On the Co-Structure of k Paths In a Random Binary Tree

W. Gutjahr

Institut für Statistik und Informatik
Universität Wien
A1010 Wien, Universitätsstrasse 5/9
AUSTRIA

Abstract. Consider the paths $\pi_t(i_1), \ldots, \pi_t(i_k)$ from the root to the leaves $i_1, \ldots, i_k$ in a random binary tree t with n internal nodes, where all such trees are assumed equally likely and the leaves are enumerated from left to right. We investigate, for fixed $i_1, \ldots, i_k$ and n, the average size of $\pi_t(i_1) \cup \cdots \cup \pi_t(i_k)$ resp. of $\pi_t(i_1) \cap \cdots \cap \pi_t(i_k)$ (the latter corresponding to the average depth of the smallest subtree containing $i_1, \ldots, i_k$). By a rotation argument, both problems are reduced to the case $k = 1$, for which a solution is known. Furthermore, formulas for the probability distributions of the depth of leaf i, the distance between leaf i and j and the length of $\pi_t(i) \cap \pi_t(j)$ are derived.

1. Introduction and definitions

Since A. MEIR's and J.W. MOON's work on the average number of nodes at a fixed level in a binary tree ([7]), several other results on the shape of a random binary tree of size n have been found: P. FLAJOLET and A. ODLYZKO established the average height of the whole tree ([1]); F. RUSKEY ([8]) and P. KIRSCHENHOFER ([4]) investigated the average depth of the leaf with number i, where the leaves are enumerated from left to right; H. PRODINGER ([6]) determined the average value of the so called register pathlength of the binary tree; etc.

The problem examined in this paper is a generalization of RUSKEY's and KIRSCHENHOFER's, considering k leaves instead of only one. This generalization has some relevance for Computer Science: The case of successive leaves is crucial for the investigation of stack oscillations (cf. [3]) and can possibly be useful for the complexity analysis of parsers; the case of arbitrary leaves allows of an analytical treatment of a software reliability model for the so called linearly domained programs (see [2]).

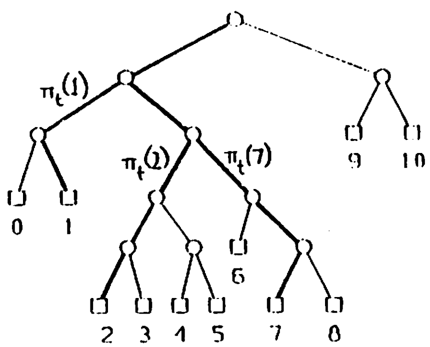
Let $\mathcal{B}_n$ be the family of extended binary trees with n internal nodes and $n + 1$ leaves, and let $t \in \mathcal{B}_n$. The leaves of t can be enumerated from left to right with the numbers $0, \ldots, n$. In the sequel, each leaf will be identified with its number in t.

If $i (0 \leq i \leq n)$ is a leaf of t, then let $\pi_t(i)$ denote the path from the root to i. The length of this path, i.e. the number of its internal nodes, shall be denoted by $h_t(i)$; this is simply the depth of leaf i in t.

Further, if $i_1, \ldots, i_k$ are leaves of $t \in \mathcal{B}_n (0 \le i_1, \ldots, i_k \le n; 1 \le k \le n+1)$, then the union $\pi_t(i_1) \cup \cdots \cup \pi_t(i_k)$ and the intersection $\pi_t(i_1) \cap \cdots \cap \pi_t(i_k)$—defined in an obvious way—are (not necessarily binary) subtrees of t, the intersection being a path again. We consider the numbers

$$u_t(i_1, \ldots, i_k) = \text{number of internal nodes of } t, \text{ contained in } \pi_t(i_1) \cup \cdots \cup \pi_t(i_k),$$
$$s_t(i_1, \ldots, i_k) = \text{number of internal nodes of } t, \text{ contained in } \pi_t(i_1) \cap \cdots \cap \pi_t(i_k).$$

Example 1.1: Let t be the following tree $\in \mathcal{B}_{10}$:



Then (setting $i_1 = 1, i_2 = 2, i_3 = 7$)

$$h_t(1) = 3, \quad h_t(2) = 5, \quad h_t(7) = 5,$$
$$u_t(1,2,7) = 8, \quad s_t(1,2,7) = 2.$$

∎

Remark 1: In the case $k = 1$,

$$u_t(i) = s_t(i) = h_t(i) \qquad (0 \le i \le n). \tag{1.1}$$

Remark 2: If $i_1 < \cdots < i_k$, i.e. each leaf i_κ lies on the left side of leaf $i_{\kappa+1} (\kappa = 1, \ldots, k-1)$, it is evident that $\pi_t(i_1) \cap \cdots \cap \pi_t(i_k) = \pi_t(i_1) \cap \pi_t(i_k)$, and so

$$s_t(i_1, \ldots, i_k) = s_t(i_1, i_k). \tag{1.2}$$

Obviously, $s_t(i_1, \ldots, i_k) - 1$ is the depth of the root of the smallest binary subtree containing the leaves $i_1, \ldots, i_k$. Following the terminology in [5], this root can be called the "i_1-th $(i_k - i_1 + 1)$-turn" of t.

The aim of the present paper is to determine the average values of the numbers $u_t(i_1, \ldots, i_k)$ resp. $s_t(i_1, \ldots, i_k)$, where $i_1, \ldots, i_k$ are fixed, the binary tree t is taken randomly from $\mathcal{B}_n$, and all binary trees in $\mathcal{B}_n$ are assumed to be equally likely. This leads to the following definitions:

For $0 \leq i_1, \ldots, i_k \leq n$, $1 \leq k \leq n+1$, let

$$h(i_1; n) = \frac{1}{c_n} \sum_{t \in \mathcal{B}_n} h_t(i_1), \tag{1.3}$$

$$u(i_1, \ldots, i_k; n) = \frac{1}{c_n} \sum_{t \in \mathcal{B}_n} u_t(i_1, \ldots, i_k), \tag{1.4}$$

$$s(i_1, \ldots, i_k; n) = \frac{1}{c_n} \sum_{t \in \mathcal{B}_n} s_t(i_1, \ldots, i_k). \tag{1.5}$$

Therein,

$$c_n = \operatorname{card} \mathcal{B}_n = \frac{1}{n+1} \binom{2n}{n} \tag{1.6}$$

denotes the n-th Catalan number.

It will be shown that for $i_1 < \cdots < i_k$,

$$u(i_1, \ldots, i_k; n) = \frac{1}{2} \left[\sum_{\kappa=0}^{k} h(i_{\kappa+1} - i_\kappa - 1; n) - k + 1 \right]$$
$$(i_0 = -1, i_{k+1} = n+1),$$
$$s(i_1, \ldots, i_k; n) = \frac{1}{2} [h(i_1; n) + h(i_k; n) - h(i_k - i_1 - 1; n) + 1] .$$

2. Cases $k=1$ and $k=2$, and the average distance between leaf i and leaf j

Let us start with the case $k = 1$. Because of (1.1), in this case the solution of our problem is given by KIRSCHENHOFER's formula ([4]) on the average depth of leaf i:

$$u(i; n) = s(i; n) = h(i; n) = 4(n+1)(2n+1)(n+2)^{-1} \binom{n}{i}^2 \binom{2n+2}{2i+1}^{-1} - 1. \tag{2.1}$$

For n, i, $n - i \to \infty$, KIRSCHENHOFER found

$$h(i; n) = 8 \left(\frac{i}{\pi}\right)^{1/2} \left(1 - \frac{i}{n}\right)^{1/2} - 1 + 0 \left(\max(i^{-1/2}, (n-i)^{-1/2})\right) . \tag{2.2}$$

At the end of Section 3, a possible derivation of (2.1) will be indicated.

Assume now $k = 2$. We define the distance $\rho_t(i, j)$ between two different leaves i, j in t as the number of internal nodes on the unique path $\overline{ij}$ connecting i

with j in t. If $i = j$, we set $\rho_t(i,j) = 0$. (Note that this definition of distance is slightly different from the usual one counting the number of edges between i and j; in our notation, the latter number is $\rho_t(i,j) + 1$ for $i \neq j$.)

The average distance between i and j is then defined by

$$\rho(i,j;n) = \frac{1}{c_n} \sum_{t \in \mathcal{B}_n} \rho_t(i,j). \tag{2.3}$$

It should be mentioned that both $\rho_t(\cdot,\cdot)$ and $\rho(\cdot,\cdot;n)$ fulfil the properties of a metric on $\{0,\dots,n\}$. These metrics can even be extended to metrics on $\mathbf{Z}_{n+2}$, the residual ring modulo $(n+2)$.

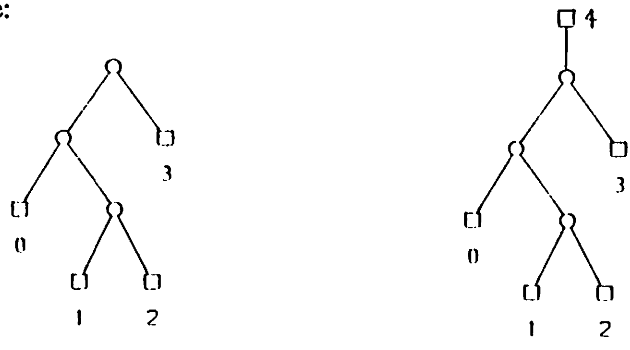
For $i \neq j$, obviously

$$u_t(i,j) = \frac{1}{2}(h_t(i) + h_t(j) + \rho_t(i,j) - 1), \tag{2.4}$$

$$s_t(i,j) = \frac{1}{2}(h_t(i) + h_t(j) - \rho_t(i,j) + 1), \tag{2.5}$$

and analogous formulas hold for the average values $u(i,j;n)$ resp. $s(i,j;n)$. So still $\rho(i,j;n)$ needs to be determined.

For this purpose, we use the well known respresentation of a binary tree as a "planted tree". Let $\tilde{\mathcal{B}}_n$ be the family of all plane trees $\tilde{t}$ with n internal nodes, each of degree 3, and $n+2$ leaves, enumerated in counter-clockwise direction with $0,\dots,n+1$. Then from each tree $t \in \mathcal{B}_n$, a corresponding tree $\tilde{t} \in \tilde{\mathcal{B}}_n$ can be constructed by adding an edge and a leaf to the root in upward direction and assigning to the leaf the number $n+1$. Conversely, if $\tilde{t} \in \tilde{\mathcal{B}}_n$, remove the leaf with number $n+1$ and the incident edge, and mark the other node that was incident with the removed edge as the root of the remaining tree; this yields again t. For example:



t $\tilde{t}$

Proposition 2.1. *Let*

$$\alpha(i,j,d;n) = card\{t \in \mathcal{B}_n | \rho_t(i,j) = d\},$$
$$\beta(i,d;n) = card\{t \in \mathcal{B}_n \mid h_t(i) = d\}$$
$$(0 \leq i,j \leq n; 0 \leq d \leq n; n \geq 0) \ .$$

Then for $d \geq 1$

$$\alpha(i,j,d;n) = \begin{cases} \beta(|i-j|-1,d;n), & i \neq j, \\ 0, & i = j. \end{cases}$$

Proof: Because of the above correspondence between $t \in \mathcal{B}_n$ and $\bar{t} \in \bar{\mathcal{B}}_n$, it suffices to show that

$$\bar{\alpha}(i,j,d;n) = \begin{cases} \bar{\beta}(|i-j|-1,d;n), & i \neq j, \\ 0, & i = j, \end{cases}$$

where

$$\bar{\alpha}(i,j,d;n) = \text{card}\ \{\bar{t} \in \bar{\mathcal{B}}_n \mid \rho_{\bar{t}}(i,j) = d\},$$
$$\bar{\beta}(i,d;n) = \text{card}\ \{\bar{t} \in \bar{\mathcal{B}}_n \mid h_{\bar{t}}(i) = d\}.$$

$\rho_{\bar{t}}(i,j)$ is defined analogously as $\rho_t(i,j)$, and $h_{\bar{t}}(i) = \rho_{\bar{t}}(n+1,i)$.

Consider the function $\phi_r : \bar{\mathcal{B}}_n \to \bar{\mathcal{B}}_n (r \in \mathbf{Z})$ which effects an $|r|$step cyclic renumeration of the leaves of a given tree in the direction indicated by the sign of r; i.e. leaf i in t gets the number $i - r \pmod{(n+2)}$ in $\phi_r(\bar{t})$. Since ϕ_r^{-1} exists (it is equal to ϕ_{-r}), ϕ_r is a permutation of the elements of $\bar{\mathcal{B}}_n$.

Now let i,j with $0 \leq i \leq j \leq n$ be fixed.. Then ϕ_{i+1} effects a renumeration of the leaves of t in such a way, that the path connecting leaf i with leaf j in the original tree t corresponds to the path connecting leaf $n+1$ with leaf $j-i-1$ in the renumerated tree $\phi_{i+1}(t)$.

In particular,

$$\rho_{\bar{t}}(i,j) = p_{\phi_{i+1}(\bar{t})}(n+1, j-i-1) = h_{\phi_{i+1}(\bar{t})}(j-i-1).$$

So we get

$$\bar{\alpha}(i,j,d;n) = \text{card}\ \{\bar{t} \in \bar{\mathcal{B}}_n \mid h_{\phi_{i+1}(\bar{t})}(j-i-1) = d\}$$
$$= \text{card}\ \{\bar{t} \in \bar{\mathcal{B}}_n \mid h_{\bar{t}}(j-i-1) = d\} = \bar{\beta}(|i-j|-1,d;n).$$

If, conversely, $j < i$, the assertion follows from the symmetry of $\bar{\alpha}$ in i and j. The case $i = j$ is trivial. ∎

Corollary. *For* $0 \le i, j \le n$,

$$\rho(i,j;n) = \begin{cases} h(|i-j|-1;n), & i \ne j, \\ 0, & i = j. \end{cases}$$

Proof: Let $i \ne j$. Then

$$\rho(i,j;n) = \frac{1}{c_n}\sum_{d\ge1} d\alpha(i,j,d;n) = \frac{1}{c_n}\sum_{d\ge1} d\beta(|i-j|-1,d;n) = h(|i-j|-1;n).$$

∎

With the aid of the last Corollary and (2.4) resp. (2.5), the solution for the case $k = 2$ can now be stated:

Proposition 2.2.

$$\begin{aligned} u(i,j;n) &= \frac{1}{2}[h(i;n) + h(j;n) + h(|i-j|-1;n) - 1], \\ s(i,j;n) &= \frac{1}{2}[h(i;n) + h(j;n) - h(|i-j|-1;n) + 1] \\ &\qquad (0 \le i,j \le n; i \ne j) \end{aligned}$$

where $h(i;n)$ *is given by (2.1).*

Proposition 2.3. *For* $i \to \infty, j \to \infty, n \to \infty, \frac{i}{n} \to x, \frac{j}{n} \to y (0 < x, y < 1, x \ne y)$, *the following asymptotic relations hold:*

$$\begin{aligned} h(i;n) &= \sqrt{n}\,\overline{h}(x) - 1 + 0(n^{-1/2}), \\ u(i,j;n) &= \sqrt{n}\overline{u}(x,y) - 1 + 0(n^{-1/2}), \\ s(i,j;n) &= \sqrt{n}\overline{s}(x,y) - 1 + 0(n^{-1/2}), \end{aligned}$$

with

$$\begin{aligned} \overline{h}(x) &= 8\pi^{-1/2}\sqrt{x(1-x)}, \\ \overline{u}(x,y) &= \frac{1}{2}[\overline{h}(x) + \overline{h}(y) + \overline{h}(|x-y|)], \\ \overline{s}(x,y) &= \frac{1}{2}[\overline{h}(x) + \overline{h}(y) - \overline{h}(|x-y|)]. \end{aligned} \tag{2.6}$$

Proof: Use of (2.2) and of Proposition 2.2. ∎

3. The probability distribution of the distance between leaf i and leaf j

It should be noted that Proposition 2.1 not only makes it possible to establish the average distance between the two leaves i and j, but beyond that yields the whole probability distribution of the distances $\rho_t(i,j)\,(t \in \mathcal{B}_n)$:

$$P\{\rho_t(i,j) = d \mid t \in \mathcal{B}_n\} = \frac{1}{c_n}\alpha(i,j,d;n) = \frac{1}{c_n}\beta(|i-j|-1,d;n) \quad (i \neq j)$$

with c_n as is (1.6). So it seems worthwhile to compute the numbers $\beta(i,d;n)$.

Proposition 3.1.

a) *The generating function of the numbers $\beta(i,d;n)$ is given by*

$$G(z,v,u) = \sum_{n\geq 0}\sum_{d\leq n}\sum_{i\leq n}\beta(i,d;n)z^n v^d u^i = [1 - zv(uy(zu)+y(z))]^{-1}, \tag{3.1}$$

where

$$y(z) = \sum_{n\geq 0} c_n z^n = \frac{1-\sqrt{1-4z}}{2z} \tag{3.2}$$

is the generating function of the Catalan numbers.

b) *The numbers $\beta(i,d;n)$ $(i \leq n, d \leq n, n \geq 0)$ satisfy the following recursions:*

$$\beta(i,d;n) = \sum_{\substack{0\leq k\leq i\\ 0\leq d-k\leq n-i}} \binom{d}{k}\beta(0,k;i)\beta(0,d-k;n-i) \tag{3.3}$$

$$\beta(0,d;n) = \begin{cases} \sum_{j=0}^{n-d} c_j\beta(0,d-1;n-1-j), & d\geq 1,\\ \delta_{n0}, & d=0. \end{cases} \tag{3.4}$$

Proof: Let $i = 0$, and d be fixed. Then $\beta(i,d;n)$ is the number of binary trees with n internal nodes, whose first leaf from the left has depth d.

If we remove the path $\pi_t(0)$, we get a forest of d binary trees with $n-d$ internal nodes in total. The generating function of the numbers of such forests is given by $z^d y(z)^d$, so

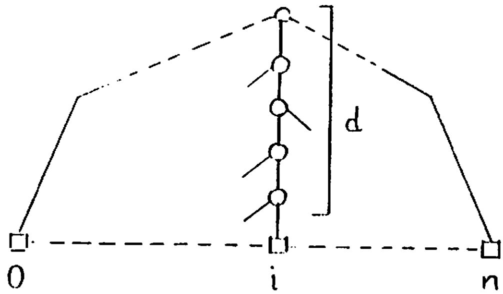
$$\sum_{n\geq d}\beta(o,d;n)z^n = (zy(z))^d. \tag{3.5}$$

Expansion of the right side yields

$$\beta(0,d;n) = \sum_{k_1+\cdots+k_d=n-d} c_{k_1}\ldots c_{k_d}, \tag{3.6}$$

and from this (3.4) can be derived.

Now let $i \geq 0$. The path $\pi_t(i)$ with d internal nodes divides the tree t into two parts:



Let

d_1 = number of the internal nodes v on $\pi_t(i)$, where the right successor of v belongs to $\pi_t(i)$ ("nodes of first kind"),

d_2 = number of the internal nodes v on $\pi_t(i)$, where the left successor of v belongs to $\pi_t(i)$ ("nodes of second kind").

Then $d_1 + d_2 = d$, and there are (for fixed d_1) exactly $\binom{d}{d_1}$ possibilities to select d_1 nodes of first kind from the d nodes of $\pi_t(i)$.

If the nodes of first resp. second kind are counted to the subtree t_1 resp. t_2 on the left resp. on the right side of $\pi_t(i)$, then t_1 and t_2 are complete binary trees with i resp. with $n-i$ internal nodes (leaves $0,\ldots,i$ resp. $i,\ldots,n$; the leaf i belongs to both t_1 and t_2).

In t_1, $\pi_t(i)$ is the path connecting the root with the rightmost leaf, so there are $\beta(i, d_1; i) = \beta(0, d_1; i)$ possibilities to choose t_1.

In t_2, $\pi_t(i)$ is the path connecting the root with the leftmost leaf, so there are $\beta(0, d_2; n-i)$ possibilities to choose t_2.

In total, we have as many possibilities for constructing a tree t with $h_t(i) = d$ as indicated in (3.3) ($k = d_1, d - k = d_2$).

It remains to prove that (3.1) holds. For fixed $d \geq 0$,

$$\begin{aligned}
&\sum_{n\geq d}\sum_{i\leq n}\beta(i,d;n)z^n u^i \\
&= \sum_{n\geq d}\sum_{i_1+i_2=n}\sum_{\substack{d_1+d_2=d\\ d_1\leq i_1\\ d_2\leq i_2}}\binom{d}{d_1}\beta(0,d_1;i_1)z^{i_1}\beta(0,d_2;i_2)z^{i_2}u^{i_1} \\
&= \sum_{d_1+d_2=d}\binom{d}{d_1}\left[\sum_{i_1\geq d_1}\beta(0,d_1;i_1)(zu)^{i_1}\right]\left[\sum_{i_2\geq d_2}\beta(0,d_2;i_2)z^{i_2}\right] \\
&= \sum_{d_1+d_2=d}\binom{d}{d_1}(zuy(zu))^{d_1}(zy(z))^{d_2} = [zuy(zu) + zy(z)]^d.
\end{aligned}$$

From that, (3.1) follows by multiplication by v^d and summation over $d \geq 0$. ∎

Formula (3.3) was already given by Ruskey in [8]; our derivation slightly simplifies his proof.

Remark: The partial derivative $\frac{\partial}{\partial v} G(z, v, u) \mid_{v=1}$ of (3.1) yields the generating function $H(z, u) = \sum_{n \geq 0} \sum_{i \leq n} (\sum_{t \in \mathcal{B}_n} h_t(i)) z^n u^i$ of the sums of depths of leaf i in trees $t \in \mathcal{B}_n$. A short computation and the use of $y(z) = 1 + z(y(z))^2$ leads again to Kirschenhofer's formula in [4],

$$H(z,u) = \left[\frac{y(z) - uy(zu)}{1-u}\right]^2 - \frac{y(z) - uy(zu)}{1-u}, \qquad (3.7)$$

from which his result (2.1) is obtained by expansion of $H(z, u)$.

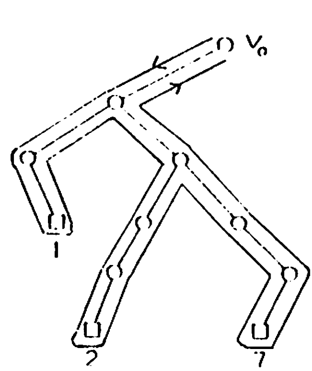
4. The general case

Now the restriction to one or two paths shall be removed; we consider the k paths to the leaves $i_1, \ldots, i_k$ and assume $i_1 < \cdots < i_k$ without loss of generality.

Proposition 4.1. *Let* $t \in \mathcal{B}_n$, *and* $0 \leq i_1 < \cdots < i_k \leq n$. *Then*

$$u_t(i_1, \ldots, i_k) = \frac{1}{2}[h_t(i_1) + \rho_t(i_1, i_2) + \rho_t(i_2, i_3) + \cdots + \rho_t(i_{k-1}, i_k) + h_t(i_k) - k + 1].$$

Proof: Surround the subtree $\pi_t(i_1) \cup \cdots \cup \pi_t(i_k)$ in counter-clockwise direction, beginning and ending with the root v_0, as it is shown in the following illustration for the case of Example 1.1:



This closed walk consists of the paths $\overline{v_0 i_1}, \overline{i_1 i_2}, \overline{i_2 i_3}, \ldots, \overline{i_{k-1} i_k}, \overline{i_k v_0}$.

Their respective lengths (numbers of internal nodes) are $h_t(i_1), \rho_t(i_1, i_2), \rho_t(i_2, i_3), \ldots, \rho_t(i_{k-1}, i_k), h_t(i_k)$. Each internal node of $\pi_t(i_1) \cup \cdots \cup \pi_t(i_k)$ is contained in exactly two of the above paths, with the exception of the nodes

$v_\kappa(\kappa = 1, \ldots, k-1)$, where v_κ is the deepest node of $\pi_\kappa(i_k) \cap \pi_t(i_{\kappa+1})$; these $k-1$ nodes are contained in exactly three of the above paths.

So

$$2\,u_t(i_1, \ldots, i_k) + (k-1) = h_t(i_1) + \rho_t(i_1, i_2) + \cdots + \rho_t(i_{k-1}, i_k) + h_t(i_k).$$

∎

Now we can state the general result:

Proposition 4.2. *For* $0 \le i_1 < \cdots < i_k \le n \quad (1 \le k \le n+1, n \ge 0)$

$$u(i_1, \ldots, i_k; n) = \frac{1}{2}\left[\sum_{\kappa=0}^{k} h(i_{\kappa+1} - i_\kappa - 1; n) - k + 1\right], \tag{4.1}$$

$$s(i_1, \ldots, i_k; n) = \frac{1}{2}[h(i_1; n) + h(i_k; n) - h(i_k - i_1 - 1; n) + 1], \tag{4.2}$$

where $i_0 = -1, i_{k+1} = n+1$, *and* $h(i; n)$ *is given by (2.1).*

Proof: (4.1) is an immediate consequence of Proposition 4.1 and the Corollary to Proposition 2.1, using additionally the symmetry $h(i_k; n) = h(n - i_k; n)$.

(4.2) follows from (1.2) and Proposition 2.2. ∎

Again, the asymptotic behaviour can be derived:

Proposition 4.3. *For k fixed,* $n \to \infty, i_1 \to \infty, \ldots, i_k \to \infty, \frac{i_1}{n} \to x_1, \ldots, \frac{i_k}{n} \to x_k$, *and* $0 < x_1 < \cdots < x_k < 1$, *the following asymptotic approximations hold:*

$$u(i_1, \ldots, i_k; n) \sim \sqrt{n}\,\overline{u}(x_1, \ldots, x_k),$$
$$s(i_1, \ldots, i_k; n) \sim \sqrt{n}\,\overline{s}(x_1, \ldots, x_k),$$

with

$$\overline{u}(x_1, \ldots, x_k) = \frac{1}{2}\sum_{\kappa=0}^{k} \overline{h}(x_{\kappa+1} - x_\kappa), \tag{4.3}$$

$$\overline{s}(x_1, \ldots, x_k) = \frac{1}{2}[\overline{h}(x_1) + \overline{h}(x_k) - \overline{h}(x_k - x_1)], \tag{4.4}$$

where $x_0 = 0, x_{k+1} = 1$, *and* $\overline{h}(x)$ *is given by (2.6).* ∎

From (4.3), it can easily be verified that for fixed k and n (n large), $u(i_1, \ldots, i_k; n)$ takes its maximum in the case of equidistant leaves.

5. The probability distribution of $s_t(i,j)$

It may be of interest to know not only the average value $s(i,j;n)$ of the numbers $s_t(i,j)$ (given by Proposition 2.2), but also their distribution, i.e. the probabilities

$$P\{s_t(i,j)=s \mid t\in \mathcal{B}_n\}=\frac{1}{c_n}\text{card }\{t\in\mathcal{B}_n \mid s_t(i,j)=s\} \quad (0\le i\le j\le n, 1\le s\le n)$$

(with c_n as in (1.6)). Consider the numbers $\gamma(i,j,s;n) = \text{card}\{t \in \mathcal{B}_n \mid s_t(i,j) = s\}$.

For each $t\in\mathcal{B}_n$ and fixed $i,j(0\le i<j\le n)$, let $p(t)$ be the number of the leftmost leaf of the smallest binary subtree t' of t containing the leaves i and j, and let $m(t)$ be the number of internal nodes in the subtree t'' obtained from t by contracting t' to a single leaf.

Then by classification of all trees $t\in\mathcal{B}_n$ with $s_t(i,j) = s$ with regard to $p=p(t)$ and $m=m(t)$, it can be seen that

$$\gamma(i,j,s;n)=\sum_{p=0}^{i}\sum_{m=p}^{n-j+p}\delta(i-p,j-p;n-m)\beta(p,s-1;m). \tag{5.1}$$

Therein,

$$\delta(i,j;n)=\sum_{k=i}^{j-1}c_kc_{n-1-k} \tag{5.2}$$

is the number of trees $t\in\mathcal{B}_n$ where leaf i lies in the left principal subtree and leaf j lies in the right principal subtree of t, and $\beta(i,d;n)$ is defined as in Proposition 2.1. Thus, $\gamma(i,j,s;n)$ can be computed numerically by means of (5.1), (5.2) and Proposition 3.1.

At least in the case $s = 1$, the asymptotic behaviour of $\gamma(i,j,s;n)$ for $i\to\infty, j\to\infty, n\to\infty, \frac{i}{n}\to x, \frac{j}{n}\to y (0<x<y<1)$ can be specified. Clearly,

$$\gamma(i,j,1;n)=\delta(i,j;n).$$

By Stirling approximation,

$$c_k=\pi^{-1/2}4^k k^{-3/2}\left(1+0\left(\frac{1}{k}\right)\right).$$

Inserted in (5.2), this yields

$$\delta(i,j;n)=\frac{4^{n-1}}{\pi}\sum_{k=i}^{j-1}[k(n-1-k)]^{-3/2}\left(1+0(\frac{1}{n})\right). \tag{5.3}$$

With $g_n(w) = [w(n-1-w)]^{-3/2}$, we have

$$\sum_{k=i}^{j-1} g_n(k) = \int_i^j g_n(w)\,dw + 0(|g_n(j) - g_n(i)|), \tag{5.4}$$

considering the fact that g_n is symmetric around $\frac{n-1}{2}$, decreasing for $0 < w < \frac{n-1}{2}$ and increasing for $\frac{n-1}{2} < w < n-1$.

The integral in (5.4) can be solved:

$$\begin{aligned}\int_i^j g_n(w)\,dw &= \frac{1}{(n-1)^2}\int_{i/(n-1)}^{j/(n-1)} [u(1-u)]^{-3/2}\,du \\ &= \frac{2}{(n-1)^2}\left\{\varphi\left(\frac{j}{n-1}\right) - \varphi\left(\frac{i}{n-1}\right)\right\}\end{aligned} \tag{5.5}$$

with

$$\varphi(u) = (2u-1)[u(1-u)]^{-1/2}, \tag{5.6}$$

and the expression $\{\ldots\}$ in (5.5) tends to the constant $\varphi(y) - \varphi(x) > 0$ for $n \to \infty$, so the integral is of order n^{-2}.

The error term $|g_n(j) - g_n(i)|$ in (5.4) is equal to $|g_n'(\xi_n)|(j-i)$ for some $\xi_n \in [i, j]$; with $c = \min(x, 1-y)$,

$$\frac{c}{2}n < \xi_n < 1 - \frac{c}{2}n$$

for sufficiently large n, hence

$$|g_n'(\xi_n)| \leq |g_n'(\frac{1}{2}cn)| = 0(n^{-4}) \text{and}$$

$$|g_n(j) - g_n(i)| = 0(n^{-3}).$$

Therefore,

$$\delta(i, j; n) = \frac{4^n}{2\pi n^2}\left(\varphi\left(\frac{j}{n}\right) - \varphi\left(\frac{i}{n}\right)\right)\left(1 + 0\left(\frac{1}{n}\right)\right). \tag{5.7}$$

As a consequence, the probability $\delta(nx, ny; n)/c_n$ that the leaves nx and ny lie in different principal subtrees tends to zero like $n^{-1/2}$ as $n \to \infty$.

6. Conclusion

The intention of this paper is a methodological one in so far as it was pointed out that diverse problems involving path lengths in random binary trees can be solved by two means:

a) the rotation principle of the proof of Proposition 2.1,
b) the knowledge of the generating function (3.1) of the path length distribution.

Since the used rotation argument can be generalized to arbitrary simply generated families of trees (including t-ary trees and ordered trees), the same approach could turn out to be helpful in this more general context. This will possibly open a more direct access to the problem of average hyperoscillations of trees (cf. [3] and [5]) and to similar combinatorial problems arising in Computer Science.

Acknowledgment

The author is indebted to G. Ch. Pflug for various stimulating discussions on the subject of this paper, to P. Kirschenhofer for some useful communications, and to I. Bomze for reading the typescript.

References

1. P. Flajolet and A. Odlyzko, *The average height of binary trees and other simple trees*, INRIA Rapports de Recherche **56** (1981), 171–213. also J. Comput. System Sci. **25** (1982), 171–213.
2. W. Gutjahr, *A Binary Tree Model for Software Reliability*, University of Vienna. preprint
3. R. Kemp, *On the average oscillation of a stack*, Combinatorica **2** (2) (1982), 157–176.
4. P. Kirschenhofer, *On the Height of Leaves in Binary Trees*, J. of Comb. Inf. & Syst. Sci. **8**(1) (1983), 44–60.
5. P. Kirschenhofer and H. Prodinger, *On the average hyperoscillations of planted plane trees*, Combinatorica 2 (2) (1982), 177–186.
6. H. Prodinger, *Some recent results on the register function of a binary tree*, M. Karonski, Z. Palka (ed.) (1987). Random Graphs 1985, North-Holland
7. A. Meir and J. W. Moon, *On the altidude of nodes in random trees*, Canad. J. Math. **30** (1978), 997–1015.
8. F. Ruskey, *On the average shape of binary trees*, SIAM J. Algebraic Discrete Methods **1** (1980), 43–50.