M-ary Anti - Uniform Huffman Codes for Infinite Sources With Geometric Distribution
Daniela Tarniceriu, Valeriu Munteanu, Gheorghe Zaharia

To cite this version:

HAL Id: hal-00873922
https://hal.archives-ouvertes.fr/hal-00873922
Submitted on 16 Oct 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
M-ary Anti - Uniform Huffman Codes for Infinite Sources With Geometric Distribution

Daniela Tarniceriu¹, Valeriu Munteanu¹, Gheorghe Zaharia²,
¹Faculty of Electronics, Telecommunications and Information Technology, Gheorghe Asachi Technical University of Iasi, Romania, e-mail: tarniced@etti.tuiasi.ro
²IETR – 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 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 letter cost 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 \( \xi : (s_1, s_2, \ldots) \) and associated ordered probability distribution \( P : (p_1, p_2, \ldots) \), where \( p_1 \geq p_2 \geq \ldots \geq p_i \geq \ldots \). 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 \( a_1, a_2, \ldots, a_M \), respectively. These letters belong to the code alphabet \( A = \{a_1, a_2, \ldots, a_M\} \). 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 \( v_k, k = 1, 2, \ldots \), 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 \( p_i > p_j \).

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

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

\[
\sum_{i=1}^{\infty} p_i \leq 1, \quad i \geq 1
\]

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 encoding 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]:

\[
\xi = \left\{ s_1^{(i)}, s_2^{(i)} \ldots s_{M-1}^{(i)}, s_{M-1}^{(i)}, s_{M-1}^{(i+1)}, \ldots, s_{M-1}^{(i+1)}, \ldots \right\}, \quad q, q_p q_p q_p_{i}, \ldots, q_{i+1}_{(i+1)}, \ldots
\]

where \(q = 1-p\) and \(0 < p < 1\). We note that the message probability is \(p_i = \left(\prod_{j=1}^{i} p_j\right)q^i\). For the sake of simplicity, in the following, we use the superscript \((i)\) to indicate a terminal node in the tree, corresponding to a source message. To indicate a message \(s_n^{(i)}\) 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

\[
\sum_{n=1}^{\infty} p_n^{(i)} = 1. \tag{3}
\]

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 \(k\) is \(p_k^{(i)} = p_k^{(i)}\). 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 \(1 - p - p^k > 0\). Note that, at limit, when \(M \to \infty\), the condition becomes totally unrestrictive, \(p < 1\).

The entropy of the coded source \(\xi\) is [21]:

\[
H(\xi) = -\sum_{n=1}^{\infty} p_n^{(i)} \log p_n^{(i)}. \tag{4}
\]

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

\[
H(\xi) = -\left(\log(1-p) + \frac{p}{1-p} \log p\right). \tag{5}
\]

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, \(s_{(k-1)(M-1)+j}^{(i)}\), \(j = 1, \ldots, M-1\), represents a leaf or a terminal node in the graph, on level \(k\), corresponding to the message \(s_{(k-1)(M-1)+j}^{(i)}\) and \(s_{(i)}^{(i)}\) represents the intermediate node on level “\(k\)”. Only this intermediate node diversifies on this level.

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

- \(s_1^{(i)} \to v_1 \to a_1\)
- \(s_2^{(i)} \to v_2 \to a_2\)
- \(s_{M-1}^{(i)} \to v_{M-1} \to a_{M-1}\)
- \(s_{(k-1)(M-1)+j}^{(i)} \to v_{(k-1)(M-1)+j} \to a_1a_2 \ldots a_{k-1} a_j\)
- \(s_{(i)}^{(i)} \to v_i^{(i)} \to a_1a_2 \ldots a_{k-1} a_j a_{k-1} \)

The probabilities of terminal nodes are equal to the probabilities of the source messages \(p_i^{(i)}\). 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 \(p_i^{(i)}\) are obtained recursively, as

\[
p_i^{(i)} = 1 - \sum_{j=1}^{k} \sum_{j=1}^{M-1} p_i^{(i)} - \sum_{j=1}^{k} \sum_{j=1}^{M-1} p_i^{(i)}. \tag{6}
\]

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

\[
p_i^{(i)} = q^{(i-k)(M-1)+j}. \tag{7}
\]

The probabilities of the source messages \(p_i^{(i)}\) are

The graph of M-ary Huffman encoding for the source \(\xi\) with distribution in (2)

Figure 1. The graph of M-ary Huffman encoding for the source \(\xi\) with distribution in (2)
The length $l_o$ of the codeword associated with the messages $s_{k+1}^{(i)}$ on the level $k$ is the number of edges on the path between the root and the node $s_{k+1}^{(i)}$ in the Huffman tree.

\[ l_{k+1}^{(i)} = k, \quad k = 1, 2, \ldots; \quad j = 1, \ldots, M - 1. \]  

(9)

The average codeword length is determined with

\[
\bar{L} = \frac{1}{n} \sum_{k=1}^{n} \sum_{j=1}^{M} \text{kp}_{k+1}^{(i)} = \frac{1}{1 - p^{M-1}}. 
\]  

(10)

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

\[
\bar{L} = \frac{1}{n} \sum_{k=1}^{n} \sum_{j=1}^{M} \text{kp}_{k+1}^{(i)} = \frac{1}{1 - p^{M-1}}. 
\]  

(11)

For a sequence of messages of the source $\xi$, 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 $A = \{a_1, a_2, \ldots, a_M\}$ becomes a source with memory. When, at a certain moment, a terminal node is reached, the source $\xi$ will deliver another message and the source with memory $A$ will deliver another sequence of messages $a_{j_1}, j = 1, 2, \ldots, M$. 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 $\xi$ (Fig. 1), as follows:

a) We link the terminal nodes in the graph of the source $\xi$ with the graph root;

b) The branches between successive 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 $S_{k+1}^{(i)}$ or $S_k^{(i)}$, $k = 1, 2, \ldots, j = 1, \ldots, M - 1$, (as represented in Fig. 2).

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

The probability of delivering symbol $a_j$, $j = 1, 2, \ldots, M - 1$ from the state $S_k^{(i)}$, $k = 1, 2, \ldots$ is equal to the probability of transition to the state $S_{k+1}^{(i)}$.

\[
p(a_j | S_k^{(i)}) = p(S_{k+1}^{(i)} | S_k^{(i)}) = \frac{p_{j+1}^{(i)}}{p^{(i)}} = p^{j-1}, \quad j = 1, \ldots, M - 1.
\]

(12)

The probability of delivering the symbol $a_M$, from the state $S_k^{(i)}$, $k = 1, 2, \ldots$ is equal to the probability of transition to the state $S_{k+1}^{(i)}$.
Considering (7) into (18) and (19), we get the stationary state probabilities:

\[ \pi_{s_{n}}^{(i)} = \frac{1}{l} q p^{n-1} \]  
(20)

\[ \pi_{s_{M}}^{(i)} = \frac{1}{l} p^{(M-1)} \]  
(21)

III. ENTROPY AND AVERAGE COST OF AUH M-ARY CODES

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

\[ H(A) = -\sum_{k=1}^{\infty} \sum_{j=1}^{M-1} \pi_{s_{n}}^{(i)} p(a_{j} | s_{n}^{(i)}) \log p(a_{j} | s_{n}^{(i)}) \]  
(22)

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

\[ H(A) = \frac{1}{l} \left( \log(1 - p) + \frac{p}{1 - p} \log p \right) \]  
(23)

Let \( c_{1}, c_{2}, \ldots, c_{M} \) be the costs associated to the code alphabet letters \( a_{1}, a_{2}, \ldots, a_{M} \), respectively. The average cost of a code is defined by [13]

\[ \overline{C} = \sum_{n=1}^{\infty} \sum_{j=1}^{M} n_{j}(n) c_{j} \]  
(24)

where we denote by \( n_{j}(n) \) the number of symbols \( a_{j} \) in the codeword corresponding to the source symbol \( s_{n}^{(i)} \).

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

\[ \overline{C} = \sum_{k=1}^{\infty} \sum_{j=1}^{M-1} q p^{(j-k(M-1))-1} (c_{j} + (k-1)c_{M}) \]  
(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 \( M \rightarrow \infty \), 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 \( M \rightarrow \infty \), the entropy of the source with memory becomes equal to the entropy of the initial source.

REFERENCES