NEW RESULTS ON THE COSTS OF HUFFMAN TREES

Xiaoji Wang

Department of Computer Science, The Australian National University,
GPO Box 4, Canberra, ACT 2601, Australia

Abstract.

We determine some explicit expressions for the costs of Huffman trees with several classes of weight sequences.

1. Introduction.

A *binary tree* consists of a *root* and two disjoint subtrees; either of which, or both, could be empty. We will deal with extended binary trees. An *extended binary tree* is obtained from a binary tree by adding square nodes to a binary tree whenever a null subtree was present. Thus, in an extended binary tree, an internal node has two sons and a square node (called a *leaf* of the tree) has no sons.

If we consider the arcs of a binary tree as *directed downward*, then there is a unique directed path from the root to every node. The number of arcs in the path to a node is called the *path length* of the node.

Let $w_1, w_2, \ldots, w_n$ are all positive real numbers, which is called *weight sequence*. Let T denote an extended binary tree with leaves $v_1, v_2, \ldots, v_n$ where v_i is associated with weight w_i $(1 \leq i \leq n)$ and let l_i denote the path length of v_i. We will call $\sum_{i=0}^{n} w_i l_i$ the *weighted path length*. For a given weight sequence, an extended binary tree with the minimal weighted path length is called a *Huffman tree* for the weight sequence and the minimal weighted path length is referred to as the *cost* of the Huffman tree. As we know, Huffman tree can be constructed by the Huffman algorithm described below.

Huffman algorithm: We begin with n nodes whose weights are $w_1, w_2, \ldots, w_n$ respectively. Create a new node which is the father of the two nodes with the smallest weights. Do this recursively for the $n-1$ nodes other than the two sons of the new node. The final single node with weight $w_1 + w_2 + \cdots + w_n$ is the root of a binary tree. This binary tree will be a Huffman tree defined by the weight sequence $w_1, w_2, \ldots, w_n$.

Since Huffman trees do not in general possess explicit expressions for their costs, it is of great value to find the formulations of the costs for the Huffman

trees with special kinds of weight sequences. Some progress has been made by A. C. Tucker [5], F. K. Hwang [2], [3] and M. Sandelius [4]. In this paper, explicit expressions of costs of the Huffman trees for several new classes of weight sequences are obtained.

Let W: $w_1 \leq w_2 \leq \cdots \leq w_n$ $(w_1 > 0)$ denote a weight sequence, $H(W)$ be a Huffman tree defined by weight sequence W, l_i $(1 \leq i \leq n)$ be the path length of the leaf of $H(W)$ associated with the weight w_i, $C(W)$ be the cost of $H(W)$. We call l_i $(1 \leq i \leq n)$ the path length of w_i for convenience. For other terminology and notation, we follow [1].

2. Main results.

Lemma 1. *(Lemma in [2]) $w_i < w_j$ implies that $l_j \leq l_i$ for $1 \leq i < j \leq n$. $w_i = w_j$ implies that $|l_i - l_j| \leq 1$ in a Huffman tree.*

Theorem 2. *If $w_1 + w_2 > w_n$, $n = 2^p + q$, $0 \leq q < 2^p$, then*

$$C(W) = (p+1)\sum_{i=1}^{2q} w_i + p\sum_{i=2q+1}^{n} w_i.$$

Proof. Because $w_1 + w_2 > w_n$, as an immediate consequence of Lemma 1 we have that the path lengths of any two leaves can differ by at most one. Hence, there exist $2s$ leaves of $H(W)$ whose path lengths are all equal to $t+1$, and the path lengths of the other $2^t - s$ leaves are all equal to t and it is obvious that $0 \leq s < 2^t$. Since each integer n has unique expression as $n = 2^p + q$ where $0 \leq q < 2^p$, we have $t = p$ and $s = q$. Since $H(W)$ is a binary tree with the minimal weighted path length among all the binary trees with the weight sequence W, the theorem follows.

Corollary 3. *Let $w_i = x$, $1 \leq i \leq n$, $n = 2^p + q$, $0 \leq q < 2^p$, then*

$$C(W) = nxp + 2qx.$$

Theorem 4. *If $w_1 + w_2 < w_3$, $w_1 + w_2 + w_3 > w_n$, $n - 1 = 2^p + q$, $0 \leq q < 2^p$, then*

$$C(W) = \begin{cases} (p+1)(w_1+w_2) + p\sum_{i=3}^{n} w_i, & q = 0; \\ (p+2)(w_1+w_2) + (p+1)\sum_{i=3}^{2q+1} w_i + p\sum_{i=2q+2}^{n} w_i, & q > 0. \end{cases}$$

Proof. Let $\overline{W}$ be the weight sequence $w_1+w_2, w_3, \cdots, w_n$. By Huffman algorithm, we know that $H(W)$ can be obtained by adding two sons with weight w_1 and w_2 at the leaf with weight w_1+w_2 in $H(\overline{W})$.

If $q=0$, by the proof of Theorem 2, we have that the path length of every leaf is p in $H(\overline{W})$. Thus the path length of w_1 and w_2 are $p+1$ while the others are p in $H(W)$.

If $q>0$, by the proof of Theorem 2, we have that there exist $2q$ leaves with path lengths $p+1$ and the other 2^p-q leaves have their path lengths p in $H(\overline{W})$. Thus by Lemma 1, we have that the path length of w_1+w_2 must be $p+1$ in $H(\overline{W})$. Therefore, w_1 and w_2 have their path lengths $p+2$ and among the $n-2$ other weights, $2q-1$ of them have their path lengths $p+1$ while the others have their path lengths p in $H(W)$.

Because $H(W)$ is the binary tree with the weight sequence W which minimizes the weighted path length, the theorem follows.

Theorem 5. *Suppose that W can be divided into t segments as follows: $w_{i_0} \le w_2 \le \cdots \le w_{i_1} \le w_{i_1+1} \le \cdots \le w_{i_{t-1}} \le w_{i_{t-1}+1} \le \cdots \le w_{i_t}$, where $w_{i_0}=1, w_{i_t}=n$. For $0 \le k \le t-1$, let*

$$i_{k+1}-i_k+1 = 2^{p_k}+q_k, 0 \le q_k < 2^{p_k},$$

$$s_k = \sum_{f=1}^{i_k} w_f,$$

$$s_k + w_{i_k+1} > w_{i_{k+1}},$$

$$\delta_k = \begin{cases} 0, & q_k = 0, \\ 1, & q_k > 0, \end{cases}$$

and for $1 \le k \le t-1$, $s_k < w_{i_k+1}$. Then $C(W)$ is

$$\sum_{i=0}^{t-1}(p_i+\delta_i)w_1 + \sum_{k=0}^{t-1}\Big(\sum_{i=k}^{t-1}(p_i+\delta_i)\sum_{j=i_k+1}^{i_k+2q_k-1} w_j + (\sum_{i=k}^{t-1} p_i + \sum_{i=k+1}^{t-1}\delta_i)\sum_{j=i_k+2q_k}^{i_{k+1}} w_j\Big).$$

Proof. Let W_k be the weight sequence $s_k, w_{i_k+1}, \ldots, w_{i_{k+1}}$. By Lemma 1 and the proof of Theorem 2, we have that s_k has its path length $p_k+\delta_k$ for $(1 \le k \le t-1)$ in $H(W_k)$. By Huffman algorithm, $H(W)$ can be obtained by connecting the root of $H(W_{k-1})$ at the leaf s_k of $H(W_k)$ for $1 \le k \le t-1$. By applying the proof of Theorem 2 and analyzing the path length of each leaf, the theorem follows.

Theorem 6. *Suppose that W can be divided into t segments as in Theorem 5. For $0 \le k \le t-1$, let*

$$i_{k+1} - i_k = 2^{p_k} + q_k, 0 \le q_k < 2^{p_k},$$

$$s_k = \sum_{f=1}^{i_k} w_f,$$

$$s_k + w_{i_k+1} < w_{i_k+2},$$

$$s_k + w_{i_k+1} + w_{i_k+2} > w_{i_{k+1}},$$

$$\sigma_k = \begin{cases} 0, & q_k > 0, \\ 1, & q_k = 0, \end{cases}$$

$$r_k = i_k + 2q_k,$$

and for $1 \le k \le t-1$, $s_k - w_{i_k} < w_{i_k+1}$. Then $C(W)$ is

$$\sum_{i=0}^{t-1}(p_i + 2 - \sigma_i)w_1 + \sum_{k=0}^{t-1}\Big(\sum_{i=k}^{t-1}(p_i + 2 - \sigma_i)\sum_{j=i_k+1}^{i_{k+1}} w_j - \sum_{j=i_k+2}^{r_k} w_j + (\sigma_k - 2)\sum_{j=r_k+1}^{i_{k+1}} w_j\Big).$$

Proof. Let W_k be the weight sequence $s_k, w_{i_k+1}, \ldots, w_{i_{k+1}}$, then the path lengths of leaves s_k and w_{i_k+1} are $p_k + 2 - \sigma_k$ for $0 \le k \le t-1$ in $H(W_k)$. By the proof of Theorem 4 and analyzing the path lengths of the leaves and simple calculating, the theorem follows.

As metioned before, we can obtain $H(W)$ by executing the Huffman algorithm on W. We start with weight sequence W of n weights. Just before the ith step ($i > 1$), we have a weight sequence of $n - i + 1$ weights which is obtained from the weight sequence with $n - i + 2$ weights by combining the two smallest weights into one (their sum is the new weight). Let $W(i)$ be the weight sequence just before the ith step and $|W(i)|$ the cardinality of $W(i)$. If s is the smallest number such that the sum of the two smallest weights exceeds the largest weight in $W(s)$, then, for any m ($1 \le m \le n-1$), define

$$|W(s)| = 2^p + q, 0 \le q < 2^p,$$

$$u = \min(m, 2^p - q),$$

$$r = \Big\lfloor \log_2\Big((n-m)\Big/\big(1 - u/2^p - (m-u)/2^{p+1}\big)\Big)\Big\rfloor,$$

and

$$g = 2^{r+1}\big(1 - u/2^p - (m-u)/2^{p+1}\big) - n + m,$$

then we refer to (p, q, u, r, g) as (W, n, m)-critical parameters.

Corollary 3 shows the cost of a Huffman tree in which all the weights are the same. If W can be divided into two segments such that the weights in each segment are the same, an explicit expression for the cost of $H(W)$ was obtained in [2]. For the case where W can be divided into three such segments, we have the following results.

Theorem 7. *Let W be*

$$\overbrace{x, \ldots, x}^{a}, \overbrace{y, \ldots, y}^{b}, \overbrace{z, \ldots, z}^{c}.$$

If $2x > y$, $x < y < z$, $v = \min(a + b - g, a)$, a, b, c are positive integers and (p, q, u, r, g) are $(W, a + b + c, c)$-critical parameters, then

$$C(W) = (ar + v)x + \big(b(r + 1) + a - g - v\big)y + \big(up + (c - u)(p + 1)\big)z.$$

Proof. As in [2], we can prove that there exist u leaves associated with weight z with path length p and the other $c - u$ leaves associated with weight z have all their path lengthes $p+1$. By Lemma 1, we know that the path length of y is less or equal to that of x. As mentioned above, $W\big(\lfloor a/2 \rfloor + 1\big)$ denotes the weight sequence after $\lfloor a/2 \rfloor$ combinations by Huffman algorithm. Since $2x > y$, the path length of $2x$ is less or equal to that of y in $H\big(W(\lfloor a/2 \rfloor + 1)\big)$. By Huffman algorithm, we know that we can obtain the $H(W)$ by choosing some $\lfloor a/2 \rfloor$ leaves with weights $2x$ in $H\big(W(\lfloor a/2 \rfloor + 1)\big)$ and making each of them as the father of two leaves with the same weight x. Therefore, there exists an integer r such that the path length of any x or y is equal to either r or $r + 1$ in $H(W)$. Let g be the number of x and y which have their path lengths r, then the $a + b - g$ other x or y have their path lengths $r + 1$. We can prove that r and g are $(W, a + b + c, c)$-critical parameters in a similar manner to [2]. Let t be the number of x which has path length r, hence, the $a - t$ other x have path lengths $r + 1$ and the number of y which has path length r is $g - t$ and the number of y which has path length $r + 1$ is $b - (g - t) = b + t - g$. Thus, $C(W)$ is

$$\begin{aligned} &trx + (g - t)ry + (a - t)(r + 1)x + (b + t - g)(r + 1)y + \big(up + (c - u)(p + 1)\big)z \\ &\quad = a(r + 1)x + b(r + 1)y - gy + t(y - x) + \big(up + (c - u)(p + 1)\big)z. \end{aligned}$$

Because the minimal possible value of t is $a - v$ and $H(W)$ is the binary tree with minimal weighted path length, we have

$$C(W) = a(r + 1)x + b(r + 1)y - gy + (a - v)(y - x) + \big(up + (c - u)(p + 1)\big)z$$

$$= (ar + v)x + \big(b(r+1) + a - g - v\big)y + \big(up + (c-u)(p+1)\big)z.$$

Theorem 8. *Let W be*

$$x, \overbrace{y, \ldots, y}^{b}, \overbrace{z, \ldots, z}^{c}.$$

If $x < y < z$, b, c are positive integers, (p, q, u, r, g) are $(W, b + c + 1, c)$-critical parameters, and

$$\sigma = \begin{cases} 1, & b + 1 - g > 0, \\ 0, & b + 1 - g = 0. \end{cases}$$

Then

$$C(W) = \big(up + (c-u)(p+1)\big)z + \big(x + (b-g)y\big)(r + \sigma) + gry.$$

Proof. When we execute the Huffman algorithm on W, the first step is combining the x and a y. Thus, x and the y have equal path length. By Lemma 1, $|l_x - l_y| \leq 1$ for any y. We can prove the Theorem by the similar method used in [2].

Theorem 9. *(1) Let W and W' be as follows:*

$$W : \overbrace{x, \ldots, x}^{2a_1}, \overbrace{y, \ldots, y}^{b_1}, \overbrace{z, \ldots, z}^{c_1},$$

$$W' : \overbrace{y, \ldots, y}^{a_1 + b_1}, \overbrace{z, \ldots, z}^{c_1}.$$

If $y = 2x$, $y < z$, a_1, b_1, c_1 are positive integers, and $(p_1, q_1, u_1, r_1, g_1)$ are $(W', a_1 + b_1 + c_1, c_1)$-critical parameters, then

$$C(W) = \Big(u_1 p_1 + (c_1 - u_1)(p_1 + 1)\Big)z + \Big(a_1(r_1 + 2) + b_1(r_1 + 1) - g_1\Big)y.$$

(2) Let $\overline{W}$ and $\overline{W'}$ be as follows:

$$\overline{W} : \overbrace{x, \ldots, x}^{2a_2 + 1}, \overbrace{y, \ldots, y}^{b_2}, \overbrace{z, \ldots, z}^{c_2},$$

$$\overline{W'} : x, \overbrace{y, \cdots, y}^{a_2 + b_2}, \overbrace{z, \ldots, z}^{c_2}.$$

If $y = 2x$, $y < z$, a_2, b_2, c_2 are positive integers, $(p_2, q_2, u_2, r_2, g_2)$ are $(\overline{W'}, a_2 + b_2 + c_2 + 1, c_2)$-critical parameters, and

$$\sigma = \begin{cases} 1, & a_2 + b_2 + 1 - g_2 > 0, \\ 0, & a_2 + b_2 + 1 - g_2 = 0. \end{cases}$$

Then

$$C(\overline{W}) = (u_2p_2 + (c_2 - u_2)(p_2 + 1))z + (x + (b_2 - g_2)y)(r_2 + \sigma) + g_2r_2y + a_2y.$$

Proof. (1). By the result in [2], we know that the cost of $H(W')$ is

$$(u_1p_1 + (c_1 - u_1)(p_1 + 1))z + g_1r_1y + (a_1 + b_1 - g_1)(r_1 + 1)y.$$

When we execute the Huffman algorithm on W, we can get W' after a_1 combinations, and these combinations produce a_1 internal nodes with weight $2x = y$. As we know, the cost of a Huffman tree is equal to the sum of its internal nodes. The a_1 internal nodes and the internal nodes of $H(W')$ are all the internal nodes in $H(W)$. Therefore, $C(W) = a_1y + C(W')$, and (1) follows.

(2). By the result of Theorem 8 and similar method used in the proof of (1), we can get (2).

3. Remark.

For monotonically increasing weight sequence, Huffman algorithm for Huffman tree coincides with the T–C algorithm for optimal alphabetic binary tree. Thus, all above results are also valid for optimal alphabetic binary trees.

4. Acknowledgement.

I wish to thank Brendan McKay for his valuable comments.

References.

[1] T. C. Hu, Combinatorial algorithms, *Addison-Wesley, 1982.*

[2] F. K. Hwang, An explicit expression for the cost of a class of Huffman trees, *Discrete Math.* **32** (1980) 163–165.

[3] F. K. Hwang, On finding a single defective in binomial group testing, *J. Amer. Statist. Assoc.* **69** (1974) 1091–1101.

[4] M. Sandelius, On an optimal search procedure, *Amer. Math. Monthly,* **68**(1961) 133–134.

[5] A. C. Tucker, The cost of a class of optimal binary tree, *J. Combinatorial Theory, Ser. A* **16** (1974) 259–263.