(Un)expected Path Lengths of Asymmetric Binary Search Trees

Uwe Trier

Johann Wolfgang Goethe Universität
Fachbereich Informatik
D-6000 Frankfurt/Main
Germany

Abstract. The average path length over the family of binary search trees with $2r+1$ nodes, built from randomly choosen permutations of r distinct keys, is $O(r \log(r))$. On the other hand, if the probability that the r keys are already sorted, tends towards 1, the resulting family of binary search trees degenerates, until it consists of the single linear list - tree with a worst case path length of $r(r+1)$.

Subject of the paper is the gap between the orders of growth of the expected path lengths of these two models. We systematically deform the probability model of binary search trees using three different kinds of deformers which are controlled by a real constant c and, using singularity analysis, we obtain the surprising result, that under two deformations the expected path length of a tree of size $2r+1$ is either $O(r \log(r))$ or $O(r^2)$. The same result is valid under the third deformation for special values of c.

Keyword: Algorithms and Data Structures

1. Introduction and Basic Definitions

Let $T = (I(T), L(T), r(T))$ be an unlabelled rooted planar tree with the set of internal nodes $I(T)$, the nonempty set of leaves $L(T)$ and the root $r(T)$. The one node tree is denoted by "$\square$". For any two nodes $u, v \in I(T) \cup L(T)$, let $d(u, v)$ be the *distance* from u to v, which is defined as the length of the shortest path from u to v ($=$ number of nodes on the path minus 1). A node $w \in I(T) \cup L(T)$ with $d(r(T), w) = l$ is said to be at *level* l. The tree T has *height* h, if the maximum level of a node in the tree is equal to h. The *total path length* $pl(T)$ of the tree T is defined as $pl(T) := \sum_{v \in I(T) \cup L(T)} level(v)$, where $level(v)$ denotes the level of node v.

A t-ary tree, $t \geq 2$, of size $tr + 1$ is an unlabelled rooted planar tree with $r \in I\!N_0 := I\!N \cup \{0\}$ internal nodes and $(t - 1)r + 1$ leaves, in which each internal node is of degree t. The family of t-ary trees, $t \geq 2$, is denoted by $\mathcal{F}_t$, its subfamily of trees of size $tr + 1$ is denoted by $\mathcal{F}_t(tr + 1)$, $t \geq 2, r \in I\!N_0$. A λ-tree, $\lambda \in I\!N_0$, is a tree with root degree λ.

The traditional model of binary search trees (BST) of size $2r + 1$, $r \in I\!N_0$, is characterized by the family $\mathcal{F}_2(2r + 1)$, in which the probability $prob(T)$ of a tree $T \in \mathcal{F}_2(2r + 1)$ is recursively defined as follows:

$$prob(T) := \text{if } T = \square \quad \text{then} \quad 1$$
$$\text{else} \quad \tfrac{1}{r} \, prob(T_1) \, prob(T_2),$$

where T_1 and T_2 are the two subtrees of T of sizes $2r_i + 1$, $i \in \{1, 2\}$, respectively, with $2(r_1 + r_2 + 1) = 2r$ (see [8]). Each of the r internal nodes holds one of the r

distinct keys of a randomly choosen permutation of r distinct keys. The $r + 1$ leaves represent the $r + 1$ NIL-pointers of the r internal nodes.

Under the BST-model, balanced trees are more likely than long skinny trees, which is the reason for the "good" average path length over the family of binary search trees of size $2r + 1$, which is $O(r \log(r))$. This is better than under the uniform model, where all trees of size $2r + 1$ are assumed to be equally likely, and where the average path length is $O(r^{\frac{3}{2}})$ (see [8]).

Our aim is to systematically deform the BST - model and investigate the influence of this deformation upon the average total path length under the deformed model. This deformation is controlled by an external nonnegative real number c, which can range from 0 to ∞. In order to explain how the deformation is done, let us first summarize the basics of a concept introduced by Kemp [6] in 1987, which can be viewed as an alternative way to describe the BST-model and even more sophisticated families of trees with a nonregular probability distribution:

Definition 1. *Let $\mathcal{F}_t(n) \subset \mathcal{F}_t$ be the family of t-ary trees of size n. A tree $T \in \mathcal{F}_t(n)$ with subtrees $T_i, 1 \le i \le t$, is said to be of type $< (n,t); (m_1, d_1), \ldots, (m_t, d_t) >$, if subtree T_i is a d_i-tree, $d_i \in \{0, t\}$, with m_i nodes, $1 \le i \le t$. Let $\mathcal{F}_{n,t}(\vec{m}, \vec{d})$ be the family of trees $T \in \mathcal{F}_t$ of type $< (n,t); (m_1, d_1), \ldots, (m_t, d_t) >$. Let $p_t(n, \vec{m}, \vec{d})$ be the probability that a tree $T \in \mathcal{F}_t(n)$ is of type $< (n,t); (m_1, d_1), \ldots, (m_t, d_t) >$, and let us assume that $p_t(n, \vec{m}, \vec{d})$ has a representation of the form*

$$p_t(n, \vec{m}, \vec{d}) = |q(n,t)|^{-1} \prod_{1 \le j \le t} q_j^{(t)}(m_j, d_j), \tag{1}$$

where $q_j^{(t)}(m, 0) = \delta_{m,1}$. The generating functions $Q_{j,l}^{(t)}(z) := \sum_{m \ge l+1} q_j^{(t)}(m, l) z^m$, $l \in \{0, t\}$, $1 \le j \le t$, are called characteristics of $\mathcal{F}_t$. Note that always $Q_{j,0}^{(t)}(z) = z$.

Some basic properties of the characteristics shown in [6] are summarized in the following

Proposition 1. *Let $Q_t(z) := \sum_{n \ge t+1} q(n,t) z^n$ be the generating function of the numbers $q(n,t)$ of Definition 1. Then*

$$Q_t(z) = z \prod_{1 \le j \le t} \sum_{l \in \{0,t\}} Q_{j,l}^{(t)}(z). \tag{2}$$

Each tree $T \in \mathcal{F}_t(n)$ belongs to exactly one class $C(T) := \mathcal{F}_{n,t}(\vec{m}, \vec{d})$, with $\vec{d} \in \{0, t\}^t$, and $\sum_{1 \le i \le t} m_i = n - 1$, and the probability $p_t(T)$ of a tree $T \in \mathcal{F}_t(n)$ is given by $p_t(T) = \prod_{T' \in SUB(T)} p(C(T'))$, where $p(C(T'))$ is the probability of the class $C(T')$, and $SUB(T)$ denotes the set of all subtrees of T (including T itself). $\square$

In addition to the definitions above we need the concept of additive weights (see [7]).

Definition 2. *Let $a \in [0, 1]$ be a real number, and let $g : \mathbb{N}_0^2 \to \mathbb{R}$ be a given*

mapping, the so-called weight function. *For a given tree $T \in \mathcal{F}_t$, the weight $w_{a,g}(T)$ is recursively defined by*

$$w_{a,g}(T) := \text{if } T = \square \quad \text{then} \quad 0$$
$$\text{else} \quad \sum_{1 \leq i \leq t} [a \ w_{a,g}(T_i) + g(m_i, d_i)],$$

where subtree T_i of T is a d_i-tree, $d_i \in \{0, t\}$, with m_i nodes, $1 \leq i \leq t$.

Choosing appropriate weight functions g and real numbers a, the resulting weights $w_{a,g}(T)$ of a tree $T \in \mathcal{F}_t$ correspond to well known parameters like

- the total path length (choose $g(m, d) := m$, $a := 1$),
- the total degree path length (choose $g(m, d) := tm$, $a := 1$).

The following relation holds (see [6]):

Proposition 2. *Let $t \in I\!N$, $t \geq 2$, $\mathcal{F}_t$ be the family of trees with the characteristics $Q_{j,l}(z)$, $l \in \{0, t\}$, $1 \leq j \leq t$, let $a \in I\!R$, and g be a given weight function. Furthermore, let $E(z) := \sum_{n \geq t+1} I\!E[w_{a,g}(n)]z^n$ be the generating function of the average weight $I\!E[w_{a,g}(n)]$ over the family $\mathcal{F}_t(n)$. Then*

$$E(z) \odot Q_t(z) = \sum_{1 \leq j \leq t} \frac{Q_t(z)}{z + Q_{j,t}(z)} [g(1, 0)z + \{a \ E(z) + G_t(z)\} \odot Q_{j,t}(z)], \quad (3)$$

where $G_t(z) = \sum_{n \geq t+1} g(n, t)z^n$, and $h_1(z) \odot h_2(z)$ denotes the Hadamard product of the two series $h_i(z) = \sum_{n \geq 0} h_{in}z^n$, $i \in \{1, 2\}$, which is defined by $h_1(z) \odot h_2(z) := \sum_{n \geq 0} h_{1n}h_{2n}z^n$. $\square$

Choosing appropriate characteristics, it is possible to describe well known probability models such as (see [6]):

- Binary search trees (choose $Q_{1,2}(z) = Q_{2,2}(z) := \frac{z^3}{1-z^2}$),
- t-ary digital search trees (choose $Q_{1,t}(z) = \ldots = Q_{t,t}(z) := z \exp(z^t) - z$),
- Patricia trees (choose $Q_{1,2}(z) = Q_{2,2}(z) := \frac{1}{2}(\exp(z^2) - 1) - z$),
- Regularly distributed t-ary trees (choose $Q_{1,2}(z) = \ldots = Q_{t,t}(z) := T(z) - z$, where $T(z)$ denotes the enumerator of the family $\mathcal{F}_t$).

For instance, in [2], R. Casas, J. Díaz and C. Martínez presented a probability model for simple families of trees, under which balanced trees are more likely than long skinny trees. Under this *Balanced Probability Model (BPM)*, the average behaviour of two parameters, a so-called *occupancy* and the size of the intersection of two trees, are analyzed and compared with the uniform probability model. In order to get a better feeling about the concepts introduced above, let us summarize from [2] the basic definitions of the *BPM* in the following

Example 1. Let $\mathcal{F}$ be an arbitrary family of simply generated trees [10]. For any tree $T \in \mathcal{F}$, in which the degree of the root $r(T)$ is equal to $k \in I\!N_0$, the "frequency" $f(T)$ is recursively defined as follows:

$$f(T) := \text{if } T = \square \quad \text{then} \quad 1$$
$$\text{else} \quad \frac{1}{|T|} f(T_1) \cdots f(T_k),$$

where $|T|$ denotes the number of internal nodes and $T_1, \ldots, T_k$ denote the k subtrees of T.

In order to obtain a probability distribution over the subfamily of $\mathcal{F}$ of trees with the same number of internal nodes as T has, the probability $p(T)$ of T is defined by $p(T) := \frac{f(T)}{\sum_{\tau \in \mathcal{F}, |\tau| = |T|} f(\tau)}$. Furthermore, let $F^{(t)}(z) = \sum_{T \in \mathcal{F}_t} f(T) z^{|T|} = \sum_{r \geq 0} f_r^{(t)} z^r$, $t \geq 2$, be the "frequency" - generating function of the trees $T \in \mathcal{F}_t$. Then, the frequency - function is given by $F^{(t)}(z) = (1 - (t-1)z)^{-\frac{1}{t-1}}$. For $t = 2$, the BPM coincides with the BST-model. $\qquad\Box$

A full average case analysis of additive weights of unlabelled rooted planar t-ary trees with polynomial weight functions under the BPM is presented in [11], using the characteristics $Q_{1,t}(z) = \ldots = Q_{t,t}(z) = zF^{(t)}(z^t) - z$.

In the sequel let $[z^r]T(z)$ denote the r-th coefficient of the series of $T(z)$ around $z = 0$. Symbols like $\mathcal{L}$ or $\mathcal{F}$ are always used to denote families of trees with some probability distribution.

2. Deformation of the BST - Model

Let L_r be the list - tree with r internal nodes, in which each internal node (with the exception of the root) is the right son of its father. Furthermore, let $\mathcal{L}_2(2r+1)$ be the family of binary trees of size $2r+1$, $r \in I\!N_0$, in which the linear list - tree L_r appears with probability 1 and all other trees appear with probability 0. This "$\mathcal{L}$ - model" is induced by the classical binary tree insertion algorithm, if the probability that r distinct keys are already sorted, is equal to 1. The average total path length $pl(2r+1)$ over the family $\mathcal{L}_2(2r+1)$ is equal to $pl(L_r) = r(r+1)$. This rather trivial result can also be obtained by the methods presented above:

Proposition 3. *Let $Q_{1,0}(z) = Q_{2,0}(z) := z$, $Q_{1,2}(z) \equiv 0$ and $Q_{2,2}(z) := \frac{z^3}{1-z^2}$ be given characteristics. Then the probability model induced by these characteristics coincides with the $\mathcal{L}$-model, and, under this model, the expected total path length $\underline{pl}(2r+1)$ of a tree of size $2r+1, r \in I\!N_0$, is equal to the total path length $pl(L_r)$ of the linear list - tree L_r, which is equal to $r(r+1)$.*

Proof. The generating function $Q_2(z)$, which is defined in Proposition 1, is equal to $Q_{2,2}(z)$. This means, that the probability $p_2(2r+1, (1, 2r-1), (0, d_2))$, that a tree of size $2r+1$ is of type $< (2r+1, 2); (1,0), (2r-1, d_2) >$ (the left subtree consists of one leaf, the right subtree has $2r-1$ nodes), is equal to 1. With $G_2(z) := \sum_{r \geq 1}(2r+1) z^{2r+1}$, we find by relation (3) of Proposition 2:

$$E(z) = \frac{z}{1-z^2} \sum_{r \geq 1} 2r \, z^{2r} = \frac{z^2}{1-z^2} \frac{\partial}{\partial z} \frac{z^2}{1-z^2} = \frac{2z^3}{(1-z^2)^3}.$$

The expected total path length $\underline{pl}(2r+1)$ is then equal to the coefficient $[z^{2r+1}]E(z)$ of z^{2r+1} of the generating function $E(z)$, hence, equal to $r(r+1)$. $\qquad\Box$

The $\mathcal{L}$-model is a "worst case" - model, which differs from the BST - model by the fact that $Q_{1,2}(z) \equiv 0$ instead of $\frac{z^3}{1-z^2}$. This observation leads to the following

Definition 3. *Let c be a nonnegative real number. Let $Q_{1,0}(z) = Q_{2,0}(z) := z$ and $Q_{2,2}(z) := \frac{z^3}{1-z^2}$ be characteristics.*

I) Let the $BST^{(I)}(c)$ - model be defined by choosing $Q_{1,2}(z) := \dfrac{cz^3}{1-z^2}$;

II) Let the $BST^{(II)}(c)$ - model be defined by choosing $Q_{1,2}(z) := \dfrac{z}{1-cz^2} - z$;

III) Let the $BST^{(III)}(c)$ - model be defined by choosing $Q_{1,2}(z) := \dfrac{z}{(1-z^2)^c} - z$.

Obviously, if $c = 0$, all models coincide with the $\mathcal{L}$ - model, and, if $c = 1$, all models correspond to the traditional BST - model.

Using Formula (1) in Definition 1, it is a simple matter to obtain expressions for the probability $p_2(2r+1, (2r_1+1, 2r_2+1), (d_1, d_2))$, that a tree of size $2r+1$, $r \geq 2$, has $2r_1 + 1$ nodes in its left subtree and $2r_2 + 1$ nodes in its right subtree, a parameter, which makes the nature of each particular model clear:

Proposition 4. *Let c be a nonnegative real number. Then the probability $p_2(2r + 1, (2r_1 + 1, 2r_2 + 1), (d_1, d_2))$, $r \geq 2$, $d_1, d_2 \in \{0, 2\}$, that a tree of size $2r + 1$ has a left subtree of size $2r_1 + 1$ and a right subtree of size $2r_2 + 1$, under the three models defined above, is given by:*

I) $p_2(2r + 1, (2r_1 + 1, 2r_2 + 1), (d_1, d_2)) = \dfrac{c^{1-\delta_{r_1,0}}}{1 + c(r - 1)}$,

II) $p_2(2r + 1, (2r_1 + 1, 2r_2 + 1), (d_1, d_2)) = \dfrac{1 - c}{1 - c^r} c^{r_1}$,

III) $p_2(2r + 1, (2r_1 + 1, 2r_2 + 1), (d_1, d_2)) = \dfrac{\Gamma(r)\Gamma(c+1)}{\Gamma(c+r)} \dfrac{\Gamma(c + r_1)}{\Gamma(r_1 + 1)\Gamma(c)}$,

where r_1 can range from 0 to $r - 1$. Here δ denotes the Kronecker symbol and Γ denotes the Gamma function [1].

Proof. In each of the cases we have to compute $Q_2(z)$ and to plug the required coefficients into formula (1):

I) $Q_2(z) = \dfrac{z^3}{1 - z^2} + \dfrac{cz^5}{(1-z^2)^2} = \sum_{r \geq 1} (1 + c(r - 1)) z^{2r+1}$. The coefficient $q_1(2r_1 + 1, d_1)$ is equal to $c^{1-\delta_{r_1,0}}$, and $q_2(2r_2 + 1, d_2) = 1$.

II) Here, $Q_2(z) = z^3 \dfrac{1}{1 - z^2} \dfrac{1}{1 - cz^2} = \sum_{r \geq 1} \dfrac{1 - c^r}{1 - c} z^{2r+1}$, and $q_1(2r_1 + 1, d_1) = c^{r_1}$.

III) By [1, formula 6.1.21],

$$Q_2(z) = \sum_{r \geq 1} \frac{\Gamma(c+r)}{\Gamma(r)\Gamma(c+1)} z^{2r+1}, \text{ and } Q_{1,2}(z) = \sum_{r \geq 1} \frac{\Gamma(c+r)}{\Gamma(r+1)\Gamma(c)} z^{2r+1}. \qquad \square$$

Corollary 1. *Let c be a nonnegative real number. The probability $p_{s,r}^{(\circ)}(c)$ $\left(p_{l,r}^{(\circ)}(c)\right)$, $\circ \in \{I, II, III\}$, that, under the $BST^{(\circ)}(c)$ - model, the first of $r \geq 2$ distinct keys is the smallest (largest) one, is given by:*

$$I) \quad p_{s,r}^{(I)}(c) \quad = \frac{1}{1 + c(r-1)}, \quad \left(p_{l,r}^{(I)}(c) \quad = c \, p_{s,r}^{(I)}(c)\right);$$

$$II) \quad p_{s,r}^{(II)}(c) \quad = \frac{1 - c}{1 - c^r}, \quad \left(p_{l,r}^{(II)}(c) \quad = c^{r-1} p_{s,r}^{(II)}(c)\right);$$

$$III) \quad p_{s,r}^{(III)}(c) \quad = \frac{\Gamma(r)\Gamma(c+1)}{\Gamma(c+r)}, \quad \left(p_{l,r}^{(III)}(c) = \frac{c}{c+r-1}\right).$$

Proof. The corollary follows from Proposition 4 by setting $r_1 := 0$ $(r_1 := r - 1)$. $\square$

Note that $p_{s,r}^{(\circ)}(0) = 1$ and $p_{l,r}^{(\circ)}(0) = 0$, $\circ \in \{I, II, III\}$, because if $c = 0$, a given sequence of r distinct keys is assumed to be already sorted. If $c = 1$, we find that $p_{s,r}^{(\circ)}(1) = p_{l,r}^{(\circ)}(1) = \frac{1}{r}$, $\circ \in \{I, II, III\}$, which corresponds to the traditional BST - model, under which all $r!$ permutations of r distinct are assumed to be equally likely $(p_{s,r}^{(II)}(1) = \frac{1}{r}$ follows by Hospital's rule). Clearly, if $r = 1$, then all the probabilities defined in Corollary 1 are equal to 1.

Figure 1 shows the graphs of the probability functions $p_{\bullet,5}^{(\circ)}(c)$, $(\bullet, \circ) \in \{s, l\} \times \{I, II, III\}$. The functions $p_{s,5}^{(\circ)}(c)$ are monotonically decreasing, the functions $p_{l,5}^{(\circ)}(c)$ are monotonically increasing, $\circ \in \{I, II, III\}$. As the graph shows, model $BST^{(II)}(c)$ causes the strongest deformation, an observation which we shall verify in the next section. For instance, consider the $BST^{(I)}(c)$ - model, and let $c = \frac{1}{2}$. Furthermore, let the r distinct keys be choosen from $\{i \mid 1 \leq i \leq r\}$. A short computation shows, that the probability, that the first key is 1, is equal to $\frac{2}{r+1}$. The probability, that the first key is i, $i \in [2:r]$, is equal to $\frac{1}{r+1}$. This causes a "deformation" of the BST - model "to the right side".
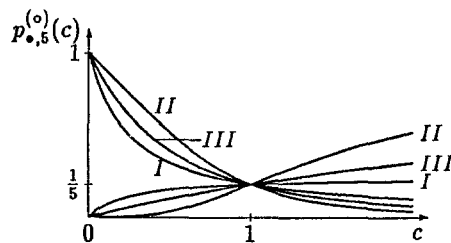
Figure 1. The monotonically decreasing probability functions $p_{s,5}^{(\circ)}(c)$ and their corresponding monotonically increasing probability functions $p_{l,5}^{(\circ)}(c)$, $\circ \in \{I, II, III\}$.

3. The Expected Total Path Length

In this section, we shall consequently make use of relation (3) of Proposition 2. In each particular case, the function $Q_2(z)$ can be picked from the corresponding proof of Proposition 4. We shall focus our attention to the main order of growth of the total path length, because the equations that are satisfied by the intervening generating functions cannot always be solved explicitly.

3.I. The $BST^{(I)}(c)$ - model

The main result is, that, unless $c = 0$, the value of c has no influence upon the dominant term of the asymptotic equivalent to the expected total path length:

Theorem 1. *Let c be a positive real number. Then, under the $BST^{(I)}(c)$ - model the expected total path length $\underline{pl}(2r+1)$ of a tree of size $2r+1$, $r \to \infty$, is asymptotically given by:*

$$\underline{pl}(2r+1) \; = 4r \log(r) + O(r).$$

Proof. Plugging the characteristics $Q_{i,j}(z)$, $i \in \{1,2\}$, $j \in \{0,2\}$, the function $Q_2(z)$ and the function $G_2(z) := \sum_{r \geq 1}(2r+1) \, z^{2r+1}$ into Formula (3), dividing the whole resulting expression by z, replacing z^2 by z, setting $e_r := \underline{pl}(2r+1)$ and using the identity $\sum_{r \geq 1}(2r+1) \, z^r = \frac{2z}{(1-z)^2} + \frac{z}{1-z}$, we obtain by a lengthy computation:

$$czH'(z) \; + \; \frac{1 - c - 2z + z^2 - cz^2}{1 - z} \; H(z) \; = \; \frac{2z(1 - z + 2cz)}{(1-z)^3}, \tag{4}$$

where $H(z) := \sum_{r \geq 1} e_r \, z^r$, with $H(0) = 0$.

The solution of the homogeneous equation (4) is $a \, \frac{e^{z(\frac{1}{c}-1)}z^{1-\frac{1}{c}}}{(1-z)^2}$, where a is an arbitrary constant. The classical *variation of constant* - method [4, p. 99] yields

$$H(z) \; = \; a \, \frac{e^{z(\frac{1}{c}-1)}z^{1-\frac{1}{c}}}{(1-z)^2} \; + \; \frac{e^{z(\frac{1}{c}-1)}z^{1-\frac{1}{c}}}{(1-z)^2} \, \frac{2}{c} \int \frac{e^{z(1-\frac{1}{c})}z^{\frac{1}{c}-1}(1 - z + 2cz)}{1 - z} dz, \tag{5}$$

where the constant a must be set to 0, because $H(z)$ is equal to 0 and analytic at $z = 0$. Unfortunately, there doesn't seem to be a closed solution of (5) for an arbitrary value of c. However, it is not hard to see, that the integrand has exactly one simple pole at $z = 1$. The expansion of the integrand around this dominant singularity is of the form $2ce^{(1-\frac{1}{c})}(1 - z)^{-1} + O(1)$. Hence, $H(z)$ has a singular expansion of the form

$$H(z) \; = \; \frac{4}{(1-z)^2} \log \left(\frac{1}{1-z} \right) \; + \; O \left(\frac{1}{(1-z)^2} \right).$$

The theorem follows by the fact that $[z^r]\frac{1}{(1-z)^2} \log(\frac{1}{1-z}) = (r+1)(H_{r+1} - 1)$, where H_r denotes the r-th harmonic number $H_r := \sum_{1 \leq i \leq r} \frac{1}{i}$ (see [3, Eq. 7.43]). $\qquad \square$

It would be nice to have a better estimate of the O-term, however, this would require the knowledge about the coefficient of $(1 - z)^0$ in the expansion of the solution of the integral around $z = 1$.

3.II. The $BST^{(II)}(c)$ - model

As we shall see, this model resembles to the $\mathcal{L}$ - model, if $c \neq 1$:

Theorem 2. *Let c be a nonnegative real number, $c \neq 1$. Then, under the $BST^{(II)}(c)$ - model, the expected total path length $\underline{pl}(2r + 1)$ of a tree of size $2r + 1$, $r \to \infty$, is asymptotically given by:*

$$\underline{pl}(2r+1) \;=\; \begin{cases} (1-c)r^2 + O(r), & \text{if } c < 1, \\[2mm] (1-\frac{1}{c})r^2 + O(r), & \text{if } c > 1. \end{cases}$$

Proof. Let $c < 1$. By an analogous computation as in the proof of Theorem 1, we find with $H(z) := \sum_{r \geq 1} e_r \, z^r$:

$$H(z) \;=\; \frac{2(1-c)z(1-cz^2)}{(1-z)^3(1-cz)} \;+\; \frac{(1-cz)^2}{(1-z)^2}\, H(cz).$$

Iterating this equation $i \in I\!N$ times yields

$$H(z) \;=\; H(c^i z)\frac{(1-c^i z)}{(1-z)^2} \;+\; \frac{2z(1-c)}{(1-z)^2} \sum_{0 \leq s \leq i-1} \frac{c^s(1-c^{2s+1}z^2)}{(1-c^s z)(1-c^{s+1}z)},$$

and for $i \to \infty$ we obtain

$$H(z) \;=\; \frac{2z(1-c)}{(1-z)^2}\left[\frac{1-cz^2}{(1-z)(1-cz)} \;+\; \sum_{s \geq 1} \frac{c^s(1-c^{2s+1}z^2)}{(1-c^s z)(1-c^{s+1}z)} \right].$$

The first part of the theorem follows by Darboux's Theorem [4].

For $c > 1$, we repeat the whole procedure, but now with $H(cz)$ instead of $H(z)$. $\square$

3.III. The $BST^{(III)}(c)$ - model

In this section we first derive a difference differential equation, using relation (3). This equation turns out to be an ordinary m-th order differential equation, if c is equal to some positive integer m. An asymptotic solution of this differential equation can be found for odd m or for $m = 2$.

Lemma 1. *Let $F(c, z)$ be the generating function $F(c, z) := \sum_{r \geq 1} e_r \frac{\Gamma(c+r)}{\Gamma(1+r)} z^r$, where $e_r := pl(2r + 1)$ is the unknown expected total path length under the $BST^{(III)}(c)$ - model. Then, if $c > 0$, the function $F(c, z)$ satisfies the following equation:*

$$\frac{\partial}{\partial z} F(c, z) \;=\; \frac{\Gamma(c+1)2(1+cz)}{(1-z)^{2+c}} \;+\; \frac{c}{1-z} F(c, z) \;+\; \frac{\Gamma(c+1)}{(1-z)^c} F(1, z). \quad (7)$$

Proof. The lemma follows by relation (3) with [1, Formula 6.1.21]. $\square$

Corollary 2. *Let c be equal to some positive integer $m \geq 1$, and let $H(z)$ be the generating function $H(z) := \sum_{r \geq 1} e_r z^{r+m-1}$. Then, under the $BST^{(III)}(c)$ - model, $H(z)$ satisfies the following ordinary differential equation (ODE):*

$$\frac{\partial^m}{\partial z^m} H(z) - \frac{m}{(1-z)} \frac{\partial^{m-1}}{\partial z^{m-1}} H(z) - \frac{m!}{(1-z)^m z^{m-1}} H(z) = \frac{2\,m!\,(1+mz)}{(1-z)^{m+2}}. \quad (8)$$

Proof. The proof follows from Formula (7) by repeated applications of the Gamma Function's recurrence formula $\Gamma(z+1) = z\Gamma(z)$ (see [1, Formula 6.1.15]) and the fact that $\Gamma(z+1) = z!$, if z is an integer value. □

If $m = 1$, the solution of ODE (8) is (see[6]):

$$H(z) = \frac{4}{(1-z)^2} \log\left(\frac{1}{1-z}\right) - \frac{2z}{(1-z)^2},$$

and the expected total path length $\underline{pl}(2r+1)$ of a tree of size $2r+1$ is

$$\underline{pl}(2r+1) = 4(r+1)(H_{r+1} - 1) - 2r.$$

For larger values of m, there doesn't seem to be a closed expression for the solution of ODE (8), with the exception of $m = 2$:

Theorem 3. *Let $m = 2$. Then, under the $BST^{(III)}(2)$ - model, the expected total path length $\underline{pl}(2r+1)$ of a tree of size $2r+1$, $r \to \infty$, is asymptotically given by:*

$$\underline{pl}(2r+1) = 4r \log(r) + O(r).$$

Proof. Two linear independant solutions of the homogeneous equation of ODE (8) are given by:

$$h_1(z) := \frac{z}{(1-z)^2} \quad \text{and} \quad h_2(z) := \frac{z}{(1-z)^2}\left(z - \frac{1}{z} - 2\log(z)\right).$$

The corresponding Wronski - Determinant $w(z)$ is given by $(1-z)^{-2}$. By the *variation of constant* - method, we find that the solution of ODE (8) for $m = 2$ is:

$$H(z) = a_1 h_1(z) + a_2 h_2(z) + 2\frac{2z \log\left(\frac{1}{1-z}\right) - 1 - 4z - 2z\log(z)}{(1-z)^2},$$

where a_1 and a_2 are some well choosen numbers. This means that the local expansion of $H(z)$ around $z = 1$ is of the form

$$H(z) = \frac{4}{(1-z)^2} \log\left(\frac{1}{1-z}\right) + \frac{b_1}{(1-z)^2} + \frac{4}{1-z} \log\left(\frac{1}{1-z}\right) + O\left(\frac{1}{1-z}\right),$$

where b_1 is a constant depending on a_1 (we need not worry about a_2, because $h_2(z)$ is analytic at $z = 1$). □

Note that under the $BST^{(III)}(2)$ - model, the dominant term of the expansion of $H(z)$ is the same as under the BST - model. This is not the case for larger m, m odd. However, before we shall prove this, let us first state the following lemma, which can be proved by elementary computations.

Lemma 2. *Let a be an arbitrary constant, and let $L^{[m]}$ be the linear operator*

$$L^{[m]}[A(z)] := \frac{\partial^m}{\partial z^m}A(z) - \frac{m}{1-z}\frac{\partial^{m-1}}{\partial z^{m-1}}A(z) - \frac{m!}{(1-z)^m z^{m-1}}A(z), \quad m \in I\!N.$$

Then the following identities hold:

a) $L^{[m]}\left[\dfrac{az^{m-1}}{(1-z)^2}\log\left(\dfrac{1}{1-z}\right)\right] = a\dfrac{(m-1)!(1+mH_{m-1})}{(1-z)^{m+2}} + \displaystyle\sum_{3 \le \nu \le m+1}\dfrac{a_\nu}{(1-z)^\nu},$

where H_r is the r-th harmonic number and the a_ν are computable constants;

b) $L^{[m]}\left[\dfrac{az^{m-1}}{(1-z)^2}\right] \equiv 0;$ *c)* $L^{[m]}\left[\dfrac{az^{m-1}}{1-z}\right] = -a\dfrac{m!}{(1-z)^{m+1}};$

d) $L^{[m]}\left[az^{m-1}(1-z)^k\right] = -a\dfrac{m!}{(1-z)^{k-m}} + \displaystyle\sum_{-1 \le \nu \le k}b_{k,\nu}(1-z)^\nu, \quad 0 \le k \le m-2,$

where the $b_{k,\nu}$ are computable constants;

e) $L^{[m]}\left[az^{m-1}(1-z)^{m-1}\right] = \begin{cases} -a\dfrac{2m!}{1-z} + \displaystyle\sum_{0 \le \nu \le m-1}c_{o,\nu}(1-z)^\nu, & m \text{ odd,} \\[3mm] \displaystyle\sum_{0 \le \nu \le m-1}c_{e,\nu}(1-z)^\nu, & m \text{ even,} \end{cases}$

where the $c_{\bullet,\nu}$, $\bullet \in \{o,e\}$, are computable constants;

f) $L^{[m]}\left[az^{m-1}(1-z)^k\right] =$

$$= a\left((-1)^m\prod_{2-m \le i \le 1}(k+i) - m!\right)(1-z)^{k-m} + \sum_{k-m+1 \le \nu \le k-1}d_{k,\nu}(1-z)^\nu,$$

where $k \ge m$, and where the $d_{k,\nu}$ are computable constants. $\square$

With Lemma 2 we are able to construct approximations to a particular solution of ODE (8), if m is odd, and from these approximations we obtain the following

Theorem 4. *Let $m \ge 3$ be an odd integer. Then, under the $BST^{(III)}(m)$ - model, the expected total path length $\underline{pl}(2r+1)$ of a tree of size $2r+1$, $r \to \infty$, is asymptotically given by:*

$$\underline{pl}(2r+1) = \frac{2m(m+1)}{1+mH_{m-1}}\,r\log(r) + O(r),$$

where H_r denotes the r-th harmonic number.

Proof. Readers who are unfamiliar with the theory of ODE's are referred to relevant books like [5,12]. We find that the homogeneous equation of ODE (8) has a so-called "regular-singular" point at $z = 1$. The corresponding "indicial equation in α" is:

$$(\alpha + 1)\alpha(\alpha - 1)\ldots(\alpha - m + 2) = 1, \text{ if } m \text{ is even,}$$
$$(\alpha + 1)\alpha(\alpha - 1)\ldots(\alpha - m + 2) = -1, \text{ if } m \text{ is odd.}$$

Similar equations appear in [9]. The solutions of them have the following properties:
- $\alpha = -2$ is always a solution, $\alpha = m - 1$ is a solution, if m is even;
- other solutions are conjugated complex numbers with real part larger than -2;
- if m is odd, no pair of solutions differs by an integer.

This means, that, if m is odd, the general solution of ODE (8) is of the form:

$$H(z) = \sum_{1 \le i \le m} \frac{e_i}{(1 - z)^{\alpha_i}} \sum_{\nu \ge 0} f_{i,\nu}(1 - z)^\nu + P(z),$$

where the e_i and the $f_{i,\nu}$, $i \in [1 : m]$, are constants, the α_i, $i \in [1 : m]$, are the m solutions of the indicial equation, and $P(z)$ is a particular solution of ODE (8). If m is even, we do not know, whether this is true, because, if two solutions of the indicial equation differ by an integer, there could be non-logarithm-free solutions of the homogeneous equation of ODE (8). Therefore, let m be odd for the rest of the proof.

In order to obtain a particular solution of ODE (8), we have to solve

$$I^{[m]}[P(z)] := L^{[m]}[P(z)] - \frac{2(m + 1)!}{(1 - z)^{m+2}} + \frac{2m\, m!}{(1 - z)^{m+1}} = 0. \tag{9}$$

Approximations to the solution $P(z)$ of (9) can be constructed step by step using Lemma 2. In each step j, $j \in \mathbb{N}_0$, we have to achieve cancellation of the coefficient of $\frac{(1-z)^j}{(1-z)^{m+2}}$ left by the step $j - 1$.

The construction algorithm works as follows:
    i) set $j := -1$; and set $P_j(z) :\equiv 0$;
    ii) set $j := j + 1$;
    iii) compute $I^{[m]}[P_{j-1}(z)]$, and extract its coefficient of $\frac{(1-z)^j}{(1-z)^{m+2}}$;
    iv) depending on j, set $R(z)$ to:
        $\frac{z^{m-1}}{(1-z)^2} \log\left(\frac{1}{1-z}\right)$, if $j = 0$, and $z^{m-1}(1 - z)^{j-2}$, if $j \ge 1$;
    v) set $P_j(z) := P_{j-1}(z) + aR(z)$, where the constant a
        is choosen such that the coefficient of $\frac{(1-z)^j}{(1-z)^{m+2}}$ in $I^{[m]}[P_j(z)]$ vanishes;
    vi) goto ii).

The theorem follows by the fact that $P(z) = P_\infty(z)$. $\square$

Note that the algorithm doesn't work for even m, because case e) of Lemma 2 shows, that we cannot achieve cancellation of the coefficient of $\frac{1}{1-z}$. Note also that Theorem 4 remains valid for $m = 1$ and $m = 2$, because the 0-th harmonic number H_0 is equal to 0.

4. Algorithms

The first three sections are molded by mathematical considerations. At this point, it is worth summarizing the process employed above and giving an interpretation of our models and results.

We started with the traditional model of binary search trees BST. Under this model, all $r!$ permutations of r distinct keys are assumed to be equally likely. For instance, if $r = 3$, the 6 permutations are $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$, and the probabilities of the resulting search trees are given by the first line of Figure 2. The probability, that the left tree is of size $2r_1 + 1$, $0 \leq r_1 \leq r - 1$, is equal to $\frac{1}{3}$ in this special case, and $\frac{1}{r}$ in the case of r distinct keys. This term $\frac{1}{r}$ can be viewed as the *granularity of the tree type probability distribution*, where the term *tree type* is used in the sense of Definiton 1.

The BST - model can be described by generating functions, the characteristics. Using an external nonnegative real number c, one of these characteristics has been deformed. The deformation induces a deformation of the distribution of the tree type probabilities.

For instance, let us consider the $BST^{(I)}(c)$ - model, and let $c = \frac{1}{m}$ or $c = m$, where $m \in I\!N$ is an arbitrary positive integer. In the case $c = \frac{1}{m}$, the probability that a tree of size $2r + 1$, $r \geq 2$, has a left subtree of size 1 is equal to $\frac{m}{m+r-1}$, and the probability that the left subtree is of size $2r_1 + 1$, $1 \leq r_1 \leq r - 1$, is equal to $\frac{1}{m+r-1}$.
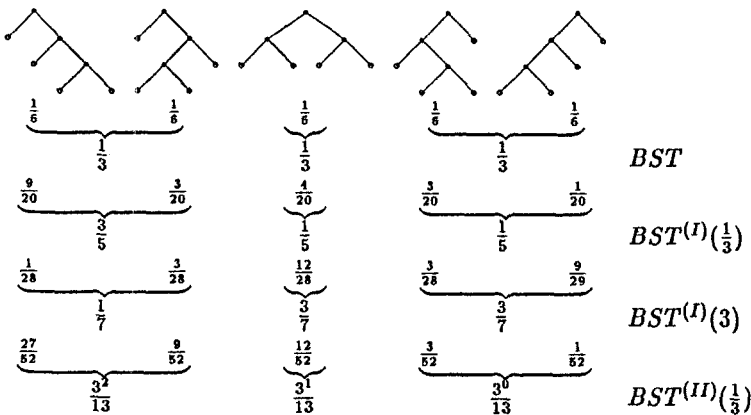

Figure 2. Binary search trees built from multisets of permutations of $\{1, 2, 3\}$ together with their probabilities and "tree type probabilities" under various models.

This recursive "splitting law on tree types" can be translated in a natural way into an algorithm. For a given a real number $c = \frac{1}{m}$, $m \in I\!N$, Program 1 generates a multiset of permutations of r distinct integer keys, such that the probability distribution over the resulting family of binary search trees is equal to that under the $BST^{(I)}(\frac{1}{m})$ - model, provided that all permutations of this multiset are choosen equally likely. In this program, the data type "permutation of length l", $l \in I\!N_0$, is an object of the form $(l, i_1, \ldots, i_l)$, where the $i_j \in I\!N$ are pairwise distinct integers. [1]

If we start Program 1 with $m := \frac{1}{3}$ and $keyset := \{1, 2, 3\}$, we finally obtain the following multiset ("$i * (\ldots)$" means, that permutation $(\ldots)$ occures i times):

$$\{1 * (3, 2, 1), 3 * (3, 1, 2), 1 * (2, 3, 1), 3 * (2, 1, 3), 3 * (1, 3, 2), 9 * (1, 2, 3)\},$$

and this induces the probability distribution given in the second line of Figure 2. This means, that the probability that the left subtree is of size 1, is equal to $\frac{3}{5}$, and the probability, that it is of size $2r_1 + 1$, $r_1 \in \{1, 2\}$, is equal to $\frac{1}{5}$, a finer granularity than under the BST - model.

```
program models;

type pset = multiset of permutation;
     iset = set of integer;

function model (m: integer; keyset: iset): pset;

var j,r: integer;
    ptmp: pset;

begin r := |keyset|;

    if r = 0 then model := {(0)};
    else if keyset = {k₁} then model := {(1,k₁)}
        else begin

            model := ∅;
            sort the r keys k_{i_j} of keyset, such that k₁ < k₂ < ... < k_r;

            for j:=1 to r do begin
                ptmp := model (m, keyset\{k_j});
                ptmp := m^{δ_{j,1}} copies of ptmp; *)
                for each (r − 1, k_{i_1}, ..., k_{i_{r−1}}) ∈ ptmp do
                    model := model ∪ {(r, k_j, k_{i_1}, ..., k_{i_{r−1}})}
            end
        end
end;

var modelset: pset;
    keyset: iset;

begin keyset := {k | 1 ≤ k ≤ r};
      modelset := model (m, keyset)
end.
```

Program 1 generates a multiset of key sequences inducing the $BST^{(I)}(\frac{1}{m})$ - model.

---

[1]The purpose of the first (additional) l is to improve the readability of the program. The resulting multiset of permutations is finally to be understood without the l.

If we replace the star-marked line of Program 1 by the line

$$ptmp := m^{1-\delta_{j,1}} \text{ copies of } ptmp; \ ,$$

then we obtain the $BST^{(I)}(m)$ - model, and, if $m = 3$, this leads to the probability distribution presented in the third line of Figure 2, and its granularity is $\frac{1}{7}$.

In both cases, $c = \frac{1}{m}$ and $c = m$, we observe that, with the exception of the case $r_1 = 0$, the probability that a tree of size $2r + 1$, $r \geq 3$, has a left subtree of size $2r_1 + 1$ and a right subtree of size $2r_2 + 1$, $r_2 \geq 1$, is equal to the probability that the left subtree is of size $2(r_1 + 1) + 1$ and the right subtree is of size $2(r_2 - 1) + 1$. In other words, with the exception of the case that the first key is the smallest one, the "tree type probabilities" remain equally distributed under the $BST^{(I)}(\circ)$ - model, $\circ \in \{\frac{1}{m}, m\}$. This is the reason for the fact, that the expected total path length doesn't increase significantly, as $r \to \infty$.

Under the $BST^{(II)}(\frac{1}{m})$ - model, we are faced with a totally different situation. An algorithm, which corresponds to this model, can be constructed by replacing the star-marked line of Program 1 by the line

$$ptmp := m^{r-j} \text{ copies of } ptmp; \ .$$

Instead of making m copies of the recursively computed multiset of permutations of $r - 1$ distinct keys, iff the key, which has been splitted off, is the smallest one, as the original Program 1 does, the modified program now makes m^{r-j} copies, iff the key, which has been splitted off, is the j-th smallest one.

If we start the modified program with $m := 3$ and $keyset := \{1, 2, 3\}$, we finally obtain the multiset

$$\{1 * (3, 2, 1), 3 * (3, 1, 2), 3 * (2, 3, 1), 9 * (2, 1, 3), 9 * (1, 3, 2), 27 * (1, 2, 3)\},$$

and the resulting probability distribution is given in the last line of Figure 2. We observe, that the tree type probability grows exponentially in the size of the right subtree instead of remaining constant, as it is the case under the $BST^{(I)}(\frac{1}{m})$ - model (with the one exception mentioned above), which causes the $O(r^2)$ order of growth of the expected total path length.

If we replace the star-marked line of Program 1 by the line

$$ptmp := m^{j-1} \text{ copies of } ptmp; \ ,$$

we obtain the $BST^{(II)}(m)$ - model, which behaves symmetrically to the $BST^{(II)}(\frac{1}{m})$ - model.

Finally, let us consider the $BST^{(III)}(m)$ - model, $m \in I\!N$. An inspection of Proposition 4 shows, that the probability, that a tree of size $2r + 1$, $r \in I\!N$, has a left subtree of size $2r_1 + 1$, $r_1 \in I\!N_0$, and a right subtree of size $2r_2 + 1$, $r_2 \in I\!N_0$, is given by:

$$p_2(2r + 1, (2r_1 + 1, 2r_2 + 1), (d_1, d_2)) = \frac{m}{r} \prod_{1 \leq k \leq m-1} \frac{r_1 + k}{r + k}.$$

This formula translates into the star-marked program line

$$ptmp := \prod_{0 \le k \le m-2}(j + k) \text{ copies of } ptmp; \quad .$$

This means, that the tree type probability grows polynomially in the size of the left subtree, where the degree of the polynomial is equal to $m - 1$, a moderate growth, compared with the $BST^{(II)}(m)$ - model, causing only a modification of the dominant coefficient of the asymptotic equivalent to the expected total path length, but not a general modification of the order itself.

5. Concluding Remarks

We introduced three new probability models by choosing different characteristics. Clearly, the choice of these characteristics is artificial, and many other characteristics can be selected. However, as we have seen in Section 4, the three characteristics used in this paper represent three totally different probability distributions, while they are simple enough, so that we are able to compute the resulting expected path length. They give us a feeling about the question, what must happen to the BST - model in order to get "bad". Note that although we are relatively free in the selection of a characteristics, the choice must be done carefully. For instance, we cannot use $Q_{1,2}(z) := \frac{z^3}{c-z^2} - (1 - c)z$, $0 < c < 1$, because its expansion is $Q_{1,2}(z) = (c-1)z + \sum_{r \ge 1} c^{-r} z^{2r+1}$, but Definition 1 requires that $Q_{1,2}(z)$ is of the form $Q_{1,2}(z) = \sum_{r \ge 1} a_r z^{2r+1}$.

References

[1] Abramowitz, M., Stegun, I.A.: *Handbook of Mathematical Functions.* New York: Dover Publications, (1970)

[2] Casas, R., Díaz, J., Martínez, C.: *Average-case Analysis on Simple Families of Trees Using a Balanced Probability Model.* Series Formelles Et Combinatoire Algebrique, Actes de Colloque, 133-143, Bordeaux, (2.-4. Mai 1991)

[3] Graham, R.L., Knuth, D.E., Patashnik, O.: *Concrete Mathematics.* Addison-Wesley, Reading, Mass., (1988)

[4] Henrici, P.: *Applied and Computational Complex Analysis, Vol. 2.* Wiley Interscience, New York, (1977)

[5] Kamke, E.: *Differentialgleichungen.* Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, (1967)

[6] Kemp, R.: *Additive Weights of Non-Regularly Distributed Trees.* Annals of Discrete Mathematics 33, 129-155, (1987)

[7] Kemp, R.: *The Expected Additive Weight of Trees.* Acta Inf. 26, 711-740, (1989)

[8] Knuth, D.E.: *The Art of Computer Programming, Vol. 1, (2nd ed.).* Addison-Wesley, Reading, Mass., (1973)

[9] Mahmoud, H., Pittel, B.: *Analysis of the Space of Search Trees Under the Random Insertion Algorithm.* Journal of Alg. 10, 52-75, (1989)

[10] Meir, A., Moon, J.W.: *On the Altitude of Nodes in Random Trees.* Can. J. Math. 30, 997-1015, (1978)

[11] Trier, U.: *Additive Weights Under the Balanced Probability Model.* Johann Wolfgang Goethe Universität Frankfurt, Fachbereich Informatik (20), Preprint, (1992)

[12] Wasow, W.: *Asymptotic Expansions for Ordinary Differential Equations.* Wiley Inter-science, New York, (1965)