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

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

Submitted August 2, 1999; Accepted August 1, 2000.

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 five 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.

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 $X = x_1x_2x_3\ldots$ 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 extendible 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.

[*] 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, Purdue Grant GIFG, and contract 1419991431A from sponsors of CERIAS at Purdue.

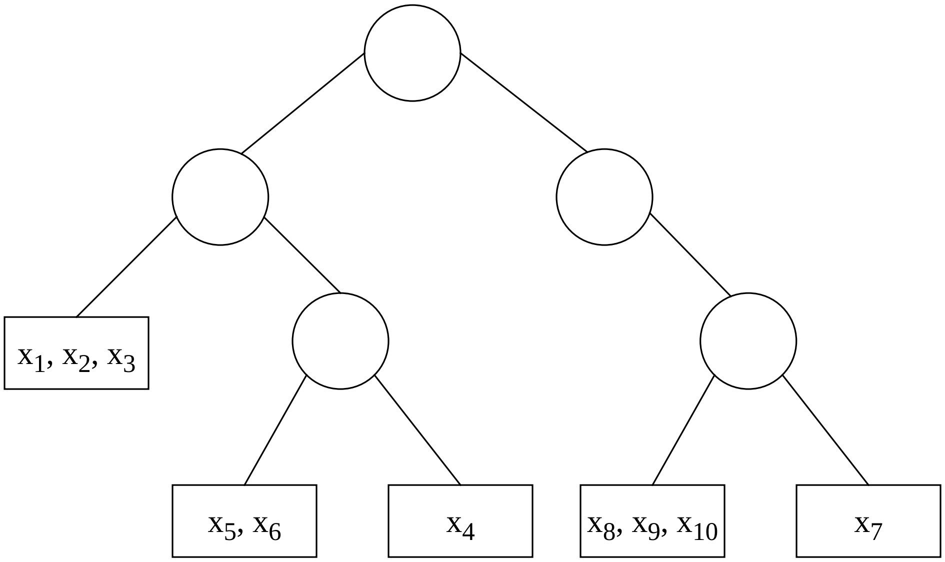

Figure 1: A b-trie with $b = 3$ built from the following ten strings: $X_1 = 11000\ldots$, $X_2 = 11100\ldots$, $X_3 = 11111\ldots$, and $X_4 = 1000\ldots$, $X_5 = 10111\ldots$, $X_6 = 10101\ldots$, $X_7 = 00000\ldots$, $X_8 = 00111\ldots$, $X_9 = 00101\ldots$, $X_4 = 00100\ldots$.

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 [17]) 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 (cf. [4, 7, 6, 14]), 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 five 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 distribution can be approximated by (for fixed large k and $n, b \to \infty$)

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

where $a = \sqrt{b}(1 - n2^{-k}) \to \infty$ (cf. Theorem 2 and the Appendix). 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, $\mathsf{Pr}\{\mathcal{H}_n = k_1\} = 1 - o(1)$. 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, Flajolet [4], Devroye [2], Jacquet and Régnier [6], and Pittel [14] established the asymptotic distribution for b-tries with *fixed* b using probabilistic and analytic tools (cf. also [7]). 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 [4] and [7] asymptotic formulas were presented that apply for n large with b fixed, for various ranges of k. For purposes of comparison, we repeat these results below.

Theorem 1 *The distribution of the height of b-tries has the following asymptotic expansions for fixed b:*

(i) Right-Tail Region*: $k \to \infty$, $n = O(1)$:*

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

(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,

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

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:

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

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 five cases of (n, k) to consider, and we summarize our final results below. The necessity of treating the five 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) $b, k \to \infty$, $(n-b)2^{-k} \to 0$, $b \geq \delta n$ $(\delta > 0)$

$$1 - h_n^k = \binom{n}{b+1} 2^{-kb}\left[1 + O\left(\frac{n-b-1}{2^k}\right)\right].$$

(b) $b, n, k \to \infty$, $(n-b)2^{-k} \to \infty$, $nb^{-1}2^{-k} \leq \delta_1 < 1$

$$1 - h_n^k = \binom{n}{b}\frac{(1-2^{-k})^{n-b}}{2^{k(b-1)}}\left[\frac{b2^k}{n-b} - 1\right]^{-1}[1 + O(b(n-b)^{-2}4^k) + O(2^{-k})]$$

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

$$h_n^k = \frac{K_0}{\sqrt{1 - a(a+\zeta_0)}}\exp(2^k\Psi_0)\left[1 + O\left(\frac{1}{\sqrt{b}}\right)\right]$$

where

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

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

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

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

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

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

$$h_n^k = \sqrt{2\pi b 2^k}\left(\frac{1}{\sqrt{2\pi b}}\right)^{2^k}\frac{2^{kj}}{j!}[1 + O(2^{-k}j^2)]$$

for $j \geq 0$.

We observe that for cases (c), (d) and (e), h_n^k is exponentially small, while for cases (a) and (b), $1-h_n^k$ is exponentially small. From the definition of ζ_0 in part (c), we can easily show that

$$\zeta_0(a) = -a + \frac{1}{\sqrt{2\pi}} e^{-a^2/2} + O(e^{-a^2}), \quad a \to +\infty \tag{2.8}$$

$$\zeta_0(a) = \frac{1}{a} - 2a + O(a^3), \quad a \to 0^+. \tag{2.9}$$

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 (b) and (c) may be obtained by evaluating (c) in the limit $a \to \infty$. Using (2.8) we are led to (see the Appendix for the derivation)

$$h_n^k \sim \exp\left(-\frac{2^k}{\sqrt{2\pi}} \frac{e^{-a^2/2}}{a} \right). \tag{2.10}$$

This result applies to the limit where $b, n, k \to \infty$ with $a = \sqrt{b}(1 - n2^{-k}/b) \to \infty$ but $n2^{-k}/b \to 1^-$. Observe that (2.10) asymptotically matches with the result in Theorem 2(b), if a is sufficiently large so that $2^k e^{-a^2/2} a^{-1} \to 0$. 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 (2.10) applies when n, b and k are such that h_n^k is neither close to 0 nor to 1.

The result (2.10) 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 let n and k vary. We first define

$$k_0 = \lceil \log_2(n/b) \rceil,$$

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 (e) 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 β $(0 < \beta < 1)$, then we can easily show that j, γ and a are all asymptotically large, so that parts (c)-(e) of Theorem 2 do not apply, and we must use part (b) (or the intermediate result in (2.10)) to compute h_n^k. We thus have $h_n^{k_0-1} = 0$ and $h_n^{k_0} \sim 1$ so that the mass accumulates at $k = k_0 = \lfloor \log_2(n/b) \rfloor + 1$. In passing we should point out that if we consider a sequence of n, b such that $\beta \to 1^-$, then the conditions where parts (c) and (d) of Theorem 2 are valid may be satisfied.

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,*

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

as $n \to \infty$. If $\beta \to 1^-$ in such a way that $2^k e^{-a^2/2} a^{-1}$ is bounded, then the height distribution concentrates on two consecutive points.

3 Derivation of Results

We establish the five 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.

The distribution $h_n^k = \mathsf{Pr}\{\mathcal{H}_n \le k\}$ is given by the Cauchy integral (2.5). Observe that

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

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

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

(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 $0 < u \leq A/b$ 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 $u \geq A/b$ 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 $(n - b - 1)2^{-k} \to 0$. 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

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

and we obtain part (a) of Theorem 2.

Now consider the limit where $(n - b)2^{-k} \to \infty$ and $nb^{-1}2^{-k} \leq \delta_1 < 1$. Using Lemma 1(i) we obtain

$$1 - h_n^k = \frac{n!}{2\pi i}\int_{|z|=(n-b)/(1-2^{-k})} \frac{e^z}{z^{n+1}}2^k e^{-z2^{-k}}\frac{z^b}{2^{kb}b!}\frac{1}{b2^k/z - 1}[1 + O(bz^{-2}4^k)]dz. \tag{3.3}$$

The above has a saddle where

$$\frac{d}{dz}[z + (b-n)\log z] = 0 \Rightarrow z = n - b$$

and then the standard saddle point approximation to (3.3) yields

$$1 - h_n^k \sim \frac{n!}{(n-b)!b!}\frac{(1-2^{-k})^{n-b}}{2^{k(b-1)}}\left[\frac{b2^k}{n-b} - 1\right]^{-1}. \tag{3.4}$$

We have thus obtained Theorem 2 part (b). The error term therein follows from (3.3).

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,

$$\begin{aligned} z + 2^k \log\left(\int_{z2^{-k}}^{\infty} e^{-w}\frac{w^b}{b!}dw\right) &= 2^k b\log(z2^{-k}) - 2^k\log(b!) \\ &- 2^k\log\left(1 - \frac{b}{z2^{-k}}\right) + O(b8^k z^{-2}). \end{aligned} \tag{3.5}$$

We furthermore scale $z = 4^k bt$ and then (2.5) with (3.5) becomes

$$\begin{aligned} h_n^k &= n! e^{-2^k\log(b!)} e^{2^k b\log(2^{-k})} \frac{1}{2\pi i}\oint z^{j-1}\exp\left(-2^k\log\left(1 - \frac{b}{z2^{-k}}\right)\right)[1 + O(b8^k z^{-2})]dz \\ &= n!(4^k b)^j e^{-2^k\log(b!)} e^{2^k b\log(2^{-k})}\frac{1}{2\pi i}\oint t^{j-1}e^{1/t}[1 + O(2^{-k}b^{-1}t^{-2} + 2^{-k}t^{-2})]dt \\ &= \frac{n!}{j!}(4^k b)^j e^{-2^k\log(b!)} e^{2^k b\log(2^{-k})}[1 + O(j^2 2^{-k})]. \end{aligned} \tag{3.6}$$

Using Stirling's formula to approximate $n!$ and $b!$ and replacing n by $b2^k - j$, we see that (3.6) is asymptotically equivalent to Theorem 2(e).

Next we take b, n, k large with $b - n2^{-k} = \gamma$ fixed. We may still use the approximation (3.5). We now set $z = 2^k b\tau$ and obtain from (2.5) and (3.5)

$$\begin{aligned} h_n^k &= n!\left(\frac{1}{2^k b}\right)^{n - 2^k b} e^{-2^k\log(b!)}\left(\frac{1}{2^k}\right)^{2^k b} J \\ J &= \frac{1}{2\pi i}\oint e^{(2^k b - n)\log\tau - 2^k\log(1 - 1/\tau)}\frac{d\tau}{\tau}[1 + O(b^{-1})]. \end{aligned} \tag{3.7}$$

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

$$J \sim \frac{1}{\sqrt{2\pi}}\exp\left[(2^k b - n)\log\left(1 + \frac{1}{\gamma}\right) + 2^k\log(1 + \gamma)\right]\frac{1}{\sqrt{(2^k b - n)(1 + b - n2^{-k})}}. \tag{3.8}$$

Using (3.8) in (3.7) along with Stirling's formula, and writing the result in terms of b, k and γ, we obtain Theorem 2(d).

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

$$\begin{aligned}\log(1-I) &= \log\left[\frac{1}{\sqrt{2\pi}}\int_B^\infty e^{-x^2/2}dx - \frac{1}{\sqrt{2\pi}}\frac{B^2+2}{3\sqrt{b}}e^{-B^2/2} + O(b^{-1})\right] \qquad (3.9)\\ &= \log\left(\frac{1}{\sqrt{2\pi}}\int_B^\infty e^{-x^2/2}dx\right) + \frac{B^2+2}{3\sqrt{b}}\frac{e^{-B^2/2}}{\int_B^\infty e^{-x^2/2}dx} + O(b^{-1}).\end{aligned}$$

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

$$\begin{aligned} n!e^z z^{-n} &= \exp\left[2^k(b+\sqrt{b}\zeta) - n\log(2^k b) - n\log\left(1+\frac{\zeta}{\sqrt{b}}\right)\right]n! \qquad (3.10)\\ &= \sqrt{2\pi n}\exp\left[2^k\frac{(a+\zeta)^2}{2} + \frac{2^k}{\sqrt{b}}\left(\frac{a^3}{6}-\frac{a\zeta}{2}-\frac{\zeta^3}{3}\right) + O\left(\frac{2^k}{b}\right)\right].\end{aligned}$$

Here we have again used Stirling's formula and recalled that $n = 2^k b(1-a/\sqrt{b})$. Using (3.9) and (3.10), (2.5) becomes

$$h_n^k = \frac{\sqrt{2\pi n}}{2\pi i}\frac{1}{\sqrt{b}}\oint K(\zeta;b)e^{2^k\Psi(\zeta)}d\zeta \qquad (3.11)$$

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 (3.11) by the saddle point method. The equation locating the saddle points is $\Psi'(\zeta)=0$, i.e.,

$$a+\zeta = \frac{e^{-\zeta^2/2}}{\int_\zeta^\infty e^{-x^2/2}dx}. \qquad (3.12)$$

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 (3.12),

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

Then the standard Laplace estimate of (3.11) leads to part (c) of Theorem 2.

We comment that a more uniform result than that in (c) can be given. We have

$$h_n^k \sim \frac{n!}{\sqrt{2\pi}}\left[\frac{n}{z_*^2} - \frac{2^{-k}}{1-I_*}\left(I_*''+\frac{(I_*')^2}{1-I_*}\right)\right]^{-1/2}\frac{e^{z_*}}{z_*^{n+1}}(1-I_*)^{2^k} \qquad (3.13)$$

where $z_* = z_*(n, b, k)$ is the solution to

$$1 - \frac{n}{z} - \frac{I'(z2^{-k}; b)}{1 - I(z2^{-k}; b)} = 0,$$

and I is defined in Lemma 1. Also, $I'_* = I'(z_*2^{-k}; b)$ and $I''_* = I''(z_*2^{-k}; b)$. The above is more general than Theorem 2(c) in that the condition $2^k = O(\sqrt{b})$ is not required. This can be obtained by writing

$$h_n^k = \frac{n!}{2\pi i} \oint \frac{1}{z} \exp[z - n \log z + 2^k \log(1 - I(z2^{-k}; b))]dz$$

and using the saddle point method, without using Lemma 1 to approximate I. The saddle point equation is

$$\frac{d}{dz}[z - n \log z + 2^k \log(1 - I)] = 1 - \frac{n}{z} - \frac{I'}{1 - I} = 0$$

and we ultimately obtain (3.13).

4 Numerical Studies

We determine the numerical accuracy of the results in Theorem 2, and also demonstrate the necessity of treating the five different scales. To do so, it is best to fix b and k, and vary n. We consider the range $b + 1 \leq n \leq b2^k$, 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 (e) of Theorem 2. We also comment that for a fixed large b and n, the conditions under which (c)–(e) 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 part (c), is (numerically) satisfied. Table 1 gives the exact values of $1 - h_n^k$ and the approximations from Theorem 2, parts (a) and (b). 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 $n \geq 18$.

In Table 2 we retain $b = 16$ and $k = 2$, but now take $46 \leq n \leq 64$. We tabulate the exact h_n^k along with the asymptotic results in parts (c)–(e) of Theorem 2. We also give the 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, γ and j are $O(1)$, respectively. When $n = 64$, approximation (e) is accurate to within 2%. When $n = 63$, (e) is more accurate than (d), but (d) becomes superior for $n \leq 62$. When n is further decreased to $n = 54$, (c) becomes more accurate than (d). We also recall that when h_n^k is not close to either 0 or 1, then part (c) 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) and (b) and in Table 4 we give h_n^k for cases (c)–(e) (again tabulating the values of a, γ and j). When $n = 65 = b + 1$, (a) is superior to (b), but (b) is the better approximation for $n \geq 66$. Table 4 considers $400 \leq n \leq 512 = b2^k$ and demonstrates the transition between cases (c) and (d) and then (d) and (e). 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 $b \to \infty$.

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

n	$1 - h_n^k$ (exact)	$1 - h_n^k$ (a)	$1 - h_n^k$ (b)
17	$.233\ (10^{-9})$	$.233\ (10^{-9})$	$.188\ (10^{-9})$
18	$.320\ (10^{-8})$	$.419\ (10^{-8})$	$.259\ (10^{-8})$
19	$.232\ (10^{-7})$	$.398\ (10^{-7})$	$.187\ (10^{-7})$
20	$.118\ (10^{-6})$	$.265\ (10^{-6})$	$.951\ (10^{-7})$
21	$.475\ (10^{-6})$	$.139\ (10^{-5})$	$.381\ (10^{-6})$
22	$.160\ (10^{-5})$	$.613\ (10^{-5})$	$.128\ (10^{-5})$
23	$.469\ (10^{-5})$		$.374\ (10^{-5})$
24	$.123\ (10^{-4})$		$.980\ (10^{-5})$
26	$.652\ (10^{-4})$		$.516\ (10^{-4})$
28	$.263\ (10^{-3})$		$.207\ (10^{-3})$
30	$.863\ (10^{-3})$		$.676\ (10^{-3})$
32	$.240\ (10^{-2})$		$.187\ (10^{-2})$
34	$.585\ (10^{-2})$		$.453\ (10^{-2})$
36	$.127\ (10^{-1})$		$.139\ (10^{-2})$
38	$.253\ (10^{-1})$		$.194\ (10^{-1})$
40	$.462\ (10^{-1})$		$.352\ (10^{-1})$
42	$.790\ (10^{-1})$		$.599\ (10^{-1})$
44	.127		$.958\ (10^{-1})$
46	.193		.146
48	.278		.211
50	.383		.294

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

n	h_n^k (exact)	(a)	h_n^k (c)	(γ)	h_n^k (d)	(j)	h_n^k (e)
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)	$.198\ (10^{-2})$

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

n	$1 - h_n^k$ (exact)	$1 - h_n^k$ (a)	$1 - h_n^k$ (b)
65	.159 (10^{-57})	.159 (10^{-57})	.142 (10^{-57})
66	.922 (10^{-56})	.105 (10^{-55})	.821 (10^{-56})
67	.271 (10^{-54})	.352 (10^{-54})	.241 (10^{-54})
68	.538 (10^{-53})	.798 (10^{-53})	.479 (10^{-53})
69	.814 (10^{-52})	.138 (10^{-51})	.724 (10^{-52})
70	.100 (10^{-50})	.193 (10^{-50})	.889 (10^{-51})
100	.176 (10^{-32})		.156 (10^{-32})
150	.564 (10^{-19})		.492 (10^{-19})
200	.118 (10^{-11})		.102 (10^{-11})
250	.468 (10^{-7})		.395 (10^{-7})
300	.522 (10^{-4})		.431 (10^{-4})
350	.583 (10^{-2})		.468 (10^{-2})
400	.130		.103

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

n	h_n^k (exact)	(a)	h_n^k (c)	(γ)	h_n^k (d)	(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})

These data also suggest that in some cases it may be desirable to calculate some of the higher order terms in the asymptotic series. For case (c) 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 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 part (c) of Theorem 2 (see also (3.13)).

ACKNOWLEDGMENT: We thank the referee for useful suggestions.

Appendix

We discuss the asymptotic matching region between cases (b) and (c) of Theorem 2. In particular, we establish (2.10) in the matching region directly from the integral representation (2.5).

We use an approximation to I that applies for $A-b=o(b)$ but with $(A-b)/\sqrt{b}=B\to -\infty$. In this range we have, from Lemma 1,

$$1-I\sim 1-\frac{e^{-B^2/2}}{\sqrt{2\pi}}\left(\frac{1}{B}-\frac{B^2}{3\sqrt{b}}+\cdots\right)$$

and hence

$$\begin{aligned} h_n^k &\sim \frac{n!}{2\pi i}\oint \frac{e^z}{z^{n+1}}\left[1+\frac{e^{-B^2/2}}{B\sqrt{2\pi}}\left(1-\frac{B^3}{3\sqrt{b}}+\cdots\right)\right]^{2^k} dz \\ &\sim \frac{n!}{2\pi i}\oint \frac{e^{2^k b}e^{2^k\sqrt{b}B}\sqrt{b}2^{-kn}}{(b+\sqrt{b}B)^{n+1}}\exp\left[\frac{e^{-B^2/2}}{B}\frac{2^k}{\sqrt{2\pi}}\left(1-\frac{B^3}{3\sqrt{b}}\right)\right]dB. \qquad \text{(A.1)}\end{aligned}$$

Next we use

$$\begin{aligned}\exp(2^k\sqrt{b}B)\left(1+\frac{B}{\sqrt{b}}\right)^{-n} &\sim \exp\left[\left(2^k\sqrt{b}-\frac{n}{\sqrt{b}}\right)B+\frac{n}{2b}B^2\right] \\ &= \exp\left[2^k\left(aB+\frac{n}{2^k b}\frac{B^2}{2}\right)\right]\end{aligned}$$

in (A.1) and note that in the matching region $n/(2^k b)\sim 1$. We recall that $a=\sqrt{b}(1-n2^{-k}/b)$.

The integrand in (A.1) has a saddle point at $B=-a$ and a standard application of the steepest descent method yields

$$h_n^k\sim \frac{n!}{b^n\sqrt{b}}e^{2^k b}\frac{1}{\sqrt{2\pi}}2^{-kn}2^{-k/2}e^{-2^k a^2/2}\exp\left[-\frac{e^{-a^2/2}}{a\sqrt{2\pi}}2^k\right]. \qquad \text{(A.2)}$$

But in this limit we have

$$\frac{n!}{b^n}e^{2^k b}2^{-kn} \sim \left(\frac{n}{b2^k}\right)^n\sqrt{2\pi n}e^{2^k b-n}$$

$$
\begin{aligned}
&= \sqrt{2\pi n}\left(1-\frac{a}{\sqrt{b}}\right)^n e^{2^k\sqrt{b}a}\\
&\sim \sqrt{2\pi n}\exp\left[\left(2^k\sqrt{b}-\frac{n}{\sqrt{b}}\right)a-\frac{n}{2b}a^2\right]\\
&\sim \sqrt{2\pi b}2^{k/2}\exp\left(2^k\frac{a^2}{2}\right).
\end{aligned}
$$

Using the above in (A.2) establishes (2.10) and shows that there are no “gaps” in the asymptotics between cases (b) and (c).

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, *Proc. LATIN'2000*, Punta del Este, Uruguay, Lecture Notes in Computer Science, **1776**, 298-307, 2000.

[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, *SIAM J. Computing*, 2000.

[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.