Limit Laws for Heights in Generalized Tries and PATRICIA Tries

April 7, 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 consider digital trees such as (generalized) tries and PATRICIA tries, built from n random strings generated by an unbiased memoryless source (i.e., all symbols are equally likely). We study limit laws of the height which is defined as the longest path in such trees. It turns out that this height also represents the number of random questions required to recognize n distinct objects. We shall identify three natural regions of the height distributions. For tries, in the region where most of the probability mass is concentrated, the asymptotic distribution is of extreme value type (i.e., double exponential distribution). Surprisingly enough, the height of the PATRICIA trie behaves quite differently in this region: It exhibits an exponential of a Gaussian distribution (with an *oscillating term*) around the most probable value $k_1 = \lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \rfloor + 1$. In fact, the asymptotic distribution of PATRICIA height concentrates on one or two points. For most n all the mass is concentrated at k_1, however, there exist subsequences of n such that the mass is on the two points $k_1 - 1$ and k_1, or k_1 and $k_1 + 1$. We derive these results by a combination of analytic methods such as generating functions, Mellin transform, the saddle point method and ideas of applied mathematics such as linearization, asymptotic matching and the WKB method. We present also some numerical verification of our results.

Key Words: digital trees, b-tries, PATRICIA trie, height distribution, Mellin transform, saddle point method, matched asymptotics, WKB method.

*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

Data structures and algorithms on words have experienced a new wave of interest due to a number of novel applications in computer science, communications, and biology. These include 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. To satisfy these diversified demands various data structures were proposed for these algorithms. Undoubtly, the most popular data structures for algorithms on words are digital trees [18, 21] (e.g., tries, PATRICIA tries, digital search trees), and suffix trees [14, 30]. The importance of digital trees stem from their applications in sorting and searching [5, 10, 15, 18, 21, 23, 26, 28, 29, 31], data compression [16, 30, 32, 33] pattern matching [14], the shortest common superstring problem [12], searching for a leader [9], estimating the number of questions necessary to identify many distinct objects [25], prediction [8], and so forth. These problems recently became very important due to the need for efficient storage and transmission of multimedia, and applications to DNA sequencing (cf. [14]).

The most basic digital tree is known as a *trie* (the name comes from re*trie*val). The primary purpose of a trie is to store a set $\mathcal{S}$ of strings (words, keys), say $\mathcal{S} = \{X_1, \ldots, X_n\}$. Each word $X = x_1x_2x_3\ldots$ is a finite or infinite string of symbols taken from a finite alphabet. Throughout the paper, we deal only with the binary alphabet $\{0, 1\}$, but all our results should be extendable to a general finite alphabet. A string will be stored in a leaf of the trie. The trie over $\mathcal{S}$ is built recursively as follows: For $|\mathcal{S}| = 0$, the trie is, of course, empty. For $|\mathcal{S}| = 1$, $trie(\mathcal{S})$ is a single node. If $|\mathcal{S}| > 1$, $\mathcal{S}$ is split into two subsets $\mathcal{S}_0$ and $\mathcal{S}_1$ so that a string is in $\mathcal{S}_j$ if its first symbol is $j \in \{0, 1\}$. The tries $trie(\mathcal{S}_0)$ and $trie(\mathcal{S}_1)$ are constructed in the same way except that at the k-th step, the splitting of sets is based on the k-th symbol of the underlying strings. Figure 1 illustrates such a construction.

There are many possible variations of the trie. One such variation is the b-trie, in which a leaf is allowed to hold as many as b strings (cf. [21, 30]). A second variation of the trie, the *PATRICIA trie*, eliminates the waste of space caused by nodes having only one branch. This is done by collapsing one-way branches into a single node. In a *digital search tree* (in short DST) strings are directly stored in nodes, and hence external nodes are eliminated. The branching policy is the same as in tries. These digital trees are also shown in Figure 1. The reader is referred to [14, 18, 21] for a detailed description of digital trees.

In this paper, we consider tries and PATRICIA tries built over n randomly generated strings of binary symbols. We assume that every symbol is equally likely, thus we are within

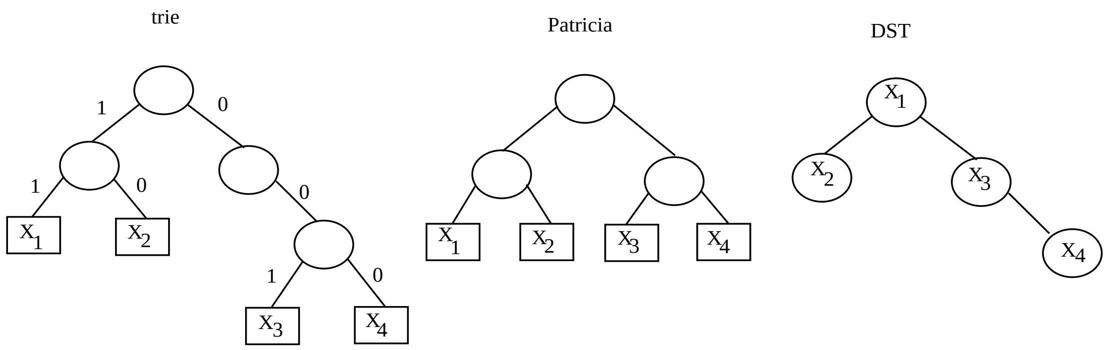

Figure 1: A trie, PATRICIA trie and a digital search tree (DST) built from the following four strings $X_1 = 11100\ldots$, $X_2 = 10111\ldots$, $X_3 = 00110\ldots$, and $X_4 = 00001\ldots$.

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 asymptotic distributions of the heights for random tries and PATRICIA tries. The height is the longest path in such trees, and its distribution is of considerable interest for several applications.

Let us further motivate our study by describing a generalization of Rényi's problem [27] analyzed by Pittel and Rubin [25]. The problem is to identify n distinct objects by asking random questions. More formally, let $\Phi : \mathcal{X} \to \mathcal{A}$ be a bijective mapping from set $\mathcal{X}$ into set $\mathcal{A}$, both sets having the same cardinality n. The goal is to identify all elements of $\mathcal{X}$ by asking questions like: for a given $\mathcal{B} \in \mathcal{A}$ what is $\Phi^{-1}(\mathcal{B})$. If $\mathcal{B}$ is chosen uniformly at random, how many questions must one ask to recognize $\mathcal{X}$? What happens if we do not allow $\mathcal{B}$ to be empty, or to be the same as in the previous attempt? Pittel and Rubin [25] proved that the former problem is equivalent to estimating the height in a trie, while the latter can be reduced to the height in a PATRICIA trie.

We now summarize our main results. As mentioned before, we are aiming at establishing asymptotic distributions for the b-trie height $\mathcal{H}_n^T$ and the PATRICIA height $\mathcal{H}_n^P$. We obtain asymptotic expansions of the distributions $\Pr\{\mathcal{H}_n^T \le k\}$ and $\Pr\{\mathcal{H}_n^P \le k\}$ for three ranges of n and k. For b-tries we consider: (i) the "right–tail region" $k \to \infty$ and $n = O(1)$; (ii) the "central region" $n, k \to \infty$ with $\xi = n2^{-k}$ and $0 < \xi < b$; and (iii) the "left–tail region" $k, n \to \infty$ with $n - b2^k = O(1)$. We prove that most probability mass is concentrated in between the right tail and the central region. In particular, for real x

$$\Pr\left\{\mathcal{H}_n^T \le \frac{1+b}{b}\log_2 n + x\right\} \sim \exp\left(-\frac{1}{(b+1)!}2^{-bx+b\langle\frac{1+b}{b}\log_2 n + x\rangle}\right),$$

where $\langle r\rangle = r - \lfloor r\rfloor$ is the fractional part of r.[1] In words, the asymptotic distribution

[1]The fractional part $\langle r\rangle$ is often denoted as $\{r\}$, but in order to avoid confusion we adopt the above

of b-tries height around its most likely value $\frac{1+b}{b}\log_2 n$ resembles a double exponential (extreme value) distribution. In fact, due to the oscillating term $\langle \frac{1+b}{b}\log_2 n+x\rangle$ the limiting distribution does *not* exist, but one can find $\liminf$ and $\limsup$ of the distribution (cf. Corollary 1 and Figure 2 in the next section).

The height of PATRICIA tries behaves differently in the central region (i.e., where most of the probability mass is concentrated). It is concentrated at or near the most likely value $k_1 = \lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \rfloor + 1$. We shall prove that the asymptotic distribution around k_1 resembles an exponential of a Gaussian distribution, with an oscillating term (cf. Theorem 3). In fact, there exist subsequences of n such that the asymptotic distribution of PATRICIA height concentrates only on k_1, or on k_1 and one of the two points $k_1 - 1$ or $k_1 + 1$. Later, we characterize precisely these subsequences.

With respect to previous results, Devroye [5] and Pittel [24] established the asymptotic distribution in the central regime for tries and b-tries, respectively, using probabilistic tools. Jacquet and Régnier [15] obtained similar results by analytic methods. The most probable value, $\log_2 n$, of the height for PATRICIA was first proved by Pittel [23]. This was then improved to $\log_2 n + \sqrt{2\log_2 n}(1+o(1))$ by Pittel and Rubin [25], and independently by Devroye [6]. No results concerning the asymptotic distribution for PATRICIA height were reported.

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 two non-linear recurrence equations. The distribution $\bar{h}_n^k = \mathsf{Pr}\{\mathcal{H}_n^T \le k\}$ of the height of b-tries satisfies

$$\bar{h}_n^{k+1} = 2^{-n}\sum_{i=0}^{n}\binom{n}{i}\bar{h}_i^k\bar{h}_{n-i}^k, \qquad k \ge 0$$

with the initial condition $\bar{h}_n^0 = 1$ for $n = 0, 1, 2, \ldots, b$ and $\bar{h}_n^0 = 0$ for $n > b$. For PATRICIA tries the distribution $h_n^k = \mathsf{Pr}\{\mathcal{H}_n^P \le k\}$ satisfies

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

with the initial conditions $h_0^0 = h_1^0 = 1$ and $h_n^0 = 0$ for $n \ge 2$. More importantly, the PATRICIA recurrence satisfies an additional boundary condition, namely, $h_n^k = 1$ for $k \ge n-1$, which does not hold for tries. This boundary condition seems to have an enormous impact on the final solution, as shown by our results (cf. Theorem 2 and Theorem 3).

We use two different methodologies to solve these recurrences. The trie recurrence is first analyzed by analytic methods (i.e., generating functions, Mellin transform, saddle

notation.

point method). We then re-derive the results by methods of applied mathematics such as matched asymptotics and the WKB method. These are also analytic methods and are especially suitable for problems that cannot be solved exactly by transform methods, such as the PATRICIA model. The approach we suggest works also for other problems (e.g., height of digital search trees and quicksort [19]).

The paper is organized as follows. In the next section, we present and discuss our main results for tries (cf. Theorem 1), b-tries (cf. Theorem 2), and PATRICIA tries (cf. Theorem 3). In Section 3 we derive the results for tries and b-tries, while in Section 4 we deal with the PATRICIA trie. Finally, the last section presents some numerical results.

2 Summary of Results

As before, we let $\mathcal{H}_n^T$ and $\mathcal{H}_n^P$ denote, respectively, the height of a b-trie and a PATRICIA trie. Their probability distributions are

$$\bar{h}_n^k = \Pr\{\mathcal{H}_n^T \leq k\} \qquad \text{and} \qquad h_n^k = \Pr\{\mathcal{H}_n^P \leq k\}. \tag{2.1}$$

We note that for tries $\bar{h}_n^k = 0$ for $n > b2^k$ (corresponding to a balanced tree), while for PATRICIA tries $h_n^k = 0$ for $n > 2^k$. In addition, for PATRICIA we have the following boundary condition: $h_n^k = 1$ for $k \geq n$. It asserts that the height in a PATRICIA trie cannot be bigger than n (due to the elimination of all one-way branches).

The distribution of b-tries satisfies the recurrence relation

$$\bar{h}_n^{k+1} = 2^{-n} \sum_{i=0}^{n} \binom{n}{i} \bar{h}_i^k h_{n-i}^k, \qquad k \geq 0 \tag{2.2}$$

with the initial condition(s)

$$\bar{h}_n^0 = 1, \ n = 0, 1, 2, \ldots, b; \qquad \text{and} \qquad \bar{h}_n^0 = 0, \ n > b. \tag{2.3}$$

This follows from $\mathcal{H}_n^T = \max\{\mathcal{H}_i^{LT}, \mathcal{H}_{n-i}^{RT}\} + 1$, 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 $2^{-n}\binom{n}{i}$. Similarly, for PATRICIA tries we have

$$h_n^{k+1} = 2^{-n+1} h_n^{k+1} + 2^{-n} \sum_{i=1}^{n-1} \binom{n}{i} h_i^k h_{n-i}^k, \quad k \geq 0 \tag{2.4}$$

with the initial conditions

$$h_0^0 = h_1^0 = 1 \qquad \text{and} \qquad h_n^0 = 0, \quad n \geq 2. \tag{2.5}$$

Unlike b-tries, in a PATRICIA trie the left and the right subtrees cannot be empty (which occurs with probability 2^{-n+1}).

We shall analyze these problems asymptotically, in the limit $n \to \infty$. Despite the similarity between (2.2) and (2.4), we will show that even asymptotically the two distributions behave very differently.

We first consider ordinary tries (i.e., 1-tries). It is relatively easy to solve (2.2) and (2.3) explicitly and obtain the integral representation (cf. Section 3)

$$\begin{aligned} \bar{h}_n^k &= \frac{n!}{2\pi i} \oint (1+z2^{-k})^{2^k} z^{-n-1} dz \\ &= \begin{cases} 0, & n > 2^k \\ \frac{(2^k)!}{2^{nk}(2^k-n)!}, & 0 \le n \le 2^k. \end{cases} \end{aligned} \tag{2.6}$$

Here the loop integral is for any closed circle surrounding $z = 0$.

Using asymptotic methods for evaluating integrals, or applying Stirling's formula to the second part of (2.6), we obtain the following.

Theorem 1 *The distribution of the height of tries has the following asymptotic expansions:*

(i) RIGHT-TAIL REGION*:* $k \to \infty$, $n = O(1)$

$$\Pr\{\mathcal{H}_n^T \le k\} = \bar{h}_n^k = 1 - n(n-1)2^{-k-1} + O(2^{-2k}).$$

(ii) CENTRAL REGION*:* $k, n \to \infty$ *with* $\xi = n2^{-k}$, $0 < \xi < 1$

$$\bar{h}_n^k \sim A(\xi) e^{n\phi(\xi)},$$

where

$$\begin{aligned} \phi(\xi) &= \left(1 - \frac{1}{\xi}\right) \log(1-\xi) - 1, \\ A(\xi) &= (1-\xi)^{-1/2}. \end{aligned}$$

(iii) LEFT-TAIL REGION*:* $k, n \to \infty$ *with* $2^k - n = j = O(1)$

$$\bar{h}_n^k \sim n^j \frac{e^{-n-j}}{j!} \sqrt{2\pi n}.$$

This shows that there are three ranges of k and n where the asymptotic form of $\bar{h}_n^k$ is different.

We next consider the "asymptotic matching" (see [20]) between the three expansions. If we expand (i) for n large, we obtain $1-\bar{h}_n^k \sim n^2 2^{-k-1}$. For $\xi \to 0$ we have $A(\xi) \sim 1$ and $\phi(\xi) \sim -\xi/2$ so that the result in (ii) becomes

$$A(\xi)e^{n\phi(\xi)} \sim e^{-n\xi/2} = \exp\left(-\frac{1}{2}n^2 2^{-k}\right) \sim 1 - \frac{1}{2}n^2 2^{-k} \tag{2.7}$$

where the last approximation assumes that $n, k \to \infty$ in such a way that $n^2 2^{-k} \to 0$. Since (2.7) agrees precisely with the expansion of (i) as $n \to \infty$, we say that (i) and (ii) asymptotically match. To be precise, we say they match the leading order; higher order matchings can be verified by computing higher order terms in the asymptotic series in (i) and (ii). We can easily show that the expansion of (ii) as $\xi \to 1^-$ agrees with the expansion of (iii) as $j \to \infty$, so that (ii) and (iii) also asymptotically match. The matching verifications imply that, at least to leading order, there are no "gaps" in the asymptotics. In other words, one of the results in (i)-(iii) applies for any asymptotic limit which has k and/or n large. We recall that $\bar{h}_n^k = 0$ for $n > 2^k$ so we need only consider $k \geq \log_2 n$.

The asymptotic limits where (i)-(iii) apply are the three "natural scales" for this problem. We can certainly consider other limits (such as $k, n \to \infty$ with k/n fixed), but the expansions that apply in these limits would necessarily be limiting cases of one of the three results in Theorem 1. In particular, if we let $k, n \to \infty$ with $k - 2\log_2 n = O(1)$, we are led to

$$\bar{h}_n^k \sim \exp\left(-\frac{1}{2}n^2 2^{-k}\right) = \exp\left(-\frac{1}{2}\exp(-k\log 2 + 2\log n)\right). \tag{2.8}$$

This result is well-known (see [5, 15]) and corresponds to a limiting double exponential (or extreme value) distribution. However, according to our discussion, $k = 2\log_2 n + O(1)$ is *not* a natural scale for this problem. The scale $k = \log_2 n + O(1)$ (where (ii) applies) is a natural scale, and the result in (2.8) may be obtained as a limiting case of (ii), by expanding (ii) for $\xi \to 0$.

To obtain Theorem 1, we have identified the natural scales from the representations of $\bar{h}_n^k$ in (2.6), and made no use of asymptotic matching. However, for problems such as (2.4) (with (2.5)), which have not been solved exactly, we shall show that asymptotic matching is a very useful tool for analyzing the model as $k, n \to \infty$. For such problems we must identify the natural scales using only the recursion equation, such as (2.4). We discuss this in more depth in Sections 3 and 4.

We next generalize Theorem 1 to arbitrary b, and obtain the following. The proof can be found in Section 3.

Theorem 2 *The distribution of the height of b-tries has the following asymptotic expansions for fixed b:*

(i) RIGHT-TAIL REGION*: $k \to \infty$, $n = O(1)$:*

$$\mathsf{Pr}\{\mathcal{H}_n^T \le k\} = \bar{h}_n^k \sim 1 - \frac{n!}{(b+1)!(n-b-1)!}2^{-kb}.$$

(ii) CENTRAL REGIME*: $k, n \to \infty$ with $\xi = n2^{-k}$, $0 < \xi < b$:*

$$\bar{h}_n^k \sim A(\xi; b)e^{n\phi(\xi;b)},$$

where

$$\begin{aligned}
\phi(\xi; b) &= -1 - \log \omega_0 + \frac{1}{\xi}\left(b \log(\omega_0 \xi) - \log b! - \log\left(1 - \frac{1}{\omega_0}\right)\right), \\
A(\xi; b) &= \frac{1}{\sqrt{1 + (\omega_0 - 1)(\xi - b)}}.
\end{aligned}$$

In the above, $\omega_0 = \omega_0(\xi; b)$ is the solution to

$$1 - \frac{1}{\omega_0} = \frac{(\omega_0 \xi)^b}{b!\left(1 + \omega_0 \xi + \frac{\omega_0^2 \xi^2}{2!} + \cdots + \frac{\omega_0^b \xi^b}{b!}\right)}.$$

(iii) LEFT-TAIL REGION*: $k, n \to \infty$ with $j = b2^k - n$*

$$\bar{h}_n^k \sim \sqrt{2\pi n}\frac{n^j}{j!}b^n \exp\left(-(n+j)\left(1 + b^{-1}\log b!\right)\right)$$

where $j = O(1)$.

We note that when $b = 1$ or 2, we can obtain ω_0 explicitly:

$$\omega_0(\xi; 1) = \frac{1}{1 - \xi}, \tag{2.9}$$

$$\omega_0(\xi; 2) = \frac{2}{1 - \xi + \sqrt{1 + 2\xi - \xi^2}}. \tag{2.10}$$

For arbitrary b, we have $\omega_0 \to \infty$ as $\xi \to b^-$ and $\omega_0 \to 1$ as $\xi \to 0^+$. More precisely,

$$\omega_0 = 1 - \frac{\xi^b}{b!} + O(\xi^{b+1}), \qquad \xi \to 0 \tag{2.11}$$

$$\omega_0 = \frac{1}{b - \xi} + \frac{b-1}{b} + O(b - \xi), \quad \xi \to b. \tag{2.12}$$

When $b = 1$ we can easily show (using (2.9)) that Theorem 2 reduces to Theorem 1.

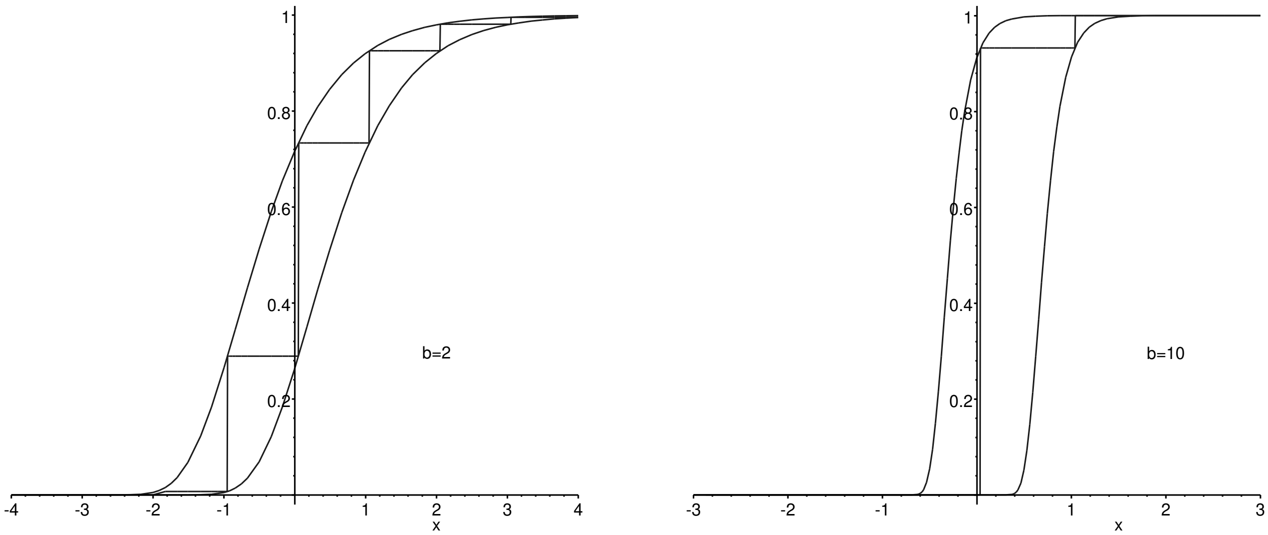

Figure 2: Asymptotic distributions for the height and their corresponding lower and upper bound for tries with $b = 2$ (left) and $b = 10$ (right).

Using (2.11) we can also show that the three parts of Theorem 2 asymptotically match. In particular, by expanding part (ii) as $\xi \to 0$ we obtain

$$\begin{aligned} \Pr\{\mathcal{H}_n^T \le k\} &\sim A(\xi)e^{n\phi(\xi)} \sim \exp\left(-\frac{n\xi^b}{(b+1)!}\right) \qquad \xi \to 0 \\ &= \exp\left(-\frac{n^{1+b}2^{-kb}}{(b+1)!}\right). \end{aligned} \tag{2.13}$$

This yields the well-known (see [15, 24]) asymptotic distribution of b-tries. We note that, for $k, n \to \infty$, (2.13) is O(1) for $k - (1 + 1/b)\log_2 n = O(1)$. More precisely, let us estimate the probability mass of $\mathcal{H}_n^T$ around $(1 + 1/b)\log_2 n + x$ where x is a real fixed value. We observe from (2.13) that

$$\begin{aligned} \Pr\{\mathcal{H}_n^T \le (1+1/b)\log_2 n + x\} &= \Pr\{\mathcal{H}_n^T \le \lfloor(1+1/b)\log_2 n + x\rfloor\} \\ &\sim \exp\left(-\frac{1}{(1+b)!}2^{-bx+b\langle(1+b)/b\cdot\log_2 n+x\rangle}\right), \end{aligned} \tag{2.14}$$

where $\langle x\rangle$ is the fractional part of x, that is, $\langle x\rangle = x - \lfloor x\rfloor$. Since $\langle\log_2 n\rangle$ is dense in $[0, 1]$ but not uniformly dense, the limit of (2.14) does not exit as $n \to \infty$. We can, however, conclude the following.

Corollary 1 *While the limiting distribution of the height for b-tries does not exist, the following lower and upper envelopes can be established*

$$\liminf_{n\to\infty} \Pr\{\mathcal{H}_n^T \le (1+1/b)\log_2 n + x\} = \exp\left(-\frac{1}{(1+b)!}2^{-b(x-1)}\right),$$

$$\limsup_{n\to\infty} \Pr\{\mathcal{H}_n^T \le (1+1/b)\log_2 n + x\} = \exp\left(-\frac{1}{(1+b)!}2^{-bx}\right)$$

for fixed real x.

In Figure 2 we plot the asymptotic distribution (2.14) (the stair-wise function) together with the lower and the upper envelopes for $b = 2$ and $b = 10$. We observe that for large b the asymptotic distribution becomes more concentrated on one or two points.

We next turn our attention to PATRICIA tries. Using ideas of applied mathematics, such as linearization and asymptotic matching, we obtain the following. The derivation of this result is presented in Section 4 where we make certain assumptions about the forms of the asymptotic expansions, as well as the asymptotic matching between the various scales.

Theorem 3 *The distribution of PATRICIA tries has the following asymptotic expansions:*
(i) RIGHT-TAIL REGIME*: $k, n \to \infty$ with $n - k = j = O(1)$, $j \ge 2$*

$$\Pr\{\mathcal{H}_n^P \le n-j\} = h_n^{n-j} \sim 1 - \rho_0 K_j \cdot n! \cdot 2^{-n^2/2+(j-3/2)n},$$

where

$$K_j = \frac{1}{j!}2^{-j^2/2}2^{3j/2}C_j, \tag{2.15}$$

$$C_j = \frac{j!}{2\pi i}\oint \frac{z^{1-j}e^z}{2}\prod_{m=0}^{\infty}\left(\frac{1-\exp(-z2^{-m-1})}{z2^{-m-1}}\right)dz \tag{2.16}$$

and $\rho_0 = \prod_{\ell=2}^{\infty}(1-2^{-\ell})^{-1} = 1.73137\ldots$

(ii) CENTRAL REGIME*: $k, n \to \infty$ with $\xi = n2^{-k}$, $0 < \xi < 1$*

$$h_n^k \sim \sqrt{1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)}\,e^{-n\Phi(\xi)}.$$

We know $\Phi(\xi)$ analytically only for $\xi \to 0$ and $\xi \to 1$. In particular, for $\xi \to 0$

$$\Phi(\xi) \sim \frac{1}{2}\rho_0 e^{\varphi(\log_2 \xi)}\xi^{3/2}\exp\left(-\frac{\log^2 \xi}{2\log 2}\right), \qquad \xi \to 0^+, \tag{2.17}$$

with

$$\begin{aligned}\varphi(x) &= \frac{\log 2}{2}x(x+1) + \sum_{\ell=0}^{\infty}\log\left(\frac{1-\exp(-2^{x-\ell})}{2^{x-\ell}}\right) + \sum_{\ell=1}^{\infty}\log(1-\exp(-2^{x+\ell}))\\ &= \Psi(x) - \frac{\log 2}{12} + \frac{1}{\log 2}\left(\frac{\gamma^2}{2} + \gamma(1) - \frac{\pi^2}{12}\right),\end{aligned} \tag{2.18}$$

$$\Psi(x) = \sum_{\substack{\ell=-\infty\\ \ell\neq 0}}^{\infty} \frac{1}{2\pi i\ell}\Gamma\left(1-\frac{2\pi i\ell}{\log 2}\right)\zeta\left(1-\frac{2\pi i\ell}{\log 2}\right)e^{2\pi i x\ell}. \tag{2.19}$$

In the above, $\Gamma(\cdot)$ *is the Gamma function,* $\zeta(\cdot)$ *is the Riemann zeta function,* $\gamma = -\Gamma'(1)$ *is the Euler constant, and* $\gamma(1)$ *is defined by the Laurent series* $\zeta(s) = 1/(s-1) + \gamma - \gamma(1)(s-1) + O((s-1)^2)$*. The function* $\Psi(x)$ *is periodic with a very small amplitude, i.e.,* $|\Psi(x)| < 10^{-5}$*. Moreover, for* $\xi \to 1$ *the function* $\Phi(\xi)$ *becomes*

$$\Phi(\xi) \sim D_1 + (1-\xi)\log(1-\xi) - (1-\xi)(1+\log D_2), \qquad \xi \to 1^-$$

where $D_1 = 1 + \log(K_0^*)$ *and* $D_2 = K_1^* K_0^*/e$ *with*

$$\begin{aligned} K_0^* &= \prod_{\ell=1}^{\infty} \left(1 - 2^{-2^{\ell}+1}\right)^{2^{-\ell}} = .68321974\ldots, \\ K_1^* &= \prod_{\ell=1}^{\infty} \prod_{m=1}^{\infty} \left(1 - 2^{-2^{\ell+1}+2}\right)^{-1} \left[1 - 2^{-2^{\ell+m}+1}\right]^{2^{-m}} = 1.2596283\ldots \end{aligned}$$

(iii) Left-Tail Regime*:* $k, n \to \infty$ *with* $2^k - n = M = O(1)$

$$h_n^k \sim \frac{\sqrt{2\pi}}{M!} D_2^M n^{M+1/2} e^{-D_1 n}$$

where D_1 *and* D_2 *are defined above.*

The expressions for h_n^k in parts (i) and (iii) are completely determined. However, the expression in part (ii) involves the function $\Phi(\xi)$. We have not been able to determine this function analytically, except for its behaviors as ξ approaches 0 or 1. In Section 5, we discuss the numerical computation of $\Phi(\cdot)$ and sketch this function in Figure 5. The behavior of $\Phi(\xi)$ as $\xi \to 1^-$ implies the asymptotic matching of parts (ii) and (iii), while the behavior as $\xi \to 0^+$ implies the matching of (i) and (ii). As $\xi \to 0$, this behavior involves the periodic function $\varphi(x)$, which satisfies $\varphi(x+1) = \varphi(x)$. In part (ii) we give two different representations for $\varphi(x)$; the latter (which involves $\Psi(x)$) is a Fourier series. In Appendix A we show the equivalence of these two representations.

Since $\Phi(\xi) > 0$, we see that in (ii) and (iii), the distribution is exponentially small in n, while in (i), $1 - h_n^k$ is super–exponentially small (the dominant term in $1 - h_n^k$ is $2^{-n^2/2}$). Thus, (i) applies in the right tail of the distribution while (ii) and (iii) apply in the left tail. We wish to compute the range of k where h_n^k undergoes the transition from $h_n^k \approx 0$ to $h_n^k \approx 1$, as $n \to \infty$. This must be in the asymptotic matching region between (i) and (ii). We shall show in Section 4 that C_j, defined in Theorem 3(i), becomes as $j \to \infty$

$$C_j \sim \frac{j^{5/2}}{2} e^{\varphi(\alpha)} \exp\left(-\frac{1}{2}\frac{\log^2 j}{\log 2}\right), \tag{2.20}$$

where $\alpha = \langle \log_2 j \rangle$ is the fractional part. With (2.20), we can verify the matching between parts (i) and (ii), and the limiting form of (ii) as $\xi \to 0^+$ is

$$\begin{aligned} h_n^k &\sim \exp\left(-\frac{\rho_0}{2} e^{\varphi(\log_2 n)} \exp\left(-\frac{\log 2}{2}\left(\left(k+\frac{3}{2}-\log_2 n\right)^2 - 2\log_2 n - \frac{9}{4}\right)\right)\right) \\ &= \exp\left(-\rho_0 e^{\varphi(\log_2 n)} 2^{1/8} n \exp\left(-\frac{\log 2}{2}\left(k+1.5-\log_2 n\right)^2\right)\right) \qquad (2.21) \\ &= \exp\left(-\rho_0 \cdot n \cdot \exp\left(-\frac{\log 2}{2}\left(k+1.5-\log_2 n\right)^2 + \theta + \Psi(\log_2 n)\right)\right) \qquad (2.22) \end{aligned}$$

where ρ_0 is defined in Theorem 3(i) and

$$\theta = \frac{1}{\log 2}\left(\frac{\gamma^2}{2} + \gamma(1) - \frac{\pi^2}{12}\right) + \frac{\log 2}{24} = -1.022401\ldots$$

while $|\Psi(\log_2 n)| \le 10^{-5}$. We have written (2.21) in terms of k and n, recalling that $\xi = n2^{-k}$. We also have used $\sqrt{1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)} \sim 1$ as $\xi \to 0$.

We now set, for an integer ℓ,

$$\begin{aligned} k_\ell &= \left\lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \right\rfloor + \ell \qquad (2.23) \\ &= \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} + \ell - \beta_n, \end{aligned}$$

where

$$\beta_n = \left\langle \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \right\rangle \in [0,1). \qquad (2.24)$$

In terms of ℓ and β_n, (2.22) becomes

$$\Pr\{\mathcal{H}_n^P \le \lfloor \log_2 n + \sqrt{2\log_2 n} - 1.5 \rfloor + \ell\} \sim \exp\left(-\rho_0 e^{\theta + \Psi(\log_2 n)} 2^{-(\ell-\beta_n)^2/2 - (\ell-\beta_n)\sqrt{2\log_2 n}}\right). \qquad (2.25)$$

For $0 < \beta_n < 1$ and $n \to \infty$ the above is small for $\ell \le 0$, and it is close to one for $\ell \ge 1$. This shows that asymptotically, as $n \to \infty$, all the mass accumulates when $k = k_1$ is given by (2.23) with $\ell = 1$. Now suppose $\beta_n = 0$ for some n, or more generally that we can find a sequence n_i such that $n_i \to \infty$ as $i \to \infty$ but $\sqrt{2\log_2 n_i}\left\langle \log_2 n_i + \sqrt{2\log_2 n_i} - \frac{3}{2} \right\rangle$ remains bounded. Then, the expression in (2.25) would be O(1) for $\ell = 0$ (since $\beta_n\sqrt{2\log_2 n} = O(1)$). For $\ell = 1$, (2.25) would then be asymptotically close to 1. Thus, now the mass would accumulate at two points, namely, $k_0 = k_1 - 1$ and k_1. Finally, if $\beta_n = 1 - o(1)$ such that $(1-\beta_n)\sqrt{2\log_2 n} = O(1)$, then the probability mass is concentrated on k_1 and $k_1 + 1$.

In order to verify the latter assertions, we must either show that $\beta_n = 0$ for an integer n or that there is a subsequence n_i such that $\sqrt{2\log_2 n_i}\beta_{n_i} = O(1)$. The former is false,

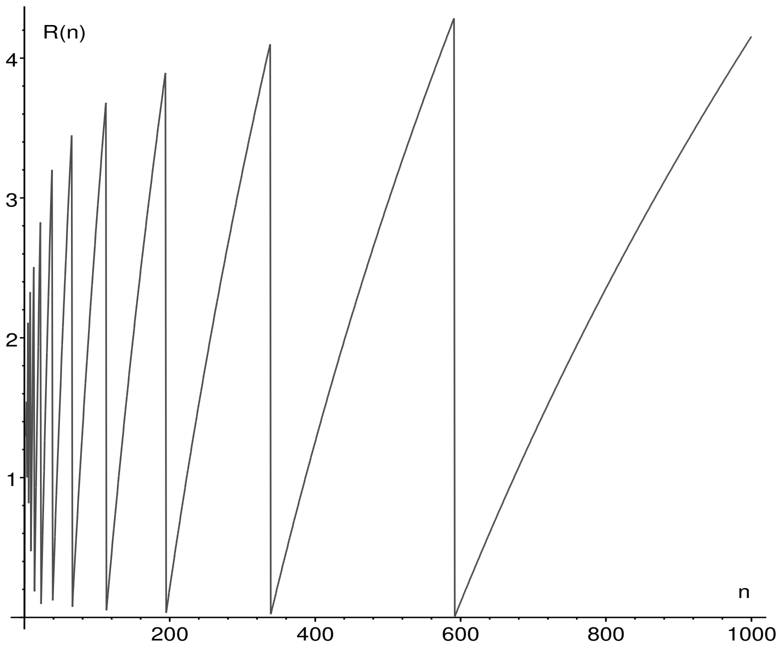

Figure 3: The function $R(n) = \sqrt{2\log_2 n}\langle \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2}\rangle$ versus n.

while the latter is true. To prove that $\beta_n = 0$ is impossible for integer n, let us assume the contrary. If there exists an integer N such that

$$\log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} = N,$$

then

$$n = 2^{N+5/2-\sqrt{4+2N}}.$$

But this is impossible since this would require that $4+2N$ is odd.

To see that there exists a subsequence such that $R(n_i) = \beta_{n_i}\sqrt{2\log_2 n_i} = O(1)$, we first refer to Figure 3 which indicates that the function $R(n)$ fluctuates from zero to $\sqrt{2\log_2 n}$. In Section 5, we show that if $n_i = \lfloor 2^{i+5/2-\sqrt{2i+4}} \rfloor + 1$, then $R(n_i) \to 0$ as $i \to \infty$. Note that this subsequence corresponds to the minima of $R(n)$ in Figure 3.

We summarize our findings in the following corollary.

Corollary 2 *The asymptotic distribution of PATRICIA height is concentrated among the three points $k_1 - 1$, k_1 and $k_1 + 1$ where $k_1 = \lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} \rfloor + 1$, that is,*

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

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

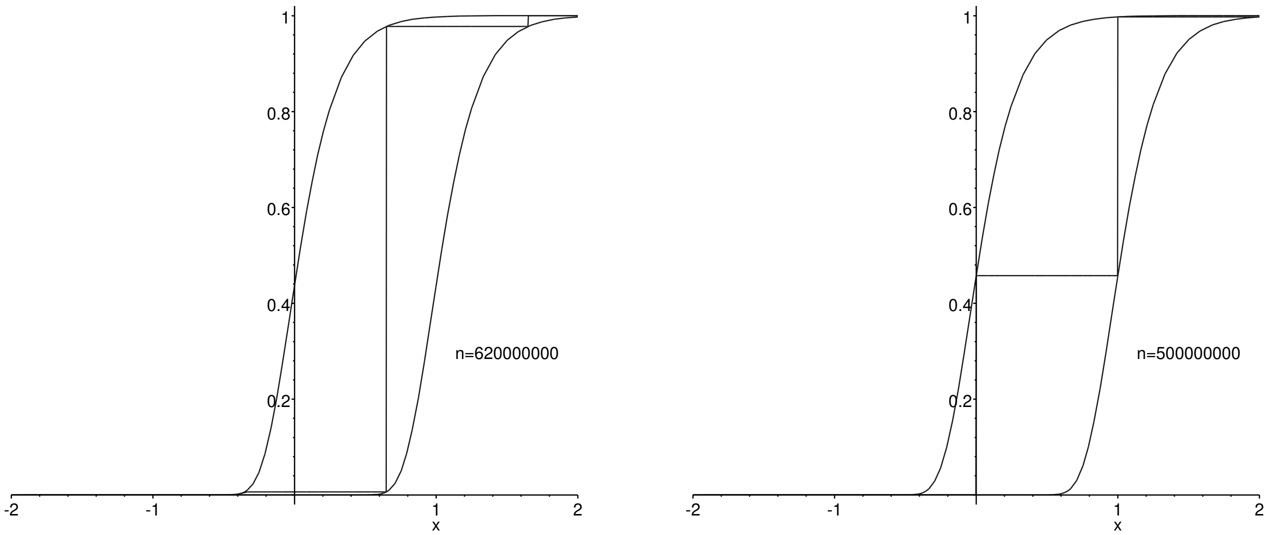

Figure 4: Asymptotic distributions for the height and their corresponding lower and upper bound for PATRICIA tries with $n = 6.2 \cdot 10^8$ (one-point distribution) and $n = 5 \cdot 10^8$ (two-points distribution).

provided that

$$R(n) = \sqrt{2\log_2 n}\left\langle \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2}\right\rangle \to \infty$$

as $i \to \infty$*; (ii) there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i}^P = k_1 - 1 \text{ or } k_1\} = 1 - o(1)$ *provided that* $R(n_i) = O(1)$*; (iii) finally, there are subsequences* n_i *for which* $\Pr\{\mathcal{H}_{n_i}^P = k_1 \text{ or } k_1 + 1\} = 1 - o(1)$ *provided that* $\sqrt{2\log_2 n_i} - R(n_i) = O(1)$.

As in the case of b-tries, we can study the asymptotic distribution of $\Pr\{\mathcal{H}_n^P \le \log_2 n + \sqrt{2\log_2 n} - 1.5 + x\}$ for x real. We plot the asymptotic distribution of $\mathcal{H}_n^P$ in Figure 4 and its lower and upper envelopes. The figure illustrates the concentration at one point or two points (cf. also Table 4 and Table 5). We also observe that the asymptotic distribution for the PATRICIA height qualitatively resembles the asymptotic distribution for b-tries with large b (cf. Figure 2 and Figure 4).

To summarize the PATRICIA results, we have obtained an explicit formula (i.e., (2.21) or (2.25)) that can be used to compute the distribution h_n^k where there is appreciable mass, while Theorem 3 also describes in detail the left and right tails of the distribution. In Section 5 we give detailed numerical comparisons between our asymptotic results and the exact values of h_n^k, obtained by numerically iterating the recursion (2.4).

3 Asymptotics of Tries and b-Tries

We analyze (2.2) and (2.3) for k and/or $n \to \infty$ by two independent approaches. First, we solve the recurrence exactly and then evaluate the result asymptotically. Second, we try to obtain asymptotic information using only the recurrence (2.2). The latter method will form the basis for analyzing the PATRICIA model.

3.1 Transform Method

To solve (2.2), let us define $H^k(z) = \sum_{n\geq 0} \bar{h}_n^k \frac{z^n}{n!}$. Then, (2.2) implies that

$$H^k(z) = \left(H^0(z2^{-k})\right)^{2^k}$$

with $H^0(z) = 1 + z + \cdots + z^b/b!$. By Cauchy's formula, we obtain

$$\bar{h}_n^k = \frac{n!}{2\pi i} \oint \left(1 + z2^{-k} + \frac{z^2}{2!}2^{-2k} + \cdots + \frac{z^b}{b!}2^{-bk}\right)^{2^k} z^{-n-1} dz \tag{3.1}$$

where the contour integral is a loop around the origin in the z-plane. When $b = 1$, (3.1) reduces to (2.6) and then the integral may be explicitly evaluated.

Let us define

$$F(t) \equiv 1 + t + \frac{t^2}{2!} + \cdots + \frac{t^b}{b!} = e^t \int_t^\infty e^{-w} \frac{w^b}{b!} dw \tag{3.2}$$

and we note that $F(t) - F'(t) = t^b/b!$. From (3.2) it follows that

$$\log[F(t)] - t = \log\left(1 - \int_0^t e^{-w} \frac{w^b}{b!} dw\right) = -\frac{t^{b+1}}{(b+1)!} + O(t^{b+2}), \qquad t \to 0. \tag{3.3}$$

We first consider the limit $n, k \to \infty$, with $n2^{-k} \equiv \xi$ fixed and $0 < \xi < b$. Scaling $z = n\omega$, (3.1) becomes

$$\bar{h}_n^k = \frac{n!}{n^n} \frac{1}{2\pi i} \oint \frac{e^{nf(\omega;\xi)}}{\omega} d\omega \tag{3.4}$$

where

$$f(\omega;\xi) = -\log\omega + \frac{1}{\xi}\log\left(1 + \omega\xi + \frac{\omega^2}{2!}\xi^2 + \cdots + \frac{\omega^b}{b!}\xi^b\right). \tag{3.5}$$

We evaluate (3.4) by the saddle point method (cf. [3, 22]). We can easily show that the equation $\frac{d}{d\omega} f(\omega;\xi) = 0$, i.e.,

$$-\frac{1}{\omega} + \frac{1 + \omega\xi + \cdots + (\omega\xi)^{b-1}/(b-1)!}{1 + \omega\xi + \cdots + (\omega\xi)^b/b!} = 1 - \frac{1}{\omega} - \frac{(\omega\xi)^b}{b!} \frac{1}{1 + \omega\xi + \cdots + (\omega\xi)^b/b!} = 0, \tag{3.6}$$

has a unique solution on the real ω-axis. We call it $\omega_0 = \omega_0(\xi; b)$. It satisfies

$$\begin{aligned} \omega_0 &\sim 1 - \frac{\xi^b}{b!} \to 1 \qquad \xi \to 0, \\ \omega_0 &\sim \frac{1}{b-\xi} \to \infty \qquad \xi \to b. \end{aligned}$$

Using Stirling's formula to simplify $n!$ in (3.4) and evaluating the integral by the standard saddle point method (cf. [3]) yields

$$\bar{h}_n^k \sim \sqrt{2\pi n} e^{-n} \frac{1}{2\pi} \frac{1}{\omega_0} \sqrt{\frac{2\pi}{f''(\omega_0; \xi) n}} e^{n f(\omega_0; \xi)}. \tag{3.7}$$

Also, by (3.4) and (3.5) we can show that

$$f''(\omega_0; \xi) = \frac{1}{\omega_0^2} + \left(1 - \frac{1}{\omega_0}\right)\left(\xi - \frac{b}{\omega_0}\right) - \xi\left(1 - \frac{1}{\omega_0}\right)^2 = \frac{1}{\omega_0^2}[1 + (\xi - b)(\omega_0 - 1)] > 0 \tag{3.8}$$

with which (3.7) becomes the same as Theorem 2, part (ii).

Next we consider the limit $n = O(1)$ with $k \to \infty$. In view of (2.3) we have $\bar{h}_n^k = 1$ for $0 \le n \le b$ and any $k \ge 0$. We return to (3.1) and note that for $k \to \infty$

$$\begin{aligned} H^k(z) &= \exp\left(2^k \log[F(z2^{-k})]\right) \\ &= \exp\left(2^k \left[z2^{-k} + \log\left(\int_{z2^{-k}}^{\infty} e^{-w} \frac{w^b}{b!} dw\right)\right]\right) \\ &= e^z \exp\left(-\frac{z^{b+1}}{(b+1)!} 2^{-kb} + O(z^{b+2} 2^{-k(b+1)})\right) \\ &= e^z \left(1 - \frac{z^{b+1}}{(b+1)!} 2^{-kb} + O(z^{b+2} 2^{-k(b+1)})\right) \end{aligned} \tag{3.9}$$

where we have used (3.3). Using the last expression of (3.9) in (3.1) and evaluating explicitly the resulting integral(s) leads to Theorem 2, part (i).

Finally, we consider the limit $k, n \to \infty$ with $2^k - n$ fixed. Setting $2^k - n = j$ and scaling $z = (n+j)\tau/b = 2^k \tau/b$ in (3.1) leads to

$$\bar{h}_n^k = n! \left(\frac{n+j}{b}\right)^{-n} \frac{1}{2\pi i} \oint \tau^{j-1} \left(\frac{1}{\tau^b} + \frac{1}{\tau^{b-1}} + \cdots + \frac{1}{(b-1)!} \frac{1}{\tau} + \frac{1}{b!}\right)^{\frac{n+j}{b}} d\tau. \tag{3.10}$$

We furthermore expand the integrand for τ large (to be precise, for $\tau = O(n)$) to get

$$\begin{aligned} \bar{h}_n^k &= n! \left(\frac{n+j}{b}\right)^{-n} \frac{1}{2\pi i} \oint \tau^{j-1} \exp\left(\frac{n+j}{b} \log\left[\frac{1}{b!} + \frac{1}{(b-1)!}\frac{1}{\tau} + O(\tau^{-2})\right]\right) d\tau \\ &\sim n! \left(\frac{n+j}{b}\right)^{-n} \exp\left[-\frac{n+j}{b} \log(b!)\right] \frac{1}{2\pi i} \oint \tau^{j-1} \exp\left(\frac{n+j}{\tau}\right) d\tau \\ &\sim n! \left(\frac{b}{n+j}\right)^n \exp\left(-\frac{n+j}{b} \log b!\right) \frac{n^j}{j!}. \end{aligned} \tag{3.11}$$

Theorem 2 part (iii) follows from (3.11) and Stirling's formula. In summary, to obtain (ii) from (3.1) the appropriate scaling was $z = O(n)$, the scaling for (iii) was $z = O(n^2)$.

3.2 Direct Method

What part of Theorem 2 can be obtained directly from the recurrence (2.2)? We first consider the limit $n = O(1)$, $k \to \infty$, and gradually increase the relative size of n to k. We note that (2.3) implies that $\bar{h}_n^k = 1$ for $0 \le n \le b$, for all $k \ge 0$, and we set $\bar{h}_n^k = 1 - G_n^k$. Then G_n^k satisfies

$$\begin{aligned} G_n^{k+1} &= 2\left(\frac{1}{2}\right)^n \sum_{i=0}^{n} \binom{n}{i} G_{n-i}^k - \left(\frac{1}{2}\right)^n \sum_{i=0}^{n} \binom{n}{i} G_i^k G_{n-i}^k \\ &= 2\left(\frac{1}{2}\right)^n \sum_{i=0}^{n-b-1} \binom{n}{i} G_{n-i}^k - \left(\frac{1}{2}\right)^n \sum_{i=b+1}^{n-b-1} \binom{n}{i} G_i^k G_{n-i}^k. \end{aligned} \tag{3.12}$$

Here we have used the fact that $G_n^k = 0$ for $0 \le n \le b$. For $b+1 \le n \le 2b+1$, the non-linear term in (3.12) vanishes and we are left with a *linear* recurrence. We can solve (3.12) exactly by first solving the linear problem for $n \in [b+1, 2b+1]$. Then, using this solution to compute explicitly the nonlinear term in (3.12) for $n \in [2b+2, 3b+2]$, we obtain an inhomogeneous linear problem. Thus we can solve the nonlinear problem (2.2) by solving a sequence of linear problems involving the n-intervals $[0,b]$, $[b+1, 2b+1]$, $[2b+2, 3b+2], \ldots$. However, the resulting expressions become complicated and may not be preferable to solving (2.2) by the transform method.

Our focus is on the asymptotics of G_n^k (hence $\bar{h}_n^k$). For ranges of k, n where $\bar{h}_n^k$ is *asymptotically close to* 1, we can replace (3.12) by the asymptotic relation

$$G_n^{k+1} \sim 2^{1-n} \sum_{i=0}^{n-b-1} \binom{n}{i} G_{n-i}^k. \tag{3.13}$$

This is "constant-coefficient" in k so we seek solutions of the form $G_n^k = a^k f(n)$. Replacing $\sim$ by $=$ in (3.13) we obtain

$$a f(n) = 2^{1-n} \sum_{i=b+1}^{n} \binom{n}{i} f(i), \qquad n \ge b+1. \tag{3.14}$$

Setting $n = b+1$ determines a as $a = 2^{-b}$ and then (3.14) becomes

$$f(n) = 2^{b+1-n} \sum_{i=b+1}^{n} \binom{n}{i} f(i), \quad n \ge b+1 \tag{3.15}$$

whose solution is

$$f(n) = \binom{n}{b+1} = \frac{n!}{(b+1)!(n-b-1)!}. \tag{3.16}$$

Here we have used $f(b+1) = 1$, which follows from (2.3). We have thus obtained Theorem 2(i) using only the recurrence relation. We also note that for $b+1 \le n \le 2b+1$, the exact answer is

$$G_n^k = \frac{n!}{b!} \sum_{\ell=0}^{n-b-1} \frac{(-1)^\ell}{\ell!(n-b-1-\ell)!} \frac{2^{-k(\ell+b)}}{\ell+1+b}$$

which certainly satisfies $G_n^k \sim 2^{-kb} f(n)$ as $k \to \infty$.

Next we consider k and n simultaneously large. We argue that the asymptotic relation $\bar{h}_n^k \sim 1 - 2^{-kb} f(n)$ becomes invalid when n is sufficiently large so to make $2^{-kb} f(n) = O(1)$. From (3.16) we see that $f(n) = O(n^{b+1})$ as $n \to \infty$ and thus $2^{-kb} f(n) = O(1)$ for $n^{b+1} 2^{-kb} = O(1)$, or $k = (1 + 1/b) \log_2 n + O(1)$. Let us set

$$k = \frac{b+1}{b} \log_2 n + \beta, \qquad \beta = O(1) \tag{3.17}$$

and re-examine (2.2) on the β-scale. Here we expect that it will no longer be permissible to neglect the non-linear term in (3.12) (and thus to linearize (2.2)).

We set

$$\bar{h}_n^k = F(\beta; n) = F(k - (1 + 1/b) \log_2 n; n) \tag{3.18}$$

and obtain from (2.2)

$$\begin{aligned} F(\beta+1; n) = \left(\frac{1}{2}\right)^n \sum_{i=0}^{n} \binom{n}{i} \quad &\cdot \; F\left(\beta - \left(1 + \frac{1}{b}\right) \log_2 \left(\frac{i}{n}\right); i\right) \\ &\cdot \; F\left(\beta - \left(1 + \frac{1}{b}\right) \log_2 \left(1 - \frac{i}{n}\right);\; n - i\right). \end{aligned} \tag{3.19}$$

Now we postulate that $F(\beta; n)$ assumes an expansion of the form

$$F(\beta; n) = F_0(\beta) + \frac{1}{n} F_1(\beta) + O(n^{-2}). \tag{3.20}$$

Then (3.19) yields the following equation for the leading term in (3.20)

$$F_0(\beta+1) = [F_0(\beta + 1 + 1/b)]^2. \tag{3.21}$$

Here we have used the fact that $2^{-n}\binom{n}{i} \sim \delta(i, n/2)$ as $n \to \infty$ (cf. (3.27) for a more precise estimate), where δ is the Kronecker delta function. Note that in order to obtain (3.21), the assumption (3.20) can be weakened to $F(\beta; n) \to F_0(\beta)$ as $n \to \infty$.

On the β-scale, we have thus approximated the non-linear problem (2.2) by another non-linear problem (3.21), which is much easier to solve. By using (3.20) and refining the approximation $2^{-n}\binom{n}{i} \sim \delta(i, n/2)$, as in (3.27), we can obtain equations for the correction terms $F_\ell(\beta)$, $\ell \geq 1$, in (3.20). These will be linear equations and the linear operator that will appear will be the linearization of (3.21).

The general solution of (3.21) is

$$F_0(\beta) = \exp(-c2^{-b\beta}) = \exp(-ce^{-b\beta \log 2}) \tag{3.22}$$

where c is a constant. To determine this constant we require that (3.20), which applies for $\beta = O(1)$, asymptotically matches to $\bar{h}_n^k \sim 1 - 2^{-kb} f(n)$, which applies for $k \to \infty$ and $n = O(1)$. Here, we use the principle of *matched asymptotic expansions* (cf. [20]). Symbolically, we write this condition as

$$F_0(\beta)|_{\beta \to \infty} \sim 1 - 2^{-kb} f(n)|_{n \to \infty}. \tag{3.23}$$

The left side of (3.23) is, using (3.17), $1 - c2^{-b\beta} = 1 - cn^{b+1}2^{-bk}$. Since $f(n) \sim n^{b+1}/(b+1)!$ as $n \to \infty$, the matching is accomplished with

$$c = \frac{1}{(b+1)!}. \tag{3.24}$$

The matching condition was necessary to uniquely determine the expansion on the β-scale. We also comment that the most general solution to (3.21) is $F_0(\beta) = \exp(-c(\beta)2^{-b\beta})$ where $c(\beta)$ is a periodic function with period $1/b$ (i.e., $c(\beta + 1/b) = c(\beta)$). However, the matching condition implies that $c(\beta)$ is in fact a constant. In summary, our analysis of the β-scale yielded the asymptotic distribution in (2.8) and (2.13), with no recourse to the exact solution.

We next consider $k, n \to \infty$ with $\xi = n2^{-k}$ fixed. We set

$$\bar{h}_n^k = G(\xi; n) = G(n2^{-k}; n)$$

and note that $\bar{h}_i^{k-1} = G(2i\xi/n; i)$ and $\bar{h}_{n-i}^{k-1} = G(2\xi(1 - i/n); n - i)$. From (2.2) we obtain

$$G(\xi; n) = \left(\frac{1}{2}\right)^n \sum_{i=0}^{n} \binom{n}{i} G\left(\frac{2i}{n}\xi; i\right) G\left(2\left(1 - \frac{i}{n}\right)\xi; n - i\right). \tag{3.25}$$

Note that the initial condition (2.3) does not apply on the ξ-scale, since k is assumed large.

We analyze (3.25) by the WKB method [2, 13]. That is, we seek an asymptotic solution of the form

$$G(\xi; n) \sim e^{n\phi(\xi)} \left[A(\xi) + \frac{1}{n}A^{(1)}(\xi) + \frac{1}{n^2}A^{(2)}(\xi) + \cdots\right] \tag{3.26}$$

The *ansatz* (3.26) may be viewed as a generalized saddle point approximation. By symmetry, the major contribution to the sum will come from $i \approx n/2$. We also note that Stirling's formula yields, for $i = xn$ $(0 < x < 1)$,

$$2^{-n}\binom{n}{i} = \frac{e^{nf_0(x)}}{\sqrt{2\pi n}}\frac{1}{\sqrt{x(1-x)}}\left[1+\frac{1}{12n}\left(1-\frac{1}{x}-\frac{1}{1-x}\right)+O(n^{-2})\right], \tag{3.27}$$

where $f_0(x) = -\log 2 - x\log x - (1-x)\log(1-x)$. For $x = 1/2 + y/\sqrt{n}$ (i.e., $i = n/2 + O(\sqrt{n})$), (3.27) simplifies to the Gaussian form

$$2^{-n}\binom{n}{n/2+y\sqrt{n}} = \sqrt{\frac{2}{\pi n}}e^{-2y^2}\left[1+\frac{1}{n}\left(-\frac{1}{4}+2y^2-\frac{4}{3}y^4\right)+O(n^{-2})\right]. \tag{3.28}$$

Using (3.28) and (3.26) in (3.25) and retaining only leading order terms, we are led to

$$\begin{aligned} A(\xi)e^{n\phi(\xi)} \quad &\sim \quad \sum_{i=0}^{n}\sqrt{\frac{2}{\pi n}}e^{-2y^2}A\left(\frac{2i}{n}\xi\right)A\left(\frac{2(n-i)}{n}\xi\right) \\ &\times \quad \exp\left[i\phi\left(\frac{2i}{n}\xi\right)+(n-i)\phi\left(2\left(1-\frac{i}{n}\right)\xi\right)\right]. \end{aligned} \tag{3.29}$$

Now we set $\psi(x) = x\phi(2x\xi) + (1-x)\phi(2(1-x)\xi)$ and expand this function about $x = 1/2$. We have $\psi(1/2) = \phi(\xi)$, $\psi(1/2) = 0$ and $\psi''(1/2) = 8\xi\phi'(\xi) + 4\xi^2\phi''(\xi)$. Approximating the sum in (3.29) by an integral and using the Laplace method to evaluate the integral for $n \to \infty$, we obtain

$$e^{n\phi(\xi)}A(\xi) = \sqrt{\frac{2}{\pi}}e^{n\phi(\xi)}A^2(\xi)\int_{-\infty}^{\infty}\exp\left(-2y^2+y^2(4\xi\phi'(\xi)+2\xi^2\phi''(\xi)\right)dy. \tag{3.30}$$

In (3.30) the exponential factors $e^{n\phi}$ cancel and we have

$$1 = \sqrt{\frac{2}{\pi}}A(\xi)\sqrt{\frac{\pi}{2-4\xi\phi'(\xi)-2\xi^2\phi''(\xi)}}. \tag{3.31}$$

We have thus determined the function $A(\xi)$ in terms of $\phi(\xi)$, though we have not determined the latter. By continuing the expansion of (3.25) (using (3.26) and (3.28)) to higher orders we can express $A^{(1)}$ in terms of A and ϕ, then $A^{(2)}$ in terms of $A^{(1)}$, A and ϕ, etc. Thus, the asymptotic series (3.26) is known, up to the function $\phi(\xi)$. It does not seem to be possible to determine ϕ using only the recursion (2.2). This function is apparently very sensitive to the initial condition(s) (2.3). By comparing Theorems 1 and 2, we see that ϕ depends on the parameter b, which enters the problem only through the initial condition (2.3).

We verify that $A(\xi) = \sqrt{1-2\xi\phi'(\xi)-\xi^2\phi''(\xi)}$ is consistent with Theorem 2(ii). By differentiating $\phi(\xi;b)$ in Theorem 2(ii) we obtain

$$(\xi\phi)' = -1 - \log\omega_0 + \frac{b}{\xi} + \left(b - \xi - \frac{1}{\omega_0 - 1}\right)\frac{\omega_0'}{\omega_0}. \tag{3.32}$$

By differentiating equation (3.6) we obtain, after some simplification,

$$\omega_0' = \left(\frac{b}{\xi} - 1\right)(\omega_0 - 1)(\omega_0 + \xi\omega_0'). \tag{3.33}$$

From (3.32) and (3.33) it follows that

$$(\xi\phi)' = -\log\omega_0 \tag{3.34}$$

so that $(\xi\phi)'' = 2\phi' + \xi\phi'' = -\omega_0'/\omega_0$ and thus

$$1 - 2\xi\phi' - \xi^2\phi'' = 1 + \xi\frac{\omega_0'}{\omega_0} = [1 + (\omega_0 - 1)(\xi - b)]^{-1},$$

which agrees with $A(\xi; b)$ of Theorem 2(ii) and verifies the relation between $A(\xi)$ and $\phi(\xi)$ in (3.31).

Finally, we comment that it is possible to obtain the (complete) result in Theorem 2(iii) using only the recurrence (2.2). We omit that analysis since it is completely analogous to that presented for PATRICIA tries (for $2^k - n = O(1)$) in Section 4. By using this result and asymptotic matching, we can infer the behavior of $\phi(\xi)$ as $\xi \to b$. By matching (3.26) to either the expansion on the β-scale, or the result for $k \to \infty$, $n = O(1)$, we can obtain the behavior of $\phi(\xi)$ as $\xi \to 0$.

To summarize, we have obtained a fair amount of asymptotic information for $\bar{h}_n^k$ using only (2.2) and (2.3), along with some mild assumptions about the forms of the expansions in the various regions of (n, k) space. The main deficiency in this method is the failure to determine the function ϕ, except for its behaviors as $\xi \to 0$ and $\xi \to b$. However, for ranges of k and n where appreciable mass accumulates (i.e., where $\bar{h}_n^k$ undergoes the transition from 0 to 1), we have been able, using asymptotic matching, to completely determine the asymptotic solution (cf. (3.22) and (3.24)).

4 Asymptotics of PATRICIA Tries

We analyze the PATRICIA model asymptotically for $n \to \infty$, and also give some exact results when n is close to k or 2^k. Since we do not have an exact expression for h_n^k, we use the ideas developed in the previous section and analyze the recurrence (2.4). We first discuss the right tail asymptotics and then deal with the left tail approximation.

4.1 Right Tail Asymptotics

From the definition of the PATRICIA model, it follows that $\Pr\{\mathcal{H}_n^P = k\} = h_n^k = 1$ for $k \geq n - 1$ and $h_n^k = 0$ for $n > 2^k$. It thus suffices to consider the range $k + 2 \leq n \leq 2^k$.

We analyze here the "right tail" (we think of plotting h_n^k as a function of k for a fixed n), where h_n^k is asymptotically close to 1. We set

$$H_n^k = 1 - h_n^k, \tag{4.1}$$

and then $H_n^k \equiv L_{n-k}(n) = L_j(n)$ for $j = n-k$. Using (4.1) in (2.4) we obtain

$$(2-2^n)L_{j-1}(n) + 2\sum_{i=1}^{j-i}\binom{n}{i}L_{j-i}(n-i) = \sum_{i=n-j+2}^{j-2}\binom{n}{i}L_{i+j-n}(i)L_{j-i}(n-i). \tag{4.2}$$

The above holds for $n \geq j$ and we have used the fact that $L_1(n) = L_0(n) = L_{-1}(n) = L_{-2}(n) = \cdots = 0$. The boundary condition (2.5) implies that $H_n^0 = 0$ for $n = 0,1$ and $H_n^0 = 1$ for $n \geq 2$. Hence, the boundary condition for $L_j(n)$ is

$$L_n(n) = 0, \quad n = 0,1; \tag{4.3}$$

$$L_n(n) = 1, \quad n \geq 2. \tag{4.4}$$

We first compute $L_j(n)$ exactly for $j = 2, 3$ and 4. We set $j = 3$ in (4.2) and obtain

$$(2-2^n)L_2(n) + 2nL_2(n-1) = 0, \qquad n \geq 3. \tag{4.5}$$

Since $L_2(2) = 1$, we can easily solve this linear recurrence and thus obtain

$$L_2(n) = n!2^{-n^2/2}2^{n/2}\prod_{m=3}^{n}\left(\frac{1}{1-2^{1-m}}\right), \qquad n \geq 2. \tag{4.6}$$

If the upper limit in any product exceeds the lower limit, we define the product to be 1.

Setting $j = 4$ in (4.2) we see that the sum in the right-hand side is void for $n \geq 5$, which yields

$$(2-2^n)L_3(n) + 2nL_3(n-1) + 2\binom{n}{2}L_2(n-2) = 0, \qquad n \geq 5 \tag{4.7}$$

and when $n = j = 4$ we obtain (using $L_2(2) = L_3(3) = 1$)

$$(2-2^4)L_3(4) + 20 = 6.$$

It follows that $L_3(4) = 1$ and then (4.7) is readily solved, using (4.6), to give

$$L_3(n) = n!2^{-n^2/2}2^{3n/2}\left(\frac{1}{4} - \frac{n2^{-n}}{2} - \frac{1}{4}2^{-n}\right)\prod_{m=3}^{n}\left(\frac{1}{1-2^{1-m}}\right), \qquad n \geq 4. \tag{4.8}$$

Next we set $j = 5$ in (4.2) and note that the sum is void for $n \geq 7$. By examining (4.2) with $j = 5$ and $n = 5, 6$ we find that $L_4(5) = L_4(6) = 1$. For $n \geq 7$ we solve (4.2) and obtain, after some calculation,

$$L_4(n) = n!2^{-n^2/2}2^{5n/2} \cdot \left[\frac{5}{288} + \frac{1-n}{8}\left(\frac{1}{2}\right)^n + \left(\frac{n^2}{8} - \frac{n}{12} - 8 + \frac{1}{72}\right)\left(\frac{1}{4}\right)^n\right] \cdot \prod_{m=3}^{n}\left(\frac{1}{1-2^{1-m}}\right), \qquad n \geq 6 \tag{4.9}$$

Using this method we can solve for $L_j(n)$ for any fixed j, but the calculations become tedious as j become large. For $n \geq 2j-3$ the recurrence (4.2) becomes linear:

$$(2-2^n)L_{j-1}(n) + 2\sum_{i=1}^{j-2}\binom{n}{i}L_{j-i}(n-i) = 0, \qquad n \geq 2j-3. \tag{4.10}$$

We can (in theory) use (4.10) to express $L_j(n)$ in terms of $L_j(2j-2)$, but the latter is unknown.

Now consider arbitrary j and the limit $n \to \infty$. From (4.6), (4.8) and (4.9) it is easy to see that

$$\begin{aligned} L_j(n) &\sim \rho_0 K_j n! 2^{-n^2/2}2^{(j-3/2)n}, \quad j \text{ fixed}, \\ \rho_0 &= \prod_{\ell=2}^{\infty}(1-2^{-\ell})^{-1}. \end{aligned} \tag{4.11}$$

To compute the constants K_j, we use (4.11) in (4.10) and obtain, to leading order in n,

$$K_j(1-2^{2-j}) = \sum_{\ell=1}^{j-2}\frac{K_{j-\ell}}{(\ell+1)!}2^{2+\ell(\ell+3)/2-(\ell+1)j} \tag{4.12}$$

for $j \geq 3$ with $K_2 = 1$, in view of (4.6). The recurrence (4.12) may be somewhat simplified by setting

$$K_j = \frac{1}{j!}2^{-j^2/2}2^{3j/2}C_j \tag{4.13}$$

which leads to

$$C_j = 4\sum_{m=2}^{j}\binom{j}{m}\frac{2^{-j}}{j-m+1}C_m, \qquad j \geq 2 \tag{4.14}$$

with $C_2 = 1$.

To solve (4.14) we introduce the exponential generating function

$$C(z) = \sum_{j=0}^{\infty}\frac{z^j}{j!}C_j \tag{4.15}$$

to obtain

$$C(z) = 4C\left(\frac{z}{2}\right)\frac{e^{z/2}-1}{z/2}. \tag{4.16}$$

Setting $\widetilde{C}(z) = e^{-z}C(z)$ (so that $\widetilde{C}(z)$ is the Poisson transform of C_j), (4.16) becomes

$$\widetilde{C}(z) = \frac{8}{z}(1 - e^{-z/2})\widetilde{C}\left(\frac{z}{2}\right) \tag{4.17}$$

where $\widetilde{C}(z) \sim z^2/2$ as $z \to 0$. The latter follows from $C_2 = 1$ and we define $C_0 = C_1 = 0$. Next we set $\widetilde{C}(z) = z^2 G(z)/2$ and replace z by $2z$ in (4.17), which gives

$$G(2z) = G(z)\left(\frac{1 - e^{-z}}{z}\right) \tag{4.18}$$

with $G(0) = 1$. Taking the logarithm of (4.18) with $F(z) = \log G(z)$, we obtain

$$F(2z) - F(z) = \log\left(\frac{1 - e^{-z}}{z}\right), \tag{4.19}$$

with $F(0) = 0$.

Functional equations of the type (4.19) are often encountered in the analysis for algorithms and they are usually handled by the Mellin transform [11]. The interested reader can find more on Mellin transform in a recent survey [11]. The Mellin transform $F^*(s)$ of a real valued function $F(z)$ is defined as

$$F^*(s) = \int_0^\infty F(z)z^{s-1}dz = \mathcal{M}[F; s]$$

and its inverse is

$$F(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s}F^*(s)ds = \mathcal{M}^{-1}[F^*; z]$$

where c belongs to the so called *fundamental strip* where the Mellin transform is analytic (cf. [11]).

Taking the Mellin transform of (4.19) yields

$$F^*(s) = \frac{2^s}{1 - 2^s}\int_0^\infty z^{s-1}\log\left(\frac{1 - e^{-z}}{z}\right)dz \tag{4.20}$$

for $\Re(s) \in (-1, 0)$. We now evaluate the inverse Mellin transform of $F^*(s)$ by two methods. First, we observe that

$$\mathcal{M}^{-1}\left[\frac{2^s}{1 - 2^s}; z\right] = \sum_{m=0}^{\infty} 2^{m+1}\delta(z - 2^{m+1}) \tag{4.21}$$

where $\delta(\cdot)$ is the Dirac delta function. The convolution theorem for Mellin transforms $f^*(s)$ and $g^*(s)$ of functions $f(z)$ and $g(z)$ is

$$\mathcal{M}^{-1}[f^*g^*; z] = \int_0^\infty f(\xi)g(z/\xi)\frac{d\xi}{\xi}. \tag{4.22}$$

We apply (4.22) with f given by the integral in (4.20) and $g = 2^s/(1-2^s)$. Using (4.21) then yields

$$
\begin{aligned}
F(z) &= \int_0^\infty \frac{1}{\xi} \log\left(\frac{1-e^{-\xi}}{\xi}\right) \sum_{m=0}^{\infty} 2^{m+1} \delta\left(\frac{z}{\xi} - 2^{m+1}\right) d\xi \qquad (4.23)\\
&= \int_0^\infty \frac{1}{u} \log\left(\frac{1-e^{-z/u}}{z/u}\right) \sum_{m=0}^{\infty} 2^{m+1} \delta(u - 2^{m+1}) du \\
&= \sum_{m=0}^{\infty} \log\left(\frac{1-\exp(-z2^{-m-1})}{z2^{-m-1}}\right).
\end{aligned}
$$

With (4.23), we can easily compute $C(z) = e^z z^2 e^{F(z)}/2$ and then invert the generating function in (4.15) by Cauchy's theorem to get

$$
C_j = \frac{j!}{2\pi i} \oint z^{-j-1} \frac{z^2}{2} e^z \prod_{m=0}^{\infty} \left(\frac{1-\exp(-z2^{-m-1})}{z2^{-m-1}}\right) dz, \qquad (4.24)
$$

which establishes (2.16) of Theorem 3(i).

An alternate expression for C_j can be established by evaluating the integral in (4.20) in a different way. Integrating by parts, we find

$$
\begin{aligned}
\mathcal{M}\left[\log\left(\frac{1-e^{-z}}{z}\right)\right] &= \int_0^\infty z^{s-1} \log\left(\frac{1-e^{-z}}{z}\right) dz \qquad (4.25)\\
&= -\frac{1}{s} \int_0^\infty z^{s-1} \left(\frac{z}{e^z - 1} - 1\right) dz,
\end{aligned}
$$

where the last integral converges for $-1 < \Re(s) < 0$. We now use

$$
\mathcal{M}\left(\frac{z}{e^z-1}\right) = \Gamma(s+1)\zeta(s+1), \qquad 0 < \Re(s) < \infty \qquad (4.26)
$$

where $\Gamma(\cdot)$ and $\zeta(\cdot)$ are the Gamma and Riemann zeta functions, respectively. Shifting the fundamental strip to $\Re(s) \in (-1, 0)$ we obtain (cf. [11])

$$
\mathcal{M}\left[\frac{z}{e^z-1} - 1\right] = \Gamma(s+1)\zeta(s+1), \qquad -1 < \Re(s) < 0. \qquad (4.27)
$$

We use (4.27) to evaluate the right side of (4.25) and then (4.20) becomes

$$
F^*(s) = -\frac{\Gamma(s+1)\zeta(s+1)}{s(2^{-s}-1)}
$$

for $\Re(s) \in (-1, 0)$. It follows that an alternate representation for the Poisson transform $\widetilde{C}(z)$ of C_j is

$$
\widetilde{C}(z) = \frac{z^2}{2} \exp\left(\frac{1}{2\pi i} \int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \frac{z^{-s}\Gamma(s+1)\zeta(s+1)}{s(1-2^{-s})} ds\right). \qquad (4.28)
$$

In Appendix A we prove that (4.28) and (4.24) are equivalent.

In view of (4.11) we see that $1 - h_n^k = 1 - h_n^{n-j}$ is exponentially small (roughly of order $O(2^{-n^2/2})$) for $n \to \infty$ and j fixed. We next study the asymptotics of C_j (and hence K_j and $L_j(n)$) for $j \to \infty$. If j is sufficiently large so as to make $h_n^{n-j} = O(1)$, then this will give us a rough estimate of the range of k where h_n^k changes from $h_n^k \approx 0$ to $h_n^k \approx 1$.

To accomplish the above goal, we apply the "principle of matched asymptotics" and assume that (4.24) holds also for $j \to \infty$. The Poisson transform $\widetilde{C}(z)$ of C_j is given in (4.28). Our goal is to "depoissonize" it, that is, to extract C_j for $j \to \infty$ from the behavior of $\widetilde{C}(z)$ as $z \to \infty$ in a cone around real axis. We expect that, under some mild growth assumptions of $\widetilde{C}(z)$, $C_j \sim \widetilde{C}(j)$ since the Poisson process is well concentrated around its mean $z = j$. To be more precise, we appeal to recent depoissonization results of Jacquet and Szpankowski [17], applying the following.

Theorem 4 (Jacquet and Szpankowski 1998) *Let g_n be a sequence whose Poisson transform is $\widetilde{G}(z) = e^{-z} \sum_{n \geq 0} g_n \frac{z^n}{n!}$ where z is complex. Consider a linear cone $\mathcal{S}_\theta = \{z : \arg(z) \leq \theta, \ |\theta| < \pi/2\}$. Assume for $z \to \infty$ that:*
(I) For $z \in \mathcal{S}_\theta$

$$|\widetilde{G}(z)| \leq A \exp(B|z|^\beta) \tag{4.29}$$

where $0 \leq \beta < 1/2$, and $A, B > 0$ are constants.
(O) For $z \notin \mathcal{S}_\theta$

$$|\widetilde{G}(z) e^z| \leq A_1 \exp(\omega |z|) \tag{4.30}$$

for $\omega < 1$ and $A_1 > 0$. Then

$$g_n = \widetilde{G}(n) + O\left(n^{-(1-2\beta)} \exp(Bn^\beta)\right) \tag{4.31}$$

for $n \to \infty$.

We apply Theorem 4 to find C_j for large j. We present two alternate derivations. Let us start with (4.28). Using (4.14) and the method of "increasing domains" proposed in [17] we can easily show that condition (4.30) of Theorem 4 is satisfied. To verify condition (4.29), let us evaluate $\widetilde{C}(z)$ asymptotically for $z \to \infty$ in the cone $\mathcal{S}_\theta$. We first compute asymptotically the exponent of (4.28), that is,

$$F(z) = \frac{1}{2\pi i} \int_{-\frac{1}{2} - i\infty}^{-\frac{1}{2} + i\infty} \frac{z^{-s} \Gamma(s+1) \zeta(s+1)}{s(1 - 2^{-s})} ds \tag{4.32}$$

as $z \to \infty$. We note that the integrand in (4.32) has a triple pole at $s = 0$ and simple poles along the imaginary axis, at $s = -2\pi ik/\log(2) = s_k$, $k = \pm 1, \pm 2, \ldots$. A lengthy computation shows that the negative of the residue at $s = 0$ is

$$-\frac{\log^2(z)}{2\log 2} + \frac{1}{2}\log(z) - \frac{1}{12}\log(2) + \frac{1}{\log 2}\left(\gamma(1) + \frac{1}{2}\gamma^2 - \frac{\pi^2}{12}\right).$$

and the residues at $s = s_k$ are

$$\mathrm{Res}_{s=s_k}[F(z)] = -\frac{1}{2\pi ik}\Gamma(1+s_k)\zeta(1+s_k).$$

It follows that

$$\exp[F(z)] \sim \sqrt{z}2^{-1/12}\exp\left(\frac{\gamma(1) + \gamma^2/2 - \pi^2/12}{\log 2}\right)\exp\left[-\frac{1}{2}\frac{\log^2(z)}{\log 2} + \Psi(\log_2 z)\right]$$

where

$$\Psi(\log_2 z) = \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \frac{1}{2\pi i\ell}\Gamma\left(1 - \frac{2\pi i\ell}{\log 2}\right)\zeta\left(1 - \frac{2\pi i\ell}{\log 2}\right)e^{2\pi i\ell\log_2 z}.$$

Thus, $\widetilde{C}(z) = O(z^{5/2}\exp(\log^2 z))$ in the cone $\mathcal{S}_\theta$ (by analytic continuation). In view of this, we can apply Theorem 4 to get

$$C_j \sim \frac{1}{2}j^{5/2}2^{-1/12}\exp\left(\frac{\gamma(1) + \gamma^2/2 - \pi^2/12}{\log 2}\right)\exp\left[-\frac{1}{2}\frac{\log^2(j)}{\log 2} + \Psi(\alpha)\right] \tag{4.33}$$

where $\alpha = \langle\log_2 j\rangle$ with $\langle\cdot\rangle$ denoting the fractional part.

An alternate representation for the above can be obtained by using (4.24). Following the footsteps of the depoissonization verification above, we can show that (I) and (O) of Theorem 4 hold, and hence

$$C_j \sim \frac{j^2}{2}\exp\left[\sum_{m=1}^{\infty}\log\left(\frac{1-\exp(-j2^{-m})}{j2^{-m}}\right)\right], \qquad j \to \infty. \tag{4.34}$$

We can further simplify the above by setting

$$m = \log_2(j) - \alpha + \ell, \tag{4.35}$$

where $\alpha = \langle\log_2 j\rangle$. Using (4.35) in (4.34) we have

$$\begin{aligned}\sum_{\ell=1-\lfloor\log_2(j)\rfloor}^{\infty}\log\left(\frac{1-\exp(-2^{\alpha-\ell})}{2^{\alpha-\ell}}\right) &= \sum_{\ell=0}^{\infty}\log\left(\frac{1-\exp(-2^{\alpha-\ell})}{2^{\alpha-\ell}}\right) \\ &+ \sum_{\ell=1}^{\infty}\log[1-\exp(-2^{\alpha+\ell})] \\ &- \sum_{N=1}^{\lfloor\log_2(j)\rfloor-1}(N+\alpha)\log(2) + O(\rho^j)\end{aligned} \tag{4.36}$$

for some $\rho < 1$. The last sum in the right-hand side of (4.36) is

$$
\begin{aligned}
-\alpha \log 2(\lfloor \log_2(j) \rfloor - 1) - \frac{1}{2} \log 2(\lfloor \log_2(j) \rfloor - 1)\lfloor \log_2(j) \rfloor \\
= -\frac{1}{2}\frac{\log^2(j)}{\log 2} + \frac{1}{2}\log(j) + \frac{1}{2}\log 2\alpha(\alpha+1).
\end{aligned}
\tag{4.37}
$$

Using (4.36) and (4.37) in (4.34) we find that

$$
C_j \sim \frac{1}{2} j^{5/2} \exp\left(-\frac{\log^2(j)}{2\log 2}\right) e^{\varphi(\alpha)}, \qquad \alpha = \langle \log_2(j) \rangle \tag{4.38}
$$

where $\varphi(\alpha+1) = \varphi(\alpha)$ is given by

$$
\varphi(\alpha) = \frac{1}{2}\alpha(\alpha+1)\log 2 + \sum_{\ell=0}^{\infty} \log\left(\frac{1-\exp(-2^{\alpha-\ell})}{2^{\alpha-\ell}}\right) + \sum_{\ell=1}^{\infty} \log\left[1 - \exp(-2^{\alpha+\ell})\right]. \tag{4.39}
$$

By comparing (4.38) and (4.33), it follows that $\varphi(\alpha)$ and $\Psi(\alpha)$ are related as in Theorem 3(ii).

We are now in position to continue our analysis and extend the validity of (4.11) to $j \to \infty$. Using (4.38) and (4.13), we see that the right side of (4.11) becomes for $j \to \infty$

$$
L_j(n) = \frac{n!}{j!} 2^{-(j-n)^2/2} 2^{3(j-n)/2} j^{5/2} \exp\left(-\frac{\log^2(j)}{2\log 2}\right) \frac{\rho_0}{2} e^{\varphi(\alpha)}. \tag{4.40}
$$

We set $j = n - k$ and find for what range of k is (4.40) $O(1)$ as $n \to \infty$. Taking the logarithm of (4.40), this condition is the same as

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

Using $j\log j = (n-k)\log(n-k) = (n-k)[\log n - k/n + O(k^2 n^{-2})]$, the above becomes

$$
k\log n - \frac{\log 2}{2}k^2 - \frac{3}{2}(\log 2)k - \frac{\log^2 n}{2\log 2} + \frac{5}{2}\log n = O(1)
$$

and this implies that

$$
k = \log_2 n + \sqrt{2\log_2 n} - \frac{3}{2} + o(1).
$$

Thus, we expect that if k satisfies the above condition, then the asymptotic expression for h_n^k in Theorem 3(i) breaks down, as the second term becomes comparable in magnitude to the first term. This completes our discussion of the right tail of h_n^k, where h_n^k is asymptotically close to 1.

4.2 Left Tail Asymptotics

We next consider the left tail. Here h_n^k will be asymptotically small. We first consider the limit where $k, n \to \infty$ with $2^k - n \equiv M = O(1)$ and $M \geq 0$. This corresponds to the "left-most" tail of the distribution, since $h_n^k = 0$ for $2^k < n$. We shall derive exact results for h_n^k if $M = 0$ or 1 and then establish part(iii) of Theorem 3, which applies for arbitrary $M = O(1)$.

Let us define the exponential generating function $H^k(z)$ of h_n^k as

$$H^k(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!} h_n^k \tag{4.41}$$

with which (2.4) becomes

$$H^{k+1}(z) = 2H^{k+1}\left(\frac{z}{2}\right) + \left[H^k\left(\frac{z}{2}\right)\right]^2 - 2H^k\left(\frac{z}{2}\right), \quad k \geq 0 \tag{4.42}$$

and $H^0(z) = 1 + z$ by (2.5). We can simplify (4.42) by setting $H^k(z) = 1 + \tilde{H}^k(z)$ to get

$$\tilde{H}^{k+1}(z) - 2\tilde{H}^{k+1}\left(\frac{z}{2}\right) = \left[\tilde{H}^k\left(\frac{z}{2}\right)\right]^2, \quad k \geq 0 \tag{4.43}$$

and $\tilde{H}^0(z) = z$.

From our previous discussion, we can truncate the sum in (4.41) at $n = 2^k$ so that $H^k(z)$ (and thus $\tilde{H}^k(z)$) will be a polynomial of degree 2^k. We will identify the two leading coefficients in this polynomial by writing

$$\tilde{H}^k(z) = a(k)z^{2^k} + b(k)z^{2^k-1} + \cdots + z \tag{4.44}$$

where

$$\begin{aligned} a(k) &= h_{2^k}^k \frac{1}{(2^k)!}, \\ b(k) &= \frac{h_{2^k-1}^k}{(2^k-1)!}. \end{aligned}$$

From (4.44) it follows that

$$\left[\tilde{H}^k\left(\frac{z}{2}\right)\right]^2 = \frac{a^2(k)}{2^{2^{k+1}}} z^{2^{k+1}} + \frac{4a(k)b(k)}{2^{2^{k+1}}} z^{2^{k+1}-1} + \cdots + \frac{z^2}{4} \tag{4.45}$$

and

$$\begin{aligned} \tilde{H}^{k+1}(z) - 2\tilde{H}^{k+1}\left(\frac{z}{2}\right) &= \left(1 - \frac{2}{2^{2^{k+1}}}\right) a(k+1) z^{2^{k+1}} \\ &+ \left(1 - \frac{4}{2^{2^{k+1}}}\right) b(k+1) z^{2^{k+1}-1} + O(z^{2^{k+1}-2}). \end{aligned} \tag{4.46}$$

By comparing (4.45) to (4.46) we obtain the recurrences

$$a(k+1)\left[1-2\left(\frac{1}{2}\right)^{2^{k+1}}\right]=a^2(k)\frac{1}{2^{2^{k+1}}} \tag{4.47}$$

and

$$b(k+1)\left[1-4\left(\frac{1}{2}\right)^{2^{k+1}}\right]=\frac{4}{2^{2^{k+1}}}a(k)b(k). \tag{4.48}$$

Solving (4.47) subject to $a(0)=1$ yields

$$a(k)=2^{-k2^k}\prod_{\ell=1}^{k}\left(1-2\left(\frac{1}{2}\right)^{2^\ell}\right)^{-2^{k-\ell}}. \tag{4.49}$$

It follows that

$$h_{2^k}^k=(2^k)!2^{-k2^k}\exp\left[-2^k\sum_{\ell=1}^{k}2^{-\ell}\log\left(1-2\left(\frac{1}{2}\right)^{2^\ell}\right)\right]. \tag{4.50}$$

Evaluating (4.50) as $k\to\infty$ yields

$$h_{2^k}^k\sim\sqrt{2\pi}\sqrt{2^k}\exp[2^k(-1-\log K_0^*)],\quad k\to\infty, \tag{4.51}$$

where

$$K_0^*=\prod_{\ell=1}^{\infty}\left[1-2\left(\frac{1}{2}\right)^{2^\ell}\right]^{2^{-\ell}}$$

and the numerical value of K_0^* is given in Theorem 3.

Having computed $a(k)$ we can easily solve the linear recurrence (4.48) for $b(k)$, subject to $b(1)=1$. This yields

$$b(k)=\prod_{\ell=1}^{k-1}\frac{4}{1-4^{1-2^\ell}}\frac{a(\ell)}{2^{2^{\ell+1}}}. \tag{4.52}$$

The expression (4.49) may be rewritten as

$$a(k)=2^{-k2^k}\left(\frac{1}{K_0^*}\right)^{2^k}\prod_{m=1}^{\infty}\left(1-2\left(\frac{1}{2}\right)^{2^{k+m}}\right)^{2^{-m}}$$

with which (4.52) becomes

$$\begin{aligned} b(k) &= \prod_{\ell=1}^{k-1}\frac{4\cdot 2^{-\ell 2^\ell}(K_0^*)^{-2^\ell}}{1-4\cdot 2^{-2^{\ell+1}})2^{2^{\ell+1}}}\prod_{m=1}^{\infty}\left(1-2\left(\frac{1}{2}\right)^{2^{\ell+m}}\right)^{2^{-m}} \\ &= \prod_{\ell=1}^{k-1}\frac{4\cdot 2^{-(\ell+2)2^\ell}}{1-4\cdot 2^{-2^{\ell+1}}}\prod_{m=1}^{\ell}\left(\frac{1}{1-2\cdot 2^{-2^m}}\right)^{2^{\ell-m}}. \end{aligned} \tag{4.53}$$

As $k \to \infty$ we obtain from (4.53)

$$b(k) \sim 2^{2k} 2^{-k2^k} (K_0^*)^{2-2^k} K_1^*, \qquad k \to \infty \tag{4.54}$$

where

$$K_1^* = \prod_{\ell=1}^{\infty} \prod_{m=1}^{\infty} \frac{(1 - 2 \cdot 2^{-2^{\ell+m}})^{2^{-m}}}{1 - 4 \cdot 2^{-2^{\ell+1}}} = \prod_{N=2}^{\infty} \frac{(1 - 2 \cdot 2^{-2^N})^{1-2^{1-N}}}{1 - 4 \cdot 4^{-2^N}}.$$

Using (4.54) and approximating $(2^k - 1)!$ by Stirling's formula we are led to

$$h_{2^k-1}^k \sim \sqrt{2\pi} 2^k \sqrt{2^k} K_1^* (K_0^*)^2 \exp[2^k(-1 - \log K_0^*)], \qquad k \to \infty. \tag{4.55}$$

We next consider (2.4) for values of n that are close to 2^k. We set $M = 2^k - n$, with

$$h_n^k = W(2^k - n; n) = W(M; n) \tag{4.56}$$

and note that

$$h_i^{k-1} = W\left(\frac{M}{2} + \frac{n}{2} - i; i\right) \quad \text{and} \quad h_{n-i}^{k-1} = W\left(\frac{M}{2} - \frac{n}{2} + i; n - i\right)$$

Replacing k by $k - 1$ in (2.4) and dropping the first term in the right-hand side (which is clearly negligible as $n \to \infty$) we obtain

$$\begin{aligned} W(M; n) &\sim \sum_{i=1}^{n-1} \binom{n}{i} 2^{-n} W\left(\frac{M+n}{2} - i; i\right) W\left(\frac{M-n}{2} + i; n - i\right) \\ &= \sum_{i=(n-M)/2}^{(n+M)/2} \binom{n}{i} 2^{-n} W\left(\frac{M+n}{2} - i; i\right) W\left(\frac{M-n}{2} + i; n - i\right). \end{aligned} \tag{4.57}$$

Here we have used the fact that $W(M; n) = 0$ for $M < 0$ (since $h_n^k = 0$ for $2^k < n$) to truncate the limits on the summation in (4.57). For $i = n/2 + O(1)$ we have

$$\binom{n}{i} 2^{-n} \sim \sqrt{\frac{2}{\pi n}}$$

so that (4.57) can be rewritten as

$$W(M; n) \sim \sqrt{\frac{2}{\pi n}} \sum_{\ell=0}^{M} W\left(\ell; \frac{n+M}{2} - \ell\right) W\left(M - \ell; \frac{n-M}{2} + \ell\right). \tag{4.58}$$

Setting $M = 0$, (4.58) becomes

$$W(0; n) \sim \sqrt{\frac{2}{\pi n}} \left[W\left(0; \frac{n}{2}\right)\right]^2$$

that admits an asymptotic solution in the form

$$W(0;n) \sim \sqrt{2\pi n}e^{-D_1 n} \tag{4.59}$$

where D_1 is at this point undetermined. By considering $M = 1$ and proceeding inductively, we find that (4.58) admits an asymptotic solution

$$W(M;n) \sim n^{M+1/2}A(M)e^{-D_1 n} \tag{4.60}$$

where $A(M)$ satisfies

$$A(M) = \frac{1}{\sqrt{2\pi}}2^{-M}\sum_{\ell=0}^{M} A(\ell)A(M-\ell), \quad M \geq 0. \tag{4.61}$$

The solution to (4.61) is

$$A(M) = \frac{\sqrt{2\pi}}{M!}(D_2)^M, \qquad M \geq 0 \tag{4.62}$$

where D_2 is also undetermined. Combining (4.59) – (4.62) yields

$$h_n^k \sim \frac{\sqrt{2\pi}}{M!}(D_2)^M n^{M+1/2}e^{-D_1 n}, \qquad M = 2^k - n = O(1). \tag{4.63}$$

It remains only to determine D_1 and D_2.

These constants are fixed by comparing (4.63) for $M = 0$ and $M = 1$ to (4.51) and (4.55). Setting $M = 0$ in (4.63) and noting that $2^k = n$, we see that (4.51) agrees with (4.63) provided that

$$D_1 = 1 + \log(K_0^*). \tag{4.64}$$

Setting $M = 1$ in (4.63), noting that $2^k = n + 1$ and comparing to (4.55) determines D_2 as

$$D_2 = K_1^* K_0^*/e. \tag{4.65}$$

This complete the analysis of the scale $M = O(1)$.

We have thus obtained the asymptotics of h_n^k for $j = n-k = O(1)$ (the right tail) and for $M = 2^k - n = O(1)$ (the left tail). However, these expansions do not asymptotically match, which indicates there must be at least one additional natural scale in the problem, as was the case for tries. The recurrences (2.2) and (2.4) differ only slightly. Furthermore, as $n \to \infty$, the term $2^{1-n}h_n^{k+1}$ in (2.4) is exponentially small compared to the left side ($= h_n^{k+1}$) of the equation. The two boundary terms in the sum in (2.4) (i.e., $i = 0$ and $i = n$) are absent, but our analysis of the scale $M = O(1)$ shows that they are asymptotically negligible (and in some cases *exactly* equal to zero).

As in the analysis of tries, we consider the scale $n, k \to \infty$ with $\xi = n2^{-k}$ fixed and $0 < \xi < 1$. We set $h_n^k = F(\xi; n) = F(n2^{-k}; n)$ and assume an asymptotic solution of (2.4) in the WKB form

$$F(\xi; n) \sim e^{-n\Phi(\xi)} \left[A(\xi) + \frac{1}{n} A^{(1)}(\xi) + \cdots \right]. \tag{4.66}$$

The calculation is essentially identical to that in Section 3, and we find that

$$A(\xi) = \sqrt{1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)}, \tag{4.67}$$

which expresses A in terms of Φ. Once again it seems that Φ is very sensitive to the initial conditions(s) (2.5), and we cannot analytically determine this function. However, in Section 5 we discuss the numerical calculation of this function; the numerical results also provide partial justification of the "ansatz" (4.66).

We next show that (4.66) can asymptotically match to the expansions on the j and M scales. This will also yield the local behavior of $\Phi(\xi)$ as $\xi \to 0^+$ and $\xi \to 1^-$.

The matching of the M and ξ scales requires that

$$\left. \frac{\sqrt{2\pi}}{M!} (D_2)^M e^{-D_1 n} n^{M+1/2} \right|_{M \to \infty} \sim \left. A(\xi) e^{-n\Phi(\xi)} \right|_{\xi \to 1^-}. \tag{4.68}$$

The left side of (4.68) is easily evaluated. We simply expand $M!$ by Stirling's formula and note that $n/M = \xi/(1-\xi)$, which yields

$$\sqrt{\frac{\xi}{1-\xi}} \exp\left(n \left[\frac{M}{n} \log\left(\frac{n}{M}\right) + \frac{M}{n} + \frac{M}{n}\log(D_2) - D_1 \right] \right)$$

so that the matching condition implies that

$$\Phi(\xi) \sim D_1 + (1-\xi)\log(1-\xi) - (1-\xi)(1 + \log(D_2)), \qquad \xi \to 1 \tag{4.69}$$

and

$$A(\xi) \sim (1-\xi)^{-1/2}, \qquad \xi \to 1. \tag{4.70}$$

Given (4.69), we can also use (4.67) to infer the behavior of $A(\xi)$ as $\xi \to 1$. We have $\Phi'(\xi) \sim -\log(1-\xi)$ and $\Phi''(\xi) \sim 1/(1-\xi)$ with which (4.67) implies (4.70). It follows that Φ is finite at $\xi = 1$ (with $\Phi(1) = D_1$), but its derivate has a logarithmic singularity at $\xi = 1$.

Now consider the matching of the j and ξ scales. This requires that the large j asymptotics of $h_n^k = 1 - L_j(n)$ agrees with the expansion of $Ae^{-n\Phi}$ as $\xi \to 0^+$. We must have $A \to 1$ and $\Phi \to 0$ as $\xi \to 0^+$. The asymptotic matching region between the j and ξ scales will be where the probability mass accumulates as $n \to \infty$. If we let $\xi \to 0^+$ in such a way that $n\Phi(\xi) \to 0$, then $Ae^{-n\Phi} \sim 1 - n\Phi$ and the matching is satisfied provided that

$$n\Phi(\xi)|_{\xi \to 0^+} \sim L_j(n)|_{j \to \infty}. \tag{4.71}$$

We have already computed the right side of (4.71) (cf. (4.11), (4.13) and (4.38)). In (4.71) $L_j(n)$ is understood to be replaced by its large n expansion (cf. (4.11)), which is then evaluated for $j \to \infty$. The matching condition thus becomes

$$n\Phi(\xi)|_{\xi \to 0^+} \sim \frac{n!}{j!} 2^{-(n-j)^2/2} 2^{3(j-n)/2} \frac{1}{2} \rho_0 e^{\varphi(\alpha)} \cdot j^{5/2} \exp\left[-\frac{\log^2(j)}{2\log 2}\right]. \tag{4.72}$$

To evaluate (4.72) we recall that $k = n - j$ and $\log \xi = \log n - k \log 2$. For $n \to \infty$ with $k = o(n)$ we can replace $\alpha = \langle \log_2(j) \rangle = \langle \log_2(n-k) \rangle$ by $\langle \log_2 n \rangle + o(1)$, since $\varphi(\alpha)$ is periodic with period one. Also, by periodicity $\varphi(\langle \log_2 n \rangle) = \varphi(\log_2 n) = \varphi(k + \log_2 \xi) = \varphi(\log_2 \xi)$. By expanding $n!$ and $j!$ in (4.72) by Stirling's formula and rewriting the result in terms ξ, we see that the matching condition is satisfied provided that

$$n\Phi(\xi) \sim \frac{1}{2} \rho_0 e^{\varphi(\log_2(\xi))} n \xi^{3/2} \exp\left(-\frac{\log^2 \xi}{2 \log 2}\right), \quad \xi \to 0^+. \tag{4.73}$$

This yields the behavior of $\Phi(\xi)$ as $\xi \to 0$ in Theorem 3, whose derivation is now complete.

It follows that $\Phi(\xi)$ and all its derivatives vanish as $\xi \to 0^+$. Also, (4.73) shows that the structure of Φ for the PATRICIA model is more complicated than the corresponding function ($= -\phi$) for b-tries, as the latter did not have the oscillatory behavior of the former. The behavior as $\xi \to 1^-$ is similar to that of b-tries, except that the constants D_1 and D_2 differ for the two models.

5 Numerical Results

In this section we discuss some numerical calculations. They will give an idea of the accuracy of the asymptotic formulas for h_n^k for the PATRICIA model, and also will verify some of the assumptions we made in our analysis.

We define, for some integer ℓ,

$$k^* = \begin{cases} \log_2 n & \text{if} \quad n = 2^\ell \\ \lfloor \log_2 n \rfloor + 1 & \text{if} \quad n \neq 2^\ell \end{cases} \tag{5.1}$$

and

$$k_1 = \left\lfloor \log_2 n + \sqrt{2 \log_2 n} - \frac{3}{2} \right\rfloor + 1, \tag{5.2}$$

Table 1: Right-Tail Comparison

n	$j = n - k$	$1 - h_n^k$ (Numerical)	$1 - h_n^k$ (Theorem 3(i))
10	2	$1.782(10^{-7})$	$1.786(10^{-7})$
	3	$4.469(10^{-5})$	$4.571(10^{-5})$
	4	$3.040(10^{-3})$	$3.251(10^{-3})$
	5 $(= n - k_1)$	$7.071(10^{-2})$	$8.322(10^{-2})$
20	2	$2.684(10^{-39})$	$2.684(10^{-39})$
	5	$1.343(10^{-24})$	$1.343(10^{-24})$
	10	$3.463(10^{-8})$	$3.517(10^{-8})$
	14 $(= n - k_1)$	0.2040	0.2676
30	2	$5.176(10^{-99})$	$5.176(10^{-99})$
	5	$2.781(10^{-75})$	$2.781(10^{-75})$
	10	$8.200(10^{-44})$	$8.200(10^{-44})$
	15	$5.411(10^{-21})$	$5.417(10^{-21})$
	20	$2.050(10^{-6})$	$2.108(10^{-6})$
	23 $(= n - k_1)$	0.1288	0.1616

as in Section 2. For a fixed n, h_n^k will satisfy $0 < h_n^k < 1$ for $k^* \le k \le n-2$, and our analysis predicts that

$$h_n^k \to 1 \quad \text{as} \quad n \to \infty \quad \text{if} \quad k \ge k_1, \tag{5.3}$$

$$h_n^k \to 0 \quad \text{as} \quad n \to \infty \quad \text{if} \quad k < k_1 - 1, \tag{5.4}$$

i.e., all the mass concentrates at $k = k_1$ if $\sqrt{2\log_2 n}\langle \log_2 n + \sqrt{2\log_2 n} - 1.5\rangle \to \infty$. If $\sqrt{2\log_2 n}\langle \log_2 n+\sqrt{2\log_2 n}-1.5\rangle = O(1)$ or $\sqrt{2\log_2 n}-\sqrt{2\log_2 n}\langle \log_2 n+\sqrt{2\log_2 n}-1.5\rangle = O(1)$, then the probability mass is concentrated on two points: either $k_1 - 1$ and k_1 or k_1 and $k_1 + 1$, as discussed in Corollary 2.

We first consider the right tail of the distribution, where Theorem 3(i) applies. In Table 1 we compare the exact (numerical) values of $1 - h_n^k$ to the asymptotic formula in Theorem 3(i). To evaluate the latter we computed C_j recursively using (4.14). We consider $n = 10, 20$ and 30. For each value of n, we start with $j = n - k = 2$ and increase j to $n - k_1$. For $j > n - k_1$, $1 - h_n^k$ may become negative, as then we are clearly out off the range of validity of this asymptotic result. Table 1 shows good agreement between asymptotic

Table 2: Left-Tail Comparison

n	k	M	h_n^k (Numerical)	h_n^k (Theorem 3(iii))
8	3	0	$5.062(10^{-2})$	$5.010(10^{-2})$
16	4	0	$5.032(10^{-4})$	$5.006(10^{-4})$
15	4	1	$4.734(10^{-3})$	$4.275(10^{-3})$
32	5	0	$3.544(10^{-8})$	$3.534(10^{-8})$
31	5	1	$6.667(10^{-7})$	$6.341(10^{-7})$
30	5	2	$6.197(10^{-6})$	$5.324(10^{-6})$
64	6	0	$1.248(10^{-16})$	$1.246(10^{-16})$
63	6	1	$4.694(10^{-15})$	$4.579(10^{-15})$
62	6	2	$8.780(10^{-14})$	$8.148(10^{-14})$
61	6	3	$1.089(10^{-12})$	$9.353(10^{-13})$

and exact results. As expected, the further we get into the tail, the more accurate is the asymptotic formula.

Next, we test the left tail approximation from Theorem 3(iii). For a fixed large n, the condition $2^k - n = O(1)$ cannot be satisfied unless n is close to a power of 2. However, for a fixed large k, this condition can be satisfied. It is thus convenient to do the comparisons when n is a power of 2, and then decrease n. In Table 2 we consider $k \in [3, 6]$ and various values of $M = 2^k - n$. For a fixed M we see that the accuracy of the asymptotic result increases with n. Also, as n becomes larger, we can allow for larger values of M and still get good agreement.

The formula that applies for $k \approx k_1$, which is where there is significant mass, is given by (2.21), which corresponds to $A(\xi)e^{-n\Phi(n)}$, with $\Phi(\xi)$ replaced by its small ξ expansion and $A(\xi) \sim 1$. We can refine this by using $A(\xi) = \sqrt{1 + 2\xi\Phi'(\xi) + \xi^2\Phi''(\xi)}$ with $\Phi(\xi)$ computed from (4.73). In Table 3 we compare the exact values of h_n^k to the expression in (2.21) and also the refinement $Ae^{-n\Phi}$ discussed above, for $n = 10, 20, 30, 50, 100, 150, 250$ and 500. We also tabulated $\xi = n2^{-k}$, since (2.21) assumes that ξ is small. For each n we consider $k = k_1 - 1$, k_1 and $k_1 + 1$. When $n = 10$ the probability mass at k_1 is about .49, and this increases to .71 when $n = 150$. Our asymptotic result predicts the values .53 and .73 when $n = 10$ and 150, respectively. Table 3 is consistent with our prediction that the mass accumulates at k_1, but the agreement between the exact and asymptotic result (2.21) is not

Table 3: Central Regime Comparison

n	k	ξ $(= n2^{-k})$	h_n^k (Numerical)	h_n^k (2.21)	h_n^k ((2.17), $\xi \to 0$)
10	4	.63	.43975	.30024	.37067
	5	.31	.92929	.82862	.87673
	6	.16	.99696	.98542	.99145
20	5	.63	.12641	.09015	.11129
	6	.31	.79599	.68661	.72648
	7	.16	.98564	.97105	.97699
30	6	.47	.24414	.16249	.18627
	7	.23	.87115	.80820	.82855
	8	.12	.99283	.98760	.98948
50	7	.39	.23770	.15836	.17423
	8	.20	.87927	.83529	.84681
	9	.10	.99401	.99125	.99210
100	8	.39	.04845	.02508	.02759
	9	.20	.75483	.69772	.70733
	10	.10	.98623	.98259	.98342
150	9	.29	.15798	.10126	.10619
	10	.15	.87248	.84558	.84961
	11	.07	.99491	.99388	.99409
250	10	.24	.18706	.12842	.13209
	11	.12	.89950	.88226	.88426
	12	.06	.99668	.99618	.99627
500	11	.24	.03336	.01649	.01696
	12	.12	.80512	.77838	.78014
	13	.06	.99313	.99238	.99247

Table 4: Probability mass at k_0 and k_1.

n	ℓ	h_n^k (Numerical)
14	0	.6507
	1	.9675
23	0	.6432
	1	.9674
39	0	.6162
	1	.9665
66	0	.6095
	1	.9692
113	0	.6016
	1	.9719

particularly good. However, the refined results are better as long as ξ is reasonable small. Also, in order to see the importance of the left tail approximation in parts (ii) and (iii) of Theorem 3, we need for $k_1 - k^*$ to be fairly large. However, when $n = 150$, $k_1 = 10$ and $k^* = 8$ so that the "left tail" really consists of the two points $k = 8, 9$. In order to have $k_1 - k^*$ as large as, say, 10, n would have to be about 10^{20} (cf. (5.1) and (5.2))!

To better see the convergence of mass to one or two points, it is useful to consider subsequences n_i of n that correspond to β_n nearly constant. To optimize the mass at $k_0 = k_1 - 1$ we want to minimize β_n. From Figure 3 we see that the local minima of $R(n)$ (and β_n) occur at

$$n_i = \left\lfloor 2^{i+5/2-\sqrt{2i+4}} \right\rfloor + 1. \tag{5.5}$$

The first few integers in this subsequence are: $3, 4, 6, 9, 14, 23, 39, 66, 113$ and 195. By using (5.5) in (2.23) and evaluating the result asymptotically, we obtain

$$R(n_i) = \sqrt{2 \log_2 n_i} \beta_{n_i} = \frac{(1 - I_i)\sqrt{2i+4}}{\log 2} 2^{-i-5/2+\sqrt{2i+4}} (1 + O(n_i^{-1})) \tag{5.6}$$

where

$$I_i = \langle 2^{i+5/2-\sqrt{2i+4}} \rangle.$$

It follows that along n_i, $R(n_i)$ is not only bounded, but approaches zero, roughly like $O(n_i^{-1})$.

Thus along this subsequence (2.25) becomes, for $\ell = 0$ and $n \to \infty$

$$\mathsf{Pr}\{\mathcal{H}_n^P \leq \lfloor \log_2 n + \sqrt{2\log_2 n} - 1.5 \rfloor\} \sim \exp\left(-\rho_0 e^{\theta + \Psi(\log_2 n)}\right). \tag{5.7}$$

While the above does not approach a limt as $n \to \infty$, the numerically small value of $\Psi(\cdot)$ allows us to approximate (5.7) by

$$\mathsf{Pr}\{\mathcal{H}_n^P \leq \lfloor \log_2 n + \sqrt{2\log_2 n} - 1.5 \rfloor\} \approx 0.536426\ldots$$

which is obtained by neglecting $\Psi(\cdot)$. This yields the optimal mass at $k_0 = k_1 - 1$ and the remaining mass (≈ 0.464) will be at k_1. In Table 4 we compare the exact values of h_n^k along this subsequence, for $k = k_0$ and k_1. We see that $h_n^{k_0}$ is slowly "converging" to the theoretical value.

By choosing the subsequence $n_i - 1$, we achieve local maxima of β_n and $R(n)$. We can similarly show that $\beta_{n_i - 1} \to 1^-$ and furthermore $\sqrt{2\log_2(n_i - 1)} - R(n_i - 1) \to 0$. Now, we will have about .536 of the mass at k_1 and the remaining .464 at $k_2 = k_1 + 1$. By choosing other subsequences we can achieve any value of mass at k_0 in the range $[0, 0.536\ldots]$ and any value of mass at k_2 in the range $[0, 0.463\ldots]$.

For most n_i the mass will be at a single point k_1. To optimize the convergence we choose n_i so that $\beta_{n_i} \approx 0.5$. This simultaneously avoids the mass at $k_1 - 1$ and $k_1 + 1$. We accomplish this by selecting $\{n_i^*\}_{i=1}^{\infty}$ such that

$$n_i^* = \left\lfloor 2^{i+3-\sqrt{2i+5}} \right\rfloor.$$

In Table 5 we consider a few values of n_i^* and show the exact h_n^k for $k = k_0$ and $k = k_1$. When $n = 10(= n_4^*)$ the mass at k_1 is about 0.49 and it gradually increases to 0.76 when $n = 446(= n_{11}^*)$. We also note that if $\beta_n = 0.5$ and $\ell = 0$, then (2.25) becomes

$$\begin{aligned} \mathsf{Pr}\{\mathcal{H}_n^P \leq \log_2 n + \sqrt{2\log_2 n} - 2\} &\sim \exp\left(-\rho_0 e^{\theta + \Psi(\log_2 n)} 2^{-1/8} 2^{\sqrt{2\log_2 n}/2}\right) \\ &\approx \exp\left(-0.571 \cdot 2^{\sqrt{2\log_2 n}/2}\right) \end{aligned} \tag{5.8}$$

where in the approximation we have neglected the oscillatory term. To make (5.8) less than 0.01 requites $n > 288000$! The last two entries in Table 5 are computed according to the approximate formula (5.8). These results confirm our prediction that all mass concentrates at the one point k_1.

Finally, we discuss the numerical computation of the function $\Phi(\xi)$. We define Φ_{NUM} by

$$\Phi_{NUM}(\xi; k) = \Phi_{NUM}(n2^{-k}; k) \equiv -\frac{1}{n}\log(h_n^k). \tag{5.9}$$

Table 5: Probability mass at k_1 only.

n	ℓ	h_n^k (Numerical)
10	0	.4397
	1	.9293
17	0	.3559
	1	.9049
29	0	.2946
	1	.8904
49	0	.2674
	1	.8904
85	0	.2169
	1	.8839
147	0	.1847
	1	.8847
256	0	.1546
	1	.8863
446	0	.1340
	1	.8916
...	...	...
80226	0	.0161
	1	.9839
1571598	0	.0051
	1	.9948

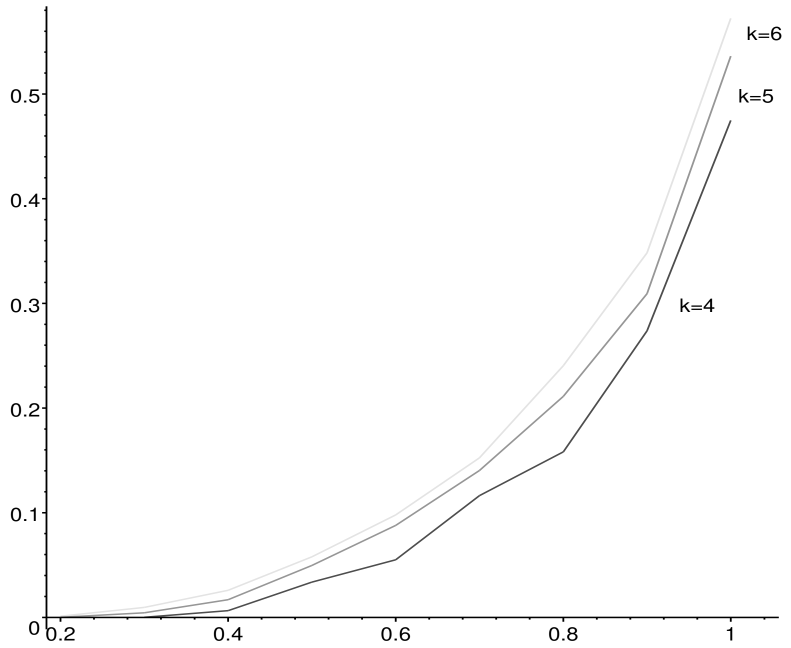

Figure 5: The functions $\Phi_{NUM}(\xi;k)$ versus ξ for $k=4,5$ and 6,

According to our analysis, as $k\to\infty$ for each fixed $0<\xi\le 1$, we should have $\Phi_{NUM}(\xi;k)\to \Phi(\xi)$. We have also analytically computed

$$\Phi(1) = 1+\log(K_0^*) = .61906125\ldots\ . \tag{5.10}$$

In Table 6 we evaluate $\Phi_{NUM}(1;k)$ for $k=4,5,6$ and 7. This sequence certainly appears to be converging to the theoretical value in (5.10). Except for $\xi=1$ we do not have the exact values of $\Phi(\xi)$, however, it is clear from Figure 5 that the sequence of functions $\Phi_{NUM}(\xi;k)$ is indeed converging to a limit. To give an approximation to $\Phi(\xi)$ we plot $\Phi_{NUM}(\xi,k)$ in Figure 5 for $k=4,5$ and 6.

A Proof of the equivalence of (4.24) and (4.28)

In this Appendix we prove the equivalence of the two representations (4.24) and (4.28) for C_j.

The expression in (4.28) may be simplified by closing the contour of integration in the left half-plane and noting that the integrand has simple poles at $s=-1,-2,\ldots$. Denoting the integral by $F(z)$, we have

$$F(z) = \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \frac{z^{-s}\Gamma(s+1)\zeta(s+1)}{s(1-2^{-s})}ds = \sum_{m=1}^{\infty} \frac{z^m\zeta(1-m)}{m(2^m-1)}\frac{(-1)^{m+1}}{(m-1)!}, \tag{A.1}$$

Table 6: Numerical Evaluation of $\Phi(1)$.

n	k	$\Phi_{NUM}(1;k)$	$\Phi(1)$
16	4	.47466	.61906
32	5	.53611	.61906
64	6	.57219	.61906
128	7	.59292	.61906

for z sufficiently small. But, $\zeta(0) = -1/2$ and

$$\zeta(-2n) = 0,\ n = 1, 2, \ldots;\ \zeta(1-2n) = -\frac{B_{2n}}{2n}, n = 1, 2, \ldots$$

where B_n are the Bernoulli numbers (cf. [1]). Thus, (A.1) becomes

$$F(z) = \mathcal{M}^{-1}\left[\frac{\Gamma(s+1)\zeta(s+1)}{s(1-2^{-s})}\right] = -\frac{z}{2} + \sum_{\ell=1}^{\infty} \frac{z^{2\ell}}{(2\ell)!}\frac{B_{2\ell}}{2\ell}\frac{1}{2^{2\ell}-1}.$$

To show the equivalence of (4.24) and (4.28), we represent the Bernoulli numbers as integrals:

$$\frac{B_{2\ell}}{(2\ell)!} = \frac{1}{2\pi i}\oint \frac{1}{e^t - 1}\frac{1}{t^{2\ell}}dt, \qquad 0 < |t| < 2\pi.$$

Using the above, expanding $1/(2^{2\ell}-1) = \sum_{k=0}^{\infty} 2^{-2\ell(k+1)}$, and noting that $B_3 = B_5 = \cdots = 0$, we obtain

$$\begin{aligned} F(z) &= \sum_{k=0}^{\infty}\sum_{m=1}^{\infty} \frac{z^m}{m} 2^{-m(k+1)}\left[\frac{1}{2\pi i}\oint \frac{t^{-m}}{e^t-1}dt\right] \\ &= \frac{1}{2\pi i}\sum_{k=0}^{\infty}\oint \frac{-\log(1 - z2^{-k-1}/t)}{e^t - 1}dt. \end{aligned}$$

Setting $t = z2^{-k-1}\xi$ and integrating by parts yields

$$F(z) = \frac{1}{2\pi i}\sum_{k=0}^{\infty}\oint \frac{1}{\xi(\xi-1)}\log\left(\frac{1-\exp(-z\xi 2^{-k-1})}{z\xi 2^{-k-1}}\right)d\xi \qquad (A.2)$$

where $|\xi| > 1$ on the loop of the integration. For any k the integrand in (A.2) is analytic at $\xi = 0$ and has a simple pole at $\xi = 1$. By evaluating the residue at $\xi = 1$, we see that $\exp[F(z)]$ becomes the same as the infinite product in (4.24). This establishes the equivalence of (4.24) and (4.28).

References

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

[2] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGrew-Hill, New York 1978.

[3] N. Bleistein and R. Handelsman, *Asymptotic Expansions of Integrals*, Dover Publications, New York 1986.

[4] E. R. Canfield, From Recursions to Asymptotics: On Szekeres' Formula for the Number of Partitions, *Electronic J. of Combinatorics*, 4, RR6, 1997.

[5] L. Devroye, A Probabilistic Analysis of the Height of Tries and the complexity of Trie Sort, *Acta Informatica*, 21, 229–237, 1984.

[6] L. Devroye, A Note on the Probabilistic Analysis of Patricia Tries, *Random Structures & Algorithms*, 3, 203–214, 1992.

[7] L. Devroye, A Study of Trie-Like Structures Under the Density Model, *Ann. Appl. Probability*, 2, 402–434, 1992.

[8] A. Ehrenfeucht and J. Mycielski, A Pseudorandom Sequence — How Random Is It? *Amer. Math. Monthly*, 99, 373-375, 1992

[9] J. Fill. H. Mahmoud, and W. Szpankowski, On the Distribution for the Duration of a Randomized Leader Election Algorithm, *Ann. Appl. Probab.*, 6, 1260–1283, 1996.

[10] P. Flajolet, On the Performance Evaluation of Extendible Hashing and Trie Searching, *Acta Informatica*, 20, 345–369, 1983.

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

[12] A. Frieze and W. Szpankowski, Greedy Algorithms for the Shortest Common Superstring Problem that are Asymptotically Optimal, *Algorithmica*, 21, 21-36, 1998.

[13] N. Froman and P. Froman, *JWKB Approximation*, North-Holland, Amsterdam 1965.

[14] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997.

[15] P. Jacquet and M. Régnier, Trie Partitioning Process: Limiting Distributions, Lecture Notes in Computer Science, **214**, 196-210, Springer Verlag, New York 1986.

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

[17] P. Jacquet and W. Szpankowski, Analytic Depoissonization and Its Applications, *Theoretical Computer Science*, 201, No. 1-2, 1–62, 1998.

[18] D. Knuth, *The Art of Computer Programming. Sorting and Searching*, Second Edition, Addison-Wesley, 1998.

[19] C. Knessl and W. Szpankowski, Quicksort Algorithm Again Revisited, *Discrete Mathematics and Theoretical Computer Science*, 3, 43-64, 1999.

[20] P. Lagerstrom, *Matched Asymptotic Expansions: Ideas and Techniques*, Springer-Verlag, New York 1988.

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

[22] A. Odlyzko, Asymptotic Enumeration, in *Handbook of Combinatorics*, Vol. II, (Eds. R. Graham, M. Götschel and L. Lovász), Elsevier Science, 1063-1229, 1995.

[23] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Ann. of Probab.*, 13, 414–427, 1985.

[24] B. Pittel, Path in a Random Digital Tree: Limiting Distributions, *Adv. Appl. Prob.*, 18, 139–155, 1986.

[25] B. Pittel and H. Rubin, How Many Random Questions Are Necessary to Identify n Distinct Objects?, *J. Combin. Theory*, Ser. A., 55, 292–312, 1990.

[26] M. Régnier, On the Average Height of Trees in Digital Searching and Dynamic Hashing, *Inform. Processing Lett.*, 13, 64–66, 1981.

[27] A. Rényi, On Random Subsets of a Finite Set, *Mathematica (Cluj)*, 3, 355–362, 1961.

[28] W. Szpankowski, Patricia Tries Again Revisited, *Journal of the ACM*, 37, 691–711, 1990.

[29] W. Szpankowski, On the Height of Digital Trees and Related Problems, *Algorithmica*, 6, 256–277, 1991.

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

[31] A. Yao, A Note on the Analysis of Extendible Hashing, *Inform. Processing Lett.*, 11, 84–86, 1980.

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

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