The Distribution of heights of binary trees and other simple trees

Philippe Flajolet, Zhicheng Gao, Andrew M. Odlyzko, Bruce Richmond

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.: (1) 39 63 55 11

Rapports de Recherche

1992 25ème anniversaire

N° 1749

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

THE DISTRIBUTION OF HEIGHTS OF BINARY TREES AND OTHER SIMPLE TREES

Philippe FLAJOLET
Zhicheng GAO
Andrew ODLYZKO
Bruce RICHMOND

Septembre 1992

RR-1749

The Distribution of Heights of Binary Trees and Other Simple Trees

Philippe Flajolet, Zhicheng Gao,
Andrew Odlyzko, Bruce Richmond

Abstract. *The number of binary trees of fixed size and given height is estimated asymptotically near the peak of the distribution. There, a local limit theorem with convergence to a theta law is established. Large deviation bounds corresponding to large heights and small heights are also derived. The methods based on the analysis of singular iterations apply to any simple family of trees.*

La distribution de hauteur dans les arbres binaires et autres familles simples d'arbres

Résumé. Le nombre d'arbres binaires de taille fixe et de hauteur bornée est estimé asymptotiquement au voisinage du pic de la distribution. Là est prouvé un théorème de limite locale avec convergence vers une loi theta. Des bornes de grandes déviations correspondant à des hauteurs petites ou grandes sont aussi obtenues. Les méthodes fondées sur l'analyse d'itérations singulières s'appliquent à toute famille simple d'arbres.

The Distribution of Heights of Binary Trees and Other Simple Trees

Philippe Flajolet
Algorithms Project
INRIA Rocquencourt
78153 Le Chesnay
France

Zhicheng Gao
Dept. of Mathematics and Statistics
Carleton University
Ottawa, Ontario K1S5B6
Canada

Andrew Odlyzko
AT&T Bell Laboratories
Murray Hill, New Jersey 07974
United States

Bruce Richmond
Dept. of Combinatorics and Optimization
University of Waterloo
Waterloo, Ontario, N2L3G1
Canada

September 12, 1992

The number, $B_n^{[h]}$, of the binary trees of height $\leq h$ with n internal nodes is shown to satisfy

$$\frac{B_n^{[h]} - B_n^{[h-1]}}{\sum_h (B_n^{[h]} - B_n^{[h-1]})} \sim \begin{cases} 2\sqrt{\pi/n}\beta^4 \sum_{m\geq 1} (m\pi)^2 (2(m\pi\beta)^2 - 3) e^{-(m\pi\beta)^2} \\ 2/(\beta\sqrt{n}) \sum_{m\geq 1} m^2 (2(m/\beta)^2 - 3) e^{-(m/\beta)^2}, \end{cases}$$

uniformly for $\beta, 1/\beta \leq \delta\sqrt{\log n}$, where $\beta = 2\sqrt{n}/h$ and δ is a positive constant. An asymptotic formula for $B_n^{[h]} - B_n^{[h-1]}$ is derived for $h = cn$ where $0 < c < 1$. Bounds for $B_n^{[h]}$ are also derived for large heights and small heights. The methods apply to any simple family of trees and the general asymptotic results are stated.

1 Introduction

We study rooted plane trees. The height of a tree is the number of nodes in a longest path from the root node to another node. Height is an important parameter of a tree, and so it has often been investigated [BS, BKR, FO, L, MM, RS]. For example, in some tree traversal algorithms, the height of the tree is the size of the stack used by the algorithm. In most situations in combinatorics and computer science the average height of a tree has been of greatest interest. However, it is often important to obtain more detailed results about the distribution of heights, for example to be able to estimate how often pathologically bad cases arise. This paper proves several results in this area.

A tree is called *binary* if all internal nodes have two successors. Let $B_n^{[h]}$ be the number of binary trees with n internal nodes and height $\leq h$. It is well known that $B_n = \sum_h (B_n^{[h]} - B_n^{[h-1]})$ is the nth Catalan number and that

$$B_n = \frac{1}{\sqrt{\pi n^3}} 4^n (1 + O(1/n)).$$

We prove a local limit theorem for the distribution defined by the $B_n^{[h]} - B_n^{[h-1]}$, assuming all binary trees with n internal nodes are equally likely.

Theorem 1 *The numbers $B_n^{[h]}$ satisfy*

$$\frac{B_n^{[h]} - B_n^{[h-1]}}{B_n} \sim \begin{cases} 2\sqrt{\pi/n}\beta^4 \sum_{m\geq 1}(m\pi)^2(2\beta^2(\pi m)^2 - 3)e^{-\pi^2 m^2 \beta^2} \\ 2/(\beta\sqrt{n}) \sum_{m\geq 1} m^2(2m^2/\beta^2 - 3)e^{-m^2/\beta^2}, \end{cases}$$

uniformly for $\beta, 1/\beta \leq \delta\sqrt{\log n}$, where $\beta = 2\sqrt{n}/h$.

Here and in the rest of the paper, δ denotes a positive constant, not necessarily the same at each occurrence. Note that the formulas in Theorem 1 are equivalent by Poisson's formula (cf. [3, p. 52]).

Corollary 1 (Flajolet and Odlyzko [FO, Theorem B]) *The average height, $\bar{H}_n(B)$, of binary trees with n internal nodes satisfies*

$$\bar{H}_n(B) \sim 2\sqrt{\pi n} \quad \text{as } n \to \infty .$$

In [FO], Flajolet and Odlyzko showed how their proof of Corollary 1 could be extended by using the method of moments to a proof of Theorem 1 for β and $1/\beta$ bounded. However, that approach is inherently incapable of yielding a wider range of validity for the approximation. Approximations such as those of Theorem 1 for β and $1/\beta$ bounded had been obtained also for various families of trees by ex-Soviet mathematicians (see [K] for results and references), but it is not clear to what extent those methods can be extended to cover wider ranges of β.

Brown and Shubert [BS] claimed to prove that the approximation of Theorem 1 is valid for $n^{3/8+\delta} \leq h \leq n$. However, there are doubts about the validity of their proof. One problem is that they use the approach of Rényi and Szekeres [RS], which is based on the paper of Szekeres [S] on iteration of analytic functions. Szekeres assumes in [S] that the analytic functions he deals with have real Taylor series coefficients. That is not true for some of the functions in [BS, RS]. Presumably the results of [S] hold more generally, but there is no proof in the literature, although Rényi and Szekeres [RS] outline how to do this. A more serious problem with the Brown-Shubert paper [BS] is that the proof of the key inequality (2.19) is wrong, and that the inequality itself is incorrect. For example, taking $\epsilon < 1/12$ and $\varphi = 0$ in that inequality implies that the $f_k(1/4)$ converge to $\omega(1/4) = 2$ exponentially in k, whereas it is proved in [FO] that the difference $f_k(1/4) - \omega(1/4)$ decreases at the rate of $1/k$. (The proof of this fact is simple, since for $\varphi = 0$ only iterations of real functions are involved.)

Our methods are a refinement of the estimates in [FO] and an adaptation of the arguments of Rényi and Szekeres [RS]. Our proof of Theorem 1 is simpler than those of Rényi and Szekeres [RS] and Brown and Shubert [BS] in that we do not need accurate estimation for the iterations near a fixed point. We prefer to develop the self-contained analysis of Flajolet and Odlyzko [FO] since little more needs be done. It seems that one could derive the results of Szekeres [S, Lemma 5] for the iteration of the generating functions of simple families of trees by developing the analysis in [FO] (using techniques of De Bruijn [BD, p. 157]). It is not necessary to do this for our theorems so we have not done so. The main advantage of the analysis of [FO] is that it is clear how the extension to simple families of trees, as defined by Meir and Moon [MM], is done. For our paper

we may extend their definition as follows : Let $y_n^{[h]}$ be the number of trees in a family with n nodes and height $\leq h$. The family is simple if the generating functions

$$y^{[h]}(z) = \sum_n y_n^{[h]} z^n$$

satisfy

$$y^{[0]}(z) = 0,\ y^{[h+1]}(z) = z\phi(y^{[h]}(z)).$$

We assume that the coefficients $[z^r]\phi(z)$ are bounded. Some families of trees that satisfy such functional equations are (examples from [FO])

(a) the family of binary trees (in Theorem 1).

(b) the family of general planar trees; the analysis of De Bruijn, Knuth and Rice [BKR] gives the average height.

(c) the family of unary-binary trees; they appear as shapes of expression trees when unary as well as binary operations are allowed, and are counted by the Motzkin numbers.

(d) the family of 2-3 trees (unbalanced); their balanced counterparts are a useful data structure and have been counted by Odlyzko [O].

(e) the family of t-ary trees (which appear in digital search).

(f) the family of nonplanar labelled trees (when the derivative of the exponential generating function is considered, this is not a simple family in the sense of Meir and Moon).

We show how the proof of Theorem 1 can be extended to derive the following result.

Theorem 2 *For simple families of trees corresponding to the equation*

$$y = z\phi(y),\ \phi(y) = \sum c_r y^r$$

and for

$$n \equiv 1 \pmod{d} \text{ with } d = \gcd\{r : c_r \neq 0\}$$

the numbers $y_n^{[h]}$ *satisfy, with* $y_n = \sum_h y_n^{[h]}$, τ *the smallest positive solution of*

$$\phi(\tau) - \tau\phi'(\tau) = 0$$

and

$$c = \sqrt{2\phi\phi''}/\phi' \text{ and } \beta = 2\sqrt{n}/(c\tau),$$

the relation

$$\frac{y_n^{[h]} - y_n^{[h-1]}}{y_n} \sim \begin{cases} 2c\sqrt{\pi/n}\beta^4 \sum_{m\geq 1}(m\pi)^2(2(m\pi\beta)^2 - 3)e^{-(m\pi\beta)^2} \\ 2c/(\beta\sqrt{n}) \sum_{m\geq 1} m^2(2(m/\beta)^2 - 3)e^{-(m/\beta)^2} \end{cases}$$

uniformly for $\beta, 1/\beta \leq \delta\sqrt{\log n}$.

Remarks :

1. If $n \not\equiv 1 \pmod{d}$ then $y_n^{[h]} = 0$ for all h.

2. From [FO],
$$y_n \sim dc_1\rho^{-n}n^{-3/2}, \rho = \tau/\phi(\tau), c_1 = \sqrt{\phi(\tau)/(2\pi\phi''(\tau))}.$$

3. For binary trees counted according to n nodes (internal plus external) $\phi(y) = 1 + y^2$, $y(z) = zB(z^2)$ [FO], so $\tau = 1$ and $c = \sqrt{2}$. Also, n internal nodes gives $2n + 1$ nodes in total. Our formula in Theorem 2 with n replaced by $2n + 1$ gives the formula of Theorem 1.

4. For labelled general trees, $y_n = n^{n-2}$, $\phi(y) = e^y$, so $\tau = 1$ and $c = \sqrt{2}$ and we get the Rényi-Szekeres formula [RS, (3.31)].

Corollary 2 (Flajolet and Odlyzko [FO, Theorem S]) *The average height, $\bar{H}_n$, satisfies*
$$\bar{H}_n \sim \lambda\sqrt{n}, \quad \lambda = \sqrt{2\pi/(\phi(\tau)\phi''(\tau))}\phi'(\tau)$$

Theorems 1 and 2 deal with the distribution of heights near the average. In Section 3 we consider large and small heights and prove upper bounds for them.

Theorem 3 *The number of binary trees with n internal nodes and height h, for $1 \le h \le n$, satisfies*
$$B_n - B_n^{[h]} = O\left(B_n n^{3/2} e^{-h^2/(4n)}\right),$$
and
$$B_n^{[h]} = O\left(B_n n^{3/2} e^{-\delta n/h^2}\right).$$

Remarks :

1. By Theorem 1, the number of trees of height h for $\delta/\log n \le n/h^2 = o(1)$ is
$$\sim \frac{B_n h^3}{2n^2} e^{-h^2/(4n)},$$
so the bound in Theorem 3 is quite sharp. When $h = n$ however, the bound is poor. The bounds in Theorem 4 are much better for $h = cn$, $0 < c < 1$.

2. The proof of Theorem 3 generalizes easily to obtain similar bounds for any simple family.

3. It is possible to derive comparable bounds for large h using the relation between height and the number of nodes at a given altitude. The number of nodes at a given altitude is easier to analyze than height, see [MM] for example. It is possible to derive an accurate estimate for the mean height of a simple family of trees.

We also mention in Section 2 that it is possible to prove that $B_n^{[h]} - B_n^{[h-1]}$ is monotonic increasing then decreasing near the peak (that is for h^2/n, $n/h^2 \le \delta \log n$). This answers a question raised by Wimp [W].

There are other approaches to studying extremely large or small heights. B. Pittel in an unpublished work obtained upper bounds somewhat weaker than those of Theorem 3 for trees of small heights by probabilistic and combinatorial arguments. Luczak [L] found a method for rigorously extrapolating the Rényi and Szekeres results [RS] about general labeled trees to all values of h such that $h/\sqrt{n} \to \infty$. In Section 4 we will present yet another method and will show how to apply local limit theorems to obtain results such as the following.

Theorem 4

$$B_n^{[h]} - B_n^{[h-1]} \sim \frac{4\epsilon^2 A(\epsilon)}{(1-\epsilon)^2\sqrt{\pi(1+\epsilon)n}} \left((1-\epsilon)^{(1-\epsilon)}(1+\epsilon)^{(1+\epsilon)}\right)^{-h/2\epsilon} 4^n$$

uniformly for all h such that $h/n = 2\epsilon/(1+\epsilon)$ with $\epsilon \in [\delta',\ 1-\delta']$, where δ' is a positive constant which can be arbitrarily small and $A(\epsilon)$ is a positive and continuous function for $\epsilon \in [\delta',\ 1-\delta']$.

2 Densities for heights near the mean

We begin with some notation from [FO]. Let

$$B^{[h]}(z) = \sum_{n\geq 0} B_n^{[h]} z^n, \quad B(z) = \sum_{n\geq 0} B_n z^n,$$

and

$$e_h(z) = (B(z) - B^{[h]}(z))/2B(z)\ . \tag{1}$$

Then

$$e_{h+1}(z) = (1-\epsilon(z))e_h(z)(1-e_h(z)),\ e_0(z) = 1/2\ , \tag{2}$$

and

$$B(z) = (1-\epsilon)/2z\ , \tag{3}$$

where $\epsilon = \epsilon(z) = \sqrt{1-4z}$, the determination of $\sqrt{1-4z}$ being positive for real $z < 1/4$. It is an easy consequence of [FO, Lemmas 1–7] that

$$|e_h(z)| \leq c|1-\epsilon(z)|^h \tag{4}$$

for some constant c and all $|z| \leq 1/4$. As Brown and Shubert [BS] point out, it is easily seen that if $z = e^{it}/4$, then $|1-\epsilon(z)|$ is a decreasing function of t on $0 \leq |t| \leq \pi$ with maximum 1 at $t = 0$ and minimum $\sqrt{2}-1$ at $|t| = \pi$.

We shall, with Brown and Shubert [BS], follow Rényi-Szekeres [RS] and investigate $e_h(z)$ for $z = e^{it}/4$. We first consider

$$|t| \leq \rho^2/h^2\ , \tag{5}$$

where $\rho = \rho(h) = h^\delta$. For this range of t the first equality in the proof of Lemma 8 of [FO] is

$$(1-\epsilon)^h/e_h = (1-(1-\epsilon)^h)/\epsilon + O(\log|\epsilon|^{-1})\ , \tag{6}$$

and this estimate is uniform in $\{z = e^{it}/4 : |t| \leq \delta\}$. All the bounds below will be uniform in the range of t in (5). We have

$$\epsilon = \sqrt{-it}\sqrt{1+it/2+\cdots} = \sqrt{-it}(1+O(\rho^2/h^2)) = O(\rho/h)\ . \tag{7}$$

Now

$$(1-\epsilon)^h = \exp(-h\epsilon + O(h\epsilon^2)) = \exp(-h\epsilon)\left(1+O(h|\epsilon|^2)\right)\ . \tag{8}$$

From (6), (7), and (8), we see that for $h^{-20} \le |t| \le \rho^2/h^2$,

$$\begin{aligned} e_h(z) &= \frac{e^{-h\epsilon}(1+O(\rho^2/h))}{(1-e^{-h\epsilon}(1+O(h|\epsilon|^2))/\epsilon + O(\log|\epsilon|^{-1})} \\ &= \frac{\epsilon e^{-h\epsilon}}{1-e^{-h\epsilon}} + O(\rho^2/h^2)\,. \end{aligned} \tag{9}$$

Setting $it = 4\tau^2/h^2$, we obtain from (7) and (9), again for $h^{-20} \le |t| \le \rho^2/h^2$,

$$\begin{aligned} e_h(z) &= \frac{i2\tau}{h}\frac{e^{-i2\tau}}{1-e^{-i2\tau}} + O(\rho^2/h^2) \\ &= \frac{\tau}{h}(\cot(\tau) - i) + O(\rho^2/h^2)\,. \end{aligned} \tag{10}$$

Using (2), (7), and (10), we obtain the estimate, valid for $h^{-20} \le |t| \le \rho^2/h^2$,

$$e_{h-1}(e^{it}/4) - e_h(e^{it}/4) = \frac{\tau^2}{h^2\sin^2(\tau)} + O(\rho^3/h^3)\,. \tag{11}$$

Noting

$$B(z) = 2e^{-it}(1-\epsilon) = 2(1 - i2\tau/h) + O(\rho^2/h^2)\,, \tag{12}$$

we obtain from (1) and (11) that for $h^{-20} \le |t| \le \rho^2/h^2$,

$$B^{[h]}(e^{it}/4) - B^{[h-1]}(e^{it}/4) = \frac{4\tau^2}{h^2\sin^2(\tau)} + O(\rho^3/h^3)\,. \tag{13}$$

This agrees with [BS, (2.33)]. (Note our $\tau = \xi/2$.) Now we write

$$\begin{aligned} B_n^{[h]} - B_n^{[h-1]} &= \frac{1}{2\pi i}\int_{|z|=1/4}\left(B^{[h]}(z) - B^{[h-1]}(z)\right) z^{-n-1}dz \\ &= \frac{4^n}{2\pi}\int_{-\pi}^{\pi}\left(B^{[h]}(e^{it}/4) - B^{[h-1]}(e^{it}/4)\right) e^{-int}dt\,. \end{aligned}$$

Using (4) and the comments immediately following (4), we obtain

$$\begin{aligned} & B_n^{[h]} - B_n^{[h-1]} \\ = \; & \frac{4^n}{2\pi}\int_{|t|\le\rho^2/h^2}\left(B^{[h]}(e^{it}/4) - B^{[h-1]}(e^{it}/4)\right) e^{-int}dt + O(4^n(1-\rho/(2h))^h) \\ = \; & \frac{4^n}{2\pi}\int_{|t|\le\rho^2/h^2}\left(B^{[h]}(e^{it}/4) - B^{[h-1]}(e^{it}/4)\right) e^{-int}dt + O(4^n e^{-\rho/2}). \end{aligned} \tag{14}$$

We now use (13) for $h^{-20} \le |t| \le \rho^2/h^2$, and the bound (4) for $|t| \le h^{-20}$. (Eq. (13) can be shown to be valid for all $|t| \le \rho^2/h^2$, but we do not need to use this.) When we substitute $4\tau^2/h^2 = it$, we can rewrite (14) as

$$\begin{aligned} & B_n^{[h]} - B_n^{[h-1]} \\ = \; & \frac{4^{n+2}}{\pi i h^4}\int_\Gamma \tau^3 e^{-4n\tau^2/h^2}/\sin^2(\tau)d\tau + O(4^n(\rho^5/h^5 + e^{-\rho/2}))\,, \end{aligned}$$

where

$$\Gamma = \{xe^{-i\pi/4} : x \text{ from } \rho/2 \text{ to } 0\} \cup \{xe^{i\pi/4} : x \text{ from } 0 \text{ to } \rho/2\} .$$

It now follows by a standard argument that

$$\begin{aligned} & B_n^{[h]} - B_n^{[h-1]} \\ = \; & \frac{32 \cdot 4^n}{h^4} \sum_{p=1}^{[\rho/2\pi]} \mathrm{Res}(\tau^3 e^{-\beta^2\tau^2} / \sin^2 \tau, \tau = p\pi) \\ & + O(4^n(\rho^5/h^5 + e^{-\rho} + e^{-\beta^2\rho}/h^4)) \\ = \; & \frac{32 \cdot 4^n}{h^4} \sum_{m=1}^{[\rho/2\pi]} (2\beta^2(m\pi)^4 - 3(m\pi)^2)e^{-(m\pi\beta)^2} \qquad (15) \\ & + O(4^n(\rho^5/h^5 + e^{-\rho} + e^{-\beta^2\rho}/h^4)), \end{aligned}$$

where $\beta = 2\sqrt{n}/h$ and $O(e^{-\beta^2\rho})$ comes from the integral along the arc

$$\{(\rho/2)e^{i\theta} : -\pi/4 \le \theta \le \pi/4\}.$$

Noting $\rho = h^\delta$, we obtain

$$B_n^{[h]} - B_n^{[h-1]} \sim \frac{32 \cdot 4^n}{h^4} \sum_{m \ge 1} (2\beta^2(m\pi)^4 - 3(m\pi)^2)e^{-(m\pi\beta)^2} \qquad (16)$$

uniformly for $\delta \le \beta \le \delta\sqrt{\log n}$. Using Poisson's formula (cf. [BD, p. 52]), we have

$$B_n^{[h]} - B_n^{[h-1]} \sim \frac{4^n 32}{\sqrt{\pi}\beta^5 h^4} \sum_{m \ge 1} m^2(2(m/\beta)^2 - 3)e^{-(m/\beta)^2} \qquad (17)$$

uniformly for $\delta \le 1/\beta \le \delta\sqrt{\log n}$. Now Theorem 1 follows from (16), (17) and the classical formula

$$B_n \sim \frac{4^n}{\sqrt{\pi n^3}}(1 + O(1/n)).$$

Using (15) and Poisson's formula, one can easily deduce

$$\frac{B_n^{[h]} - B_n^{[h-1]}}{B_n} = O\left(n^{-1/2-\delta}\right) \qquad (18)$$

for $h \ge \delta\sqrt{\log n}$ or $h \le \delta\sqrt{\log n}$, which, together with Theorem 1, gives Corollary 1.

We now give the proof of Theorem 2, relying heavily on the results of [FO, Section 6]. There we find that

$$[z^n]y(z) \sim dc_1\rho^{-n}n^{-3/2}, \; c_1 = \sqrt{\phi(\tau)/(2\pi\phi''(\tau))}.$$

Furthermore, if

$$e_h(z) = y(z) - y^{[h]}(z)$$

then

$$e_{h+1}(z) = (1 - \epsilon)e_h(z)(1 - \phi''(\tau)e_h(z)/2\phi'(\tau)) + O(|e_h^2(z)| + |e_h(z)||y - \tau|),$$

where

$$\epsilon(z) = 1 - z\phi'(y) = dc(1 - z/\rho)^{1/2} + O((y-\tau)^2),$$
$$c = \sqrt{2\phi''\phi}/\phi', \text{ and } (y-\tau)^2 = O(1 - z/\rho).$$

The analysis is now similar to that for binary trees. We find that

$$e_h(z) = c_2 \frac{(1-\epsilon(z))^h}{(1-(1-\epsilon(z))^h)/\epsilon(z) + O(\log|\epsilon(z)|^{-1})},$$

where $c_2 = 2\phi'(\tau)/\phi''(\tau)$. The analysis of this paper is now easily extended. We consider $z = \rho e^{it}$, $|t| \le h^{\delta-2}$, $\beta = 2\sqrt{n}/(ch)$ and find that

$$\frac{y_n^{[h]} - y_n^{[h-1]}}{y_n} \sim 2c\sqrt{\pi/n}\beta^4 \sum_{m\ge 1} (m\pi)^2(2(m\pi\beta)^2 - 3)e^{-(m\pi\beta)^2}$$

which gives Theorem 2.

The proof of Corollary 2 is essentially the same as that of Corollary 1.

3 Bounds for large and small heights

First we make some observations about extending the range of h for which we can obtain asymptotic estimates. Note that given an approximation for $y^{[h]}(z)$ of the form

$$\begin{aligned} & c_0 + c_1(\tau)/h + (c_2(\tau)\log h + c_3(\tau))/h^2 \\ + \; & (c_4(\tau)\log^2 h + c_5(\tau)\log h + c_6(\tau))/h^3 + \cdots \end{aligned} \tag{19}$$

it is possible to use the method of variable coefficients as do Rényi and Szekeres to find $c_i(\tau)$ explicitly. One can then obtain an approximation for $y^{[h]} - y^{[h-1]}$ to an accuracy of

$$(\text{polynomial in } \log h \text{ of degree } l-1)/h^l \; .$$

From such an approximation it is possible to obtain asymptotic estimates for $y^{[h]}(z) - y^{[h-1]}(z)$ for h in the range

$$h^2/n,\; n/h^2 \le c_h \log n, \quad c_h \to \infty \text{ with } h.$$

It is possible to derive the required approximations for $y^{[h]}(z)$, but it is tedious to do so. The approximations seem complicated.

It may be more interesting to observe that if such an approximation is determined to $O((\log h)^m h^{-3})$ accuracy it is easy to show the second difference of $y_n^{[h]}$ has the same sign as the derivative of the density functions in Theorems 1 and 2. To do this is rather routine so we simply sketch the proof. Eq. (19) allows us to refine (13) and hence (16). (It suffices to know the h^{-2} term in (19).) The

$$(c_2(\tau)\log h + c_3(\tau))/h^2$$

term in (19) gives a complicated expression that can be evaluated as a sum of residues when it is used to refine (13). The higher terms in (19) are negligible as we shall see.

If h is incremented by one then the summation in (16) is altered by its derivative plus a second derivative term. The derivative of β^2 is of the form cnh^{-3}, c a constant. For the relevant h, that is near $\sqrt{n}$, such a term introduces a factor of h^{-1}, a second derivative term a factor of h^{-2}. Thus

$$(y_n^{[h]} - 2y_n^{[h-1]} + y_n^{[h-2]})/y_n$$

is asymptotically the derivative of the density function of the Θ-distribution of Theorem 2 plus terms smaller by a factor of h^{-1}. Since the limiting Θ-distribution is unimodal, see the graph in Brown and Shubert [BS], our claim about the unimodality follows. It would be possible in principle to derive asymptotic expressions to any accuracy of the form n^{-M} for $\bar{H}_n$ from (19). These expressions become complicated.

It is possible to derive upper bounds for $y_n^{[h]}$ for h far away from $\sqrt{n}$. We illustrate with $B_n^{[h]}$. Since the coefficients of $B(z) - B^{[h]}(z)$ are ≥ 0, we find that for every real z, $0 < z < 1/4$,

$$B_n - B_n^{[h]} \leq (B(z) - B^{[h]}(z))z^{-n} ,$$

and so by (4),

$$B_n - B_n^{[h]} = O((1 - \epsilon(r))^h (1 - \epsilon^2(r))^{-n} 4^n) \tag{20}$$

for $0 < r < 1/4$. To minimize the expression in (20) we set the logarithmic derivative with respect to ϵ to zero. We find

$$h/(1 - \epsilon) = 2n\epsilon/(1 - \epsilon^2), \quad \epsilon = h/(2n - h).$$

So

$$(1 - \epsilon)^h (1 - \epsilon^2)^{-n} = O\left((1 - \epsilon)^{(1-\epsilon)/2\epsilon} (1 + \epsilon)^{(1+\epsilon)/2\epsilon}\right)^{-h} .$$

Now the first part of Theorem 3 follows from

$$(1 - \epsilon)^{(1-\epsilon)/2\epsilon} (1 + \epsilon)^{(1+\epsilon)/2\epsilon} \geq e^{\epsilon/2} .$$

The second part can be derived by an argument similar to that of Wright, Richmond, Odlyzko and McKay [WROM]. For $h \geq 2$, let x_h be defined by

$$B^{[h]}(x_h) = B(1/4) = 2 .$$

It is clear that $1 = x_2 > x_3 > \cdots > 1/4$. We use $Df(u)$ to denote the derivative of $f(z)$ at $z = u$.

Lemma 1 *There exist positive constants h_0 and $c > 0$ such that for $h \geq h_0$,*

$$DB^{[h]}(1/4) \geq ch .$$

Proof: The proof is by induction on h. The claim is clearly true for $h \leq 8$ if c is small enough. Suppose it is true for h. From [FO, Eq. (8)]

$$1/(4h) < e_h(1/4) < 1/h ,$$

and hence

$$1/h < 2 - B^{[h]}(1/4) < 4/h . \tag{21}$$

Using (21) and

$$B^{[h+1]}(z) = 1 + z(B^{[h]}(z))^2 ,$$

we obtain

$$\begin{aligned} DB^{[h+1]}(1/4) &= (B^{[h]}(1/4))^2 + (1/2)B^{[h]}(1/4)DB^{[h]}(1/4) \\ &\geq (2-4/h)^2 + (1-2/h)DB^{[h]}(1/4) \geq 4 - 16/h + (1-2/h)ch \geq c(h+1) \end{aligned}$$

provided $h \geq h_0 = 8$ and $c \leq 1/6$. ∎

Lemma 2 *There exist positive constants C and h_1 such that for $h \geq h_1$*

$$DB^{[h]}(x_h) \leq Ch \tag{22}$$

and

$$x_h \geq 1/4 + 1/(Ch^2)\ . \tag{23}$$

Proof: Using Lemma 1, (21) and

$$2 - B^{[h]}(1/4) = B^{[h]}(x_h) - B^{[h]}(1/4) \geq DB^{[h]}(1/4)(x_h - 1/4)\ ,$$

we have

$$x_h - 1/4 \leq 4/(hDB^{[h]}(1/4)) \leq 4/(ch^2)\ . \tag{24}$$

Now for $h_1 \geq 8/c$ and $C \geq \max\{8, DB^{[h_1]}(1/4)/h_1\}$,

$$\begin{aligned} DB^{[h+1]}(x_{h+1}) &\leq 4 + 4x_h DB^{[h]}(x_h) \\ &\leq 4 + (1 + 4/(ch^2))Ch \leq C(h+1). \end{aligned}$$

So (22) follows by induction. Now (23) follows from (21), (22) and

$$2 - B^{[h]}(1/4) = B^{[h]}(x_h) - B^{[h]}(1/4) \leq DB^{[h]}(x_h)(x_h - 1/4)\ .$$ ∎

We now complete the proof of Theorem 3. From Lemma 2

$$x_h \geq 1/4 + 1/(Ch^2)\ .$$

Hence

$$\begin{aligned} B_n^{[h]} &\leq B^{[h]}(x_h)x_h^{-n} \\ &= O\left(4^{-n}(1 + 4/(Ch^2))^{-n}\right) = O\left(B_n n^{3/2} e^{-\delta n/h^2}\right). \end{aligned}$$

4 Trees with large heights

In this section, we use the local limit theorem of Bender and Richmond [BR, Th. 2] to obtain an asymptotic formula for the number of trees with n internal nodes and heights $h = cn$ where $0 < c < 1$. Let δ and θ be some positive constants which can be arbitrarily small, and define

$$R = [\delta,\ 1/4 - \delta] \quad \text{and} \quad N(R) = \{z = re^{it} : r \in R, |t| \leq \theta\}\ .$$

It was shown in [FO] that

$$e_h(1/4) = O(1/h)\ .$$

For $|z| \le 1/4$, we have

$$2|B(z)e_h(z)| = |B(z) - B^{[h]}(z)| \le B(1/4) - B^{[h]}(1/4) ,$$

and hence

$$|e_h(z)| \le B(1/4)e_h(1/4)/|B(z)| .$$

Since $|B(z)| > 0$ for $|z| \le 1/4$,

$$e_h(z) = O(1/h) \text{ uniformly for } |z| \le 1/4 . \tag{25}$$

It is easy to check that

$$|1 - \epsilon(z)| < |1 - \epsilon(|z|)| < 1 \text{ for } z \ne |z|, |z| < 1/4 . \tag{26}$$

Using (25), (26) and [FO, (7)], we obtain

$$(1-\epsilon)^h/e_h \sim 1/\epsilon + 2 + \sum_{j\ge 0} \frac{e_j}{1-e_j}(1-\epsilon)^j \tag{27}$$

uniformly for $|z| \in N(R)$. Let

$$A(z) = \left(1/\epsilon + 2 + \sum_{j\ge 0} \frac{e_j}{1-e_j}(1-\epsilon)^j\right)^{-1} . \tag{28}$$

We know from [FO] that $A(z)$ is continuous in $N(R)$. It follows from (27) and (1) that

$$B(z) - B^{[h]}(z) \sim 2A(z)B(z)(1-\epsilon(z))^h$$

and hence

$$B^{[h]}(z) - B^{[h-1]}(z) \sim \frac{2\epsilon(z)}{1-\epsilon(z)} A(z)B(z)(1-\epsilon(z))^h \tag{29}$$

uniformly for $z \in N(R)$.

Now we compute

$$\mu = \frac{d\log(1-\epsilon)}{d\log z} = \frac{1+\epsilon}{2\epsilon} \tag{30}$$

and

$$\sigma^2 = \frac{d^2\log(1-\epsilon)}{(d\log z)^2} = \frac{1-\epsilon^2}{4\epsilon^3} . \tag{31}$$

It follows from (26), (29)–(31) and [BR, Theorem 2] that

$$B_n^{[h]} - B_n^{[h-1]} \sim 2A(z)B(z)\frac{\epsilon}{1-\epsilon}(1-\epsilon)^h z^{-n}/\sqrt{2\pi h\sigma} \tag{32}$$

uniformly for all h and n such that

$$\frac{n}{h} = \frac{1+\epsilon}{2\epsilon} \text{ for } \epsilon \in [\delta',\ 1-\delta'] , \tag{33}$$

where δ' is any small positive constant. Noting $B(z) = 2/(1+\epsilon)$ and $z = (1-\epsilon^2)/4$, we can simplify (32) as

$$\begin{aligned} & B_n^{[h]} - B_n^{[h-1]} \\ \sim\ & \frac{4\epsilon^2 A(\epsilon)}{(1-\epsilon)^2\sqrt{\pi(1+\epsilon)n}} \left((1-\epsilon)^{(1-\epsilon)}(1+\epsilon)^{(1+\epsilon)}\right)^{-h/2\epsilon} 4^n , \end{aligned}$$

which gives Theorem 4.

References

[BR] E. A. Bender and L. B. Richmond, Central and local limit theorems applied to asymptotic enumeration II: Multivariate generating functions, *J. Combin. Theory Ser. B* **34** (1983), 255–265.

[BS] G. G. Brown and B. O. Shubert, On random binary trees, *Math. Op. Res.* **9** (1984), 43–65.

[BD] N. De Bruijn, *Asymptotic Methods in Analysis*, North-Holland, Amsterdam, 1961.

[BKR] N. De Bruijn, D. Knuth and S. Rice, The average height of planted plane trees, pp. 15–22 in *Graph Theory and Computing*, R. C. Read, ed., Academic Press, New York, 1972.

[FO] P. Flajolet and A.M. Odlyzko, The average height of binary trees and other simple trees, *J. Comp. Sys. Sci.* **25** (1982), 171–213.

[K] V. F. Kolchin, *Random Mappings*, (English translation of 1984 Russian original), Optimization Software, 1986.

[L] T. Luczak, The number of trees with a large diameter, preprint.

[MM] A. Meir and J. W. Moon, On the altitude of nodes in random trees, *Canad. J. Math.* **30** (1978), 997–1015.

[O] A. Odlyzko, Periodic oscillation of coefficients of power series that satisfy functional equations, *Adv. in Math.* **44** (1982), 180–205.

[RS] A. Rényi and G. Szekeres, On the height of trees, *Aust. J. Math.* **7** (1967), 497–507.

[S] G. Szekeres, Regular iteration of real and complex functions, *Acta Math.* **100** (1958), 203–258.

[W] J. Wimp, Current trends in asymptotics: some problems and some solutions, *J. Comp. and Appl. Math.* **35** (1991), 53–79.

[WROM] R. A. Wright, B. Richmond, A. Odlyzko and B. D. Mckay, Constant time generation of free trees, *SIAM. J. Comput.* **15** (1986), 540–548.

ISSN 0249-6399