selected.

The typical depth D_n^{LZ} in the Lempel-Ziv model can be estimated as follows (cf. [23]):

$$\Pr\{D_n^{LZ} = k\} = \sum_{m=m_L}^{m_U} \Pr\{D_n^{LZ} = k | M_n = m\} \Pr\{M_n = m\} \tag{19}$$

where m_L and m_U are lower and upper bounds for the number of phrases. It is easy to see that there exist constants $\alpha_1 > 0$ and $\alpha_2 < \infty$ 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 $n = \sum_{i=1}^{M_n} D_n^{LZ}(i) \le b \sum_{i=1}^{M_n} i$, whence 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 $O((n/b)/\log_2(n/b))$.

To estimate the probabilities appearing in (19) one seeks the limiting distribution of M_n. This is a difficult problem even for $b = 1$, and only recently Jacquet and Szpankowski [14] "cracked" it down by showing that M_n appropriately normalized weakly converges to the standard normal distribution. The case $b > 1$ is still unsolved until now, however, the technique of [14] can handle this case, too. To see this, we first reduce the problem to another one on the digital tree model. Indeed, observe that the following relationship between M_n (Lempel-Ziv model) and $L_m = \sum_{i=1}^{m} D_m(i)$ (digital tree model) takes place:

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

which immediately implies

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

The above relationship is known as the *renewal equation*, and from Theorem 17.3 of [2] we conclude the central limit theorem for M_n knowing it holds for L_m. The latter is easier to handle but far from trivial, and the reader is referred to [14] for details.

One finds a similar situation for the case $b > 1$, thus a central limit theorem for the internal path length L_m should hold. The exponential generating function $L(z,u) = \sum_{m=0}^{\infty} Eu^{L_m} \frac{z^m}{m!}$ satisfies the following partial-functional differential equation

$$\frac{\partial^b L(z,u)}{\partial z^b} = L(pzu,u)L(qzu,u) \ . \tag{21}$$

The arguments from [14] can be extended to $b > 1$, after some tedious labor, and one can solve asymptotically the above equation. We formulate our conclusions in a form of a fact that follows from [14] but without providing any detailed derivation.

Fact 1 *Consider the* **asymmetric** *Bernoulli model. Let* $c_1 = 1/h_1$ *and* $c_2 = (h_2 - h_1^2)/h_1^3$.
(i) *The path length* L_m *in a b-digital search tree possesses the following limiting distribution*

$$\frac{L_m - EL_m}{\sqrt{\mathrm{Var} L_m}} \to N(0,1) \tag{22}$$

where $N(0,1)$ *denotes the standard normal distribution,* $EL_m = mED_m$, *and* Var $L_m = c_2 m \log m + O(m)$, *and the convergence is also in moments.*

(ii) *The number of phrases* M_n *of the generalized Lempel-Ziv parsing scheme satisfies the following*

$$\frac{M_n - EM_n}{\sqrt{\mathrm{Var} M_n}} \to N(0,1) \tag{23}$$

where $EM_n \sim nh_1/\log n$ *and* $\mathrm{Var} M_n \sim c_2 h_1^3 n/\log^2 n$ *where* c_2 *is defined in Theorem 1. Moreover, all moments of* M_n *converge to the appropriate moments of the normal distribution.*

Having settled this, we can return to evaluating the limiting distribution of the phrase length D_n^{LZ}. According to (19), one needs to estimate the conditional probability $\Pr\{D_n^{LZ} = k | M_n = m\}$. It is tempting to assume that it is equal to $\Pr\{D_m = k\}$ (the latter refers to the probability of 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. A mathematical form of this dependency is actually written down in (20). We can

use exactly the same arguments as in Louchard and Szpankowski [23] (cf. Section III-B of [23]) to show that for $b > 1$ the following holds

$$\Pr\{D_n^{LZ} = k | M_n = m\} = (1 + O(\sqrt{\log n/n}))\Pr\{D_m = k\} \tag{24}$$

as long as $k = O(ED_m) = O(\log m)$. This would particularly imply that for complex ϑ

$$Ee^{\vartheta D_n^{LZ}} \sim Ee^{\vartheta D_{\lfloor nh_1/\log n \rfloor}} .$$

as $n \to \infty$ (cf. [23] for details).

In summary, the second main result concerning the Lempel-Ziv model is presented below (for simplicity we formulate it only for the asymmetric case).

Theorem 3 *Consider the* **asymmetric** *Bernoulli model. Let D_n^{LZ} be the length of a randomly selected phrase in the generalized Lempel-Ziv scheme that partitions a string of length n. Then*

$$\frac{D_n^{LZ} - c_1 \log(nh_1/\log n)}{\sqrt{c_2 \log(nh_1/\log n)}} \to N(0,1) . \tag{25}$$

More precisely, for complex ϑ

$$e^{-\vartheta c_1 \sqrt{\log(nh_1/\log n)}} E\left(e^{\vartheta D_n^{LZ}/\sqrt{\log(nh_1/\log n)}}\right) = e^{c_2\vartheta^2/2}\left(1 + O\left(1/\sqrt{\log(n/\log n)}\right)\right) . \tag{26}$$

Furthermore, there exist two positive constants A' and $\alpha_1 < 1$ such that

$$\Pr\left\{\left|\frac{D_m^{LZ} - c_1 \log(nh_1/\log n)}{\sqrt{c_2 \log(nh_1/\log n)}}\right| > k\right\} \le A'\alpha_1^k \tag{27}$$

uniformly in k for large m, where c_1 and c_2 are defined above.

The symmetric Bernoulli model can be handled in a similar manner, but its formulation is too complicated to be presented in a compact form. It is described by a similar formula as for the digital tree model with m replaced by $n/\log_2 n$. Naturally, the limiting distribution does not exist as such but some limiting theorem can be formulated as in the case of the digital tree model (cf. Theorem 2(iii)).

Finally, in order to find an optimal b that asymptotically minimize the Lempel-Ziv code length, we shall deal with the *average redundancy* $\bar{r}_n$ which is defined as follows:

$$\bar{r}_n = \frac{E\ell_n - nh_1}{n}$$

where ℓ_n is the length of the generalized Lempel-Ziv code, and the expectation is taken with respect to the underlying probability measure. As explain in Section 1, the data compression

code for the generalized Lempel-Ziv scheme consists of pairs of numbers: one being a pointer to the previous occurrence of the prefix of the phrase, and the second number either contains the last symbol of a new phrase in the case it is the first phrase among b identical phrases or otherwise it is empty, as discussed in Section 1. Clearly, the length ℓ_n of such a code depends on two parameters, namely: the number of phrases M_n, and the number of *distinct* phrases M'_n, and it can be computed (for a binary alphabet) as follows

$$\ell_n(X_1^n) = M_n(\log M'_n + I) \tag{28}$$

where I is equal to one (for the binary alphabet, otherwise it is equivalent to the number of bits representing a symbol) if the phrase consists of a previously occurred prefix and an additional symbol (i.e., a bit in our case), and zero otherwise. It is not difficult to see that

$$EI = \frac{E(M'_n - 1)}{EM_n} .$$

Thus, we must evaluate EM_n and $E \log EM'_n$. But, as in [25] we notice that $EM_n \log M'_n = EM_n \log EM'_n + O(1/\log n)$. To estimate EM_n we observe that by (20) it is related to the average path length in the *digital tree model*, and $EL_m = mED_m$. As in Louchard and Szpankowski [25], and using Fact 1, we conclude that $EM_n \sim x_n$ where x_n is a solution of $EL_{x_n} = n$, that is:

$$x_n = \frac{nh_1}{\log n}\left(1 + \frac{\log\log n}{\log n} + \frac{A - \log h_1}{\log n} + O\left(\frac{(\log\log n)^2}{\log^2 n}\right)\right) .$$

where $-A$ is the $O(1)$ term divided by h_1 of (8) in Theorem 1(i), that is,

$$A = 1 + H_{b-1} + \Delta(b,p) - \frac{h_2}{2h_1} - \gamma - \delta_1(m) .$$

In a similar fashion we can estimate EM'_n, however, one should observe that M'_n is related to the size of the associated b-DST. More precisely, if S_m is the size of a b-DST built form m strings, then according to Flajolet and Richmond [8] (symmetric case), and Louchard [22] (asymmetric case):

$$ES_m = m\left(q_0(b) + \delta_2(m)\right) + O(1)$$

where $q_0(b)$ is a constant that can be computed explicitly. For example, Flajolet and Richmond [8] proved that

$$q_0(b) = \frac{1}{\log 2}\int_0^\infty \left(\frac{1+t}{Q(t)}\right)^b \frac{dt}{1+1}$$

where $Q(t) = \prod_{j=0}^{\infty}(1 + t2^{-j})$. Clearly, $q_0(1) = 1$, and the authors of [8] computed $q_0(2) = 0.5747$, $q_0(3) = 0.4069$, and so on. For large b one derives that $q_0(b) \sim 1/(b \log 2)$.

Putting everything together, and using the approach from [25], we finally arrive at the following formula for the average redundancy of the gencralized Lempel-Ziv code

$$\bar{r}_n(b) = h_1 \frac{1 - \gamma - \frac{h_2}{2h_1} + \Delta(b,p) + H_{b-1} + q_0(b) + \log q_0(b) - \delta(n)}{\log n} + O\left(\frac{\log\log n}{\log^2 n}\right) , \quad (29)$$

where $\delta(n)$ is a fluctuating function with a small amplitude, and the other quantities are defined as above.

It may be interesting to compare the average redundancy for different values of b hoping that there exists an optimal value of b. For example, for the symmetric Bernoulli model with a binary alphabet we obtain

$$\begin{aligned}
\bar{r}_n(1) &= \frac{2.27 + \delta(n)}{\log_2 n} + O\left(\frac{\log\log n}{\log^2 n}\right) , \\
\bar{r}_n(2) &= \frac{1.98 + \delta(n)}{\log_2 n} + O\left(\frac{\log\log n}{\log^2 n}\right) , \\
\bar{r}_n(3) &= \frac{1.89 + \delta(n)}{\log_2 n} + O\left(\frac{\log\log n}{\log^2 n}\right) , \\
\bar{r}_n(\infty) &= \frac{1.71 + \delta(n)}{\log_2 n} + O\left(\frac{\log\log n}{\log^2 n}\right) .
\end{aligned}$$

Thus there exists an optimal value of b which minimizes the average redundancy. Some recent preliminary experimental results (cf. [13]) carried out on structured ASCII files seem to confirm that a practical saving can be achieved for $b > 1$, and this is particularly true for large alphabets (e.g., image), as already seen in Section 1. We should point out that these experimental findings are very sensitive to implementation issues. A particular implementation can add $O(M_n)$ bits which contribute $O(1/\log n)$ to the expected redundancy.

3. ANALYSIS OF THE ASYMMETRIC BERNOULLI MODEL

In this section, we prove Theorem 1 concerning the digital tree model in the asymmetric Bernoulli model. After establishing recurrences for the mean and variance, we proceed to derive the asymptotics of these quantities. We first deal with the Poisson model (Section 3.1), and then depoissonize the results (Section 3.2). Special attention is devoted to computing some constants arising in the analysis (Section 3.3). Finally, we show how to obtain the limiting distribution for D_m (Section 3.4).

3.1 Analysis of Moments in the Poisson Model

As defined in Section 2.1, $B_m(u)$ 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 $B_m''(1)/m = E\{D_m(D_m - 1)\} =$

$\mathrm{Var} D_m - ED_m + (ED_m)^2$. Thus,

$$\text{Var } D_m = \frac{B''_m(1)}{m} + \frac{B'_m(1)}{m} - \left(\frac{B'_m(1)}{m}\right)^2 . \tag{30}$$

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 of the average profile in the Poisson model, that is, the first and the second derivative with respect to u at $u = 1$ of $\widetilde{B}(u, z)$. One obtains:

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

Let $\widetilde{B}_u(1, z) = \widetilde{X}(z)$, $\widetilde{B}_{uu}(1, z) = \widetilde{W}(z)$ which suffice to compute the mean and the variance of D_m, as indicated above. Then,

$$\left(1 + \frac{\partial}{\partial z}\right)^b \widetilde{X}(z) = z + \widetilde{X}(pz) + \widetilde{X}(qz) , \tag{31}$$
$$\left(1 + \frac{\partial}{\partial z}\right)^b \widetilde{W}(z) = 2\left(\widetilde{X}(pz) + \widetilde{X}(qz)\right) + \left(\widetilde{W}(pz) + \widetilde{W}(qz)\right) . \tag{32}$$

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 known that equations like these are amiable to attack by the Mellin transform. To recall, for a function $f(x)$ of real valued x, we define its Mellin transform $F(s)$ as

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

In some of our arguments (e.g., depoissonization of Section 3.2 and singularity analysis of Section 4.1), we could either use Mellin transform of a complex variable function $f(z)$ or we could use an analytical continuation argument. It is known (cf. [5, 16]) that as long as $\arg(z)$ belongs to some cone around the real axis, the Mellin transform $f(s)$ of a function of a real argument and its corresponding function of a complex argument is the same. Therefore, we work most of the time with the Mellin transform of a function of real variable as defined above.

Let now

$$X(s) = \mathcal{M}[\widetilde{X}(t); s] = \Gamma(s)\gamma(s) , \tag{33}$$
$$Y(s) = \mathcal{M}[\widetilde{W}(t); s] = \Gamma(s)\beta(s) \tag{34}$$

where $\Gamma(s)$ is the classical Gamma function, and we aim at computing $\gamma(s)$ and $\beta(s)$. They exist in a proper strip as proved in the lemma below:

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

(ii) *Furthermore,* $\gamma(-1-i)=0$ *for* $i=1,...,b-1$, $\gamma(-1-b)=(-1)^{b+1}$, *and* $\beta(-1-i)=0$ *for* $i=1,...,b$, *and* $\gamma(s)$ *has simple poles at* $s=-1,0,1,\ldots$.

Proof: By recurrence (4), we have $B_i(u)=i$ for $i=0,1,\ldots,b$ and thus $B_i(u)=b+(i-b)u$ for $i=1+b,\ldots,2b$. Taking derivatives, we obtain $\frac{\partial B_i(u)}{\partial u}=0$ for $i=0,1,...,b$ and $\frac{\partial B_i(u)}{\partial u}=i-b$ for $i=b,1+b,...,2b$. Furthermore, the second derivative becomes $\frac{\partial^2 B_i(u)}{\partial u^2}=0$ for $i=0,1,...,2b$. Hence, for $z\to 0$

$$\begin{aligned}\tilde{X}(z) &= \left(z^{(b+1)}/(b+1)!+2z^{(b+2)}/(b+2)!+3z^{(b+3)}/(b+3)!+O(z^{b+4})\right)e^{-z}\\ &= z^{(b+1)}/(b+1)!+O(z^{b+2}) \;\; as \;\; z\to 0\\ \widetilde{W}(z) &= O(z^{2b+1}) \;\; as \;\; z\to 0\end{aligned}$$

On the other hand, for $z\to\infty$ we conclude from (31) and (32) that $\tilde{X}(z)=O(z\log z)$ and $\widetilde{W}(z)=O(z\log^2 z)$. Thus, the first part of the lemma is proven. Part (ii) follows directly from the lemma below and (39). ■

Lemma 2 *Let* $\{f_n\}_{n=0}^{\infty}$ *be a sequence of real numbers, and suppose that its Poisson generating function* $\tilde{F}(z)=\sum_{n=0}^{\infty}f_n\frac{z^n}{n!}e^{-z}$ *is an entire function. Furthermore, let its Mellin transform* $F(s)$ *has the following factorization* $F(s)=\mathcal{M}[\tilde{F}(z);s]=\Gamma(s)\gamma(s)$, *and assume* $F(s)$ *exist for* $\Re(s)\in(-2,-1)$ *and* $\gamma(s)$ *is analytical at* $s=-2,-3,\ldots$. *Then*

$$\gamma(-n)=\sum_{k=0}^{n}\binom{n}{k}(-1)^k f_k, \quad \text{for} \quad n\geq 2. \tag{35}$$

Proof: Let a sequence $\{g_n\}_{n=0}^{\infty}$ be such that $\tilde{F}(z)=\sum_{n=0}^{\infty}g_n\frac{z^n}{n!}$, that is (cf. [8]),

$$g_n=\sum_{k=0}^{n}\binom{n}{k}(-1)^{n-k}f_k \; .$$

Define now for some fixed $N\geq 2$, the function $\tilde{F}_N(z)=\sum_{n=0}^{N-1}g_n\frac{z^n}{n!}$. Due to our assumptions, we can analytically continue $F(s)$ to the whole complex plane except $s=-2,-3,\ldots$. In particular, for $\Re(s)\in(-N,-N+1)$ we have

$$F(s)=\mathcal{M}[\tilde{F}(z)-\tilde{F}_N(z);s]$$

since a polynomial in z such as $\tilde{F}_N(z)$ can only shift the fundamental strip of the Mellin transform but cannot change its value (cf. [9]). As $s\to -N$, due to the assumed factorization

$F(s) = \Gamma(s)\gamma(s)$, we have

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

thus, by the inverse Mellin transform, we have

$$\widetilde{F}(z) - \widetilde{F}_N(z) = \frac{(-1)^N}{N!}\gamma(-N)z^N + O(z^{N+1}) \quad \text{as} \quad z \to 0 \ . \tag{36}$$

But,

$$\widetilde{F}(z) - \widetilde{F}_N(z) = \sum_{i=N}^{\infty} g_n \frac{z^n}{n!} = g_N \frac{z^N}{N!} + O(z^{N+1}) \ . \tag{37}$$

Thus, by comparing (36) and (37), we prove that $\gamma(-N) = (-1)^N g_N = \sum_{k=0}^{N} \binom{N}{k}(-1)^k f_k$ for $N \geq 2$. ■

Now, we are in a position to compute the Mellin transforms of $\widetilde{X}(z)$ and $\widetilde{W}(z)$. From (31) and (32), after taking Mellin transforms and using (33)-(34), 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 by Lemma 1 $\gamma(s)$ exists in $\Re(s) \in (-b-1, -1)$, while $\beta(s)$ is well defined in the strip $\Re(s) \in (-2b-1, -1)$. To simplify the above, we define for any function $g(s)$:

$$\widehat{g}(s) = \sum_{i=1}^{b} \binom{b}{i}(-1)^{i+1} g(s-i) \tag{38}$$

provided $g(s-1), ..., g(s-b)$ exist. Then

$$\gamma(s) = \frac{1}{1-p^{-s}-q^{-s}} \sum_{i=1}^{b} \binom{b}{i}(-1)^{i+1}\gamma(s-i) = \frac{1}{1-p^{-s}-q^{-s}}\widehat{\gamma}(s) \tag{39}$$

$$\begin{aligned}
\beta(s) &= \frac{1}{1-p^{-s}-q^{-s}} \sum_{i=1}^{b} \binom{b}{i}(-1)^{i+1}\beta(s-i) + \frac{2(p^{-s}+q^{-s})}{1-p^{-s}-q^{-s}}\gamma(s) \ , \\
&= \frac{1}{1-p^{-s}-q^{-s}}\widehat{\beta}(s) + \frac{2(p^{-s}+q^{-s})}{(1-p^{-s}-q^{-s})^2}\widehat{\gamma}(s) \ .
\end{aligned} \tag{40}$$

Let now s_k, $k = 0, \pm 1, \pm 2, \ldots$, be roots of $1 - p^{-s} - q^{-s} = 0$. Observe that $s_0 = -1$. At $s = s_k$ we have

$$\frac{1}{1-p^{-s}-q^{-s}} = -\frac{1}{h(s_k)}\frac{1}{s-s_k} + \frac{h_2(s_k)}{2h^2(s_k)} + O(s-s_k) \ , \tag{41}$$

where

$$h(t) = -p^{-t}\log p - q^{-t}\log q\ , \tag{42}$$

$$h_2(t) = p^{-t}\log^2 p + q^{-t}\log^2 q\ . \tag{43}$$

Expanding $\Gamma(s)\widehat{\gamma}(s)$ around $s = s_k$, we find

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

Therefore, since $X(s) = \Gamma(s)\gamma(s) = \frac{1}{1-p^{-s}-q^{-s}}\Gamma(s)\widehat{\gamma}(s)$, we derive around $s = s_k \neq -1$

$$\begin{aligned} X(s) &= -\frac{1}{s-s_k}\frac{\Gamma(s_k)}{h(s_k)}\widehat{\gamma}(s_k) + \frac{h_2(s_k)}{2h^2(s_k)}\Gamma(s_k)\widehat{\gamma}(s_k) \\ &- \frac{1}{h(s_k)}\left(\Gamma(s_k)\widehat{\gamma}'(s_k) + \Gamma'(s_k)\widehat{\gamma}(s_k)\right) + O(s-s_k)\ . \end{aligned} \tag{44}$$

In a similar manner, from (40) we have

$$\begin{aligned} Y(s) &= -\frac{1}{s-s_k}\frac{\Gamma(s_k)}{h(s_k)}\widehat{\beta}(s_k) + 2\Gamma(s_k)\left(\frac{1}{h^2(s_k)}\frac{1}{(s-s_k)^2} - \frac{h_2(s_k)-h^2(s_k)}{h^3(s_k)}\frac{1}{s-s_k}\right) \\ &\quad (\widehat{\gamma}(s_k) + (s-s_k)\widehat{\gamma}'(s_k)) + \frac{2\Gamma'(s_k)\widehat{\gamma}(s_k)}{h^2(s_k)}\frac{1}{s-s_k} + O(1) \\ &= \frac{2\Gamma(s_k)\widehat{\gamma}(s_k)}{h^2(s_k)}\frac{1}{(s-s_k)^2} + \left(\frac{2\Gamma'(s_k)\widehat{\gamma}(s_k)}{h^2(s_k)} - \frac{\Gamma(s_k)}{h(s_k)}\widehat{\beta}(s_k)\right. \\ &\quad \left. -2\Gamma(s_k)\frac{h_2(s_k)-h^2(s_k)}{h^3(s_k)}\widehat{\gamma}(s_k) - \frac{2\Gamma(s_k)\widehat{\gamma}'(s_k)}{h^2(s_k)}\right)\frac{1}{s-s_k} + O(1)\ . \end{aligned} \tag{45}$$

On the other hand, from (41) and (39) at $s = s_0 = -1$, we find

$$\begin{aligned} \gamma(s) &= \left(-\frac{1}{h_1}\frac{1}{s+1} + \frac{h_2}{2h_1^2}\right)(\widehat{\gamma}(-1) + (s+1)\widehat{\gamma}'(-1)) + O(s+1) \\ &= -\frac{1}{h_1}\frac{1}{s+1} + \frac{h_2}{2h_1^2} - \frac{\widehat{\gamma}'(-1)}{h_1} + O(s+1)\ , \end{aligned} \tag{46}$$

$$\begin{aligned} \beta(s) &= \left(-\frac{1}{h_1}\frac{1}{s+1} + \frac{h_2}{2h_1^2}\right)\left(\widehat{\beta}(-1) + (s+1)\widehat{\beta}'(-1)\right) \\ &\quad +2\left(\frac{1}{h_1^2}\frac{1}{(s+1)^2} - \frac{h_2-h^2}{h_1^3}\frac{1}{s+1} + O(1)\right)(\widehat{\gamma}(-1) + (s+1)\widehat{\gamma}'(-1)) \\ &= \frac{2}{h_1^2}\frac{1}{(s+1)^2} + \left(-2\frac{h_2-h^2}{h_1^3} + 2\widehat{\gamma}'(-1)\frac{1}{h_1^2}\right)\frac{1}{s+1} + O(1)\ . \end{aligned} \tag{47}$$

In the above, we used the fact that $\widehat{\gamma}(-1) = 1$ and $\widehat{\beta}(-1) = 0$, which follow directly from Lemma 1. 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

$$X(s)=\Gamma(s)\gamma(s)=\frac{1}{h_1}\frac{1}{(s+1)^2}-\left(\frac{h_2}{2h_1^2}-\frac{1}{h_1}\widehat{\gamma}'(-1)+\frac{\gamma-1}{h_1}\right)\frac{1}{s+1}+O(1)\ . \tag{48}$$

In order to obtain an asymptotic expansion of $\widetilde{X}(z)$ for large z we use well known arguments (cf. [6, 9, 14, 23]) of the inverse Mellin transform, that is,

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

(The evaluation of this integral is quite standard (e.g., see [26]): we extend the line of integration to a big rectangle right to the integration line, and observe that bottom and top lines contribute negligibly due to the Gamma function behavior at imaginary argument, and the right side positioned at, say d, contributes x^{-d} for $d\to\infty$.) However, to estimate the error term we must note, as observed in Lemma 1, that $\gamma(s)$ has additional simple poles at $s=0,1,\ldots$. The pole at $s=0$ is a double pole of $X(s)=\Gamma(s)\gamma(s)$, thus its contribution to $\widetilde{X}(z)$ is $O(\log z)$. Putting together everything, we finally arrive at

$$\widetilde{X}(z) \;=\; \frac{1}{h_1}z\log z+\left(\frac{h_2}{2h_1^2}-\frac{1}{h_1}\widehat{\gamma}'(-1)+\frac{\gamma-1}{h_1}\right)z+\sum_{k\neq 0}\frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h(s_k)}z^{-s_k}+O(\log z) \tag{49}$$

Similarly, at $s=-1$,

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

In addition, there is a double pole at $s=0$, hence by the inverse Mellin transform we obtain

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

for $z\to\infty$, where $O(z)$ comes from the term $O((s+1)^{-1})$. This formula will allow us to infer asymptotics of the variance of D_m.

3.2 Depoissonization

The above asymptotic formulæ concern the behavior of the Poisson mean and the second factorial moment as $z\to\infty$. More precisely, we must restrict the growth of z to a cone $S_\theta=\{z:\ |\arg(z)|\le\theta\}$ for some $|\theta|<\pi/2$. But, our original goal was to derive asymptotics of the mean ED_m and the variance $\mathrm{Var}\,D_m$ 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 $mED_m \sim \widetilde{X}(m)$ and $mED_m(D_m-1) \sim \widetilde{W}(m)$ under some weak conditions that are easy to verify in our case. The reader is referred to [16, 31] for more details about depoissonization lemma. For completeness, however, we review some depoissonization results that are useful for our problem.

Let us consider a more general problem: For a random variable X_n define g_n as a characterization of X_n, e.g., $g_n = EX_n$ or $g_n = EX_n^2$, etc. We may also need to consider the generating function $G_n(u) = Eu^{X_n}$ for a complex u which can be viewed as g_n when u belongs to a compact set. Define the Poisson transform of g_n as $\widetilde{G}(z) = \sum_{n=0}^{\infty} g_n \frac{z^n}{n!} e^{-z}$ (or more generally: $\widetilde{G}(z,u) = \sum_{n=0}^{\infty} G_n(u) \frac{z^n}{n!} e^{-z}$ for u in a compact set). Assume that we know the asymptotics of $\widetilde{G}(z)$ for z large and belonging to a cone $S_\theta = \{z : \ |\arg(z)| \le \theta\}$ for some $|\theta| < \pi/2$. How can we infer asymptotics of g_n from $\widetilde{G}(z)$? One possible answer is given in the depoissonization lemma below (cf. [16, 31]):

Lemma 3 (DEPOISSONIZATION LEMMA)
(i) *Let $\widetilde{G}(z)$ be the Poisson transform of a sequence g_n that is assumed to be an entire function of z. We postulate that for $0 < |\theta| < \pi/2$ the following two conditions simultaneously hold for some numbers $A, B, \xi > 0$, β, and $\alpha < 1$:*
(I) *For $z \in S_\theta$*

$$|z| > \xi \quad \Rightarrow \quad |\widetilde{G}(z)| \le B|z|^{\beta} \Psi(|z|) \ , \tag{50}$$

where $\Psi(z)$ is a slowly varying function (e.g., $\Psi(z) = \log^d z$ for some $d > 0$),
(O) *For $z \notin S_\theta$*

$$|z| > \xi \quad \Rightarrow \quad |\widetilde{G}(z)e^z| \le A \exp(\alpha|z|) \ . \tag{51}$$

Then, for large n

$$g_n = \widetilde{G}(n) + O(n^{\beta-1}\Psi(n)) \ , \tag{52}$$

or more precisely:

$$g_n = \widetilde{G}(n) - \frac{1}{2}\widetilde{G}''(n) + O(n^{\beta-2}\Psi(n)) \ .$$

(ii) *If the above two conditions, namely (I) and (O), hold for $\widetilde{G}(z,u)$ for u belonging to a compact set $\mathcal{U}$, then*

$$G_n(u) = \widetilde{G}(n,u) + O(n^{\beta-1}\Psi(n)) \tag{53}$$

for large n and uniformly in $u \in \mathcal{U}$.

To apply the above lemma to $\widetilde{X}(z)$ and $\widetilde{W}(z)$ one must check conditions (I) and (O). But condition (I) inside the cone S_θ is automatically satisfied due to the asymptotics of $\widetilde{X}(z)$

and $\widetilde{W}(z)$ just derived. Formally, we must either use complex variable Mellin transform or use analytical continuation to establish $\widetilde{X}(z) = O(z \log z)$ and $\widetilde{W}(z) = O(z \log^2 z)$. Thus, it suffices to check condition (O) outside the cone (in fact, the arguments below work fine also for verifying condition (I)).

We only consider $\widetilde{X}(z)$ since $\widetilde{W}(z)$ can be treated in a similar manner. Let $X(z) = \widetilde{X}(z)e^z$. Then, functional equation (31) transforms into

$$X^{(b)}(z) = X(zp)e^{zq} + X(zq)e^{zp} + ze^z$$

where $X^{(b)}(z)$ denotes the bth derivative of $X(z)$. Observe that the above equation can be alternatively represented as

$$X(z) = \underbrace{\int_0^z \int_0^{w_2} \cdots \int_0^{w_b}}_{b \text{ times}} \left(X(w_1 p))e^{w-1q} + X(w_1 q)e^{w_1 p} + w_1 e^{w_1} \right) dw_1 dw_2 \cdots dw_b \qquad (54)$$

We now prove $|X(z)| \leq e^{\alpha|z|}$ for $z \notin S_\theta$ for $\alpha < 1$. We use induction over the so called *increasing domains* defined as follows (cf. [16, 26]): For all positive integers $m \geq 1$ and constants $\xi, \delta > 0$, let

$$\mathcal{F}_m = \{z = \rho e^{i\vartheta} : \ \rho \in [\xi\delta, \xi\nu^{-m}], \ \ 0 \leq \vartheta < 2\pi\}$$

where $\max\{p, q\} \leq \nu < 1$ and $\delta \leq \min\{p, q\}$. The point to observe is that if $z \in \mathcal{F}_{m+1} - \mathcal{F}_m$ then $zp, zq \in \mathcal{F}_m$ provided $|z| \geq \xi$ which is assumed to hold.

To carry out the induction, we first define $\bar{\mathcal{F}}_m = \mathcal{F}_m \cap \bar{S}_\theta$ where $\bar{S}_\theta$ denotes points in the complex plane outside S_θ. Since $X(z)$ is bounded for $z \in \bar{\mathcal{F}}_1$, thus the initial step of induction holds. Let us now assume that for some $m > 1$ and for $z \in \bar{\mathcal{F}}_m$ we have $|X(z)| \leq e^{\alpha|z|}$ with $\alpha < 1$. We intend to prove that $|X(z)| \leq e^{\alpha|z|}$ for $z \in \bar{\mathcal{F}}_{m+1}$. Indeed, let $z \in \bar{\mathcal{F}}_{m+1}$. If also $z \in \mathcal{F}_m$, then the proof is completed. So let us now assume that $z \in \bar{\mathcal{F}}_{m+1} - \bar{\mathcal{F}}_m$. Since then $zp, zq \in \bar{\mathcal{F}}_m$, we can use our induction hypothesis together with the integral equation (54) to obtain the following estimate for $|z| > \xi$ where ξ is large enough

$$|X(z)| \leq |z|^{b+1} \left(e^{|z|(p\alpha + q\cos\theta)} + e^{|z|(q\alpha + p\cos\theta)} + e^{|z|\cos\theta} \right) .$$

Let us now define $1 > \alpha > \cos\theta$ such that the following three inequalities are simultaneously fulfilled

$$\begin{aligned} |z|^b e^{|z|(p\alpha + q\cos\theta)} &\leq \frac{1}{3} e^{\alpha|z|} , \\ |z|^b e^{|z|(q\alpha + p\cos\theta)} &\leq \frac{1}{3} e^{\alpha|z|} , \\ |z|^{b+1} e^{|z|\cos\theta} &\leq \frac{1}{3} e^{\alpha|z|} . \end{aligned}$$

Then, for $z \in \tilde{\mathcal{F}}_{m+1}$ we have $|X(z)| \le e^{\alpha|z|}$, as needed to verify condition (O) of the depoissonization lemma.

In view of the above, we can apply the depoissonization lemma to $\widetilde{X}(z)$, and using our previous asymptotics on $\widetilde{X}(z)$ we immediately obtain

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

Another application of the depoissonization lemma leads to a formula on the second factorial moment gives

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

But

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

hence finally we arrive at

$$\begin{aligned} \text{Var } D_m &= ED_m(D_m-1) + ED_m - (ED_m)^2 \\ &= \frac{h_2-h_1^2}{h_1^3}\log m + 2\sum_{k\neq 0}\frac{\Gamma(s_k)\widehat{\gamma}(s_k)}{h(s_k)}\left(\frac{1}{h(s_k)} - \frac{1}{h_1}\right)m^{-1-s_k}\log m + O(1). \end{aligned}$$

If $\Re(s_k) = -1$ for all k, then $h(s_k) = h_1$, and

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

If $\Re(s_k) > -1$, then $m^{-1-s_k}\log m = o(1)$. Therefore, Var $D_m = \frac{h_2-h_1^2}{h_1^3}\log m + O(1)$. We know that $\Re(s_k) = -1$ whenever $(\log p)/(\log q)$ is rational (cf. [15]), otherwise $\Re(s_k) > -1$. To complete the proof of Theorem 1(i) we must evaluate the constant $\gamma'(-1)$, which is discussed below.

3.3 Evaluation of Some Constants

In several applications (e.g., the computation of the average code redundancy discussed at the end of Section 2) the constant appearing in ED_m plays a very important role. Therefore,

knowing its value, or providing a numerical algorithm to compute it, is of prime interest. In the previous subsection, we have shown that the value of the constant is given in terms of $\widehat{\gamma}'(-1)$, which can be also expressed as

$$\widehat{\gamma}'(-1) = \sum_{i=1}^{b} \binom{b}{i} (-1)^{i+1} \gamma'(-1-i) \ ,$$

where $\gamma(s)\Gamma(s) = \mathcal{M}[\tilde{X}(t); s]$, and $\tilde{X}(z) = \sum_{n=1}^{\infty} f_n \frac{z^n}{n!} e^{-z}$. We recall from Theorem 1 that f_n is defined as

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

Clearly, $\bar{f}_i > 0$ for any $i > b$ since $f_i \geq i - b$ for $i \geq b$.

Throughout this section we assume $b > 1$. To compute $\widehat{\gamma}'(-1)$ we must find $\gamma(s)$ in terms of computable quantities such as f_n. We proceed as follows

$$\begin{aligned} \gamma(s) &= \frac{1}{\Gamma(s)} \mathcal{M}[\sum_{n=b+1}^{\infty} f_n \frac{z^n}{n!} e^{-z}; s] = \sum_{n=b+1}^{\infty} \frac{f_n}{n!} \frac{\Gamma(s+n)}{\Gamma(s)} \\ &= \sum_{n=b+1}^{\infty} \frac{f_n}{n!} s(s+1)...(s+n-1) \ . \end{aligned} \tag{55}$$

We assume above that $\Re(s) \in (-b-2, -1)$ to assure the existence of the Mellin transform and the convergence of the series. Then

$$\gamma'(s) = \sum_{n=b+1}^{\infty} \frac{f_n}{n!} s(s+1) \cdots (s+n-1) \sum_{i=0}^{n-1} \frac{1}{s+i} \qquad s \notin \{-2, -3, \ldots, -b\} \ . \tag{56}$$

After some algebra, we arrive at the following

$$\begin{aligned} \gamma'(-k) &= (-1)^k \sum_{n=b+1}^{\infty} \frac{f_n}{n!} k!(n-k-1)! \quad \text{for} \quad k = 2, ..., b, \\ \gamma'(-b-1) &= (-1)^b H_{b+1} + (-1)^{b+1} \sum_{n=b+2}^{\infty} \frac{f_n}{n!} (b+1)!(n-b-2)! \ . \end{aligned}$$

Let us first assume $b > 1$. Then, to estimate $\widehat{\gamma}'(-1)$ we proceed as follows

$$\begin{aligned} \widehat{\gamma}'(-1) &= \sum_{i=1}^{b} \binom{b}{i} (-1)^{i+1} \gamma'(-i-1) \\ &= -H_{b+1} + \frac{1}{b+1} \sum_{i=1}^{b-1} \frac{i+1}{b-i} + \sum_{i=1}^{b} \binom{b}{i} \sum_{n=b+2}^{\infty} \frac{f_n}{n!} (i+1)!(n-i-2)! \end{aligned}$$

Table 1: Numerical values of $\Delta(b,p)$ and $ED_m - \frac{1}{h_1}\log m$ for $p = 0.3$

b	$\Delta(b,p)$	$ED_m - \frac{1}{h_1}\log m$
1	1.25	-2.05
2	0.96	-3.27
3	0.91	-3.94
5	0.83	-4.76
8	0.76	-5.49
20	0.60	-6.90
50	0.36	-7.92
90	0.12	-8.49

$$= -H_{b+1} - H_{b-1} - \frac{b-1}{b+1} + \sum_{n=b+2}^{\infty}(n-b+\bar{f}_n)\sum_{i=1}^{b}\frac{(i+1)b!}{(b-i)!n(n-1)...(n-i-1)}$$

$$= -\frac{1}{b} - \frac{b}{b+1} + A + \Delta(b,p)$$

where

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

$$A = \sum_{n=b+2}^{\infty}(n-b)\sum_{i=1}^{b}\frac{(i+1)b!}{(b-i)!n(n-1)...(n-i-1)}.$$

The above series converge since the summands are $O(\log n/n^2)$. Finally, observe that $\bar{f}_{m+b} = 0$ for $m = 1,2,...,b$ and $\bar{f}_i > 0$ for $i > 2b$, hence

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

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

After a long and tedious algebra (cf. [36]) we can prove that $A = H_b + b(1+b)^{-1}$, hence $\widetilde{\gamma}'(-1) = H_{b-1} + \Delta(b,p)$ as presented in Theorem 1, and this completes the proof of part (i) of the theorem for $b > 1$.

For $b = 1$ we have

$$\widetilde{\gamma}'(-1) = \gamma'(-2) = -H_2 + \Delta(1,p) + \sum_{n=3}\frac{2}{n(n-2)}$$

$$= \Delta(1,p)$$

since the above series is equal to $3/2$ which is cancelled by $-H_2 = -3/2$. Thus, Theorem 1 is also proved for $b = 1$. Actually, in this case we may also conclude from [34] that

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

In Table 1 we present numerical values of $\Delta(b,p)$ and $ED_m - \frac{1}{h_1}\log m$ as a function of b. While $\Delta(b,p)$ is relatively easy to compute numerically, we must point out that the rate of convergence for this series is only $O(\log N/N)$ where N is the cut-off value of the series computation.

3.4. Limiting Distribution

In this section, we will prove part (ii) of Theorem 1, that is, we establish the central limit theorem for D_m. We recall that $\widetilde{B}(u,z) = \sum_{i=0}^{\infty} B_i(u)\frac{z^i}{i!}e^{-z}$, and

$$\left(1+\frac{\partial}{\partial z}\right)^b \widetilde{B}(u,z) = b + u\left(\widetilde{B}(u,pz) + \widetilde{B}(u,qz)\right) \ . \tag{57}$$

Let for some function $\omega(u,s)$ the Mellin transform of $\widetilde{B}(u,z)$ be given by

$$Z(u,s) = \mathcal{M}\left(\widetilde{B}(u,z) - z; s\right) = \Gamma(s)\omega(u,s) \ . \tag{58}$$

The existence of the Mellin transform $Z(u,s)$ is proved in the lemma below:

Lemma 4 (i) *The Mellin $Z(u,s)$ exists for $\Re(s) \in (-b-1,-1)$.*

(ii) *For $i = 1,\ldots,b-1$ we have $\omega(u,-1-i) = 0$ and $\omega(u,-1-b) = (-1)^{b+1}(u-1)$.*

Proof: The proof uses the same arguments as in Lemma 1. In particular,

$$\begin{aligned}\widetilde{B}(u,z) &= \left(z + z^2 + z^3/2! + \ldots + z^b/(b-1)! + (u+b)z^{b+1}/(b+1)! + O(z^{b+2})\right)e^{-z} \\ &= z + (u-1)z^{b+1}/(b+1)! + O(z^{b+2}) \ ,\end{aligned}$$

thus as $z \to 0$ one obtains $\widetilde{B}(u,z) - z = O(z^{(b+1)})$. For fixed u, we also have $\widetilde{B}(u,z) = O(z\log z)$ for $z \to \infty$. Therefore, part (i) is proved. Part (ii) follows from Lemma 2. ■

The plan for this section is similar to the previous one. We first use Mellin transform technique to derive asymptotics of $\widetilde{B}(z,u) - z$ for $z \to \infty$ in a cone S_θ, then depoissonize this result by Lemma 3. In fact, we follow the footsteps of Jacquet and Szpankowski [14]. We start with taking the Mellin transform to (57). After some algebra, we obtain

$$\sum_{i=0}^{b}\binom{b}{i}(-1)^i\omega(u,s-i) = u(p^{-s}+q^{-s})\omega(u,s) \ ,$$

which further leads to

$$\omega(u,s) = \frac{1}{1-u(p^{-s}+q^{-s})}\widehat{\omega}(u,s) \ .$$

Let now $s_k(u), k = 0, \pm 1, \pm 2, \ldots$ be the roots of the equation $1-u(p^{-s}+q^{-s}) = 0$ for fixed u. Then, for $s = s_k(u)$,

$$\frac{1}{1-u(p^{-s}+q^{-s})} = \frac{1}{s-s_k(u)}\frac{u^{-1}}{-h(s_k(u))} \ .$$

In addition, one must consider two poles of the Gamma function $\Gamma(s)$ at $s_{-1} = -1$ and $s_0 = 0$. The latter pole contribute $O(1)$ while the former $-z\omega(u,-1)$. But, by Lemma 4 we know that $\omega(u,-1) = 1$, thus the total contribution of these two poles is $-z + O(1)$.

Summing up,

$$\widetilde{B}(u,z) = \frac{u^{-1}}{h(s_0(u))}\Gamma(s_0(u))\widehat{\omega}(u,s_0(u))z^{-s_0(u)} + \sum_{k\neq 0}\frac{u^{-1}}{h(s_k(u))}\Gamma(s_k(u))\widehat{\omega}(u,s_k(u))z^{-s_k(u)} + O(1) \ .$$

We now set $u = e^t$ for some complex t in the vicinity of zero. Similar algebra to the ones in [14, 23] leads to the following for $t \to 0$:

$$\begin{aligned} s_0(t) &= -1 - \frac{t}{h_1} - \frac{\alpha t^2}{2} + O(t^3) \ , \\ \Gamma(s_0(t)) &= \frac{h_1}{t} + O(t^2) \ , \\ \frac{e^{-t}}{h(s_0(t))} &= \frac{1}{h_1} + O(t) \ , \\ \widehat{\omega}(t, s_0(t)) &= e^t - 1 + O(t^2) = t + O(t^2) \ . \end{aligned} \tag{59}$$

The rest is a matter of depoissonization. But, the depoissonization conditions (I) and (O) of Lemma 3 are easy to verify for u belonging to a compact set around $u = 1$, as we already shown in the case of $\widetilde{X}(z)$. Thus, an application of (53) provides the following estimate

$$B_m(t) = \frac{1}{h_1}\frac{h}{t}\widehat{\omega}(t,s_0(t))m^{-s_0(t)} + e^{-t}\sum_{k\neq 0}\frac{1}{h(s_k(t))}\Gamma(s_k(t))\widehat{\omega}(u,s_k(t))m^{-s_k(t)} + O(1) \ .$$

Then, the generating function $G_m(t) = Ee^{tD_m}$ becomes

$$\begin{aligned} G_m(t) &= \frac{B_m(t)}{m} \\ &= \frac{1}{t}\widehat{\omega}(t,s_0(t))m^{-1-s_0(t)} + e^{-t}\sum_{k\neq 0}\frac{1}{h(s_k(t))}\Gamma(s_k(t))\widehat{\omega}(u,s_k(t))m^{-1-s_k(t)} + O\left(\frac{1}{m}\right) \\ &= \frac{1}{t}(t-1)m^{-1-s_0(t)} + e^{-t}\sum_{k\neq 0}\frac{1}{h(s_k(t))}\Gamma(s_k(t))\widehat{\omega}(u,s_k(t))m^{-1-s_k(t)} + O\left(\frac{1}{m}\right) \\ &= m^{-1-s_0(t)} + e^{-t}\sum_{k\neq 0}\frac{1}{h(s_k(t))}\Gamma(s_k(t))\widehat{\omega}(u,s_k(t))m^{-1-s_k(t)} + O\left(\frac{1}{m}\right) \ . \end{aligned}$$

As the final step, we set $t = \frac{\tau}{\sigma_m}$ for some fixed τ and $\sigma_m = \text{Var } D_m$. Then, using (59) $m^{-1-s_0(t)} = e^{\tau\mu_m/\sigma_m + \frac{\tau^2}{2}}$, as well as

$$\begin{aligned}
& e^{-\tau\mu_m/\sigma_m} G_m(e^{\tau/\sigma_m}) = \\
= \; & e^{-\tau\mu_m/\sigma_m} \left(\frac{1}{t} t e^{\tau\mu_m/\sigma_m + \frac{\tau^2}{2}} + e^{-t} m^{-1-s_0(t)} \sum_{k\neq 0} \frac{1}{h(s_k(t))} \Gamma(s_k(t)) \widehat{\omega}(u, s_k(t)) m^{s_0(t)-s_k(t)} \right) \\
= \; & e^{\frac{\tau^2}{2}} \left(1 + O\left(\sum_{k\neq 0} \frac{1}{h(s_k(t))} \Gamma(s_k(t)) \widehat{\omega}(u, s_k(t)) m^{s_0(t)-s_k(t)} \right) \right) \\
= \; & e^{\frac{\tau^2}{2}} \left(1 + O\left(\frac{1}{\sqrt{\log m}} \right) \right)
\end{aligned}$$

since , as in [14], we prove that (cf. [36])

$$\sum_{k\neq 0} \frac{1}{h(s_k(t))} \Gamma(s_k(t)) \widehat{\omega}(u, s_k(t)) m^{s_0(t)-s_k(t)} = O(t) = O\left(\frac{1}{\sqrt{\log m}} \right)$$

for $t = \tau/\sigma_m = O(1/\sqrt{\log m})$). In summary, part (ii) of Theorem 1 as well as the theorem as a whole is proven.

4. ANALYSIS OF THE SYMMETRIC BERNOULLI MODEL

In this section we prove Theorem 2 concerning the asymptotic behavior of a b-digital search tree in an unbiased (symmetric) Bernoulli model.

4.1 The Variance

The average value ED_m follows directly from formula (7) of Theorem 1(i). But, in the symmetric case $h_2 = h_1^2 = \log 2$, and therefore from (8) we deduce that Var $D_m = O(1)$. Our goal is to compute it precisely. In this case, an extension of a Flajolet and Richmond technique [8] works fine, and we apply it in this subsection. We follow Hubalek [12] to derive our results. We omit most detailed calculations, and the reader is referred to [8, 12].

First of all, we observe that our differential functional equation (6) becomes in this case

$$\left(1 + \frac{\partial}{\partial z} \right)^b \widetilde{B}(u,z) = b + 2u\widetilde{B}(u, z/2) \; .$$

The coefficients of $\widetilde{B}(u,z)$ can be computed by solving a linear recurrence of type (1). Unfortunately, there is no easy way to solve such a recurrence unless $b = 1$ (cf. [19, 34]). To circumvent this difficulty, Flajolet and Richmond [8] reduced it to a certain functional equation on an *ordinary* generating function that is easier to solve. We proceed along this path.

Let $\widetilde{B}(u,z) = \sum_{k=0}^{\infty} g_k(u)\frac{z^k}{k!}$, and $G(u,z) = \sum_{k=0}^{\infty} g_k(u)z^k$. We also define an ordinary generating function of $B_k(u)$ as $F(u,z) = \sum_{k=0}^{\infty} B_k(u)z^k$. Observe that $B_n(u) = \sum_{k=0}^{n} \binom{n}{k} g_k(u)$, hence as in [8] we obtain

$$F(u,z) = \frac{1}{1-z}G\left(u, \frac{z}{1-z}\right) . \tag{60}$$

Indeed,

$$\begin{aligned}\frac{1}{1-z}G\left(u, \frac{z}{1-z}\right) &= \sum_{m=0}^{\infty} g_m(u)z^m \frac{1}{(1-z)^{m+1}} = \sum_{m=0}^{\infty} g_m(u)z_m \sum_{j=0}^{\infty}\binom{m+j}{j}z^j \\ &= \sum_{n=0}^{\infty} z^n \sum_{k=0}^{\infty}\binom{n}{k} g_k(u) = F(u,z) .\end{aligned}$$

Certainly, (60) further implies that

$$F_u^{(n)}(u,z) = \frac{1}{1-z}G_u^{(n)}\left(u, \frac{z}{1-z}\right) ,$$

where $f_u^{(k)}(z,u)$ denotes the kth derivative of $f(z,u)$ with respect to u. Then

$$G(u,z)(1+z)^b = z(1+z)^b - z^{b+1} + 2uz^b G(u, \frac{z}{2}) , \tag{61}$$

$$G'_u(u,z)(1+z)^b = 2z^b G(u, \frac{z}{2}) + 2uz^b G'_u(u, \frac{z}{2}) , \tag{62}$$

$$G''_u(u,z)(1+z)^b = 4z^b G'_u(u, \frac{z}{2}) + 2uz^b G''_u(u, \frac{z}{2}) . \tag{63}$$

In order to compute the variance, we compute $L^1(z) := G'_u(u,z)|_{u=1}$ and $L^2(z) := G''_u(u,z)|_{u=1}$, and then use (60). From (62) and (63) we immediately obtain

$$\begin{aligned}L^1(z)(1+z)^b &= z^{b+1} + 2z^b L^1(\frac{z}{2}), \\ L^2(z)(1+z)^b &= 4z^b L^1(\frac{z}{2}) + 2z^b L^2(\frac{z}{2}).\end{aligned}$$

Iterating these equations we easily find (cf. [8, 12])

$$L^1(z) = \sum_{k=0}^{\infty} \frac{(2z^b)(2(\frac{z}{2})^b)\cdots(2(\frac{z}{2^k})^b)}{\left((1+z)(1+\frac{z}{2})\cdots(1+\frac{z}{2^k})\right)^b}\frac{z}{2^{k+1}} , \tag{64}$$

$$L^2(z) = \sum_{k=0}^{\infty} \frac{(2z^b)(2(\frac{z}{2})^b)\cdots(2(\frac{z}{2^k})^b)}{\left((1+z)(1+\frac{z}{2})\cdots(1+\frac{z}{2^k})\right)^b} 2L^1(\frac{z}{2^{k+1}}) . \tag{65}$$

The next step is to transform the above sums (64)–(65) into certain harmonic sums (cf. [9]). For this, we set $z = 1/t$ and define $Q(t) = \prod_{k=0}^{\infty}(1+\frac{t}{2^k})$. Then (64)–(65) become

$$\frac{tL^1(\frac{1}{t})}{(Q(\frac{t}{2}))^b} = \sum_{k=0}^{\infty}\frac{1}{(Q(2^k t))^b} , \tag{66}$$

$$\frac{tL^2(\frac{1}{t})}{(Q(\frac{t}{2}))^b} = 2\sum_{k=0}^{\infty}\frac{2^{k+1}tL^1(\frac{1}{2^{k+1}t})}{\left(Q(\frac{2^{k+1}t}{2})\right)^b} . \tag{67}$$

Both sums are of the following form $\sum_{k\geq 0}\lambda_k f(\mu_k x)$ for some function $f(\cdot)$ and sequences λ_k, μ_k, that is, they fall under the so called *harmonic sums* (cf. [9]). It is well known that the Mellin transform of such a sum is $f(s)\sum_{k\geq 0}\lambda_k\mu_k^{-s}$ (cf. [9]). In our case, we have

$$\begin{aligned}\mathcal{M}\left[\frac{tL^1(\frac{1}{t})}{Q^b(\frac{t}{2})};s\right] &= \frac{1}{1-2^{-s}}I(s)\ ,\\ \mathcal{M}\left[\frac{tL^2(\frac{1}{t})}{Q^b(\frac{t}{2})};s\right] &= \frac{2^{1-s}}{(1-2^{-s})^2}I(s)\ ,\end{aligned}$$

where

$$\begin{aligned}I(s) &= \int_0^\infty \frac{t^{s-1}}{Q^b(t)}dt = \frac{\pi}{\sin\pi s}J(s)\ ,\\ J(s) &= \frac{1}{2\pi i}\int_{\mathcal{H}}\frac{(-t)^{s-1}}{Q^b(t)}dt\end{aligned}$$

with $\mathcal{H}$ being the Hankel contour (cf. [9, 12]).

The rest is easy. Applying standard arguments of the inverse Mellin transform we can derive asymptotic expansions of $L^1(\frac{1}{t})$ and $L^2(\frac{1}{t})$ as $t\to 0$. We find

$$\begin{aligned}L^1(\frac{1}{t}) &= \frac{1}{t}k(t) + bk(t) + O(t\log t^{-1}),\\ L^2(\frac{1}{t}) &= \frac{1}{t}K(t) + bK(t) + O(t\log^2 t^{-1})\ ,\end{aligned}$$

where

$$\begin{aligned}k(t) &= \frac{1}{L}\log\frac{1}{t} + \frac{1}{2} + \frac{J'(0)}{L} - \frac{1}{L}\sum_{k=0}\frac{I(s_k)}{s_k}t^{-s_k},\\ K(t) &= \frac{1}{L^2}\log^2\frac{1}{t} + \frac{2J'(0)}{L^2}\log\frac{1}{t} - (\frac{1}{6} + \frac{J''(0)}{L^2} - \frac{\pi^2}{3L^2}) + 8bt\\ &- \frac{2}{L^2}\sum_{k=0}\frac{I(s_k)}{s_k}t^{-s_k}\log\frac{1}{t} + \frac{2}{L^2}\sum_{k=0}(\frac{I(s_k)}{s_k^2} - \frac{I'(s_k)}{s_k})t^{-s_k}\ ,\end{aligned}$$

with $s_k = 2\pi ik/\log 2$ for $k = 0, \pm 1, \ldots$ are roots of $1-2^{-s}=0$, and $L = \log 2$. Finally, applying the *singularity analysis* of Flajolet and Odlyzko [7], after somewhat tedious algebra we prove formula (15) of Theorem 2(i).

4.2 Exact and Limiting Distribution

We need another approach to establish exact and asymptotic distributions in the symmetric case since as shown above Var $D_m = O(1)$. We also point out that – even it is possible in principle – using recurrence (5) or functional equation (6) may be quite troublesome. Therefore, we devised another, more combinatorial and probabilistic approach.

Let us fix $j \geq 1$, and consider a *particular path*, say $\mathcal{P}$, from the root to a node at level j on $\mathcal{P}$. Let $T_{j,r}$ be the number of strings needed to be added to the tree (after the first b) to assure that a node at level j contains exactly r strings ($1 \leq r \leq b$). Since the first b strings are stored in the root, thus we observe the following:

$$\Pr\{T_{j,r} \leq m-b\} = \Pr\{\text{node at level } j \text{ contains at least } r \text{ strings when } m \text{ strings are in the tree}\} .$$

Observe that $P[j,r] := \Pr\{\text{exactly } r \text{ strings are in a node at level } j \text{ when } m \text{ strings are added}\} = \Pr\{T_{j,r} \leq m-b\} - \Pr\{T_{j,r+1} \leq m-b\}$.

Then, the distribution of D_m can be computed as

$$\Pr\{D_m = j\} = \frac{2^j}{m} \sum_{r=1}^{b} P[j,r] \cdot r = \frac{2^j}{m} \sum_{r=1}^{b} \Pr\{T_{j,r} \leq m-b\} . \tag{68}$$

In view of the above, to compute the exact distribution of D_m one needs the distribution of $T_{r,j}$. But, the number of strings, say X_i, that one must insert into the tree in order to fill up a node at level $i < j$ on the path $\mathcal{P}$ (when the node on $\mathcal{P}$ at level $i-1$ is full) is distributed as the sum of b independent random variables geometrically distributed with success probability $\pi(i) = 2^{-i}$. Let $X_i(z) = Ez^{X_i}$ be the probability generating function. Then

$$X_i(z) = \left(\frac{\pi(i)z}{1-(1-\pi(i))z}\right)^b \quad \text{for} \quad i < j ,$$

Similarly, the probability generating function for the number of strings needed to get exactly r strings in a given node at level j (when the node on $\mathcal{P}$ at level $j-1$ is full) is given by

$$X_j(z) = \left(\frac{\pi(j)z}{1-(1-\pi(j))z}\right)^r .$$

Summing up, the probability generating function $T_{j,r}(z)$ of $T_{j,r}$ is

$$T_{j,r}(z) = X_j(z) \prod_{i=1}^{j-1} X_i(z) . \tag{69}$$

To compute the required probabilities, we first use the Cauchy formula

$$\Pr\{T_{j,r} = \ell\} = \frac{1}{2\pi i} \oint \frac{T_{j,r}(z)}{z^{\ell+1}} dz ,$$

and then the residue theorem. The calculations are rather straightforward but quite tedious. We find

$$\begin{aligned} m\Pr\{D_m \leq j\} &= b - \frac{1}{(b-1)!} \sum_{k=1}^{j} \left(\frac{\pi(k)}{\pi(k)-1}\right)^b \frac{\partial^{b-1}}{\partial z^{(b-1)}} \left\{ \frac{z^{2b}}{(z-1)^2}(z^{-b}-z^{-m}) \right. \\ &\quad \left. \cdot \prod_{v=1, v\neq k}^{j} \left(\frac{\pi(v)z}{1-(1-\pi(v))z}\right)^b \right\}_{z=z^*(k)} , \end{aligned} \tag{70}$$

where $z^*(k) = (1-\pi(k))^{-1}$. They lead to formula (18) in Theorem 2(ii) for the exact distribution for D_m.

The asymptotic formula of part (iii) of Theorem 2 follows from the above after some algebra that we summarize below. We set throughout this derivation $j = \log_2 m + \eta$ with $\eta = O(1)$, and $k = j + O(1)$ which we justify below. After substituting $\eta = \kappa - \{\log_2 m\}$ we prove part (iii) of Theorem 2.

Let us now analyze (70). The term involving z^{-m} in (70) becomes:

$$H_1 := \frac{\pi^b(k)}{m(b-1)!(\pi(k)-1)^b}\frac{\partial^{b-1}}{\partial z^{b-1}}\left(z^{-(m-2b)}(z-1)^{-2}\varphi_1(z)\varphi_2(z)\right)|_{z=z^*(k)} ,$$

where

$$\begin{aligned}\varphi_1(z) &= \prod_{v=1}^{k-1}\left(\frac{\pi(v)z}{1-(1-\pi(v))z}\right)^b ,\\ \varphi_2(z) &= \prod_{v=k+1}^{j}\left(\frac{\pi(v)z}{1-(1-\pi(v))z}\right)^b .\end{aligned}$$

After using Leibniz's rule for differentiation, we obtain

$$\sum_{\ell+s+s_1+s_2=b-1}\binom{b-1}{\ell,s,s_1,s_2}\cdot\frac{(-1)^{\ell+s-b}\pi^b(k)(m-2b)_\ell(-1)^s(s+1)!(1-\pi(k))^{m-3b+\ell+2+s}}{(m(b-1)!\pi(k))^{2+s}}\cdot\left(\varphi_1^{(s_1)}(z)\varphi_2^{(s_2)}(z)\right)|_{z^*(k)} , \tag{71}$$

where $f^{(k)}(z)$ denotes the kth derivative of $f(z)$, and $(m)_\ell = m(m-1)\cdots(m-\ell+1)$.

Let now, as announced above, set $j = \log_2 m + \eta$ and $i = j - k$ where $i = O(1)$. We obtain:

$$\frac{(\pi(k))^{\ell-1}(m-2b)_\ell}{m} \sim (m\pi(k))^{\ell-1} = \frac{2^{i(\ell-1)}}{2^{\eta(\ell-1)}}, \tag{72}$$

and

$$(1-\pi(k))^{m-3b+\ell+2+s} \sim e^{-2^{-(\eta-i)}} .$$

To compute the derivatives of $\varphi_1(z)$ and $\varphi_2(z)$ we observe, for example, that for any integer r,

$$Y := \frac{\partial^r}{\partial z^r}\left(\frac{\pi(v)}{1-(1-\pi(v))z}\right)^b = \frac{(\pi(v))^b(1-\pi(v))^r b_r}{(1-(1-\pi(v)z)^{b+r}} .$$

Setting now $z = z^*(k)$ and $v = k + O(1)$, we find

$$Y \sim \frac{b-r}{(\pi(k))^r(1-2^{-u})^{b+r}2^{ur}} \qquad \text{in the } \varphi_1 \text{ case} \tag{73}$$

where $u = k - v > 0$, and

$$Y \sim \frac{b_r 2^{ur}}{(\pi(k))^r (1-2^u)^{b+r}} \qquad \text{in the } \varphi_2 \text{ case .} \tag{74}$$

To deal with expressions like (73) or (74), we define

$$H(s) = \frac{\partial^s}{\partial z^s} \prod_{k=1}^{\infty} \left(\frac{1}{1-\pi(k)z} \right)^b |_{z=1} \ ,$$

and with $R(0,s) = -1$,

$$R(i,s) = -\frac{\partial^s}{\partial z^s} \prod_{k=1}^{i} \left(\frac{1}{1-\frac{z}{\pi(k)}} \right)^b \Bigg|_{z=1} \ ,$$

which are exactly (13) and (14) from Section 2.

These expressions are the b-equivalent of $Q^{-1}(t)$ (cf. (12) of Section 2), and function $|R_i|$ used in [21, 23] (cf. (30) of [23]) parametrized by s. Clearly $R(i,s)$ decreases exponentially with i and $H(s)$ is uniformly bounded, which justify our choice $k = j + O(1)$ for asymptotic analysis. Moreover, any term $(1-\pi(v))^r$ $(v < k)$ leads to a contribution $(1-2^{u-k})^r 2^{-ur}(1-2^{-u})^{-r}$. The sum of all these contributions is $O(1)$ which shows that we can asymptotically take $(1-\pi(v)) \sim 1$.

Let us return to (70). We can extract a term $(\pi(k))^{b-2-s-s_1-s_2} = (\pi(k))^{\ell-1}$ and, with (72), after summing over k we obtain

$$H_1 \sim \sum_{l+s+s_1+s_2=b-1} \frac{(-1)^{s_1+s_2}}{l! s_1! s_2!} \sum_{i=0}^{\infty} (s+1) R(i,s_2) H(s_1) \frac{2^{i(\ell-1)}}{2^{\eta(\ell-1)}} e^{-2^{-(\eta-i)}} \ .$$

Similar analysis is valid for the term at z^{-b} of (70). Finally, after substituting $\eta = \kappa - \{\log_2 m\}$ we prove part (iii) of Theorem 2, which completes the proof of Theorem 2.

ACKNOWLEDGEMENT

We thank Professor Helmut Prodinger for many valuable comments regarding this research. We are particularly obliged to one of the referees whose very careful reading of the paper allow us to eliminate many inaccuracies and led to a better presentation of our results.

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

[4] B. Davies, *Integral Transforms and Their Applications*, Springer–Verlag, New York (1985).

[5] G. Doetsch, *Handbuch der Laplace Transformation*, Verlag Birkhäuser, Basel (1950).

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

[7] P. Flajolet and A. Odlyzko, Singularity Analysis of Generating Functions, *SIAM J. Disc. Methods*, 3, 216–240 (1990).

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

[9] P. Flajolet, X. Gourdon, P. Dumas, Mellin Transforms and Asymptotics: Harmonic Sums, *Theoretical Computer Science*, 144, 3-58, (1995).

[10] P. Flamant, Sur une Equation Différentielle Fonctionelle Linéaire, *Rendiconti del Circolo Mathematico di Palermo*, XLVIII, 135—208 (1924).

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

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

[13] K. Hummelsheim and C. Kleiner, Project in CS 543 "Analysis of a Data Compression Algorithm", Purdue University, Department of Computer Science (1996).

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

[15] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161–197 (1995).

[16] P. Jacquet and W. Szpankowski, Analytical Depoissonization Lemma and Its Applications, Purdue University, CSD-TR-96-62 (1996).

[17] S. Janson and W. Szpankowski, Analysis of an Asymmetric Leader Election Algorithm, Purdue University, CSD-TR-96-049 (1996).

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

[19] D. Knuth, *The Art of Computer Programming. Sorting and Searching.* Vol. 3., Addison-Wesley (1973).

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

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

[22] G. Louchard, Digital Search Trees Revisited, *Cahiers du CERO*, 36, 259–27 (1995).

[23] 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).

[24] 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).

[25] G. Louchard and W. Szpankowski, On the Average Redundancy Rate of the Lempel-Ziv Code, *IEEE Information Theory*, 43, 1–7 (1997).

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

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

[28] B. Pittel, Asymptotic Growth of a Class of random Trees, *Ann. Probab.*, 13, 414 – 427 (1985).

[29] H. Prodinger, Approximate Counting via Euler Transform, *Math. Slovaka*, 44, 569–574 (1994).

[30] H. Prodinger, Digital Search Trees and Basic Hypergeometric Functions, *EATCS Bulletin*, 56, 112–115 (1995).

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

[32] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. Information Theory*, 43, (1997).

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

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

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

[36] J. Tang, *Probabilistic Analysis of Digital Search Trees*, Ph.D. Thesis, Purdue University (1996).

[37] 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).

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

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