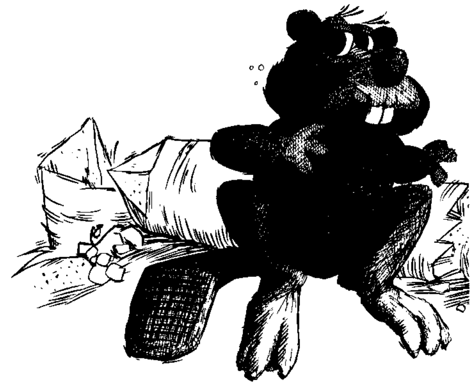
UNIVERSITY OF WATERLOO COMPUTER SCIENCE DEPARTMENT



On the Path Length of Binary Trees

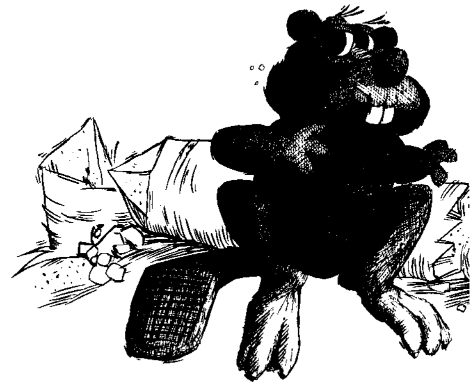
Rolf Klein
Derick Wood

Data Structuring Group

Research Report
CS-87-06

February, 1987

UNIVERSITY OF WATERLOO COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF WATERLOO COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF WATERLOO COMPUTER SCIENCE DEPARTMENT
UNIVERSITY OF WATERLOO COMPUTER SCIENCE DEPARTMENT



On the Path Length of Binary Trees

Rolf Klein
Derick Wood

Data Structuring Group

Research Report
CS-87-06

February, 1987

On the Path Length of Binary Trees*

Rolf Klein † Derick Wood‡

February 16, 1987

Abstract

We show that the external path length of a binary tree is closely related to the ratios of means of certain integers and establish the upper bound

$$\text{External Path Length} \leq N(\log_2 N + \triangle - \log_2 \triangle - 0.6623)$$

where N denotes the number of external nodes in the tree and $\triangle$ is the difference in length between a longest and a shortest path. Then we prove that this bound is (almost) achieved if N and $\triangle$ are arbitrary integers that satisfy $\triangle \leq \sqrt{N}$. If $\triangle > \sqrt{N}$, we construct binary trees whose external path length is at least as large as $N(\log_2 N + \phi(N, \triangle)\triangle - \log_2 \triangle - 4)$, where $\phi(N, \triangle) = (1 + \Theta(\frac{\triangle}{N}))^{-1}$.

Keywords: Binary trees, path length, comparison cost, node visit cost, ratio of means.

1 Introduction

The time taken by a search operation in a search tree depends on the length of the path from the root to the node that contains the desired information. More generally, the execution time an algorithm needs to reach a certain state from its initial state is related to the length of the corresponding path in the decision tree. Therefore, the path length of a tree is a cost measure of great importance for the analysis of algorithms.

We consider the *external path length* EPL(T) of an extended binary tree T, that is, the total number of edges along all the paths from the root to the

*This work was partially supported by a Natural Sciences and Engineering Research Council of Canada Grant A-5692. It was done during the first author's stay with the Data Structuring Group in Waterloo in 1986/87.

†Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Universität Karlsruhe, Postfach 6980, 7500 Karlsruhe, West Germany

‡Data Structuring Group, Department of Computer Science, University of Waterloo, WATERLOO, Ontario N2L 3G1, CANADA

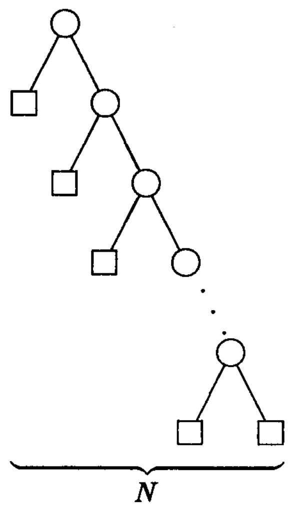


Figure 1: A snake.

external nodes of T. If N denotes the number of external nodes (the *weight* of T) then $\mathrm{EPL}(T)/N$ is just the average length of a path from the root of T to an external node. It is well known that the external path length is a minimum if and only if all paths in T differ in length by at most 1. In this case

$$EPL(T) = N(\log_2 N + 1 + \theta - 2^\theta) \tag{1}$$

holds, where $\theta = \lceil \log_2 N \rceil - \log_2 N \in [0, 1)$; see Knuth [6], p. 194. This formula establishes a lower bound for the external path length. On the other hand, the path length takes its maximum value

$$\frac{N(N+1)}{2} - 1$$

if the tree is a "snake" as shown in Figure 1[1]. Here the shortest path is $N-2$ levels shorter than the longest path.

In this paper we present an upper bound for the external path length in terms of the weight N and the maximal path length difference $\triangle$ (see Figure 2) by proving that

$$EPL(T) \leq N(\log_2 N + \triangle - \log_2 \triangle - \Psi(\triangle)) \tag{2}$$

holds for all binary trees, where

$$\Psi(\triangle) = 0.9139 - o(1) \geq 0.6623$$

and $o(1)$ tends to zero as $\triangle$ tends to infinity. For the tree in Figure 2, for example, we obtain the value 31.56 whereas its actual path length is equal to 30.

[1]This and the following figures have been produced using TreeTEX; see [1]

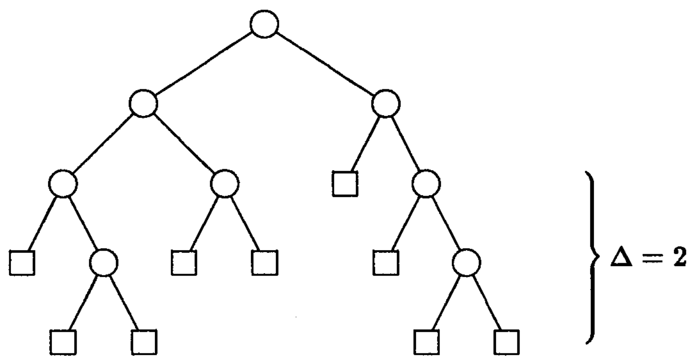


Figure 2: An example tree.

In order to establish this upper bound we first show, via the Kraft Inequality, that the external path length of a binary tree is related to the ratio of the geometric and the harmonic means of the integers 2^{l_i}, where l_i denotes the length of the i-th path. It is also related to the ratio of the arithmetic and the geometric mean of the integers 2^{h-l_i}, where h denotes the height of the tree. Either relation can be used to obtain an upper bound for the path length by applying a general theorem by Specht that imposes an upper bound on the ratio of means of arbitrary real numbers. However, we shall also give a direct proof for our result, in order to keep this paper self-contained.

In Section 4 we discuss the tightness of our upper bound. If $\triangle = 1$, then expression (2) is exactly equal to the maximum value the lower bound in (1) can take. If $\triangle$ and N are integers of arbitrary, independent orders of magnitude, we can build a tree whose path length is greater than

$$N(\log_2 N + \triangle - \log_2 \triangle - 4)$$

if $\triangle \leq \sqrt{N}$, and greater than

$$N\left(\log_2 N + \frac{1}{1+\Theta(\frac{\triangle}{N})}\triangle - \log_2 \triangle - 4\right)$$

if $\triangle > \sqrt{N}$. This shows that the upper bound for the external path length obtained here is tight if $\triangle \leq \sqrt{N}$ and quite sharp if $\triangle > \sqrt{N}$.

2 Path length and ratios of means

Let T be an extended binary tree. We count the *level* of nodes starting with level 0 at the root. The *access path* to a node at level j is of *length* j, because it consists of j edges. The *height* h of the tree T is the maximum level number or, equivalently, the length of a longest path in T. A node at level i is said to be at *height $h - i$ with respect to T*. Furthermore, $N = weight(T)$ denotes the number of external nodes of T. Finally, we let

$$EPL(T) = \sum_{i=1}^{N} l_i$$

where l_i is the length of the path to the i-th external node.

First, we recall the definition of means. Let $a_1, \ldots, a_N$ and $q_1, \ldots, q_N$ be sequences of positive real numbers such that $q_1 + q_2 + \ldots + q_N = 1$. Then

$$M_N^{[-1]}(a,q) = (\sum_{i=1}^{N} \frac{q_i}{a_i})^{-1}$$

is the *weighted harmonic mean*,

$$M_N^{[0]}(a,q) = \prod_{i=1}^{N} a_i^{q_i}$$

is the *weighted geometric mean*, and

$$M_N^{[1]}(a,q) = \sum_{i=1}^{N} q_i a_i$$

is the *weighted arithmetic mean* of the numbers $a_1, \ldots, a_N$ with weights $q_1, \ldots, q_N$.

Lemma 2.1 *Let T be a binary tree of weight N whose paths to the external nodes are of length $l_1, \ldots, l_N$. Let $a_i = 2^{l_i}$ and $q_i = \frac{1}{N}, 1 \leq i \leq N$. Then*

$$\frac{M^{[0]}(a,q)}{M^{[-1]}(a,q)} = \frac{2^{\frac{EPL(T)}{N}}}{N}$$

Proof: By Kraft's Theorem (usually referred to as the Kraft Inequality, see [3]) there exists a binary tree whose paths to the external nodes are of length $l_1, \ldots, l_N$ if and only if

$$\sum_{i=1}^{N} 2^{-l_i} = 1$$

Hence,

$$\begin{aligned} M_N^{[-1]}(a,q) &= \left(\sum_{i=1}^{N} \frac{2^{-l_i}}{N}\right)^{-1} \\ &= N \end{aligned}$$

Furthermore,

$$\begin{aligned} M^{[0]}(a,q) &= \left(\prod_{i=1}^{N} 2^{l_i}\right)^{\frac{1}{N}} \\ &= 2^{\frac{1}{N}EPL(T)} \end{aligned}$$

□

We can compute the external path length and the height of a binary tree of weight N if we know only the heights $h_1, \ldots, h_N$ of the external nodes in the tree. This leads to

Lemma 2.2 *Let T be a binary tree of weight N whose external nodes are of height $h_1, \ldots, h_N$ in T. Let $b_i = 2^{h_i}$ and $q_i = \frac{1}{N}, 1 \leq i \leq N$. Then*

$$\frac{M_N^{[1]}(b,q)}{M_N^{[0]}(b,q)} = \frac{2^{\frac{EPL(T)}{N}}}{N}$$

Proof: After multiplying by 2^h, the Kraft Inequality becomes

$$2^h = \sum_{i=1}^{N} 2^{h-l_i} = \sum_{i=1}^{N} 2^{h_i}$$

Therefore,

$$EPL(T) = Nh - \sum_{i=1}^{N} h_i$$

$$\begin{aligned} \frac{EPL(T)}{N} &= \log_2\left(\sum_{i=1}^{N} 2^{h_i}\right) - \log_2\left(2^{\frac{1}{N}\sum_{i=1}^{N} h_i}\right) \\ &= \log_2\left(\sum_{i=1}^{N} b_i\right) - \log_2\left(\prod_{i=1}^{N} b_i^{\frac{1}{N}}\right) \end{aligned}$$

and

$$2^{\frac{EPL(T)}{N}} = \frac{N M_N^{[1]}(b,q)}{M_N^{[0]}(b,q)}$$

□

Inequalities involving means were first studied by the Pythagoreans and Euclid, and many interesting results have been obtained since. For example, it is well known that

$$M_N^{[-1]}(a,q) \leq M_N^{[0]}(a,q) \leq M_N^{[1]}(a,q)$$

holds, for any sequences of numbers a_i and weights q_i. Either inequality, combined with the corresponding Lemma above, yields immediately

$$EPL(T) \geq N \log_2 N$$

In the next section we will use an upper bound for the ratios of these means discovered by Specht [7] in order to derive a new upper bound for the external path length of binary trees.

3 An upper bound for the external path length

Throughout this paper, $\triangle(T)$ denotes the difference between the length of a longest path of T and the length of a shortest path to an external node. We also refer to $\triangle$ as to the *thickness of the fringe* of T.

Theorem 3.1 *Let T be a binary tree of weight N whose fringe is of thickness $\triangle$. Then*

$$EPL(T) \leq N(\log_2 N + \triangle - \log_2 \triangle - \Psi(\triangle))$$

where

$$\begin{aligned} \Psi(\triangle) &= \log_2 e - \log_2 \log_2 e - \frac{\triangle}{2^{\triangle} - 1} - \log_2 \left(1 - \frac{1}{2^{\triangle}}\right) \\ &= 0.91392867 - o(1) \\ &\geq 0.66229950 \end{aligned}$$

and e denotes the basis of the natural logarithm.

Proof: (first version) By Lemma 2.1,

$$\frac{2^{\frac{EPL(T)}{N}}}{N} = \frac{M_N^{[0]}(a,q)}{M_N^{[-1]}(a,q)}$$

where $a_i = 2^{l_i}$, l_i = length of the path to the i-th external node, and $q_i = \frac{1}{N}$, for $i = 1, \ldots, N$. By a theorem by Specht (Satz 1, (5.4) in [7]) we have

$$\frac{M_N^{[0]}(a,q)}{M_N^{[-1]}(a,q)} \leq \left(\frac{\frac{1}{B} - 1}{-\ln B}\right) e^{\left(-1 - \frac{\ln B}{\frac{1}{B} - 1}\right)} \tag{3}$$

if $B = \frac{M}{m}$ is such that $m \leq a_1, \ldots, a_N \leq M$. If $l = \min_i l_i$, then $B = \frac{2^h}{2^l} = 2^{\triangle}$ will do. The above exponential term equals

$$e^{-1} B^{1 + \frac{1}{B-1}}$$

whereas the left hand factor is equal to

$$\frac{1-\frac{1}{B}}{(\log_2 e)^{-1}\triangle}$$

observing that $\ln x = (\log_2 e)^{-1}\log_2 x$ holds for the natural logarithm. Taking logs yields

$$\begin{aligned}\frac{EPL(T)}{N} - \log_2 N \quad \leq \quad & \log_2\left(1-\frac{1}{2^\triangle}\right) + \log_2\log_2 e - \log_2\triangle \\ & -\log_2 e + \left(1+\frac{1}{2^\triangle-1}\right)\triangle\end{aligned}$$

In order to complete the proof we note that the function $\log_2\left(1-\frac{1}{2^\triangle}\right) + \frac{\triangle}{2^\triangle-1}$ takes its maximum value among all integer arguments $\triangle \geq 1$, if $\triangle = 2$. $\square$

Another proof for the theorem used in the above proof was given by Cargo and Shisha in [2]. In addition, they showed for which values of $a_1,\ldots,a_N$ and B the inequality (3) becomes an equality. However, we now give a direct proof of Theorem 3.1.

Proof: (second version) We want to determine the maximum value of $EPL(T) = \sum_{i=1}^N l_i$ under the condition that $\sum_{i=1}^N 2^{-l_i} = 1$ (the Kraft Inequality), where $\max_i l_i - \min_i l_i = \triangle$. To this end, we let $l_i = X_0 + (\sin X_i)^2\triangle$. Here X_0 denotes the (unknown) length of a shortest path in T. The value of $(\sin X_i)^2$ oscillates in [0,1] as X_i varies in $\Re$, thereby leading to a total path length l_i that lies between X_0 and $X_0+\triangle$.
We consider the function

$$f(X_0,X_1,\ldots,X_N) = \sum_{i=1}^N (X_0 + (\sin X_i)^2\triangle)$$

under the condition that $g(X_0,X_1,\ldots,X_N) = 0$, where

$$g(X_0,X_1,\ldots,X_N) = \sum_{i=1}^N 2^{-(X_0+(\sin X_i)^2\triangle)} - 1$$

For each constrained maximum $p = (a_0,a_1,\ldots,a_N)$ of f there must be a real number λ such that all partial derivatives of $f-\lambda g$ vanish in p, by the Lagrange Multiplier Theorem (see [4], for example). This means

$$0 = \frac{\partial(f-\lambda g)}{\partial X_0}(p) = N + \lambda\ln 2\sum_{i=1}^N 2^{-(a_0+(\sin a_i)^2\triangle)} = N + \lambda\ln 2$$

and

$$0 = \frac{\partial(f - \lambda g)}{\partial X_i}(p) = 2\sin a_i \cos a_i \triangle \left(1 + \lambda \frac{\ln 2}{2^{a_0 + (\sin a_i)^2 \triangle}}\right)$$

for $i = 1, \ldots, N$. The latter equalities imply

$$(\sin a_i)^2 \in \{0, 1\} \text{ or } N = 2^{a_0 + (\sin a_i)^2 \triangle}$$

due to the first equality. Therefore,

$$l_i = a_0 + (\sin a_i)^2 \triangle \in \{a_0, a_0 + \triangle, \log_2 N\}$$

for $i = 1, \ldots, N$. In order to determine how often each of these three values occurs we consider the constrained maxima of

$$f(X, V, W, R) = V^2 X + W^2(X + \triangle) + R^2 \log_2 N$$

subject to the conditions $g_1(X, V, W, R) = 0$ and $g_2(X, V, W, R) = 0$, where

$$g_1(X, V, W, R) = V^2 \frac{1}{2^X} + W^2 \frac{1}{2^{X+\triangle}} + \frac{R^2}{N} - 1$$

represents the Kraft Inequality and

$$g_2(X, V, W, R) = V^2 + W^2 + R^2 - N$$

is because we are considering trees of weight N. Again, for each maximum $p = (x, v, w, r)$ of f_1 there must be real numbers λ and μ such that the partial derivatives of the function $f - \lambda g_1 - \mu g_2$ with respect to the variables X, V, W, and R vanish at p. This means

$$\begin{aligned} 0 &= v^2 + w^2 + \lambda \ln 2 \left(v^2 \frac{1}{2^x} + w^2 \frac{1}{2^{x+\triangle}}\right) \\ &= (N - r^2)\left(1 + \frac{\lambda \ln 2}{N}\right) \end{aligned} \tag{4}$$

due to the constraint conditions, and

$$0 = 2v\gamma(x) \tag{5}$$
$$0 = 2w\gamma(x + \triangle) \tag{6}$$
$$0 = 2r\gamma(\log_2 N) \tag{7}$$

where

$$\gamma(Z) = Z - \lambda \frac{1}{2^Z} - \mu$$

If $r^2 = N$, then according to g_1, $v = w = 0$ and f takes the value $N \log_2 N$ at p—the minimum! Therefore, we must have $\lambda = \frac{-N}{\ln 2}$, due to (4).

The function $\gamma(Z)$ takes its unique minimum if $Z = \log_2 N$, because $\frac{d\gamma}{dZ}(Z) = 1 - \frac{N}{2^Z}$. Hence, $\gamma(y) = \gamma(\log_2 N)$ implies $y = \log_2 N$, for arbitrary real numbers y. If we assume $r \neq 0$ then (7) implies $\gamma(\log_2 N) = 0$.Therefore, due to (5) and (6), v or w must be equal to zero because $\triangle > 0$. Assume $v = 0$ and $w \neq 0$. Then $\gamma(x + \triangle) = \gamma(\log_2 N)$ implies $x + \triangle = \log_2 N$ and again, f takes its minimum at p, a contradiction. The same holds if we assume $v \neq 0$ and $w = 0$ or $v = w = 0$.

Therefore, r must be equal to zero. This yields $v \neq 0$ and $w \neq 0$ (otherwise f would take a minimum), hence $\gamma(x) = 0 = \gamma(x + \triangle)$. So,

$$x + \frac{N}{\ln 2}\frac{1}{2^x} = x + \triangle + \frac{N}{\ln 2}\frac{1}{2^x + \triangle}$$

or

$$x = \log_2\left(\frac{N}{\triangle \ln 2}\right) + \log_2\left(1 - \frac{1}{2^\triangle}\right) \tag{8}$$

The constraint conditions now read as $v^2 + w^2 = N$ and $v^2 2^\triangle + w^2 = \frac{N}{\triangle \ln 2}(2^\triangle - 1)$, the solution of these linear equations being

$$v^2 = N\left(\frac{1}{\triangle \ln 2} - \frac{1}{2^\triangle - 1}\right) \tag{9}$$

$$w^2 = N\left(\frac{2^\triangle}{2^\triangle - 1} - \frac{1}{\triangle \ln 2}\right) \tag{10}$$

Now combining (8) and (10) yields

$$\begin{aligned} f(p) &= v^2 x + w^2(x + \triangle) \\ &= Nx + w^2\,\triangle \\ &= N\,(\log_2 N - \log_2 \ln 2 - \log_2 \triangle \\ &\quad + \log_2\left(1 - \frac{1}{2^\triangle}\right) + \triangle\frac{2^\triangle}{2^\triangle - 1} - \frac{1}{\ln 2}\right) \end{aligned}$$

Therefore, for all binary trees T of weight N and fringe thickness $\triangle$ we have

$$EPL(T) \leq f(p) = N(\log_2 N + \triangle - \log_2 \triangle - \Psi(\triangle))$$

□

We observe a certain similarity between the formula in Theorem 3.1 and the tight upper bound for the path length of AVL trees recently obtained by Klein and Wood [5], caused by the term $\log_2 \triangle$. Namely, for each AVL tree T we have

$$\triangle \leq \frac{1}{2}h(T) \leq \frac{1}{2}1.4404\log_2 N$$

which makes Theorem 3.1 read as

$$EPL(T) \leq N(1.7202 \log_2 N - \log_2 \log_2 N) + O(N)$$

This bound is bigger than the tight upper bound

$$1.4404N(\log_2 N - \log_2 \log_2 N) + O(N)$$

in [5], but the presence of the term $\log_2 \log_2 N$ in the above equation is surprising!

Equations (8), (9), and (10) in the above proof seem to indicate how, for given integers $\triangle$ and N, a binary tree of maximal external path length has to look. Namely, $N\left(\frac{1}{\triangle \ln 2} - \frac{1}{2^{\triangle}-1}\right)$ external nodes should appear at level $x = \log_2\left(\frac{N}{\triangle \ln 2}\right) + \log_2\left(1 - \frac{1}{2^{\triangle}}\right)$, and the rest of them at level $x + \triangle$. But these numbers are reals, not integers, a difficulty that in this case cannot be overcome by rounding! For example, if $N = 320000$ and $\triangle = 14427$ then these formulae yield that $31.99989\ldots$ external nodes should be located at level $4.999995\ldots$. But there is no 2-level-tree that has 32 external nodes at level 5, because it couldn't have any internal node at level 5. Moreover, there is no 2-level-tree at all, if $N < 2^{\triangle}$! Nevertheless, in the next paragraph we will construct binary trees whose external path length comes very close to the upper bound established in Theorem 3.1.

4 Binary trees of high external path length

In order to investigate how close to reality the upper bound established in Theorem 3.1 is we have to allow the parameters N and $\triangle$ to vary independently. If $\triangle = 1$, then $\Psi(\triangle) = \log_2 e - \log_2 \log_2 e = (1 + \ln\ln 2)/\ln 2$, and our upper bound takes the form

$$N(\log_2 N + 1 - (1 + \ln\ln 2)/\ln 2)$$

This is exactly the maximum value of the lower bound $N(\log_2 N + 1 + \theta - 2^{\theta})$ for the external path length, for $\theta = -(\ln\ln 2)/\ln 2$, see formula (1) in Section 1, and Knuth [6], p. 194.

Next we consider the case $2^{\triangle} \leq N$.

Lemma 4.1 *Let $\triangle = 2^a \geq 1$. Then for each integer $s \geq 0$, there exists a binary tree T of weight $N = \Theta(2^{\triangle+s})$ whose fringe is of thickness $\triangle$ such that*

$$EPL(T) \geq N(\log_2 N + \triangle - \log_2 \triangle - 2)$$

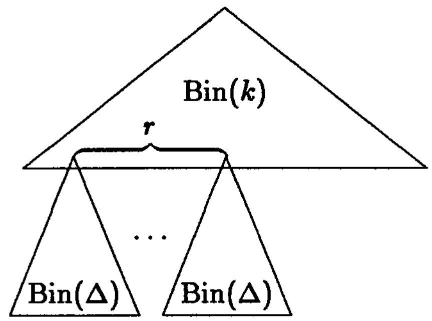


Figure 3: The tree $T_1(r, k, \triangle)$.

Proof: Consider the tree $T = T_1(r, k, \triangle)$ displayed in Figure 3, a complete binary tree of height k, r of whose external nodes are the roots of complete binary trees of height $\triangle$.
Since a complete binary tree $Bin(h)$ of height h has 2^h external nodes and external path length $h2^h$ we have

$$weight(T) = N = 2^k + (2^{\triangle} - 1)r$$

and

$$\begin{aligned} EPL(T) &= k(2^k - r) + r(k + \triangle)2^{\triangle} \\ &= kN + r \triangle 2^{\triangle} \end{aligned}$$

Now let $k = \triangle - \log_2 \triangle + s = 2^a - a + s$ and $r = \triangle 2^{k-\triangle} = 2^s \geq 1$. Then

$$EPL(T) = kN + \triangle^2 2^k$$

and

$$\begin{aligned} N &= \triangle 2^k \frac{\triangle 2^{\triangle} + 2^{\triangle} - \triangle}{\triangle 2^{\triangle}} \\ &= \Theta(\triangle 2^k) \\ &= \Theta(2^{\triangle + s}) \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{EPL(T)}{N} &= k + \triangle \frac{\triangle 2^{\triangle}}{\triangle 2^{\triangle} + 2^{\triangle} - \triangle} \\ &\geq k + \triangle - 1 \end{aligned} \tag{11}$$

On the other hand,

$$\begin{aligned} \log_2 N &= k + \log_2 \triangle + \log_2 \frac{\triangle 2^{\triangle} + 2^{\triangle} - \triangle}{\triangle 2^{\triangle}} \\ &\leq k + \log_2 \triangle + 1 \end{aligned} \tag{12}$$

Combining equations (11) and (12) completes the proof. □

In the above construction we could build a 2-level-tree because $2^{\triangle} \leq N$. But if the fringe grows thicker (in relation to the weight) we have to place the external nodes at more than 2 levels in the tree. This tends to keep the maximal possible external path length smaller than the value of the upper bound. However, if $\triangle \leq \sqrt{N}$, then the difference from $EPL(T)/N$ is only a small additive constant, as the following lemma shows.

Lemma 4.2 *Let $\triangle = 2^a \geq 1$. Then, for each integer k in $[1, \triangle - a]$, there exists a binary tree T of weight $N = \Theta(\triangle 2^k)$ whose fringe is of thickness $\triangle$ such that*

A. $EPL(T) \geq N(\log_2 N + \triangle - \log_2 \triangle - 4)$

holds if $\triangle \leq \sqrt{N}$ and

B. $EPL(T) \geq N\left(\log_2 N + \frac{1}{1+\Theta(\frac{\triangle}{N})} \triangle - \log_2 \triangle - 4\right)$

holds otherwise.

Proof: We consider the tree $T = T_2(k, s, t)$ shown in Figure 4, a complete binary tree of height $k \geq 1$ in one of whose external nodes a "snake" of length s originates that leads to another complete binary tree of height t. Clearly, $\triangle = s + t$, $N = 2^k + s - 1 + 2^t$ and

$$\begin{aligned} EPL(T) &= k(2^k - 1) + sk + \frac{s(s+1)}{2} + (k + s + t)2^t \\ &\geq kN + \triangle 2^t \end{aligned}$$

Now let $t = k + a$ and $s = \triangle - t$. Then $2^t = \triangle 2^k$ and

$$N = (\triangle + 1)2^k + s - 1 \tag{13}$$

Therefore

$$\frac{EPL(T)}{N} \geq k + \frac{\triangle^2 2^k}{(\triangle + 1)2^k + s - 1}$$

Because of

$$s - 1 = \triangle - t - 1 = 2^a - 1 - t \leq 2^{2a-t} 2^{t-a} = \triangle^2 2^{-t} 2^k \tag{14}$$

we obtain

$$\begin{aligned} \frac{EPL}{N} &\geq k + \triangle \frac{\triangle}{\triangle + 1 + \triangle^2 2^{-t}} \\ &\geq k + \triangle \frac{1}{1 + \triangle 2^{-t}} - 1 \end{aligned} \tag{15}$$

The latter inequality can be verified easily by crossmultiplying. Because

$$\left(1+\frac{1}{\triangle}\right)2^t \le \left(1+\frac{1}{\triangle}\right)2^t + s = N+1 \le \left(1+\frac{1}{\triangle}\right)2^t + \triangle$$

we have $\frac{N}{2^t} \le 1 + \frac{1}{\triangle} + \frac{\triangle}{2^t} \le 1 + \frac{1}{\triangle} + \frac{1}{2}$ and

$$N = \Theta(2^t) = \Theta(\triangle 2^k)$$

The former, applied to (15), yields

$$\frac{EPL(T)}{N} \ge k + \triangle \frac{1}{1+\Theta(\frac{\triangle}{N})} - 1 \tag{16}$$

where $\Theta\left(\frac{\triangle}{N}\right) = \frac{3}{2}\frac{\triangle}{N} + \frac{1}{N}$. If $\triangle^2 \le N$, then

$$\left(1+\frac{3}{2}\frac{\triangle}{N}+\frac{1}{N}\right)^{-1} \ge 1 - \frac{3}{2\triangle}$$

hence

$$\frac{EPL(T)}{N} \ge k + \triangle - \frac{5}{2} \tag{17}$$

On the other hand, equation (13) yields

$$\begin{aligned} \log_2 N &= \log_2\left(\triangle 2^k\left(1+\frac{2^k+s-1}{\triangle 2^k}\right)\right) \\ &\le \log_2 \triangle + k + \frac{3}{2} \end{aligned} \tag{18}$$

because

$$\begin{aligned} 1+\frac{2^k+s-1}{\triangle 2^k} &\le 1+\frac{1+\triangle^2 2^{-t}}{\triangle} \\ &\le \frac{5}{2} \end{aligned}$$

according to (14). Now assertions A and B follow by combining (17) and (16) with (18), correspondingly. □

Lemma 4.2 covers the case where $2^\triangle > N$ holds (by orders of magnitude). According to assertion A, the upper bound for EPL established in Theorem 3.1 is tight up to a small additive $O(N)$ term if $\triangle \le \sqrt{N}$. If $\triangle$ is increased beyond $\sqrt{N}$, then the coefficient

$$\rho = \frac{1}{1+\Theta(\frac{\triangle}{N})}$$

of $\triangle$ in B begins to decrease. However, as long as $\triangle = O(N^\alpha)$ holds for some real number $\alpha > 1$ the value of ρ still comes arbitrarily close to 1 for large integers N. Only if $\triangle = \Theta(N)$ is the decreasing of ρ substantial — but bounded, nevertheless. In fact, in the extreme case where the tree T is a snake (see Figure 1) we have $\triangle = N - 2$ and

$$\begin{aligned} EPL(T) &= \frac{N(N+1)}{2} - 1 \\ &= N\frac{1}{2}\triangle + O(N) \\ &= N\left(\log_2 N + \frac{1}{2}\triangle - \log_2 \triangle\right) + O(N) \end{aligned}$$

This indicates that there is a difference between the world of reals where the upper bound of Theorem 3.1 is tight for all values of N and $\triangle$ and the real world of trees — but only a small one! (See the end of Section 3.)

5 Concluding remarks

We have used the relationship between the external path length of a binary tree and the ratio of means of certain integers to derive an upper bound for the path length in terms of the weight N and the thickness of the fringe, $\triangle$, namely

$$EPL(T) \leq N(\log_2 N + \triangle - \log_2 \triangle + O(1))$$

Then we have constructed binary trees that have a high external path length to show that this bound is tight if $\triangle \leq \sqrt{N}$ and reasonably sharp otherwise.

The result obtained here raises a number of interesting problems for further research. Does our result extend to weighted binary trees? To multiway trees? What does a tight upper bound look like in the case $\triangle > \sqrt{N}$? And finally, how much better a bound can be established if more information about the fringe is available?

References

[1] A. Brüggemann-Klein and D. Wood. *Drawing Trees Nicely with TEX.* Technical Report , Department of Computer Science, University of Waterloo, 1987.

[2] G.T. Cargo and O. Shisha. Bounds on the ratios of means. *Journal of Research of the National Bureau of Standards*, 66B:169–170, 1962.

[3] R.W. Hamming. *Coding and Information Theory.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

[4] W. Kaplan. *Advanced Calculus.* Addison-Wesley Publishing Co., Reading, Mass., 1959.

[5] R. Klein and D. Wood. *A Tight Upper Bound for the Path Length of AVL Trees.* Technical Report , Department of Computer Science, University of Waterloo, 1987.

[6] D.E. Knuth. *The Art of Computer Programming, Vol.3: Sorting and Searching.* Addison-Wesley Publishing Co., Reading, Mass., 1973.

[7] W. Specht. Zur Theorie der Elementaren Mittel. *Mathematische Zeitschrift*, 74:91–98, 1960.

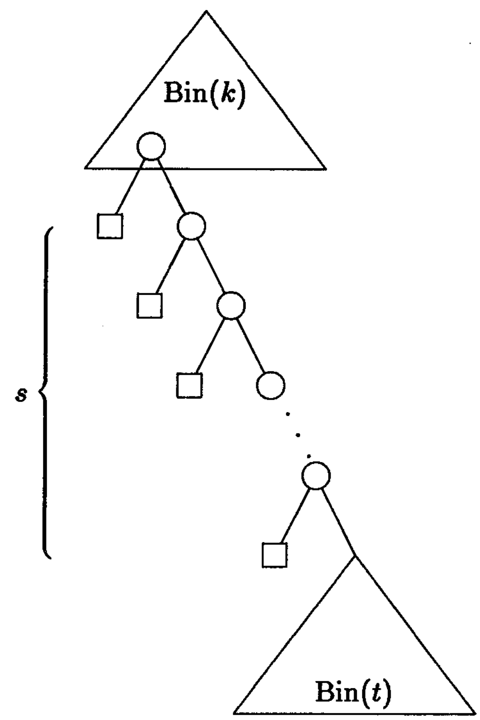


Figure 4: The tree $T_2(k, s, t)$.