Acta Mathematica Academiae Scientiarum Hungaricae
Tomus 23 (3—4), (1972), pp. 439—442.

SOME REMARKS ON THE CONSTRUCTION OF OPTIMAL CODES

By

G. O. H. KATONA (Budapest) and M. A. LEE (El Paso)

Let X be a random variable taking on values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ respectively. Suppose C is a decodable binary code which assigns to x_i a word of length l_i. There is a well-known procedure due to Huffman for constructing such a code C which minimizes

$$\bar{l}_C = \sum_{i=1}^{n} p_i l_i ,$$

the average code word length. The procedure, however, does not lend itself to an explicit formulation of such a code even in some very special cases, for example the case where $p_1 = p_2 = \ldots = p_n = 1/n$. It is the object of this paper to prove a few theorems which do provide such a formula under suitable, but rather simple, restrictions on the probabilities p_i. Theorems 1 and 2 deal with general binary codes and Theorem 3 concerns itself with alphabetic codes.

THEOREM 1. *Suppose that* $p_1 \geqq p_2 \geqq \ldots \geqq p_n$ *and that*

$$p_n + kp_{n-1} > p_1$$

for some positive integer k. *Then for an optimal instantaneous code*

$$l_n - l_1 \leqq k ,$$

i.e. the number of different code word lengths is at most $k+1$.

PROOF. Let C be an optimal instantaneous code and suppose that $l_n - l_1 > k$. Since C is optimal there exist two code words of length l_n which agree in all positions except the last. We may further assume that these are the words associated with x_{n-1} and x_n. Thus suppose the code words of these two variables are $a0$ and $a1$ where a is some sequence of 0's and 1's. Let b be the code word of x_1.

We construct now a new code C' as follows. For $i \neq 1, n-1, n$ let the code word associated with x_i in C' be the same as in C. For x_1, x_{n-1}, x_n let the assignment be as follows.

$$x_1 \to b0, \quad x_{n-1} \to b1, \quad x_n \to a .$$

One now has

$$\bar{l}_C - \bar{l}_{C'} = p_1 l_1 + p_{n-1} l_n + p_n l_n - p_1(l_1 + 1) - p_{n-1}(l_1 + 1) - p_n(l_n - 1) =$$
$$= - p_1 + p_n + p_{n-1}(l_n - l_1 - 1).$$

We have assumed that $l_n - l_1 > k$. Therefore

$$\bar{l}_C - \bar{l}_{C'} \geqq -p_1 + p_n + kp_{n-1}.$$

Since $p_n + kp_{n-1} > p_1$ the right hand side of this inequality is positive contradicting the optimality of C. This finishes the proof.

THEOREM 2. *Suppose* $p_1 \geqq p_2 \geqq \ldots \geqq p_n$ *and that*

$$p_n + p_{n-1} > p_1.$$

Then for an optimal code C

$$\bar{l}_C = \{\log n\} - \sum_{i=1}^{s} p_i$$

where $\{x\}$ *denotes the least integer greater than or equal to* x *and* $s = 2^{\{\log n\}} - n$.

PROOF. Let l be the maximum of the code word lengths for C. We may assume that C is instantaneous. Thus by Theorem 1 $l - 1$ is the only other possible code word length. Let s $(0 \leqq s < n)$ be the number of words of length $l - 1$.

If $a0$ is any code word of length l then $a1$ must also appear as a code word; otherwise we could decrease the average code word length by replacing $a0$ by a. On the other hand if b is an arbitrary sequence of 0's and 1's of length $l - 1$ then either b is a code word of C or both $b0$ and $b1$ are code words of C. For example, if $b0$ is a code word then by the preceding argument $b1$ must also be a code word. If neither $b0$ nor $b1$ is a code word then b must be a code word for otherwise one could decrease the average code word length by reassigning to some variable with code word length l the word b. It follows that

$$s + \frac{n-s}{2} = 2^{l-1} \quad \text{or} \quad n + s = 2^l.$$

On the other hand $2^{l-1} < n \leqq 2^l$. Thus $l = \{\log n\}$. One now has for the average code word length

$$\bar{l}_C = \sum_{i=1}^{n} p_i l_i = \left(\sum_{i=1}^{s} p_i\right)(l-1) + \left(\sum_{i=s+1}^{n} p_i\right) l = l - \sum_{i=1}^{s} p_i = \{\log n\} - \sum_{i=1}^{s} p_i.$$

This completes the proof of the theorem.

REMARK 1. We point out that the proof of Theorem 2 actually provides a method of constructing explicitly the optimal code.

REMARK 2. The results for the special case $p_1 = p_2 = \ldots = p_n = 1/n$ were discovered independently by SANDELIUS [1] and SOBEL [2]. Other special cases of Theorem 2 are also discussed by SOBEL [2].

A code is *alphabetic* if the words assigned to $x_1, x_2, \ldots, x_n$ are in lexicographic order. In contrast to the general situation where one has the Huffman algorithm there are no efficient algorithms for determining optimal alphabetic codes. There are algorithms leading to good alphabetic codes ([3], [4], [5]).

Under certain assumptions on the probabilities one can obtain results analogous to those of Theorem 2. In particular we prove the following

THEOREM 3. *Suppose*

$$p_i + p_{i-1} > \max_{1 \leqq j \leqq n} \{p_j\} \quad \textit{for} \quad i = 1, 2, \ldots, n-1 .$$

Then in an optimal alphabetic code there can be at most two code word lengths, $\{\log n\}$ *and* $\{\log n\} - 1$. *The number of words of length* $\{\log n\} - 1$ *is* $s = 2^{\{\log n\}} - n$ *and they are associated with the variables* $x_{i_1}, x_{i_2}, \ldots, x_{i_s}$ *where* $i_1 < i_2 < \ldots < i_s$ *are chosen such that* $i_{k+1} - i_k$ *is odd* $(1 \leqq k < s)$ *and* $\sum_{k=1}^{s} p_{i_k}$ *is maximal with respect to this restriction on the indices.*

PROOF. Let C be an optimal instantaneous alphabetic code and let l be the maximum of the code word lengths. It is easy to see that the words of length l must occur in pairs of the form $a0, a1$. On the other hand since the code is alphabetic these pairs must correspond to consecutive variables x_i, x_{i+1}.

Now suppose that there exist more than two code word lengths. Then there exist indices i, j such that

$$l_j < l_i - 1$$

and $l_i = l$. Let I be the set of indices such that $l_i = l_{i+1} = l$ and let J be the set of indices j which satisfy the above inequality for some $i \in I$. Choose $i \in I$ and $j \in J$ such that $\min \{|i-j|, |i+1-j|\}$ is minimal. Then $l_k = l - 1$ for $k = i+2, i+3, \ldots, j-1$ if $i < j$ (and for $k = j+1, j+2, \ldots, i-1$ if $i > j$). We construct now a new alphabetic code C' as follows. Let $a0, a1, b_{i+2}, b_{i+3}, \ldots, b_{j-1}, c$ be the code words of C assigned to $x_i, x_{i+1}, \ldots, x_j$ respectively if $i < j$ ($c, b_{j+1}, b_{j+2}, \ldots, b_{i-1}, a0, a1$ the words assigned to $x_j, x_{j+1}, \ldots, x_{i+1}$ if $i > j$). The code words of C' assigned to $x_i, x_{i+1}, \ldots, x_j$ will be $a, b_{i+2}, b_{i+3}, \ldots, b_{j-1}, c0, c1$ in that order if $i < j$ ($c0, c1, b_{j+1}, \ldots, b_{i-1}, a$ if $i > j$). The words of C' assigned to the other x_k's are to be the same as the words of C. It is easy to check that C' is instantaneous and alphabetic. One has

$$\bar{l}_C - \bar{l}_{C'} = p_i + p_{i+1} + p_{j-1}\{l - 1 - (l_j + 1)\} - p_j .$$

By assumption $p_i + p_{i+1} - p_j > 0$. On the other hand $l - 1 > l_j$ and hence $l - 1 - (l_j + 1) > 0$. It follows then that $\bar{l}_C - \bar{l}_{C'}$ is positive which contradicts the optimality of C. This establishes the fact that there are at most two code word lengths.

Arguing now as in the proof of Theorem 2 one can show that these two code word lengths must be $\{\log n\}$ and $\{\log n\}-1$ and that there are exactly $s=2^{\{\log n\}}-n$ words of length $\{\log n\}-1$. It follows that $\bar{l}_C$ is given by

$$\bar{l}_C = \{\log n\} - \sum_{k=1}^{s} p_{i_k}$$

where $i_1 < i_2 < \ldots < i_s$ are the indices such that the word of x_{i_k} is of length $\{\log n\} - 1$. Since the words of length $\{\log n\}$ occur in consecutive pairs it is clear that $i_{k+1} - i_k$ is odd $(1 \leqq k < s)$. On the other hand for $\bar{l}_C$ to be a minimum $\sum_{k=1}^{s} p_{i_k}$ must be a maximum. This completes the proof of the theorem.

NOTE. Remark 1 applies also to Theorem 3.

(Received 16 February 1971)

MTA MATEMATIKAI KUTATÓ INTÉZETE
BUDAPEST, V., REÁLTANODA U. 13—15

THE UNIVERSITY OF TEXAS AT EL PASO
DEPARTMENT OF MATHEMATICS
EL PASO, TEXAS 79968
U.S.A.

References

[1] M. SANDELIUS, On an optimal search procedure, *Amer. Math. Monthly*, **68** (1961), pp. 138—154.

[2] M. SOREL, Binomial and Hypergeometric Group-Testing, *Studia Sci. Math. Hungar.*, **3** (1968), pp. 19—42.

[3] E. N. GILBERT and E. F. MOORE, Variable-length Binary Encodings, *Bell System Techn. J.*, **38** (1959), pp. 933—967.

[4] D. E. KNUTH, Optimum Binary Search Trees, *Acta Informatica*, **1** (1971), pp. 14—15.

[5] T. C. HU and A. C. TUCKER, Optimum Binary Search Trees, *Combinatorial Math. and its Appl.*, Univ. of North Carolina (Chapel Hill, 1969).