M-ary Anti - Uniform Huffman Codes for Infinite Sources With Geometric Distribution

Daniela Tarniceriu[1], Valeriu Munteanu[1], Gheorghe Zaharia[2],
[1] Faculty of Electronics, Telecommunications and Information Technology, Gheorghe Asachi Technical University of Iasi, Romania,
e-mail: tarniced@etti.tuiasi.ro
[2] IETR – INSA, UMR CNRS 6164 Rennes, France, e-mail: Gheorghe.Zaharia@insa-rennes.fr

***Abstract*—In this paper we consider the class of generalized anti-uniform Huffman (AUH) codes for sources with infinite alphabet and geometric distribution. This distribution leads to infinite anti – uniform sources for some ranges of its parameters. Huffman coding of these sources results in AUH codes. We perform a generalization of binary Huffman encoding, using a M-letter code alphabet and prove that as a result of this encoding, sources with memory are obtained. For these sources we attach the graph and derive the transition probabilities between states, as well as the state probabilities. The entropy and the average cost for AUH codes are derived.**

I. INTRODUCTION

Anti-uniform Huffman (AUH) sources appear in a wide variety of situations in the real world, because this class of sources have the property of achieving minimum redundancy in different situations and minimal average cost in highly unbalanced cost regime [1]-[3]. Unequal cost letter problem modeling situations in which different characters have different transmission times or storage costs was addressed in [4], [5]. One example is the telegraph channel with the alphabet {. -} in which dashes are twice as long as dots [6]. Another example is the {a, b} run – length – limited codes used in magnetic and optical storage, in which the binary codewords are constrained so that each 1 must be preceded by at least *a*, and at most *b*, 0's [7]. There is a large literature addressing the problem of cost of prefix-free codes with unequal letter cost encoding alphabet [8] and references herein.

Consider a discrete memoryless information source with infinite size $\xi:(s_1 \;\; s_2 \;\; \cdots \;\; s_k \cdots)$ and associated ordered probability distribution $P:(p_1 \; p_2 \cdots p_k \cdots)$, where $p_1 \geq p_2 \geq ... \geq p_k \geq ....$ We assume the general case of an alphabet consisting of M letters, ($M \geq 2$). After Huffman encoding [9] a tree graph results, whose leaves are terminal nodes and correspond to the source messages. The M edges emanating from each intermediate tree node are labeled by a_1, a_2 and a_M, respectively. These letters belong to the code alphabet $A=\{a_1, a_2, ..., a_M\}$. The length between the root and a leaf is the length of the codeword associated with the corresponding message.

In [10] it is shown that, for minimum redundancy Huffman codes, the average codeword length becomes minimum, if the source distribution is chosen so that on each level of the tree only one node diversifies.

Assuming that v_k, $k=1,2,...$, is the codeword representing the message s_k, we denote the length of v_k by l_k. The optimality of Huffman coding implies that $l_k \leq l_j$, if $p_k > p_j$.

Anti uniform Huffman (AUH) codes were firstly introduced in [11] and they are characterized by the fact that $l_k = k$, for $k=1,2,...$.

For this, the following condition has to be fulfilled [11]:

$$\sum_{k=i+2}^{\infty} p_k \leq p_i, \; i \geq 1 \tag{1}$$

The AUH codes have been extensively analyzed, concerning bounds on average codeword length, entropy and redundancy for different types of probability distribution. In [12] it has been shown that these codes maximize the average length and the entropy. Tight lower and upper bounds on average codeword length, entropy and redundancy of finite and infinite AUH codes in terms of alphabet size are derived. Related topics are addressed in [13]-[15]. The problem of M-ary Huffman codes is analyzed in [16] and it is shown that for AUH codes, by a proper choice of the source probabilities, the average codeword length can be made closed to unity. In [17] and [18] a general treatment and an information analysis of M-ary Huffman encoding are performed. The problem of bounding the average length of an optimal Huffman code is considered in [19], when only limited knowledge of the source symbol probability distribution is available.

AUH sources can be generated by several probability distributions**.** It has been shown that geometric distribution lays in the class of AUH sources for some regimes of their parameters [3], [20].

The rest of the paper is organized as follows. In Section II we consider the AUH sources with geometric distribution and infinite alphabet. For this source we compute the entropy, perform a M-ary Huffman encoding and compute the average codeword length. We show that, in general, employing Huffman coding, a source with memory results. The graph of the source with memory resulting by M-ary Huffman encoding

of the AUH source is also built. For this source with memory we compute the state probabilities and the transition probabilities between states. In Section III we compute the code entropy, representing the average information per symbol, as well as the average cost for AUH codes corresponding to sources with geometric distribution for infinite source alphabet. Finally, we conclude the paper in Section IV.

II. M-ARY HUFFMAN ENCODING AS SOURCE WITH MEMORY

Let us consider a discrete source with infinite alphabet, characterized by the geometric distribution [20]:

$$\xi:\begin{pmatrix} s_1^{(t)} & s_2^{(t)} & \cdots & s_{M-1}^{(t)} & s_{(M-1)+1}^{(t)} & \cdots & s_{2(M-1)}^{(t)} \\ q & qp & \cdots & qp^{M-2} & qp^{(M-1)} & \cdots qp^{2(M-1)-1} \end{pmatrix}$$
$$\begin{pmatrix} \cdots & s_n^{(t)} & \cdots & s_{k(M-1)+1}^{(t)} & \cdots & s_{(k+1)(M-1)}^{(t)} & \cdots \\ \cdots & qp^{n-1} & \cdots & qp^{k(M-1)} & \cdots & qp^{(k+1)(M-1)-1} & \cdots \end{pmatrix}, \quad (2)$$

where q=1-p and 0 < p < 1. We note that the message probability is $p(s_n^{(t)}) = p_n^{(t)} = qp^{n-1}$.

For the sake of simplicity, in the following, we use the superscript (*t*) to indicate a terminal node in the tree, corresponding to a source message. To indicate a message $s_n^{(t)}$ on a level *k* in the coding tree, we use the index $n = (k-1)(M-1)+j$, $j = 1,\ldots,M-1$. The source is complete, that is

$$\sum_{n=1}^{\infty} p_n^{(t)} = 1. \quad (3)$$

For this source to be anti – uniform, any message probability on a level *k* has to be greater than the sum of all message probabilities placed on next levels. In other words, the smallest message probability on a level *k*, $p_{(k-1)(M-1)}^{(t)} = qp^{(k-1)(M-1)-1}$, has to be greater than the probability of the intermediate node on that level, $p_k^{(i)} = p^{k(M-1)}$. This is because the probability of an intermediate node is equal to the sum of all terminal nodes placed on all next levels. Under these circumstances, the source is AUH, if $1-p-p^M > 0$. Note that, at limit, when $M \to \infty$, the condition becomes totally unrestrictive, $p < 1$.

The entropy of the coded source ξ is [21]:

$$H(\xi) = -\sum_{n=1}^{\infty} p\left(s_n^{(t)}\right) \log p\left(s_n^{(t)}\right). \quad (4)$$

Considering probabilities in (2), the source entropy becomes:

$$H(\xi) = -\left(\log(1-p) + \frac{p}{1-p}\log p\right). \quad (5)$$

After a M-ary Huffman encoding of this source, the graph in Fig. 1 results, that is, an infinite anti – uniform code. This means that on each level there are *M*-1 code words and only one node diversifies. The length of words on level *k* is equal to *k*. Therefore, $s_{(k-1)(M-1)+j}^{(t)}$, $j = 1,\ldots,M-1$, represents a leaf or a terminal node in the graph, on level *k*, corresponding to the message $s_{(k-1)(M-1)+j}^{(t)}$ and $s_k^{(i)}$ represents the intermediate node

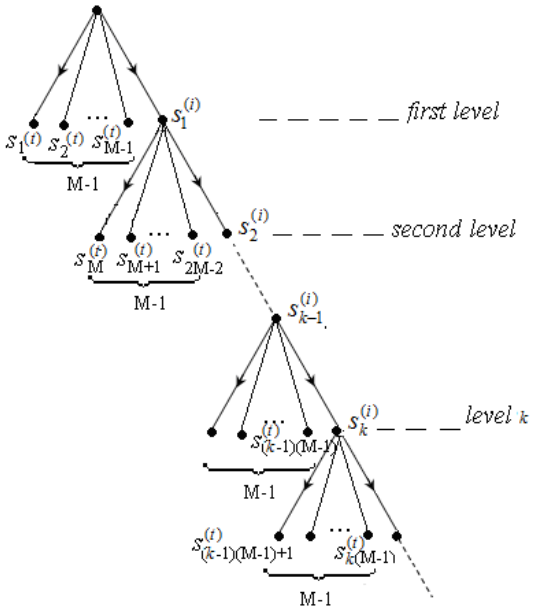

Figure 1. The graph of M-ary Huffman encoding for the source ξ with distribution in (2)

on level "*k*". Only this intermediate node diversifies on this level.

The probabilities of terminal nodes are equal to the probabilities of the source messages $p_n^{(t)}$. Unlike a leaf, an intermediate node is not corresponding to a source message, therefore no probability mass is associated. However, with slight abuse we can call the weight of the intermediate node also probability.

Considering (3), the probabilities of intermediate nodes $p\left(s_k^{(i)}\right) = p_k^{(i)}$ are obtained recursively, as

$$p_k^{(i)} = 1 - \sum_{i=1}^{k} \sum_{j=1}^{M-1} p_{(i-1)(M-1)+j}^{(t)}. \quad (6)$$

Considering (2) and (6), the probabilities of terminal and intermediate nodes are obtained by:

$$p_{(i-1)(M-1)+j}^{(t)} = qp^{(i-1)(M-1)+j-1},\ p_k^{(i)} = p^{k(M-1)} \quad (7)$$

The structure of codewords resulting by M-ary Huffman encoding is:

$$\begin{aligned} & s_1^{(t)} \to v_1 \to a_1 \\ & s_2^{(t)} \to v_2 \to a_2 \\ & s_{M-1}^{(t)} \to v_{M-1} \to a_{M-1} \\ & \ldots\ldots\ldots\ldots\ldots\ldots \\ & s_{(k-1)(M-1)+1}^{(t)} \to v_{(k-1)(M-1)+1} \to \underbrace{a_M a_M \ldots a_M}_{k-1} a_1 \\ & s_{k(M-1)}^{(t)} \to v_{k(M-1)} \to \underbrace{a_M a_M \ldots a_M}_{k-1} a_{M-1} \\ & \ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \quad (8)$$

The length l_n of the codeword associated with the messages $s^{(t)}_{(k-1)(M-1)+j}$ on the level k is the number of edges on the path between the root and the node $s^{(t)}_{(k-1)(M-1)+j}$ in the Huffman tree.

$$l_{(k-1)(M-1)+j} = k, \quad k = 1,2,...; \; j = 1,...,M-1. \tag{9}$$

The average codeword length is determined with

$$\bar{l} = \sum_{n=1}^{\infty} p_n^{(t)} l_n = \sum_{k=1}^{\infty} \sum_{j=1}^{M-1} k p_{(k-1)(M-1)+j}. \tag{10}$$

The average codeword length is obtained considering (2) into (10)

$$\bar{l} = \sum_{k=1}^{\infty} \sum_{j=1}^{M-1} k q p^{(k-1)(M-1)+j-1} = \frac{1}{1-p^{M-1}}. \tag{11}$$

For a sequence of messages of the source ξ, a sequence of symbols from the code alphabet A will be transmitted. As long as the probabilities of these symbols depend, generally, on the node from which they are generated, the set $A = \{a_1, a_2, ..., a_M\}$ becomes a source with memory. When, at a certain moment, a terminal node is reached, the source ξ will deliver another message and the source with memory A will deliver another sequence of messages a_j, $j = 1,2,...M$. Its states correspond to terminal or intermediate nodes (excepting the root) in the graph in Fig. 1. When a terminal node is reached, the M-ary encoding Huffman procedure is resumed from the graph root. Since the source with the distribution in (2) is complete, the probability of the root is equal to 1.

The graph attached to the source with memory A is shown in Fig. 2 and it can be obtained from the Huffman encoding graph of the source ξ (Fig. 1), as follows:

a) We link the terminal nodes in the graph of the source ξ with with the graph root;

b) The branches between succesive nodes have the probabilities equal to the ratio between the probability of the node in which the branch ends and the probability of the node from which it starts, excepting the branches linking the terminal nodes with graph root, whose probabilities are equal to unity;

c) Each terminal or intermediate node (excepting the graph root) will represent a state $S^{(t)}_{(k-1)(M-1)+j}$ or $S^{(i)}_k$, $k = 1,2,...$ and $j = 1,...,M-1$, (as represented in Fig. 2).

Let $S = \{S_1^{(t)}, S_2^{(t)}, ..., S_{M-1}^{(t)}, S_1^{(i)}, S_M^{(t)}, S_{M+1}^{(t)}, ..., S_{2M-2}^{(t)}, S_2^{(i)}, ...\}$ be the state set of the source with memory.

The probability of delivering symbol a_j, $j = 1,2,...M-1$ from the state $S^{(i)}_k$, $k = 1,2,...$ is equal to the probability of transition to the state $S^{(t)}_{k(M-1)+j}$, $j = 1,2,...M-1$.

$$p(a_j \mid S_k^{(i)}) = p(S^{(t)}_{k(M-1)+j}) \mid p(S_k^{(i)}) = \frac{q p^{kM-1)+j-1}}{p^{k(M-1)}} = q p^{j-1}, \tag{12}$$

$j = 1,...,M-1$

The probability of delivering the symbol a_M, from the state $S^{(i)}_k$, $k = 1,2,...$ is equal to the probability of transition to the state $S^{(i)}_{k+1}$

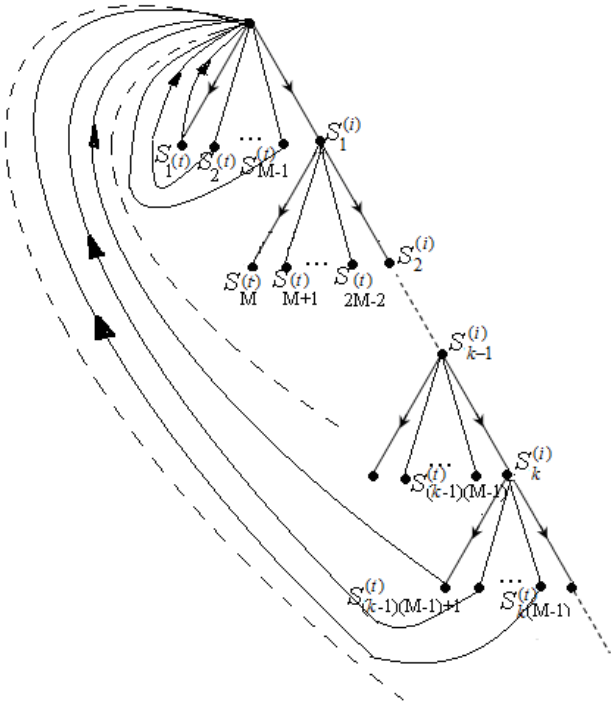

Figure 2. The graph of the source with memory

$$p(a_M \mid S_k^{(i)}) = p(S_{k+1}^{(i)} \mid S_k^{(i)}) = \frac{p^{(k+1)(M-1)}}{p^{k(M-1)}} = p^{M-1}. \tag{13}$$

The probability of delivering symbol a_j, $j = 1,2,...M-1$ from the state $S_n^{(t)}$, $n = (k-1)(M-1)+j$ is equal to the probability of transition to the state $S_j^{(t)}$

$$p(a_j \mid S_n^{(t)}) = p(S_j \mid S_n^{(t)}) = p_j^{(t)} = q p^{j-1}, \; j = 1,...,M-1 \tag{14}$$

and probability of delivering the symbol a_M from the state $S_n^{(t)}$, $n = (k-1)(M-1)+j$ is

$$p(a_M \mid S_n^{(t)}) = p(S_1^{(i)} \mid S_n^{(t)}) = p(s_1^{(i)}) = p^{M-1}. \tag{15}$$

The transition probabilities (12) – (15) can be organized in the transition matrix between states, **T**.

Let $\pi_n^{(t)}$ and $\pi_k^{(i)}$, denote the state probabilities of the source with memory. They can be determined by means of [20]:

$$[\pi_1^{(t)} ... \pi_n^{(t)} ... \pi_1^{(i)} ... \pi_k^{(i)} ...] = [\pi_1^{(t)} ... \pi_k^{(t)} ... \pi_1^{(i)} ... \pi_k^{(i)} ...]\mathbf{T} \tag{16}$$

$$\sum_{n=1}^{\infty} \pi_n^{(t)} + \sum_{k=1}^{\infty} \pi_k^{(i)} = 1. \tag{17}$$

Considering (10) and (12) – (15), from (16) and (17) we get the state probabilities:

$$\pi_n^{(t)} = \frac{1}{\bar{l}} p_n^{(t)}, n = (k-1)(M-1)+j, \quad k = 0,1,2,...; \; j = 1,...,M-1 \tag{18}$$

$$\pi_k^{(i)} = \frac{1}{\bar{l}} p_k^{(i)} = \frac{1}{\bar{l}} \left(1 - \sum_{i=1}^{k} \sum_{j=1}^{M-1} p_{(i-1)(M-1)+j}^{(t)} \right). \tag{19}$$

Considering (7) into (18) and (19), we get the stationary state probabilities:

$$\pi_n^{(t)} = \frac{1}{\bar{l}} qp^{n-1} \tag{20}$$

$$\pi_k^{(i)} = \frac{1}{\bar{l}} p^{k(M-1)} \tag{21}$$

III. Entropy and average cost of AUH M-ary codes

Generally, the entropy of the source with memory is computed by [21]

$$H(A) = -\sum_{n=1}^{\infty}\sum_{j=1}^{M-1} \pi_n^{(t)} p\left(a_j \middle| S_n^{(t)}\right) \log p\left(a_j \middle| S_n^{(t)}\right) - \sum_{k=1}^{\infty}\sum_{j=1}^{M-1} \pi_k^{(i)} p\left(a_j \middle| S_k^{(i)}\right) \log p\left(a_j \middle| S_k^{(i)}\right). \tag{22}$$

Substituting (12) - (15) and (20), (21) into (22), we get the entropy of the source with memory, as

$$H(A) = -\frac{1}{\bar{l}}\left(\log(1-p) + \frac{p}{1-p}\log p\right). \tag{23}$$

Let $c_1, c_2, ..., c_M$ be the costs associated to the code alphabet letters $a_1, a_2, ..., a_M$, respectively. The average cost of a code is defined by [13]

$$\overline{C} = \sum_{n=1}^{\infty} p_n^{(t)} \sum_{j=1}^{M} n_j(n) c_j \ , \tag{24}$$

where we denote by $n_j(n)$ the number of symbols a_j in the codeword corresponding to the source symbol $s_n^{(t)}$.

Considering (2) and (9), the average cost is

$$\overline{C} = \sum_{k=1}^{\infty}\sum_{j=1}^{M-1} qp^{(k-1)(M-1)+j-1}\left(c_j + (k-1)c_M\right) \tag{25}$$

IV. Conclusions

In this paper we have considered the case of infinite AUH sources with infinite alphabet, generated by geometric distribution. We have showed that by M-ary Huffman encoding of these sources, we obtained a source with memory A. We have specified the rules for drawing the graph of the source with memory *A*, and the calculation of state probabilities and transition probabilities between states. We have calculated the entropy of the encoded source, the average length of the M-ary Huffman code, the entropy of the source with memory obtained as result of M-ary Huffman encoding, as well as the average cost of codes in this situation. From (11) we note that as the cardinality of the set *A* increases, the average length of codewords decreases. At limit, when $M \to \infty$, the average length tends to unity. From (23), we note that with increasing the cardinality of the set *A*, the entropy of the source with memory increases. At limit, when $M \to \infty$, the entropy of the source with memory becomes equal to the entropy of the initial source.

References

[1] O. Johnsen, "On the redundancy of binary Huffman codes," *IEEE Trans. Inf. Theory*. vol. 26, pp.220-222, 1980.

[2] P. Bradford, M Golin, L. L. Larmore and W. Rytter, "Optimal prefix free codes for unequal letter costs and dynamic programming with the Monge property," *J. Algorithms*, vol. 42, pp. 277 – 303, 2002.

[3] E. N. Gilbert, "Coding with digits of unequal costs," *IEEE Trans. Inf. Theory*. vol. 41, pp. 596-600, 1995.

[4] N. M. Blachman, Minimum cost coding of information," *IRE Trans. Inf. Theory*, vol. PGIT-, pp. 139-149, Mar. 1954.

[5] B. Varn, "Optimal variable length codes (arbitrary symbol cost and equal code word probability)," *Inf. Contr*. Vol. 19, pp. 289-301, 1971.

[6] R. M. Krause, "Channels which transmit letters of unequal duration," *Inf. Contr*., Vol. 55, pp. 13 – 24, 1962.

[7] M. Golin, G. Rote, "A dynamic programming algorithm for constructing optimal prefix-free codes for unequal costs," *IEEE Trans. IT* 44(5):1770-1781, 1998.

[8] M. Golin, J. Li, More efficient algorithms and analyses for unequal letter cost prefix-free coding, *IEEE Trans. Inf. Theory*. vol. 54, no. 8, pp. 3412-3424, Aug. 2008.

[9] R. Huffman, A method for the construction of minimum – redundancy codes. *Proc. IRE* 1952; **40:** 1098 – 1101.

[10] V. Munteanu, D. Tarniceriu, "Bounds on codeword lengths of optimal codes for noiseless channels", Proc. of International Symposium ISSCS'07, 2007, Iasi, pp, 493-496 , ISBN 0-7803-7979-9.

[11] Esmaeili, M., Kakhbod, A.: On antiuniform and partially antiuniform sources. *Proc. IEEE ICC*, pp. 1611 – 1615 June 2006

[12] Mohajer, S., Kakhbod, A.: Anti – uniform Huffman codes. *IET Commun*., vol. 5, pp. 1213 – 1219 (2011)

[13] S. Mohajer, A. Kakhbod, "Tight bounds on the AUH codes," Information Sciences and Sysyems, 2008. CISS 2008. 42nd Annual Conference on, pp. 1010 – 1014, 19 – 21 March

[14] Mohajer, S., Pakzad, P., Kakhbod, A.: Tight bounds on the redundancy of Huffman codes. *Proc. IEEE ITW,* , pp. 131 – 135 (March, 2006)

[15] Esmaeili, M., Kakhbod, A.: On information theory parameters of infinite anti-uniform sources. *IET Commun.*, vol.1, pp. 1039 – 1041 (2007)

[16] G. Zaharia, V. Munteanu, D. Tarniceriu, "Tight bounds on the codeword length and average codeword length for D-ary Huffman codes" Part I, II, *Proc. IEEE, ISSCS*, 2009, Iasi, Romania

[17] V. Munteanu, D. Tarniceriu, Gh. Zaharia, Information quantities attached to M-ary antiuniform Huffman codes, *IEEE ISSCS Proceedings*, volume 2, pp. 549-552, July 9-10, 2009, Iasi, Romania.

[18] V. Munteanu, D. Tarniceriu, Gh. Zaharia, Information analysis for a large class of discrete sources, *IEEE ISSCS Proceedings*, volume 2, pp. 553-556, July 9-10, 2009, Iasi, Romania.

[19] F. Cicalese, U. Vaccaro, "Bounding the Average Length of Optimal Source Codes Via Majorization Theory," *Trans. Inf. Theory*. vol. 50, no. 4, pp. 633-637, 2004.

[20] R. Gallager, D. Van Voorhis, "Optimal source coding for geometrically distributed integer alphabets," *IEEE Trans. Inf. Theory*, vol. 21, No. 2, pp. 228 – 230, 1975.

[21] T. M. Cover, J. A. Thomas, Elements of Information Theory. *John Wiley and Sons, Inc.* New York, 1991.