DOCUMENT RESUME

ED 095 924 IR 001 111

AUTHOR	Pradels, Jean Louis
TITLE	Two Upper Bounds for the Weighted Path Length of Binary Trees. Report No. UIUCDCS-R-73-565.
INSTITUTION	Illinois Univ., Urbana. Dept. of Computer Science.
SPONS AGENCY	Ministry of Foreign Affairs, Paris (France).
REPORT NO	UIUCDCS-R-73-565
PUB DATE	Jan 73
NOTE	16p.
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	Codification; *Computer Science; Information Retrieval; *Information Science; Information Storage; Information Theory; *Mathematical Models; Statistics
IDENTIFIERS	*Binary Trees

ABSTRACT

Rooted binary trees with weighted nodes are structures encountered in many areas, such as coding theory, searching and sorting, information storage and retrieval. The path length is a meaningful quantity which gives indications about the expected time of a search or the length of a code, for example. In this paper, two sharp bounds for the total path length of general weighted node trees are derived. (Author)

Report No. UIUCDCS-R-73-565

TWO UPPER BOUNDS FOR THE WEIGHTED PATH LENGTH OF BINARY TREES*

by

Jean Louis Pradels

January 1973

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

*This work was supported in part by the Department of Computer Science at the University of Illinois at Urbana-Champaign and by the Ministry of Foreign Affairs of France. It is submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science in the Graduate College of the University of Illinois, June 1971.

ACKNOWLEDGMENT

I wish to express my sincere gratitude and appreciation to Professor Jurg Nievergelt for his helpful suggestions and guidance in this work.

Special thanks are also due to Miss Sue Cook for her time and excellent typing.

Finally, I wish to express my gratitude to the Department of Computer Science of the University of Illinois and to the Ministry of Foreign Affairs of France for their support.

TABLE OF CONTENTS

I. INTRODUCTION

Rooted binary trees with weighted nodes are structures encountered in many areas, such as coding theory, searching and sorting, information storage and retrieval.

A common quantity of great importance is the weighted path length. An estimate of it gives indications about the expected time of a search or the length of a code, for example. The knowledge of lower and upper bounds would permit such estimates.

The noiseless coding theorem in Information Theory provides a lower bound for the weighted root-leaf path length. But, until recently, upper bounds which are the more meaninfgul for many applications, were still lacking. They were first obtained for unweighted trees, introducing the concept of structural balance of a tree [1]. Then upper bounds for various weighted trees have been derived using structural balance and the new concept of weight balance of a tree [2].

In this paper, using a different definition of the weight balance of a tree, we derive two upper bounds for the total path length of general weighted node trees.

The first one introduces two parameters γ and β_T which sharpen the bound but complicate its expression. The parameters γ will be useful every time we have a certain knowledge of the weight distribution. However, if the estimation of γ or β_T requires too much computation, we can take them equal to zero and derive from this upper bound a simpler one.

The second upper bound, using a different normalization, will be useful when the entropy of the weights of the nodes is not known.

II. DEFINITIONS AND NOTATION

A binary tree T_n is a finite set of n nodes which is either empty (if $n = 0$) or else is partitioned into the following three classes: A single node r called the root of T_n, a binary tree T_g on the g nodes 1, .., r-1 called the left subtree of the root and a binary tree T_d on the d nodes r+1, .., n called the right subtree of the root ($r = g+1$, $g + d + 1 = n$).

The subscripts g and d will always refer respectively to the left and right subtrees of T_n. In the above definition they have two meanings. They indicate both subtrees and number of nodes.

A weighted binary tree is a binary tree T_n such that a non-negative real number w_k, called a weight, is assigned to each node k of T_n. We denote a weighted binary tree of n nodes by the (n+1)-tuple $(T_n; w_1, .., w_n)$.

W, W_g, W_d will denote respectively the sum of the weights of T_n, T_g, and T_d.

Weight distribution. We will restrict the weight distribution to the following case: At least one of the two sons of each non-terminal node must have a strictly positive weight.

The weighted path length $|T_n|$ of a tree T_n is defined as the sum over all the nodes of the product of the weight of the node and the level of the node. The weighted path length satisfies the following equalities:

$$|T_1| = 0$$

$$|T_n| = |T_g| + |T_d| + W_g + W_d \quad (n > 1)$$

Weighted root balance $\rho(T_n)$. The two upper bounds depend on a parameter which measures the "balance" of a tree, in the sense of how close the total weights of the left and right subtrees of the root are to each other. The following quantity:

$\rho(T_n) = \frac{1}{2}$ if $n = 1$; otherwise, $\rho(T_n) = \min\left(\frac{W_g}{W-w_r}, \frac{W_d}{W-w_r}\right)$ serves partially this purpose. This definition implies $0 \leq \rho(T_n) \leq \frac{1}{2}$. According to the restriction made previously about the weight distribution, $\rho(T_n) = 0$ whenever T_n has a right or left subtree of one node, the weight of which is equal to zero. $(W_g \cdot W_d = 0)$. We will see later that such a value must be avoided because it would give a bad estimate of the weighted balance of T_n. We notice that the weighted path length $|T_n|$ remains the same if the weights of the two sons of T_n are interchanged. Let $\overline{W}$ denote the positive weight of the two sons of the root of T_n when the condition $W_g \cdot W_d = 0$ arises, then the following definition:

$\rho(T_n) = \frac{1}{2}$ if $n = 1$, otherwise

if $W_g \cdot W_d \neq 0$, then $\rho(T_n) = \min\left(\frac{W_g}{W-w_r}, \frac{W_d}{W-w_r}\right)$

if $W_g \cdot W_d = 0$, then $\rho(T_n) = \min\left(\frac{W_g + W_d - \overline{W}}{W-w_r}, \frac{\overline{W}}{W-w_r}\right)$

makes the weighted root balance strictly positive, except for the weighted binary trees of 3 nodes when one son of the root has a weight equal to zero.

The weighted balance $\beta(T_n)$ is equal to $\frac{1}{2}$ if $n = 1$ or $n = 3$, otherwise $\beta(T_n) = \min[\rho(T_n), \beta(T_g), \beta(T_d)]$. Although the weighted root balance can be equal to zero for trees of three nodes, we notice that the weighted balance $\beta(T_n)$ for $n \geq 1$ is always strictly positive. We deduce from this definition the following inequality.

for $n = 1$ or $n > 3$: $0 < \beta(T_n) \leq \rho(T_n) \leq \frac{1}{2}$.

Terminal weighted balance $\beta_T(T_n)$:

$\beta_T(T_n) = \rho(T_n)$ if $n = 1$ or 3, otherwise:

$\beta_T(T_n) = \min[\beta_T(T_g), \beta_T(T_d)]$

this definition implies that $0 \leq \beta_T(T_n) \leq \frac{1}{2}$. This parameter appears in the first upper bound. Its evaluation is easy and it sharpens the bound significantly. In particular, it makes the bound equal to the weighted path length for trees of one or three nodes. However, in any case, if we don't want to estimate it we can take it equal to zero.

Definition of γ: This parameter appears also in the first upper bound. It can be verified easily that its value does not matter for trees of one or three nodes. Moreover, the bound is equal to the path length for every value of γ. Therefore, the value of γ needs only to be estimated when $n > 3$.

$$\text{if } n = 3 \quad \gamma(T_3) = \delta(T_3)$$

$$\text{otherwise} \quad \gamma(T_n) = \min(\delta(T_n), \gamma(T_g), \gamma(T_d))$$

$$\text{with} \quad \delta(T_n) = (\alpha + 1)\log(\alpha + 1) - \alpha \log \alpha$$

$$\text{and where} \quad \alpha = \frac{W - w_r}{w_r}$$

The assumption made previously about the weight distribtution implies that for $n > 3$, $\gamma(T_n)$ has always a finite value, even if $w_r = 0$.

The expression of the first upper-bound shows that the parameter γ will sharpen strongly the bound if its value is high. However, this parameter will be useful only whenever we have a sufficient knowledge of the weight distribution of T_n, because its estimation implies some computation. Nevertheless, γ as well as β_T, can in any case be taken equal to zero. This corresponds to the fact that if $n \geq 1$, then $W - w_r \geq 0$.

Entropy H(x). We will introduce the following quantity in the two bounds:

$$H(x) = -[x \log x + (1-x)\log(1-x)] \quad \text{for } 0 \leq x \leq 1.$$

L or $L(T_n)$ will denote the sum of the weights of the terminal nodes of T_n.

All the logarithms in this paper are taken to the base two.

III. [illegible]

1. Inequalities Used

Let $\{x_1, x_2, .., x_n\}$ be a set of n non-negative real numbers such that $S = \sum_{i=1}^{n} x_i \geq 1$. Let x_k denote any one of these n numbers and α_k a real positive number such that $S \geq (1 + \alpha_k)x_k$. Then, we have the following inequality:

(1) $(S-x_k) \log (S-x_k) \leq S \log S - x_k \log x_k - \delta x_k$ where

$$\delta = (\alpha_k + 1) \log (\alpha_k + 1) - \alpha_k \log \alpha_k.$$

Proof: Let $f(x) = x \log x$ and a and b be two real numbers such that $1 \leq a$, $0 \leq b$. Then we have:

$$f(1 + b) - f(1) \leq f(a + b) - f(a).$$

Applying this relation with $a = \frac{S - x_k}{\alpha_k x_k}$, $b = \frac{x_k}{\alpha_k x_k}$, we obtain the previous inequality (1).

2. First Upper Bound

Let $(T_n, w_1, .., w_n)$ be a weighted binary tree, then the weighted path length $|T_n|$ satisfies the following inequality:

$$|T_n| \leq \frac{1}{H(B)} \left[(W-w_r) \log(W-w_r) + \left(\sum_{k=1}^{n} w_k \log \frac{1}{w_k} - w_r \log \frac{1}{w_r} \right) - \gamma (W-w_r) \right]$$

$$+ (\gamma + 1 - H(\beta_T))L$$

where β, β_T, γ, L have the definitions given in (II).

Proof: Case (i). The weight distribution verifies the restriction introduced previously and we assume that if $w_k \neq 0$ then $w_k \geq 1$ for all k.

a) For n = 1, $W = w_1$ and $L(T_1) = 0$, the assertion is true. We can also write:

(2) $$|T_1| \leq \frac{1}{H(\frac{1}{2})} [W \log W + w_1 \log \frac{1}{w_1} - \gamma w_1] + (\gamma + 1 - H(\tfrac{1}{2})) w_1$$

This inequality which holds for every value of γ will be useful in the following part of the proof.

b) Assume that the assertion is true for all $i < n$ and let $(T_g; w'_1, .., w'_g)$, $(T_d; w''_1, .., w''_d)$ be the left and right subtrees of T_n. Hence $w'_k = w_k$ for $1 \leq k \leq r-1$ and $w''_k = w_{k+r}$ for $1 \leq k \leq n-r$ $(g = r-1$ and $d = n-r)$. Let β_g, β_{T_g}, γ_g; β_d, β_{T_d}, γ_d be respectively the parameters of T_g and T_d defined in (II). Then, according to the relation $|T_n| = |T_g| + |T_d| + W_g + W_d$ we write:

$$|T_n| \leq \frac{1}{H(\beta_g)} [W_g - w'_r) \log (W_g - w'_r) + \sum_{k=1}^{g} w'_k \log \frac{1}{w'_k} - w'_r \log \frac{1}{w'_r} - \gamma_g (W_g - w'_r)$$

$$+ \frac{1}{H(\beta_d)} [W_d - w''_r) \log (W_d - w''_r) + \sum_{k=1}^{d} w''_k \log \frac{1}{w''_k} - w''_r \log \frac{1}{w''_r} - \gamma_d (W_d - w''_r)]$$

$$+ (\gamma_g + 1 - H(\beta_{T_g})) L_g + (\gamma_d + 1 - H(\beta_{T_d})) L_d + W_g + W_d$$

Using the equality $L = L_g + L_d$ and the three relations:

$\beta(T_n) = \min [\rho(T_n), \beta_g, \beta_d]$, $\beta_T(T_n) = \min[\beta_{T_g}, \beta_{T_d}]$, $\gamma(T_n) = \min[\delta(T_n), \gamma_g, \gamma_d]$

defined in (II), we write:

$$|T_n| \leq \frac{1}{H(\beta)} [W_g - w'_r) \log (W_g - w'_r) + (W_d - w''_r) \log (W_d - w''_r) + \sum_{k=1}^{g} w'_k \log \frac{1}{w'_k}$$
$$+ \sum_{k=1}^{d} w''_k \log \frac{1}{w''_k} - \gamma (W_g - w'_r + W_d - w''_r)] + (\gamma + 1 - H(\beta_T)) L + W_g + W_d.$$

If we assume that T_g and T_d have both more than one node ($g > 1$ and $d > 1$) then $W_g - w'_r \geq 1$ and $W_d - w''_r \geq 1$. We can apply the inequality (1) defined at the beginning of (III). Moreover, using the relation $\gamma = \min[\delta(T_n), \gamma(T_g), \gamma(T_d)]$, we obtain the pair of inequalities:

$$(3)\quad (W_g - w'_r)\log(W_g - w'_r) \le W_g \log W_g + w'_r \log \frac{1}{w'_r} - \gamma w'_r$$

$$(W_d - w''_r)\log(W_d - w''_r) \le W_d \log W_d + w''_r \log \frac{1}{w''_r} - \gamma w''_r$$

We now obtain the following inequality:

$$(4)\quad |T_n| \le \frac{1}{H(\beta)} [W_g \log W_g + W_d \log W_d + \sum_{k=1}^{n} w_k \log \frac{1}{w_k} - w_r \log \frac{1}{w_r}$$

$$- \gamma(W - w_r)] + (\gamma + 1 - H(\beta_T))L + W_g + W_d$$

If, however, T_g and T_d are such that $g = 1$ or $d = 1$, we can't apply inequality (1) because $W_g - w'_r$ or $W_d - w''_r$ is equal to zero. We will use instead the expression (2) derived at the beginning of the proof when $n = 1$. Assume that T_g and T_d are such that $g > 1$ and $d = 1$, then

$$|T_d| \le \frac{1}{H(\beta_d)} [W_d \log W_d + w''_r \log \frac{1}{w''_r} - \gamma_d w''_r] + (\gamma_d + 1 - H(\beta_{T_d}))w''_r$$

where $\beta_d = \beta_{T_d} = \frac{1}{2}$, γ_d having any value, therefore, after similar steps as before, we obtain:

$$|T_n| \le \frac{1}{H(\beta)} [W_g - w'_r)\log(W_g - w'_r) + W_d \log W_d + \sum_{k=1}^{g} w'_k \log \frac{1}{w'_k} - w'_r \log \frac{1}{w'_r}$$

$$+ w''_r \log \frac{1}{w''_r} - \gamma(W_g - w'_r - w''_r)] + (\gamma + 1 - H(\beta_T))L + W_g + W_d$$

If now we apply the first of the pair of inequalities (2) we obtain the inequality (4) already derived.

Hence for every left and right subtree, we have:

$$|T_n| \le \frac{1}{H(\beta)} [W_g \log W_g + W_d \log W_d + \sum_{k=1}^{n} w_k \log \frac{1}{w_k} - w_r \log \frac{1}{w_r} - \gamma(W - w_r)]$$

$$+ (\gamma + 1 - H(\beta_T))L + W_g + W_d.$$

The quantity $W_g \log W_g + W_d \log W_d$ can be expressed in terms of the weighted root balance $\rho(T_n)$

$$W_g \log W_g + W_d \log W_d$$

$$= \rho(W_g + W_d)\,[\log(W_g + W_d) + \log \rho\,] + (1 - \rho)(W_g + W_d)[\log(1-\rho) + \log(W_g + W_d)]$$

$$= (W-w_r) \log (W-w_r) + (W-w_r)[\rho \log \rho + (1 - \rho) \log (1 - \rho)]$$

$\rho \log(\frac{1}{\rho}) + (1 - \rho) \log (\frac{1}{1-\rho})$ is the entropy of the weighted root balance of T_n. Then if we recall that $0 < \beta \leq \rho \leq \frac{1}{2}$, we have $0 < H(\beta) \leq H(\rho) \leq 1$.

$$|T_n| \leq \frac{1}{H(\beta)} [W-w_r) \log(W-w_r) + \sum_{k=1}^{n} w_k \log \frac{1}{w_k} - w_r \log \frac{1}{w_r} - \gamma(W-w_r)]$$

$$+ (\gamma + 1 - H(\beta_T))L + (1 - \frac{H(\rho)}{H(\beta)}) (W-w_r)$$

Hence the induction hypothesis is verified. The weighted path length $|T_n|$ satisfies the following inequality:

$$|T_n| \leq \frac{1}{H(\beta)} [(W-w_r) \log (W-w_r) + \sum_{k=1}^{n} w_k \log \frac{1}{w_k} - w_r \log \frac{1}{w_r} - \gamma(W-w_r)]$$

$$+ (\gamma + 1 - H(\beta_T))L$$

Case (ii): In the general case, let $\overline{w}_k = \frac{w_k}{w_{min}}$ where w_{min} is the minimum of all the positive weights. Then $\overline{w}_k \geq 1$ for all k such that $w_k \neq 0$. The result obtained in case (i), applied to this new weight distribution, gives immediately the desired result.

Remark: Except for particular distributions of the weights like a descending one, for example, it is difficult to obtain an upper bound which is both sharp and simple. If we don't want to introduce the parameters γ and β_T, we can set both of them equal to zero. This leads us to the simpler bound:

$$|T_n| \leq \frac{1}{H(\beta)} [W-w_r) \log(W-w_r) + \sum_{k=1}^{n} w_k \log \frac{1}{w_k} - w_r \log \frac{1}{w_r}] + L$$

3. Second Upper Bound

Although the entropy $\sum_{k=1}^{n} w_k \log \frac{1}{w_k}$ is a natural quantity, it may not always be known. For such cases, the following bound would be useful.

$$|T_n| \leq \frac{1}{H(\beta)} [(W-w_r) \log (\frac{W-w_r}{w_{min}}) - 2 (W-w_r)] + 2L$$

w_{min} is the minimum of all thepositive weights of T_n. The proof is quite similar. We would have the following substitutions in part (III).

Inequality Used:

(1) $(S - x_k) \log (S - x_k) \leq S \log S - 2 x_k$

$S > x_k$, $x_k = 0$ or $x_k \geq 1$

Proof: case (i)

(2) $|T_1| \leq \frac{1}{H(\frac{1}{2})} [W \log W - 2 W_1] + 2 W_1$

(3) $(W_g - w_r') \log (W_g - w_r') \leq W_g \log W_g - 2 w_r'$

$(W_d - w_r'') \log (W_d - w_r'') \leq W_d \log W_d - 2 w_r''$

(4) $|T_n| \leq \frac{1}{H(\beta)} [W_g \log W_g + W_d \log W_d - 2 (W_g + W_d)] + 2 L_g + 2 L_d + W_g + W_d$

case (ii) remains the same.

IV. BIBLIOGRAPHY

[1] Hievergelt, J. and Wong, C. K. (1970) Upper Bounds for the Total Path Length of Binary Trees, IBM Research Report RC-3075.

[2] Wong, C. K. and Yue, P. C. (1971) Upper-Bounds for Various Types of Path Lengths of Binary Trees, IBM Research Report.

BIBLIOGRAPHIC DATA SHEET	1. Report No. UIUCDCS-R-73-565	2.	3. Recipient's Accession No.

4. Title and Subtitle
Two Upper Bounds for the Weighted Path Length of Binary Trees

5. Report Date
January 1973

6.

7. Author(s)
Jean Louis Pradels

8. Performing Organization Rept. No. UIUCDCS-R-73-565

9. Performing Organization Name and Address
University of Illinois at Urbana-Champaign
Department of Computer Science
Urbana, Illinois 61801

10. Project/Task/Work Unit No.

11. Contract/Grant No.

12. Sponsoring Organization Name and Address
University of Illinois at Urbana-Champaign
Department of Computer Science
Urbana, Illinois 61801

13. Type of Report & Period Covered
Master's Thesis

14.

15. Supplementary Notes

16. Abstracts

Rooted binary trees with weighted nodes are structures encountered in many areas, such as coding theory, searching and sorting, information storage and retrieval. The path length is a meaningful quantity which gives indication about the expected time of a search or the length of a code for example. In this paper, two sharp bounds for the total path length of general weighted node trees are derived.

17. Key Words and Document Analysis. 17a. Descriptors

Bound
Path-length
Binary tree
Weight

17b. Identifiers/Open-Ended Terms

17c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Release Unlimited	UNCLASSIFIED	14
	20. Security Class (This Page) UNCLASSIFIED	22. Price ------