On the Average Profile of Symmetric Digital Search Trees*

Charles Knessl
Dept. Mathematics, Statistics & Computer Science
University of Illinois at Chicago
University of Illinois at Chicago
Chicago, Illinois 60607-7045
U.S.A
knessl@uic.edu

Wojciech Szpankowski
Department of Computer Science
Purdue University
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

A digital search tree (DST) – one of the most fundamental data structures on words – is a digital tree in which keys (strings, words) are stored directly in (internal) nodes. The *profile* of a digital search tree is a parameter that counts the number of nodes at the same distance from the root. It is a function of the number of nodes and the distance from the root. Several tree parameters, such as height, size, depth, shortest path, and fill-up level, can be uniformly analyzed through the profile. In this note we analyze asymptotically the *average* profile for a *symmetric digital search tree* in which strings are generated by an unbiased memoryless source. We show that the average profile undergoes several phase transitions: initially it resembles a full tree until it starts growing algebraically with the number of nodes, and then it decays first algebraically, then exponentially, and finally quadratic exponentially. We derive these results by a combinational of analytic techniques, such as the saddle point method.

1 Introduction

Digital trees [6, 14] have experienced a new wave of interest due to a number of novel applications in computer science and telecommunications. For example, recent developments in the context of a distributed hash table leads to the analysis of digital trees [9]. Partial matching of multidimensional data provides another application. In telecommunications, recent developments in conflict resolution algorithms and data compression have also brought a new interest in digital trees [1, 4, 14]. The three primary digital tree search methods are [6, 14]: tries, PATRICIA tries, and digital search trees (DST). In a digital search tree, the subject of this paper, strings are directly stored in nodes. More precisely, the root contains the first string, and the next string occupies the right or the left child of the root depending on whether its first symbol of the next string is "0" or "1". The remaining strings are stored in available nodes which are directly attached to nodes already existing in the tree. The search for an available node follows the prefix structure of a string. That is, if the next symbol in a string is "0" we move to the right, otherwise we move to the left. This is illustrated in Figure 1.

Throughout the paper, we write X_n^k to denote the number of nodes at distance k from the root. We call it the *profile* of the tree. We shall analyze it in a digital search tree built on n strings generated by an unbiased memoryless source. More precisely, we assume that the input is a sequence of n independent and identically distributed random variables, each being composed of an infinite sequence of Bernoulli random variables with probability $p = 1/2$ of generating a "1". The corresponding DST constructed from these n bit-strings

*The work was supported by NSF Grants DMS-05-03745, CCF-0513636, DMS-0800568, and CCF-0830140, and NSA Grant H98230-08-1-0092, and the AFOSR Grant FA8655-08-1-3018.

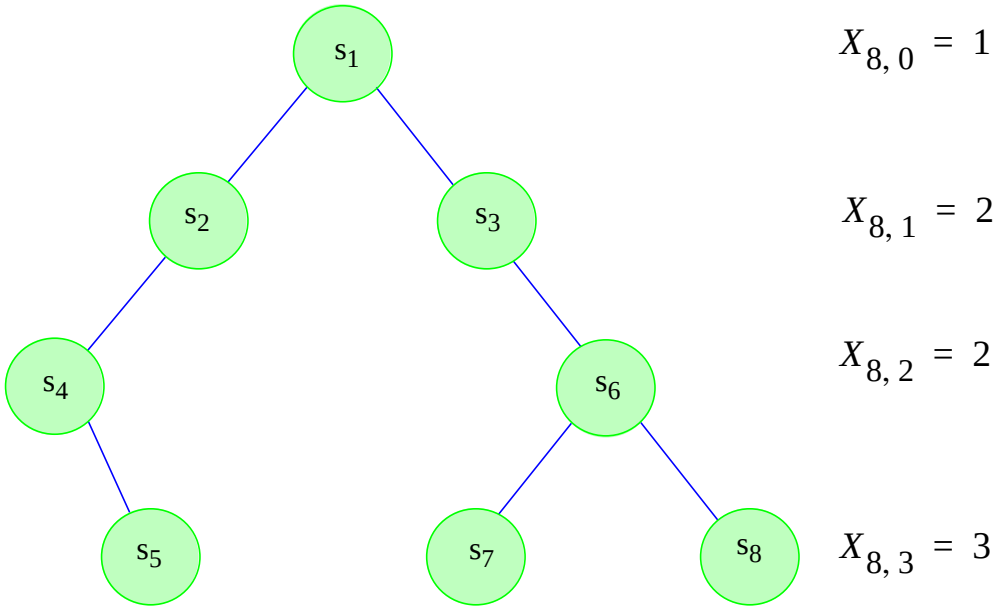

Figure 1: A digital search tree built on eight strings $s_1, \ldots, s_8$ (i.e., $s_1 = 0\ldots$, $s_2 = 1\ldots$, $s_3 = 01\ldots$, $s_4 = 11\ldots$, etc.) and its profile.

is called a binary *random DST*. This simple model may seem too idealized for practical purposes, however, the typical behaviors under such a model often hold under more general models such as Markovian or dynamical sources, although the technicalities are usually more involved. For example, recently [2] extended our analysis to *asymmetric* digital trees built over strings generated by a biased memoryless source.

The motivation for studying the profiles is multifold. The profile is a fine shape measure closely connected to many other cost measures on DST; it is used in the analysis of many important algorithms (e.g., in the popular data compression scheme Lempel-Ziv'78 the profile X_n^k enumerates the number of phrases of length k with n LZ'78 phrases [4]). Most interesting DST parameters can be directly computed and analyzed through the profile; for example, the height is $\max\{k : \ X_n^k > 0\}$ while the total path length is $\sum_j X_n^j$, where X_n^k is the profile at level k.

In this paper we study the *expected* profile, $x_n^k = \mathbf{E}[X_n^k]$ for a wide range of values of k and $n \geq k$. From the analytic point of view, we study here the following recurrence:

$$x_{n+1}^{k+1} = 2^{1-n} \sum_{\ell=0}^{n} \binom{n}{\ell} x_\ell^k \ ; \qquad k \geq 0, n \geq 0$$

(along with some initial conditions). To see how to construct such a recurrence, we observe that two subtrees of a DST are digital search trees themselves with levels reduced to k, and with sizes, respectively, ℓ and $n - \ell$, where ℓ is binomially distributed $\binom{n}{\ell}2^{-n}$ (i.e., ℓ strings start, say with a zero). Thus both of these subtree profiles are x_ℓ^k and $x_{n-\ell}^k$, but symmetry allows us to consider only x_ℓ^k (and multiply the sum above by 2). We also notice that this recurrence depends on two parameters n and k which makes the analysis quite challenging, as we will demonstrate.

Our main result concerns the asymptotic expansion of the average profile for various

ranges of k and n. We identify five ranges k, from $k = O(1)$ up to $k \sim n$. We shall show that the average profile undergoes several phase transitions: initially it resembles a full tree (i.e., $x_n^k \sim 2^k$) until it starts growing algebraically with the number of nodes, and then it decays first algebraically, then exponentially, and finally quadratic exponentially. We derive these results by a combinational of analytic techniques such as the saddle point method.

To the best of our knowledge these are new results, except for the range $k = \log_2 n + O(1)$, which was previously analyzed by Louchard [7], Szpankowski [13], Prodinger [12], and Louchard and Szpankowski [8]. Recently, Drmota and Szpankowski [2] presented a detailed analysis of the DST profile for the *asymmetric* case, that is, when strings are generated by a biased memoryless sources. We also mention the recent complete analysis of (external and internal) profiles of tries [10, 11]. However, tries are easier to analyze than DST, due to their simpler structure [6, 14].

The paper is organized as follows. In the next section we summarize our main results. In Section 3 we derive an exact formula for the expected profile, and in Section 4 we derive our asymptotic results.

2 Summary of Results

We let x_n^k be the mean profile in a digital search tree (DST) built over n binary strings, at a distance k from the root. The binary strings are independent and a "zero" and "one" occur with equal probabilities 1/2, that is, strings are generated by an unbiased memoryless source.

The mean profile satisfies the recurrence

$$x_{n+1}^{k+1} = 2^{1-n} \sum_{\ell=0}^{n} \binom{n}{\ell} x_\ell^k \ ; \ k \geq 0, n \geq 0 \tag{2.1}$$

and the initial conditions in k are

$$x_0^0 = 0, \ x_n^0 = 1 \text{ for } n \geq 1. \tag{2.2}$$

It follows from (2.1) that

$$x_1^1 = 0, \ x_n^1 = 2 - 2^{2-n} \text{ for } n \geq 2. \tag{2.3}$$

Since the height of a DST can be at most $n - 1$, we have

$$x_n^k = 0 \text{ for } k \geq n. \tag{2.4}$$

This result also follows from (2.1) and (2.2), and use of (2.4) allows us to truncate the sum in (2.1), thus obtaining

$$x_{n+1}^{k+1} = 2^{1-n} \sum_{\ell=k+1}^{n} \binom{n}{\ell} x_\ell^k, \ k < n. \tag{2.5}$$

If we change variables from (n, k) to (n, L) where $L = n - k$ and

$$x_n^k = Y(n, L) = Y(n, n - k) \tag{2.6}$$

and shift the summation index in (2.5), we are led to the new recurrence

$$Y(n+1,L) = 2^{1-n} \sum_{j=1}^{L} \binom{n}{n-L+j} Y(n-L+j,j), \tag{2.7}$$

which applies for $L \geq 1$.

When $L = 1$ we obtain from (2.7) $Y(n+1,1) = 2^{1-n} Y(n,1)$ and thus, since $Y(1,1) = x_1^0 = 1$,

$$Y(n,1) = x_n^{n-1} = 2 \cdot 2^{-n^2/2} 2^{3n/2}, \ n \geq 1. \tag{2.8}$$

Then by using (2.8) we obtain

$$Y(n,2) = x_n^{n-2} = 2^{-n^2/2} 2^{5n/2} \left[\frac{n}{8} - \frac{1}{4} + \frac{1}{2} \cdot 2^{-n} \right], \ n \geq 2 \tag{2.9}$$

and

$$Y(n,3) = x_n^{n-3} = 2^{-n^2/2} 2^{7n/2} \left[\frac{n^2}{128} - \frac{5n}{128} + \frac{1}{24} + 2^{-n} \left(\frac{n}{8} - \frac{1}{4} \right) + \frac{1}{3} 4^{-n} \right], \ n \geq 3. \tag{2.10}$$

The expressions in (2.8)-(2.10) give the mean profiles when the height of the tree is near its maximum possible value, which is clearly very unlikely.

In the next section, we derive a general expression for $Y(n,k)$:

$$\begin{aligned} Y(n,L) = x_n^{n-L} &= 2^{-k^2/2} 2^{k/2} \sum_{p=0}^{L-1} 2^{-np} \sum_{j=0}^{L-1-p} \binom{n}{L-j-1-p} \\ &\times \ 2^{Lp} (-1)^j \left[\prod_{\nu=1}^{j} \frac{2\nu}{2^\nu - 1} \right] G_0(p), \end{aligned} \tag{2.11}$$

where

$$G_0(p) = \prod_{i=1}^{p} \frac{1}{2^i - 1} = 1 \cdot \frac{1}{3} \cdot \frac{1}{7} \cdot \frac{1}{15} \cdots \frac{1}{2^p - 1}, \tag{2.12}$$

and we define $G_0(0) = 1$.

We contrast (2.11) to the form of the solution found by Louchard [7] as

$$x_n^k = 2^k \left[1 + \sum_{\ell=1}^{k} \frac{R(k-\ell)}{Q(\ell-1)} (1 - 2^{-\ell})^{n-1} \right], \tag{2.13}$$

where

$$Q(\ell) = \prod_{\nu=1}^{\ell} [1 - 2^{-\nu}], \qquad R(\ell) = (-1)^{\ell+1} \prod_{\nu=1}^{\ell} \frac{1}{2^\nu - 1}, \tag{2.14}$$

with $Q(0) = 1$ and $R(0) = -1$. The form (2.13) is most useful for smaller values of k, while (2.11) is most useful for $k \approx n$.

Now consider the asymptotic limit $n \to \infty$. Below we give five ranges of k that lead to different asymptotic expansions of x_n^k:

(i) **Case**: $k = O(1)$, $n \to \infty$

$$x_n^k \sim 2^k - \frac{2^k}{Q(k-1)}(1-2^{-k})^{n-1}. \tag{2.15}$$

(ii) **Case**: $k, n \to \infty$ with $k - \log_2 n = \theta = O(1)$

$$\begin{aligned} x_n^k &\sim 2^k \left\{ 1 + \frac{1}{Q(\infty)} \sum_{i=0}^{\infty} (-1)^{i+1} \left[\prod_{\ell=1}^{i} \frac{1}{2^\ell - 1} \right] \exp(-2^{i-\theta}) \right\} \\ Q(\infty) &= \prod_{\nu=1}^{\infty} (1-2^{-\nu}) = .2887880950\cdots . \end{aligned} \tag{2.16}$$

(iii) **Case**: $k, n \to \infty$ with $k = \alpha \log_2 n$, $1 < \alpha < \infty$ ($\theta = (\alpha - 1)\log_2 n$)

$$\begin{aligned} x_n^k &\sim 2^k \sqrt{\pi} 2^{-5/8} \theta^{-1} 2^{-\theta^2/2} 2^{-\theta/2} e^{\theta} e^{-\theta \log \theta} \\ &\times 2^{-\log_2^2(\theta)/2} \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} 2^{u^2/2} \tilde{A}\left(u + \frac{1}{2} - \theta - \log_2 \theta \right) du. \end{aligned} \tag{2.17}$$

Here $\tilde{A}$ satisfies $\tilde{A}(z+1) = \tilde{A}(z)$, $\tilde{A}(z) = \tilde{A}(-z)$ and has the explicit form

$$\tilde{A}(z) = \frac{2^{-z^2/2} 2^{z/2} (1-2^{-z})}{Q(\infty)\sin(\pi z)} \prod_{m=1}^{\infty} (1-2^{z-m})(1-2^{-z-m}).$$

(iv) **Case**: $k, n \to \infty$ with $0 < k/n < 1$

$$\begin{aligned} x_n^k &\sim 2^{-k^2/2} 2^{k/2} \frac{n^n}{(n-k)^{n-k} k^k} \frac{\sqrt{n}\sqrt{n-k}}{\sqrt{2\pi} k^{3/2}} J(\beta), \\ J(\beta) &= \sum_{j=0}^{\infty} (-\beta)^j \left[\prod_{\nu=1}^{j} \frac{2^\nu}{2^\nu - 1} \right], \ \beta = \frac{n}{k} - 1, \ 0 < \beta < 1, \\ J(\beta) &= \frac{1}{2i} \int_{\frac{1}{2} - i\infty}^{\frac{1}{2} + i\infty} \frac{\beta^{-z}}{\sin(\pi z)} A(z) dz, \ 0 < \beta < \infty, \\ A(z) &= \frac{1}{Q(\infty)} \prod_{m=1}^{\infty} (1-2^{z-m}) = \prod_{\ell=1}^{\infty} \exp\left[\frac{1-2^{\ell z}}{\ell(2^\ell - 1)} \right]. \end{aligned} \tag{2.18}$$

We note that $\beta \to \infty$ as $k/n \to 0$.

(v) **Case**: $k, n \to \infty$ with $n - k = L$, $L \geq 1$

$$x_n^k \sim 2^{-n^2/2} 2^{\left(L+\frac{1}{2}\right)n} 2^{-L^2/2} 2^{-L/2} \frac{n^{L-1}}{(L-1)!}.$$

The result in (ii) was obtained by Louchard in [7], and gives the asymptotic form of the mean DST profile in the form of an infinite mixture of double-exponentials, in the important range where x_n^k transitions from being approximately 2^k to being a fraction of this. Note that

2^k is the maximum number of strings that can be stored at level k. Extensions of Louchard's results to asymmetric DST can be found in [8, 13].

The results in items (iii)-(iv) give ranges of k where x_n^k is very small, as these levels will tend to contain very few strings. Note also that x_n^k becomes asymptotically $O(1)$ for $k - \log_2 n \sim \sqrt{2\log_2 n}$, as then in (2.17) the term 2^k is balanced by $2^{-\theta^2/2}$.

3 Exact Representation

We shall derive the result presented in (2.11). First we note that by examining the cases L = 1,2 and 3 in (2.8)-(2.10), it is clear that $Y(n, L)$ takes the form

$$Y(n,L) = 2^{-n^2/2}2^{\left(L+\frac{1}{2}\right)n}[P(n,L) + 2^{-n}Q(n,L) + 4^{-n}R(n,L) + 8^{-n}S(n,L) + \cdots] \tag{3.1}$$

where $P, Q, R, S, \ldots$ are polynomials in n of degrees $L-1$, $L-2$, $L-3$, $L-4, \ldots$. The series in (3.1) truncates with the term $2^{-(L-1)n}$, which is multiplied by a constant polynomial. We thus set

$$Y(n,L) = x_n^k = 2^{-k^2/2}2^{k/2}A(n,n-k) = 2^{-n^2/2}2^{\left(L+\frac{1}{2}\right)n}2^{-L^2/2}2^{-L/2}A(n,L) \tag{3.2}$$

with which (2.5) becomes

$$2^L A(n+1,L) = 2\sum_{\ell=1}^{L}\binom{n}{n-L+\ell}A(n-L+\ell,\ell). \tag{3.3}$$

Since $x_n^0 = 1$ for $n \geq 1$, we also have

$$A(n,n) = 1,\ n \geq 1. \tag{3.4}$$

Comparing (3.1) to (3.2) let us write the solution in the form

$$\begin{aligned} A(n,L) &= \sum_{j=0}^{L-1}\binom{n}{L-j-1}F_j^L(0) + 2^{-n}\sum_{j=0}^{L-2}\binom{n}{L-j-2}F_j^L(1) \\ &+ 4^{-n}\sum_{j=0}^{L-3}\binom{n}{L-j-3}F_j^L(2) + \cdots \\ &= \sum_{p=0}^{L-1}2^{-np}\sum_{j=0}^{L-1-p}\binom{n}{L-j-p-1}F_j^L(p), \end{aligned} \tag{3.5}$$

where the coefficients $F_j^L(p)$ must be determined from (3.3) and (3.4). Thus the highest power of n in the first sum in (3.5) comes from $F_0^L(0)\ n(n-1)\cdots(n-L+2)/(L-1)!$, and thus the coefficient of n^{L-1} in $P(n,L)$ in (3.1) is $2^{-L^2/2}2^{-L/2}F_0^L(0)/(L-1)!$.

We use (3.5) in (3.3) and the identities

$$\binom{n+1}{L-j-p-1} = \binom{n}{L-j-p-1} + \binom{n}{L-j-p-2} \tag{3.6}$$

and

$$\binom{n}{n-L+\ell}\binom{n-L+\ell}{\ell-p-j-1} = \binom{n}{L-j-p-1}\binom{L-j-p-1}{\ell-j-p-1}. \tag{3.7}$$

First we compare coefficients of 2^{-np} to get

$$2^{L-p}\sum_{j=0}^{L-p-1}\binom{n+1}{L-j-p-1}F_j^L(p)=2\sum_{\ell=1}^{L}\sum_{j=0}^{\ell-p-1}\binom{n}{n-L+\ell}\binom{n-L+\ell}{\ell-p-j-1}2^{(L-\ell)p}F_j^\ell(p) \tag{3.8}$$

and then use (3.6), (3.7) and reverse the order of the double summation in (3.8). After some calculation this yields

$$\begin{aligned}&2^{L-p}[F_j^L(p)+F_{j-1}^L(p)]\\&\quad=2\sum_{\ell=j+p+1}^{L}2^{(L-\ell)p}\binom{L-j-p-1}{\ell-j-p-1}F_j^\ell(p),\ 0\le j\le L-p-1.\end{aligned} \tag{3.9}$$

Here we also equated like polynomials in n in (3.8). We note that, in view of (2.8)-(2.10), $A(n,1)=1$, $A(n,2)=n-2+4\cdot 2^{-n}$ and $A(n,3)=\frac{1}{2}n(n-1)-2n+\frac{8}{3}+8(n-2)2^{-n}+\frac{64}{3}4^{-n}$, so that

$$\begin{gathered}F_0^1=1, F_0^2(0)=1, F_1^2(0)=-2, F_0^2(1)=4,\\F_0^3(0)=1, F_1^3(0)=-2, F_2^3(0)=\tfrac{8}{3}, F_0^3(1)=8, F_1^3(1)=-16, F_0^3(2)=\tfrac{64}{3}.\end{gathered} \tag{3.10}$$

Setting $p=0$, we see that (3.9) admits a solution $F_j^L(0)=F_j(0)$ (independent of L) with

$$2^L[F_j(0)+F_{j-1}(0)]=2F_j(0)2^{L-j-1} \tag{3.11}$$

so that, using $F_0(0)=1$, we have

$$F_j(0)=(-1)^j\prod_{\ell=1}^{j}\frac{2^\ell}{2^\ell-1}. \tag{3.12}$$

Similarly, for $p\ge 1$, (3.9) admits a solution in the form

$$F_j^L(p)=2^{Lp}G_j(p) \tag{3.13}$$

where $G_j(p)$ satisfies (3.11) for any p. Thus we write

$$G_j(p)=(-1)^j\left[\prod_{\nu=1}^{j}\frac{2^\nu}{2^\nu-1}\right]G_0(p). \tag{3.14}$$

With (3.14), (3.5) and (3.2) we have established (2.11).

It remains only to determine the sequence of constants $G_0(p)$. Using (3.4), (3.13) and setting $L=n$ in (3.5), we can characterize $G_0(p)$ recursively from

$$1=\sum_{p=0}^{n}G_0(p)\left[\sum_{j=0}^{n-p}\binom{n+1}{n-p-j}(-1)^j\prod_{\nu=1}^{j}\frac{2^\nu}{2^\nu-1}\right] \tag{3.15}$$

with $G_0(0)=1$. We have verified from numerical experiments that $G_0(n)$ has the form (2.12), and H. Prodinger [personal communication; cf. also [12]] pointed out to us that (3.15) with

(2.12) can be derived using basic hypergeometric functions. Indeed, to show (3.15) we shall follow Prodinger's derivation using Euler's q-binomial theorem [3], where we define

$$(q)_n = (1-q)(1-q^2)\cdots(1-q^n)$$

(e.g., $Q(n) = (1/2)_n$). In general, for any $0 < q < 1$ (3.15) becomes

$$S_n = \sum_{k=0}^{n-1} \frac{q^{\binom{k+1}{2}}}{(q)_k} \sum_{j=0}^{n-k-1} \binom{n}{k+1+j} \frac{(-1)^j}{(q)_j} = 1,$$

provided that $G_0(p)$ is given by (2.12), that is, in q-notation $G_0(p) = (1/2)_p$. Re-arranging the above sum we find that

$$\begin{aligned}
S_n &:= \sum_{s=0}^{n-1} \binom{n}{s+1} \sum_{k+j=s} \frac{q^{\binom{k+1}{2}}}{(q)_k} \frac{(-1)^j}{(q)_j} \\
&= \sum_{s=0}^{n-1} \binom{n}{s+1} [t^s] \sum_{k\geq 0} \frac{q^{\binom{k}{2}}(qt)^k}{(q)_k} \sum_{j\geq 0} \frac{(-t)^j}{(q)j} \\
&= \sum_{s=0}^{n-1} \binom{n}{s+1} [t^s] \prod_{k\geq 1} (1+q^k t) \prod_{j\geq 0} \frac{1}{1+q^j t} \\
&= \sum_{s=0}^{n-1} \binom{n}{s+1} [t^s] \frac{1}{1+t} \\
&= \sum_{s=0}^{n-1} \binom{n}{s+1} (-1)^s = 1.
\end{aligned}$$

where $[t^n]f(t)$ denotes the coefficient at t^n of $f(t)$. In the above we used twice Euler's formula [3], namely

$$\sum_{n\geq 0} \frac{t^n}{(q)_n} = \frac{1}{(t)_n}.$$

Our (3.15) follows from the above after setting $q = 1/2$.

4 Asymptotic Expansions

We consider x_n^k in various limiting cases. For moderate values of k the form (2.13) obtained by Louchard [7] is convenient for deriving asymptotic results. For a fixed k and $n \to \infty$ we have $x_n^k \sim 2^k$ and the difference $x_n^k - 2^k$ is asymptotically given by the term with $\ell = k$ in the sum in (2.13), which yields (2.15). For n and k large with $n2^{-k} = O(1)$, we define θ by $2^{-\theta} = n2^{-k}$. Then the sum in (2.13) becomes

$$\sum_{i=0}^{k-1} \frac{R(i)}{Q(k-i-1)} \exp[(n-1)\log(1-2^{i-k})]. \tag{4.1}$$

But in this limit $Q(k-i-1) \sim Q(\infty)$,

$$(n-1)\log(1-2^{i-k}) \sim -n2^{i-k} = -2^{i-\theta},$$

and we can extend the upper limit in the sum in (4.1) to $i=\infty$. We thus obtain (2.16). On the θ-scale the terms with $\ell = k - O(1)$ in (2.13) all contribute to the expansion of x_n^k.

Once $\theta = k - \log_2 n$ becomes large it proves difficult to obtain the asymptotics of x_n^k from (2.13), due to the alternating signs arising from $R(\ell)$ in (2.14). We thus use the alternate form in (2.11) to obtain the asymptotic results in (2.17)-(2.19), in decreasing ranges of k/n.

First we let $k, n \to \infty$ with $L = n - k = O(1)$. Now only the term with $p = 0$ and $j = 0$ in (2.11) contributes to the leading order asymptotics. Thus,

$$Y(n,L) \sim 2^{-k^2/2} 2^{k/2} \binom{n}{L-1} \sim 2^{-(n-L)^2/2} 2^{(n-L)/2} \frac{n^{L-2}}{(L-1)!}$$

which establishes (2.19).

Next we let k and $n \to \infty$ at the same rate, letting $\beta \equiv (n-k)/k \in (0,\infty)$. Now the $p = 0$ term in (2.11) dominates the terms for $p \geq 1$ (which are exponentially small in n due to the factor(s) 2^{-np}), but all the terms in the j-sum contribute to the leading term for $Y(n,L)$. We thus have

$$Y(n,L) \sim 2^{-k^2/2} 2^{k/2} \sum_{j=0}^{L-1} \binom{n}{L-j-1} (-1)^j \left[\prod_{\nu=1}^{j} \frac{2^\nu}{2^\nu - 1} \right]. \tag{4.2}$$

We can further simplify (4.2) by using Stirling's formula in the form

$$\begin{aligned} \binom{n}{L-j-1} &= \frac{n!}{(k+j+1)!(n-k-j-1)!} \sim \frac{n!}{k!(n-k)!} \left(\frac{n}{k} - 1\right)^{j+1} \\ &\sim \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sqrt{k}\sqrt{n-k}} \frac{n^n}{k^k (n-k)^{n-k}} \frac{n-k}{k} \beta^j, \end{aligned} \tag{4.3}$$

which holds for $n, k \to \infty$ with $\beta = O(1)$. Using (4.3) in (4.2) and extending the upper limit on the sum in (4.2) from $j = L-1$ to $j = \infty$, we obtain (2.18) with the first representation of $J(\beta)$, as an infinite series. However, the series converges only for $|\beta| < 1$, so that the result applies for $k/n \in \left(\frac{1}{2}, 1\right)$.

To obtain a result for $k/n < \frac{1}{2}$, we continue the series into the range $\beta > 1$. Defining as in [5]

$$A(z) = \frac{1}{Q(\infty)} \prod_{m=1}^{\infty} (1 - 2^{z-m})$$

we see that A vanishes at $z = 1, 2, 3, ...$ and at the negative integer values we have

$$A(-N) = \frac{1}{Q(\infty)} \prod_{m=1}^{\infty} (1 - 2^{-N-m}) = \prod_{m=1}^{N} \frac{2^m}{2^m - 1}, \ N \geq 1. \tag{4.5}$$

Also, $A(0) = 1$. Thus we can use a Watson transformation to represent the series for $J(\beta)$ as the following contour integral

$$J(\beta) = \frac{1}{2\pi i} \int_{Br} \frac{\pi}{\sin(\pi z)} \beta^{-z} A(z) dz, \tag{4.6}$$

where $\Re(z) > 0$ on the vertical Bromwich contour Br. We note that the integrand is analytic in the right half-plane $\Re(z) > 0$, but has simple poles at $z = 0, -1, -2, \cdots$. Closing the

contour in the left half-plane, which is permissible if $\beta \in (0,1)$, and evaluating the integral as a residue series regains the series representation for $J(\beta)$. But, (4.6) converges for all $\beta > 0$.

We next examine the behavior of $J(\beta)$ and the approximation in (2.18) for $\beta \to \infty$, which corresponds to $k = o(n)$. For $\beta \to \infty$ it is desirable to shift the contour Br in (4.6) toward the right and to have an alternate representation of the integrand. In [5] we showed that $A(z)$ may be written in the form

$$A(z) = e^{-\pi i z} 2^{z^2/2} 2^{-z/2} e^{p(z)} \left[\prod_{m=0}^{\infty} \frac{1}{1-2^{-z-m}} \right] \tag{4.7}$$

where $p(z+1) = p(z)$ and this function may be expressed in terms of the Jacobi theta function as

$$e^{p(z)} = 2^{-z^2/2} e^{\pi i z} (-i) 2^{1/8} [Q(\infty)]^{-2} \Theta_1 \left(\frac{i \log 2}{2} z \right) \tag{4.8}$$

where

$$\Theta_1(u) = \Theta_1(u|\tau) = 2q^{1/4} \sin(u) \prod_{m=1}^{\infty} (1 - 2\cos(2u) q^{2m} + q^{4m})(1-q^{2m}). \tag{4.9}$$

Here τ and q are related by $e^{\pi i \tau} = q$ and in the present application $q = 1/\sqrt{2}$ and $u = i(\log 2) z/2$. Let us also define $\tilde{A}$ by

$$\tilde{A}(z) = \frac{e^{-\pi i z}}{\sin(\pi z)} e^{p(z)} = \frac{2^{-z^2/2} 2^{z/2} (1-2^{-z})}{Q(\infty) \sin(\pi z)} \prod_{m=1}^{\infty} (1-2^{z-m})(1-2^{-z-m}) \tag{4.10}$$

and we note that A and $\tilde{A}$ are related by

$$A(z) = \frac{2^{z^2/2} 2^{-z/2} \sin(\pi z)}{\prod_{m=0}^{\infty} (1-2^{-z-m})} \quad \tilde{A}(z). \tag{4.11}$$

From (4.10) we see that $\tilde{A}$ is an entire function of z that satisfies $\tilde{A}(z) = \tilde{A}(-z)$ and $\tilde{A}(z+1) = \tilde{A}(z)$. In terms of $\tilde{A}$ we have

$$J(\beta) = \frac{1}{2i} \int_{Br} \beta^{-z} \frac{2^{z^2/2} 2^{-z/2}}{\prod_{m=0}^{\infty} (1-2^{-z-m})} \tilde{A}(z) dz \tag{4.12}$$

where $\Re(z) > 0$ on Br.

The asymptotics of $J(\beta)$ as $\beta \to \infty$ are readily obtained by shifting the contour far to the right ($\Re(z) \gg 1$) and approximating the infinite product in (4.12) by $1 + O(2^{-z})$. There is a saddle point in (4.12) when

$$\frac{d}{dz} \left[\frac{z^2}{2} \log 2 - z \log \beta \right] = 0 \qquad \Rightarrow \qquad z = \log_2 \beta.$$

Setting $z = \log_2 \beta + \frac{1}{2} + w$, (4.12) becomes

$$J(\beta) \sim 2^{-1/8} \frac{1}{\sqrt{\beta}} 2^{-(\log_2 \beta)^2/2} \frac{1}{2i} \int_{Br} 2^{w^2/2} \tilde{A} \left(w + \frac{1}{2} + \log_2 \beta \right) dw, \quad \beta \to \infty. \tag{4.13}$$

Due to the periodic behavior of $\tilde{A}$ we cannot simplify (4.13) any further, though we can replace $\log_2\beta$ in the argument of $\tilde{A}$ by its fractional part $\{\log_2\beta\} = \log_2\beta - \lfloor\log_2\beta\rfloor$. For $k/n \to 0$ we have $\sqrt{n}\ \sqrt{n-k}\ k^{-3/2}\beta^{1/2} \sim \sqrt{n}k^{-1}$ and for $k^2/n \to 0$ we also have $n^n(n-k)^{k-n}k^{-k} \sim \exp[k - k\log n + k\log n]$. Thus for $\beta \to \infty$ (with $k = o(\sqrt{n})$) (2.18) becomes

$$\begin{aligned} x_n^k \sim\ & 2^{-k^2/2}2^{k/2}e^{k-k\log k}e^{k\log n}2^{-[\log_2(n/k)]^2/2} \\ & \times\frac{2^{-5/8}\sqrt{n\pi}}{k(2\pi i)}\int_{-i\infty}^{i\infty}2^{w^2/2}\tilde{A}\left(w+\frac{1}{2}+\log_2\left(\frac{n}{k}\right)\right)dw. \end{aligned} \tag{4.14}$$

This result was obtained by first fixing n/k, and then letting $\beta \to \infty$ in (2.18), and we will show that (2.18) asymptotically matches to (2.17), which applies for $\alpha = k/\log_2 n > 1$.

Now consider $k, n \to \infty$ with $k = \alpha\log_2 n$ and $\alpha \in (1,\infty)$. In this range again only the $p = 0$ term in (2.11) is important, but now we must re-examine (4.2). We change the summation index from j to $M = L-1-j$ and the right side of (4.2) becomes

$$\begin{aligned} & 2^{-k^2/2}2^{k/2}\sum_{M=0}^{L-1}\binom{n}{M}(-1)^M e^{\pi i(L-1)}A(M+1-L) \\ & = 2^{-k^2/2}2^{k/2}n!(-1)^{n+L+1}\frac{1}{2\pi i}\int_{Br(0,1)}\frac{\Gamma(z+L-n-1)}{\Gamma(z+L)}A(z)dz, \end{aligned} \tag{4.15}$$

where $Br(0,1)$ denotes a Bromwich contour on which $0 < \Re(z) < 1$. We have again represented a sum by a contour integral. The factor $\Gamma(z+L-n-1)/\Gamma(z+L)$ in (4.15) has poles at $z = -L+1+j$ for $0 \le j \le n$. Since $L \ge 1$, $n-L = k$ and $A(z)$ has zeros at all positive integer values, the integrand in (4.15) is analytic for $\Re(z) > 0$. We set $z = n+1-L+s = k+1+s$ and use

$$\frac{\Gamma(s)n!}{\Gamma(n+1+s)} \sim \Gamma(s)n^{-s},\ n \to \infty.$$

Then we shift the contour in (4.15) far to the right and also note that, as $z \to \infty$, $A(z) \sim 2^{z^2/2}2^{-z/2}\sin(\pi z)\tilde{A}(z)$, which follows from (4.11). Thus (4.15) yields

$$x_n^k \sim 2^k\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\frac{\pi}{\Gamma(1-s)}2^{s^2/2}2^{s/2}\left(\frac{2^k}{n}\right)^2\tilde{A}(s)ds, \tag{4.16}$$

where we also used $\Gamma(s)\Gamma(1-s)\sin(\pi s) = \pi$.

Now for $k = \alpha\log_2 n$ with $\alpha > 1$ we have $2^k n^{-1} = n^{\alpha-1}$, which is still large. Then we can simplify (4.16) by the saddle point method. There is a saddle with $s \to -\infty$ and we approximate $\Gamma(1-s)$ by $\sqrt{2\pi(-s)}(-s)^{(-s)}e^s$. The saddle point will satisfy

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

so that $s \sim -(\alpha-1)\log_2 n - \log_2[(\alpha-1)\log_2 n]$. We get

$$s = -(\alpha-1)\log_2 n - \log_2[(\alpha-1)\log_2 n] + \xi$$

and note that

$$\frac{1}{\Gamma(1-s)}2^{s^2/2}2^{s/2}\left(\frac{2^k}{n}\right)^s \sim \frac{1}{\sqrt{2\pi}}[(\alpha-1)\log_2 n]^{-1/2}2^{f(\alpha,n)}e^{g(\alpha,n)}2^{\xi^2/2}2^{\xi/2}, \tag{4.17}$$

where

$$f = -\frac{1}{2}\log_2^2[(\alpha-1)\log_2 n] - \frac{1}{2}(\alpha-1)^2(\log_2 n)^2 - \frac{1}{2}\log[(\alpha-1)\log_2 n],$$

$$g = (\alpha-1)\log_2 n - \frac{1}{2}\log[(\alpha-1)\log_2 n] - (\alpha-1)(\log_2 n)\log[(\alpha-1)\log_2 n].$$

A further shift of integration variables with $\xi = -\frac{1}{2} + u$, with which $2^{\xi^2/2}2^{\xi/2} = 2^{-1/8}2^{u^2/2}$, ultimately leads to (2.17). Here we also used $\tilde{A}(u+1) = \tilde{A}(u)$, $\exp\left\{-\frac{1}{2}\log[(\alpha-1)\log_2 n]\right\} = \theta^{-1/2}$, and wrote the results in terms of $\theta = k - \log_2 n = (\alpha-1)\log_2 n$, even though θ is large in this asymptotic range.

Next we discuss the asymptotic matching between (2.17), as $\alpha \to \infty$, and (2.18) as $\beta \to \infty$. We already expanded (2.18) for $k = o(\sqrt{n})$ which led to (4.14). By periodicity of $\tilde{A}$ we have

$$\begin{aligned}\tilde{A}\left(w + \frac{1}{2} + \log_2\left(\frac{n}{k}\right)\right) &= \tilde{A}\left(w + \frac{1}{2} + \log_2 n - k - \log_2 k\right)\\ &= \tilde{A}\left(w + \frac{1}{2} - (\alpha-1)\log_2 n - \log_2(\alpha\log_2 n)\right)\end{aligned}$$

so that the integrals in (4.14) and (2.17) agree (since $\log_2(\alpha\log_2 n) \sim \log_2[(\alpha-1)\log_2 n]$ for $\alpha \to \infty$) and we must only show that the other factors match. Apart from the factors $\sqrt{\pi}2^{-5/8}$, which appear in both (4.14) and (2.17), we must show that

$$\begin{aligned}&2^{-k^2/2}2^{k/2}e^{k-k\log k}e^{k\log n}\frac{\sqrt{n}}{k}2^{-[\log_2(n/k)]^2/2}\Big|_{\frac{k}{n}\to 0}\\ &\sim 2^k\theta^{-1}2^{-\theta^2/2}2^{-\theta/2}e^{\theta-\theta\log\theta}2^{-\log_2^2(\theta)/2}\Big|_{\alpha\to\infty}.\end{aligned} \tag{4.18}$$

But in this limit $\theta^{-1} = (k-\log_2 n)^{-1} \sim k^{-1}$ and $2^k 2^{-\theta/2} = 2^{k/2}\sqrt{n}$. Furthermore, for $\alpha \gg 1$,

$$\theta - \theta\log\theta = k - \log_2 n - (k-\log_2 n)\log[k-\log_2 n] \sim k - k\log k + (\log k)\log_2 n,$$

$$-\frac{\theta^2}{2} = -\frac{k^2}{2} + k\log_2 n - \frac{1}{2}\log_2^2 n,$$

$$-\frac{1}{2}\left[\log_2\left(\frac{n}{k}\right)\right]^2 = -\frac{1}{2}\log_2^2 n + (\log_2 n)\log_2 k - \frac{1}{2}\log_2^2 k,$$

and $\log_2^2(\theta) = \log_2^2(k-\log_2 n) \sim \log_2^2(k)$ if $k \gg \log_2 n$. Using the above in (4.18) the matching condition is easily verified.

Finally, we verify the asymptotic matching between (2.16) and (2.17). We must let $\theta \to \infty$ in (2.16) and $\alpha \to 1$ in (2.17). Since $\theta = (\alpha-1)\log_2(n)$ the latter amounts to doing nothing. We thus expand (2.16) as $\theta \to \infty$, rewriting the expression as

$$x_n^k \sim \frac{2^k}{Q(\infty)}\sum_{i=0}^{\infty}(-1)^{i+1}\left[\prod_{\ell=1}^{i}\frac{1}{2^\ell-1}\right]\left[\exp\left(-2^{i-\theta}\right)-1\right]. \tag{4.19}$$

We shall use the identity [7]

$$\prod_{\ell=1}^{\infty}\left(1-\frac{u}{2^\ell}\right)=\sum_{i=0}^{\infty}\ u^i\ (-1)^i\ 2^{-i(i+1)/2}\ \left[\prod_{\ell=1}^{i}\left(1-2^{-\ell}\right)\right]^{-1}. \tag{4.20}$$

Setting $u=2^{-s}$ in (4.20) we then multiply (4.20) by $\Gamma(s)2^{\theta s}$ and integrate over a vertical Bromwich contour on which $\Re(s)\in(-1,0)$. Using the definition of $R(i)$ in (2.14) we have

$$\begin{aligned}
-\frac{1}{2\pi i}\int_{Br(-1,0)}\Gamma(s)2^{\theta s}\prod_{\ell=1}^{\infty}(1-2^{-s-\ell})ds &= =\frac{1}{2\pi i}\int_{Br(-1,0)}\Gamma(s)2^{\theta s}\left[\sum_{\ell=0}^{\infty}2^{s\ell}R(\ell)\right]ds\\
&= \sum_{\ell=0}^{\infty}R(\ell)\left[\frac{1}{2\pi i}\int_{Br(-1,0)}\Gamma(s)2^{(\theta-\ell)s}ds\right]\\
&= \sum_{\ell=0}^{\infty}R(\ell)\left[\sum_{m=1}^{\infty}\frac{(-1)^m}{m!}2^{(\ell-\theta)m}\right]\\
&= \sum_{\ell=0}^{\infty}R(\ell)[\exp(-2^{\ell-\theta})-1].
\end{aligned}$$

Thus to expand the right side of (4.19) as $\theta\to\infty$ we must expand the left side of (4.21) in this limit and multiply the result by $2^k/Q(\infty)$. We replace $\Gamma(s)$ by $\pi/[\Gamma(1-s)\sin(\pi s)]$ and use (cf. (4.10))

$$\frac{\prod_{m=0}^{\infty}(1-2^{-s-m})}{\sin(\pi s)}=\frac{2^{s^2/2}2^{-s/2}Q(\infty)}{\prod_{m=1}^{\infty}(1-2^{s-m})}\tilde{A}(s).$$

But the expansion of

$$-\frac{\pi}{2\pi i}\int_{Br(-1,0)}\frac{2^{s^2/2}2^{-s/2}Q(\infty)}{\Gamma(1-s)(1-2^{-s})}\ \ \frac{2^{\theta s}\tilde{A}(s)}{\prod_{m=1}^{\infty}(1-2^{s-m})}ds$$

as $\theta\to\infty$ is essentially the same calculation as (4.16). There is a saddle point where $s\to-\infty$, and in this range $-2^{-s/2}/(1-2^{-s})\sim 2^{s/2}$ and $\prod_{m=1}^{\infty}(1-2^{s-m})\sim 1$. We find, setting $s=\theta+\log_2\theta-v-\frac{1}{2}$, that (4.22) is asymptotically the same as

$$Q(\theta)\sqrt{\pi}2^{-5/8}\theta^{-1}2^{-\theta^2/2}2^{-\theta/2}e^{\theta}e^{-\theta\log\theta}2^{-\log_2^2(\theta)/2}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty}2^{v^2/2}\tilde{A}\left(\theta+\log_2\theta-v-\frac{1}{2}\right)dv,$$

which when multiplied by $2^k/Q(\infty)$ is the same as (2.17). This completes the analysis.

It is interesting to note that the scale $k=\alpha\log_2 n$ with $1<\alpha<\infty$ is needed, as (2.16) and (2.18) do not asymptotically match. Scalings of the form $k=O(\log_2 n)$ often arise in non-symmetric digital trees (tries, PATRICIA tries, DST's), where the probability of a "one" appearing in the string is p and the probability of a zero is $q=1-p$. However, such a scale is frequently not needed in symmetric models, though we did previously encounter this scale in studying the height distribution in DST's [5].

Acknowledgment

The authors thank Helmut Prodinger for helping us to verify that (2.12) satisfies (3.15).

References

[1] D. Aldous and P. Shields, A Diffusion Limit for a Class of Random-Growing Binary Trees, *Probab. Th. Rel. Fields*, 79, 509–542, 1988.

[2] M. Drmota and W. Szpankowski, (Un)Expected Behavior of Digital Search Tree Profile, *Proc. SODA 2009*, 130-138, New York, 2009.

[3] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, 1990.

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

[5] 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*, 30, 923-964, 2000.

[6] D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

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

[8] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Trans. Information Theory*, 41, 478–488, 1995.

[9] M. Naor and U. Wieder, Novel Architectures for P2P Applications: The Continuous-discrete Approach, *Proceedings of the 15th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 2003)*, 50–59, 2003.

[10] G. Park, Profile of Tries, Ph.D. Thesis, Purdue University, 2006.

[11] G. Park, H.K. Hwang, P. Nicodeme, and W. Szpankowski, Profile of Tries, *SIAM J. Computing*, 38, 5, 1821-1880, 2009.

[12] Helmut Prodinger, Digital Search Trees and Basic Hypergeometric Functions, *Bulletin of the EATCS*, 56, 1995.

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

[14] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.