

Huffman - type Codes for
Infinite Source Distributions

Julia Abrahams
Mathematical Sciences Division
Office of Naval Research
Arlington, VA 22217-5660

Abstract

A new sufficient condition is given for an infinite source distribution to share a minimum average codeword length code with the geometric distribution. Thus some new examples of parametric families of infinite source distributions can be optimally encoded by Huffman - type codes.

I. Introduction

In the case of infinite sources for which the constructive Huffman algorithm does not apply, an indirect approach has been taken to obtain the minimum average codeword length code for certain parametric families of source distributions. For infinite sources, Gallager and Van Voorhis [3] find the optimal code for the geometric source, completing the solution to a problem first addressed by Golomb [4]. Humblet [6] solves the problem for the Poisson distribution. These results and the methods used to obtain them will be addressed further in the next section.

It also appears that a problem equivalent to coding the infinite geometric source and a finite distribution related to it arises in the statistics literature on group testing and in a search problem in the operations research literature where independent results appear. In particular, Hwang [7] gives explicit expressions for the minimum average binary codeword length for source probabilities $p_k = a^{k-1}(1-a)/(1-a^n)$ for n both finite and infinite where $0 < a < 1$. The codeword lengths themselves seem to be implicit. Explicit codeword lengths for the finite source are derived in Yao and Hwang [8] but only for $M = 2$ where M is the unique positive integer solution to $a^M + a^{M+1} \leq 1 < a^M + a^{M-1}$, that is for $0.618 \leq a \leq 0.755$. The case of $M = 1$, that is $a \leq 0.618$, is well known to be solved by $l_k = k$, $k < n$, $l_n = n-1$. Hassin [5] also discusses these problems in part.

In this note, a new sufficient condition for infinite sources to have minimum average codeword length codes of particular form are presented together with its application to example parametric source distributions. This generalizes the results of Gallager and Van Voorhis [3] and Humblet [6].

II. Sufficient Conditions for a Family of Infinite Codes to
Minimize Average Codeword Length

Humblet [6] shows for probabilities $p_0 \geq p_1 \geq \dots$ that for $p_k \geq p_{k+2} + p_{k+3} + \dots$ for all k the corresponding codeword lengths in the optimal binary code are given by $l_k = k+1$. In addition, if Humblet's sufficient condition holds only for $k \geq m$, the optimal codeword lengths can be obtained from the Huffman algorithm applied to the reduced source with probabilities $p_0, p_1, \dots, p_m, p_{m+1}+p_{m+2}+\dots$; if the lengths of these codewords are $l'_k, k = 0, 1, \dots, m+1$, the lengths for the original source are $l_k = l'_k, k = 0, 1, \dots, m, l_k = l'_{m+1} + k - m, k = m+1, m+2, \dots$. In particular, the geometric distribution $p_k = a^k(1-a)$ satisfies the sufficient condition for all k for $a \leq (\sqrt{5}-1)/2 \approx 0.618$ (that is for a satisfying $a+a^2 \leq 1$) as does the Poisson distribution $p_k = (a^k e^{-a})/k!$ for $a \leq 1$. For $a > 1$, the Poisson distribution satisfies the condition for $k \geq m$ where m depends on a .

Gallager and Van Voorhis's [3] reduced source approach to the geometric distribution is the basis for Humblet's analysis as well as for the more extensive generalization to be given in this note. For the geometric distribution with M the unique integer solution to

$$a^M + a^{M+1} \leq 1 < a^M + a^{M-1}, \quad (1)$$

Gallager and Van Voorhis construct the reduced source with probabilities $p_0, p_1, \dots, p_m, p_{m+1}+p_{m+1+M}+p_{m+1+2M}+\dots, p_{m+2}+p_{m+2+M}+p_{m+2+2M}+\dots, \dots, p_{m+M}+p_{m+2M}+\dots$ when $p_k = a^k(1-a)$. They binary Huffman code the reduced source making use of inequalities among the probabilities to determine the two lowest weight symbols to merge at each stage of the Huffman procedure and then examine the limiting situation as $m \rightarrow \infty$. The result is that l_0, l_1, \dots, l_{m-1} are determined according to the Huffman code for $p_0+p_M+p_{2M}+\dots, p_1+p_{1+M}+p_{1+2M}+\dots, \dots, p_{M-1}+p_{2M-1}+p_{3M-1}+\dots$, respectively and $l_k = l_\beta + \alpha$ for $k = \alpha M + \beta$.

The optimal codes have n_i codewords of length $i, i = 1, 2, \dots$, given in Table 1. For convenient reference, denote a family of codes with these lengths for each M as M -codes.

The contribution of this note is to point out that Gallager and Van Voorhis's argument goes through not only for the geometric distribution but more generally. In particular, we have the

Proposition: The set of inequalities

$$p_{k+M-1} + p_{k+2M-1} + p_{k+3M-1} + \dots \geq p_k \geq p_{k+M+1} + p_{k+2M+1} + p_{k+3M+1} + \dots, \quad (2)$$

$k = 0, 1, 2, \dots$, where $p_0 \geq p_1 \geq \dots$, holding for some M is sufficient for the optimal binary code to be given by an M -code. (Note that this condition (2) for arbitrary probabilities reduces to Gallager and Van Voorhis's (1) for the geometric distribution.)

The derivation exactly follows Gallager and Van Voorhis. For the reduced source with $m+1+M$ symbols, the Huffman algorithm first combines p_m and $p_{m+M}+p_{m+1+2M}+\dots$ since $p_{m+M}+p_{m+2M}+\dots$

$\leq p_{m-1}$ and $p_{m-1+N}+p_{m-1+2N}+\dots \geq p_m$, because of the generalized sufficient condition (2). The combination gives a composite symbol with probability $p_m+p_{m+N}+p_{m+2N}+\dots$, and the situation is now that of the reduced source with $m+N$ letters. The process continues, and a limiting step as in Gallager and Van Voorhis completes the demonstration.

We also have as in Humblet's generalization of Gallager and Van Voorhis for $M = 1$, the

Corollary: If the generalized sufficient condition (2) holds for $k \geq m$, then the codeword lengths for the optimal binary code are determined according to the Huffman code for $p_0, p_1, \dots, p_m, p_{m+1+N}+p_{m+1+2N}+\dots, p_{m+2+N}+p_{m+2+2N}+\dots, \dots, p_{m+N}+p_{m+2N}+\dots$; if the lengths of these codewords are $l'_k, k = 0, 1, \dots, m+N$, the lengths for the original source are $l_k = l'_k, k = 0, 1, \dots, m, l_k = l'_{m+\beta} + \alpha$, for $k = \alpha M + \beta \geq m+1$.

III. New Examples of Optimal Codes for Infinite Sources

Because there are so few examples of infinite sources for which optimal codes are known, it is of interest to apply the sufficient condition (2) of the previous section to some parametric families. One example to which (2) applies at least in part is the distribution $p_{2k} = a^{k+1}(1-a), p_{2k+1} = a^k(1-a)^2, k = 0, 1, 2, \dots$ which is monotonically ordered as $p_0 \geq p_2 \geq p_4 \geq \dots \geq p_{2(N-1)} \geq p_1 \geq p_{2N} \geq p_3 \geq p_{2(N+1)} \geq p_5 \geq \dots$ where N is the unique positive integer determined by $a^{N+1} \leq 1-a < a^N$.

For $N = 1$, (2) holds with $M = 2$ whenever $(1-a)^2 \geq a^3$, and the vector of number of codewords of each length $(0, 2, 2, 2, \dots)$ is optimal for this distribution for a satisfying $0.5 \leq a \leq 0.57$. For $N = 2$, (2) holds with $M = 3$ whenever $(1-a^2-a^3)/(1-a^3) \geq 1-a \geq (1-a-a^3)/(1-a^3), (a-a^2+a^3)/(1-a^3) \geq 1-a \geq (a^2-a^3+a^6)/(1-a^3)$, (this is the constraining pair of inequalities), and $(1-a+a^4)/(1-a^3) \geq a \geq (a-a^2+a^3)/(1-a^3)$; thus the length distribution vector $(0, 1, 3, 3, \dots)$ is optimal for this distribution for a satisfying $0.57 \leq a \leq 0.66$. The example distribution not been verified to satisfy (2) for other parameter values due to the complexity of the calculations involved. A mathematical computer package with symbolic manipulation capability should be effective in carrying out these calculations; this has not been pursued here.

However Golomb's [4] plausibility argument for the geometric distribution also can be used to suggest promising candidate codes for this example distribution for certain values of a . The candidate codes would need to be verified through (2). The argument is that for $a^j = 1/2$, the symbol with probability $a^{k+j}(1-a)$ is half as likely as the symbol with probability $a^k(1-a)$ and should require a codeword one bit longer. Similarly the symbol with probability $a^{k+j}(1-a)^2$ should require a codeword one bit longer than that with

probability $a^k(1-a)^2$. Indeed this argument, together with the imposition of equality in the Kraft-McMillan inequality, $\sum_n 2^{-l_n} = 1$, yields the codeword length distribution vectors $(0, 2, 2, 2, \dots)$ for $a = 0.5$, which has already been verified to satisfy (2), $(0, 1, 2, 4, 4, \dots)$ for $a = 0.707$, that is the $M = 4$ case for $m \geq 3$, and $(0, 0, 3, 4, 6, 6, \dots)$ for $a = 0.794$, that is the $M = 6$ case for $m \geq 5$, and (2) should be able to be verified for these parameter values.

This particular example distribution arises in a run length coding problem. Assume the binary source generates 0's and 1's independently with probabilities $a \geq 0.5$ and $1-a$ respectively. In standard run length coding, runs of k 0's followed by a single 1 are each assigned a codeword. The runs are geometrically distributed so that M-codes encode them with minimum average codeword length. Consider instead runs of 0's punctuated by the occurrence of either the string 10 or 11. This nonstandard run length coding scheme leads to the example distribution discussed above. Clearly many highly structured nonstandard run length coding schemes lead to infinite source distributions of parametric form, however in general it appears difficult to identify parameter values for which (2) is satisfied.

Another example infinite source distribution to which (2) also applies at least in part is the two sided geometric distribution $p_k = a^{|k|} (1-a)/(1+a)$, $k = 0, \pm 1, \pm 2, \dots$ used by Cheung et al. [1,2] in an image compression application. They propose a variant of M-codes in which $l_0 = 1$, l_k for $k > 0$ is obtained from the M-code plus one additional bit, and l_k for $k < 0$ is equal to l_k for $k > 0$. They find the minimum average codeword length possible over the class of M-code variants for the two sided geometric distribution. When the average codeword length for the Huffman code for a finite approximation to the two sided geometric distribution, obtained numerically, matches the minimum over the class of M-code variants, they conclude on the basis of numerical evidence that the optimal code has been found for particular values of a to be in fact one of the M-code variants. By using the corollary, it can be confirmed, for example, that the M-code variant for $M = 2$ is in fact the minimum average codeword length code for $0.618 \leq a \leq 0.707$ as the numerical evidence suggests, and, a new result, it can be seen that the M-code itself for $M = 4$ is the optimal code for $0.707 \leq a < 0.755$ for the two sided geometric distribution. Again, while numerical evidence in [1,2] suggests that the M-code variants are optimal for additional parameter ranges and while it may be that M-codes themselves are optimal for parameter ranges where the numerical evidence indicates that M-code variants are not optimal, the two sided geometric distribution has not been verified to satisfy (2) for other parameter values due to the complexity of the calculations involved.

If two sources satisfy sufficient conditions for the same optimal code, it is immediate that their convex combination satisfies the same conditions and therefore is optimally coded by the same code. Thus, for example, the distribution given by

$$p_{2k} = c a^{k+1}(1-a) + (1-c)b^{2k}(1-b),$$

$$p_{2k+1} = c a^k(1-a)^2 + (1-c)b^{2k+1}(1-b)$$

for $a \in [0.5, 0.57]$, $b \in [0.618, 0.775]$, and $c \in [0, 1]$ is optimally coded by the M-code with $M = 2$.

IV. Optimal Ternary Coding

The same approach can be used to find sufficient conditions for optimal full ternary codes. First, a plausibility argument based on Golomb's using $(a^2)^M = 1/3$ finds us candidate optimal codes for geometric sources and particular values of a . These codes suggest the form of the reduced source for which an argument along the lines of Gallager and Van Voorhis goes through both for geometric sources and the more general case addressed in this note for binary. The reduced sources are of the form

$$p_0, p_1, \dots, p_m, p_{m+1}+p_{m+2}+p_{m+2M+1}+p_{m+2M+2}+\dots,$$

$$p_{m+3}+p_{m+4}+p_{m+2M+3}+p_{m+2M+4}+\dots, \dots, p_{m+2M-1}+p_{m+2M}+p_{m+4M-1}+p_{m+4M}+\dots$$

The combination in the Huffman algorithm of p_{m-1} , p_m , and $p_{m+2M-1}+p_{m+2M}+p_{m+4M-1}+p_{m+4M}+\dots$, which leads to a reduced source with two fewer symbols, is ensured if

$$p_{m+2M-1}+p_{m+2M}+p_{m+4M-1}+p_{m+4M}+\dots \leq p_{m-2}$$

and

$$p_{m+2M-3}+p_{m+2M-2}+p_{m+4M-3}+p_{m+4M-2} \geq p_{m-1}.$$

The overall sufficient conditions which imply this pair of inequalities are

$$p_{k+2M-2}+p_{k+2M-1}+p_{k+4M-2}+p_{k+4M-1}+\dots \geq p_k \geq p_{k+2M+1}+p_{k+2M+2}+p_{k+4M+1}+p_{k+4M+2}, \quad (3)$$

$k = 0, 1, 2, \dots$ For the special geometric case, M is determined according to

$$a^{2(M-1)} \geq (1-a^{2M})/(1+a) \geq a^{2M+1}.$$

The optimal codes have n_i codewords of length i , $i = 1, 2, \dots$, given in Table 2.

A ternary run length coding problem leads to a nongeometric distribution which can also be shown to satisfy (3) for certain parameter values. Here runs of $k > 0$ 0's followed by a single 1 or 2 are each assigned a codeword. If 0's occur with probability a_0 , 1's with probability a_1 , and 2's with probability a_2 , where $a_0 + a_1 + a_2 = 1$, then the infinite source probabilities are $a_1, a_2, a_0a_1, a_0a_2, a_0^2a_1, a_0^2a_2, \dots$ in monotonically nonincreasing order whenever $a_1 \geq a_2 \geq a_0a_1$. For example, for $a_0 = .1, a_1 = .7, a_2 = .2$, we find that (3) is satisfied for $M = 1$, and the optimal code has the codeword lengths given in Table 2.

These seem to be the only nonbinary optimal codes known for infinite sources. It may well be that other parametric source distributions also satisfy the ternary sufficient conditions (3), but that has not been pursued here. Nor has

the D-ary case been pursued, but the same general approach should apply.

V. Conclusion

A new sufficient condition which is useful in optimally binary coding some infinite source distributions is presented. Some new infinite source distributions are optimally coded using this condition. The generalization to ternary coding, particularly for the geometric distribution, is given.

References

1. K.- M. Cheung and P. Smyth, "A high-speed distortionless predictive image compression scheme," Proceedings 1990 International Symposium on Information Theory and Its Applications, Honolulu, Hawaii, pp. 467-470, Nov. 27-30, 1990.
2. K.- M. Cheung, P. Smyth, and H. Wang, "A high-speed distortionless predictive image - compression scheme," TDA Progress Report 42-102, Jet Propulsion Laboratory, Pasadena, CA, Aug. 15, 1990.
3. R. G. Gallager and D.C. Van Voorhis, "Optimal source codes for geometrically distributed alphabets," IEEE Trans. Inform. Theory, vol. IT-21, no. 2, pp. 228-230, Mar. 1975.
4. S. W. Golomb, "Run-length encodings," IEEE Trans. Inform. Theory, vol. IT-12, no. 4, pp. 339-401, Jul. 1966.
5. R. Hassin, "A dichotomous search for a geometric random variable," Operations Research, vol. 32, no. 2, pp. 423-439, Mar.-Apr. 1984.
6. P. Humblet, "Optimal source coding for a class of integer alphabets," IEEE Trans. Inform. Theory, vol. IT-24, no. 1, pp. 110-112, Jan. 1978.
7. F. K. Hwang, "On finding a single defective in binomial group testing," J. American Statistical Association, vol. 69, no. 345, pp. 146-150, Mar. 1974.
8. Y. C. Yao and F. K. Hwang, "On optimal nested group testing algorithms," J. Statistical Planning and Inference, vol. 24, pp. 167-175, 1990.

M	$(n_1, n_2, n_3, n_4, \dots)$
1	(1, 1, 1, 1, ...)
2	(0, 2, 2, 2, ...)
3	(0, 1, 3, 3, ...)
4	(0, 0, 4, 4, ...)
5	(0, 0, 3, 5, 5, ...)
6	(0, 0, 2, 6, 6, ...)
7	(0, 0, 1, 7, 7, ...)
8	(0, 0, 0, 8, 8, ...)
9	(0, 0, 0, 7, 9, 9, ...)
...	...

Table 1: Optimal Binary Codes for Geometric Sources

M	$(n_1, n_2, n_3, n_4, \dots)$
1	(2, 2, 2, 2, ...)
2	(1, 4, 4, 4, ...)
3	(0, 6, 6, 6, ...)
4	(0, 5, 8, 8, ...)
5	(0, 4, 10, 10, ...)
6	(0, 3, 12, 12, ...)
7	(0, 2, 14, 14, ...)
8	(0, 1, 16, 16, ...)
9	(0, 0, 18, 18, ...)
...	...

Table 2: Optimal Ternary Codes for Geometric Sources