# A Note on the Asymptotic Behavior of the Heights in $b$-Tries for $b$ Large

August 2, 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 limiting distribution of the height in a generalized trie in which external nodes are capable to store up to $b$ items (the so called $b$-tries). We assume that such a tree is built from $n$ random strings (items) generated by an unbiased memoryless source. In this paper, we discuss the case when $b$ and $n$ are both large. We shall identify six natural regions of the height distribution that should be compared to three regions obtained for *fixed* $b$. We prove that for most $n$, the limiting distribution is concentrated at the single point $k_1 = \lfloor \log_2(n/b) \rfloor + 1$ as $n, b \to \infty$. We observe that this is quite different than the height distribution for fixed $b$, in which case the limiting distribution is of an extreme value type concentrated around $(1 + 1/b) \log_2 n$. We derive our results by analytic methods, namely generating functions and the saddle point method. We also present some numerical verification of our results.

*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, and Purdue Grant GIFG.

## 1 Introduction

We study here the most basic digital tree 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 $X = x_1 x_2 x_3 \ldots$ 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 (an external node) 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.

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. [5, 9, 11, 17]). In Figure 1 we show an example of a 3-trie constructed over $n = 10$ strings. The $b$-trie is particularly useful in algorithms for extendible hashing in which the capacity of a page or other storage unit is $b$. Also, in lossy compression based on an extension of Lempel-Ziv lossless schemes (cf. [10, 18]), $b$-tries (or more precisely, $b$-suffix trees) are very useful. In these applications, the parameter $b$ is quite large, and may depend on $n$. There are other applications of $b$-tries in computer science, communications and biology. Among these are 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, and genome maps.

In this paper, we consider $b$-tries with a *large* parameter $b$, that may depend on $n$. Such a tree is built over $n$ randomly generated strings of binary symbols. We assume that every symbol is equally likely, thus the strings are emitted by an *unbiased memoryless* source. Our interest lies in establishing the asymptotic distribution of the height, which is the longest path in such a $b$-trie. We also compare our results to those for $b$-tries with fixed $b$, PATRICIA tries (cf. [7, 9, 11, 13]) and digital search trees (cf. [8, 9, 11]).

We now briefly summarize our main results. We obtain asymptotic expansions of the distribution $\Pr\{\mathcal{H}_n \leq k\}$ of the height $\mathcal{H}_n$ for six ranges of $n$, $k$, and $b$ (cf. Theorem 2). This should be compared to three regions of $n$ and $k$ for fixed $b$ (cf. Theorem 1). We shall prove that in the region where most of the probability mass is concentrated, the height

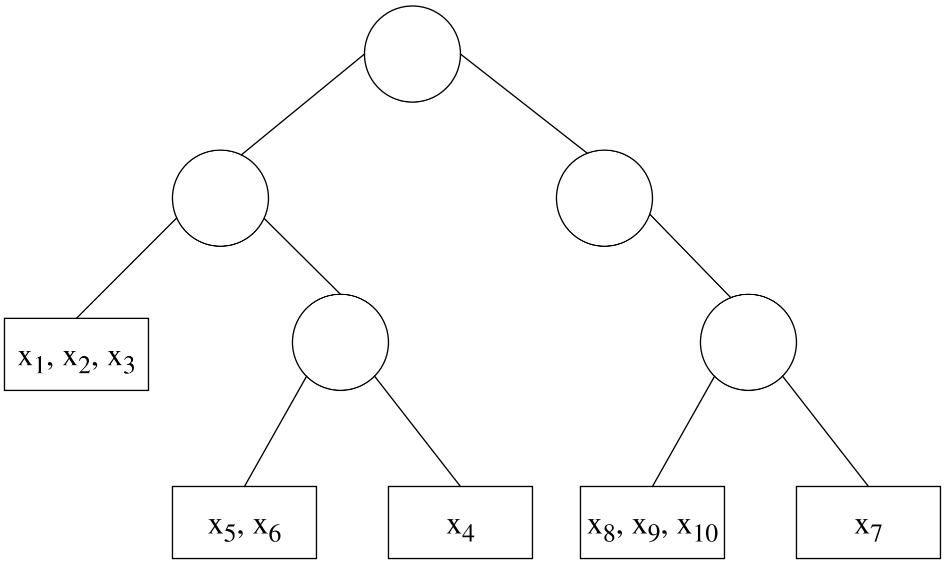


Figure 1: A $b$-trie with $b = 3$ built from the following ten strings: $X_1 = 11000\ldots$ , $X_2 = 11100\ldots$ , $X_3 = 11111\ldots$ , and $X_4 = 1000\ldots$, $X_5 = 10111\ldots$, $X_6 = 10101\ldots$, $X_7 = 00000\ldots$, $X_8 = 00111\ldots$, $X_9 = 00101\ldots$, $X_4 = 00100\ldots$.

distribution can be approximated by (for fixed large $k$ and $n, b \to \infty$)

$$\Pr\{\mathcal{H}_n \le k\} \sim \exp\left(-\frac{2^k}{\sqrt{2\pi}}\frac{e^{-a^2/2}}{a}\right)$$

where $a = \sqrt{b}(1 - n2^{-k}) \to \infty$. This resembles an exponential of a Gaussian distribution. However, a closer look reveals that the asymptotic distribution of the height is concentrated (for fixed large $n$ and $k, b \to \infty$) on the point $k_1 = \lfloor \log_2(n/b) \rfloor + 1$, that is, $\Pr\{\mathcal{H}_n = k_1\} = 1 - o(1)$. This should be contrasted with the height distribution of $b$-tries with fixed $b$, in which cases the limiting distribution is of extreme value type, and is concentrated around $(1+1/b)\log_2 n$. We observe that the height distribution of $b$-tries with large $b$ resembles the height distribution for a PATRICIA trie (cf. [7, 13, 17]). In fact, in [13, 17] the probabilistic behavior of the PATRICIA height was obtained through the height of $b$-tries after taking the limit with $b \to \infty$.

With respect to previous results, Devroye [2] and Pittel [14] established the asymptotic distribution (in the region where most of the probability mass is concentrated) for $b$-tries with fixed $b$ using probabilistic tools. Jacquet and Régnier [6] obtained similar results by analytic methods. We extended these results in [7] to other regimes of $n$ and $k$, but still for fixed $b$. To the best of our knowledge, there are no reported results in literature for large $b$.

The paper is organized as follows. In the next section we present and discuss our main results for $b$-tries for large $b$ (cf. Theorem 2). The proof is delayed until Section 3. It is based on an asymptotic evaluation of a certain integral.

## 2 Summary of Results

We let $\mathcal{H}_n$ be the height of a $b$-trie of size $n$. We denote its probability distribution by

$$h_n^k = \mathsf{Pr}\{\mathcal{H}_n \le k\}. \tag{2.1}$$

This function satisfies the non-linear recurrence

$$h_n^k = \sum_{i=0}^{n} \binom{n}{i} 2^{-n} h_i^{k-1} h_{n-i}^{k-1}, \qquad k \ge 1 \tag{2.2}$$

with the initial condition

$$\begin{aligned} h_n^0 &= 1, \qquad n = 0, 1, \ldots, b; & (2.3)\\ h_n^0 &= 0, \qquad n > b. & (2.4) \end{aligned}$$

By using exponential generating functions, we can easily solve (2.2) and (2.3)-(2.4). Indeed, let us define $H^k(z) = \sum_{n\ge 0} h_n^k \frac{z^n}{n!}$. 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 the following representation of $h_n^k$ as a complex contour integral:

$$h_n^k = \frac{n!}{2\pi i} \oint z^{-n-1} \left[1 + z2^{-k} + \frac{z^2 4^{-k}}{2!} + \cdots + \frac{z^b 2^{-bk}}{b!}\right]^{2^k} dz. \tag{2.5}$$

Here the loop integral is around any closed loop about the origin.

To gain more insight into the structure of this probability distribution, it is useful to evaluate (2.5) in the asymptotic limit $n \to \infty$. In [7] we derived asymptotic formulas that apply for $n$ large with $b$ fixed, for various ranges of $k$. For purposes of comparison, we repeat these results below.

**Theorem 1 (Knessl and Szpankowski 1999)** *The distribution of the height of $b$-tries has the following asymptotic expansions for fixed $b$:*

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

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

(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 also observed that the probability mass is concentrated in the central region when $\xi \to 0$. In particular,

$$\begin{aligned}
\Pr\{\mathcal{H}_n \le k\} &\sim A(\xi) e^{n\phi(\xi)} \sim \exp\left(-\frac{n\xi^b}{(b+1)!}\right), \qquad \xi \to 0 \\
&= \exp\left(-\frac{n^{1+b} 2^{-kb}}{(b+1)!}\right). \qquad (2.6)
\end{aligned}$$

In fact, most of the probability mass is concentrated around $k = (1 + 1/b)\log_2 n + x$ where $x$ is a fixed real number. More precisely:

$$\begin{aligned}
\Pr\{\mathcal{H}_n \le (1 + 1/b)\log_2 n + x\} &= \Pr\{\mathcal{H}_n \le \lfloor (1 + 1/b)\log_2 n + x \rfloor\} \\
&\sim \exp\left(-\frac{1}{(1+b)!} 2^{-bx + b\langle (1+b)/b \cdot \log_2 n + x \rangle}\right), \qquad (2.7)
\end{aligned}$$

where $\langle x \rangle$ is the fractional part of $x$, that is, $\langle x \rangle = x - \lfloor x \rfloor$. Due to the term $\langle \log_2 n \rangle$ the limit of (2.7) does not exit as $n \to \infty$.

We next consider the limit $b \to \infty$. We now find that there are six cases of $(n, k)$ to consider, and we summarize our final results below. The necessity of treating the six cases in Theorem 2 is better understood by viewing the problem as first fixing $k$ and $b$, and then varying $n$ (cf. Section 4).

**Theorem 2** *For $b \to \infty$ the distribution of the height of $b$-tries has the following asymptotic expansions:*

(a) $b, k \to \infty$, $\ell = n - b - 1 \geq 0$, $\ell = O(1)$

$$1 - h_n^k \sim \frac{b^\ell}{\ell!} 2^{-kb}.$$

(b) $b, n \to \infty$, $2^k = O(\sqrt{b})$, $n/b > 1$ *and fixed*

$$1 - h_n^k \sim \frac{1}{\sqrt{2\pi}} 2^{-kb} \sqrt{\frac{n}{b}} \frac{\sqrt{n-b}}{b} \left(\frac{n}{b}\right)^n \left(\frac{n}{b} - 1\right)^{b-n} \exp\left((b-n)2^{-k}\left(1 + 2^{-k-1}\right)\right).$$

(c) $b, n, k \to \infty$, $2^k = O(\sqrt{b})$, $0 < n2^{-k}/b < 1$

$$\begin{aligned} 1 - h_n^k \quad &\sim \quad \frac{n}{\sqrt{2\pi b}} \frac{1}{(b - n2^{-k})} \left(\frac{n}{2^k b}\right)^b \exp(b - n2^{-k}) \\ &\times \quad \exp\left(-\frac{1}{2n}(b - n2^{-k})^2 - \frac{b}{2n^2}(b - n2^{-k})^2\right). \end{aligned}$$

(d) $b, n, k \to \infty$, $2^k = O(\sqrt{b})$, $a \equiv \sqrt{b}(1 - n2^{-k}/b)$ *fixed*

$$h_n^k \sim \frac{K_0}{\sqrt{1 - a(a + \zeta_0)}} \exp(2^k \Psi_0)$$

*where*

$$\begin{aligned} K_0 \quad &= \quad \exp\left[\frac{2^k}{6\sqrt{b}}(a + \zeta_0)(a^2 - a\zeta_0 + 4)\right], \\ \Psi_0 \quad &= \quad \frac{1}{2}(a + \zeta_0)^2 + \log Q(\zeta_0) \\ Q(\zeta_0) \quad &= \quad \frac{1}{\sqrt{2\pi}} \int_{\zeta_0}^{\infty} e^{-x^2/2} dx, \end{aligned}$$

*and* $\zeta_0 = \zeta_0(a)$ *is the solution to the transcendental equation*

$$a + \zeta_0 = \frac{e^{-\zeta_0^2/2}}{\sqrt{2\pi} Q(\zeta_0)}.$$

(e) $b, n, k \to \infty$ *with* $b - n2^{-k} = \gamma$ *fixed*

$$\begin{aligned} h_n^k \quad &\sim \quad \sqrt{\frac{b}{\gamma(1+\gamma)}} \left(\frac{1}{\sqrt{2\pi b}}\right)^{2^k} e^{2^k \varphi(\gamma)}, \\ \varphi(\gamma) \quad &= \quad \gamma \log\left(1 + \frac{1}{\gamma}\right) + \log(1 + \gamma). \end{aligned}$$

(f) $b, n, k \to \infty$ *with* $b2^k - n = j$ *fixed*, $j \geq 0$

$$h_n^k \sim \sqrt{2\pi b 2^k} \left(\frac{1}{\sqrt{2\pi b}}\right)^{2^k} \frac{2^{kj}}{j!}.$$

We observe that for cases (d), (e) and (f), $h_n^k$ is exponentially small, while for cases (a)-(c), $1 - h_n^k$ is exponentially small. From the definition of $\zeta_0$ in part (d), we can easily show that

$$
\begin{aligned}
\zeta_0(a) &= -a + \frac{1}{\sqrt{2\pi}} e^{-a^2/2} + O(e^{-a^2}), \quad a \to +\infty \\
\zeta_0(a) &= \frac{1}{a} - 2a + O(a^3), \qquad\qquad\qquad a \to 0^+
\end{aligned}
\tag{2.8}
$$

We also note that from the definition of a $b$-trie we have $h_n^k = 0$ for $n > b2^k$ and $h_n^k = 1$ for $0 \le n \le b$, $k \ge 0$.

The asymptotic formula for $h_n^k$ in the matching region between (c) and (d) may be obtained by evaluating (d) in the limit $a \to \infty$. Using (2.8) we are led to

$$
h_n^k \sim \exp\left( -\frac{2^k}{\sqrt{2\pi}} \frac{e^{-a^2/2}}{a} \right). \tag{2.9}
$$

This result applies to the limit where $b, n, k \to \infty$ with $a = \sqrt{b}(1 - n2^{-k}/b) \to \infty$ but $n2^{-k}/b \to 1^-$. We note that for fixed large $n$ the condition $a = O(1)$, with $0 < a < \infty$, as $b \to \infty$ may not be satisfied for any $k$. However, for fixed large $b$ and $k$, we can clearly find $n$ so that $a = \sqrt{b}(1 - n2^{-k}/b) = O(1)$ for some range of $n$ (see also numerical studies in Section 4). The expansion (2.9) applies when $n, b$ and $k$ are such that $h_n^k$ is neither close to 0 nor to 1.

The result (2.9) has roughly the form of an exponential of a Gaussian, and it should be contrasted with the double exponential in (2.6), which applies for $b$ fixed. The large $b$ result is somewhat similar to the corresponding one for PATRICIA trees analyzed by us in [7] and digital search trees discussed in [8].

Next, we apply Theorem 2 for a fixed (large) $b$ and $n$ and $k$ vary. We first define

$$
k_0 = \begin{cases} \lfloor \log_2(n/b) \rfloor & \text{if} \quad n/b = \text{power of } 2 \\ \lfloor \log_2(n/b) \rfloor + 1 & \text{if} \quad n/b \neq \text{power of } 2 \end{cases}
$$

and note that $h_n^k = 0$ for $k < k_0$. We furthermore set

$$
k = \lfloor \log_2(n/b) \rfloor + \ell = \log_2(n/b) + \ell - \beta
$$

where $\beta = \langle \log_2(n/b) \rangle$ (as before $\langle \cdot \rangle$ denotes the fractional part). If $n/b$ is a power of 2 then $\beta = 0$ and for $\ell = 0$ part (f) of Theorem 2 yields (with $j = 0$)

$$
h_n^{k_0} \sim \sqrt{2^{k_0}} \left( \frac{1}{\sqrt{2\pi b}} \right)^{2^{k_0} - 1}, \qquad 2^{k_0} = n/b
$$

which is asymptotically small. On the other hand if $\beta = 0$ and $\ell = 1$ then

$$a = \sqrt{b}(1 - n2^{-k}/b) = \sqrt{b}(1 - 2^{\beta-1}) = \frac{1}{2}\sqrt{b}$$

which is large, so that $h_n^{k_0+1} \sim 1$. This shows that when $n/b$ is a power of 2, all the mass accumulates at $k_0 + 1 = \log_2(n/b) + 1$.

When $n/b$ is not a power of 2 (with $\ell = 1, 2, \ldots$) and we consider a fixed $\beta$ ($0 < \beta < 1$), then we can easily show that $j, \gamma$ and $a$ are all asymptotically large, so that parts (d)-(f) of Theorem 2 do not apply, and we must use part (c) (or the intermediate result in (2.9)) to compute $h_n^k$.[1] We thus have $h_n^{k_0-1} = 0$ and $h_n^{k_0} \sim 1$ so that the mass accumulates at $k = k_0 = \lfloor \log_2(n/b) \rfloor + 1$.

We summarize this analysis in the following corollary.

**Corollary 1** *For any fixed $0 \le \beta < 1$ and $n, b \to \infty$, the asymptotic distribution of the b-trie height is concentrated on the one point $k_1 = \lfloor \log_2(n/b) \rfloor + 1$, that is,*

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

*as $n \to \infty$.*

## 3 Derivation of Results

We establish the six parts of Theorem 2. Since the analysis involves a routine use of the saddle point method (cf. [1, 12]), we only give the main points of the calculations.

We first note that

$$1 + z2^{-k} + \cdots + \frac{z^b 2^{-kb}}{b!} = e^{z2^{-k}} \int_{z2^{-k}}^{\infty} e^{-w} \frac{w^b}{b!} dw = e^{z2^{-k}} \left[ 1 - \int_0^{z2^{-k}} e^{-w} \frac{w^b}{b!} dw \right]. \qquad (3.1)$$

It will thus prove useful to have the asymptotic behavior of the integral(s) in (3.1), and this we summarize below.

**Lemma 1** *We let*

$$I = I(A, b) = \frac{1}{b!} \int_0^A e^{b \log w - w} dw = \frac{e^{-b} b^{b+1}}{b!} \int_0^{A/b} e^{b(\log u - u + 1)} du.$$

*Let $\alpha = b/A$. Then, the asymptotic expansions of $I$ are as follows:*

(i) $b, A \to \infty$, $\alpha = b/A > 1$

$$I = e^{-A} \frac{A^b}{b!} \left[ \frac{1}{b/A - 1} - \frac{b}{A^2} \frac{1}{(b/A - 1)^3} + O(A^{-2}) \right].$$

[1] We should point, however, that if we consider a sequence of $n, b$ such that $\beta \to 1^-$, then the conditions where parts (d) and (e) of Theorem 2 are valid may be satisfied.

(ii) $b, A \to \infty,\ b/A < 1$

$$I = 1 - e^{-A}\frac{A^b}{b!}\left[\frac{1}{1-b/A} - \frac{b}{A^2}\frac{1}{(1-b/A)^3} + O(A^{-2})\right].$$

(iii) $b, A \to \infty,\ A - b = \sqrt{b}B,\ B = O(1)$

$$\begin{aligned} I &= \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{B} e^{-x^2/2}dx - \frac{1}{3\sqrt{b}}(B^2+2)e^{-B^2/2}\right. \\ &+ \left.\frac{1}{b}\left(-\frac{B^5}{18} - \frac{B^3}{36} - \frac{B}{12}\right)e^{-B^2/2} + O(b^{-3/2})\right). \end{aligned}$$

**Proof**. To establish Lemma 1 we note that $I$ is a Laplace-type integral [1, 12]. Setting $f(u) = \log u - u + 1$ we see that $f$ is maximal at $u = 1$. For $A/b < 1$ we have $f'(u) > 0$ for $0 < u \le A/b$ and thus the major contribution to the integral comes from the upper endpoint (more precisely, from $u = A/b - O(b^{-1})$). Then, the standard Laplace method yields part (i) of the Lemma 1. If $A/b > 1$ we write $\int_0^{A/b}(\cdots) = \int_0^\infty(\cdots) - \int_{A/b}^\infty(\cdots)$, evaluate the first integral exactly and use Laplace's method on the second integral. Now $f'(u) < 0$ for $u \ge A/b$ and the major contribution to the second integral is from the lower endpoint. Obtaining the leading two terms leads to (ii) in the Lemma 1.

To derive part (iii), we scale $A - b = \sqrt{b}B$ to see that the main contribution will come from $u - 1 = O(b^{-1/2})$. We thus set $u = 1 + x/\sqrt{b}$ and obtain

$$\begin{aligned} I &= \frac{e^{-b}b^{b+1}}{b!}\int_{-\sqrt{b}}^{B}\exp\left(b\left[\log\left(1+\frac{x}{\sqrt{b}}\right) - \frac{x}{\sqrt{b}}\right]\right)\frac{dx}{\sqrt{b}} \\ &= \frac{b^b\sqrt{b}e^{-b}}{b!}\int_{-\infty}^{B} e^{-x^2/2}\left[1 + \frac{x^3}{3\sqrt{b}} + \frac{1}{b}\left(-\frac{x^4}{4} + \frac{x^6}{18}\right) + O(b^{-3/2})\right]dx. \end{aligned} \tag{3.2}$$

Evaluating explicitly the integrals in (3.2) and using Stirling's formula in the form $b! = \sqrt{2\pi b}b^b e^{-b}(1 + (12b)^{-1} + O(b^{-2}))$, we obtain part (iii) of the Lemma. ∎

We return to (2.5) and first consider the limit $b \to \infty$ with $n - b - 1 = \ell$ fixed. Now we have

$$2^k \log\left(1 - \int_0^{z2^{-k}} e^{-w}\frac{w^b}{b!}dw\right) \sim -\frac{z^{b+1}}{(b+1)!}2^{-kb}$$

which when used in (2.5) yields

$$\begin{aligned} 1 - h_n^k &= \frac{n!}{2\pi i}\oint \frac{e^z}{z^{n+1}}\left(1 - \exp\left[2^k\log\left(1 - \int_0^{z2^{-k}} e^{-w}\frac{w^b}{b!}dw\right)\right]\right)dz \\ &\sim \frac{n!}{2\pi i}\oint e^z z^{b-n}\frac{2^{-kb}}{(b+1)!}dz \\ &= \frac{n!}{(n-b-1)!}\frac{2^{-kb}}{(b+1)!}. \end{aligned}$$

Setting $n = b+1+\ell$ and noting that $(b+1+\ell)!/(b+1)! \sim b^\ell$ we obtain part (a) of Theorem 2.

Next we consider the limit $n, b \to \infty$ with $n/b > 1$ and $2^k = O(\sqrt{b})$. We first use Lemma 1(i) and Stirling's formula to approximate

$$\begin{aligned} 1 \quad &- \quad \exp\left[2^k \log\left(1 - \frac{b^{b+1}}{b!}\int_0^x e^{-b\tau}\tau^b d\tau\right)\right] \qquad (3.3) \\ &= \quad 2^k\sqrt{\frac{b}{2\pi}}(1+O(b^{-1}))\int_0^x e^{b(\log\tau+1-\tau)}dt \\ &\sim \quad \frac{2^k}{\sqrt{2\pi b}}\frac{x^{b+1}}{1-x}e^{b(1-x)}. \end{aligned}$$

Using (3.3) and setting $z = b2^k x$ in (2.5) we obtain

$$1 - h_n^k \sim \frac{n!}{(2^k b)^n}\frac{2^k}{\sqrt{2\pi b}}\frac{1}{2\pi i}\oint e^{(b-n)\log x}e^{2^k bx}e^{b(1-x)}\frac{dx}{1-x}. \qquad (3.4)$$

The integral in (3.4) is easily evaluated by setting $x = 2^{-k}y$ and noting that there is a saddle point where

$$\frac{d}{dy}[(b-n)\log y + by] = \frac{b-n}{y} + b \qquad (3.5)$$

so that $y = y_0 \equiv n/b - 1$. Expanding the integrand in (3.4) near the saddle point and noting that the steepest descent directions are $\arg(y - y_0) = \pm\pi/2$, we obtain

$$\begin{aligned} \frac{1}{2\pi i}\oint e^{(b-n)\log y + by}e^{-2^{-k}by}dy \quad &\sim \quad \left(\frac{n}{b}-1\right)^{b-n} e^{(n-b)(1-2^{-k})} \qquad (3.6) \\ &\times \quad \frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\exp\left(-2^{-k}b\xi + \frac{b^2}{2(n-b)}\xi^2\right)d\xi \\ &= \quad \left(\frac{n}{b}-1\right)^{b-n} e^{(n-b)(1-2^{-k})}\frac{1}{\sqrt{2\pi}}\frac{\sqrt{n-b}}{b}\exp\left(-\frac{1}{2}4^{-k}(n-b)\right). \end{aligned}$$

Using (3.6) in (3.4) immediately yields part (b) of Theorem 2, after we approximate $n!$ by Stirling's formula.

Now consider $b, n, k$ all large with $2^k = O(\sqrt{b})$ and $n2^{-k}/b$ fixed, with $0 < n2^{-k}/b < 1$. We again use approximation (3.3), which applies for a fixed $x \in (0, 1)$. With this scaling we have $n = O(b^{3/2})$ and to leading order the equation locating the saddle point(s) is

$$\frac{d}{dx}\left(2^k bx - n\log x\right) = 2^k b - \frac{n}{x} = 0$$

so the saddle is at $x = x_0 \equiv n2^{-k}/b$. Expanding the integrand in (3.4) about $x_0$ using

$$2^k bx \quad - \quad n\log x + b(1 - x + \log x)$$

$$
\begin{aligned}
&= n - n\log(n2^{-k}/b) + b(1 - x_0 + \log x_0) + b\left(\frac{1}{x_0} - 1\right)(x - x_0) \\
&+ \frac{n-b}{2x_0^2}(x - x_0)^2 + O(n(x - x_0)^3)
\end{aligned}
$$

we are led to

$$
\begin{aligned}
1 - h_n^k &\sim \frac{n!}{(b2^k)^n}\frac{2^k}{\sqrt{2\pi b}}\frac{1}{1 - n2^{-k}/b}\exp\Big(n - n\log(n2^{-k}/b) + b(1 - x_0 + \log x_0)\Big) \\
&\times \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{b(1/x_0 - 1)\eta} e^{(n-b)\eta^2/2x_0^2} d\eta. \qquad (3.7)
\end{aligned}
$$

We explicitly evaluate the integral in (3.7), use Stirling's formula to approximate $n!$ and note that in the indicated limit $1/(n-b) = n^{-1}[1+b/n+b^2/n^2+O(b^{-3/2})]$. Then Theorem 2 (c) follows. Note that as $x_0 \to 1$ the approximation (3.3) ceases to be valid near the saddle point. This completes our consideration of the "right tail" of the distribution, where $h_n^k$ is asymptotically close to one.

We proceed to analyze the left tail of the distribution. First, we consider the limit $b, n, k \to \infty$ with $b2^k - n = j$ fixed, and $j \geq 0$. We use part (ii) of Lemma 1 to approximate (3.1). Thus,

$$
\begin{aligned}
z + 2^k \log\left(\int_{z2^{-k}}^{\infty} e^{-w}\frac{w^b}{b!}dw\right) &\sim 2^k b\log(z2^{-k}) - 2^k\log(b!) \qquad (3.8) \\
&- 2^k\log\left(1 - \frac{b}{z2^{-k}}\right).
\end{aligned}
$$

We furthermore scale $z = 4^k bt$ and then (2.5) with (3.8) becomes

$$
\begin{aligned}
h_n^k &\sim n! e^{-2^k\log(b!)} e^{2^k b\log(2^{-k})}\frac{1}{2\pi i}\oint z^{j-1}\exp\left(-2^k\log\left(1 - \frac{b}{z2^{-k}}\right)\right)dz \qquad (3.9) \\
&\sim n!(4^k b)^j e^{-2^k\log(b!)} e^{2^k b\log(2^{-k})}\frac{1}{2\pi i}\oint t^{j-1}e^{1/t}d\tau \\
&= \frac{n!}{j!}(4^k b)^j e^{-2^k\log(b!)} e^{2^k b\log(2^{-k})}.
\end{aligned}
$$

Using Stirling's formula to approximate $n!$ and $b!$ and replacing $n$ by $b2^k - j$, we see that (3.9) is asymptotically equivalent to Theorem 2(f).

Next we take $b, n, k$ large with $b - n2^{-k} = \gamma$ fixed. We may still use the approximation (3.8) We now set $z = 2^k b\tau$ and obtain from (2.5) and (3.8)

$$
\begin{aligned}
h_n^k &\sim n!\left(\frac{1}{2^k b}\right)^{n - 2^k b} e^{-2^k\log(b!)}\left(\frac{1}{2^k}\right)^{2^k b} \times J \qquad (3.10) \\
J &= \frac{1}{2\pi i}\oint e^{(2^k b - n)\log\tau - 2^k\log(1 - 1/\tau)}\frac{d\tau}{\tau}.
\end{aligned}
$$

The integral $J$ is easily evaluated by the saddle point method. The saddle point equation is

$$\frac{d}{d\tau}[(2^k b - n)\log\tau - 2^k\log(1-1/\tau)] = 0$$

so there is a saddle at $\tau = \tau_0 \equiv 1 + 1/(b - n2^{-k}) = 1 + 1/\gamma$. Then the standard leading order estimate for $J$ is

$$J \sim \frac{1}{\sqrt{2\pi}}\exp\left[(2^k b - n)\log\left(1+\frac{1}{\gamma}\right) + 2^k\log(1+\gamma)\right]\frac{1}{\sqrt{(2^k b - n)(1 + b - n2^{-k})}}. \quad (3.11)$$

Using (3.11) in (3.10) along with Stirling's formula, and writing the result in terms of $b, k$ and $\gamma$, we obtain Theorem 2(e).

Finally we consider $b, n, k$ large with $a = \sqrt{b}(1 - n2^{-k}/b)$ fixed. Now we must use part (iii) of Lemma 1 to approximate the integrand in (2.5). Setting $B = (z2^{-k} - b)/\sqrt{b}$ and using Lemma 1(iii) we obtain

$$\begin{aligned}\log(1-I) &= \log\left[\frac{1}{\sqrt{2\pi}}\int_B^\infty e^{-x^2/2}dx - \frac{1}{\sqrt{2\pi}}\frac{B^2+2}{3\sqrt{b}}e^{-B^2/2} + O(b^{-1})\right] \\ &= \log\left(\frac{1}{\sqrt{2\pi}}\int_B^\infty e^{-x^2/2}dx\right) + \frac{B^2+2}{3\sqrt{b}}\frac{e^{-B^2/2}}{\int_B^\infty e^{-x^2/2}dx} + O(b^{-1}).\end{aligned} \quad (3.12)$$

Setting $\zeta = (z2^{-k} - b)/\sqrt{b}$ we find that

$$\begin{aligned}n!e^z z^{-n} &= \exp\left[2^k(b+\sqrt{b}\zeta) - n\log(2^k b) - n\log\left(1+\frac{\zeta}{\sqrt{b}}\right)\right]n! \\ &= \sqrt{2\pi n}\exp\left[2^k\frac{(a+\zeta)^2}{2} + \frac{2^k}{\sqrt{b}}\left(\frac{a^3}{6} - \frac{a\zeta}{2} - \frac{\zeta^3}{3}\right) + O\left(\frac{2^k}{b}\right)\right].\end{aligned} \quad (3.13)$$

Here we have again used Stirling's formula and recalled that $n = 2^k b(1 - a/\sqrt{b})$. Using (3.12) and (3.13), (2.5) becomes

$$h_n^k = \frac{\sqrt{2\pi n}}{2\pi i}\frac{1}{\sqrt{b}}\oint K(\zeta; b)e^{2^k\Psi(\zeta)}d\zeta \quad (3.14)$$

where

$$\Psi(\zeta) = \frac{1}{2}(a+\zeta)^2 + \log\left(\frac{1}{\sqrt{2\pi}}\int_\zeta^\infty e^{-x^2/2}dx\right)$$

and

$$\begin{aligned}K(\zeta; b) = \exp&\left(\frac{2^k}{\sqrt{b}}\left(\frac{a^3}{6} - \frac{a\zeta^2}{2} - \frac{\zeta^3}{3} + \frac{(\zeta^2+2)e^{-\zeta^2/2}}{3\int_\zeta^\infty e^{-x^2/2}dx}\right)\right) \\ &\times[1 + O(b^{-1/2}, 2^k b^{-1})].\end{aligned}$$

For $k \to \infty$ in such a way that $a$ is fixed and $2^k/b \to 0$, we evaluate (3.14) by the saddle point method. The equation locating the saddle points is $\Psi'(\zeta) = 0$, i.e.,

$$a + \zeta = \frac{e^{-\zeta^2/2}}{\int_\zeta^\infty e^{-x^2/2} dx}. \tag{3.15}$$

This defines $\zeta = \zeta_0(a)$, which satisfies $\zeta_0 \to -\infty$ as $a \to +\infty$ and $\zeta_0 \to +\infty$ as $a \to 0^+$. We note that $n2^{-k}/b \sim 1$ and, in view of (3.15),

$$\begin{aligned}
\Psi''(\zeta_0) &= 1 + \frac{\zeta_0 e^{-\zeta_0^2/2}}{\int_{\zeta_0}^\infty e^{-x^2/2} dx} - \frac{e^{-\zeta_0^2}}{\left(\int_{\zeta_0}^\infty e^{-x^2/2} dx\right)^2} \\
&= 1 - a^2 - a\zeta_0.
\end{aligned}$$

Then the standard Laplace estimate of (3.14) leads to part (d) of Theorem 2.

To summarize the calculations, we have evaluated (2.5) by the saddle point method. The saddle point that determined the asymptotics of the integral corresponded to the scaling $z = O(b)$, $z = O(b2^k)$, $z = b2^k + O(\sqrt{b})$, $z = O(b2^k)$ and $z = O(b4^k)$ for cases (b), (c), (d), (e) and (f), respectively. For the different ranges of $z$, we needed to use different approximations to the integral $I$ in Lemma 1.

## 4 Numerical Studies

We determine the numerical accuracy of the results in Theorem 2, and also demonstrate the necessity of treating the six different scales. To do so, it is best to fix $b$ and $k$, and vary $n$. We consider the range $b + 1 \le n \le b2^k$, since otherwise $h_n^k = 1$ or $h_n^k = 0$. We note that as we increase $n$, we gradually move from case (a) to case (f) of Theorem 2. We also comment that for a fixed large $b$ and $n$, the conditions under which (d)–(f) apply may not be satisfied for any $k$. However, for a fixed large $b$ and $k$, we can always find a range of $n$ such that each of the parts of Theorem 2 apply.

In Table 1 we consider $b = 16$ and $k = 2$. We thus have $2^k = \sqrt{b}$ so that the condition $2^k = O(\sqrt{b})$, which appears in parts (b)–(d), is (numerically) satisfied. Table 1 gives the exact values of $1 - h_n^k$ and the approximations from Theorem 2, parts (a)–(c). The part (a) approximation is denoted by $1 - h_n^k$ (a), etc. We see that when $n = 17$, (a) is a better approximation than (b), but (b) is superior when $n \ge 19$. We also see that (c) gradually becomes a better approximation than (b), though the former always over estimates the true value by a factor of about 2.

In Table 2 we retain $b = 16$ and $k = 2$, but now take $46 \le n \le 64$. We tabulate the exact $h_n^k$ along with the asymptotic results in parts (d)–(f) of Theorem 2. We also give the

Table 1: $b = 16$, $k = 2$

| $n$ | $1 - h_n^k$ (exact) | $1 - h_n^k$ (a) | $1 - h_n^k$ (b) | $1 - h_n^k$ (c) |
|---|---|---|---|---|
| 17 | .233 $(10^{-9})$ | .233 $(10^{-9})$ | .203 $(10^{-9})$ | |
| 18 | .320 $(10^{-8})$ | .373 $(10^{-8})$ | .265 $(10^{-8})$ | |
| 19 | .232 $(10^{-7})$ | .298 $(10^{-7})$ | .187 $(10^{-7})$ | |
| 20 | .118 $(10^{-6})$ | .159 $(10^{-6})$ | .936 $(10^{-7})$ | |
| 21 | .475 $(10^{-6})$ | .636 $(10^{-6})$ | .369 $(10^{-6})$ | |
| 22 | .160 $(10^{-5})$ | .203 $(10^{-5})$ | .122 $(10^{-5})$ | |
| 23 | .469 $(10^{-5})$ | .543 $(10^{-5})$ | .353 $(10^{-5})$ | |
| 24 | .123 $(10^{-4})$ | .124 $(10^{-4})$ | .913 $(10^{-5})$ | |
| 26 | .652 $(10^{-4})$ | .441 $(10^{-4})$ | .469 $(10^{-4})$ | |
| 28 | .263 $(10^{-3})$ | .103 $(10^{-3})$ | .183 $(10^{-3})$ | .467 $(10^{-3})$ |
| 30 | .863 $(10^{-3})$ | | .582 $(10^{-3})$ | .148 $(10^{-2})$ |
| 32 | .240 $(10^{-2})$ | | .157 $(10^{-2})$ | .405 $(10^{-2})$ |
| 34 | .585 $(10^{-2})$ | | .368 $(10^{-2})$ | .975 $(10^{-2})$ |
| 36 | .127 $(10^{-1})$ | | .773 $(10^{-2})$ | .211 $(10^{-2})$ |
| 38 | .253 $(10^{-1})$ | | .147 $(10^{-1})$ | .420 $(10^{-1})$ |
| 40 | .462 $(10^{-1})$ | | .259 $(10^{-1})$ | .774 $(10^{-1})$ |
| 42 | .790 $(10^{-1})$ | | .423 $(10^{-1})$ | .134 |
| 44 | .127 | | .650 $(10^{-1})$ | .220 |
| 46 | .193 | | .943 $(10^{-1})$ | .346 |
| 48 | .278 | | .130 | .524 |
| 50 | .383 | | .172 | .773 |

Table 2: $b = 16,\ k = 2$

| $n$ | $h_n^k$ (exact) | $(a)$ | $h_n^k$ (d) | $(\gamma)$ | $h_n^k$ (e) | $(j)$ | $h_n^k$ (f) |
|---|---|---|---|---|---|---|---|
| 46 | .807 | (1.125) | .960 | | | | |
| 48 | .722 | (1.000) | .873 | | | | |
| 50 | .617 | (.875) | .763 | | | | |
| 52 | .497 | (.750) | .635 | | | | |
| 54 | .370 | (.563) | .497 | (2.50) | .581 | | |
| 56 | .247 | (.500) | .359 | (2.00) | .335 | | |
| 58 | .142 | (.375) | .238 | (1.50) | .171 | | |
| 60 | .643 $(10^{-1})$ | | | (1.00) | .716 $(10^{-1})$ | | |
| 61 | .378 $(10^{-1})$ | | | (.75) | .412 $(10^{-1})$ | (3) | .211 $(10^{-1})$ |
| 62 | .193 $(10^{-1})$ | | | (.50) | .208 $(10^{-1})$ | (2) | .159 $(10^{-1})$ |
| 63 | .778 $(10^{-2})$ | | | (.25) | .864 $(10^{-2})$ | (1) | .794 $(10^{-2})$ |
| 64 | .195 $(10^{-2})$ | | | | | (0) | .198 $(10^{-2})$ |

corresponding values of $a = \sqrt{b}(1 - n2^{-k}/b)$, $\gamma = b - n2^{-k}$ and $j = b2^k - n$, since these results assume that $a$, $\gamma$ and $j$ are $O(1)$, respectively. When $n = 64$, approximation (f) is accurate to within 2%. When $n = 63$, (f) is more accurate than (e), but (e) becomes superior for $n \leq 62$. When $n$ is further decreased to $n = 54$, (d) becomes more accurate than (e). We also recall that when $h_n^k$ is not close to either 0 or 1, then part (d) applies.

In Tables 3 and 4 we increase $b$ and $k$ to $b = 64$ and $k = 3$ (thus retaining $2^k = \sqrt{b}$). In Table 3 we consider $1 - h_n^k$ for cases (a)–(c) and in Table 4 we give $h_n^k$ for cases (d)–(f) (again tabulating the values of $a$, $\gamma$ and $j$). When $n = 65 = b+1$, (a) is superior to (b), but (b) is the better approximation for $n \geq 66$. Approximation (c) becomes better than (b) for some $n$ in the range [150,200]. Table 4 considers $400 \leq n \leq 512 = b2^k$ and demonstrates the transition between cases (d) and (e) and then (e) and (f). In general, the results in Tables 3 and 4 are more accurate than those in Tables 1 and 2, as one would expect, since the asymptotics apply for $b \to \infty$.

These data also suggest that in some cases (especially (c)), it may be desirable to calculate some of the higher order terms in the asymptotic series. In most of the cases, these are likely to be of order $O(b^{-1/2})$ relative to the leading term, for $2^k = O(\sqrt{b})$. The overall accuracy of the asymptotic results is also consistent with $O(b^{-1/2})$ error terms. Finally, we

Table 3: $b = 64,\ k = 3$

| $n$ | $1 - h_n^k$ (exact) | $1 - h_n^k$ (a) | $1 - h_n^k$ (b) | $1 - h_n^k$ (c) |
|---|---|---|---|---|
| 65 | $.159\ (10^{-57})$ | $.159\ (10^{-57})$ | $.154\ (10^{-57})$ | |
| 66 | $.922\ (10^{-56})$ | $.102\ (10^{-55})$ | $.853\ (10^{-56})$ | |
| 67 | $.271\ (10^{-54})$ | $.326\ (10^{-54})$ | $.247\ (10^{-54})$ | |
| 68 | $.538\ (10^{-53})$ | $.696\ (10^{-53})$ | $.487\ (10^{-53})$ | |
| 69 | $.814\ (10^{-52})$ | $.111\ (10^{-51})$ | $.732\ (10^{-52})$ | |
| 70 | $.100\ (10^{-50})$ | $.142\ (10^{-50})$ | $.895\ (10^{-51})$ | |
| 100 | $.176\ (10^{-32})$ | | $.149\ (10^{-32})$ | $.326\ (10^{-33})$ |
| 150 | $.564\ (10^{-19})$ | | $.437\ (10^{-19})$ | $.329\ (10^{-19})$ |
| 200 | $.118\ (10^{-11})$ | | $.824\ (10^{-12})$ | $.109\ (10^{-11})$ |
| 250 | $.468\ (10^{-7})$ | | $.288\ (10^{-7})$ | $.511\ (10^{-7})$ |
| 300 | $.522\ (10^{-4})$ | | $.275\ (10^{-4})$ | $.611\ (10^{-4})$ |
| 350 | $.583\ (10^{-2})$ | | | $.718\ (10^{-2})$ |
| 400 | .130 | | | .177 |

Table 4: $b = 64,\ k = 3$

| $n$ | $h_n^k$ (exact) | $(a)$ | $h_n^k$ (d) | $(\gamma)$ | $h_n^k$ (e) | $(j)$ | $h_n^k$ (f) |
|---|---|---|---|---|---|---|---|
| 400 | .870 | (1.75) | .924 | | | | |
| 420 | .690 | (1.44) | .743 | | | | |
| 440 | .416 | (1.13) | .454 | | | | |
| 460 | .145 | (.81) | .164 | | | | |
| 480 | $.166\ (10^{-1})$ | (.48) | $.204\ (10^{-1})$ | (4) | $.337\ (10^{-1})$ | | |
| 500 | $.980\ (10^{-4})$ | (.17) | $.202\ (10^{-3})$ | (1.50) | $.111\ (10^{-3})$ | | |
| 508 | $.702\ (10^{-6})$ | | | (.500) | $.733\ (10^{-6})$ | (5) | $.370\ (10^{-6})$ |
| 509 | $.256\ (10^{-6})$ | | | (.375) | $.268\ (10^{-6})$ | (4) | $.185\ (10^{-6})$ |
| 510 | $.772\ (10^{-7})$ | | | (.250) | $.816\ (10^{-7})$ | (2) | $.694\ (10^{-7})$ |
| 511 | $.172\ (10^{-7})$ | | | (.125) | $.188\ (10^{-7})$ | (1) | $.174\ (10^{-7})$ |
| 512 | $.215\ (10^{-8})$ | | | | | (0) | $.217\ (10^{-8})$ |

comment that by calculating higher order terms in the expansions in Lemma 1, it may be possible to relax the condition $2^k = O(\sqrt{b})$, that appears in some parts of Theorem 2.

## References


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

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

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

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

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

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

[7] 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`.

[8] C. Knessl and W. Szpankowski, Asymptotic Behavior of the Height in a Digital Search Tree and the Longest Phrase of the Lempel-Ziv Scheme, Purdue University CSD-TR–99-019; also available on `http://www.cs.purdue.edu/homes/spa/reports.html`.

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

[10] T. Łuczak and W. Szpankowski, A Suboptimal Lossy Data Compression Based in Approximate Pattern Matching, *IEEE Trans. Information Theory*, 43, 1439–1451, 1997.

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

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

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

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

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

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

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

[18] E.H. Yang, and J. Kieffer, On the Performance of Data Compression Algorithms Based upon String Matching, *IEEE Trans. Information Theory*, 44, 47-65, 1998.