BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Tomul LVIII (LXII), Fasc. 1, 2012
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ

INFINITE ANTI-UNIFORM SOURCES WITH GEOMETRIC DISTRIBUTION

BY

DANIELA TĂRNICERIU[1,*], VALERIU MUNTEANU[1] and GHEORGHE ZAHARIA[2]

[1]“Gheorghe Asachi” Technical University of Iaşi
Faculty of Electronics, Telecommunications and Information Technology
[2]Institute for Electronics and Telecommunications,
Rennes, France

Received: January 24, 2012
Accepted for publication: February 19, 2012

Abstract. In this paper we consider the class of anti-uniform Huffman (AUH) codes for sources with infinite alphabet generated by geometric distribution. Huffman encoding of this source results in AUH codes. As a result of this encoding, in general, we obtain sources with memory. The entropy and the average cost of this source with memory are derived. We perform an analogy between sources with memory and discrete memoryless channels, showing that the entropy of the source with memory is similar to the mean error of the discrete memoryless channel. The information quantity, $I(X,S)$, specifies for AUH codes whether they are with memory or not, as it differs from zero or is equal to zero, respectively.

Key words: average codeword length; average cost; entropy; Huffman coding.

1. Introduction

Consider a discrete source with infinite size alphabet, $\xi:(s_1\ s_2\ \cdots s_k\ \cdots)$, and associated ordered probability distribution, $P:(p_1\ p_2\ \cdots p_k\ \cdots)$, where

*Corresponding author: *e-mail*: tarniced@etti.tuiasi.ro

$p_1 \geq p_2 \geq ... \geq p_k \geq ...$ It is well known that the Huffman encoding algorithm (Huffman, 1952) provides an optimal prefix-free code for this source. A binary Huffman code is usually represented using a binary tree, whose leaves correspond to the source messages. The two edges emanating from each intermediate tree node (father) are labeled either 0 or 1. For related literature on Huffman coding and Huffman trees, we refer the reader to the papers published by Linder *et al*., (1997); Capocelli *et al*., (1991); Gallager, (1978); Johnsen, (1980); Khorsavifard *et al*., (2003).

In contrast with the uniform Huffman code, where $\left|l_k - l_j\right| \leq 1$, ($l_k$ denotes the length of the codeword associated with the message s_k), a code is called *anti-uniform Huffman* (AUH) if $l_k = k+1$, for $k = 0,1,2,...$ In this case the following condition has to be fulfilled (Esmaeili *et. al.,* 2006, 2007):

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

The class of AUH sources is known for their property of achieving minimum redundancy in different situations. It has been shown by Mohajer *et. al*., (2006), that AUH codes potentially achieve the minimum redundancy of a Huffman code of a source for which the probability of one of the symbols is known. The AUH codes are efficient codes with minimal average cost in highly unbalanced cost regime among all prefix-free codes (Mohajer, 2011). These properties determine a wide range of applications and motivate us to study these sources from information theoretic perspective. Such sources can be generated by a several probability distributions. It has been shown that geometric distribution is among the class of infinite alphabet anti-uniform sources (Esmaeili *et. al.,* 2006, 2007; Humblet, 1978; Gallager *et. al.,* 1975).

In this paper we consider the AUH structure and derive the average codeword length, the average information per binary symbol of the source with memory or code entropy, $H(X)$, as result of Huffman encoding of the discrete AUH source with geometric distribution, as well as the average cost of the code.

The rest of the paper is organized as follows. In Section **2** we consider an infinite source with geometric distribution and compute its entropy. For this source we perform a Huffman encoding and derive the average codeword length. We also show that employing Huffman coding, in general, a source with memory results. The average cost of the code is also derived. An analogy between sources with memory and discrete memoryless channels is proposed. The information quantity corresponding to mutual information for discrete channels, $I(X,S)$, specifies for AUH codes whether they are with memory or not, as it differs from zero or is equal to zero, respectively. We conclude the paper in Section **3**.

2. The Entropy and the Average Cost of AUH Codes for Sources with Infinite Alphabet

Let there be a discrete source with infinite alphabet, characterized by the geometric distribution

$$\xi:\begin{pmatrix} s_0^{(t)} & s_1^{(t)} & s_2^{(t)} & \cdots & s_k^{(t)} & \cdots \\ p_0^{(t)}=q & p_1^{(t)}=pq & p_2^{(t)}=p^2q & \cdots & p_k^{(t)}=p^kq & \cdots \end{pmatrix}, \tag{2}$$

where $q = 1 - p$.

Gallager (1975) has shown that geometric distribution with parameter $0 < p \le (\sqrt{5}-1)/2$ satisfies condition (1) and leads to an AUH code.

The source is complete, because (Larsen, 2001; Corduneanu, 2011)

$$\sum_{k=0}^{\infty} p^k q = 1. \tag{3}$$

After a binary Huffman encoding of this source, the graph in Fig. 1 results, that is, an infinite anti-uniform code. $s_k^{(t)}$ represents a leaf or a terminal node in the graph, corresponding to the message $s_k^{(t)}$ of the source and $s_k^{(i)}$ represents the intermediate node, "k".

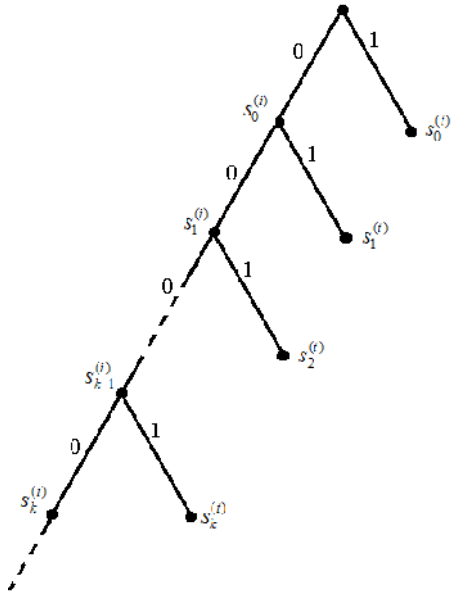

Fig. 1 – The graph of Huffman encoding for the source ξ with distribution in (2).

The probabilities of terminal nodes are equal to probabilities of the source messages, $p_k^{(t)}$. Unlike a leaf, an intermediate node is not corresponding to a source message and, 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 $p_k^{(i)}$ are obtained recursively, as the sum of the two siblings. In this way, we get

$$p_k^{(i)} = 1 - \sum_{j=0}^{k} p_j^{(t)}, \ \ (k = 0,1,2,...). \tag{4}$$

Considering (2) and (4), the probabilities of intermediate nodes are

$$p_k^{(i)} = p^{k+1}, \ \ (k = 0,1,2,...). \tag{5}$$

The structure of codewords resulted in binary Huffman encoding is

$$\left.\begin{array}{l} s_0 \to c_0 \to 1, \\ s_1 \to c_1 \to 01, \\ s_2 \to c_2 \to 001, \\ ..... \\ s_k \to c_k \to \underbrace{00...0}_{k}1, \\ ..... \end{array}\right\} \tag{6}$$

The length, l_k, of the codeword associated with the message, $s_k^{(t)}$, is the number of edges on the path between the root and the node $s_k^{(t)}$ in the Huffman tree

$$l_k = k + 1, \ \ (k = 0,1,2,...). \tag{7}$$

The average codeword length (Munteanu, 2007) is determined with

$$\bar{l} = \sum_{k=0}^{\infty} p_k^{(t)} l_k. \tag{8}$$

The average codeword length is obtained substituting (2) and (7) in (8)

$$\bar{l} = \sum_{k=0}^{\infty} (k+1) q p^k = \frac{1}{1-p}. \tag{9}$$

The entropy of the source with the distribution given in (2) is

$$H(\xi) = -\sum_{k=0}^{\infty} p_k^{(t)} \log p_k^{(t)}, \tag{10}$$

where the logarithm function "log" is in base 2.

Considering (2) and (10), the entropy of the source is

$$H(\xi) = -\log(1-p) - \frac{p}{1-p}\log p. \tag{11}$$

Relations (9) and (11) were obtained taking into account that $\sum_{k=0}^{\infty} p^k q = 1$ and $\sum_{k=0}^{\infty} k p^k q = \frac{p}{1-p}$.

We note that the probabilities to deliver the symbols $x_1 = 1$ or $x_0 = 0$ depend on the node from which they are generated. In other words, as a result of Huffman encoding of the source, a source $X = \{x_0 = 0, x_1 = 1\}$, with memory, is generally obtained. Its states correspond to terminal or intermediate nodes (excluding the root) in the graph in Fig. 1. When a terminal node is reached, the binary 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 this source, denoted by *X,* can be obtained from the Huffman encoding graph of the source ξ, as follows:

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

c) Each terminal or intermediate node will represent a state, $S_k^{(t)}$, or $S_k^{(i)}$, $(k = 0,1,2,...)$ (as it is represented in Fig. 2).

Let $S = \{S_0^{(t)}, S_1^{(t)}, ..., S_{k-1}^{(t)}, S_k^{(t)}, ..., S_0^{(i)}, S_1^{(i)}, ..., S_{k-1}^{(i)}, S_k^{(i)}, ...\}$ be the state set of the source with memory.

The probabilities of delivering the symbols $x_0 = 0$ or $x_1 = 1$ from the state $S_{k-1}^{(i)}$, $(k = 1,2,...)$, corresponding to an intermediate node $s_{k-1}^{(i)}$, $(k = 1,2,...)$, are equal to the probabilities of transition from the state $S_{k-1}^{(i)}, (k = 1,2,...)$, to the states $S_k^{(t)}$ and $S_k^{(i)}$, $(k = 1,2,...)$, respectively, *i.e.*

$$p(x_1 / S_{k-1}^{(i)}) = q = 1 - p, \ (k = 1,2,...), \tag{12}$$

and

$$p(x_0 / S_{k-1}^{(i)}) = p,\ (k = 1,2,...)\,, \tag{13}$$

respectively.

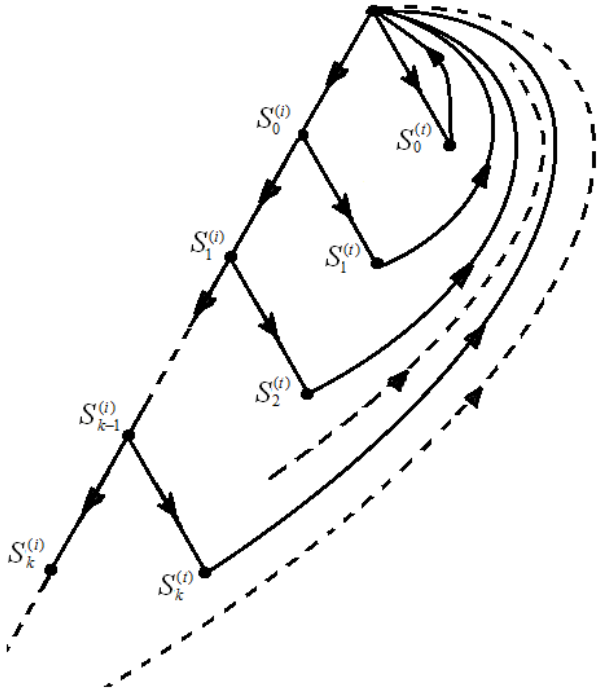

Fig. 2 – The graph of the source with memory.

The probabilities of delivering the symbols $x_1 = 1$ or $x_0 = 0$ from the state $S_k^{(t)}$, $(k = 0,1,2,...)$, corresponding to a terminal node, $s_k^{(t)}$, $(k = 0,1,2,...)$, are equal to the probabilities of transition from the state $S_k^{(t)}$, $(k = 0,1,2,...)$, in the states $S_0^{(t)}$ and $S_0^{(i)}$, respectively, *i.e.*

$$p(x_1 / S_k^{(t)}) = 1 - p,\ (k = 0,1,2,...) \tag{14}$$

and

$$p(x_0 / S_k^{(t)}) = p,\ (k = 0,1,2,...)\,, \tag{15}$$

respectively.

The transition matrix between states is

$$\mathbf{T}=\begin{array}{c} \begin{matrix} S_0^{(t)} & S_1^{(t)} & S_2^{(t)} & \cdots & S_k^{(t)} & \cdots & S_0^{(i)} & S_1^{(i)} & \cdots & S_k^{(i)} & \cdots \end{matrix} \\ \begin{bmatrix} p_0^{(t)} & 0 & 0 & \cdots & 0 & \cdots & 1-p_0^{(t)} & 0 & \cdots & 0 & \cdots \\ p_0^{(t)} & 0 & 0 & \cdots & 0 & \cdots & 1-p_0^{(t)} & 0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_0^{(t)} & 0 & 0 & \cdots & 0 & \cdots & 1-p_0^{(t)} & 0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots \\ 0 & p_1^{(t)}/p_0^{(i)} & 0 & \cdots & 0 & \cdots & 0 & p_1^{(t)}/p_0^{(i)} & \cdots & 0 & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots & p_k^{(t)}/p_{k-1}^{(i)} & \cdots & 0 & 0 & \cdots & p_k^{(t)}/p_{k-1}^{(i)} & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \end{array} \begin{matrix} S_0^{(t)} \\ S_1^{(t)} \\ \vdots \\ S_k^{(t)} \\ \vdots \\ S_0^{(i)} \\ \vdots \\ S_{k-1}^{(i)} \\ \vdots \end{matrix} . \quad (16)$$

Considering relations (12),…, (15), the transition matrix (16) becomes

$$\mathbf{T}=\begin{array}{c} \begin{matrix} S_0^{(t)} & S_1^{(t)} & S_2^{(t)} & \cdots & S_k^{(t)} & \cdots & S_0^{(i)} & S_1^{(i)} & \cdots & S_k^{(i)} & \cdots \end{matrix} \\ \begin{bmatrix} p & 0 & 0 & \cdots & 0 & \cdots & 1-p & 0 & \cdots & 0 & \cdots \\ p & 0 & 0 & \cdots & 0 & \cdots & 1-p & 0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p & 0 & 0 & \cdots & 0 & \cdots & 1-p & 0 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots \\ & 1-p & & \cdots & & \cdots & & p & \cdots & & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & & \cdots & \vdots & \cdots \\ 0 & 0 & 0 & \cdots & 1-p & \cdots & 0 & 0 & \cdots & p & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \end{array} \begin{matrix} S_0^{(t)} \\ S_1^{(t)} \\ \vdots \\ S_k^{(t)} \\ \vdots \\ S_0^{(i)} \\ \vdots \\ S_{k-1}^{(i)} \\ \vdots \end{matrix} . \quad (17)$$

Let $\pi_k^{(t)}$ and $\pi_k^{(i)}$, $(k=0,1,2,...)$, denote the state probabilities corresponding to terminal or intermediate nodes. They can be determined by means of relation (Munteanu *et. al.,* 2007; Cover, 1991)

$$[\pi_0^{(t)}\ \pi_1^{(t)}...\pi_k^{(t)}...\pi_0^{(i)}\ \pi_1^{(i)}...\pi_k^{(i)}...]=[\pi_0^{(t)}\ \pi_1^{(t)}...\pi_k^{(t)}...\pi_0^{(i)}\ \pi_1^{(i)}...\pi_k^{(i)}...]\mathbf{T}, \quad (18)$$

$$\sum_{k=0}^{\infty}\left(\pi_k^{(t)}+\pi_k^{(i)}\right)=1. \quad (19)$$

Considering (8) and (17), from (18) and (19) we get the state probabilities

$$\pi_k^{(t)} = \frac{1}{\bar{l}} p_k^{(t)},\ (k = 0,1,2,...)\,, \tag{20}$$

$$\pi_k^{(i)} = \frac{1}{\bar{l}} p_k^{(i)} = \frac{1}{\bar{l}}\left(1 - \sum_{j=0}^{k} p_j^{(t)}\right). \tag{21}$$

Substituting (2) and (4) in (20) and (21), we get the state probabilities

$$\pi_k^{(t)} = (1-p)^2 p^k,\ (k = 0,1,2,...)\,, \tag{22}$$

$$\pi_k^{(i)} = (1-p) p^{k+1},\ (k = 0,1,2,...)\,. \tag{23}$$

Generally, the entropy of the source with memory is computed with (Cover, 1991)

$$H(X) = -\sum_{k=0}^{\infty}\sum_{j=0}^{1} \pi_k^{(t)} p\left(x_j \middle| S_k^{(t)}\right) \log p\left(x_j \middle| S_k^{(t)}\right) - \sum_{k=0}^{\infty}\sum_{j=0}^{1} \pi_k^{(i)} p\left(x_j \middle| S_k^{(i)}\right) \log p\left(x_j \middle| S_k^{(i)}\right). \tag{24}$$

Substituting (12),…,(15), (22) and (23) in (24), we get the entropy of the source with memory

$$H(X) = -(1-p)\left[\log(1-p) + \frac{p}{1-p}\log p\right]. \tag{25}$$

From (7), (11) and (25) we see that *H(X),* the average information per symbol, is the ratio between the source entropy and the average codeword length

$$H(X) = \frac{H(\xi)}{\bar{l}}\,. \tag{26}$$

Let c_0 and c_1 be the costs associated to the bits 0 and 1, respectively. The average cost of a code is defined by (Mohajer *et. al.,* 2011)

$$\overline{C} = \sum_{k=0}^{\infty} p_k^{(t)} \left(n_0(k)c_0 + n_1(k)c_1\right), \tag{27}$$

where we denote by $n_0(k)$ and $n_1(k)$ the number of 0's and 1's in the codeword corresponding to the source symbol, $s_k^{(t)}$.

Considering (6), the average cost is

$$\overline{C} = \sum_{k=0}^{\infty} p_k^{(t)} \left(kc_0 + c_1 \right). \tag{28}$$

We obtain the average cost of the AUH code for the source with geometric distribution, substituting (2) in (28)

$$\overline{C} = c_0 \frac{p}{1-p} + c_1. \tag{29}$$

If we consider that the state set, *S*, of the source with memory, represents the field at a discrete memoryless channel input and the symbols generated by the source with memory represents the field at the channel output, from (24) it results that the entropy of the source with memory represents the mean error of the channel with input, *S*, and output, *X*, that is, *H(X|S)*.

Making use of this analogy, we can calculate for sources with memory the information quantities specific to discrete memoryless channels. One of them, corresponding to mutual information, indicates whether the source is with memory or not, as it is different from zero or equal to zero

$$I(X,S) = \sum_{j=0}^{1} \sum_{k=0}^{\infty} p(x_j, S_k^{(t)}) \frac{\log p(x_j, S_k^{(t)})}{p(x_j)\pi_k^{(t)}} + \sum_{j=0}^{1} \sum_{k=0}^{\infty} p(x_j, S_k^{(i)}) \frac{\log p(x_j, S_k^{(i)})}{p(x_j)\pi_k^{(i)}}. \tag{30}$$

We get the joint probabilities as

$$p(x_j, S_k^{(t)}) = \pi_k^{(t)} p(x_j \mid S_k^{(t)}), \tag{31}$$

$$p(x_j, S_k^{(i)}) = \pi_k^{(i)} p(x_j \mid S_k^{(i)}). \tag{32}$$

Substituting (12),…,(15), (22) and (23) in (31) and (32), we get the joint probabilities

$$p(x_1, S_k^{(t)}) = (1-p)^3 p^k, \ (k = 0,1,2,...), \tag{33}$$

$$p(x_1, S_k^{(i)}) = (1-p)^2 p^k, \ (k = 1,2,...), \tag{34}$$

$$p(x_0, S_k^{(t)}) = (1-p)^2 p^{1+k}, \ (k = 0,1,2,...), \tag{35}$$

$$p(x_0, S_k^{(i)}) = (1-p) p^{1+k}, \ (k = 1,2,...). \tag{36}$$

We compute the symbol probabilities as

$$p(x_j) = \sum_{k=0}^{\infty} p(x_j, S_k^{(t)}) + \sum_{k=0}^{\infty} p(x_j, S_k^{(i)}), \ (j = 0,1). \tag{37}$$

Substituting the probabilities (33),…,(36) in (37), we get the probabilities

$$p(x_0) = p, \tag{38}$$

$$p(x_1) = 1 - p\,. \tag{39}$$

Substituting (22), (23), (33),…,(36), (38) and (39) in (30), it results

$$I(S, X) = 0. \tag{40}$$

It is useful to observe that this quantity is equal to zero. This indicates that the source resulted by binary encoding of the source with geometric distribution is memoryless.

3. Conclusions

In this paper we have considered an infinite discrete memoryless AUH source with geometric distribution. Performing a binary Huffman encoding of this source, we get, in general, a source with memory, because the probabilities of delivering the symbols $x_0 = 0$ and $x_1 = 1$ in the encoding process depend on the nodes in the graph from where they are generated. The graph of the source with memory is obtained from the encoding graph by linking the terminal nodes with the graph root. The states of the source with memory correspond to the terminal or intermediate nodes in the encoding graph. We determined the state probabilities of the source with memory, as well as the transition probabilities between states. The average information and cost per binary symbol in encoding process are computed. As the entropy of the source that is to be encoded measures the average information per codeword, and the code entropy measures the average information per symbol, their ratio represents the average length of codewords.

Performing the analogy between discrete sources with memory and discrete memoryless channels, we compute the information quantity $I(X,S)$, which indicates whether the source resulted by binary Huffman encoding is with memory or not.

REFERENCES

Capocelli R.M., De Santis A., *A Note on D-ary Huffman Codes*. IEEE Trans. Inf. Theory, **37**, 174-179 (1991).

Corduneanu S.O., *Chapters of Mathematical Analysis* (in Romanian). MatrixRom, Bucharest, 2009.

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

Esmaeili M., Kakhbod A., *On Antiuniform and Partially Antiuniform Sources*. Proc. IEEE ICC, 1611-1615 (2006).

Esmaeili M., Kakhbod A., *On Information Theory Parameters of Infinite Anti-Uniform Sources*. IET Commun., **1**, 1039-1041 (2007).

Gallager R., *Variations by a Theme of Huffman*. IEEE Trans. Inf. Theory, **24**, 668-674 (1978).

Gallager R., Van Voorhis D., *Optimal Source Coding for Geometrically Distributed Integer Alphabets*. IEEE Trans. Inf. Theory, **21**, *2*, 228-230 (1975).

Gray R. M., *Source Coding Theory*. Kluwer, Boston, 1990.

Huffman R., *A Method for the Construction of Minimum-Redundancy Codes*. Proc. IRE, **40**, 1098 – 1101 (1952).

Humblet P., *Optimal Source Coding for a Class of Integer Alphabets*. IEEE Trans. Inf. Theory, **24**, *1*, 110-112 (1978).

Johnsen O., *On the Redundancy of Binary Huffman Codes*. IT, **26**, 220-222 (1980).

Khorsavifard M., Esmaeili M., Saidi H., Gulliver T.A., *A Tree Based Algorithm for Generating all Possible Binary Compact Codes with N Codewords*. IEICE Trans. Fundam. Electron. Commun. Comp. Sci., **10**, 2510-2516 (2003).

Larsen R.J., Marx M.L., *An Introduction to Mathematical Statistics and its Applications*. Upper Saddle River, Prentice Hall, NJ, 2001.

Linder T., Tarokh V., Zeger K., *Existence of Optimal Prefix Codes for Infinite Source Alphabets*. IEEE Trans. Inf. Theory, **43**, 2026-2028 (1997).

Mohajer S., Pakzad P., Kakhbod A., *Tight Bounds on the Redundancy of Huffman Codes*. Proc. IEEE ITW, 131-135 (2006).

Mohajer S., Kakhbod A., *Anti-Uniform Huffman Codes*. IET Commun., **5**, *9*, 1213-1219 (2011).

Munteanu V., Tărniceriu D., *Elements of Information Theory* (in Romanian). Cermi, Iaşi, 2007.

SURSE ANTI-UNIFORME INFINITE CU DISTRIBUȚIE GEOMETRICĂ

(Rezumat)

S-a analizat clasa codurilor Huffman antiuniforme pentru surse caracterizate de o distribuție geometrică, cu alfabet infinit.Codarea Huffman a acestor surse conduce la coduri AUH. Ca urmare a acestei codări se obțin, in general, surse cu memorie. Pentru aceste surse s-a calculat entropia şi costul mediu. S-a efectuat o analogie între sursele discrete cu memorie şi canalele discrete fără memorie, arătându-se că entropia sursei cu memorie este similară cu eroarea medie din cazul canalului discret fără memorie. Mărimea informațională, $I(X,S)$, indică pentru codurile AUH, dacă acestea sunt sau nu cu memorie, după cum această mărime este diferită de zero sau nu.