Huffman Codes and Self-Information

GYULA O. H. KATONA AND TIBOR O. H. NEMETZ

***Abstract*—In this paper the connection between the self-information of a source letter from a finite alphabet and its codeword length in a Huffman code is investigated. Consider the set of all independent finite alphabet sources which contain a source letter *a* of probability *p*. The maximum over this set of the length of a Huffman codeword for *a* is determined. This maximum remains constant as *p* varies between the reciprocal values of two consecutive Fibonacci numbers. For the small *p* this maximum is approximately equal to**

$$\left[\log_2 \frac{1+\sqrt{5}}{2}\right]^{-1} \approx 1.44$$

times the self-information.

I. Introduction

SUPPOSE that a discrete memoryless source U has a K letter alphabet $a_1, a_2, \cdots, a_K$ with probability distribution $\mathcal{P} = \{P(a_1), P(a_2), \cdots, P(a_K)\}$. We shall consider binary variable length codes satisfying the prefix con-

Manuscript received February 17, 1975; revised September 18, 1975.

The authors are with the Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

dition. For concepts and notation not defined herein, refer to Gallager [1, pp. 43–55]. The self-information and the codeword length of a letter a_i will be denoted, respectively, by $I(a_i) = -\log_2 P(a_i)$ and $n_i = n(a_i)$. For the entropy of the source we use the notation $H(\mathcal{P})$.

The major significance of the entropy comes from the source coding theorem, see [1]. Roughly speaking, this states that, if $H(\mathcal{P})$ is the entropy of a discrete memoryless source, a sequence of source letters cannot be encoded using fewer than $H(\mathcal{P})$ binary digits per source letter on the average, but can be encoded by using an average number of binary digits per source letter as close as desired to $H(\mathcal{P})$. Therefore, the entropy has an operational meaning expressed by

$$H(\mathcal{P}) \approx \bar{n} = \sum_{i=1}^{K} P(a_i) \cdot n_i$$

for optimal codes.

Although the entropy, or average self-information, has operational significance, it is hard to attribute comparable significance to the self-information of an individual source letter. One might expect in a certain sense

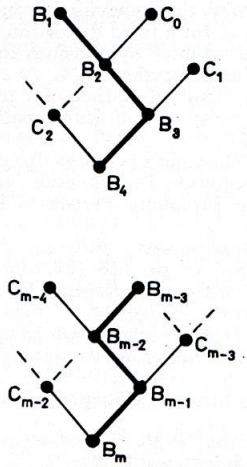

Fig. 1. Code tree used in Lemmas 1 and 2.

that for optimal codes

$$I(a_i) \approx n(a_i).$$

Considering the source with probability distribution $p_1 = P(a_1) = 1 - \epsilon$ and $P_2 = P(a_2) = \epsilon$, we have $n(a_1) = n(a_2) = 1$ for the only optimal code, showing that the ratio of $I(a_i)$ to $n(a_i)$ may be arbitrarily large. On the other hand, however, it is possible to derive a least upper bound on $n(a_i)$ in terms of $I(a_i)$. It turns out that this endeavor involves the Fibonacci numbers and that the least upper bound is a constant times the self-information for small $P(a_i)$.

We also determine the least upper bound to the difference $n(a_i) - n(a_j)$ for Huffman codes when $P(a_i)/P(a_j) \geq 1$ is fixed.

II. Analysis and Results

Let us consider a discrete memoryless source, one letter of which, say a, has probability $p = P(a)$. Fig. 1 shows the path leading from the terminal node B_1 assigned to this letter a to the root B_m of the code tree of a given binary Huffman code (cf. [1, pp. 52–55]). The intermediate nodes of this path are denoted by $B_2, B_3, \cdots, B_{m-1}$. The node which is connected with B_i but not on this path is denoted by C_{i-2}, $i = 2, 3, \cdots, m$.

Each terminal node of the code tree corresponds to a source letter. We assign to the terminal nodes the probability $P(a_i)$ of the corresponding letter a_i. The probability assigned to other nodes N of the tree is the sum of the probabilities of all terminal nodes which are connected with the root through N. The probability assigned to B_i (respectively, C_i) is denoted by b_i (respectively, c_i). In this notation $b_1 = p$ and $b_m = 1$.

Obviously, $n(a) = m - 1$. Therefore, our aim is to derive a bound on m in terms of p. Toward this end, we establish the following two lemmas which concern inequalities between the probabilities assigned to the nodes of Fig. 1.

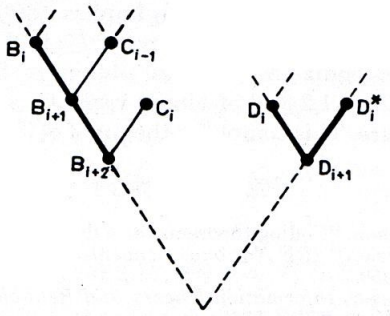

Fig. 2. Detail of code tree used in proof of Lemma 2.

Lemma 1:

$$c_i \geq b_i, \qquad i = 1, \cdots, m-2, \tag{1}$$

$$b_i > f_i p, \qquad i = 2, \cdots, m, \tag{2}$$

and

$$c_{i-1} \geq f_{i-1} p, \qquad i = 2, \cdots, m-1, \tag{3}$$

where f_i if defined by the recursion $f_{i+1} = f_i + f_{i-1}$, $f_1 = f_2 = 1$, for $i \geq 2$. That is, f_i is the ith Fibonacci number.

Proof: The essential part of our lemma is (1); (2) and (3) follow by an easy induction over i as we shall now show. For $i = 2$,

$$b_2 = b_1 + c_0 > b_1 = p = f_1 p,$$

while $c_1 \geq p$ follows from (1). According to the induction procedure we assume, for $i > 2$, that $b_{i-1} > f_{i-1} p$

and $c_{i-2} \geqslant f_{i-2}\, p$. Since $b_i = b_{i-1} + c_{i-2} > f_{i-1}p + f_{i-2}p = f_i p$, (2) is proved for all i. Equation (3) follows from (1) because $c_{i-1} \geqslant b_{i-1} > f_{i-1}p$.

Let us now prove (1). Consider the instant in the Huffman procedure when B_{i+1} is introduced as a new node. At this instant, b_i and c_{i-1} are the two smallest probabilities assigned to nodes which had been introduced earlier. Since C_i will take part in the construction of a future node,

$$c_i \geqslant \max(b_i, c_{i-1}) \geqslant b_i$$

which proves (1).

In the following lemma we shall use the assumptions and the notation of the previous lemma. D_i denotes an arbitrary node of the tree with code length $m - i - 1$, $i = 1, \cdots, m-2$. This is the same code length as for C_i. Its probability is denoted by d_i.

Lemma 2:

$$d_i \geqslant b_i, \qquad i = 1, 2, \cdots, m-2, \tag{4}$$

and

$$d_i > f_i p, \qquad i = 2, \cdots, m-2. \tag{5}$$

Proof: (5) follows from (4) and (2), so we only have to prove (4). Denoting by r the distance of D_i from the nearest B_j, we use induction over r. The case $r = 0$ is trivial. If $r = 1$, $D_i = C_i$, and (1) gives the desired in-

equality. Suppose $r > 1$ and (4) is proved for the smaller r. Let D_{i+1} be the first node along the path leading from D_i to the root and let D_i^* be the other node of code length $m - i - 1$ and connected with D_{i+1} as in Fig. 2.

We continue the Huffman procedure until either B_{i+1} or D_{i+1} is formed as a new node. We have to distinguish between two cases. 1) B_{i+1} is introduced earlier in the code tree than D_{i+1}. This means that $d_i \geq \max(b_i, c_{i-1}) \geq b_i$. 2) The node D_{i+1} has been formed before B_{i+1}. Then,

$$\max(d_i, d_i^*) \leq \min(b_i, c_{i-1}) \tag{6}$$

and it follows that

$$d_{i+1} = d_i + d_i^* \leq b_i + c_{i-1} = b_{i+1}. \tag{7}$$

The equality can hold in (7) if and only if

$$d_i = d_i^* = b_i = c_{i-1}. \tag{8}$$

However, the inductional assumption implies that

$$d_{i+1} \geq b_{i+1}$$

so (7) holds with equality. Thus (8) must be true, so (4) is satisfied with equality and the lemma is proved.

Theorem 1: If the probability p of a source letter a satisfies

$$\frac{1}{f_{s+1}} \leq p < \frac{1}{f_s}, \qquad 2 \leq s, \tag{9}$$

where f_s is the sth Fibonacci number, then the code length $n(a)$ of a in a binary Huffman code satisfies

$$n(a) \leq s - 1 \tag{10}$$

and this is the best possible bound.

Proof: 1) First we prove the inequality (10) using the left side of (9). Suppose $n(a) = s - 1 \geqslant 1$. Then, by (2), we have

$$1 = b_m > f_m p, \qquad \text{if } m \geqslant 2.$$

Comparing with the left side of (9), we obtain

$$\frac{1}{f_{s+1}} \leq p < \frac{1}{f_m}$$

which implies $m < s + 1$, i.e., $m - 1 \leq s - 1$, and (10) is proved.

2) Let us now construct a probability distribution for a source with $P(a) = p$ satisfying (9) in which the code length of a can be $s - 1$. Let

$$p_1 = f_{s-2}/f_s,\ p_2 = f_{s-3}/f_s,\ \cdots,\ p_{s-2} = f_1/f_s,$$

$$p_{s-1} = \frac{1}{f_s} - p,\ p_s = p.$$

By (9), $p_{s-1} > 0$ and the two smallest probabilities are p_{s-1} and p_s. Also, $p_{s-1} + p_s = 1/f_s$.

Continuing the Huffman procedure, we can choose p_{s-2} and $p_s + p_{s-1}$ as the two smallest probabilities.

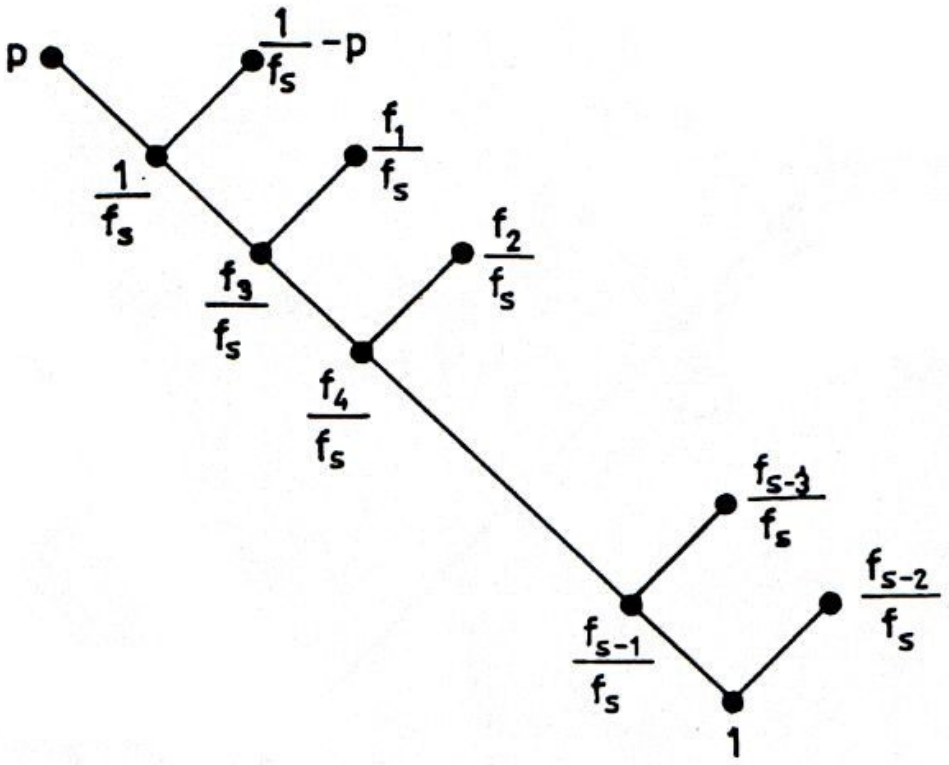

Fig. 3. Code tree for Huffman code used in proof of Theorem 1.

Since

$$p_{s-2} + p_{s-1} + p_s = \frac{f_1 + 1}{f_s} = \frac{f_3}{f_s} > \frac{f_2}{f_s},$$

in the following step of the procedure one can consider p_{s-3} and $p_{s-2} + p_{s-1} + p_s$ as the smallest probabilities and so on. In this way we get that one of the possible Huffman codes which is illustrated in Fig. 3. Obviously, the codeword length is $s - 1$ for this Huffman code.

Corollary:

$$\limsup_{P(a)\to 0} \frac{n(a)}{I(a)} = \left[\log_2 \frac{1+\sqrt{5}}{2}\right]^{-1} \approx 1.44042.$$

Proof: It is well known that

$$f_i = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^i - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^i, \qquad i = 1, 2, \cdots.$$

Substituting this into Theorem 1 and taking the limit, the corollary follows.

Theorem 2: If the probabilities p and q of the source letters a and b, respectively, satisfy

$$\frac{1}{f_{s+1}} \leq \frac{p}{q} < \frac{1}{f_s}, \qquad s \geq 2, \tag{11}$$

then the difference of the code lengths of a and b in a Huffman code is at most s, i.e.,

$$n(a) - n(b) \leq s \tag{12}$$

and this is the best possible bound.

Proof: Again the structure of Fig. 2 is used with a associated with node B_1. b is associated with D_i, for $i \geq 2$, such that $d_i = q$. Since $n(a) = m - 1$ and $n(b) = m - i - 1$, the difference of the code lengths of a and b is i. From (5) and (11),

$$\frac{1}{f_{s+1}} < \frac{1}{f_i}$$

which implies $i \leq s$ and thereby proves (12).

The probability distribution which gives equality in

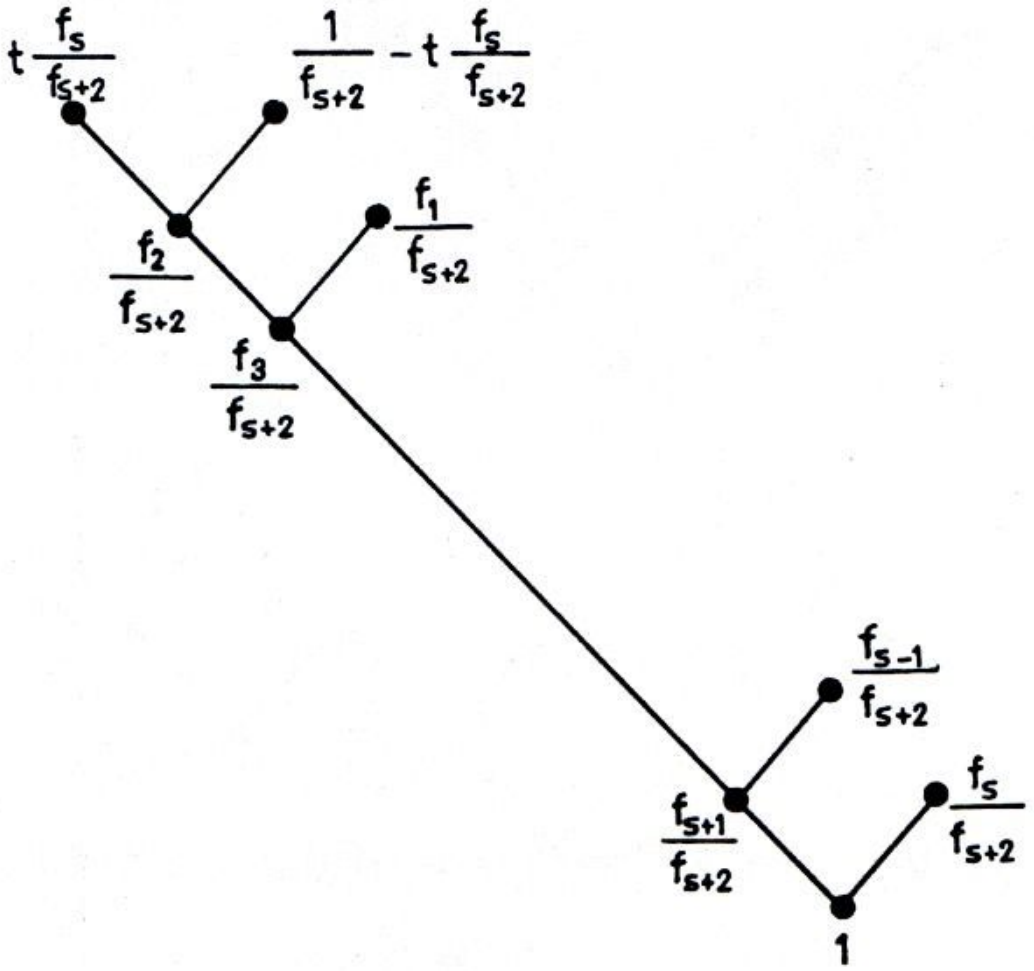

Fig. 4. Variant of Huffman code used in proof of Theorem 2.

(12) is the following:

$$p_1 = p = t\frac{f_s}{f_{s+2}}, p_2 = \frac{1}{f_{s+2}} - t\frac{f_s}{f_{s+2}}, p_3 = \frac{f_1}{f_{s+2}},$$

$$p_4 = \frac{f_2}{f_{s+2}}, \cdots, p_{s+1} = \frac{f_{s-1}}{f_{s+2}}, p_{s+2} = q = \frac{f_s}{f_{s+2}},$$

where $t = p/q$. It is easy to see that $p_i > 0$, $1 \leq i \leq s + 2$. A variant of the Huffman codes of this distribution is illustrated in Fig. 4 where the distance of the codes corresponding to p and q is exactly s. This completes the proof.

Remarks: 1) The difference of the code lengths of a and b in Theorem 2 cannot be negative, because in an optimal code a letter of smaller probability cannot be assigned a shorter codeword. However, the difference can be zero for any given $t = p/q$ as demonstrated by the probability distribution $1/(1 + t)$, $t/(1 + t)$.

2) Theorem 2 does not say anything about the case $p = q$. It is easy to see that in this case the best possible bound for the code length difference is one. The distribution {1/3, 1/3, 1/3} shows that equality can hold.

3) The connection between the self-information and the codeword length of optimal codes is dealt with from another point of view in [2].

References

[1] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[2] T. Nemetz and J. Simon, "Self-information and optimal codes," to appear in the *Proceedings of the Colloquium on Information Theory*, Keszthely, Hungary, August 1975.