Contemporary Engineering Sciences, Vol. 1, 2008, no. 4, 169 - 175

On Finite Antiuniform Probability Distributions

M. Esmaeili

Department of Mathematical Sciences
Isfahan University of Technology
84156-8311, Isfahan, Iran
emorteza@cc.iut.ac.ir

A. Kakhbod

Dept. of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA
akakhbod@umich.edu

T.A. Gulliver

Dept. of Electrical and Computer Engineering
University of Victoria
P.O. Box 3055, STN CSC
Victoria, B.C., Canada V8W 3P6
agullive@ece.uvic.ca

Abstract

A probability distribution $\{p_1, p_2, \ldots, p_n\}$ satisfying $p_{i+2} + p_{i+3} + \ldots + p_n \leq p_i,\ 1 \leq i \leq n-3$, is called an antiuniform source. It is shown that the geometric, negative binomial, quasi-geometric, Poisson and exponential distributions lie in the class of antiuniform sources. A linear representation of finite antiuniform sources and their i-dimensional, $2 \leq i \leq n-1$, Euclidean projection is given. The i-dimensional, $2 \leq i \leq n-1$, projection determines the simplex consisting of the i-tuples $(p_1, p_2, \ldots, p_i)$ for which there exists an antiuniform source $(p_1, p_2, \ldots, p_n)$. This linear representation along with linear programming can be used to compute the maximum expected codeword length.

Keywords: Antiuniform distributions; Linear representation; Simplex

1 Introduction

Let $(p_1, p_2, \ldots, p_n)$ be the probability distribution of a given n-symbol source $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$. It is known that the Huffman coding algorithm [1] [3] produces an optimal binary code (a code with minimum average codeword length) for $\mathcal{S}$. A Huffman code is usually represented by the leaves of a tree, called a Huffman tree [1]. The length between a leaf and the root is the length of the binary codeword associated with the corresponding symbol. In this paper by a Huffman code we mean a *binary* Huffman code.

Assuming that c_i is the codeword representing symbol s_i, we denote the length of c_i by l_i. The optimality of Huffman coding implies that $l_i \leq l_j$ if $p_i > p_j$. For a given source $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ we suppose that $p_1 \geq p_2 \geq \ldots \geq p_n > 0$. In this paper we study the class of finite *antiuniform* sources which are defined below.

Definition 1.1 *A Huffman code is called uniform if $|l_i - l_j| \leq 1$ for any two symbols s_i and s_j. We call a Huffman code $\mathcal{C}$ representing a finite source $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, satisfying $p_1 \geq p_2 \geq \ldots \geq p_n$, an antiuniform Huffman code if $l_1 = 1$, $l_2 = 2$, ..., $l_{n-2} = n-2$ and $l_{n-1} = l_n = n-1$. A source $\mathcal{S}$ having an antiuniform code (AC) is called an antiuniform source (AS).*

In Section 2 we characterize the class of antiuniform sources and introduce several classes of such sources. In Section 3 the i-dimensional Euclidean projection, $i \leq |\mathcal{S}| - 1$, of these sources is given in terms of a system of linear inequalities. For $i < |\mathcal{S}| - 1$, the system consists of $i + 1$ constraints while for $i = n - 1$ (when $|\mathcal{S}| = n$), there are $n - 1$ constraints. In Section 4 we apply linear programming to the linear representation of antiuniform sources to determine the maximum expected codeword length of n-symbol antiuniform sources.

2 Antiuniform Distributions

In this section we first give a simple criteria for a source $\mathcal{S}$ to be antiuniform, and then study the i-dimensional Euclidean projection of these distributions.
Characterization The Huffman coding algorithm results in the following characterization for antiuniform sources.

Theorem 2.1 *A necessary and sufficient condition for an n-symbol source $\mathcal{S}$ with distribution $p_1 \geq p_2 \geq \ldots \geq p_n$ to be an AS is*

$$p_{i+2} + p_{i+3} + \ldots + p_n \leq p_i, \qquad 1 \leq i \leq n-3. \tag{1}$$

Proof The Huffman coding algorithm is a bottom to top reduction process and the codeword length l_i corresponding to source symbol s_i is equal to the number of times s_i is amalgamated during the reduction process. ■

According to this characterization, an infinite alphabet source $\mathcal{S}$ with distribution $p_1 \geq p_2 \geq \ldots$ is also called antiuniform if and only if

$$p_{i+2} + p_{i+3} + \ldots = \sum_{k=i+2}^{\infty} p_k \leq p_i, \qquad i \geq 1. \tag{2}$$

We should mention that if $p_1 \geq p_2 \ldots$ is the distribution of an infinite antiuniform source then $(p_1,\ p_2, \ldots, p_{n-1},\ p'_n := \sum_{t=n}^{\infty} p_t)$ represents a finite AS.

Example 2.2 *Many well known distributions lie in the class of antiuniform sources. It is known that Poisson distributions [4] with parameter $\lambda \leq 1$ and geometric distributions [2] $p_i = (1-\theta)\theta^i$, $i \geq 0$, with $0 < \theta \leq \frac{\sqrt{5}-1}{2}$ are among the class of infinite alphabet antiuniform sources. In the following we show that the discrete form of the exponential distribution is also antiuniform.*

Consider the exponential distribution defined by the probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Define $F(a) = P(X \leq a) = 1 - e^{-\lambda a}$, $a \geq 0$, and $p_i = F(i) - F(i-1)$ for $i \geq 1$. It is easy to see that $p_i = e^{-\lambda i}(e^{\lambda} - 1)$. We also have $p_i + p_{i+1} + \ldots = (e^{\lambda} - 1)e^{-\lambda i}(1 + e^{-\lambda} + e^{-2\lambda i} + \ldots) = e^{-\lambda i}\frac{e^{\lambda}-1}{1-e^{-\lambda}}$. Thus, in this case, condition (2) is equivalent to $e^{-\lambda i}\frac{e^{\lambda}-1}{1-e^{-\lambda}} \leq e^{-\lambda(i-2)}(e^{\lambda} - 1) = p_{i-2}$, that is $1 \leq e^{2\lambda} - e^{\lambda}$ which is true if and only if $e^{\lambda} \geq \frac{1+\sqrt{5}}{2}$. Therefore, the discrete form of the exponential distribution is antiuniform if and only if $\lambda \geq ln(\frac{1+\sqrt{5}}{2}) \simeq 0.4812$.

A negative binomial random variable [5] with parameters (r, p) is defined by $P(X = n) = \binom{n-1}{r-1}(1-p)^r p^{n-r}$, where $1-p$ is the probability of success of a Bernoulli random variable Y, and $P(n)$ is the probability of accumulation of a total of r successes in n independent trials of Y assuming that the outcome of the last trial is a success. It is straightforward to check that for $r = 2$, relation (2) holds if and only if $k(p^3 - 2p + 1) \geq 1$, $k \geq 2$. Thus the negative binomial distribution with parameters $(2, p)$ is antiuniform if and only if $p^3 - 2p + \frac{1}{2} \geq 0$, that is $0 < p < a$ where $a \simeq 0.258652$.

Consider the quasi-geometric series $\sum_{n=1}^{\infty} nx^n = \frac{x}{(1-x)^2}$, [6]. For $0 < x < 1$, this series produces the quasi-geometric distribution $\{nx^{n-1}(1-x)^2\}_{n=1}$. It is easy to show that $\sum_{j=k+2}^{\infty} jx^j \leq kx^k$, $k \geq 1$, if and only if $(k+1)x^3 - 2x^2 - 2kx + k \geq 0$. For $k = 1$ and $|x| \leq 1$ we have $(k+1)x^3 - 2x^2 - 2kx + k \geq 0$ if $-0.8546 \leq x \leq 0.403$. We also have $(k+2)x^3 - 2x^2 - 2(k+1)x + (k+1) \geq (k+1)x^3 - 2x^2 - 2kx + k$ if and only if $x^3 - 2x + 1 \geq 0$, and this holds for $-1 \leq x \leq 0.618$. Therefore, for $0 < x \leq 0.403$ the quasi-geometric distribution $\{nx^{n-1}(1-x)^2\}_{n=1}$ is an antiuniform distribution.

For the quasi-geometric series $\sum_{n=1}^{\infty} n^2x^n = \frac{x(1+x)}{(1-x)^3}$, we have $\sum_{j=k+2}^{\infty} j^2x^j \leq k^2x^k$, $k \geq 1$, if and only if $f(k, x) := (k+1)^2x^4 - (k^2 + 6k + 3)x^3 + (-2k^2 +$

$2k+4)x^2+3k^2x-k^2 \leq 0$. *Setting* $k=1$, *this relation reduces to* $4x^4-10x^3+4x^2+3x-1 \leq 0$, *and this holds for* $-0.5 \leq x \leq 0.2928932$. *On the other hand,* $f(k+1,x) \leq f(k,x)$ *if and only if* $(2k+3)x^4-(2k+7)x^3+(6k+3)x-(2k+1) \leq 0$. *It is obvious that this inequality holds if* $0 < x \leq \frac{1}{3}$. *Therefore, if* $0 < x \leq 0.2928932$ *the quasi-geometric distribution* $\left\{\frac{n^2x^{n-1}(1-x)^3}{1+x}\right\}_{n=1}$ *is an antiuniform distribution.*

Note that if equality holds in (1), the source will also have non-antiuniform codes. For instance, the distribution $(1/3, 1/3, 1/6, 1/6)$ has both the AC $l_1 = 1,\ l_2 = 2,\ l_3 = l_4 = 3$ and the non-antiuniform code $l_1 = l_2 = l_3 = l_4 = 2$. The reason for this is that $p_4 + p_3 = p_1$.

3 Projection

Let $\mathcal{S}$ be a 5-symbol AS with distribution $p_1 \geq p_2 \geq p_3 \geq p_4 \geq p_5 > 0$. Then

$$\begin{cases} p_5+p_4+p_3 \leq p_1 \\ p_5+p_4 \leq p_2 \\ 0 < p_5 \leq p_4 \leq p_3 \leq p_2 \leq p_1 \\ p_5+p_4+p_3+p_2+p_1 = 1. \end{cases} \tag{3}$$

Relation $p_5+p_4+p_3 = 1-(p_1+p_2) \leq p_1$ is equivalent to $2p_1+p_2 \geq 1$. The condition $p_5+p_4 \leq p_2$ is equivalent to $1-(p_1+p_2+p_3) \leq p_2$ and thus is equivalent to $p_1+2p_1+p_3 \geq 1$. On the other hand, $0 < p_5 \leq p_4$ if and only if $1-(p_1+p_2+p_3) = p_4+p_5 \leq 2p_4$, and hence $p_1+p_2+p_3+2p_4 \geq 1$. Therefore, the system of inequalities given by (3) holds if and only if (p_1, p_2, p_3, p_4) satisfies the following system of equations

$$\begin{cases} 2p_1+p_2 \geq 1 \\ p_1+2p_2+p_3 \geq 1 \\ p_1+p_2+p_3+2p_4 \geq 1 \\ 0 < p_4 \leq p_3 \leq p_2 \leq p_1 \\ p_4+p_3+p_2+p_1 < 1. \end{cases} \tag{4}$$

It follows from $p_1+p_2+p_3+2p_4 \geq 1$ and $0 < p_4 \leq p_3$ that (4) holds if and only if (p_1, p_2, p_3) satisfies the following system of equations

$$\begin{cases} 2p_1+p_2 \geq 1 \\ p_1+2p_2+p_3 \geq 1 \\ p_1+p_2+3p_3 \geq 1 \\ 0 < p_3 \leq p_2 \leq p_1 \\ p_3+p_2+p_1 < 1. \end{cases} \tag{5}$$

Conditions $p_1 + 2p_2 + p_3 \geq 1$ and $p_3 \leq p_2$ in (5) imply that (5) holds if and only if (p_1, p_2) satisfies

$$\begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + 3p_2 \geq 1 \\ 0 < p_2 \leq p_1 \\ p_2 + p_1 < 1. \end{cases} \tag{6}$$

Therefore, the set of 4-tuples (p_1, p_2, p_3, p_4), 3-tuples (p_1, p_2, p_3) and 2-tuples (p_1, p_2) for which there exist 5-symbol antiuniform sources are specified by (4), (5) and (6), respectively. In fact these are the simplexes representing the 2-, 3- and 4-dimensional Euclidean projections of 5-symbol antiuniform sources.

Applying the same argument to (1), we obtain the following.

Theorem 3.1 *Given an n-symbol antiuniform source $\mathcal{S}$, the i-dimensional projection, $1 \leq i \leq n-1$, of $\mathcal{S}$ is given by the following systems of linear inequalities.*

$$\begin{array}{cc} i - \text{Dimensional Projection} & (n-1) - \text{Dimensional Projection} \\ 1 \leq i \leq n-2 & \\ \begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + 2p_2 + p_3 \geq 1 \\ p_1 + p_2 + 2p_3 + p_4 \geq 1 \\ \vdots \\ p_1 + \ldots + p_{i-2} + 2p_{i-1} + p_i \geq 1 \\ p_1 + \ldots + p_{i-1} + 3p_i \geq 1 \\ p_1 + p_2 + \ldots + p_i < 1 \\ p_i \leq \ldots \leq p_2 \leq p_1. \end{cases} & \begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + 2p_2 + p_3 \geq 1 \\ p_1 + p_2 + 2p_3 + p_4 \geq 1 \\ \vdots \\ p_1 + \ldots + p_{n-4} + 2p_{n-3} + p_{n-2} \geq 1 \\ p_1 + \ldots + p_{n-3} + p_{n-2} + 2p_{n-1} \geq 1 \\ p_1 + p_2 + \ldots + p_{n-1} < 1 \\ p_{n-1} \leq \ldots \leq p_2 \leq p_1. \end{cases} \end{array} \tag{7}$$

Conversely, any i-tuple $(p_1, \ldots, p_i)$ satisfying these constraints can be extended to an n-tuple $(p_1, \ldots, p_n)$ representing an AS. ■

Example 3.2 *The set of 2-tuples (p_1, p_2) and 3-tuples (p_1, p_2, p_3) for which there exist 4-symbol antiuniform sources are specified by the following systems of linear inequalities*

$$\begin{array}{cc} 2 - \textit{Dimensional Projection} & 3 - \textit{Dimensional Projection} \\ \begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + 3p_2 \geq 1 \\ p_1 + p_2 < 1 \\ p_2 \leq p_1. \end{cases} & \begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + p_2 + 2p_3 \geq 1 \\ p_1 + p_2 + p_3 < 1 \\ p_3 \leq p_2 \leq p_1. \end{cases} \end{array} \tag{8}$$

Fig. 1 shows these two sets. The 3-dimensional projection is the simplex with vertices $(\frac{2}{5}, \frac{1}{5}, \frac{1}{5})$, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{6})$, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(1, 0, 0)$.

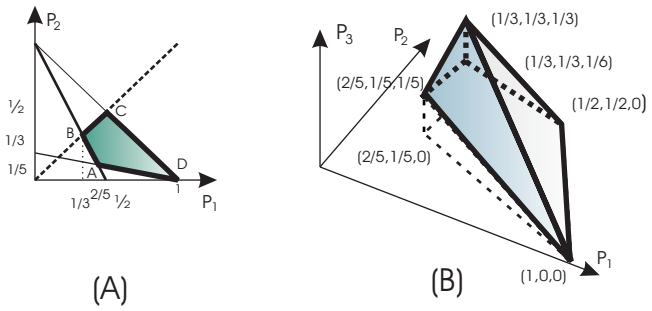

Figure 1: The set of 2-tuples (p_1, p_2) (A) and 3-tuples (p_1, p_2, p_3) (B), for which there exist 4-symbol antiuniform sources.

4 Maximum Expected Codeword Length

Consider an n-symbol antiuniform distribution $p_1 \geq p_2 \geq \ldots \geq p_{n-1} \geq p_n$. It is obvious that the expected codeword length of this distribution is

$$\begin{aligned} L &= p_1 + 2p_2 + \ldots + (n-2)p_{n-2} + (n-1)p_{n-1} + (n-1)p_n \\ &= p_1 + 2p_2 + \ldots + (n-2)p_{n-2} + (n-1)(p_{n-1} + p_n) \\ &= (n-1) - \{(n-2)p_1 + (n-3)p_2 + \ldots + 2p_{n-3} + p_{n-2}\}. \end{aligned}$$

Therefore, maximizing L is equivalent to minimizing $Z := (n-2)p_1 + (n-3)p_2 + \ldots + 2p_{n-3} + p_{n-2}$ over the $(n-2)$-dimensional Euclidean projection of the set of n-symbol antiuniform sources given by (7). Hence the problem is reduced to minimizing $Z = \sum_{i=2}^{n-1}(n-i)p_{i-1}$ subject to the following constraints:

$$\begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + 2p_2 + p_3 \geq 1 \\ p_1 + p_2 + 2p_3 + p_4 \geq 1 \\ \vdots \\ p_1 + p_2 + \ldots + p_{n-4} + 2p_{n-3} + p_{n-2} \geq 1 \\ p_1 + p_2 + \ldots + p_{n-3} + 3p_{n-2} \geq 1 \\ p_1 + p_2 + \ldots + p_{n-2} < 1 \\ p_1 \geq p_2 \geq \ldots \geq p_{n-2}. \end{cases}$$

Example 4.1 *For* 6*-symbol sources we have the problem of minimizing* $Z = 4p_1 + 3p_2 + 2p_3 + p_4$ *subject to:*

$$\begin{cases} 2p_1 + p_2 \geq 1 \\ p_1 + 2p_2 + p_3 \geq 1 \\ p_1 + p_2 + 2p_3 + p_4 \geq 1 \\ p_1 + p_2 + p_3 + 3p_4 \geq 1 \\ p_1 + p_2 + p_3 + p_4 < 1 \\ p_1 \geq p_2 \geq p_3 \geq p_4 \geq 0. \end{cases}$$

The corresponding solution is $Z = \frac{34}{13},\ p_1 = \frac{5}{13},\ p_2 = \frac{3}{13},\ p_3 = \frac{2}{13},\ p_4 = \frac{1}{13}$. This means that the expected codeword length of any 6-symbol antiuniform source is bounded above by $L = 5 - Z = \frac{31}{13}$, and this bound is achieved by the distribution $(p_1, p_2, \ldots, p_6) = (\frac{5}{13}, \frac{3}{13}, \frac{2}{13}, \frac{1}{13}, \frac{1}{13}, \frac{1}{13})$. Note that for this case we have

$$p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 5p_6 = \frac{5}{13} + \frac{6}{13} + \frac{6}{13} + \frac{4}{13} + \frac{5}{13} + \frac{5}{13} = \frac{31}{13}$$

which is consistent with $L = 5 - Z = \frac{31}{13}$.

Since the entropy H of a source with expected codeword length L satisfies $H \leq L < H + 1$, it follows from the result given above that the entropy of any 6-symbol antiuniform distribution is bounded above by $\frac{31}{13}$.

5 Summary

Several classes of antiuniform probability distributions were introduced and characterized by linear models. These models determine the i-dimensional simplexes consisting of the i-tuples $(p_1, p_2, \ldots, p_i)$ for which there exists an antiuniform source $(p_1, p_2, \ldots, p_n)$. As an application, these models together with linear programming can be used to determine the maximum expected codeword length of n-symbol antiuniform distributions, and hence a bound on their entropy.

References

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

[2] R.G. Gallager and D.C. Van Voorhis, Optimal source coding for geometrically distributed integer alphabets, *IEEE Trans. Inform. Theory* 21 (1975), 228–230.

[3] D.A. Huffman, A method for the construction of minimum-redundancy codes, *Proc. Inst. Electr. Radio Eng.*, 40 (1952), 1098–1101.

[4] P.A. Humblet, Optimal source coding for a class of integer alphabets, *IEEE Trans. Inform. Theory*, 24 (1978), 110–112.

[5] S. Ross, *A First Course in Probability* Prentice Hall, Upper Saddle River, New Jersey, 1998.

[6] A.C. Segal, Closed-form formulas for quasi-geometric series, *The Two-Year College Mathematics Journal*, 14 (1983), 118–122.

Received: January 9, 2008