Enumeration of Binary Trees, Lempel-Ziv'78 Parsings, and Universal Types*

Charles Knessl† Wojciech Szpankowski‡

Abstract

Binary unlabeled ordered trees (further called binary trees) were studied at least since Euler, who enumerated them. The number of such trees with n nodes is now known as the Catalan number. Over the years various interesting questions about the statistics of such trees were investigated (e.g., height and path length distributions for a randomly selected tree). Binary trees find an abundance of applications in computer science. However, recently Seroussi posed a new and interesting problem motivated by information theory considerations: how many binary trees of a *given path length* (sum of depths) are there? This question arose in the study of *universal types* of sequences. Seroussi declares that two sequences of length p have the same universal type if they generate the same set of phrases in the incremental parsing of the Lempel-Ziv'78 scheme. (He then proves that sequences of the same type converge to the same empirical distribution.) It turns out that the number of distinct types of sequences of length p corresponds to the number of binary (unlabeled and ordered) trees, T_p, of given path length p (and also the number of different Lempel-Ziv'78 parsings of length p sequences). We first show that the number of binary trees with given path length p is asymptotically equal to $T_p \sim 2^{2p/(\log_2 p)}$. Then we establish various limiting distributions for the number of nodes (number of phrases in the Lempel-Ziv'78 scheme) when a tree is selected randomly among all T_p trees. Throughout, we use methods of analytic algorithmics such as generating functions and complex asymptotics, as well as methods of applied mathematics such as the WKB method and matched asymptotics.

1 Introduction

Trees are the most important nonlinear structures that arise in computer science. Applications are in abundance (cf. [13, 15]); in this paper we discuss a novel application of binary unlabeled ordered trees (further called binary trees) in information theory (e.g., counting Lempel-Ziv'78 parsings and universal types). Tree structures have been the object of extensive mathematical investigations for many years, and many interesting facts have been discovered. Enumeration of binary trees, which are of principal importance to computer science, has been known already by Euler. Nowadays, the number of such trees built on n nodes is called the Catalan number.

Since Euler and Cayley, various interesting questions concerning statistics of randomly generated binary trees were investigated (cf. [6, 13, 15, 22, 24, 25]). In the standard model, one selects uniformly a tree among all binary unlabeled ordered trees, $\mathcal{T}_n$, built on n nodes (where $|\mathcal{T}_n| = \binom{2n}{n}\frac{1}{n+1}$ =Catalan number). For example, Flajolet and Odlyzko [4] and Takacs [24] established the average and the limiting distribution for the height (longest path), while Louchard [16, 17] and Takacs [23, 24, 25] derive the limiting distribution for the path length (sum of all paths from the root to all nodes). As we indicate below, these limiting distributions are expressible in terms of the Airy's function (cf. [1]).

While deep and interesting results concerning the behavior of binary trees in the standard model were uncovered, there are still many important unsolved problems of practical importance. Recently, Seroussi [20], when studying universal types for sequences and distinct parsings of the Lempel-Ziv scheme, asked for the enumeration of binary trees with a *given path length.* Let $\mathcal{T}_p$ be the set of binary trees of given path length p. Seroussi observed that the cardinality of $\mathcal{T}_p$ corresponds to the number of possible parsings of sequences of length p in the Lempel-Ziv'78 scheme, and the number of universal types (that we discuss below). We shall first prove that $T_p = |\mathcal{T}_p| \sim 2^{2p/(\log_2 p)}$ (cf. also Seroussi [21]), and then compute the limiting distribution of the number of nodes (phrases in the LZ'78 scheme) when a tree is selected uniformly among $\mathcal{T}_p$. To the best of our knowledge these problems were never addressed before, with the exception of [20]. We show below that they are much harder than the corresponding problems in the $\mathcal{T}_n$ model.

As mentioned above, the problem of enumerating

*This work was supported by the NSF Grants CCR-0208709, DMS-02-02815, NIH grant R01 GM068959-01, and NSA Grant MDA 904-03-1-0036,

†Dept. Math., Stat. & Computer Science, University of Illinois at Chicago, Chicago, Illinois 60607-7045, U.S.A

‡Department of Computer Science, Purdue University, West Lafayette, IN 47907-2066 U.S.A.

binary trees of a given path arose in Seroussi's research on universal types. The method of types [3] is a powerful technique in information theory, large deviations, and analysis of algorithms. It reduces calculations of the probability of rare events to a combinatorial analysis. Two sequences (over a finite alphabet) are of the same *type* if they have the same empirical distribution. For memoryless sources, the type is measured by the relative frequency of symbol occurrences, while for Markov sources one needs to count the number of pairs of symbols. It turns out (cf. [9]) that the number of sequences of a given Markovian type can be counted by enumerating Eulerian paths in a multigraph. Recently, Seroussi [20] introduced *universal types* (for individual sequences and/or for sequences generated by a stationary and ergodic source). Two sequences of the same length p are said to be of the same universal type if they generate the same set of phrases in the incremental parsing of the Lempel-Ziv'78 scheme. (It is proved that such sequences have the same asymptotic empirical distribution.) But, every set of phrases defines uniquely a binary tree of path length p [8, 20] (with the number of phrases corresponding to the number of nodes in the $\mathcal{T}_p$ model). For example, strings 10101100 and 01001011 have the same set of phrases $\{1, 0, 10, 11, 00\}$ and therefore the corresponding binary trees are the same. Thus, enumeration of $\mathcal{T}_p$ leads to counting universal types and different LZ'78 parsings of sequences of length p.

Let us now summarize our main results. It is easy to see that the generating function $B(z, w) = \sum_{n,p\geq 0} b(n, p)z^n w^p$ of the number $b(n, p)$ of binary trees with n nodes and path length p satisfies the following functional equation [13]

$$B(z, w) = 1 + zB^2(zw, w). \tag{1.1}$$

Observe that this equation is asymmetric with respect to z and w. When enumerating trees in $\mathcal{T}_n$, we set $w = 1$ to get the well known algebraic equation $B(z, 1) = 1 + zB^2(z, 1)$ that can be explicitly solved as $B(z, 1) = \left(1 - \sqrt{1 - 4z}\right)/(2z)$ leading to the Catalan number. However, when enumerating trees of a given path length, $\mathcal{T}_p$, we must substitute $z = 1$ in (1.1) to arrive at

$$B(1, w) = 1 + B^2(w, w)$$

which is not *algebraically* solvable. Observe that $T_p = |\mathcal{T}_p| = \sum_{n=0}^{\infty} b(n, p)$ is the coefficient of $B(1, w)$ at w^p. In fact, the functional equation (1.1) falls into the class of quicksort-like nonlinear functional equations (cf. [7, 5, 15, 10, 18, 19]) that we still do not know how to analyze precisely (with some exceptions like the linear probing algorithm [5, 14]). We shall show (and also give explicitly the error term that involves the Airy function) that

$$T_p = \frac{1}{(\log_2 p)\sqrt{\pi p}} 2^{\frac{2p}{\log_2 p}\left(1+O(\log^{-2/3} p)\right)}$$

for large p. Seroussi first conjectured the form of the leading term in the exponent of the above asymptotic result, proved an upper bound of that form (November 2003, private communication), and has recently obtained [21] a proof for the matching lower bound using information-theoretic and combinatorial arguments for t-ary trees.

In this paper we further analyze the random variable N_p representing the number of nodes in a randomly selected tree from the assembly $\mathcal{T}_p$. We prove that $(N_p - \mathbf{E}[N_p])/\mathbf{Var}\,[N_p]$ is asymptotically normal. Finally, we analyze the number of trees $b(n, p)$ with n nodes and path length p for various ranges of n and p. In passing, we point out that $T_p = |\mathcal{T}_p|$ corresponds to the number of distinct universal types in Seroussi's sense and the number of distinct parsings of binary sequences of length p, while $b(n, p)$ enumerates the number of Lempel-Ziv'78 parsings with n phrases.

Nonlinear functional equations of type (1.1) are not particularly suitable for analytic tools which work fine for linear functional equations (cf. [6, 22]). Therefore, we turn to methods of applied mathematics such as matched asymptotics and the WKB method [2]. These make certain assumptions about the forms of some asymptotic expansions and their asymptotic matching. These are analytic methods and are especially suitable for problems that cannot be solved exactly by transform methods (cf. [10, 11]).

In this conference version of the paper, we are not able to present any details of the proof which is quite long (the final paper [12] is about one hundred pages long). In the next section we present some of our main results and their consequences.

2 Summary of Results

We let $b(n, p)$ denote the number of binary trees with n nodes and path length p. This function satisfies the recurrence relation

$$b(n, p) = \sum_{k+\ell=n-1} \sum_{r+s+n-1=p} b(k, r)b(\ell, s), \quad n \geq 1$$

with the boundary conditions

$$b(0, 0) = 1; \quad b(0, p) = 0, \quad p \geq 1.$$

The generating function

$$B_n(w) = \sum_{p=0}^{\infty} b(n, p)w^p$$

satisfies

$$B_{n+1}(w) = w^n \sum_{\ell=0}^{n} B_\ell(w) B_{n-\ell}(w), \quad n \geq 0 \tag{2.2}$$

with $B_0(w) = 1$. Furthermore, the double transform

$$B(z,w) = \sum_{n=0}^{\infty}\sum_{p=0}^{\infty} b(n,p) w^p z^n = \sum_{n=0}^{\infty} z^n B_n(w)$$

satisfies the functional equation

$$B(z,w) = 1 + zB^2(zw,w). \tag{2.3}$$

We shall mostly analyze (2.2), and then obtain asymptotic results for $b(n,p)$ by expanding the Cauchy integral (cf. [22])

$$b(n,p) = \frac{1}{2\pi i}\int_C B_n(w) w^{-p-1} dw. \tag{2.4}$$

Here C is any closed loop about the origin in the w-plane.

We can solve (2.3) when $w = 1$, noting that $B(0,1) = 1$, to obtain

$$B(z,1) \equiv a(z) = \frac{1}{2z}\left[1 - \sqrt{1-4z}\right]$$

and thus

$$\sum_{p=0}^{\infty} b(n,p) = \frac{1}{2\pi i}\int_C \frac{a(z)}{z^{n+1}} dz = B_n(1) = \frac{1}{n+1}\binom{2n}{n} \tag{2.5}$$

is the Catalan number. This gives the total number of trees with n nodes, regardless of the total path length. By expanding (2.3) about $w = 1$, with

$$\begin{aligned} B_n(w) &= a_n + b_n(w-1) + \frac{1}{2}c_n(w-1)^2 \\ &+ O((w-1)^3) \\ B(z,w) &= a(z) + b(z)(w-1) + \frac{1}{2}c(z)(w-1)^2 \\ &+ O((w-1)^3) \end{aligned}$$

we are led to

$$\begin{aligned} b'(z) &= B_w(z,1) = \frac{2z^2 B(z,1) B_z(z,1)}{1 - 2zB(z,1)} \\ &= \frac{2z^2 a(z) a'(z)}{1 - 2za(z)} \end{aligned}$$

and thus

$$b_n = B_n'(1) \equiv \sum_{p=1}^{\infty} p\, b(n,p) = 4^n - \frac{3n+1}{n+1}\binom{2n}{n}, \quad n \geq 0. \tag{2.6}$$

This gives the average total path length. Higher-order moments can be obtained in a similar manner. In particular, we obtain

$$\begin{aligned} c_n &= B_n''(1) = \sum_{p=2}^{\infty} p(p-1) b(n,p) \\ &= -4^n\left(\frac{13}{2}n + 4\right) + \binom{2n}{n}\left[\frac{10}{3}n^2 + \frac{44}{3}n + 2 + \frac{2}{n+1}\right]. \end{aligned}$$

for $n \geq 0$.

Asymptotically, for $n \to \infty$, we obtain from (2.5), (2.6) and the above via Stirling's formula

$$\begin{aligned} a_n &= \frac{4^n}{\sqrt{\pi} n^{3/2}}[1 + O(n^{-1})] \\ b_n &= 4^n\left[1 - \frac{3}{\sqrt{\pi n}} + O(n^{-1})\right] \\ c_n &= 4^n\left[\frac{10}{3\sqrt{\pi}} n^{3/2} - \frac{13}{2}n + O(\sqrt{n})\right]. \end{aligned} \tag{2.7}$$

We can easily show that for each j

$$\frac{B_n^{(j+1)}(1)}{B_n^{(j)}(1)} = O(n^{3/2}), \quad n \to \infty. \tag{2.8}$$

It is known [16, 17, 23, 24] that the distribution of the total path length L_n, that is,

$$\Pr\{L_n = p\} = \frac{b(n,p)}{\sum_{p=0}^{\infty} b(n,p)} \tag{2.9}$$

follows an Airy distribution as $n \to \infty$, and most of the mass occurs in the range $p = O(n^{3/2})$.

A more difficult problem is to study the distribution of the number of nodes in trees of a fixed path length p, that is, for the assembly $\mathcal{T}_p$. Let N_p be the number of nodes for a tree uniformly generated from $\mathcal{T}_p$. It is a random variable distributed as

$$\Pr\{N_p = n\} = \frac{b(n,p)}{\sum_{n=0}^{\infty} b(n,p)}. \tag{2.10}$$

We shall compute this distribution asymptotically, and also obtain the asymptotic structure of $b(n,p)$ for various ranges of n and p. We note that the above sums are actually finite, since $b(n,p)$ is only non-zero in the range

$$\sum_{J=2}^{n} \lfloor \log_2 J \rfloor = p_{\min}(n) \leq p \leq p_{\max}(n) = \binom{n}{2}. \tag{2.11}$$

Here $p_{\min}$ and $p_{\max}$ are the minimal and maximal total path lengths possible in a tree with n nodes. If we view the problem as having p fixed and varying n, then $b(n,p)$ is non-zero in the range $n \in [n_{\min}(p), n_{\max}(p)]$ where

$$n_{\min}(p) = \min\{n : \binom{n}{2} \geq p\}$$

and

$$n_{\max}(p) = \max\left\{n : \sum_{J=2}^{n} \lfloor \log_2 J \rfloor \leq p\right\}.$$

Asymptotically, for $n \to \infty$,

$$[p_{\min}, p_{\max}] \sim \left[n \log_2 n, \frac{n^2}{2}\right]$$

and, for $p \to \infty$,

$$[n_{\min}, n_{\max}] \sim \left[\sqrt{2p}, \frac{p}{\log_2 p}\right].$$

We now summarize our main results. Our derivations are quite complicated, and are left for the final version of this paper. In passing, we should add that we use ideas of applied mathematics, such as linearization and asymptotic matching. We shall make certain assumptions about the forms of the asymptotic expansions, as well as the asymptotic matching between the various scales. In particular, we shall use the WKB method that we will briefly discuss below.

The WKB method [2, 22] was named after the physicists Wentzel, Kramers and Brillouin. It *assumes* that the solution $B(\xi; n)$ to a recurrence, functional equation or differential equation has an asymptotic solution in the following form for $n \to \infty$

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

where $\phi(\xi)$ and $A(\xi), A^{(1)}(\xi), \ldots$ are unknown functions. These functions must be determined from the equation itself, often in conjunction with another tool known as the *asymptotic matching* principle (cf. [10, 11, 22]).

We next formulate our main result concerning the cardinality of $\mathcal{T}_p$.

Result 1. *The total number of trees of path length p is, for $p \to \infty$*

$$\begin{aligned}(2.12)\quad |\mathcal{T}_p| &= \sum_{n=0}^{\infty} b(n,p) = \frac{1}{(\log_2 p)\sqrt{\pi p}} \\ &\times \exp\left(\frac{2p \log 2}{\log_2 p}\left(1 - \frac{3}{2} A_0 \frac{\log 2}{a^{1/3}} Q^{-2/3}\right.\right. \\ &+ \left.\left. M(Q) Q^{-1} + O(Q^{-4/3})\right)\right)\end{aligned}$$

where $Q = \log p$ and

$$\begin{aligned}(2.13)\quad M(Q) &= (\log Q)(1 + A_1 \log 2) - \log\log 2 \\ &+ (k_2 - A_1 \log a) \log 2,\end{aligned}$$

$$A_0 = \frac{2}{3} 4^{1/3} |r_0| = 2.4743\ldots,$$

and

$$A_1 = \frac{1}{\log 2} - \frac{1}{3} = 1.1093\ldots,$$

$$a = 2(\log 2)^2 = .96090\ldots,$$

and

$$r_0 = \max\{z : Ai(z) = 0\} = -2.3381\ldots.$$

Here $k_2 \approx 3.696$ is obtained by numerically solving a nonlinear integral equation, and $Ai(\cdot)$ is the Airy function [1] defined as a solution of the differential equation $f'' - zf = 0$ that decays as $z \to \infty$.

It follows that the exponential growth rate of the total number of trees of path length p is

$$(2.14)\qquad \log\left[\sum_n b(n,p)\right] \sim \frac{p}{\log p} 2(\log 2)^2$$

with the correction terms involving the least negative root of the Airy function. Recently, Seroussi [21] proved the same result using information theoretic arguments. We will indicate how to formally obtain further terms in the asymptotic series in (2.14).

Defining the mean and variance of N_p by

$$\mathcal{N}(p) := \mathbf{E}[N_p] = \frac{\sum_{n=0}^{\infty} n b(n,p)}{\sum_{n=0}^{\infty} b(n,p)},$$

and

$$\mathcal{V}(p) := \mathbf{Var}\ [N_p] = \frac{\sum_{n=0}^{\infty} (n - \mathcal{N}(p))^2 b(n,p)}{\sum_{n=0}^{\infty} b(n,p)}$$

we shall also obtain, for $p = e^Q \to \infty$,

(2.15)

$$\mathcal{N}(p) = \frac{p}{Q} \log 2 \left[1 - \frac{\log 2}{a^{1/3}} \frac{A_0}{Q^{2/3}} + \frac{M(Q) - A_1 \log 2}{Q} + O(Q^{-4/3})\right]$$

and

(2.16)

$$\mathcal{V}(p) = \frac{p}{Q^{5/3}} \frac{(\log 2) A_0}{6 a^{1/3}} \left[1 - \frac{3A_1}{A_0}\left(\frac{a}{Q}\right)^{1/3} + O(Q^{-2/3})\right]$$

where A_0, A_1 and $M(Q)$ are given below (2.13). Furthermore, the local limiting law is Gaussian, that is,

$$\Pr\{N_p = n\} = \frac{b(n,p)}{\sum_{n=0}^{\infty} b(n,p)} \tag{2.17}$$

$$\sim \frac{1}{\sqrt{2\pi\mathcal{V}(p)}} \exp\left[-\frac{(n-\mathcal{N}(p))^2}{2\mathcal{V}(p)}\right],$$

for $p \to \infty$ and $n - \mathcal{N}(p) = O(\mathcal{V}^{1/2}(p)) = O(\sqrt{p}(\log p)^{-5/6})$. We note that while the most important scale for (2.9) is $p = O(n^{3/2})$, that for (2.10) is

$$p = n\log_2 n + O[n(\log n)^{1/3}],$$

which is close to the lower limit $p_{\min}(n)$ (or upper limit $n_{\max}(p)$) of the support of $b(n,p)$.

The above and preceding findings are derived through the following main technical result. It gives detailed asymptotic results for the solution $B_n(w)$ to (2.2) as $n \to \infty$, for various ranges of w.

Result 2. *Consider binary trees with path length equal to p. Let $B_n(w)$ be its generating function satisfying (2.2). Then for $n \to \infty$ we have the following asymptotic expansions.*

(a) **far right region**: $n \to \infty$, $w > 1$

$$B_n(w) \sim w^{\binom{n}{2}} 2^{n-1} B_*(w) \tag{2.18}$$

where $B_*(w)$ satisfies

(2.19)

$$B_*(w) = 1 + \frac{1}{4w} + \frac{1}{2w^2} + O(w^{-3}), \quad w \to \infty;$$

(2.20)

$$B_*(w) \sim d_1\sqrt{w-1}\,\exp\left(\frac{d_0}{w-1}\right), \quad w \to 1^+,$$

$$\begin{aligned} d_0 &= \int_0^{\log 2} \frac{\xi}{e^{\xi}-1}d\xi = .58224\ldots, \\ d_1 &= \frac{4}{\sqrt{2\pi}}e^{d_0/2} = 2.1350\ldots . \end{aligned}$$

(b) **right region**: $w = 1 + \beta/n$, $0 < \beta < \infty$

$$B_n(w) \sim \sqrt{\frac{\beta}{n}}\,\hat{g}(\beta)\,\exp[n\Phi(\beta)], \tag{2.21}$$

where

$$\begin{aligned} \Phi(\beta) &= \log 2 + \frac{\beta}{2} + \frac{1}{\beta}\int_{-\log\left(1-\frac{1}{2}e^{-\beta}\right)}^{\log 2} \frac{\xi}{e^{\xi}-1}d\xi \\ &\equiv \log 2 + \frac{\beta}{2} + \phi(\beta), \\ \hat{g}(\beta) &= \frac{4}{\sqrt{\pi}}e^{-\beta^2/4}e^{-\beta/2}\left(\frac{1-e^{-\beta}}{2-e^{-\beta}}\right)^{3/2} \\ &\times \exp\left[\frac{1}{2}\beta\phi(\beta) + \frac{1}{2}\beta\log\left(1-\frac{1}{2}e^{-\beta}\right)\right]. \end{aligned}$$

(c) **central region**: $w = 1 + a/n^{3/2}$, $-\infty < a < \infty$

$$B_n(w) = \frac{1}{n+1}\binom{2n}{n} + \tag{2.22}$$

$$+\frac{4^n}{n^{3/2}}\left[C(a) + \frac{1}{\sqrt{n}}C^{(1)}(a) + O(n^{-1})\right],$$

where for $a < 0$

$$\begin{aligned} C(a) &= (-a)\bar{D}((-a)^{2/3}) = Y^{3/2}\bar{D}(Y), \quad Y = (-a)^{2/3}, \\ \bar{D}(Y) &= \frac{1}{2\pi i}\int_{Br} e^{sY}\left[2\sqrt{s} + 4^{2/3}\frac{Ai'(4^{-1/3}s)}{Ai(4^{-1/3}s)}\right]ds. \end{aligned}$$

Here Br is a vertical contour on which $Re(s) > 0$, and $\sqrt{s}$ is analytic for $Re(s) > 0$ and positive for s real and positive. An alternate expression for the leading term is, for $a = -Y^{3/2} < 0$,

(2.23)

$$B_n(w) \sim \frac{4^n}{n^{3/2}}(-a)\frac{d}{dY}\left[\frac{1}{2\pi i}\int_{Br}\frac{4^{2/3}}{s}\frac{Ai'(4^{-1/3}s)}{Ai(4^{-1/3}s)}e^{sY}ds\right]$$

$$= \frac{4^{n+1}}{n^{3/2}}(-a)\sum_{j=0}^{\infty}\exp(-|r_j|4^{1/3}Y)$$

where $0 > r_0 > r_1 > r_2 > \ldots$ and r_j are the roots of $Ai(z) = 0$. The correction term has the integral representation, for $a < 0$,

(2.24)

$$C^{(1)}(a) = -a\bar{D}_1(Y) = \frac{Y^2}{2\pi i}\int_{Br} e^{sY}\mathcal{E}_*(s)ds,$$

$$\mathcal{E}_*(s) = -\frac{5}{2}s + 8\left(\frac{h'(s)}{h(s)}\right)^2 - \frac{4}{h^2(s)}\int_s^{\infty}\frac{(h'(v))^3}{h(v)}dv$$

$$= -\frac{5}{2}s + 10\left(\frac{h'(s)}{h(s)}\right)^2 + 4\left(\frac{h'(s)}{h(s)}\right)^2\log[h(s)] - s\log[h(s)]$$

$$-\frac{1}{h^2(s)}\int_s^{\infty} h^2(v)\log[h(v)]dv, \quad h(s) = Ai(4^{-1/3}s).$$

For $a > 0$ we let $a = y^{3/2}$ with $y > 0$ and the leading term is

(2.25)

$$\begin{aligned} B_n(w) \sim\ & \frac{4^n}{n^{3/2}}\left\{\frac{a}{\pi^2 4^{1/3}}\int_0^{\infty}\frac{e^{\tau y}}{h(\omega\tau)h(\omega^2\tau)}d\tau\right. \\ & \left. - \frac{4a}{\pi}\int_0^{\infty} Re\left[e^{\pi i/6}\frac{h'(\omega\tau)}{h(\omega\tau)}e^{\omega^2\tau y}\right]d\tau\right\} \end{aligned}$$

where $\omega =$ $\exp(2\pi i/3)$.

(d) **left region**: $w = 1 - \gamma/n$, $0 < \gamma < \infty$

(2.26)

$$B_n(w) \sim \frac{4^n}{n} \exp[\nu_0 n^{1/3}\gamma^{2/3} + \nu_1 \gamma \log n] F_0(\gamma)$$

$$F_0(\gamma) = 4\gamma F_1(\gamma),$$

$$\nu_0 = 4^{1/3} r_0 = -4^{1/3}|r_0|, \quad \nu_1 = -\frac{1}{3},$$

where $F_1(\cdot)$ satisfies the non-linear integral equation

(2.27)

$$\frac{e^\gamma - 1}{\gamma} F_1(\gamma) = \int_0^1 F_1(\gamma x) F_1(\gamma - \gamma x) e^{-\gamma H(x)/3} dx,$$

with

$$H(x) = x \log x + (1-x)\log(1-x),$$

and behaves, for $\gamma \to 0^+$, as

$$F_1(\gamma) = 1 - \frac{2}{3}\gamma \log \gamma + \alpha_1 \gamma + O(\gamma)$$

where

$$\begin{aligned} \alpha_1 &= \frac{7}{2} - \gamma_E + \log[h'(s_0)] \\ &- \frac{1}{4[h'(s_0)]^2} \int_{s_0}^\infty h^2(v) \log[h(v)] dv \approx 2.9622, \\ s_0 &= 4^{1/3} r_0, \quad \gamma_E = \text{Euler's constant} = .57721\ldots . \end{aligned}$$

For $\gamma \to \infty$ we have

(2.28)

$$F_1(\gamma) \sim \frac{1}{\sqrt{2\pi \log 2}} \frac{e^{k_2\gamma}}{\sqrt{\gamma}} \exp\left[\left(\frac{1}{3} - \frac{1}{\log 2}\right)\gamma \log \gamma\right]$$

where $k_2 \approx 3.696$ is found numerically.

(e) **far left region***: $n \to \infty$, $0 < w < 1$*

(2.29)

$$\begin{aligned} B_n(w) &\sim e^{n(\log_2 n)\log w} e^{n[g(w) + B_0^*(w,n)]} n^{\log_2 w} \\ &\times (2\pi n)^{-1/2} e^{g(w)} w^{2+\frac{1}{\log 2}} e^{B_0^*(w,n) + B_1^*(w,n)} \\ &\times \sqrt{-\log_2 w - B_1^*(w,n) - \frac{1}{4} B_2^*(w,n)}. \end{aligned}$$

Here

(2.30)

$$\begin{aligned} B_0^*(w,n) &= \sum_{k=-\infty}^{\infty}{}' g_k(w) e^{2\pi i (\log_2 n) k} \\ B_1^*(w,n) &= \frac{2\pi i}{\log 2} \sum_{k=-\infty}^{\infty} k g_k(w) e^{2\pi i (\log_2 n) k} \end{aligned}$$

$$B_2^*(w,n) = \frac{2\pi i}{\log 2} \sum_{k=-\infty}^{\infty} \left[\frac{2\pi i}{\log 2} k^2 - k\right] g_k(w) e^{2\pi i (\log_2 n) k}$$

and $g(w)$ has the asymptotic expansion

$$g(w) = \log 4 + 4^{1/3} r_0 (1-w)^{2/3} +$$

$$+ \left(\frac{1}{\log 2} - \frac{1}{3}\right)(w-1)\log(1-w) - k_2(w-1) + O(w-1)$$

for $w \to 1^-$.

We comment that our analysis suggests that yet another scale exists, which has $n \to \infty$ and $w \to 0$ simultaneously, and where a different expansion for $B_n(w)$ is needed. We have not been able to analyze this scale, but it is not needed to obtain the asymptotic results for the number of trees of a given total path length. For this the important range is the asymptotic matching region between the left and far left regions, corresponding to $w \to 1^-$, but $n(1-w) = \gamma \to +\infty$. Since we have explicit analytic results for $g(w)$ as $w \to 1^-$, and $g_k(w) \to 0$ for $k \neq 0$, we can use the above results to obtain the explicit expressions in (2.12) - (2.16). To obtain the distribution of the path length in trees with n ($\to \infty$) nodes, the central region (c) is the most important, and the leading term corresponds to (the transform of) the Airy distribution.

We next give results for $b(n,p)$ for n and $p \to \infty$, and summarize the main results as items (A)-(E) below. Going from (A) to (E) corresponds to increasing n or decreasing p.

Result 3. *Consider binary trees built over n nodes with path length p. Let $b(n,p)$ denote the number of such trees. Then we have the following for $p, n \to \infty$.*

(A) $\quad n \to \infty, \quad p = \binom{n}{2} - L, \quad L = O(1), \quad L \geq 0$

(2.31) $$b(n,p) \sim 2^{n-1} \frac{1}{2\pi i} \int_C w^{L-1} B_*(w) dw$$

where C is a closed loop with $|w| > 1$ and $B_*(w)$ is as in (a) in Result 2.

(B) $\quad p, n \to \infty$ with $\Lambda = p/n^2 \in \left(0, \frac{1}{2}\right)$

(2.32)

$$b(n,p) \sim \frac{\sqrt{2}}{\pi} \beta_* e^{-\beta_*/2} \left(\frac{1 - e^{-\beta_*}}{2 - e^{-\beta_*}}\right)^{3/2} \left[1 - 2\Lambda - \frac{1}{2e^{\beta_*} - 1}\right]^{-1/2}$$

$$\times \frac{2^{n+1}}{n^2} \exp\left\{n\left[\beta_*(1 - 2\Lambda) - \log\left(1 - \frac{1}{2}e^{-\beta_*}\right)\right]\right\}$$

where $\beta_* \equiv \beta_*(\Lambda)$ is defined implicitly by

$$\beta_*^2 \left(\frac{1}{2} - \Lambda\right) + \beta_* \log\left(1 - \frac{1}{2}e^{-\beta_*}\right) = \int_{-\log\left(1 - \frac{1}{2}e^{-\beta_*}\right)}^{\log 2} \frac{\xi}{e^\xi - 1} d\xi.$$

(C) $\qquad p, n \to \infty$ *with* $\Omega = p/n^{3/2} \in (0, \infty)$

$$b(n,p) \sim -\frac{4^n}{n^3}\left(\frac{1}{3\Omega}\right)^{1/3}\sum_{j=0}^{\infty}\left\{\left[\frac{56}{9}\frac{4^{2/3}r_j^2}{\Omega^3}+\frac{64}{81}\frac{4^{5/3}r_j^5}{\Omega^5}\right]\right.$$
$$\times Ai\left(\frac{r_j^2 4^{2/3}}{3^{4/3}\Omega^{4/3}}\right)+\left(\frac{1}{3\Omega}\right)^{1/3}\left[\frac{40}{3}\frac{4^{1/3}r_j}{\Omega^2}+\frac{64}{27}\frac{4^{4/3}r_j^4}{\Omega^4}\right]$$
$$\left.\times Ai'\left(\frac{r_j^2 4^{2/3}}{3^{4/3}\Omega^{4/3}}\right)\right\}\ \exp\left(-\frac{8|r_j|^3}{27\Omega^2}\right).$$

Here $r_j < 0$ are the roots of the Airy function.

(D) $\qquad p, n \to \infty$ with $\Theta = p/n^{4/3} \in (0, \infty)$

$$\begin{aligned} b(n,p) &\sim \tfrac{4^n}{n^{13/6}} n^{-\gamma_*/3} \tfrac{2^9}{3^4\sqrt{\pi}} \tfrac{|r_0|^{9/2}}{\Theta^5} F_1(\gamma_*) \\ &\quad \exp\left[-\tfrac{16 n^{1/3}|r_0|^3}{27\Theta^2}\right], \\ \gamma_* &= \tfrac{32}{27}\tfrac{|r_0|^3}{\Theta^3} \end{aligned} \tag{2.33}$$

where $F_1(\cdot)$ satisfies (2.27) (cf. item (d) in Result 2).

(E) $\qquad p, n \to \infty$ *with* $p = n\log_2 n + \alpha n,\ \alpha = O(1)$

(2.34)
$$b(n,p) \approx \frac{n^{\log_2 w_*}}{2\pi n}\frac{w_*^{2+\frac{1}{\log 2}}}{\sqrt{\alpha + w_*^2 g''(w_*)}} e^{g(w_*)}\sqrt{-\log_2 w_*}$$
$$\times\ \exp\left[n g(w_*) - n\alpha \log w_*\right]$$

where $w_* = w_*(\alpha)$ is the solution to

$$w_* g'(w_*) = \alpha.$$

In obtaining (2.34) we used (2.29) in (2.4) and neglected the non-constant terms in the Fourier series, which are numerically small. Then we evaluated the integral by the saddle point method (cf. [22]). A refined approximation that uses also the non-constant terms in the Fourier series' in (2.30). We can also obtain an $O(n^{-1/2})$ correction term to the Airy distribution in (C) by using the correction term (i.e., $C^{(1)}$(a)) in (2.24) in asymptotically inverting (2.4). Our approximation(s) to $b(n,p)$ involve the unknown functions $B_*(w)$, $F_1(\gamma)$ and $g(w)$, whose numerical calculation we discuss in the full version of the paper.

In view of the complexity of the results in items (A)-(E), we can get more insight into the structure of $b(n,p)$ by giving formulas that apply in the asymptotic matching regions between the various scales. We summarize these below, with the notation (AB) denoting the asymptotic matching region between the scales in (A) and (B), and so on. Note that the (AB) result can be obtained by either expanding (2.31) as $L \to \infty$ (using (2.20)), or by expanding (2.32) as $\Lambda \to \left(\frac{1}{2}\right)^-$.

Result 4. *The following matching asymptotics hold:*

(AB) $\quad n, p \to \infty;\ L = \binom{n}{2} - p \to \infty,\ \Delta = p/n^2 \to \frac{1}{2}$

$$b(n,p) \sim \frac{2^n}{\pi n^2}\left(\frac{2d_0}{1-2\Delta}\right)^{1/2}$$
$$\times\ \frac{1}{\sqrt{1-2\Delta}}\exp\left[n\sqrt{2d_0(1-2\Delta)} - \frac{1}{2}\sqrt{\frac{2d_0}{1-2\Delta}}\right]$$

where d_0 is given below (2.20).

(BC) $\quad n, p \to \infty;\ \Lambda = p/n^2 \to 0,\ \Omega = p/n^{3/2} \to \infty$

$$\begin{aligned} b(n,p) &\sim \frac{4^n}{n^2}\frac{9\sqrt{3}}{2\pi}\Lambda^2\ \exp\left(-\frac{3}{4}n\Lambda^2\right) \\ &= \frac{4^n}{n^3}\frac{9\sqrt{3}}{2\pi}\Omega^2\ \exp\left(-\frac{3}{4}\Omega^2\right). \end{aligned} \tag{2.35}$$

(CD) $\quad n, p \to \infty;\ \Omega = p/n^{3/2} \to 0,\ \Theta = p/n^{4/3} \to \infty$

(2.36)
$$\begin{aligned} b(n,p) &\sim \frac{4^n}{n^{13/6}}\frac{|r_0|^{9/2}}{\Theta^5}\frac{2^9}{3^4\sqrt{\pi}}\ \exp\left[-\frac{16n^{1/3}|r_0|^3}{27\Theta^2}\right] \\ &= \frac{4^n}{n^3}\frac{|r_0|^{9/2}}{\Omega^5}\frac{2^9}{3^4\sqrt{\pi}}\ \exp\left[-\frac{16|r_0|^5}{27\Omega^2}\right]. \end{aligned}$$

(DE)
$\quad n, p \to \infty;\ \Theta = p/n^{4/3} \to 0,\ \alpha = p/n - \log_2 n \to \infty$

(2.37)
$$b(n,p) \sim \frac{1}{n^{13/6}}\frac{|r_0|^3}{\Theta^{7/2}}\frac{1}{\pi\sqrt{\log 2}}\frac{64}{9\sqrt{3}}\ \exp\left\{n\log 4 - \frac{\gamma_*}{3}\log n\right.$$
$$\left. + \left(\frac{1}{3} - \frac{1}{\log 2}\right)\gamma_*\log\gamma_* + k_2\gamma_* - \frac{16}{27}\frac{n^{1/3}|r_0|^3}{\Theta^2}\right\}$$

where γ_* is given below (2.33).

We will show in the final version of the paper that the asymptotic matching region (DE) leads to the Gaussian distribution in (2.17). Note that in each of the four matching regions, our results are completely explicit functions of n and p. The result in (BC) (resp., (CD)) gives the right (resp., left) tail of the Airy distribution in (C). The right tail has not been characterized this precisely in previous studies [16, 17, 23, 24, 25].

Acknowledgment

We thank Dr. Gadiel Seroussi (HPL, Palo Alto) for motivating the problem, and sharing with us his early results concerning the number of trees of a given path length.

References

[1] M. Abramowitz, and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.

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

[3] I. Csiszár, and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[4] P. Flajolet, and A. Odlyzko, The Average Height of Binary Trees and Other Simple Trees, *J. Computer and System Sciences*, 25, 171–213, 1982.

[5] P. Flajolet, P. Poblete, and A. Viola, On the Analysis of Linear Probing Hashing, *Algorithmica*, 22, 490–515, 1998.

[6] P. Flajolet and R. Sedgewick, *Analytical Combinatorics*, in preparation; see also INRIA TR-1888 1993, TR-2026 1993 and TR-2376 1994.

[7] P. Hennequin, Combinatorial Analysis of Quicksort Algorithm, *Theoretical Informatics and Applications*, 23, 317–333, 1989.

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

[9] P. Jacquet and W. Szpankowski, Precise Worst Case Minimax Redundancy for Markov Sources *IEEE Trans. Information Theory*, 50, 2004.

[10] C. Knessl and W. Szpankowski, Quicksort algorithm again revisited, *Discrete Math. Theor. Comput. Sci.*, 3 43–64, 1999.

[11] 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.

[12] C. Knessl and W. Szpankowski, Enumeration of Binary Trees and Universal Types, preprint 2004 (cf. `http://www.cs.purdue.edu/people/spa`).

[13] D. E. Knuth, *The Art of Computer Programming. Fundamental Algorithms,* Vol. 1, Third Edition, Addison-Wesley, Reading, MA, 1997.

[14] D. E. Knuth, Linear Probing and Graphs, *Algorithmica*, 22, 561–568, 1998.

[15] D. E. Knuth, *Selected Papers on the Analysis of Algorithms*, Cambridge University Press, Cambridge, 2000.

[16] G. Louchard, Kac's Formula, Lévy Local Time and Brownian Excursion, *J. Appl. Probab.*, 21, 479-499, 1984.

[17] G. Louchard, The Brownian Excursion Area: A Numerical Analysis, *Comp. & Maths. with Appls.*, 10, 413-417, 1984.

[18] M. Régnier, A Limiting Distribution for Quicksort, *Theoretical Informatics and Applications*, 23, 335–343, 1989.

[19] U. Rösler, A Limit Theorem for Quicksort, *Theoretical Informatics and Applications*, 25, 85–100, 1991.

[20] G. Seroussi, On Universal Types, *Proc. ISIT 2004*, pp. 223, Chicago, 2004.

[21] G. Seroussi, "On the Number of t-ary Trees with a Given Pathlength".

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

[23] L. Takács, A Bernoulli Excursion and its Various Applications, *J. Appl. Probab.*, 23, 557-585, 1991.

[24] L. Takács, Conditional Limit Theorems for Branching Processes, *J. Applied Mathematics and Stochastic Analysis*, 4, 263-292, 1991.

[25] L. Takács, The Asymptotic Distribution of the Total Heights of Random Rooted Trees, *Acta Sci. Math. (Szegad)*, 57, 613-625, 1993.