Purdue University

Purdue e-Pubs

Department of Computer Science Technical
Reports

Department of Computer Science

1999

Asymptotic Behavior of the Height in a Digital Search Tree and the Longest Phrase of the Lempel-Ziv Scheme

Charles Knessl

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

Report Number:

99-019

Knessl, Charles and Szpankowski, Wojciech, "Asymptotic Behavior of the Height in a Digital Search Tree and the Longest Phrase of the Lempel-Ziv Scheme" (1999). *Department of Computer Science Technical Reports*. Paper 1450.
<https://docs.lib.purdue.edu/cstech/1450>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries.
Please contact epubs@purdue.edu for additional information.

**ASYMPTOTIC BEHAVIOR OF THE HEIGHT
IN A DIGITAL SEARCH TREE AND THE
LONGEST PHRASE OF THE LEMPEL-ZIV SCHEME**

**Charles Knessl
Wojciech Szpankowski**

**CSD TR #99-019
May 1999**

Asymptotic Behavior of the Height in a Digital Search Tree and the Longest Phrase of the Lempel-Ziv Scheme

May 25, 1999

Charles Knessl*

Dept. Mathematics, Statistics & Computer Science
University of Illinois at Chicago

Chicago, Illinois 60607-7045

U.S.A

knessl@uic.edu

Wojciech Szpankowski†

Department of Computer Science
Purdue University

W. Lafayette, IN 47907

U.S.A.

spa@cs.purdue.edu

Abstract

We study the height of a *digital search tree* (DST in short) built from n random strings generated by an unbiased memoryless source (i.e., all symbols are equally likely). We shall argue that the height of such a tree is equivalent to the length of the longest phrase in the Lempel-Ziv parsing scheme that partitions a random sequence into n phrases. We also analyze the longest phrase in the Lempel-Ziv scheme in which a string of *fixed* length m is parsed into a *random* number of phrases. In the course of our analysis, we shall identify four natural regions of the height distribution and characterize them asymptotically for large n . In particular, for the region where most of the probability mass is concentrated, the asymptotic distribution of the height exhibits an exponential of a Gaussian distribution (with an oscillating term) around the most probable value $k_1 = \lfloor \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \rfloor + 1$. More precisely, we shall prove that the asymptotic distribution of a digital search tree is either concentrated on the one point k_1 or the two points $k_1 - 1$ and k_1 , which actually proves (slightly modified) Kesten's conjecture quoted in [2]. Finally, we compare our findings for DST with the asymptotic distributions of the height (recently obtained by us) for other digital trees such as tries and PATRICIA tries. We derive these results by a combination of analytic methods such as generating functions, Laplace transform, the saddle point method and ideas of applied mathematics such as linearization, asymptotic matching and the WKB method. We also present detailed numerical verification of our results.

Key Words: Digital search trees, Lempel-Ziv algorithm, height distribution, longest phrase distribution, Laplace transform, saddle point method, matched asymptotics, linearization, WKB method, elliptic theta function.

*This work was supported by DOE Grant DE-FG02-96ER25168.

†The work of this author was supported by NSF Grants NCR-9415491 and CCR-9804760.

1 Introduction

The heart of some universal data compression schemes is the parsing algorithm due to Lempel and Ziv [32]. It partitions a sequence into phrases (blocks) of variable sizes such that a new block is the shortest substring not seen in the past as a phrase. For example, the string 110010100010001000 is parsed into (1)(10)(0)(101)(00)(01)(000)(100). This parsing algorithm plays a crucial role in numerous applications such as efficient transmission of data discriminating between information sources, test of randomness, estimating the statistical model of individual sequences, and so forth. The parameters of interest in these applications are: the number of phrases, the number of phrases of a given size, the size of a phrase, the length of a sequence built from a given number of phrases, the length of the longest phrase, etc. Some of them have already been analyzed (e.g., number of phrases [2, 13], the size of a typical phrase [15, 23]). Here we study the distribution of the longest phrase.

The Lempel-Ziv parsing scheme can be efficiently implemented by a special digital tree called the *digital search tree* (in short DST). This finds a myriad of applications in computer science and telecommunications such as dynamic hashing, partial match retrieval of multidimensional data, searching and sorting, pattern matching, conflict resolution algorithms for broadcast communications, data compression, coding, security, genes searching, DNA sequencing, genome maps, and so forth. The digital search tree is constructed as follows (cf. Figure 1). We consider n , possibly infinite, strings of symbols from a finite alphabet Σ (however, for the simplicity we work only with the binary alphabet $\Sigma = \{0, 1\}$). The empty string is stored in the root, while the first string occupies the right or the left child of the root depending whether its first symbol is “1” or “0”. The remaining strings are stored in available nodes (that are directly attached to nodes already existing in the tree). The search for an available node follows the prefix structure of a string. The rule is simple: if the next symbol in a string is “1” we move to the right, otherwise we move to the left. The resulting tree has $n + 1$ internal nodes. The details can be found in [17] and [24].

The parsing scheme according to Lempel-Ziv with a *fixed* number, n , of phrases (cf. [11, 13, 23]) can be accomplished on the associated digital search tree as follows: We consider an infinite sequence of binary symbols and partition it until we create the first n phrases. Assuming the first phrase is an *empty* one, we store it in the root of a digital search tree, and all other phrases are stored in internal nodes. When a new phrase is created, the search starts at the root and proceeds down the tree as directed by the input symbols exactly in the same manner as in the digital search tree construction. For example, for the binary alphabet, “0” in the input string means move to the left and “1” means proceed to

Figure 1: Digital search trees built from: (left) five strings $X_1 = 11100\dots$, $X_2 = 11011\dots$, $X_3 = 00110\dots$, $X_4 = 00001\dots$, $X_5 = 001011\dots$; (right) eight Lempel-Ziv phrases of the string $11001010001000100\dots$, that is, $(1)(10)(0)(101)(00)(01)(000)(100)$.

the right. The search is completed when a branch is taken from an existing tree node to a new node that has not been visited before. Then, an edge and a new node are added to the tree. Phrases created in such a way are stored directly in the nodes of the tree (cf. Figure 1).

We also study another model, called the *Lempel-Ziv model* in which a string of *fixed* length m is parsed according to the Lempel-Ziv algorithm. We can again use the associated digital search tree to parse the string but this time the number of phrases, M_m , and hence the number of nodes in the DST, is random. It is known (cf. [2, 32]) that almost surely the number of phrases $M_m \sim mh/\log m$ (a.s.) where h is the entropy of the source.

In this paper, we consider digital search trees built over n randomly and independently generated strings of binary symbols. We assume that every symbol is equally likely, thus we are within the framework of the so called *symmetric Bernoulli* model. In other words, the strings are emitted by an *unbiased memoryless* source. Our interest lies in establishing the *asymptotic distribution* of the height, $\mathcal{H}_n$, for random DST. The height is the longest path in such trees, and its distribution is of considerable interest for several applications (e.g., the length of the longest phrase in the Lempel-Ziv scheme, maximum search time, and sorting).

We now summarize our main results. First of all, as a consequence of our analysis we prove that the average height is $\mathbb{E}[\mathcal{H}_n] = \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + O(1)$. However, our main contribution lies in establishing the asymptotic distribution $\Pr\{\mathcal{H}_n \leq k\}$ of

the height as $n \rightarrow \infty$. In Theorem 1 we shall identify four natural regions of this distribution and characterize them asymptotically as n and k become large. In particular, we show that for the region where most of the probability mass is concentrated, the asymptotic distribution of the height exhibits an exponential of a Gaussian distribution (with an oscillating term) around the most probable value $k_1 = \lfloor \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \rfloor + 1$. In fact, we shall prove that the asymptotic distribution of a digital search tree is either concentrated on the one point k_1 or the two points $k_1 - 1$ and k_1 . More precisely, we show the existence of certain subsequences of n such that the asymptotic distribution of the height concentrates only on k_1 , or on k_1 and $k_1 - 1$. This proves (slightly modified) Kesten's conjecture quoted in Aldous and Shields [2] on page 538. Finally, we obtain the asymptotic distribution of the length of the longest phrase in the Lempel-Ziv model by showing that one must replace n in the digital tree model by $m/\log_2 m$ (cf. Theorem 2).

Digital trees, that is, tries, PATRICIA tries, and digital search trees have been extensively analyzed in the past (cf. [2, 5, 7, 8, 13, 16, 17, 20, 22, 23, 24, 26, 29]). However, relatively little is known about the height of digital search trees. An exception is the paper by Pittel [26] who proved that almost surely the height $\mathcal{H}_n \sim \log_2 n$ (a.s.). Later Aldous and Shields [2] improved Pittel's result to $\mathcal{H}_n \sim \log_2 n + \sqrt{2 \log_2 n}$ (a.s.). No distributional result for the height is known to us. We fill this gap by presenting a complete characterization of the asymptotic distribution of the height.

Finally, we say a few words about our method of derivation, and put our results in a larger perspective. From a mathematical view point, we study a non-linear recurrence equation. The distribution $h_n^k = \Pr\{\mathcal{H}_n \leq k\}$ satisfies

$$h_{n+1}^{k+1} = \sum_{i=0}^n \binom{n}{i} 2^{-n} h_i^k h_{n-i}^k, \quad k \geq 0$$

with the initial condition $h_0^0 = h_1^0 = 1$ and $h_n^0 = 0$ for $n \geq 2$. We shall solve asymptotically this recurrence by methods of applied mathematics such as linearization, matched asymptotics and the WKB method. In passing, we mention that in a companion paper [19] we solved two similar recurrences: The so called b -tries recurrence satisfies

$$h_n^{k+1} = 2^{-n} \sum_{i=0}^n \binom{n}{i} h_i^k h_{n-i}^k, \quad k \geq 0$$

with the initial condition $h_n^0 = 1$ for $n = 0, 1, 2, \dots, b$ and $h_n^0 = 0$ for $n > b$. For PATRICIA tries the distribution $h_n^k = \Pr\{\mathcal{H}_n^P \leq k\}$ of the height $\mathcal{H}_n^P$ satisfies

$$h_n^{k+1} = 2^{-n+1} h_n^{k+1} + 2^{-n} \sum_{i=1}^{n-1} \binom{n}{i} h_i^k h_{n-i}^k, \quad k \geq 0$$

with the initial conditions $h_0^0 = h_1^0 = 1$ and $h_n^0 = 0$ for $n \geq 2$. Surprisingly enough, although these recurrences resemble the digital search tree recurrence, they lead to quite different solutions. The reader is referred to [19] for details.

The paper is organized as follows. In the next section, we present our main results for digital search trees (cf. Theorem 1) and the Lempel-Ziv model (cf. Theorem 2). In Section 3 we present detailed numerical results and discuss consequences of our findings. The proofs of these results are relegated to Sections 4 and 5.

2 Main Results

We let $\mathcal{H}_n$ be the height of a digital search tree and we denote its probability distribution by

$$h_n^k = \Pr\{\mathcal{H}_n \leq k\}. \quad (2.1)$$

It satisfies the following difference equation

$$h_{n+1}^{k+1} = \sum_{i=0}^n \binom{n}{i} 2^{-n} h_i^k h_{n-i}^k, \quad k \geq 0 \quad (2.2)$$

with the initial condition

$$h_0^0 = h_1^0 = 1; \quad h_n^0 = 0, n \geq 2. \quad (2.3)$$

This follows from $\mathcal{H}_{n+1} = \max\{\mathcal{H}_i^{LT}, \mathcal{H}_{n-i}^{RT}\} + 1$, where $\mathcal{H}_i^{LT}$ and $\mathcal{H}_{n-i}^{RT}$ denote, respectively, the left subtree and the right subtree of sizes i and $n-i$, which happens with probability $2^{-n} \binom{n}{i}$. The root contains one string (or an empty string).

We can easily show that $h_n^k = 0$ for $n \geq 2^{k+1}$ (i.e., a balanced tree) and $h_n^k = 1$ for $k \geq n-1$ (i.e., a skewed tree). It therefore suffices to consider the range $k+2 \leq n \leq 2^{k+1}-1$.

We also note that the exponential generating function

$$H_k(z) = \sum_{n=0}^{\infty} \frac{h_n^k z^n}{n!} = \sum_{n=0}^{2^{k+1}-1} \frac{h_n^k z^n}{n!} = \sum_{n=0}^{k+1} \frac{z^n}{n!} + \sum_{n=k+2}^{2^{k+1}-1} h_n^k \frac{z^n}{n!} \quad (2.4)$$

satisfies

$$H'_{k+1}(2z) = H_k^2(z), \quad k \geq 0 \quad (2.5)$$

with $H_0(z) = 1 + z$. Thus, for any k , the generating function $H_k(z)$ is a polynomial of degree $2^{k+1} - 1$ and the leading $k+2$ coefficients in this polynomial are the same as those in the Taylor series of e^z .

Below we give the exact expressions for h_n^k for a few values of n that are close to either k or 2^{k+1} :

$$h_n^{n-2} = 1 - 2^{-n^2/2} 2^{3n/2} 2^{-1}, \quad n \geq 2; \quad (2.6)$$

$$h_n^{n-3} = 1 - (n-2) 2^{-n^2/2} 2^{5n/2} 2^{-3}, \quad n \geq 4; \quad (2.7)$$

$$h_n^{n-4} = 1 - \left(\frac{n^2}{2} - \frac{5n}{2} + \frac{8}{3} - \frac{2}{3} 4^{5-n} \right) 2^{-n^2/2} 2^{7n/2} 2^{-6}, \quad n \geq 6; \quad (2.8)$$

and

$$h_{2^{k+1}-1}^k = (2^{k+1}-1)! 2^{-k2^{k+1}} 4^{2^k-1} \prod_{\ell=1}^k \left(\frac{1}{2^{\ell+1}-1} \right)^{2^{k-\ell}}, \quad k \geq 0, \quad (2.9)$$

$$h_{2^{k+1}-2}^k = (2^{k+1}-2)! 2^{-(2^{k+1})2^k} 2^{3^{k+1}} \times \prod_{\ell=1}^k \frac{1}{1-2^{-\ell}} \left[\prod_{m=1}^{\ell-1} (1-2^{-m-1})^{-2^{-m}} \right]^{2^{\ell-1}}, \quad k \geq 1. \quad (2.10)$$

We next consider the asymptotic limit $n \rightarrow \infty$. A singular perturbation analysis of the problem (2.2) shows that there are four cases of k that lead to different asymptotic expansions of h_n^k . These are: $n \rightarrow \infty$ with $n-k$ fixed; $n, k \rightarrow \infty$ with $0 < k/n < 1$; $n, k \rightarrow \infty$ with $\xi = n2^{-k}$ fixed and $0 < \xi < 2$; and $n, k \rightarrow \infty$ with $2^{k+1} - n = O(1)$. We note that the last limit is only possible if n is close to a power of 2.

Using ideas of applied mathematics, such as linearization and asymptotic matching, we obtain the following results, listed in decreasing size of k/n . The derivation of these results is presented in Sections 4 and 5 where we make certain assumptions about the forms of the asymptotic expansions, as well as the asymptotic matching between the various scales.

Theorem 1 *The distribution of the height for digital search trees built from n independent strings generated by an unbiased memoryless source (and thus the length of the longest phrase in the Lempel-Ziv algorithm with a fixed number, n , of phrases) has the following asymptotic expansions:*

(i) FAR RIGHT-TAIL REGIME: $n, k \rightarrow \infty$, $n-k = j = O(1)$, $j \geq 2$

$$\Pr\{\mathcal{H}_n > k\} = 1 - h_n^k \sim 2^{-j^2/2} 2^{j/2} \frac{1}{(j-2)!} n^{j-2} 2^{-n^2/2} 2^{(j-1/2)n}. \quad (2.11)$$

(ii) RIGHT-TAIL REGIME: $n, k \rightarrow \infty$, $0 < k/n < 1$ or $\alpha \equiv 1/(1-k/n) = n/(n-k) \in (1, \infty)$

$$1 - h_n^k \sim 2^{-k^2/2} 2^{-k/2} F_n^k, \quad (2.12)$$

where

$$F_n^k = \frac{n^n}{(n-k)^{n-k} k^k} \frac{\sqrt{\pi n} (n-k)^{3/2}}{\sqrt{2} k^{5/2}} I(\alpha), \quad (2.13)$$

$$\begin{aligned} I(\alpha) &= \frac{1}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^m}{(\alpha-1)^m} \left[\prod_{\ell=1}^m (1-2^{-\ell}) \right]^{-1}, \quad \alpha > 2, \\ &= \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} \frac{e^{z \log(\alpha-1)}}{\sin(\pi z)} \prod_{\ell=1}^{\infty} \exp\left(\frac{1-2^{\ell z}}{\ell(2^{\ell}-1)}\right) dz, \quad \alpha > 1. \end{aligned} \quad (2.14)$$

(iii) CENTRAL REGIME: $n, k \rightarrow \infty$, $\xi = n2^{-k} \in (0, 2)$

$$h_n^k \sim A(\xi) e^{-n\Phi(\xi)} \quad (2.15)$$

$$A(\xi) = e^{-\xi\Phi'(\xi)-\Phi(\xi)} \sqrt{1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)}, \quad (2.16)$$

where the function $\Phi(\xi)$ satisfies, for $\xi \rightarrow 2^-$,

$$\Phi(\xi) \sim \frac{1}{2} \log\left(\frac{e^2}{2C_0}\right) + \frac{1}{2}(2-\xi) \log(2-\xi) + \frac{1}{4}(\xi-2) \log(2C_0), \quad (2.17)$$

$$C_0 = \prod_{\ell=1}^{\infty} (1-2^{-\ell-1})^{-2^{-\ell}} = 1.20675\dots, \quad (2.18)$$

and for $\xi \rightarrow 0^+$

$$\begin{aligned} \Phi(\xi) &\sim 2^{-9/8} \frac{(\log 2)^{3/2}}{(\log \xi)^2} \xi^{1/2-1/\log 2} \exp\left(-\frac{[\log \xi - \log(-\log_2 \xi)]^2}{2 \log 2}\right) \\ &\times Q_*\left(\log_2 \xi - \log_2(-\log_2 \xi) + \frac{1}{2}\right). \end{aligned} \quad (2.19)$$

Here $Q_*(z)$ is a periodic function of period one

$$Q_*(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2/2} Q\left(z + \frac{is}{\sqrt{\log 2}}\right) ds$$

where $Q(z)$ has the Fourier series

$$Q(z) = 2^{13/12} \prod_{\ell=1}^{\infty} (1-2^{-\ell})^{-1} \exp\left[-\frac{\pi^2}{3 \log 2} - \sum_{m=1}^{\infty} \frac{\cos(2m\pi z)}{m \sinh(2\pi^2 m / \log 2)} \exp\left(-\frac{2\pi^2 m}{\log 2}\right)\right]. \quad (2.20)$$

An alternate representation is

$$Q(z) = \frac{e^{-\pi iz}}{\sin(\pi z)} e^{p(z)}$$

where

$$e^{p(z)} = e^{-(\log 2)z^2/2} e^{\pi i(z+\frac{3}{2})} 2^{1/8} \left[\prod_{\ell=1}^{\infty} (1-2^{-\ell})^{-2} \right] \vartheta_1\left(\frac{i}{2}z \log 2\right)$$

and $\vartheta_1(u)$ is the Jacobi theta function defined by (cf. [3])

$$\begin{aligned}\vartheta_1(u) &= \vartheta_1(u|\tau) = 2 \sum_{m=1}^{\infty} (-1)^{m+1} q^{(m-1/2)^2} \sin((2m-1)u) \\ &= 2q^{1/4} \sin(u) \prod_{m=1}^{\infty} (1 - 2\cos(2u)q^{2m} + q^{4m})(1 - q^{2m})\end{aligned}\quad (2.21)$$

where $q = 1/\sqrt{2} = e^{\pi i \tau}$ with $\tau = (i \log 2)/(2\pi)$.

(iv) LEFT-TAIL REGIME: $n \rightarrow \infty$, $2^{k+1} - n = M = O(1)$, $M \geq 1$

$$h_n^k \sim \frac{2\sqrt{\pi}}{(M-1)!} \left(\frac{2C_0}{e^2} \right)^{2^k} 2^{(M-1/2)k} \quad (2.22)$$

where C_0 is defined in (2.18).

Remark. We note that the leading term for h_n^k or $1 - h_n^k$ is completely determined for cases (i), (ii) and (iv). However, for case (iii) the result involves the function $\Phi(\xi)$. We have not been able to determine Φ completely, except for its behaviors as $\xi \rightarrow 0$ and $\xi \rightarrow 2$. In Section 3 we discuss the numerical computation of Φ and also give a refinement of (2.19). Our analytical results predict that Φ is finite at $\xi = 2$, with $\Phi(2) = 1 - \log(\sqrt{2C_0}) = .559461\dots$, but Φ' has a logarithmic singularity with $\Phi'(\xi) \sim -\log(\sqrt{2-\xi})$ as $\xi \rightarrow 2$. As $\xi \rightarrow 0^+$, (2.19) shows that Φ as well as all its derivatives vanish. The dominant term in the right side of (2.19) is $\exp[-(\log 2)(\log_2 \xi)^2/2]$. It is interesting to note that in [19] we obtained results analogous to (2.15) for b -tries and for PATRICIA trees. For b -tries the corresponding Φ satisfies $\Phi(\xi) \sim \xi^b/(b+1)!$ as $\xi \rightarrow 0$, so that this function has a zero of order b at $\xi = 0$. For the PATRICIA model the dominant term is the same as that for the DST model (i.e., $\log \Phi$ has the same leading term for the two models). More precisely, Φ in PATRICIA becomes

$$\Phi(\xi) \sim \frac{1}{2} \rho_0 e^{\varphi(\log_2 \xi)} \xi^{3/2} \exp\left(-\frac{\log^2 \xi}{2 \log 2}\right), \quad \xi \rightarrow 0^+,$$

where $\rho_0 \approx 1.73137$ and $\varphi(\cdot)$ is periodic with period one. But, by examining the second terms for $\log \Phi$ we see that Φ is flatter near $\xi = 0$ for the DST model. For both DST and PATRICIA, the function Φ shows oscillatory behavior as $\xi \rightarrow 0$, while this is not the case for b -tries. $\square$

Since $\Phi(\xi) > 0$ for $0 < \xi \leq 2$ (see also the numerical results in Section 3), h_n^k is exponentially small for cases (iii) and (iv), while for cases (i) and (ii) $1 - h_n^k$ is exponentially small. For case (i) $1 - h_n^k$ is (roughly) $O(2^{-n^2/2})$, for case (ii) $1 - h_n^k = O(2^{-k^2/2})$, while

for cases (iii) and (iv) $h_n^k = O(e^{-n\Phi(\xi)})$. For a fixed large n , most of the probability mass occurs in that range of k where h_n^k changes from $h_n^k \approx 0$ to $h_n^k \approx 1$. We think of plotting h_n^k as a function of k so that (i) and (ii) apply in the right tail of the distribution while (iii) and (iv) apply in the left tail. The mass must thus be concentrated in the asymptotic matching region between cases (ii) and (iii). If we let $\xi \rightarrow 0$ sufficiently rapidly so that $n\Phi(\xi) \rightarrow 0$, then we can approximate (2.15) by $Ae^{-n\Phi} \sim 1 - n\Phi$ and then (ii) and (iii) asymptotically match, as is shown in Section 5. We can also consider a limit where $n \rightarrow \infty$, $\xi \rightarrow 0$ with $n\Phi$ bounded, or even $n\Phi \rightarrow \infty$. Then (2.15) can be approximated by $e^{-n\Phi}$ or by $Ae^{-n\Phi}$, with Φ replaced by its expansion (2.19) as $\xi \rightarrow 0^+$.

To simplify the expression in the matching region we set

$$\begin{aligned} k_\ell &= \left\lfloor \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \right\rfloor + \ell \\ &= \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} - \beta(n) + \ell \end{aligned} \quad (2.23)$$

where ℓ is an integer and

$$\beta(n) = \left\langle \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \right\rangle. \quad (2.24)$$

Here $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x . By using (2.23) in $e^{-n\Phi}$ (with Φ replaced by (2.19)) and simplifying the expression for n large, we are led to

$$h_n^k \approx \exp \left(\frac{-Q_*}{2\sqrt{\log 2}} e^{\beta - \ell - 1/\log 2} 2^{\sqrt{2 \log_2 n}(\beta - \ell) - (\beta - \ell)^2/2 - \frac{3}{4} \log_2(2 \log_2 n)} \right) \quad (2.25)$$

where $Q_* = Q_*(\log_2 n - \log_2(\sqrt{2 \log_2 n}) + 1/2)$. In (2.25) we write $\approx$ rather than $\sim$ since we neglected a factor $1 + O(1/\sqrt{\log n})$ in the exponent. Also, we note that $Q_*(z)$ is almost constant, with very small fluctuations, as explained below (2.29).

Now examine (2.25) for $\ell = 0$ and $\ell = 1$. If $0 < \beta < 1$ and $\ell = 0$ then h_n^k is small as $n \rightarrow \infty$. If $0 < \beta < 1$ and $\ell = 1$, h_n^k is close to one. This shows that for $0 < \beta(n) < 1$ the mass becomes concentrated at a single point as $n \rightarrow \infty$ (cf. Figure 2), corresponding to $\ell = 1$ in (2.23). It also follows from our analysis that the mean height is

$$\mathbf{E}[\mathcal{H}_n] = \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + O(1), \quad n \rightarrow \infty. \quad (2.26)$$

We now show that one can find special sequences $n(i)$ such that $n(i) \rightarrow \infty$ as $i \rightarrow \infty$ and along these sequences there is probability mass concentrated at two points, corresponding to $\ell = 0$ and $\ell = 1$ in (2.23). We define

$$\Delta(n) = \beta(n) \sqrt{2 \log_2 n} - \frac{3}{2} \log_2(\sqrt{2 \log_2 n}) \quad (2.27)$$

Figure 2: Asymptotic distributions for the height of DST and their corresponding lower and upper bounds.

and note that if $\Delta(n)$ is $O(1)$ then h_n^k in (2.25) is not asymptotically small at $\ell = 0$. We show in Figure 3 and more precisely in Section 3 that we can easily find sequences $n(i)$ such that $\Delta(n(i))$ remains bounded, and in fact we can have $\Delta \rightarrow 0$ for some of these (cf. (3.15)). If we consider $n(i)$ with $\Delta(n(i)) \rightarrow 0$, then as $n \rightarrow \infty$ for $\ell = 0$ the right side of (2.25) asymptotically becomes

$$\exp \left[-\frac{Q_*}{2\sqrt{\log 2}} \exp \left(-\frac{1}{\log 2} \right) \right]. \quad (2.28)$$

This does not approach a limit as $n \rightarrow \infty$ due to the oscillatory behavior of $Q(z)$ and hence $Q_*(z)$. However, we show in Section 4 that the non-constant terms in the Fourier series for Q (to be precise the series for $\log Q$) are *numerically* very small. Thus we can use the approximation resulting by using only the constant term, which yields

$$Q(z) \approx Q^0 = 2^{13/12} \left[\prod_{m=1}^{\infty} (1 - 2^{-m})^{-1} \right] \exp \left(-\frac{\pi^2}{3 \log 2} \right) = .063716934676 \dots \quad (2.29)$$

Now, $Q_*(z)$ is not the same as $Q(z)$. However, in Appendix A we derive the following

Figure 3: The function $\text{DELTA}(n) = \Delta(n) = \beta(n)\sqrt{2\log_2 n} - 1.5\log_2 \sqrt{2\log_2(n)}$ versus n .

alternate representation for $Q_*(z)$:

$$Q_*(z) = 2^{z^2/2} \sqrt{\frac{\log 2}{2\pi}} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} \frac{(-1)^{m+1} 2^{-m^2/2} 2^{m/2} (2^{m-1/2} - 2^{-m+1/2})}{2^{m-1/2} + 2^{-m+1/2} + 2^z + 2^{-z}}. \quad (2.30)$$

We tabulate both $Q(z)$ and $Q_*(z)$ in Table 1 (cf. Section 3) and sketch these functions in Figure 4. This shows that $Q_*(z)$ is also nearly constant with $Q_*(z) \approx Q_*^0 \approx .0637169$, which agrees with (2.29) to the accuracy given. Using (2.29) in (2.28) gives an estimate of the mass at $\ell = 0$ along sequences $n(i)$ that have $\Delta(n(i)) \rightarrow 0$. Note that $\Delta \rightarrow 0$ implies that $\beta \rightarrow 0$, in view of (2.27). In Section 3 we test the accuracy of (2.28). We summarize our findings in the following corollary.

Corollary 1 *The asymptotic distribution of DST height is concentrated among the two points $k_1 - 1$ and k_1 where $k_1 = k_1(n) = \left\lfloor \log_2 n + \sqrt{2\log_2 n} - \log_2(\sqrt{2\log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \right\rfloor + 1$, that is,*

$$\Pr\{\mathcal{H}_n = k_1 - 1 \text{ or } k_1\} = 1 - o(1)$$

as $n \rightarrow \infty$. More precisely: (i) there are subsequences n_i for which $\Pr\{\mathcal{H}_{n_i} = k_1\} = 1 - o(1)$

Figure 4: (a) A plot of $(Q(z) - 0.0637169346763735400492200) \cdot 10^{24}$ over three periods; (b) A plot of $(Q_*(z) - 0.063716934676) \cdot 10^{12}$ over three periods.

provided that

$$\begin{aligned} \Delta(n_i) &= \sqrt{2 \log_2 n_i} \left\langle \log_2 n_i + \sqrt{2 \log_2 n_i} - \log_2(\sqrt{2 \log_2 n_i}) + \frac{1}{\log 2} - \frac{1}{2} \right\rangle \\ &\quad - \frac{3}{2} \log_2(\sqrt{2 \log_2 n_i}) \rightarrow \infty \end{aligned}$$

as $i \rightarrow \infty$; (ii) there are subsequences n_i for which $\Pr\{\mathcal{H}_{n_i} = k_1 - 1 \text{ or } k_1\} = 1 - o(1)$ provided that $\Delta(n_i) = O(1)$.

We are now in a position to compare our results to the corresponding results for PATRICIA trees that we recently obtained in [19]. With regards to the average $\mathbb{E}[\mathcal{H}_n]$ given by (2.26) the first two terms are the same as for PATRICIA (see also Devroye [5]), but the third term for PATRICIA is $O(1)$ while for DST it is $O(\log \log n)$. Thus (2.26) shows that the improvement of the DST over the PATRICIA models appears only at the third term in the asymptotics of $\mathbb{E}[\mathcal{H}_n]$. Since the coefficient of the $\log \log n$ term in (2.26) is negative, $\mathbb{E}[\mathcal{H}_n]$ is asymptotically smaller for DST. With respect to the limiting distribution in the central regime, both are of an exponential of a Gaussian type. However, the DST distribution function contains an additional term $O(\log \log n)$ in the double exponent. This additional term prohibits the limiting distribution for DST to be concentrated in some cases on k_1 and $k_1 + 1$, which does occur for PATRICIA height.

Finally, we shall discuss the Lempel-Ziv model in which a random string of *fixed* length, m , is partitioned into a random number, M_m , of phrases. We shall use the results of Theorem 1 to prove the asymptotic distribution of the longest phrase among the M_m random phrases. Let us first introduce for any $\varepsilon > 0$

$$\delta_m = \Pr\{M_m \notin [(1 - \varepsilon)\mu(m), (1 + \varepsilon)\mu(m)]\}$$

where $\mu(m) = m/\log_2 m$ (i.e., the typical number of phrases). It is known that $\delta_m \rightarrow 0$ as $m \rightarrow \infty$ (cf. [2, 13, 32]). In fact, a stronger result is known. Jacquet and Szpankowski [13] proved that $\delta_m = O(e^{-R\varepsilon\sqrt{m}})$ for some $R > 0$.

Now, we can formulate our second main result.

Theorem 2 *Consider the Lempel-Ziv model in which a fixed string of length m is parsed according to the Lempel-Ziv algorithm. Let $\mathcal{H}_m^{LZ}$ be the length of the longest phrase (among random number, M_m , of phrases), while $\mathcal{H}_n$ is still the height of a digital search tree built from n independently generated strings, as studied in Theorem 1. Then for all $k \geq 0$ and $\varepsilon > 0$*

$$\Pr\{\mathcal{H}_{(1+\varepsilon)\mu(m)} \leq k\} - \delta_m \leq \Pr\{\mathcal{H}_m^{LZ} \leq k\} \leq \delta_m + \Pr\{\mathcal{H}_{(1-\varepsilon)\mu(m)} \leq k\}. \quad (2.31)$$

In particular,

$$\mathbb{E}[\mathcal{H}_m^{LZ}] = \log_2(m/\log_2 m) + \sqrt{2 \log_2(m/\log_2 m)} - \frac{1}{2} \log_2 \log_2(m/\log_2 m) + O(1), \quad (2.32)$$

and most of the probability mass is concentrated either at k_1^{LZ} or $k_1^{LZ} - 1$ where

$$k_1^{LZ} = \left\lfloor \log_2(m/\log_2 m) + \sqrt{2 \log_2(m/\log_2 m)} - \frac{1}{2} \log_2 \log_2(m/\log_2 m) + \frac{1}{\log 2} - 1 \right\rfloor + 1$$

for large m .

Proof. We proceed as follows

$$\begin{aligned} \Pr\{\mathcal{H}_m^{LZ} \leq k\} &= \sum_{n=0}^m \Pr\{\mathcal{H}_n \leq k, M_m = n\} \\ &\leq \delta_m + \sum_{(1-\varepsilon)\mu(m) \leq n \leq (1+\varepsilon)\mu(m)} \Pr\{\mathcal{H}_n \leq k, M_m = n\} \\ &\leq \delta_m + \Pr\{\mathcal{H}_{(1-\varepsilon)\mu(m)} \leq k\}. \end{aligned}$$

where the last inequality is a consequence of the fact that $\mathcal{H}_m^{LZ}$ is a nondecreasing sequence with respect to m . The lower bound of (2.31) can be proved in a similar manner. The rest is a simple consequence of (2.31). ■

3 Discussion and Numerical Results

We shall discuss the accuracy of the various asymptotic results, and also numerically calculate the hitherto undetermined function $\Phi(\xi)$. We begin by making some general comments on how to use the asymptotic formulae.

It is most natural to view the problem as starting with a fixed (large) n and then varying k . We let k^* be the minimum integer such that $2^{k^*+1} - 1 \geq n$. More precisely, we set

$$k^* = \begin{cases} \log_2(n+1) - 1 & \text{if } n+1 = \text{power of 2} \\ \lfloor \log_2(n+1) \rfloor & \text{if } n+1 \neq \text{power of 2,} \end{cases} \quad (3.1)$$

and note that h_n^k is only non-zero for $k \geq k^*$. If, say, $n = 100$ we have $k^* = 6$. Then $2^{k^*+1} - n = 28$ so that we are probably out of the range where the M -scale expansion applies. For $k = k^*$ we have $\xi = n2^{-k} = 25/16$ and this is well within the range $0 < \xi < 2$, where Theorem 1(iii) applies. By increasing k to $k^* + 1, k^* + 2, k^* + 3, \dots$, we obtain the values $25/32, 25/64, 25/128, \dots$ for ξ , so that ξ becomes small rapidly and it thus may be desirable to use the WKB approximation (cf. Section 5) and further replace A and Φ by their small ξ expansions, which we have derived explicitly. When k further increases to a significant fraction of n (e.g., $k = 20$) then we should use the expansion (ii) of Theorem 1, which applies on the α -scale (where $\alpha = n/(n-k) > 1$). When k further increases to a value close to n , such as 95, we can use the expansion that applies for $j = n - k$ fixed (cf. Theorem 1(i)). Of course, for $k \geq n - 1$, we have $h_n^k = 1$.

If we start with $n = 127$, then $k^* = 6$, which correspond to $M = 1$ and $\xi = 127/64$. Thus for $k = k^*$ we can use the M -scale result (cf. Theorem 1(iv)), but increasing k to $k^* + 1 = 7$ puts us well within the region $0 < \xi < 2$, where Theorem 1(iii) applies. If $n = 128$ then $k^* = 7$ and $2^{k^*+1} - n = 128$. This corresponds to $M = 128$ and $\xi = 1$, and this indicates the ξ -scale result should be used. Since $\xi = 1$ is not close to either 0 or 2, we must use the numerical value of $\Phi(1)$.

We define, as before, k_1 by

$$k_1 = k_1(n) = \left\lfloor \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \right\rfloor + 1. \quad (3.2)$$

According to our analysis, as $n \rightarrow \infty$ the probability mass should be concentrating at the single point k_1 , or the two points $k_1 - 1, k_1$ (with the former being more likely). If $n = 100$ we have $k_1(100) = 10$ so that $k_1 - k^* = 4$, and hence the “left-tail” really consists of only the four points with $k = 6 - 9$.

In using the asymptotic results that involve the periodic functions $Q(z)$ and $Q_*(z)$, we approximate them by the numerical value in (2.29), which is correct to 11 significant digits for both of them. Figure 4 and Table 1 show that the oscillations in Q_* occur only in the 12th significant digit, while those in Q are even smaller, occurring in the 25th digit. From the Fourier series for Q given by (2.20) we can analytically estimate the oscillations by using only the term with $m = 1$ and approximating $\sinh(a)$ by $e^a/2$, which yields

$$\begin{aligned}\frac{Q(z)}{Q_0} &\approx 1 - 2e^{-2a} \cos(2\pi z), & a &= \frac{2\pi^2}{\log 2} \\ &= 1 - 3.678 \dots (10^{-25}) \cos(2\pi z).\end{aligned}$$

We note that both Q and Q_* are symmetric about the midpoint $z = 1/2$ and achieve their maxima here.

We first discuss the accuracy of the expansion for M fixed, which corresponds to case (iv) of Theorem 1, as defined in Section 2. In Table 2, we consider values of n that are of the form $2^{k+1} - 1$, or slightly smaller, and various M in the range 1 to 5. From the table we see that the asymptotic formula is highly accurate and that for larger k , we can allow M to be larger and still obtain good agreement. We recall that the error term in (2.22) is $O(2^{-k}) = O(n^{-1})$. The results in Table 2 suggest that the numerical coefficient of the correction term is fairly small.

We next consider the far right tail (cf. Theorem 1(i)), where h_n^k is close to one. In Table 3 we consider n in the range $[10, 25]$ for $2 \leq j \leq 5$. We tabulate the exact values of $1 - h_n^k$ and the result in (2.11). When $j = 2$ the agreement is excellent for all n , as it should be since (2.6) shows that (2.11) is not only asymptotic but exact! However, for $j = 3$ the error is about 25% when $n = 10$; it decreases to below 10% when $n = 25$. The situation becomes worse when j increases.

The correction term to (2.11) is $O(n^{-1})$. The data in Table 3 suggest that the numerical coefficient in the correction term is fairly large, and increases with j . This is certainly consistent with the exact results in (2.7) and (2.8), which show that the correction factor is of the form $1 - 2/n$ and $1 - 5/n$ for $j = 3$ and $j = 4$, respectively. Note also that if, say, $n = 20$ and $j = 5$, it is not apriori clear whether we should use the $j = O(1)$ result, or that for $\alpha = n/j$ fixed.

In Table 4 we test the accuracy of (2.12). We fix $\alpha = n/(n - k) = 4$ and consider n in the range $[16, 32]$. Note that since $\alpha > 2$, we can use the infinite sum in (2.13) to calculate the integral I , which yields $I(4) = .17398 \dots$. When $n = 16$ the error is about 25% and decreases to 10% when $n = 32$. While this is consistent with a correction term of $O(n^{-1})$, the data again suggest that the coefficient in the error term is fairly large, relative to one.

Table 1: Comparison of $Q(z)$ and $Q_*(z)$.

z	$Q(z)$	$Q_*(z)$
0	.06371693467637354004922003786	.0637169346763188
.05	_____ 3901	_____ 3215
.1	_____ 4234	_____ 3293
.15	_____ 4753	_____ 3414
.2	_____ 5406	_____ 3566
.25	_____ 6130	_____ 3735
.3	_____ 6854	_____ 3904
.35	_____ 7508	_____ 4056
.4	_____ 8026	_____ 4177
.45	_____ 8359	_____ 4255
.5	_____ 8474	_____ 4281
.55	_____ 8359	_____ 4255
.6	_____ 8026	_____ 4177
.65	_____ 7508	_____ 4056
.7	_____ 6854	_____ 3904
.75	_____ 6130	_____ 3735
.8	_____ 5406	_____ 3566
.85	_____ 4753	_____ 3414
.9	_____ 4234	_____ 3293
.95	_____ 3901	_____ 3215
1	_____ 3786	_____ 3188

Table 2: Left-Tail Comparison

n	k	M	h_n^k (exact)	h_n^k (2.22)
7	2	1	$7.81250(10^{-2})$	$8.06991(10^{-2})$
6	2	2	.312500	.322796
15	3	1	$1.27852(10^{-3})$	$1.29902(10^{-3})$
14	3	2	$1.02282(10^{-2})$	$1.03922(10^{-2})$
13	3	3	$4.09126(10^{-2})$	$4.15688(10^{-2})$
31	4	1	$2.36143(10^{-7})$	$2.38012(10^{-7})$
30	4	2	$3.77829(10^{-6})$	$3.80820(10^{-6})$
29	4	3	$3.02264(10^{-5})$	$3.04656(10^{-5})$
28	4	4	$1.61555(10^{-4})$	$1.62483(10^{-4})$
63	5	1	$5.62788(10^{-15})$	$5.65000(10^{-15})$
62	5	2	$1.80092(10^{-13})$	$1.80800(10^{-13})$
61	5	3	$2.88147(10^{-12})$	$2.89280(10^{-12})$
60	5	4	$3.07515(10^{-11})$	$3.08566(10^{-11})$
59	5	5	$2.46852(10^{-10})$	$2.46390(10^{-10})$

Table 3: Far Right-Tail Comparison

n	k	j	$1 - h_n^k$ (exact)	$1 - h_n^k$ (2.11)
10	8	2	$1.45519(10^{-11})$	$1.45519(10^{-11})$
	7	3	$2.98023(10^{-8})$	$3.72529(10^{-8})$
	6	4	$1.31922(10^{-5})$	$2.38419(10^{-5})$
	5	5	$1.63566(10^{-3})$	$5.08626(10^{-3})$
15	13	2	$4.03897(10^{-28})$	$4.03897(10^{-28})$
	12	3	$4.30134(10^{-23})$	$4.96308(10^{-23})$
	11	4	$1.05258(10^{-18})$	$1.52446(10^{-18})$
	10	5	$7.82575(10^{-15})$	$1.56125(10^{-14})$
20	18	2	$3.34096(10^{-52})$	$3.34096(10^{-52})$
	17	3	$1.57646(10^{-45})$	$1.75162(10^{-45})$
	16	4	$1.75253(10^{-39})$	$2.29589(10^{-39})$
	15	5	$6.09591(10^{-34})$	$1.00309(10^{-33})$
25	23	2	$8.23609(10^{-84})$	$8.23609(10^{-84})$
	22	3	$1.58906(10^{-75})$	$1.72723(10^{-75})$
	21	4	$7.32182(10^{-68})$	$9.05568(10^{-68})$
	20	5	$1.07178(10^{-60})$	$1.58259(10^{-60})$

Table 4: Right-Tail Comparison

n	k	α	$1 - h_n^k$ (exact)	$1 - h_n^k$ (2.12)
16	12	4	$2.99991(10^{-22})$	$3.74045(10^{-22})$
20	15	4	$6.09591(10^{-34})$	$7.21250(10^{-34})$
24	18	4	$2.42022(10^{-48})$	$2.77231(10^{-48})$
28	21	4	$1.88363(10^{-65})$	$2.11079(10^{-65})$
32	24	4	$2.87538(10^{-85})$	$3.17143(10^{-85})$

Table 5: Uniform Right-Tail Approximation

n	k	j	$1 - h_n^k$ (exact)	$1 - h_n^k$ (3.3)
10	8	2	$1.45519(10^{-11})$	$1.45519(10^{-11})$
	7	3	$2.98023(10^{-8})$	$2.98023(10^{-8})$
	6	4	$1.31922(10^{-5})$	$1.31925(10^{-5})$
	5	5	$1.63566(10^{-3})$	$1.63632(10^{-3})$
	4	6	$6.09603(10^{-2})$	$6.13033(10^{-2})$
	3	7	.617309	.687647
20	18	2	$3.34096(10^{-52})$	$3.34096(10^{-52})$
	16	4	$1.75253(10^{-39})$	$1.75253(10^{-39})$
	14	6	$7.43000(10^{-29})$	$7.43000(10^{-29})$
	12	8	$5.87353(10^{-20})$	$5.87353(10^{-20})$
	10	10	$1.09764(10^{-12})$	$1.09764(10^{-12})$
	8	12	$5.19019(10^{-7})$	$5.19043(10^{-7})$
	6	14	$5.92196(10^{-3})$	$5.93731(10^{-3})$
	5	15	.143413	.148697
30	4	16	.855618	>1
	10	20	$2.52277(10^{-10})$	$2.52278(10^{-10})$
	9	21	$1.24549(10^{-7})$	$1.24553(10^{-7})$
	8	22	$2.56647(10^{-5})$	$2.56706(10^{-5})$
	7	23	$2.16191(10^{-3})$	$2.16562(10^{-3})$
	6	24	$7.08821(10^{-2})$	$7.25828(10^{-2})$
	5	25	.668201	.920224
100	13	87	$3.80919(10^{-13})$	$3.80919(10^{-13})$
	12	88	$3.99988(10^{-10})$	$3.99991(10^{-10})$
	11	89	$1.86584(10^{-6})$	$1.86592(10^{-6})$
	10	90	$3.81030(10^{-5})$	$3.81128(10^{-5})$
	9	91	$3.33434(10^{-3})$	$3.34291(10^{-3})$
	8	92	.115924	.122474
	7	93	.866855	>1

In obtaining the asymptotic results summarized in (2.11)–(2.22), we have written them in the simplest possible form. However, from our analysis we can easily generate results that are more uniform and more numerically accurate. By more uniform we mean they apply to larger ranges of k than those in (2.11)–(2.22). For example, in the right tail we find that

$$1 - h_n^k \sim 2^{-k^2/2} 2^{-k/2} \tilde{F}_j(n) \quad (3.3)$$

where $\tilde{F}_j(n)$ is given precisely by (4.23) of Section 4. For j fixed and $n \rightarrow \infty$ (4.23) reduces to (2.11), for $\alpha = n/j > 1$ and $n \rightarrow \infty$ it reduces to (2.12), and the analysis in Section 4 shows that it applies even for $n \rightarrow \infty$ with $k/n \rightarrow 0$ as long as $k/\log_2 n = \nu > 1$. Thus, the approximation holds anywhere in the right tail and we call $2^{-k^2/2} 2^{-k/2} \tilde{F}_j(n)$ the “uniform right tail” (URT) approximation for $1 - h_n^k$. In Table 5 we consider $n = 10, 20, 30$ and 100 and various values of $k = n - j$. We decrease k until the exact value of $1 - h_n^k$ exceeds .5 (i.e., $h_n^k < .5$). For $n = 10$, Table 5 shows that for $k = 8$ and $k = 7$, the exact and URT results agree to 6 decimal places. As k decreases from 6 to 4, the error slowly increases, but remains under 2%. For $k = 3$ we have $1 - h_n^k > .5$, but even then URT is accurate to within 12%. By now we are well outside of the right tail. For $n = 20$ and $k \geq 6$, URT is accurate to within .3% and agrees with the exact result to 6 decimal places for $k \geq 10$. When $k = 5$ we have $1 - h_n^k$ becoming “ $O(1)$ ”, but still URT is accurate to within 5%. For $k = 4$ the asymptotic result exceeds 1, but then we are clearly outside of the right tail. Similar trends are apparent for $n = 30$ and $n = 100$. Here we need only consider relatively small values of k/n , since even here URT is accurate to 6 decimal places. We also note that in each data point in Table 5, URT *overestimates* $1 - h_n^k$. This suggests that it may be an upper bound for $1 - h_n^k$ (thus a lower bound for h_n^k) for points in the right tail.

Now we test the accuracy of the asymptotics for points where there is appreciable mass. According to our result, this corresponds to $k - \log_2 n \approx \sqrt{2 \log_2 n}$. Our approximation here is that in (2.15) with $\Phi(\xi)$ replaced by its small ξ expansion (2.19). However, the analysis in Section 4 shows that the error term in (2.19) is only smaller than the leading term by a factor $1/\log \xi$ (with possibly some $\log(-\log \xi)$ factors). While we could compute these, it proves more efficient to use the full result in (4.42) proved in Section 4. If we set $n2^{-k} = \xi$ in (4.42) we find that $2^{-k^2/2} 2^{-k/2} \times$ (4.42) has the form $n \times [\text{function of } \xi]$, so that the matching condition in (5.40) may be refined to

$$\Phi(\xi) \sim \Phi_0(\xi), \quad \xi \rightarrow 0^+ \quad (3.4)$$

where

$$\Phi_0(\xi) = \pi\xi \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{2^{s(s-1)/2} Q(s)}{\Gamma(3-s)} e^{-s \log \xi} ds. \quad (3.5)$$

While the error in (2.19) is only about $1 + O(1/\log \xi)$, we believe that in (3.4) the error is about $1 + O(\xi)$, though we have not explicitly calculated the correction. Now, (3.5) is much less insightful than (2.19), but it is much more accurate asymptotically, and, as we show, numerically. The calculations in Section 4 shows that the leading term in a saddle point expansion of (3.5) as $\xi \rightarrow 0^+$ yields precisely (2.19). We also note that the integrand in (3.5) is an entire function of s , as both $Q(\cdot)$ and $1/\Gamma(\cdot)$ are.

In Appendix B we obtain the following representation for Φ_0 as an infinite sum:

$$\Phi_0(\xi) = \frac{1}{4\xi} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{5m/2} \exp(-2\xi 2^{-m}). \quad (3.6)$$

For any fixed $\xi > 0$ this sum converges very rapidly as $m \rightarrow \infty$, and even more rapidly as $m \rightarrow -\infty$, due to the exponential factor. It is thus useful for numerical calculations. The asymptotics as $\xi \rightarrow 0^+$ are difficult to obtain from (3.6), due to the alternating sum. However, we can easily see that $\Phi_0(\xi) = o(\xi^N)$ for all $N \geq 1$. Indeed, we expand the exponent in Taylor series and exchange the order of the two summations. We have

$$\sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{5m/2} 2^{-mL} = 0 \quad (3.7)$$

for any integer L , as can be seen by the antisymmetry of the summand with respect to the shift $m \rightarrow 5 - 2L - m$. It thus follows that Φ_0 vanishes to all algebraic orders as $\xi \rightarrow 0^+$. A rough estimate of the leading behavior can be obtained by noting that for $\xi \rightarrow 0^+$ the important terms in the sum are those where $\xi 2^{-m} = O(1)$ so that $m \approx \log_2 \xi$. There the magnitude of the summand is roughly $2^{-m^2/2} = \exp\left[-\frac{1}{2}(\log 2)(\log_2 \xi)^2\right]$, which is the same as the dominant factor in (2.19).

For numerical calculations we use (3.6) to approximate Φ and A in (2.15), for $\xi \rightarrow 0^+$. Setting $z_0 = \Phi_0(\xi)$,

$$z_1 = (\xi \Phi_0)'(\xi) = -\frac{1}{2} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{3m/2} \exp(-2\xi 2^{-m}) \quad (3.8)$$

and

$$z_2 = \xi(\xi \Phi_0)''(\xi) = \xi \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{m/2} \exp(-2\xi 2^{-m}), \quad (3.9)$$

we thus obtain the asymptotic form of the (improved) WKB approximation as (cf. (2.15))

$$\sqrt{1+z_2}e^{-nz_0}e^{-z_1}. \quad (3.10)$$

We denote the above “refined” approximation by REF. In Table 6 we compare (3.10) to the exact values, for the few points where there is appreciable (numerical) mass. We also give the values of $\xi = n2^{-k}$, since (3.10) assumes that $\xi \rightarrow 0^+$. We consider n in the range [30,100].

When $n = 40$ and $k = 5$ we have $\xi = 1.25$, and here (3.10) overestimates the true value by a factor of about 3. However, here we would not expect to be able to use the small ξ approximation to (2.15), but rather should compute $\Phi(1.25)$ numerically and use this value instead. We will discuss the numerical computation of Φ shortly. When $k = 6$ we have $\xi = .625$ which is certainly not small, but nevertheless (3.10) is accurate to about 1%. For $k = 7, 8$ and 9 the two results agree to 3 or 4 decimal places. For each n we increased k until h_n^k is one to 4 decimal places. Note that by the time k reaches this value, we are fairly well in the right tail and there we showed that URT is highly accurate.

As we increase n , Table 6 shows that REF is not accurate whenever $\xi \geq 1$, but gives reasonable approximations for $\xi \leq .8$ and is very good for $\xi \leq .5$. Thus the small ξ approximation to (2.15) in (3.10) is quite robust. The numerical computation of Φ is most difficult when ξ is small, so that the numerical and asymptotic methods complement each other, as is often the case in applied problems.

We note that it was essential that we used the refined approximation in (3.10), rather than the leading term result in (2.19), which we denote by $\Phi_{LT}(\xi)$. In Table 7 we compare the values of Φ_0 and Φ_{LT} , starting at $\xi = .1$ and decreasing to $\xi = 10^{-14}$. When $\xi = .1$ we have $\Phi_{LT}/\Phi_0 = 7.31$ and even if $\xi = 10^{-14}$ this ratio is 1.79, which is far from the theoretical value (=1) as $\xi \rightarrow 0^+$. When $\xi = 10^{-14}$, both values are about 10^{-400} . By now we are so far in the right tail that we would never use the WKB result in the first place.

Finally, we discuss the accumulation of the probability mass at one or two points as $n \rightarrow \infty$, as our results predict. We define ℓ and $\beta = \beta(n)$ as in (2.23). We clearly have $0 \leq \beta < 1$. It proves easiest to discuss the limit $n \rightarrow \infty$ along subsequences $n(i)$ that correspond to β nearly constant.

If we take a fixed $0 < \beta < 1$, then (2.25) predicts that the masses at $\ell = 0, 1, 2$ (corresponding to $k = k_1 - 1, k_1, k_1 + 1$) are approximately

$$m_0 \approx \exp \left[-\frac{Q_*}{2\sqrt{\log 2}} e^{-\frac{1}{\log 2}} \frac{2^{\beta\sqrt{2\log 2}n} 2^{-\beta^2/2} e^{\beta}}{(2\log_2 n)^{3/4}} \right],$$

Table 6: Refined Approximation in the Central Regime

n	k	ξ	h_n^k (exact)	h_n^k (3.10)
30	5	.938	.3318	.3921
	6	.469	.9291	.9293
	7	.234	.9978	.9979
	8	.117	1.0000	1.0000
40	5	1.25	.0206	.0747
	6	.625	.7134	.7220
	7	.313	.9851	.9852
	8	.156	.9997	.9997
	9	.078	1.0000	1.0000
50	6	.781	.3723	.4086
	7	.391	.9438	.9440
	8	.195	.9985	.9986
	9	.098	1.0000	1.0000
60	6	.938	.1051	.1541
	7	.469	.8518	.8532
	8	.234	.9948	.9948
	9	.117	.9999	.9999
70	6	1.09	.0122	.0375
	7	.547	.6972	.7035
	8	.273	.9857	.9857
	9	.137	.9998	.9998
	10	.068	1.0000	1.0000
80	6	1.25	.0004	.0060
	7	.625	.4950	.5114
	8	.313	.9673	.9674
	9	.156	.9993	.9993
	10	.078	1.0000	1.0000
90	6	1.41	.0000	.0007
	7	.703	.2902	.3177
	8	.352	.9350	.9351
	9	.176	.9984	.9984
	10	.088	1.0000	1.0000
100	7	.781	.1331	.1644
	8	.391	.8841	.8846
	9	.195	.9967	.9967
	10	.098	1.0000	1.0000

Table 7: Different Approximations of $\Phi(\xi)$.

ξ	$\Phi_{LT}(\xi)$	$\Phi_0(\xi)$	Φ_{LT}/Φ_0
.1	3.99(10 ⁻⁶)	5.45(10 ⁻⁷)	7.31
10 ⁻²	3.58(10 ⁻¹⁵)	7.54(10 ⁻¹⁶)	4.75
10 ⁻³	6.62(10 ⁻²⁸)	1.78(10 ⁻²⁸)	3.72
10 ⁻⁴	2.95(10 ⁻⁴⁴)	9.33(10 ⁻⁴⁵)	3.16
10 ⁻⁶	1.26(10 ⁻⁸⁷)	4.94(10 ⁻⁸⁸)	2.55
10 ⁻⁸	5.12(10 ⁻¹⁴⁵)	2.29(10 ⁻¹⁴⁵)	2.23
10 ⁻¹⁰	2.81(10 ⁻²¹⁶)	1.38(10 ⁻²¹⁶)	2.03
10 ⁻¹²	2.62(10 ⁻³⁰¹)	1.38(10 ⁻³⁰¹)	1.89
10 ⁻¹⁴	4.89(10 ⁻⁴⁰⁰)	2.73(10 ⁻⁴⁰⁰)	1.79

$$\begin{aligned}
 m_1 &\approx 1 - \frac{Q_*}{2\sqrt{\log 2}} e^{-\frac{1}{\log 2}} \frac{2^{(\beta-1)\sqrt{2\log_2 n}} 2^{-(\beta-1)^2/2} e^{\beta-1}}{(2\log_2 n)^{3/4}}, \\
 m_2 &\approx \frac{Q_*}{2\sqrt{\log 2}} e^{-\frac{1}{\log 2}} \frac{2^{(\beta-1)\sqrt{2\log_2 n}} 2^{-(\beta-1)^2/2} e^{\beta-1}}{(2\log_2 n)^{3/4}}.
 \end{aligned} \tag{3.11}$$

Here $m_0 = h_n^{k_1-1} - h_n^{k_1-2} \sim h_n^{k_1-1}$, $m_1 = h_n^{k_1} - h_n^{k_1-1} \sim h_n^{k_1} (\sim 1)$ and $m_2 = h_n^{k_1+1} - h_n^{k_1} \sim 1 - m_1$. Thus all the mass should concentrate at $k = k_1$ as long as β remains bounded from 0 and 1. Even as $\beta \rightarrow 1^-$ we will have $m_2 \rightarrow 0$ and $m_1 \rightarrow 1$, due to the factor $(\log n)^{-3/4}$ in m_2 and $1 - m_1$.

Next consider subsequences along which $\beta \rightarrow 0$. As before, we define $\Delta(n)$ by

$$\Delta(n) = \beta(n)\sqrt{2\log_2 n} - \frac{3}{2}\log_2(\sqrt{2\log_2 n}). \tag{3.12}$$

Using (3.12) in (2.25) (or (3.11)) we find that if $n \rightarrow \infty$ in such a way that $\Delta(n)$ is bounded, then $h_n^{k_1-1}$ is $O(1)$ and < 1 . Thus, for such sequences there is mass at two points, corresponding to $k = k_1 - 1$ and $k = k_1$.

To construct such sequences $\Delta(n)$ we consider the equation

$$N + \frac{3\log_2(\sqrt{2\log_2 n})}{\sqrt{2\log_2 n}} = \log_2 n + \sqrt{2\log_2 n} - \log_2(\sqrt{2\log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \tag{3.13}$$

Table 8: Solutions $n_*(M)$ of (3.13)

N	$n_*(N)$	$\min \Delta(n) $
3	3.52	4
4	6.64	7
5	12.08	12
6	21.78	22
7	39.14	39
8	70.38	70
9	126.69	127
10	228.45	228
11	412.79	413
12	747.38	747
13	1355.94	1356
14	2464.88	2465
15	4489.33	4489

where N is an integer $N \geq 3$. For any N we can solve (numerically) the implicit relation (3.13) and generate a sequence of solutions, which we denote by $n_*(N)$. In view of the definition (2.24) of $\beta(n)$ we see that when $n = n_*(N)$ we have $\Delta(n) = 0$. Now, $n_*(N)$ is generally not an integer, but $\lfloor n_*(N) \rfloor$ or $\lfloor n_*(N) \rfloor + 1$ should correspond to a local minimum of the sequence $|\Delta(n)|$. In Table 8 we compute the first few $n_*(N)$ and also give the sequence of local minima of $|\Delta(n)|$. We see that $\min |\Delta(n)|$ is the integer closest to $n_*(N)$ for all $3 \leq N \leq 15$. In Figure 4 we plot the sequence $\Delta(n)$. From (3.12) and the figure we see that the upper envelope grows (roughly) like $\sqrt{\log n}$ and the lower envelope like $-\log \log n$. The sequence increases, crosses zero and then jumps back down. The jump corresponds to the fractional part β changing from 1^- to 0^+ , which occurs by increasing n . The maxima of $\Delta(n)$ corresponds to $\beta \approx 1$ while the minima have $\beta \approx 0$. Note also that β is small at the minima of $|\Delta(n)|$.

We can easily solve (3.13) asymptotically as $N \rightarrow \infty$, and this yields the estimate

$$\begin{aligned} n_*(N) &= 2^{N-\sqrt{2N}} \sqrt{2N} 2^{3/2} e^{-1} \\ &\times \left[1 + \frac{\log 2}{\sqrt{2N}} \left(\frac{1}{2} \log_2(\sqrt{2N}) - 1 \right) + O\left(\frac{\log N}{N}\right) \right]. \end{aligned} \quad (3.14)$$

Using (3.14) to compute $\beta(n)$ and then $\Delta(n)$ for $N \rightarrow \infty$, we find that for $N \rightarrow \infty$ and $n - n_* = O(1)$

$$\begin{aligned} \Delta(n) &\sim \frac{1}{n_*} (n - n_*) \left\{ \frac{3 \log_2(\sqrt{2 \log_2 n_*})}{4 \log(n_*)} - \frac{3}{4 \log 2} \frac{1}{\log(n_*)} \right. \\ &\quad \left. + \frac{\sqrt{2}}{(\log 2)^{3/2}} \left(\sqrt{\log n_*} + \sqrt{\frac{\log 2}{2}} - \frac{1}{2\sqrt{\log n_*}} \right) \right\}. \end{aligned} \quad (3.15)$$

This shows that as $n \rightarrow \infty$, $\Delta(n)$ is not only bounded but $\Delta(n) \rightarrow 0$ with (roughly) $\Delta(n_*(N)) = O(n^{-1})$. Thus, along sequences such as $n = \lfloor n_*(N) \rfloor$ or $n = \lfloor n_*(N) \rfloor + 1$, the mass m_0 should be

$$\exp \left[-\frac{Q_*}{2\sqrt{\log 2}} \exp \left(-\frac{1}{\log 2} \right) \right] \approx .990998897 \dots \quad (3.16)$$

Even though this is *asymptotically* $O(1)$ and < 1 , the small value of Q_* (cf. (2.29)) shows that *numerically* most of the mass will be at $k_1 - 1$, with the remaining mass at k_1 . Expression (3.16) has some fluctuations, but these are very small in view of Table 1.

If we choose a sequence $n(i)$ such that $\beta(n) \rightarrow 1^-$, then the mass becomes concentrated at $\ell = 1$ ($k = k_1$), and the mass at $\ell = 2$ is

$$m_2 \sim \frac{Q_*}{2\sqrt{\log 2}} e^{-\frac{1}{\log 2}} \frac{1}{(2 \log_2 n)^{3/4}}. \quad (3.17)$$

The right side of (3.17) approaches zero slowly, but in view of the small numerical value of $Q_* e^{-1/(\log 2)} 2^{-1} (\log 2)^{-1/2} \approx .009041857 \dots$, m_2 will be numerically small even for moderate values of n . We also note that choosing n to make $\beta \approx 1$ minimizes the mass at $\ell = 0$ ($k = k_1 - 1$). From an asymptotic point of view, the optimal way to choose $n(i)$ to get most rapid convergence to mass at the single point k_1 , is to minimize $|\beta(n) - 1/2|$. This makes m_0 and m_2 both small. However, for moderate values of n , it is preferable to choose $\beta \approx 1$, in order that the factor $2^{\sqrt{2 \log_2 n}}$ in the exponent in m_0 in (3.11) compensates for the numerically small value of Q_* . Thus for moderate n it is more essential to minimize m_0 rather than m_2 , to obtain $m_1 \rightarrow 1$. For very large values of n it becomes desirable to minimize m_2 , which requires $\beta \approx 0$ (but in such a way that $\Delta(n) \rightarrow -\infty$).

The numerical results in Table 6 show that for $k = 40, 50, 60, 70, 80, 90$ and 100 , the most mass occurs, respectively, at $k = 6, 7, 7, 7, 7, 8$ and 8 . The respective masses, obtained from $h_n^k - h_n^{k-1}$, are about .69, .57, .75, .69, .49, .64 and .75. The respective values of k_1 are 8, 9, 9, 9, 9, 10 and 10, so that the most mass is at $k_1 - 2$, not k_1 ! This apparent discrepancy, however, can easily be understood from the asymptotic analysis. In order to make the mass for $k \leq k_1 - 2$ ($\ell \leq -1$) about .1, even with an optimal value of $\beta \approx 1$, we need n large enough so that

$$.1 = \exp \left[-\frac{Q_*}{2\sqrt{\log 2}} \frac{1}{(2 \log_2 n)^{3/4}} 2^{2\sqrt{2 \log_2 n}} 2^{-2} e^{2-1/\log 2} \right]$$

so that $n \approx 22, 123$. To make this mass .01, we need $n \approx 252, 025$. To make the mass in the range $k \leq k_1 - 1$ ($\ell \leq 0$) .1 and .01 we need to have $n \approx 1.364(10^{10})$ and $n \approx 2.340(10^{12})$, respectively. This is well beyond the range of the values in Tables 2-6. Our asymptotic results predict that the mass should migrate from $k_1 - 2$ to $k_1 - 1$ and eventually to k_1 , at least for values of β bounded away from zero. Note, however, that the numerical mass is well predicted by our asymptotic expansions, even for moderate values of n .

To better see the convergence of mass to one or two points, it is best to consider special subsequences $n(i)$, in order to avoid the oscillations caused by the appearance of $\beta(n)$. In Table 9, we consider the first few n which correspond to local minima of $|\Delta(n)|$, as given in Table 8. For these we give the exact masses $m_{-1} = h_n^{k_1-2} - h_n^{k_1-3}$, m_0 and m_1 . As previously discussed the theory predicts that m_0 and m_1 will be close to .991 and .009. From Table 9 we see only that there is a gradual migration of mass from $k_1 - 2$ to $k_1 - 1$. However, even at $n = 127$ most of the mass is still at $k_1 - 2$.

Table 9: Probability mass at k_1 only.

n	k_1	m_{-1}	m_0	m_1
7	5	.7715	.1455	.0049
12	6	.7301	.1530	.0067
22	7	.7235	.2122	.0110
39	8	.7120	.2446	.0124
70	9	.6849	.2885	.0141
127	10	.6345	.3461	.0162

Table 10: Probability mass at k_0 and k_1 .

n	k_1	m_{-1}	m_0	m_1
7	5	.7715	.1455	.0049
13	6	.7280	.2192	.0117
24	7	.6643	.3028	.0190
44	8	.5801	.3903	.0261
80	9	.4946	.4723	.0320
145	10	.4129	.5495	.0369
264	11	.3272	.6297	.0423

Table 11: Numerical Evaluation of $\Phi(2)$.

k	$\Phi_{NUM}(2 - 2^k)$	$\Phi(2)$	3-term approx.
2	.3642	.5595	.2544
3	.4441	.5595	.3444
4	.4922	.5595	.4207
5	.5208	.5595	.4744
6	.5375	.5595	.5091
7	.5472	.5595	.5304

In Table 10 we choose a sequence $n(i)$ so as to minimize $1 - \beta(n)$. This is the optimal way to “push” the mass out of $k = k_1 - 2$ and into $k = k_1$ for moderate n . The table clearly shows a gradual migration of mass from $k_1 - 2$ into $k_1 - 1$, and an increase of mass at k_1 , though the latter is still below .05 when $n = 264$. Our asymptotic results predict that the migration from $k_1 - 1$ to k_1 along this subsequence would occur when $n \approx 10^{10}$, which is well beyond the range where it is feasible to do numerical calculations.

We conclude by discussing the numerical computation of $\Phi(\xi)$. We define

$$\Phi_{NUM}(\xi; k) \equiv -\frac{1}{n} \log(h_n^k), \quad \xi = n2^{-k}.$$

Our results predict that for each fixed $0 < \xi \leq 2$, $\Phi_{NUM}(\xi; k)$ should approach $\Phi(\xi)$ as $k \rightarrow \infty$.

When $\xi = 2$ we have the theoretical value exactly, as $\Phi(2) = 1 - \log(\sqrt{2C_0}) = .55946 \dots$. Note that $n = 2^{k+1} - 1$ corresponds to $\xi = 2 - 2^{-k}$. In view of (2.17) we should have, as $k \rightarrow \infty$,

$$\Phi(2 - 2^{-k}) = \Phi(2) - \frac{k}{2} 2^{-k} - \frac{1}{4} \log(2C_0) 2^{-k} + o(2^{-k}). \quad (3.18)$$

In Table 11 we give $\Phi_{NUM}(2 - 2^{-k}; k) = -\log(h_{2^{k+1}-1}^k)/(2^{k+1} - 1)$ for k in the range $[2, 7]$. We compare Φ_{NUM} to both $\Phi(2)$ and the 3-term approximation in (3.18). This shows that Φ_{NUM} is indeed converging to the theoretical value.

To summarize, we have shown that the asymptotic results provide, on the whole, very good approximations to h_n^k . To achieve this, however, it is sometimes necessary to obtain more uniform and/or higher order results than those in (2.11)–(2.22). The accumulation of

mass at $k = k_1$ is not in good agreement with the numerical results, but the asymptotics explain this, and estimate the size of n necessary for the mass to migrate to k_1 .

4 The Right Tail Analysis

In this section, we derive parts (i) and (ii) of Theorem 1, that is, the far right-tail approximation (2.11) and the right-tail asymptotic expansion (2.12).

4.1 The Far Right-Tail Analysis

Since we analyze h_n^k when the distribution is asymptotically close to one, it is convenient to set $j = n - k$ and

$$\begin{aligned} h_n^k &= 1 - H_n^k, \\ H_n^k &= H_n^{n-j} = \mathcal{L}_j(n). \end{aligned} \quad (4.1)$$

Using (4.1) in (2.2) we obtain

$$2^n \mathcal{L}_j(n+1) = 2 \sum_{\ell=0}^{j-2} \binom{n}{\ell} \mathcal{L}_{j-\ell}(n-\ell) - \sum_{i=n-j+2}^{j-2} \binom{n}{i} \mathcal{L}_{i+j-n}(i) \mathcal{L}_{j-i}(n-i) \quad (4.2)$$

for $n \geq j$. Here we have multiplied (2.2) by 2^n and used the fact that $h_n^k = 1$ for $k \geq n-1$, which is equivalent to $\mathcal{L}_j(n) = 0$ for $j \leq 1$. We also note that if $k+1 \geq n$, $h_n^k = 1$ is an exact solution to (2.2), since every term in the sum lies in this range.

We have thus decomposed the non-linear right side of (2.2) into a linear part and a non-linear part, corresponding to the two sums in (4.2). For $n \geq 2j-3$ the second sum is void and the equation is *exactly* linear. The initial condition (2.3) becomes

$$\begin{aligned} \mathcal{L}_j(j) &= 0, & j &= 0, 1, \\ \mathcal{L}_j(j) &= 1, & j &\geq 2. \end{aligned} \quad (4.3)$$

First we compute $\mathcal{L}_j(n)$ exactly for some small values of $j \geq 2$. Setting $j = 2$ we obtain from (4.2)

$$2^n \mathcal{L}_2(n+1) = 2 \mathcal{L}_2(n), \quad n \geq 2$$

which we can easily solve subject to $\mathcal{L}_2(2) = 1$, to get

$$\mathcal{L}_2(n) = 2^{-n^2/2} 2^{3n/2} 2^{-1}, \quad n \geq 2, \quad (4.4)$$

and this proves (2.6). Letting $j = 3$ in (4.2) yields

$$2^n \mathcal{L}_3(n+1) = 2\mathcal{L}_3(n) + 2n\mathcal{L}_2(n-1), \quad n \geq 3 \quad (4.5)$$

Using (4.4) in (4.5), we see that the above is a first order, inhomogeneous linear difference equation. To solve it, we set $\mathcal{L}_3(n) = 2^{-n^2/2} 2^{5n/2} f(n)$ to get

$$4f(n+1) = 2f(n) + \frac{n}{4}$$

so that

$$f(n) = \frac{n}{8} - \frac{1}{4} + c_1 2^{-n}$$

where c_1 is an arbitrary constant. By setting $n = 3$ in (4.5) we find that $8\mathcal{L}_3(4) = 2\mathcal{L}_3(3) + 6\mathcal{L}_2(2) = 8$ so that $\mathcal{L}_3(4) = 1$. We must thus have $f(4) = 1/4$ which forces $c_1 = 0$. Thus we have

$$\mathcal{L}_3(n) = \left(\frac{n}{8} - \frac{1}{4}\right) 2^{-n^2/2} 2^{5n/2}, \quad n \geq 4 \quad (4.6)$$

which is equivalent to (2.7). Note that (4.6) remains valid if $n = 3$ since $\mathcal{L}_3(3) = 1$ by (4.3).

Next we consider (4.2) with $j = 4$. For $n \geq 5$ the non-linear term drops out. By considering this equation with $n = 4$ and using the previous results we find that $\mathcal{L}_4(5) = 1$. We can simplify the solution of (4.2) (for arbitrary j) by setting ($k = n - j$)

$$\mathcal{L}_j(n) = 2^{-(n-j)^2/2} 2^{(j-n)/2} F_j(n) = 2^{-k^2/2} 2^{-k/2} F_j(n) \quad (4.7)$$

which yields

$$2^{j-2} F_j(n+1) = \sum_{\ell=0}^{j-2} \binom{n}{\ell} F_{j-\ell}(n-\ell), \quad 4 \leq 2j \leq n+3. \quad (4.8)$$

Our previous calculations show that

$$F_2(n) = 1, \quad F_3(n) = n - 2. \quad (4.9)$$

For $j = 4$, (4.8) becomes

$$\begin{aligned} 4F_4(n+1) &= F_4(n) + nF_3(n-1) + \binom{n}{2} F_2(n-2), \quad n \geq 5 \\ &= F_4(n) + n(n-3) + \frac{1}{2}n(n-1) \end{aligned}$$

whose general solution is

$$F_4(n) = \frac{n^2}{2} - \frac{5n}{2} + \frac{8}{3} + c_2 4^{-n}, \quad n \geq 6. \quad (4.10)$$

Since $\mathcal{L}_4(5) = 1$ we have $F_4(5) = 2$ which determines $c_2 = -2 \cdot 4^5/5 = -2048/5$. With (4.10) and (4.7) we have established the result in (2.8). Also, (4.10) remains valid when $n = 5$.

Proceeding in this manner we can compute $F_j(n)$ for any j . A lengthy computation shows that

$$F_5(n) = \frac{n^3}{6} - \frac{3n^2}{2} + 4n - \frac{64}{21} - \frac{1}{3}(n-1)4^{6-n} - \frac{125}{21}8^{6-n}, \quad n \geq 8$$

so that

$$\mathcal{L}_5(n) = 2^{-n^2/2} 2^{9n/2} 2^{-10} \left[\frac{n^3}{6} - \frac{3n^2}{2} + 4n - \frac{64}{21} - \frac{1}{3}(n-1)4^{6-n} - \frac{125}{21}8^{6-n} \right], \quad n \geq 8 \quad (4.11)$$

and we have $\mathcal{L}_5(5) = \mathcal{L}_5(6) = 1$ and $\mathcal{L}_5(7) = 59/64$. The last equality also shows that (4.11) remains valid at $n = 7$.

Next we discuss the asymptotics as $n \rightarrow \infty$. For any fixed j and $n \rightarrow \infty$, we can conclude that

$$\mathcal{L}_j(n) \sim 2^{-n^2/2} 2^{(j-1/2)n} n^{j-2} C(j). \quad (4.12)$$

To compute $C(j)$ we use (4.12) in (4.2) to obtain

$$\begin{aligned} 2^n C(j) &= (n+1)^{j-2} 2^{-n^2/2} 2^{-n} 2^{-1/2} 2^{(j-1/2)n} n^{j-1/2} \\ &\sim 2 \sum_{\ell=0}^{j-2} \binom{n}{\ell} C(j-\ell) (n-\ell)^{j-\ell-2} 2^{-(n-\ell)^2/2} 2^{(j-\ell-1/2)(n-\ell)}. \end{aligned} \quad (4.13)$$

Noting that $\binom{n}{\ell} n^{-\ell} \sim \frac{1}{\ell!}$ and $(n-\ell)^{j-\ell-2} \sim n^{j-2} n^\ell$ as $n \rightarrow \infty$ the leading terms in (4.13) yields

$$2^{j-1} C(j) = 2 \sum_{\ell=0}^{j-2} \frac{C(j-\ell)}{\ell!} 2^{\ell^2/2} 2^{\ell/2} 2^{-\ell j}, \quad j \geq 2. \quad (4.14)$$

In view of (4.4) and (4.12) we have $C(2) = 1/2$. We can simplify (4.14) by setting

$$C(j) = 2^{-j^2/2} 2^{j/2} D(j)$$

to obtain

$$2^{j-1} D(j) = 2 \sum_{\ell=0}^{j-2} \frac{D(j-\ell)}{\ell!}, \quad j \geq 2. \quad (4.15)$$

Solving (4.15) subject to $D(2) = 1$ yields

$$D(j) = \frac{1}{(j-2)!} \quad (4.16)$$

which completes the derivation of the asymptotic formula for $1 - h_n^k = H_n^k = \mathcal{L}_j(n)$ in (2.11).

4.2 The Right-Tail Analysis

We next consider n and $j = n - k$ simultaneously large and such that $0 < k/n < 1$, so that we are in the range of Theorem 1(ii). We will show that the asymptotic relation (4.12) ceases to be valid if j is as large as $O(n)$. Note also that $j = n - k$ can be as large as $n - \log_2 n$. From our consideration of small values of j , it is easy to see that $F_j(n)$ has the form

$$F_j(n) = \tilde{F}_j(n) + O(\rho^n), \quad n \geq 2j - 3 \quad (4.17)$$

where $\rho < 1$ and $\tilde{F}_j(n)$ is a polynomial of degree $j - 2$. The exponentially small term corresponds to terms that are $O(4^{-n}, 8^{-n}, 16^{-n}, \dots)$, modulo some factors algebraic in n , as $n \rightarrow \infty$. Our calculations also showed that $\tilde{F}_j(n)$ depends on the initial conditions (4.3) only through the value $\mathcal{L}_2(2) = 1$. We then used the fact that $\mathcal{L}_3(3) = 1$ to conclude that $c_1 = 0$, which implies that there are no terms of order $O(2^{-n})$ in the exponentially small term. Then the terms proportional to 4^{-n} are completely determined by the value of c_2 , which follows from $\mathcal{L}_4(4) = 1$.

For the region $n/j > 2$ in the (j, n) plane the problem (4.2) is exactly linear. For $1 < n/j < 2$ the non-linear terms are present, but as long as $\mathcal{L}_j(n)$ is asymptotically small (i.e., h_n^k is asymptotically close to 1), the non-linear term is small compared to the linear term(s). Thus in the right tail we can drop the second sum in (4.2) and replace $=$ by $\sim$. Also, the exponentially small part of the solution $F_j(n)$ depends on the values of $F_j(n)$ when $n/j = 2$, which in turn depends on the initial condition(s) (4.3). The initial data propagates from $n/j = 1$ to the range $n/j > 2$ via the region where the non-linear term is present. However, these terms are exponentially small compared to $\tilde{F}_j(n)$. This discussion shows that

$$F_j(n) \sim \tilde{F}_j(n) \quad (4.18)$$

as long as j is such that $\mathcal{L}_j(n)$ is asymptotically small.

We next compute $\tilde{F}_j(n)$ for arbitrary j and use the result to obtain the asymptotics of $\mathcal{L}_j(n)$ (and hence h_n^k) for $n, j \rightarrow \infty$ with n/j fixed and $n/j > 1$. This implies that $k, n \rightarrow \infty$ at the same rate and $0 < k/n < 1$.

Since $\tilde{F}_j(n)$ is a polynomial in n of degree $j - 2$ we can write

$$\tilde{F}_j(n) = \binom{n}{j-2} + d_1(j) \binom{n}{j-3} + d_2(j) \binom{n}{j-4} + \dots \quad (4.19)$$

where the sum truncates after a finite number of terms. Here we have set $d_0(j) = 1$, which follows from (4.12) and (4.16). For a fixed j and $n \rightarrow \infty$ the successive terms in (4.19) are asymptotically smaller by factors of n^{-1} . However, we will show that this is no longer true if n and j are both large. Using (4.19) in (4.8) yields

$$\begin{aligned} 2^{j-2} \left[\binom{n+1}{j-2} + d_1(j) \binom{n+1}{j-3} + d_2(j) \binom{n+1}{j-4} + \dots \right] \\ = \sum_{\ell=0}^{j-2} \binom{n}{\ell} \left[\binom{n-\ell}{j-\ell-2} + d_1(j-\ell) \binom{n-\ell}{j-\ell-3} \right. \\ \left. + d_2(j-\ell) \binom{n-\ell}{j-\ell-4} + \dots \right]. \end{aligned} \quad (4.20)$$

Using the identities

$$\begin{aligned} \binom{n+1}{m} &= \binom{n}{m} + \binom{n}{m-1}, \\ \binom{n}{\ell} \binom{n-\ell}{j-\ell-m} &= \binom{n}{j-m} \binom{j-m}{\ell}, \end{aligned}$$

and comparing coefficients of $\binom{n}{j-m}$ in (4.20) for $m \geq 2$, we are led to the recurrences

$$2^{j-2}[1 + d_1(j)] = \sum_{\ell=0}^{j-3} d_1(j-\ell) \binom{j-3}{\ell} \quad (4.21)$$

and

$$2^{j-2}[d_m(j) + d_{m+1}(j)] = \sum_{\ell=0}^{j-m-3} d_{m+1}(j-\ell) \binom{j-m-3}{\ell}, \quad m \geq 1. \quad (4.22)$$

From (4.21) it follows that $d_1(j) = -2$, which is independent of j . We then find from (4.22) that $d_m(j)$ is independent of j for all $m \geq 1$, and $d_m(j) = d_m$ satisfies

$$d_m + d_{m-1} = 2^{-m} d_m$$

so that $d_m = (-1)^m \prod_{N=1}^m (1 - 2^{-N})^{-1}$ and hence

$$\tilde{F}_j(n) = \sum_{m=0}^{j-2} (-1)^m \binom{n}{j-2-m} \prod_{N=1}^m (1 - 2^{-N})^{-1}. \quad (4.23)$$

We can easily check that this agrees with our previous results for $2 \leq j \leq 5$. If in any product the upper limit exceeds the lower limit, we set the product equal to 1.

Now consider the binomial coefficient in (4.23) in the limit $n, j \rightarrow \infty$ with m and $\alpha \equiv n/j$ fixed. Using Stirling's formula we obtain

$$\begin{aligned} \frac{n!}{(n+m-j+2)!} \frac{1}{(j-2-m)!} &\sim \frac{n!}{(n-j+2)!} \frac{1}{(n-j)^m} \frac{j^m}{(j-2)!} \\ &\sim \frac{n^n}{(n-j)^{n-j} j^j} \frac{j^{3/2} \sqrt{n}}{(n-j)^{5/2}} \left(\frac{1}{\alpha-1} \right)^m \frac{1}{\sqrt{2\pi}}. \end{aligned} \quad (4.24)$$

It follows that

$$\tilde{F}_j(n) \sim \frac{n^n}{(n-j)^{n-j} j^j} \frac{j^{3/2} \sqrt{n\pi}}{\sqrt{2}(n-j)^{5/2}} I(\alpha) \quad (4.25)$$

where

$$I(\alpha) \equiv \frac{1}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^m}{(\alpha-1)^m} \left[\prod_{\ell=1}^m (1-2^{-\ell}) \right]^{-1}. \quad (4.26)$$

The sum in (4.26) is absolutely convergent for $\alpha > 2$. We can extend its range of validity by writing the product as $\prod_{\ell=1}^m (\cdots) = \prod_{\ell=1}^{\infty} (\cdots) / \prod_{\ell=m+1}^{\infty} (\cdots)$ and subtracting $\prod_{\ell=1}^{\infty} (\cdots)$. The resulting sum will converge for $\alpha > 3/2$, but we will need to know the behavior of $\tilde{F}_j(n)$ as $n \rightarrow \infty$ for all $\alpha = n/j > 1$. Thus we need to obtain a different representation for $\tilde{F}_j(n)$, which will be more useful for asymptotic analysis.

We begin by considering the function

$$A(z) = \prod_{\ell=1}^{\infty} \exp \left(\frac{1-2^{\ell z}}{\ell(2^{\ell}-1)} \right), \quad \Re(z) < 1. \quad (4.27)$$

From (4.27) we can easily show that $A(\cdot)$ satisfies the functional equation

$$A(z+1) = A(z)[1-2^z], \quad (4.28)$$

which can be used to analytically continue $A(\cdot)$ into the range $\Re(z) > 1$. Observe that

$$A(-m) = \prod_{N=1}^m (1-2^{-N})^{-1} = \prod_{\ell=1}^{\infty} \exp \left(\frac{1-2^{-m\ell}}{\ell(2^{\ell}-1)} \right).$$

Indeed, the above follows from

$$\begin{aligned} \prod_{\ell=1}^{\infty} \exp \left(\frac{1-2^{-m\ell}}{\ell(2^{\ell}-1)} \right) &= \exp \left(\sum_{\ell=1}^{\infty} \frac{1}{\ell} 2^{-\ell} \frac{1-2^{-\ell m}}{1-2^{-\ell}} \right) \\ &= \exp \left(\sum_{N=1}^m \sum_{\ell=1}^{\infty} 2^{-N\ell} \frac{1}{\ell} \right) \\ &= \exp \left(- \sum_{N=1}^m \log(1-2^{-N}) \right) = \prod_{N=1}^m (1-2^{-N})^{-1} \end{aligned}$$

Then, we can rewrite the sum in (4.23) as

$$\begin{aligned}\bar{F}_j(n) &= \sum_{\ell=0}^{j-2} (-1)^\ell e^{\pi i j} \binom{n}{\ell} A(\ell+2-j) \\ &= \frac{n!}{2\pi i} (-1)^{n+j} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} \frac{\Gamma(z+j-n-2)}{\Gamma(z+j-1)} A(z) dz\end{aligned}\quad (4.29)$$

where $0 < \Re(z) < 1$ on the contour of integration. The last inequality is basically Rice's formula (cf. [7, 28]), but we can derive it directly. Indeed, in the region $\Re(z) < 1$ the integrand has simple poles at $z = 0, -1, \dots, -j+1, -j+2$ and $\text{Res}[z = -m] = \frac{(-1)^{m-j+n}}{(j-m-2)!} \frac{A(-m)}{(n+m-j-2)!}$, $m \geq 0$, where $\text{Res}[z = A]$ stands for at A for the function under the integral. Next, we close the contour in the left half-plane and the integral is equal to the finite residue sum, which is the same as (4.23).

Noting that

$$\begin{aligned}\sum_{\ell=1}^{\infty} \frac{1-2^{\ell z}}{\ell(2^{\ell}-1)} &= \sum_{\ell=1}^{\infty} \frac{2^{-\ell}}{\ell} \sum_{N=0}^{\infty} 2^{-N\ell} - \sum_{\ell=1}^{\infty} \frac{2^{\ell(z-1)}}{\ell} \sum_{N=0}^{\infty} 2^{-N\ell} \\ &= - \sum_{N=0}^{\infty} \log(1-2^{-N-1}) + \sum_{N=0}^{\infty} \log(1-2^{z-N-1})\end{aligned}$$

the expression in (4.27) becomes

$$A(z) = \prod_{\ell=1}^{\infty} (1-2^{-\ell})^{-1} \prod_{m=1}^{\infty} (1-2^{z-m}) \quad (4.30)$$

which applies for all z , and thus gives the analytic continuation of (4.27) into the half-plane $\Re(z) > 1$. Expression (4.30) also shows that $A(z)$ is an entire function, with zeros at $z = 1, 2, 3, \dots$.

In order to find another representation for $A(z)$, we define $p(z)$ by the relation

$$A(z) = e^{-\pi i z} 2^{z^2/2} 2^{-z/2} e^{p(z)} \prod_{m=1}^{\infty} \exp\left(\frac{2^{-mz}}{m(1-2^{-m})}\right) \quad (4.31)$$

and we also note that for $\Re(z) > 0$

$$\prod_{m=1}^{\infty} \exp\left(\frac{2^{-mz}}{m(1-2^{-m})}\right) = \prod_{m=0}^{\infty} \frac{1}{1-2^{-z-m}}. \quad (4.32)$$

Denoting the above product by $g(z)$, it satisfies $g(z+1) = g(z)(1-2^{-z})$ and $g(z) \sim 1$ as $z \rightarrow \infty$ with $\Re(z) > 0$. Using (4.31) in (4.28) then shows that $p(z+1) = p(z)$.

We next identify explicitly the periodic function $p(\cdot)$ in terms of the Jacobi elliptic theta function (cf. [3]):

$$\vartheta_1(u) = \vartheta_1(u|\tau) = 2q^{1/4} \sin(u) \prod_{m=1}^{\infty} (1 - 2\cos(2u)q^{2m} + q^{4m})(1 - q^{2m})$$

where $e^{\pi i \tau} = q$. Setting $q = 1/\sqrt{2}$ and $u = (i \log 2)z/2$ the above becomes

$$\begin{aligned} \vartheta\left(\frac{i}{2}(\log 2)z\right) &= i2^{7/8} \left[\prod_{m=1}^{\infty} (1 - 2^{-m}) \right] \sinh\left(\frac{\log 2}{2}z\right) \prod_{m=1}^{\infty} [1 - 2 \cosh(z \log 2)2^{-m} + 4^{-m}] \\ &= i2^{-1/8} 2^{z/2} (1 - 2^{-z}) \prod_{m=1}^{\infty} (1 - 2^{-m}) \prod_{m=1}^{\infty} (1 - 2^{z-m}) \prod_{m=1}^{\infty} (1 - 2^{-z-m}) \end{aligned} \quad (4.33)$$

Using (4.30), (4.32) and (4.33) in (4.31) we find that

$$e^{p(z)} = 2^{-z^2/2} e^{\pi i z} (-i) 2^{1/8} \left[\prod_{m=1}^{\infty} (1 - 2^{-m})^{-2} \right] \vartheta_1\left(\frac{i}{2}(\log 2)z\right). \quad (4.34)$$

The periodicity of $p(z)$ in the real direction follows from the “quasiperiodicity” of the theta function in the imaginary direction.

We return to the discussion of the asymptotics of $\tilde{F}_j(n)$ (cf. 4.29) for $n, j \rightarrow \infty$ with $\alpha > 1$. For z fixed and n, j large we have

$$\begin{aligned} \frac{\Gamma(z+j-n-2)}{\Gamma(z+j-1)} &= \frac{\pi}{\sin(\pi z)} \frac{(-1)^{n+j}}{\Gamma(n-j+3-z)} \frac{1}{\Gamma(z+j-1)} \\ &\sim \frac{\pi}{\sin(\pi z)} \frac{(-1)^{n+j}}{\Gamma(n-j+3)} \frac{1}{\Gamma(j-1)} \left(\frac{n}{j} - 1\right)^z \\ &\sim \frac{(-1)^{n+j}}{\sin(\pi z)} \frac{1}{2} \frac{e^n}{(n-j)^{n-j} j^j} \frac{j^{3/2}}{(n-j)^{5/2}} (\alpha - 1)^z. \end{aligned}$$

Using the above in (4.29) and approximating $n!$ by Stirling’s formula yields

$$\tilde{F}_j(n) \sim \frac{n^n}{(n-j)^{n-j} j^j} \frac{\sqrt{n} j^{3/2}}{(n-j)^{5/2}} \frac{\sqrt{\pi}}{\sqrt{2}} \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} \frac{(\alpha-1)^z}{\sin(\pi z)} A(z) dz \quad (4.35)$$

which yields the second representation (2.14) for $I(\alpha)$ that appears in Theorem 1(ii). For $\alpha > 2$ we can close the integration contour in the left half-plane and recover (4.25). However, (4.35) applies for all $\alpha > 1$.

4.3 Asymptotic Matching for $\alpha \rightarrow 1^+$

For purposes of asymptotic matching, we will need to know the behavior of the approximation in (4.35) as $\alpha \rightarrow 1^+$. We first observe that since $A(z)$ is an entire function with zeros

of $z = 1, 2, 3, \dots$, the integrand in (4.35) is analytic for $\Re(z) > 0$. The asymptotics of the integral as $\alpha \rightarrow 1$ will be obtained by shifting the contour to the right. Since there are no singularities in the right half-plane, the asymptotics must be governed by a saddle point (cf. [4, 25]). We use the representation (4.31) for $A(z)$ and note that the infinite product in (4.31) is $1 + O(2^{-z})$ as $z \rightarrow \infty$ in the right half-plane. We thus write the integral in (4.35) as

$$J \equiv \frac{1}{2\pi i} \int_B e^{-\lambda z} 2^{z^2/2} 2^{-z/2} Q(z) [1 + O(2^{-z})] dz \quad (4.36)$$

where B is any vertical contour with $\Re(z) > 0$ and

$$Q(z) = \frac{e^{-\pi i z}}{\sin(\pi z)} e^{p(z)}, \quad \alpha - 1 = e^{-\lambda}. \quad (4.37)$$

The periodicity of $p(\cdot)$ implies that $Q(z+1) = Q(z)$ and $\alpha \rightarrow 1^+$ corresponds to $\lambda \rightarrow +\infty$.

For $\lambda \rightarrow \infty$ the integrand in (4.36) has a saddle point where

$$\frac{d}{dz} \left[-\lambda z + \frac{z^2}{2} \log 2 - \frac{z}{2} \log 2 \right] = z \log 2 - \lambda - \frac{1}{2} \log 2 = 0.$$

Thus the saddle is at $z \approx \lambda / \log 2$ and we set

$$z = \frac{\lambda}{\log 2} + \frac{1}{2} + \zeta. \quad (4.38)$$

We now use (4.36) – (4.38) and obtain

$$J \sim e^{-\lambda/2} e^{-\lambda^2/(2 \log 2)} 2^{-1/8} \frac{1}{2\pi} \int_{-\infty}^{\infty} 2^{-t^2/2} Q\left(\frac{\lambda}{\log 2} + \frac{1}{2} + it\right) dt \quad (4.39)$$

for $\lambda \rightarrow \infty$. By periodicity, the above integral is $O(1)$ in this limit. We use (4.39) in (4.35), set $j = n - k$ and expand the result for $k/n \rightarrow 0$. Noting that $e^{-\lambda/2} = \sqrt{k/(n-k)} \sim \sqrt{k/n}$ and $\lambda = \log(n/k) + O(k/n)$ we find that

$$\begin{aligned} \bar{F}_j(n) &\sim e^{k \log n + k - k \log k} \frac{n^{3/2}}{k^2} \frac{2^{-9/8}}{\sqrt{\log 2}} Q_* \left(\log_2 n - \log_2 k + \frac{1}{2} \right) \\ &\times \exp \left[-\frac{1}{2 \log 2} (\log n - \log k)^2 \right] \end{aligned} \quad (4.40)$$

where

$$Q_*(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2/2} Q\left(z + \frac{is}{\sqrt{\log 2}}\right) ds. \quad (4.41)$$

It is important to note that (4.40) was obtained by first taking $n, j \rightarrow \infty$ with n/j fixed and then expanding the result for $n/j = \alpha \rightarrow 1$. Thus (4.40) applies for $k/n \rightarrow 0$, but we will shortly show that we also need $k/\log_2 n \rightarrow \infty$.

We next consider the limits $n \rightarrow \infty$ with $k - \log_2 n = O(1)$ and then $n \rightarrow \infty$ with $k/\log_2 n = \nu > 1$. The first limit has no significance to the distribution h_n^k , since if $k = \log_2 n + O(1)$ we are no longer in the right tail and thus we cannot linearize (4.2). However, the function (4.23) is defined for *all* k . In Section 5 we will show that the limit with $k/\log_2 n$ fixed > 1 is important for the asymptotic matching between the left and right tails.

We return to the representation of $\tilde{F}_j(n)$ in (4.29) and set $z = n - j + s = k + s$. For k, n large and $s = O(1)$ we use $\Gamma(z + j - 1) = \Gamma(n - 1 + s) \sim \Gamma(n + 1)n^{-2}n^s = n!n^{s-2}$ and the approximation for $A(z)$ as $z \rightarrow \infty$. Then the asymptotic form of (4.29) is

$$\begin{aligned}\tilde{F}_j(n) &\sim \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{\pi n^{2-s}}{\Gamma(3-s)} Q(s) \exp\left(\frac{\log 2}{2}(k+s)(k+s-1)\right) ds \\ &= \frac{1}{2\pi i} \pi 2^{k^2/2} 2^{-k/2} n^2 \int_{-i\infty}^{i\infty} \frac{Q(s)}{\Gamma(3-s)} 2^{s^2/2} 2^{-s/2} e^{(k \log 2 - \log n)s} ds\end{aligned}\quad (4.42)$$

where the integral is $O(1)$ for $k = \log_2 n + O(1)$. If we use (4.42) in (4.7), we would find that $\mathcal{L}_j(n) = O(n)$ which corresponds to h_n^k large and negative!

Next we show that (4.42) asymptotically matches to (4.40) in an intermediate limit where $n, k \rightarrow \infty$ with $k/n \rightarrow 0$ but $k - \log_2 n \rightarrow \infty$. We also obtain an approximation to $\tilde{F}_j(n)$ that applies for $k/\log_2 n \in (1, \infty)$. We set $k = \nu \log_2 n$ with $\nu > 1$ and expand (4.42) for $n \rightarrow \infty$. To leading order, the saddle point equation for (4.42) is

$$\frac{d}{ds} \left[\frac{s^2}{2} \log 2 + (\nu - 1)s \log n \right] = s \log 2 + (\nu - 1) \log n = 0$$

so the saddle is at $s \approx -(\nu - 1) \log_2 n$.

Since this is asymptotically large, the factor $1/\Gamma(3-s)$ also affects the location of the saddle. For $s \rightarrow -\infty$ we thus approximate (cf. [1])

$$\frac{1}{\Gamma(3-s)} = (-s)^{-5/2} \frac{1}{\sqrt{2\pi}} e^{s \log(-s) - s} \left(1 + O\left(\frac{1}{s}\right) \right). \quad (4.43)$$

Using (4.43) in (4.42) and expanding the integrand near the saddle $s = s_0 \equiv -(\nu - 1) \log_2 n$ yields

$$\begin{aligned}\tilde{F}_j(n) &\sim \sqrt{\frac{\pi}{2}} 2^{k^2/2} 2^{-k/2} n^2 \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} Q(s) e^{F(s_0)} e^{F'(s_0)(s-s_0)} e^{F''(s_0)(s-s_0)^2/2} \\ &\times (-s)^{-5/2} \left[1 + O\left(\frac{1}{s}, F'''(s_0)(s-s_0)^3\right) \right] ds\end{aligned}\quad (4.44)$$

where

$$F(s) = -s + s \log(-s) + \frac{\log 2}{2} s^2 - \frac{\log 2}{2} s + (\nu - 1)s \log n$$

so that

$$F'(s) = \log(-s) + (\log 2)s - \frac{1}{2} \log 2 + (\nu - 1) \log n$$

and hence $F'(s_0) = \log(-s_0) - (\log 2)/2$, $F''(s_0) = 1/s_0 + \log 2 \sim \log 2$. Setting $s = s_0 - \log_2(-s_0) + \frac{1}{2} + it$, the leading term for (4.44) becomes

$$\begin{aligned} \tilde{F}_j(n) &\sim \sqrt{\frac{\pi}{2}} 2^{k^2/2} 2^{-k/2} n^2 (-s_0)^{-2} e^{F(s_0)} \exp\left(-\frac{1}{2}(\log 2)[\log_2(-s_0)]^2\right) 2^{-1/8} \\ &\times \frac{1}{2\pi} \int_{-\infty}^{\infty} 2^{-t^2/2} Q\left(s_0 - \log_2(-s_0) + \frac{1}{2} + it\right) dt \\ &= \frac{2^{-9/8}}{\sqrt{\log 2}} (k - \log_2 n)^{-2} \exp\left(k \log n - \frac{1}{2} \frac{\log^2 n}{\log 2} + \frac{3}{2} \log n\right) \\ &\times \exp\left[-(k - \log_2 n) \log(k - \log_2 n) + k - \log_2 n - \frac{\log^2(k - \log_2 n)}{2 \log 2}\right] \quad (4.45) \\ &\times Q_*\left(\log_2 n - \log_2(k - \log_2 n) + \frac{1}{2}\right) \end{aligned}$$

where $Q_*(z)$ is given by (4.41). Expression (4.45) holds in the limit $k, n \rightarrow \infty$ with $k/\log_2 n$ fixed and > 1 . We can easily show that it matches to (4.40) in the intermediate limit where $k/n \rightarrow 0$, $k/\log_2 n \rightarrow \infty$. In this limit

$$\begin{aligned} &-(k - \log_2 n) \log(k - \log_2 n) + k - \log_2 n - \frac{\log^2(k - \log_2 n)}{2 \log 2} \\ &= -k \log k + k + (\log_2 n)(\log k) - \frac{\log^2 k}{2 \log 2} + O(1) \end{aligned}$$

so that (4.45) becomes the same as (4.40). This also shows that (4.40) only applies to the range where $k/\log_2 n \rightarrow \infty$. In order to asymptotically match the right and left tails, we shall need to use (4.45). We also note that if we multiply (4.45) by $2^{-k^2/2} 2^{-k/2}$ (thus obtaining the corresponding approximation to $1 - h_n^k = H_n^k$), the result has the form $n \times [\text{function of } (k - \log_2 n)]$. This is precisely what is needed in order to match to the WKB expansion that we will derive in Section 5.

For $k, n \rightarrow \infty$ with $k/\log_2 n = \nu > 1$, we are still in the right tail. We can estimate the value of k where there is appreciable mass by using (4.45). We argue that when $H_n^k = 2^{-k^2/2} 2^{-k/2} F_j(n)$ becomes $O(1)$ in n , then we are no longer in the tail. Since Q_* is clearly $O(1)$, this condition is equivalent to

$$\begin{aligned} &-\frac{k^2}{2} \log 2 - \frac{k}{2} \log 2 + k \log n + \frac{3}{2} \log n - \frac{1}{2} \frac{\log^2 n}{\log 2} \quad (4.46) \\ &-(k - \log_2 n) \log(k - \log_2 n) + k - \log_2 n - \frac{\log^2(k - \log_2 n)}{2 \log 2} - 2 \log(k - \log_2 n) = O(1). \end{aligned}$$

The largest terms in (4.46) are $-\frac{\log 2}{2}(k - \log_2 n)^2 + \log n$ which balance when $k - \log_2 n \sim \sqrt{2 \log_2 n}$. This gives a rough argument that the mean $\mathbb{E}[\mathcal{H}_n]$ behaves as $\mathbb{E}[\mathcal{H}_n] - \log_2 n \sim \sqrt{2 \log_2 n}$. A more careful analysis of (4.46) shows that the left side is $O(1)$ for

$$\begin{aligned} k &= \log_2 n + \sqrt{2 \log_2 n} - \log_2(\sqrt{2 \log_2 n}) + \frac{1}{\log 2} - \frac{1}{2} \\ &\quad - \frac{3 \log_2(\sqrt{2 \log_2 n})}{2 \sqrt{2 \log_2 n}} + O((\log n)^{-1/2}). \end{aligned} \quad (4.47)$$

We discussed the condition in (4.47) in more detail in Section 3.

4.4 Another Representation for $Q(z)$

We next obtain another representation for the periodic function $p(z)$ (and hence $Q(z)$), which involves a Fourier series that we need in some computations. We let $\tilde{A}(z)$ be a solution of (4.28) and set $\tilde{A}(z) = \exp[B(z)]$. Then $B(z)$ satisfies

$$B(z+1) - B(z) = \log(1 - 2^z). \quad (4.48)$$

Introducing the two-sided Laplace transform

$$B^*(s) = \int_{-\infty}^{\infty} B(z) e^{-sz} dz$$

we find that

$$\begin{aligned} (e^s - 1)B^*(s) &= \int_{-\infty}^{\infty} \log(1 - 2^z) e^{-sz} dz \\ &= \int_{-\infty}^0 \log(1 - 2^z) e^{-sz} dz + \int_0^{\infty} (\pm \pi i + \log(2^z - 1)) e^{-sz} dz \\ &= \int_0^{\infty} e^{sz} \log(1 - 2^{-z}) dz \pm \frac{\pi i}{s} + \frac{\log 2}{s^2} + \int_0^{\infty} e^{-sz} \log(1 - 2^{-z}) dz \\ &= \pm \frac{\pi i}{s} + \frac{\log 2}{s^2} + \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} \frac{1}{m} \frac{1}{s - m \log 2} \end{aligned} \quad (4.49)$$

where both integrals converge in the strip $-\log 2 < \Re(s) < \log 2$, and the left side is analytic for $0 < \Re(s) < \log 2$. The standard inversion formula then yields

$$B(z) = \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} \frac{e^{sz}}{e^s - 1} \left[\frac{\log 2}{s^2} \pm \frac{\pi i}{s} + \frac{\pi}{s} \left(\cot \left(\frac{\pi s}{\log 2} \right) - \frac{\log 2}{\pi s} \right) \right] ds \quad (4.50)$$

where $0 < \Re(s) < \log 2$ on the contour of integration. Here we have used the partial fractions expansion of $\cot(z)$ to evaluate the sum in (4.49).

The integrand $h(s)$ in (4.50) has a triple pole at $s = 0$, simple poles at $s = 2k\pi i$, $k = \pm 1, \pm 2, \dots$, and simple poles where $s = \ell \log 2$, $\ell = \pm 1, \pm 2, \dots$. To evaluate $B(z)$ as $z \rightarrow \infty$ we shift the contour to the left. A lengthy calculation shows that the residue at $z = 0$ is

$$\text{Res}[h(s); s = 0] = \log 2 \left(\frac{z^2}{2} - \frac{z}{2} + \frac{1}{12} \right) \pm \pi i \left(z - \frac{1}{2} \right) - \frac{\pi^2}{3 \log 2}. \quad (4.51)$$

The residues along the imaginary axis combine with (4.51) to yield

$$\begin{aligned} B(z) &= \frac{\log 2}{2} (z^2 - z) \pm i\pi \left(z - \frac{1}{2} \right) + \frac{\log 2}{12} - \frac{\pi^2}{3 \log 2} \\ &+ \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} e^{2m\pi iz} \left[\pm \frac{1}{2m} - \frac{1}{2m} \coth \left(\frac{2\pi^2 m}{\log 2} \right) \right] + O(2^{-z}) \end{aligned} \quad (4.52)$$

as $z \rightarrow \infty$. By shifting the contour to the right we find that $B(z) \sim -2^z$ as $z \rightarrow -\infty$.

We observe that $\tilde{A} = e^B$ and A in (4.27) (or (4.30)) are related by $\tilde{A}(z)/\tilde{A}(0) = A(z)$. For z real we have

$$\sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} \frac{1}{m} e^{2m\pi iz} = -2\pi i \left(\langle z \rangle - \frac{1}{2} \right)$$

where, to recall, $\langle z \rangle$ is the fractional part of z . By comparing (4.52) with (4.31) we find that

$$\begin{aligned} Q(z) &= \frac{e^{-\pi iz}}{\sin(\pi z)} e^{p(z)} = \frac{1}{\sin(\pi z)} 2^{1/12} \exp \left(-\frac{\pi^2}{3 \log 2} \pm i\pi [z] \right) \prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-1} \\ &\times \exp \left[- \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} \frac{e^{2m\pi iz}}{2m} \coth \left(\frac{2\pi^2 m}{\log 2} \right) \right]. \end{aligned} \quad (4.53)$$

By periodicity it suffices to consider $0 < z < 1$. Writing

$$[\sin(\pi z)]^{-1} = e^{-\log(\sin \pi z)} = \exp \left[\log 2 + \frac{1}{2} \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} \text{sgn}(m) \frac{e^{2m\pi iz}}{m} \right]$$

($\text{sgn}(x)$ is 1 for $x > 0$, -1 for $x < 0$ and 0 for $x = 0$) and noting that

$$\coth(am) - \text{sgn}(m) = \frac{e^{-a|m|}}{\sinh(am)}, \quad a > 0$$

we find from (4.52), (4.33) and (4.34) that

$$\begin{aligned} Q(z) &= 2^{13/12} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-1} \right] \exp \left[-\frac{\pi^2}{3 \log 2} - \sum_{m=1}^{\infty} \frac{e^{-am}}{m \sinh(am)} \cos(2m\pi z) \right] \\ &= \frac{2^{-z^2/2+z/2}(1-2^{-z})}{\sin(\pi z)} \prod_{m=1}^{\infty} (1-2^{-m})(1-2^{z-m})(1-2^{-z-m}) \end{aligned} \quad (4.54)$$

where

$$a = \frac{2\pi^2}{\log 2}.$$

This coincides with the representation of $Q(z)$ given in Theorem 1(iii). From the Fourier series we see that $Q(z)$ is nearly constant since

$$\left| \sum_{m=1}^{\infty} \frac{e^{-am}}{m \sinh(am)} \cos(2m\pi z) \right| \leq \sum_{m=1}^{\infty} \frac{e^{-am}}{m \sinh(am)} = 3.678 \dots (10^{-25}).$$

To summarize the calculations, we have obtained the leading term for $1 - h_n^k$ for the limits (i) $n \rightarrow \infty$, $n - k = j = O(1)$, (ii) $n, k \rightarrow \infty$ with $0 < k/n = 1 - 1/\alpha < 1$ and (iii) $k, n \rightarrow \infty$ with $k/\log_2 n = \nu > 1$. The last corresponds to the “left-most” right tail region. The third region involves correction terms that are $O(1/\log n)$ while the first two had correction terms $O(1/n)$. Thus we expect that the third case will result in the worst numerical agreement (this was discussed in more detail in Section 3).

5 The Left Tail and Central Regime Analyses

We analyze here the distribution when h_n^k is asymptotically small. Then we can no longer linearize (2.2). Whereas the right tail was treated using purely linear analysis, the left tail requires different techniques such as the WKB method and matched asymptotics.

5.1 The Left Tail Analysis

We first prove part (iv) of Theorem 1. We set $2^{k+1} - n = M$, $M \geq 1$ and first show how to compute h_n^k explicitly for small values of M . Since the generating function in (2.4) is a polynomial of degree $2^{k+1} - 1$ we isolate the two leading coefficients by writing

$$H_k(z) = a(k)z^{2^{k+1}-1} + b(k)z^{2^{k+1}-2} + \dots + z + 1. \quad (5.1)$$

It follows that

$$H_k^2(z) = a^2(k)z^{2^{k+2}-2} + 2a(k)b(k)z^{2^{k+2}-3} + O(z^{2^{k+2}-4}) \quad (5.2)$$

and

$$\begin{aligned} H'_{k+1}(2z) &= \frac{1}{2} \left[a(k+1)2^{2^{k+2}-1}(2^{k+2}-1)z^{2^{k+2}-2} \right. \\ &\quad \left. + b(k+1)2^{2^{k+2}-2}(2^{k+2}-2)z^{2^{k+2}-3} + O(z^{2^{k+2}-4}) \right]. \end{aligned} \quad (5.3)$$

By using (2.5), (5.2) and (5.3) we obtain the recurrences

$$2^{2^{k+2}}(2^{k+2} - 1)a(k+1) = 4a^2(k), \quad a(0) = 1 \quad (5.4)$$

$$2^{2^{k+2}}(2^{k+2} - 2)b(k+1) = 16a(k)b(k), \quad b(0) = 1. \quad (5.5)$$

Equation (5.4) is non-linear but (5.5) is linear, once we compute $a(k)$.

To solve (5.4) we set $a(k) = e^{\alpha(k)}$ to get

$$\alpha(k+1) - 2\alpha(k) = -\log \left[\frac{1}{4} 2^{2^{k+2}} (2^{k+2} - 1) \right]. \quad (5.6)$$

If $\alpha(k) = 2^k \tilde{\alpha}(k)$ then (5.6) becomes

$$\tilde{\alpha}(k+1) - \tilde{\alpha}(k) = -2^{-k-1} \log \left[\frac{1}{4} 2^{2^{k+2}} (2^{k+2} - 1) \right]. \quad (5.7)$$

We solve (5.7) subject to $\tilde{\alpha}(0) = 0$ to obtain

$$\alpha(k) = -2^k \sum_{\ell=0}^{k-1} 2^{-\ell-1} \log \left[\frac{1}{4} 2^{2^{\ell+2}} (2^{\ell+2} - 1) \right]. \quad (5.8)$$

Exponentiating (5.8) and some rearrangement yields

$$\begin{aligned} a(k) &= \prod_{\ell=1}^k \left(\frac{4}{2^{2^{\ell+1}} (2^{\ell+1} - 1)} \right)^{2^{k-\ell}} \\ &= 2^{-k2^{k+1}} 4^{2^k-1} \prod_{\ell=1}^k \left(\frac{1}{2^{\ell+1} - 1} \right)^{2^{k-\ell}}. \end{aligned} \quad (5.9)$$

We next obtain the asymptotics of $a(k)$ as $k \rightarrow \infty$. We write

$$2^k \sum_{\ell=1}^k 2^{-\ell} \log(2^{\ell+1} - 1) = (3 \cdot 2^k - k + 3) \log 2 + 2^k \sum_{\ell=1}^k 2^{-\ell} \log(1 - 2^{-\ell-1})$$

so that

$$\begin{aligned} a(k) &= 2^{-(2k+3)2^k} 2^{2^{k+1}+k+1} \left\{ \prod_{\ell=1}^k (1 - 2^{-\ell-1})^{-2^{-\ell}} \right\}^{2^k} \\ &= 2^{-(2k+1)2^k} 2^{k+1} C_0^{2^k} [1 + O(2^{-k})], \quad k \rightarrow \infty, \end{aligned} \quad (5.10)$$

where

$$C_0 = \prod_{\ell=1}^{\infty} (1 - 2^{-\ell-1})^{-2^{-\ell}} \quad (5.11)$$

whose numerical value was given in (2.18) of Section 2.

Since $h_{2^{k+1}-1}^k = (2^{k+1} - 1)!a(k) = \Gamma(2^{k+1})a(k)$, we use Stirling's formula to get

$$h_{2^{k+1}-1}^k = 2\sqrt{\pi}2^{k/2} \left(\frac{2C_0}{e^2} \right)^{2^k} [1 + O(2^{-k})], \quad k \rightarrow \infty. \quad (5.12)$$

Using our result for $a(k)$ in (5.5) and solving the linear recurrence yields

$$b(k) = \prod_{\ell=1}^k \frac{8a(\ell-1)}{2^{2^{\ell+1}}(2^\ell - 1)}. \quad (5.13)$$

Using the (exact) expression for $a(k)$ in (5.10) in (5.13), and noting that

$$\prod_{\ell=1}^k \frac{8 \cdot 2^\ell}{2^{2^{\ell+1}}(2^\ell - 1)} 2^{-(\ell+1/2)2^\ell} 2^{2^\ell} = 8^k 2^{1-(2k+1)2^k} \prod_{\ell=1}^k \left(\frac{1}{1-2^{-\ell}} \right),$$

we obtain

$$b(k) = 2^{-(2k+1)2^k} 2^{3k+1} \prod_{\ell=1}^k \left(\frac{1}{1-2^{-\ell}} \left(\prod_{j=1}^{\ell-1} (1-2^{-j-1})^{-2^{-j}} \right)^{2^{\ell-1}} \right). \quad (5.14)$$

We write the double product in (5.14) as

$$\prod_{\ell=1}^k C_0^{2^{\ell-1}} \frac{1}{1-2^{-\ell}} \left(\prod_{j=\ell}^{\infty} (1-2^{-j-1})^{2^{-j}} \right)^{2^{\ell-1}} \sim C_0^{2^k-1} C_1, \quad k \rightarrow \infty \quad (5.15)$$

where C_0 is given in (5.11) and

$$C_1 = \prod_{\ell=1}^{\infty} \left(\frac{1}{1-2^{-\ell}} \prod_{j=\ell}^{\infty} (1-2^{-j-1})^{2^{\ell-j-1}} \right).$$

In view of (5.15) we have

$$b(k) \sim 2^{-(2k+1)2^k} 2^{3k+1} C_0^{2^k-1} C_1, \quad k \rightarrow \infty \quad (5.16)$$

and then $h_{2^{k+1}-2}^k = (2^{k+1} - 2)!b(k)$ has the expansion

$$h_{2^{k+1}-2}^k \sim \frac{C_1}{C_0} \sqrt{\pi} 2^{3k/2} \left(\frac{2C_0}{e^2} \right)^{2^k}, \quad k \rightarrow \infty. \quad (5.17)$$

The constant C_1 may be simplified by noting that

$$\begin{aligned} C_1 &= \prod_{\ell=1}^{\infty} \frac{1}{1-2^{-\ell}} \left[\prod_{m=1}^{\infty} (1-2^{-m-\ell})^{2^{-m}} \right] \\ &= \left[\prod_{\ell=1}^{\infty} \frac{1}{1-2^{-\ell}} \right] \left(\prod_{N=2}^{\infty} \left[\prod_{m=1}^{N-1} (1-2^{-N})^{2^{-m}} \right] \right) \\ &= \left[\prod_{N=2}^{\infty} (1-2^{-N})^{1-2^{1-N}} \right] \left[\prod_{\ell=1}^{\infty} \frac{1}{1-2^{-\ell}} \right] \\ &= 2 \prod_{N=2}^{\infty} (1-2^{-N})^{-2^{1-N}} \\ &= 2C_0. \end{aligned}$$

Hence, $h_n^k \sim 2\sqrt{\pi}2^{3k/2}(2C_0e^{-2})^{2^k}$ for $k \rightarrow \infty$ if $n = 2^{k+1} - 2$.

Next we solve the recurrence (2.2) for $M = O(1)$ and $n \rightarrow \infty$, thus proving Theorem 1(iv). We change variables from (n, k) to (M, k) with $n = 2^{k+1} - M$ and

$$h_n^k = G(M, k) = G(2^{k+1} - n, k). \quad (5.18)$$

We replace k by $k - 1$ in (2.2) and note that

$$\begin{aligned} h_{n+1}^k &= G(2^{k+1} - n - 1, k) = G(M - 1, k), \\ h_i^{k-1} &= G(2^k - i, k - 1) = G\left(\frac{M}{2} + \frac{n}{2} - i, k - 1\right). \end{aligned}$$

Thus (2.2) becomes, in terms of G and M ,

$$G(M - 1, k) = \sum_{i=0}^n \binom{n}{i} 2^{-n} G\left(\frac{M}{2} + \frac{n}{2} - i, k - 1\right) G\left(\frac{M}{2} - \frac{n}{2} + i, k - 1\right). \quad (5.19)$$

But $G = 0$ for $M \leq 0$ (i.e., $n \geq 2^{k+1}$) so that the sum in (5.19) may be truncated over the range $(n - M)/2 < i < (n + M)/2$. Then setting $i = \ell + (n - M)/2$ we have

$$G(M - 1, k) = \sum_{\ell=1}^{M-1} \binom{n}{\ell + (n - M)/2} 2^{-n} G(\ell, k - 1) G(M - \ell, k - 1). \quad (5.20)$$

The expression in (5.20) is still exact. For $n, k \rightarrow \infty$ with $M = O(1)$, we have

$$\binom{n}{i} 2^{-n} = \sqrt{\frac{2}{\pi n}} [1 + O(n^{-1})] = \frac{2^{-k/2}}{\sqrt{\pi}} [1 + O(2^{-k})]$$

so in this range we replace (5.20) by the asymptotic relation

$$G(M - 1, k) \sim \frac{2^{-k/2}}{\sqrt{\pi}} \sum_{\ell=1}^{M-1} G(\ell, k - 1) G(M - \ell, k - 1). \quad (5.21)$$

Setting $M = 1$, replacing $\sim$ by $=$ and solving (5.21) for $G(1, k)$ yields

$$G(1, k) \sim 2\sqrt{\pi}2^{k/2}A_0^{2^k} \quad (5.22)$$

where A_0 is an undetermined constant. Setting $M = 2$ in (5.21) and using (5.22) leads to

$$\frac{G(2, k)}{G(2, k - 1)} \sim \frac{2}{\sqrt{\pi}} 2^{-k/2} G(1, k - 1) \sim 2\sqrt{2}A_0^{2^{k-1}} \quad (5.23)$$

so that

$$G(2, k) \sim 2\sqrt{\pi}B_02^{3k/2}A_0^{2^k} \quad (5.24)$$

where B_0 is another constant. We have scaled the factor $2\sqrt{\pi}$ out of the constant for convenience. Proceeding inductively we find that for general M

$$G(M, k) \sim 2\sqrt{\pi} B_0^{M-1} 2^{(M-1/2)k} A_0^{2^k} \mathcal{G}(M), \quad k \rightarrow \infty. \quad (5.25)$$

Using (5.25) in (5.21) we find that

$$\mathcal{G}(M-1) = 4 \cdot 2^{-M} \sum_{\ell=1}^{M-1} \mathcal{G}(\ell) \mathcal{G}(M-\ell) \quad (5.26)$$

with $\mathcal{G}(2) = 1$. Thus, $\mathcal{G}(M) = 1/(M-1)!$ which yields

$$G(M, k) \sim \frac{2\sqrt{\pi}}{(M-1)!} B_0^{M-1} 2^{(M-1/2)k} A_0^{2^k}. \quad (5.27)$$

It remains only to determine A_0 and B_0 . But our exact analysis for $M = 1$ and $M = 2$ shows, in view of (5.10) and (5.17) (with $C_1/C_0 = 2$), that

$$A_0 = \frac{2C_0}{e^2}, \quad B_0 = 1. \quad (5.28)$$

We have thus derived the result in (2.22) of Theorem 1(iv). We conclude by noting that the range $M = O(1)$ corresponds to the “left-most” tail of the distribution, and that for $n \rightarrow \infty$ the condition $M = O(1)$ can be satisfied only when n is close to a power of 2.

5.2 The Central Regime Analysis

We next analyze the scale $k, n \rightarrow \infty$ with $k - \log_2 n = O(1)$. We thus set $\xi = n2^{-k}$ and

$$h_n^k = F(\xi; n) = F(n2^{-k}; n). \quad (5.29)$$

We consider the range $0 < \xi < 2$ and note that as $\xi \rightarrow 2^-$ we are approaching the scale $M = O(1)$, which we just analyzed. We have $k - \log_2 n = -\log_2 \xi$ so that $k - \log_2 n \rightarrow \infty$ corresponds to $\xi \rightarrow 0^+$. For any fixed $\xi > 0$ we are still in the left tail as the mass is concentrated where $k - \log_2 n \sim \sqrt{2 \log_2 n}$, which corresponds to $\xi \approx 2^{-\sqrt{2 \log_2 n}}$, which is small.

We comment that the ξ -scale also arises in related models. We have previously shown in [19] that for tries, b -tries and PATRICIA trees, the limit $n, k \rightarrow \infty$ with ξ fixed is important to the asymptotic analysis. We note that

$$h_{n+1}^k = F((n+1)2^{-k}; n+1) = F\left(\xi + \frac{\xi}{n}; n+1\right)$$

and

$$h_i^{k-1} = F(i2^{-k+1}; i) = F\left(\frac{2i}{n}\xi; i\right).$$

Using the above in (2.2), after replacing k by $k - 1$, we obtain

$$F\left(\xi + \frac{\xi}{n}; n+1\right) = \sum_{i=0}^n \binom{n}{i} 2^{-n} F\left(\frac{2i}{n}\xi; i\right) F\left(2\left(1 - \frac{i}{n}\right)\xi; n-i\right). \quad (5.30)$$

We analyze (5.30) by a WKB-type expansion (cf. [10]). That is, we seek an asymptotic solution of (5.30) in the form

$$F(\xi; n) = e^{-n\Phi(\xi)} \left[A(\xi) + \frac{1}{n} A^{(1)}(\xi) + O(n^{-2}) \right]. \quad (5.31)$$

The above may be viewed as a generalized saddle-point expansion. For simpler models, such as tries and b -tries, we can obtain exact expressions for the corresponding distributions. These usually involve Cauchy integrals that can be asymptotically evaluated by the saddle-point method. This then leads to an expansion of the form (5.31) for fixed ξ . Note that for b -tries ξ takes on values in the range $[0, b]$ and for PATRICIA trees the range is $[0, 1]$. For more difficult models, which have not and/or cannot be solved exactly, we must try to obtain the asymptotic expansion directly from the equations, such as (2.2). This generally requires making an *ansatz*, such as (5.31).

Setting $x = i/n$ and $y = \sqrt{n}(x - 1/2) = \sqrt{n}(i/n - 1/2)$ and using Stirling's formula, we have

$$\binom{n}{i} 2^{-n} \sim \frac{e^{nf_0(x)}}{\sqrt{2\pi nx(1-x)}}, \quad 0 < x < 1, \quad (5.32)$$

where $f_0(x) = -\log 2 - x \log x - (1-x) \log(1-x)$. For y fixed (5.32) simplifies to the Gaussian

$$\binom{n}{i} 2^{-n} = \sqrt{\frac{2}{\pi n}} e^{-2y^2} [1 + O(n^{-1})]. \quad (5.33)$$

We use (5.31) in (5.30) and expand for large n , which yields

$$\begin{aligned} e^{-n\Phi(\xi)} \cdot e^{-(\xi\Phi)'(\xi)} [A(\xi) + O(n^{-1})] &= \sum_{i=0}^n \frac{e^{nf_0(i/n)}}{\sqrt{2\pi n}} \sqrt{\frac{n^2}{i(n-i)}} \\ &\times \exp\left(-n\left[\frac{i}{n}\Phi\left(\frac{2i}{n}\xi\right) + \left(1 - \frac{i}{n}\right)\Phi\left(2\left(1 - \frac{i}{n}\right)\xi\right)\right]\right) A\left(\frac{2i}{n}\xi\right) A\left(2\left(1 - \frac{i}{n}\right)\xi\right). \end{aligned}$$

Note that near the endpoints in the sum in (5.30), i and $n-i$ may not be large so that the expansion (5.31) does not apply. However, the major contribution to the sum comes from the range $i = n/2 + O(\sqrt{n})$. We approximate the sum in (5.34) by an integral via

Euler-MacLaurin to get

$$\begin{aligned} e^{-n\Phi(\xi)} &\sim e^{-(\xi\Phi)'(\xi)} [A(\xi) + O(n^{-1})] \\ &\sim \frac{\sqrt{n}}{\sqrt{2\pi}} \int_0^1 \frac{e^{nf_0(x)}}{\sqrt{x(1-x)}} A(2x\xi) A(2(1-x)\xi) e^{-n[x\Phi(2\xi x) + (1-x)\Phi(2\xi(1-x))]} dx. \end{aligned} \quad (5.34)$$

By symmetry, the major contribution must come from the midpoint $x = 1/2$. Setting

$$g(t) = t\Phi(2\xi t) + (1-t)\Phi(2\xi(1-t))$$

we have $g(1/2) = \Phi(\xi)$, $g'(1/2) = 0$ and

$$g''\left(\frac{1}{2}\right) = 8\xi\Phi'(\xi) + 4\xi^2\Phi''(\xi).$$

Setting $x = 1/2 + y/\sqrt{n}$ and expanding the integral by Laplace's method yields, to leading order,

$$e^{-n\Phi(\xi)} e^{-(\xi\Phi)'(\xi)} A(\xi) = \frac{1}{\sqrt{2\pi}} \frac{2\sqrt{2\pi}}{\sqrt{4 + g''(1/2)}} e^{-n\Phi(\xi)} A^2(\xi).$$

Thus the exponential factors cancel and we have

$$A(\xi) = [1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)]^{1/2} e^{-\xi\Phi'(\xi) - \Phi(\xi)}. \quad (5.35)$$

We have thus expressed $A(\cdot)$ in terms of $\Phi(\cdot)$, though the latter remains undetermined. By refining the approximation (5.33) to the “kernel” in (5.30) and obtaining higher order terms in the Laplace expansion of the integral, we can obtain relations between the terms $A^{(j)}$ in the series in (5.31). Using these we can express $A^{(j+1)}$ in terms of $\Phi, A, A^{(1)}, \dots, A^{(j)}$, but we can never determine Φ . For tries and b -tries the corresponding Φ can be determined using a standard saddle-point expansion. For PATRICIA trees we could only study Φ numerically and asymptotically, which we proceed to do for the DST model.

We next use asymptotic matching to determine the behaviors of $\Phi(\xi)$ as $\xi \rightarrow 2^-$ and $\xi \rightarrow 0^+$. We first require that the expansion on the M -scale matches to that on the ξ -scale. This amounts to assuming that both are valid on some intermediate scale where $M = 2^{k+1} - n \rightarrow \infty$ but $\xi = n2^{-k} = 2n/(n+M) \rightarrow 2$. We write this condition symbolically as

$$G(M, k)|_{M \rightarrow \infty} \sim F(\xi; n)|_{\xi \rightarrow 2}. \quad (5.36)$$

To compare the two sides of (5.36) we expand G as $M \rightarrow \infty$, which simply amounts to approximating $(M-1)!$ by Stirling's formula in (5.27). Using (5.28) then yields

$$G(M, k) \sim \sqrt{2M} \left(\frac{2C_0}{e^2}\right)^{n/\xi} e^M \left(\frac{2^k}{M}\right)^M 2^{-k/2}. \quad (5.37)$$

Noting that $n/\xi = 2^k$ and $M = n(2/\xi - 1)$ we rewrite (5.37) in terms of n and ξ . According to (5.36) this should be the expansion of $Ae^{-n\Phi}$ as $\xi \rightarrow 2$. We then find that the matching condition is satisfied provided that

$$\begin{aligned}\Phi(\xi) &= \frac{1}{2} \log \left(\frac{e^2}{2C_0} \right) + \frac{1}{2}(2 - \xi) \log(2 - \xi) \\ &+ (\xi - 2) \left[\frac{1}{4} \log \left(\frac{2C_0}{e^2} \right) + \frac{1}{2} \right] + o(2 - \xi)\end{aligned}\quad (5.38)$$

and for $\xi \rightarrow 2^-$

$$A(\xi) \sim \sqrt{2(2 - \xi)}. \quad (5.39)$$

Thus the matching condition yields the behavior of both Φ and A . We show that (5.38) and (5.39) are consistent with (5.35). From (5.38) we get $\Phi(\xi) \sim \Phi(2) = -1 - \log(\sqrt{2C_0})$, $\Phi'(\xi) \sim -\frac{1}{2} \log(2 - \xi)$ and $\Phi''(\xi) \sim 1/[2(2 - \xi)]$. Using these results in (5.35) yields $A(\xi) \sim \sqrt{4\Phi''(\xi)} e^{-2\Phi'(\xi)} e^{-\Phi(2)} \sim \sqrt{\frac{2}{2 - \xi}} (2 - \xi)$ which recovers (5.39). To obtain the above we also used the second term in the expansion of $\Phi'(\xi)$ as $\xi \rightarrow 2$.

We next match the WKB expansion to the “left-most” right tail, which we derived in Section 4. We let $\xi \rightarrow 0$ with $k/\log_2 n = \nu \rightarrow 1^+$. Furthermore we choose an intermediate limit where $n \rightarrow \infty$, $\xi \rightarrow 0$ so that $n\Phi(\xi) \rightarrow 0$ and we can approximate

$$1 - h_n^k \sim 1 - Ae^{-n\Phi} \sim 1 - e^{-n\Phi} \sim n\Phi.$$

The expansion that applies for fixed $k/\log_2 n = \nu > 1$ is given by $1 - h_n^k \sim 2^{-k^2/2} 2^{-k/2} \tilde{F}_j(n)$ where $\tilde{F}_j$ is given by (4.45). Expanding (4.45) for $k/\log_2 n = \nu \rightarrow 1^+$ amounts to doing nothing, since Q_* is periodic and the rest of (4.45) is in the simplest possible form. We have already observed that $2^{-k^2/2} 2^{-k/2} \times$ (4.45) can be written as $n \times [\text{function of } (k - \log_2 n)] = n \times [\text{function of } \xi]$. Thus the matching condition is satisfied if

$$\begin{aligned}n\Phi(\xi)|_{\xi \rightarrow 0^+} &\sim \frac{2^{-9/8}}{\sqrt{\log 2}} \frac{n}{(-\log_2 \xi)^2} Q_* \left(\log_2 \xi - \log_2(-\log_2 \xi) + \frac{1}{2} \right) \\ &\times \exp \left(-\frac{\log^2 \xi}{2 \log 2} + (\log_2 \xi) \log(-\log_2 \xi) + \left(\frac{1}{2} - \frac{1}{\log 2} \right) \log \xi - \frac{[\log(-\log_2 \xi)]^2}{2 \log 2} \right).\end{aligned}\quad (5.40)$$

The above is equivalent to (2.19) and we have used the periodicity of Q_* .

Expression (5.40) represents the leading term for Φ and we may also view it as representing the first six terms in the expansion of $\log \Phi$. These have respective orders of magnitude $O(\log^2 \xi)$, $O(\log \xi \log |\log \xi|)$, $O(\log \xi)$, $O(\log^2 |\log \xi|)$, $O(\log |\log \xi|)$, and $O(1)$. The $O(1)$ term involves the periodic function Q_* . The analysis in Section 4 suggests that there is an

error term in (5.40) of the form $[1 + O(1/\log \xi)]$. It may thus be desirable to compute higher order terms for Φ , which would involve higher order matchings (cf. Section 3). Finally we note that Φ and all its derivatives vanish as $\xi \rightarrow 0^+$. This completes the analysis of the left tail. To summarize, we have treated the M and ξ scales, determined the leading term completely for the former, and obtained partial results for the latter.

Appendix A

We derive the series representation (2.30) for the periodic function $Q_*(z)$. By using the series representation (2.21) for $\vartheta_1(\cdot)$, comparing it to the infinite product form of $\vartheta_1(\cdot)$ and also using $Q(z) = e^{-\pi iz} \csc(\pi z) e^{\mathcal{P}(z)}$ we obtain

$$Q(z) = \frac{2^{-z^2/2} 2^{-z/2}}{\sin(\pi z)} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{m/2} 2^{mz}. \quad (\text{A.1})$$

This yields a third representation for $Q(z)$, which supplements the two in (4.54). We evaluate

$$\begin{aligned} Q_*(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-s^2/2} Q\left(z + \frac{is}{\sqrt{\log 2}}\right) ds \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} e^{-i\sqrt{\log 2}zy} 2^{z^2/2} Q\left(\frac{iy}{\sqrt{\log 2}}\right) dy \end{aligned} \quad (\text{A.2})$$

where we have shifted the contour by setting $s = i\sqrt{\log 2}z + y$. Using (A.1) in (A.2) and exchanging the orders of integration and summation, we obtain

$$\begin{aligned} Q_*(z) &= \frac{1}{\sqrt{2\pi}} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] 2^{z^2/2} \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{m/2} \\ &\quad \times \int_{-\infty}^{\infty} \Re \left[\frac{e^{-iy\sqrt{\log 2}/2} e^{-i\sqrt{\log 2}zy} 2^{my/\sqrt{\log 2}}}{i \sinh(\pi y/\sqrt{\log 2})} \right] dy, \end{aligned} \quad (\text{A.3})$$

where $\Re$ denotes the real part. From tables of integrals we have

$$\int_{-\infty}^{\infty} \frac{\sin(Ax)}{\sinh(Bx)} dx = \frac{\pi}{B} \tanh\left(\frac{\pi A}{2B}\right). \quad (\text{A.4})$$

Applying (A.4) to (A.3) with $B = \pi/\sqrt{\log 2}$ and $A = (m - z - \frac{1}{2})\sqrt{\log 2}$, we get

$$Q_*(z) = \frac{\sqrt{\log 2}}{\sqrt{2\pi}} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] 2^{z^2/2} \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{m/2} \tanh\left(\frac{\log 2}{2} \left(m - z - \frac{1}{2}\right)\right). \quad (\text{A.5})$$

By changing $m \rightarrow 1 - m$, we see that $Q_+(z) = Q_+(-z)$. We can thus write $Q_+(z) = \frac{1}{2}[Q_+(z) + Q_+(-z)]$. Then we use

$$\tanh \left[\frac{\log 2}{2} \left(m - z - \frac{1}{2} \right) \right] + \tanh \left[\frac{\log 2}{2} \left(m + z - \frac{1}{2} \right) \right] = \frac{2(2^{m-1/2} - 2^{-m+1/2})}{2^{m-1/2} + 2^{-m+1/2} + 2^z + 2^{-z}}$$

with which (A.5) becomes the same as (2.30).

Appendix B

We obtain a series representation for $\Phi_0(\xi)$, defined in (3.5). We use the series form of $Q(z)$ found in (A.1) and exchange the order of summation and integration. This yields

$$\begin{aligned} \Phi_0(\xi) &= \pi \xi \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} 2^{-m^2/2} 2^{m/2} (-1)^{m+1} \\ &\times \frac{1}{2\pi i} \int_{\frac{5}{2}-i\infty}^{\frac{5}{2}+i\infty} \frac{2^{-s} 2^{ms} 2^{\Delta s}}{\Gamma(3-s) \sin(\pi s)} ds, \end{aligned} \quad (\text{B.1})$$

where $\Delta = -\log_2 \xi$. In (3.5) the contour of integration is any vertical contour, but in (B.1) we restrict to $\Re(s) > 2$. Using $\Gamma(3-s) \sin(\pi s) = \pi/\Gamma(s-2)$ (cf. [1]) we see that the integral in (B.1) has simple poles at $s = 2 - N$, $N \geq 0$ and hence

$$\begin{aligned} \frac{1}{2\pi i} \int_{\frac{5}{2}-i\infty}^{\frac{5}{2}+i\infty} \Gamma(s-2) 2^{(m+\Delta-1)s} ds &= \sum_{N=0}^{\infty} \frac{(-1)^N}{N!} 2^{(2-N)(m+\Delta-1)} \\ &= 2^{2m+2\Delta-2} \exp(-2^{-m-\Delta+1}). \end{aligned} \quad (\text{B.2})$$

Using (B.2) in (B.1) and noting that $2^{2\Delta} = \xi^{-2}$ yields

$$\Phi_0(\xi) = \frac{1}{4\xi} \left[\prod_{\ell=1}^{\infty} (1 - 2^{-\ell})^{-2} \right] \sum_{m=-\infty}^{\infty} (-1)^{m+1} 2^{-m^2/2} 2^{5m/2} \exp(-2\xi 2^{-m}) \quad (\text{B.3})$$

which establishes (3.6). The choice $\Re(s) > 2$ was somewhat arbitrary. However, if we choose *any* vertical contour, then the value of the integral in (B.2) will differ from the right side of (B.2) by the residues at a finite number of poles. Each such residue will be proportional to 2^{-Nm} , but this term will vanish after we evaluate the sum over m , in view of (3.7) and the comments below it. Thus, the final result for Φ_0 is independent of the contour chosen.

References

- [1] Abramowitz, M. and Stegun, I., eds. *Handbook of Mathematical Functions*. Wiley, New York, 1962.

- [2] D. Aldous and P. Shields, A Diffusion Limit for a Class of Random-Growing Binary Trees, *Probab. Th. Rel. Fields*, 79, 509-542, 1988.
- [3] G. Andrews, R. Askey and R. Roy, *Special Functions*, Cambridge University Press, 1999.
- [4] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York 1978.
- [5] L. Devroye, A Note on the Probabilistic Analysis of Patricia Tries, *Random Structures & Algorithms*, 3, 203-214, 1992.
- [6] L. Devroye, A Study of Trie-Like Structures Under the Density Model, *Ann. Appl. Probability*, 2, 402-434, 1992.
- [7] P. Flajolet and R. Sedgewick, Digital Search Trees Revisited, *SIAM J. Computing*, 15, 748-767, 1986.
- [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, and P. Dumas, Mellin Transforms and Asymptotics: Harmonic sums, *Theoretical Computer Science*, 144, 3-58, 1995.
- [10] N. Froman and P. Froman, *JWKB Approximation*, North-Holland, Amsterdam 1965.
- [11] E. Gilbert and T. Kadota, The Lempel-Ziv Algorithm and Message Complexity, *IEEE Trans. Information Theory*, 38, 1839-1842, 1992.
- [12] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997.
- [13] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197, 1995.
- [14] P. Jacquet and W. Szpankowski, Analytic Depoissonization and its Applications, *Theoretical Computer Science*, 201, 1-62, 1998.
- [15] P. Jacquet, W. Szpankowski and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, Purdue University, CSD-TR-98-029, 1998.
- [16] P. Kirschenhofer, H. Prodinger and W. Szpankowski, Digital Search Trees Again Revisited: The Internal Path Length Perspective, *SIAM J. Computing*, 23, 598-616, 1994.
- [17] D. Knuth, *The Art of Computer Programming. Sorting and Searching*, Second Edition, Addison-Wesley, 1998.
- [18] C. Knessl and W. Szpankowski, Quicksort Algorithm Again Revisited, *Discrete Mathematics and Theoretical Computer Science*, 3, 43-64, 1999.

- [19] C. Knessl and W. Szpankowski, Limit Laws for Heights in Generalized Tries and PATRICIA Tries, Purdue University CSD-TR-99-011; also available on <http://www.cs.purdue.edu/homes/spa/reports.html>.
- [20] A. Konheim and D.J. Newman, A Note on Growing Binary Trees, *Discrete Mathematics*, 4, 57-63, 1973.
- [21] P. Lagerstrom, *Matched Asymptotic Expansions: Ideas and Techniques*, Springer-Verlag, New York 1988.
- [22] G. Louchard, Exact and Asymptotic Distributions in Digital and Binary Search Trees, *RAIRO Theoretical Inform. Applications*, 21, 479-495, 1987.
- [23] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Trans. on Information Theory*, 41, 478-488, 1995.
- [24] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York 1992.
- [25] A. Odlyzko, Asymptotic Enumeration, in *Handbook of Combinatorics*, Vol. II, (Eds. R. Graham, M. Götschel and L. Lovász), Elsevier Science, 1063-1229, 1995.
- [26] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Ann. of Probab.*, 13, 414-427, 1985.
- [27] A. Rényi, On Random Subsets of a Finite Set, *Mathematica (Cluj)*, 3, 355-362, 1961.
- [28] W. Szpankowski, The Evaluation of an Alternative Sum with Applications to the Analysis of Some Data Structures, *Information Processing Letters*, 28, 13-19, 1988.
- [29] W. Szpankowski, A Characterization of Digital Search Trees From the Successful Search Viewpoint, *Theoretical Computer Science*, 85, 117-134, 1991.
- [30] W. Szpankowski, On the Height of Digital Trees and Related Problems, *Algorithmica*, 6, 256-277, 1991.
- [31] J. Ziv and A. Lempel, A Universal Algorithm for Sequential Data Compression, *IEEE Trans. Information Theory*, 23, 337-343, 1977.
- [32] J. Ziv and A. Lempel, Compression of Individual Sequences via Variable-rate Coding, *IEEE Trans. Information Theory*, 24, 530-536, 1978.