W. GUTJAHR

G. CH. PFLUG

The limiting common distribution of two leaf heights in a random binary tree

Informatique théorique et applications, tome 26, nº 1 (1992), p. 1-18.

<http://www.numdam.org/item?id=ITA_1992__26_1_1_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 26, n° 1, 1992, p. 1 à 18)

THE LIMITING COMMON DISTRIBUTION OF TWO LEAF HEIGHTS IN A RANDOM BINARY TREE (*)

by W. GUTJAHR ([1]) and G. CH. PFLUG ([1])

Communicated by J. E. PIN

Abstract. – *Let $\mathscr{B}_n$ be the family of extended binary trees with n internal nodes. Assume that each $t \in \mathscr{B}_n$ has equal probability. We compute the asymptotic common distribution of the heights of leaf i resp. j (where the leaves are enumerated from left to right) as i, j and n tend to infinity, such that i/n resp. j/n tends to x resp. y, $0<x<y<1$. As a corollary, the asymptotic covariance of the two heights is determined. Applications are indicated.*

Résumé. – *Soit $\mathscr{B}_n$ la famille des arbres binaires à n sommets. On suppose que chaque arbre $t \in \mathscr{B}_n$ est de même probabilité. Nous déterminons la distribution asymptotique commune des hauteurs des feuilles i et j (les feuilles étant énumérées de gauche à droite) quand i, j, $n \to \infty$, ainsi que $i/n \to x$, $j/n \to y\,(0<x<y<1)$. Comme corollaire, la covariance asymptotique des deux hauteurs est déterminée. Des applications sont indiquées.*

1. INTRODUCTION AND DEFINITIONS

Random binary trees occur as input models for the performance analysis of algorithms. Parameters of interest of particular algorithms often depend on the heights of one or several leaves of the input tree. (For this point of view, *see* e.g. [10, 2, 7, 8].) The asymptotic distribution of one single leaf height in a binary tree was recently identified by the authors as a Maxwell type distribution [5]. The purpose of this paper is to determine the two-dimensional asymptotic distribution of two leaf heights *i, j*, from which the calculation of covariances and conditional expectations is easily possible.

This common distribution turns out to be derivable from three independent Rayleigh-distributions, which are known as the speed distributions of thermo-

(*) Received March 1990, revised April 1991.
([1]) Universität Wien, Institut für Statistik und Informatik, Universitätstr. 5/9, A-1010 Wien, Austria.

0988-3754/92/01 1 18/$3.80/© AFCET-Gauthier-Villars

dynamic particles in a two-dimensional space, whereas the Maxwell distribution describes the speed distribution of such particles in the three-dimensional space. These analogies with statistical mechanics are somewhat surprising; an explanation for them has not yet been found.

Let $\mathscr{B}_n$ be the family of extended binary trees with n internal nodes $n+1$ leaves. We assume throughout this paper that all trees $t \in \mathscr{B}_n$ have equal probabilities.

Consider a special tree $t \in \mathscr{B}_n$. The leaves of t can be labelled, from left to right, with the numbers $0, \ldots, n$. Let $\pi_t(i)$ denote the (uniquely determined) path connecting the root with the leaf labelled with i, and let $h_t(i)$, called the *height* of i, denote the number of internal nodes on $\pi_t(i)$.

The number of trees $t \in \mathscr{B}_n$ with $h_t(i)=k$ will be denoted by $a(i, k, n)$, and

$$c_n := \operatorname{card} \mathscr{B}_n = \frac{1}{n+1} \begin{bmatrix} 2n \\ n \end{bmatrix}$$

denotes the Catalan numbers. Thus, $\mathbb{P}\{h_t(i)=k \mid t \in \mathscr{B}_n\} = a(i, k, n)/c_n$.

In this paper, the limiting behaviour of the probabilities

$$\mathbb{P}\{h_t(i)=k_1,\ h_t(j)=k_2 \mid t \in \mathscr{B}_n\}$$

as $n, i, j \to \infty$ and $i/n \to x$, $j/n \to y\,(0<x<y<1)$ will be investigated.

2. THE COMMON DISTRIBUTION OF TWO HEIGHTS

We start this section with some further definitions:

Let i, j be leaves of $t\,(i<j)$. Set $\pi_{1,t}(i, j) := \pi_t(i)$, $\pi_{3,t}(i, j) := \pi_t(j)$, and let $\pi_{2,t}(i,j)$ denote the (uniquely determined) path from leaf i to leaf j.

The number of internal nodes on the path $\pi_{m,t}(i, j)$ will be denoted by $h_{m,t}(i, j)$ $(m=1, 2, 3)$.

We make the following convention concerning the index m: Increments of m are always cyclic increments, *i.e.*

$$m+p \quad \text{means } (m+p-1) (\operatorname{mod} 3)+1 \qquad (p \geqq 0). \tag{2.1}$$

Obviously, for fixed i and j the following inequalities hold:

$$h_{m,t}(i, j) \leqq h_{m+1,t}(i,j) + h_{m+2,t}(i, j) \qquad (m=1, 2, 3). \tag{2.2}$$

We call them "triangle inequalities", since they require that the heights $h_{1,t}$, $h_{2,t}$, $h_{3,t}$ can be the lengths of the sides of a triangle.

Furthermore, consider the sub-paths $\hat{\pi}_{m,t}(i,j)$ of $\pi_{m,t}(i,j)$, containing all nodes which do not lie on any of the other two paths ($m=1, 2, 3$):

$$\hat{\pi}_{m,t}(i,j) := \pi_{m,t}(i,j) \setminus \pi_{m+1,t}(i,j) \qquad (m=1, 2, 3). \tag{2.3}$$

The number of internal nodes on $\hat{\pi}_{m,t}(i,j)$ will be denoted by $s_{m,t}(i,j)$.

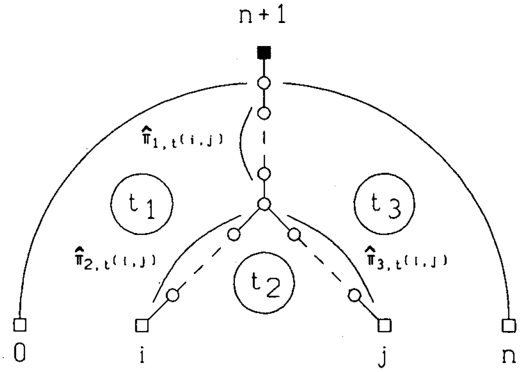
Theorem 1 establishes the asymptotic common distribution of the three numbers $s_{m,t}(i,j)$ ($m=1, 2, 3$). As a corollary, the asymptotic common distribution of the three heights $h_{m,t}(i,j)$ ($m=1, 2, 3$) will be derived; computing the marginal distribution for $m=1$ and 3 yields the asymptotic common distribution of $h_t(i)$ and $h_t(j)$, and computing the marginal distribution for $m=1$ yields the asymptotic distribution of $h_t(i)$.

For the proof of Theorem 1, it is convenient to use an alternative definition of an extended binary tree:

An extended binary tree of size n is a plane tree with n internal nodes, each of degree 3, and $n+2$ leaves, where one of the leaves is marked.

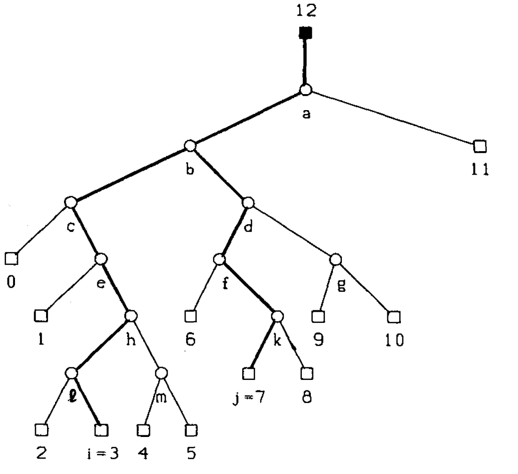
By removing the marked leaf and the incident edge, and marking the other node which was incident with this edge as the root of the remaining tree, we get an extended binary tree in the usual sense. It is clear that the correspondence is one-to-one.

The paths $\hat{\pi}_{m,t}(i,j)$ partition the given tree t into three subtrees t_1, t_2, t_3:

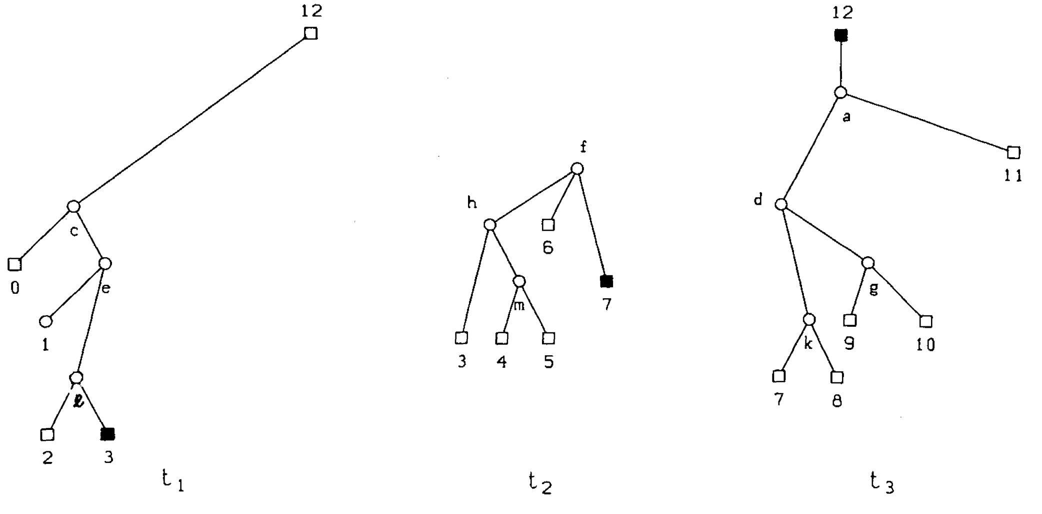


Each internal node v lying on one of the paths $\hat{\pi}_{m,t}(i,j)$ is counted to the subtree t_m, if there is an edge leading from v to the area of t_m ($m=1, 2, 3$); each of the leaves i, j, $n+1$ is counted to both adjacent subtrees. Then, by marking the leaves i, j, resp. $n+1$, the subtrees t_1, t_2, t_3 become extended binary trees in the sense of the above definition.

Example 1: The tree



is partitioned into the following subtrees:



From the number of leaves in each subtree it can be concluded that t_m ($m=1, 2, 3$) has i_m internal nodes with

$$i_1 := i, \qquad i_2 := j-i-1, \qquad i_3 = n-j. \qquad (2.4)$$

The proof of Theorem 1 relies on the knowledge of the height distribution of leaf 0 (the leftmost leaf):

Ruskey [12] shows

$$\left.\begin{aligned} a(i, k, n) &= \sum_{j=0}^{k} \begin{bmatrix} k \\ j \end{bmatrix} a(0, j, i)\, a(0, k-j, n-1) \\ \text{with} \qquad & \\ a(0, k, n) &= \frac{k}{2n-k} \binom{2\mathrm{n}-\mathrm{k}}{n-k}. \end{aligned}\right\} \tag{2.5}$$

(The last formula was generalized in [13] to the case of t-ary trees.)

In [5], the authors approximate the probability $a(0, k, n)/c_n$ that leaf 0 has height k asymptotically:

LEMMA 1:

$$\frac{a(0, k, n)}{c_n} = \frac{k}{2^{k+1}} \exp\left(-\frac{k^2}{4n} + R(k, n)\right).$$

To each $\alpha > 0$ *there are constants* N *and* M, *such that*

$$|R(k, n)| \leqq M k/n, \qquad \forall n \geqq N, \quad k \leqq \alpha \sqrt{n}. \quad \square$$

Now we are ready to prove the announced theorem.

THEOREM 1: *Let* $(s_1^{(n)})$, $(s_2^{(n)})$, $(s_3^{(n)})$, $(i^{(n)})$ *and* $(j^{(n)})$ *be sequences of positive integers with*

$$\left.\begin{aligned} & s_m^{(n)}/\sqrt{n} \to \sigma_m \qquad (m = 1, 2, 3), \\ & i^{(n)}/n \to x, \quad j^{(n)}/n \to y \qquad (n \to \infty) \\ & (\sigma_1, \sigma_2, \sigma_3 \geqq 0,\ 0 < x < y < 1). \end{aligned}\right\} \tag{2.6}$$

Then

$$n^{3/2}\, \mathbb{P}\left\{ s_{m,t}(i^{(n)}, j^{(n)}) = s_m^{(n)} (m = 1, 2, 3) \,\middle|\, t \in \mathscr{B}_n \right\}$$
$$= \frac{1}{256\pi} \prod_{m=1}^{3} x_m^{-3/2} (\sigma_m + \sigma_{m+1}) \exp\left[-\frac{(\sigma_m + \sigma_{m+1})^2}{16 x_m}\right] + o(1), \tag{2.7}$$

where $x_1 := x$, $x_2 := y - x$, $x_3 := 1 - y$, *and the increments of* m *are interpreted in a cyclic sense.*

Proof: Denote by $A(i, j, s_1, s_2, s_3, n)$ the number of trees $t \in \mathscr{B}_n$ *with* $s_{m,t}(i, j) = s_m (m=1, 2, 3)$, and classify these trees according to the numbers l_m of edges turning off from $\hat{\pi}_{m,t}(i,j)$ in the direction of the subtree $t_m (m=1, 2, 3)$.

There are $\begin{bmatrix} s_m \\ l_m \end{bmatrix}$ possibilities to select, from the s_m nodes, l_m nodes with an edge turning off in the indicated (counter-clockwise) direction. The remaining $s_m - l_m$ nodes have edges turning off in the direction of t_{m+2}, *i.e.* in clockwise direction.

That border of the tree t_m which is formed by the path $\hat{\pi}_{m,t}(i, j)$, contains $l_m + (s_{m+1} - l_{m+1})$ internal nodes. So, for fixed l_m, s_m $(m=1, 2, 3)$, i, j and n, there are $a(0, l_m + s_{m+1} - l_{m+1}, i_m)$ possibilities to choose a tree t_m, where i_m is given by (2.4).

Therefore,

$$A(i, j, s_1, s_2, s_3, n) = \sum_{l_1=0}^{s_1} \sum_{l_2=0}^{s_2} \sum_{l_3=0}^{s_3} \prod_{m=1}^{3} \begin{bmatrix} s_m \\ l_m \end{bmatrix} a(0, l_m + s_{m+1} - l_{m+1}, i_m), \quad (2.8)$$

and using Lemma 1, we can write this as

$$\frac{1}{8}\left[\prod_{k=1}^{3} c_{i_k}\right] \sum_{\substack{0 \leqq l_m \leqq s_m \\ m=1, 2, 3}} \prod_{m=1}^{3} \begin{bmatrix} s_m \\ l_m \end{bmatrix} \left[\frac{1}{2}\right]^{s_m} (l_m + s_{m+1} - l_{m+1}) \times \exp\left[-\frac{(l_m + s_{m+1} - l_{m+1})^2}{4\, i_m} + R(l_m + s_{m+1} - l_{m+1}, i_m)\right]. \quad (2.9)$$

Therein, $\prod_{m=1}^{3} \begin{bmatrix} s_m \\ l_m \end{bmatrix} [1/2]^{s_m}$ is the probability function of the common distribution of three independent, $B(s_m, 1/2)$-distributed random variables $L_m (m=1, 2, 3)$, where $B(n, p)$ denotes the Biomial distribution. The sum in (2.9) may be interpreted as the expected value of

$$\prod_{m=1}^{3} \begin{bmatrix} s_m \\ l_m \end{bmatrix} \left[\frac{1}{2}\right]^{s_m} (L_m + s_{m+1} - L_{m+1}) \times \exp\left[-\frac{(L_m + s_{m+1} - L_{m+1})^2}{4\, i_m} + \mathrm{R}(L_m + s_{m+1} - L_{m+1}, i_m)\right].$$

We normalize the variables L_m by defining $Z_m := L_m/s_m$, and consider the distributions μ_m of Z_m, which are linear transforms of $B(s_m, 1/2)$.

Taking account of the dependence on n, we provide the symbols s_m, i_m, j_m, etc. with upper index n and assume (2.6).

By the use of the probability meausres $\mu_1^{(n)}$, $\mu_2^{(n)}$, $\mu_3^{(n)}$, and by the substitutions

$$\sigma_m^{(n)} := s_m^{(n)}/\sqrt{n}, \qquad x_m^{(n)} := i_m^{(n)}/n$$
$$(\sigma_m^{(n)} \to \sigma_m,\ x_m^{(n)} \to x_m) \qquad (m=1, 2, 3),$$

the sum in (2.9) can be represented as

$$n^{3/2} \int_0^1 \int_0^1 \int_0^1 F_n(z_1, z_2, z_3)\, d\mu_1^{(n)}(z_1) \ldots d\mu_3^{(n)}(z_3) \tag{2.10}$$

with

$$F_n(z_1, z_2, z_3) := \prod_{m=1}^{3} (\sigma_m^{(n)} z_m + \sigma_{m+1}^{(n)}(1 - z_{m+1}))$$
$$\times \exp\left[-\frac{(\sigma_m^{(n)} z_m + \sigma_{m+1}^{(n)}(1-z_{m+1}))^2}{4 x_m^{(n)}} \right.$$
$$\left. + R(\sqrt{n}(\sigma_m^{(n)} z_m + \sigma_{m+1}^{(n)}(1-z_{m+1})), n x_m^{(n)}) \right]. \tag{2.11}$$

Because of $s_m^{(n)} \to \infty$ for $n \to \infty$ and the Law of Large Numbers, $Z_m \to 1/2$ in probability or

$$\mu_m^{(n)} \to \delta_{1/2} \quad \text{weakly} \qquad (m=1, 2, 3), \tag{2.12}$$

where δ_z denotes the point mass in z.

Obviously $\sigma_m^+ := \sup_{n \geqq 1} \sigma_m^{(n)} < \infty$ and $x_m^- := \inf_{n \geqq N} x_m^{(n)} > 0$ for sufficiently large N, and thus for $m = 1, 2, 3$

$$\sqrt{n}(\sigma_m^{(n)} z_m + \sigma_{m+1}^{(n)}(1 - z_{m+1})) \leqq \frac{\sigma_m^+ + \sigma_{m+1}^+}{\sqrt{x_m^-}} \sqrt{n x_m^{(n)}}.$$

Therefore, the remainder estimation of Lemma 1 can be applied to (2.11): The sum of the three remainder terms

$$R(\sqrt{n}(\sigma_m^{(n)} z_m + \sigma_{m+1}^{(n)}(1 - z_{m+1})), n x_m^{(n)}) \qquad (m=1, 2, 3)$$

in (2.11) is bounded by $M.n^{-1/2}$ with some constant M.

A rather tiresome, but quite straightforward estimation shows then that

$$F_n(z_1, z_2, z_3) - F(z_1, z_2, z_3) \to 0 \tag{2.13}$$

uniformly on $Q := [0, 1]^3$, where

$$F(z_1, z_2, z_3) := \prod_{m=1}^{3} (\sigma_m z_m + \sigma_{m+1}(1 - z_{m+1})) \times \exp\left[-\frac{(\sigma_m z_m + \sigma_{m+1}(1 - z_{m+1}))^2}{4 x_m}\right]. \tag{2.14}$$

Let $\mu^{(n)} := \mu_1^{(n)} \otimes \mu_2^{(n)} \otimes \mu_3^{(n)}$ denote the product measure of the measures $\mu_m^{(n)}$, and let $\delta_{(z_1, z_2, z_3)}$ denote the point mass in (z_1, z_2, z_3). Then by (2.13), (2.12) and the fact that F is continuous and bounded on Q,

$$\left|\int_Q \mathrm{F}_n \, d\mu^{(n)} - \int_Q \mathrm{F} \, d\delta_{(1/2,\, 1/2,\, 1/2)}\right| \leqq \int_Q |F_n - F| \, d\mu^{(n)} + \left|\int_Q \mathrm{F} \, d\mu^{(n)} - \int_Q \mathrm{F} \, d\delta_{(1/2,\, 1/2,\, 1/2)}\right| \to 0.$$

So (2.10) is asymptotically equivalent to

$$n^{3/2} F\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right). \tag{2.15}$$

Inserting (2.15) for the sum in (2.9), dividing the expression by c_n and applying Stirling's formula to the Catalan numbers yields the assertion. □

Remark 1: An inspection of the proof of Theorem 1 shows the following: If the sequences $\sigma_m^{(n)}$ in (2.6) depend on the parameter $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, and $s_m^{(n)}(\sigma)/\sqrt{n} \to \sigma_m$ uniformly in a cube $\sigma \in Q_c = [0, C]^3$ $(m = 1, 2, 3)$, then the convergence of (2.13) is uniform for $\sigma \in Q_c$, and as a consequence also the convergence of (2.7) is uniform for $\sigma \in Q_c$. In particular, this holds for $s_m^{(n)}(\sigma) := \lfloor \sigma_m \sqrt{n} \rfloor$. □

Corollary 1: *The limiting common distribution of the normalized heights* $h_{m,t}(i, j)/\sqrt{n}$ $(m = 1, 2, 3)$ *for* $i, j, n \to \infty$, $i/n \to x$, $j/n \to y$ $(0 < x < y < 1)$, *has* *the density function*

$$\gamma_{x,y}(w_1, w_2, w_3)=\begin{cases}\dfrac{1}{512\pi}\displaystyle\prod_{m=1}^{3} x_m^{-3/2}\, w_m \exp\left[-\dfrac{w_m^2}{16\,x_m}\right], \\ \text{if } w_m \leqq w_{m+1}+w_{m+2}\,(m=1,2,3), \\ 0, \quad \text{else}\end{cases} \tag{2.16}$$

(x_1, x_2, x_3 *as in Theorem* 1).

Proof: For a triple (w_1, w_2, w_3) not satisfying the triangles inequalities

$$w_1 \leqq w_{m+1}+w_{m+2} \qquad (m=1, 2, 3), \tag{2.17}$$

the asymptotic probability density is obviously zero. Assume now that (2.17) is fulfilled. It is easy to see that

$$\left.\begin{array}{c} s_{m,t}(i,j)=\dfrac{1}{2}[h_{m,t}(i,j)-h_{m+1,t}(i,j)+h_{m+2,t}(i,j)-1] \\ (m=1, 2, 3). \end{array}\right\} \tag{2.18}$$

So if for $n \to \infty$

$$h_{m,t}(i^{(n)}, j^{(n)})/\sqrt{n} \to w_m \qquad (m=1, 2, 3),$$

then the numbers

$$s_m^{(n)} := s_{m,t}(i^{(n)}, j^{(n)})$$

satisfy the condition $s_m^{(n)}/\sqrt{n} \to \sigma_m \geqq 0$ of theorem 1 with

$$\sigma_m := \frac{1}{2}(w_m - w_{m+1} + w_{m+2}) \qquad (m=1, 2, 3). \tag{2.19}$$

The equations (2.19) are equivalent to

$$w_m = \sigma_m + \sigma_{m+1} \qquad (m=1, 2, 3). \tag{2.20}$$

By Theorem 1 and Remark 1 we find that the distribution function

$$\mathbb{P}\{s_{m,t}(i^{(n)}, j^{(n)})/\sqrt{n} \leqq \sigma_m \qquad (m=1, 2, 3) \mid t \in \mathscr{B}_n\}$$

converges for every $\sigma=(\sigma_1, \sigma_2, \sigma_3)(\sigma_m \geqq 0)$, and the limiting distribution function yields the asymptotic probability density

$$\tilde{\gamma}_{x, y}(\sigma_1, \sigma_2, \sigma_3)=\frac{1}{256\pi}\prod_{m=1}^{3} x_m^{-3/2}(\sigma_m+\sigma_{m+1})\exp\left[-\frac{(\sigma_m+\sigma_{m+1})^2}{16 x_m}\right]. \quad (2.21)$$

A linear transformation of the density $\tilde{\gamma}_{x, y}$ with respect to (2.20) establishes the result. □

Remark 2: If S_m, resp. $H_m (m=1, 2, 3)$ are random variables distributed according to the limiting distribution of $s_{m, t}(i,j)/\sqrt{n}$, resp. $h_{m, t}(i,j)/\sqrt{n}$ $(m=1, 2, 3)$, then from (2.21), resp. (2.16), the moments

$$\begin{gathered}\mathbb{E}(S_1^{k_1} S_2^{k_2} S_3^{k_3}), \qquad \text{resp. } \mathbb{E}(H_1^{k_1} H_2^{k_2} H_3^{k_3}) \\ (k_1, k_2, k_3 \geqq 0)\end{gathered} \quad (2.22)$$

can be computed. In general, the limit of moments of distributions can be larger than the corresponding moment of the limiting distributions. However, it turns out that the random variables

$$\prod_{m=1}^{3}\left[\frac{s_{m, t}(i,j)}{\sqrt{n}}\right]^{k_m}, \qquad \text{resp. } \prod_{m=1}^{3}\left[\frac{h_{m, t}(i,j)}{\sqrt{n}}\right]^{k_m} \quad (2.23)$$

are uniformly integrable with respect to n. (The proof must be omitted here for the sake of brevity). As a consequence, the expected values of (2.23) converge to corresponding expected values (2.22). □

Remark 3: The probability density

$$g_{x_m}(w)=\text{const}.\, w \exp\left(-\frac{w^2}{16 x_m}\right)$$

appearing as a factor of $\gamma_{x, y}(w_1, w_2, w_3)$ in (2.16) is the density of a Rayleigh distribution (special case $c=2$ of a Weibull distribution, *see* [6], p. 251) with parameter $\alpha=4\sqrt{x_m}$. Thus we can – to each given x and y – simulate the normalized heights w_1, w_2, w_3 in the following way:

(1) Draw values w_1, w_2, w_3 independently from each other, where w_m is Rayleigh-distributed with parameter $\alpha=4\sqrt{x_m}$ $(m=1, 2, 3)$.

(2) If the drawn values satisfy the triangle inequalities (2.17), accept them; otherwise, reject them and go to (1).

We observe the noticeable fact that the random variables w_1, w_2, w_3 depend from each other only by the condition that they must satify the triangle inequalities. □

COROLLARY 2: *The limiting common distribution of the normalized heights* $h_t(i)/\sqrt{n}$ *and* $h_t(j)/\sqrt{n}$ *for* $i, j, n \to \infty$, $i/n \to x$, $j/n \to y$ $(0<x<y<1)$ *has the density function*

$$\tau_{x,y}(u, v)=\frac{1}{64\pi}[x(1-y)]^{-3/2}(y-x)^{-1/2}uv$$

$$\times \exp\left(-\frac{u^2}{16x}\right).\exp\left(-\frac{v^2}{16(1-y)}\right)$$

$$\times\left[\exp\left(-\frac{(u-v)^2}{16(y-x)}\right)-\exp\left(-\frac{(u+v)^2}{16(y-x)}\right)\right]. \qquad (2.24)$$

Proof: The triangle inequalities (2.17) are equivalent to

$$|w_1-w_3| \leqq w_2 \leqq w_1+w_3.$$

So the density of the marginal distribution of (w_1, w_3) is given by

$$\int_{|w_1-w_3|}^{w_1+w_3} \gamma_{x,y}(w_1, w_2, w_3)\,dw_2.$$

A short calculation and setting $u:=w_1$, $v:=w_3$ yields the density $\tau_{x,y}(u, v)$ above. □

COROLLARY 3: *The limiting distribution of the normalized height* $h_t(i)/\sqrt{n}$ *for* $i, n \to \infty$, $i/n \to x$ $(0<x<1)$, *has the density function*

$$\rho_x(u)=\frac{1}{16\sqrt{\pi}}[x(1-x)]^{-3/2}u^2\exp\left(-\frac{u^2}{16x(1-x)}\right). \qquad (2.25)$$

For $k \geqq 0$,

$$\mathbb{E}\left[\left[\frac{h_t(i)}{\sqrt{n}}\right]^k\right] \to \frac{2}{\sqrt{\pi}}[16x(1-x)]^{k/2}\,\Gamma\left[\frac{k+3}{2}\right]. \qquad (2.26)$$

In particular,

$$\mathbb{E}\left[\frac{h_t(i)}{\sqrt{n}}\right] \to \frac{8}{\sqrt{\pi}}\sqrt{x(1-x)}, \tag{2.27}$$

$$\operatorname{Var}\left[\frac{h_t(i)}{\sqrt{n}}\right] \to (24-64/\pi)\,x(1-x). \tag{2.28}$$

Proof: (2.25) is obtained from (2.24) by computing the marginal distribution

$$\int_0^\infty \tau_{x,y}(u, v)\,dv.$$

Computing the k-th moment about the origin of the distribution (2.25) yields the right hand side of (2.26). Because of Remark 2, this moment of the limiting distribution is equal to the limit of the moments on the left hand side of (2.26). $\square$

The density (2.25) is that of a Maxwell distribution, *i.e.* the distribution of $\sqrt{X_1^2+X_2^2+X_3^2}$, where X_i are independently normally distributed (*see* [1], pp. 32, 48). In [5], the authors derive the assertions of Corollary 3 by the consideration of only one leaf i. Result (2.27) was already found by Kirschenhofer in [9].

3. THE COVARIANCE OF TWO HEIGHTS

For abbreviation, we set $U^{(n)} := h_t(i)/\sqrt{n}$, $V^{(n)} := h_t(j)/\sqrt{n}$.

Then by Remark 2, for $i, j, n \to \infty$, $i/n \to x$, $j/n \to y\,(0<x<y<1)$,

$$E(U^{(n)} V^{(n)}) \to \int_0^\infty \int_0^\infty uv\,\tau_{x,y}(u, v)\,du\,dv, \tag{3.1}$$

where $\tau_{x,y}$ is given by (2.24). With a medium amount of computation effort, the above double integral can be solved:

$$E(U^{(n)} V^{(n)}) \to \frac{16}{\pi}\left[(3x_1 x_3 + x_2)\arctan\left(\sqrt{\frac{x_1 x_3}{x_2}}\right) + 3\sqrt{x_1 x_2 x_3}\right], \tag{3.2}$$

where, as in Theorem 1, $x_1 = x$, $x_2 = y-x$ and $x_3 = 1-y$.

Combining (3.2) with (2.27) yields:

COROLLARY 4:

$$\operatorname{Cov}(U^{(n)} V^{(n)}) \to \frac{16}{\pi}\left[(3 x_1 x_3 + x_2)\arctan\left(\sqrt{\frac{x_1 x_3}{x_2}}\right) + 3\sqrt{x_1 x_2 x_3} - 4\sqrt{x_1(1-x_1)}\sqrt{x_3(1-x_3)}\right]. \quad \square \tag{3.3}$$

If $x_2 \to 0$, (3.3) leads again to (2.28).

(3.3) also enables us to compute the asymptotic variance of the random variable $s_{1,t}(i, j)$ defined in Section 2:

By (2.18),

$$s_{1,t}(i,j) = \frac{1}{2}(h_t(i) - h_{2,t}(i,j) + h_t(j) - 1), \tag{3.4}$$

$$\mathbb{E}(2 s_{1,t}(i,j) + 1) = \mathbb{E}(h_t(i)) - \mathbb{E}(h_{2,t}(i,j)) + \mathbb{E}(h_t(j)), \tag{3.5}$$

and

$$\begin{aligned}\operatorname{Var}(2 s_{1,t}(i,j) + 1) &= \operatorname{Var}(h_t(i)) + \operatorname{Var}(h_{2,t}(i,j)) + \operatorname{Var}(h_t(j)) \\ &+ 2[\operatorname{Cov}(h_t(i), h_t(j)) - \operatorname{Cov}(h_t(i), h_{2,t}(i,j)) - \operatorname{Cov}(h_t(j), h_{2,t}(i,j))].\end{aligned} \tag{3.6}$$

By a $(i+1)$-step cyclic re-numeration of the leaves of $t \in \mathscr{B}_n$, we find

$$\mathbb{E}(h_{2,t}(i,j)) = \mathbb{E}(h_t(j-i-1)) \tag{3.7}$$

and

$$\operatorname{Var}(h_{2,t}(i,j)) = \operatorname{Var}(h_t(j-i-1)). \tag{3.8}$$

(For details *see* [3], Proposition 2.1.)

Using similar re-numerations of the leaves, one obtains

$$\operatorname{Cov}(h_t(i), h_{2,t}(i,j)) = \operatorname{Cov}(h_t(i), h_t(n+i-j+1)), \tag{3.9}$$

$$\operatorname{Cov}(h_t(j), h_{2,t}(i,j)) = \operatorname{Cov}(h_t(j), h_t(j-i-1)). \tag{3.10}$$

Now assume again $i^{(n)}/n \to x$, $j^{(n)}/n \to y$.

Then with (3.5), (3.7), (2.27) and the abreviations $x_1 := x$, $x_2 := y - x$, $x_3 := 1 - y$, $\alpha_m := \sqrt{x_m (1 - x_m)}$ $(m = 1, 2, 3)$, one finds

$$\mathbb{E}\left[\frac{s_{1,t}(i^{(n)}, j^{(n)})}{\sqrt{n}}\right] \to \frac{4}{\sqrt{\pi}}(\alpha_1 - \alpha_2 + \alpha_3). \tag{3.11}$$

With (3.6), (3.8)-(3.10), (2.28), (3.3) and the additional abreviation $\beta := \sqrt{x_1 x_2 x_2}$, we compute

$$\begin{aligned}\operatorname{Var}\left[\frac{s_{1,t}(i^{(n)}, j^{(n)})}{\sqrt{n}}\right] \to \left(6 - \frac{16}{\pi}\right) \sum_{m=1}^{3} \alpha_m^2 \\ + \frac{8}{\pi} \sum_{m=1}^{3} (-1)^m [(x_m + 3 x_{m+1} x_{m+2}) \\ \times \arctan\left(\frac{\beta}{x_m}\right) - 4 \alpha_{m+1} \alpha_{m+2}\right] - \frac{24 \beta}{\pi}. \end{aligned} \tag{3.12}$$

Note that $s_{1,t}(i, j)$ can be interpreted as the height of the root of the (uniquely determined) binary subtree $t_{i,j}$ of t with minimal height, containing all leaves $i, i+1, \ldots, j$.

4. CONDITIONAL PROBABILITIES

Consider two random variables U, V distributed according to the limiting common distribution given by Corollary 2. Then the density of the conditional distribution of V, given $U = u$ fixed, can immediately be computed from (2.24) and (2.25):

$$\begin{aligned}\tau_{x,y}(v \mid u) &= \tau_{x,y}(u, v)/\rho_x(u) \\ &= \frac{1}{4\sqrt{\pi}} [(1-x)/(1-y)]^{3/2} (y-x)^{-1/2} u^{-1} v \\ &\quad \times \exp\left(\frac{u^2}{16(1-x)}\right) \exp\left(-\frac{v^2}{16(1-y)}\right) \\ &\quad \times \left[\exp\left(-\frac{(u-v)^2}{16(y-x)}\right) - \exp\left(-\frac{(u+v)^2}{16(y-x)}\right)\right]. \end{aligned} \tag{4.1}$$

Asymptotically, the subtree t_r on the right hand side of the path $\pi_t(x\sqrt{n})$ contains $n/(1-x)$ internal nodes, and the leaf $x\sqrt{n}$, resp. $y\sqrt{n}$, of t is labelled with the number 0, resp. $(y-x)/(1-x)$ in t_r. So we may expect that

$$\tau_{x,y}(v \mid u) = \tau_{0,(y-x)/(1-x)}(v(1-x)^{-1/2} \mid u(1-x)^{-1/2}).(1-x)^{-1/2}, \quad (4.2)$$

and this can indeed be verified from (4.1).

Surprisingly enough, it turns out that the conditional distribution (4.1) already suffices for the computation of the common asymptotic distribution of *more* than two heights:

Consider k leaves $i_1, \ldots, i_k$ and assume $i_1, \ldots, i_k, n \to \infty$ with

$$i_1/n \to x_1, \ldots, i_k/n \to x_k, \qquad 0 < x_1 < \ldots < x_k < 1.$$

Then, as it will be shown in a forthcoming paper, the normalized heights $(h_t(i_1, n)/\sqrt{n}, \ldots, h_t(i_k, n)/\sqrt{n})$ converge in probability to limiting random variables $(U_1, \ldots, U_k)$ with the following Markov property:

$$\left.\begin{aligned} &\mathbb{P}\{U_k \leqq u_k \mid U_{k-1} = u_{k-1}, \ldots, U_1 = u_1\} \\ &\qquad = \mathbb{P}\{U_k \leqq u_k \mid U_{k-1} = u_{k-1}\} = \int_0^{u_k} \tau_{x,y}(v \mid u_{k-1})\, dv \\ &\qquad\qquad (u_1, \ldots, u_k \geqq 0). \end{aligned}\right\} \quad (4.3)$$

Thus (4.1) determines the common distribution of an arbitrary number of heights.

Finally, we remark without proof that the Markovian process given by (4.3) and (4.1) is identical with the process $\sqrt{B_1(x)^2 + B_2(x)^2 + B_3(x)^2}$, where B_1, B_2, B_3 are independent Brownian bridges on $[0, 1]$. This process is also known under the name of Brownian excursion.

So we can say that the asymptotic contour of a random binary tree is described by the distance from the starting point of a Brownian particle in $\mathbb{R}^3$, on the condition that it returns to the starting point.

5. APPLICATIONS

Since binary trees belong to the most frequently used data structures, there are a number of possible applications of the results of Section 2-3 in Computer Science. We mention two of them:

(1) A straightforward application is the analysis of stack oscillations during *level order traversal* of a binary tree t, which is given by the following

recursive procedure: Visit the root – traverse the left subtree – visit the root – traverse the right subtree – visit the root (*see* [8], pp. 82 f).

The size of the needed stack corresponds to the recursion depth of the procedure.

Then the so-called MAX-turns (*see* [7]) $m_t(0)$, $m_t(1)$, ..., *i.e.* the local maxima of the function describing the stack size, are identical to the heights $h_t(0)$, $h_t(1)$, ... of the leaves of t. We can state the result that level order traversal of a binary tree leads to Maxwell-distributed MAX-turns of the stack, and the common distribution, resp. the covariance of MAX-turn i and j are given by the formulae (2.24), resp. (3.3).

In the case of *postorder* traversal of a binary tree, which occurs during the evaluation of an arithmetical or boolean expression given in infix-form, the MAX-turns of the stack are the leaf heights of the corresponding *ordered tree* ([7], p. 159); so distribution results for the leaf heights in random ordered trees would be desirable.

(2) Another application concerns the theory of software reliability, espacially input domain based models (*cf.* [11]) which try to describe the correctness correlation of a program in different input points. Let P be a program with binary control flow structure $t=t_P\in\mathscr{B}_n$, assume that to each internal node v of t there corresponds a probability p of correct execution, and that the correctness in v is independent from the correctness in the other nodes, and let the random variables C_i take the value 1, if the $(i+1)$-th path of P is executed correctly, 0 else $(0\leqq i\leqq n)$. In [4], the covariance of C_i and C_j is studied on the "black box" assumption that t is unknown and all $t\in\mathscr{B}_n$ are equally likely.

Clearly,

$$\operatorname{Cov}(C_i, C_j)=\mathbb{E}(C_i C_j)-\mathbb{E}(C_i).\mathbb{E}(C_j)$$
$$=\mathbb{P}\{C_i=1, C_j=1\}-\mathbb{P}\{C_i=1\}.\mathbb{P}\{C_j=1\}.$$

For fixed $t\in\mathscr{B}_n$, the conditional correctness variables $C_{i,t}$ and $C_{j,t}$ are both 1, iff all nodes on $\pi_t(i)\cup\pi_t(j)$ are "correct". Thus

$$\mathbb{P}\{C_i=1, C_j=1\}=\frac{1}{c_n}\sum_{t\in\mathscr{B}_n}\mathbb{P}\{C_{i,t}=1, C_{j,t}=1\}=\frac{1}{c_n}\sum_{t\in\mathscr{B}_n}p^{u_t(i,j)}=\mathbb{E}(p^{u_t(i,j)}),$$

where $u_t(i, j)$ denotes the number of internal nodes on $\pi_t(i)\cup\pi_t(j)$.

Analogously,

$$\mathbb{P}\{C_i=1\}=\mathbb{E}(p^{h_t(i)}),$$

and therefore

$$\operatorname{Cov}(C_i, C_j) = \mathbb{E}[p^{u_t(i,j)}] - \mathbb{E}[p^{h_t(i)}]\,\mathbb{E}[p^{h_t(j)}]. \tag{5.1}$$

In the case of $p = 1 - \varepsilon$ with a small failure probability ε, we obtain from (5.1):

$$\operatorname{Cov}(C_i, C_j) \approx \varepsilon\,\mathbb{E}(h_t(i) + h_t(j) - u_t(i,j)) = \varepsilon\,\mathbb{E}(s_{1,t}(i,j) + 1) \sim \varepsilon\,\mathbb{E}(s_{1,t}(i,j)),$$

and the last value may be received from (3.11).

If, on the other side, ε is not small, the knowledge of the expectation of $s_{1,t}(i,j)$ does not suffice for the computation of (5.1). However, we have

$$u_t(i,j) = \frac{1}{2}\sum_{m=1}^{3} h_{m,t}(i,j) - \frac{1}{2} \sim \frac{1}{2}\sum_{m=1}^{3} h_{m,t}(i,j),$$

and since the asymptotic common distribution of the heights $h_{m,t}(i,j)$ is given by Corollary 1, the covariance (5.1) can be computed for an arbitrary correctness probability p. With this modification, the model may be extended to early phases of the software life cycle, where the failure probability cannot be expected to be small.

REFERENCES

1. W. Feller, An Introduction to Probility Theory and Its Applications, 1971, *II*, *Wiley*.
2. P. Flajolet and A. Odlyzko, Exploring Binary Trees and Other Simple Trees, Proc. 21st Annual Symposion on Foundations of Computer Science, 1980, pp. 207-216.
3. W. Gutjahr, On the co-structure of k paths in a random binary tree, *Technical Report* TR-ISI/Stamcom-71, Inst. of Statistics and Computer Science, Univ. of Vienna, 1989. Accepted for publication in *J. Comb., Math. Comb. Comput.*
4. W. Gutjahr, A Combinatorial Model for Software Testing and Reliability, *Technical Report* TR-ISI/Stamcom-74, Inst. of Statistics and Computer Science, Univ. of Vienna, 1989.
5. W. Gutjahr and G. Pflug, The Asymptotic Distribution of Leaf Heights in Binary Trees, *Technical Report* TR-ISI/Stamcom-76, Inst. of Statistics and Computer Science, Univ. of Vienna, 1990. Accepted for publication in *Graphs and Combinatorics*.
6. N. L. Johnson and S. Kotz, Continuous Univariate Distributions-1, Houghton Mifflin, 1970.
7. R. Kemp, On the Average Oscillation of a Stack, *Combinatorica*, 1982, *2*, (2), pp. 157-176.

8. R. Kemp, Fundamentals of the Average Case Analysis of Particular Algorithms, *Wiley-Teubner*, 1984.

9. P. Kirschenhofer, On the Height of Leaves in Binary Trees, *J. Combin. Inform. System Sci.*, 1983, *8* (1), pp. 44-60.

10. D. E. Knuth, The Art of Computer Programming, 1972, *1*, *Addison-Wesley*.

11. G. V. Ramamoorthy and F. B. Bastani, Software Reliability-Status and Perspectives, *I.E.E.E. Trans. Software Engrg.*, 1982, *SE-8*, pp. 354-371.

12. F. Ruskey, On the average shape of binary trees, *S.I.A.M. J. Alg. Disc. Meth.*, 1980, *1* (1), pp. 43-50.

13. J. W. Moon, On Level Numbers of t-ary Trees, *S.I.A.M. J. Alg. Disc. Meth.*, 1983, *4*, pp. 8-13.