Published in IET Communications
Received on 16th January 2010
Revised on 30th May 2010
doi: 10.1049/iet-com.2010.0049

ISSN 1751-8628

Anti-uniform Huffman codes

S. Mohajer[1] *A. Kakhbod*[2]

[1]*School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland*
[2]*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA*
E-mail: akakhbod@umich.edu

Abstract: In this study, the authors consider the class of anti-uniform Huffman (AUH) codes. The authors derived tight lower and upper bounds on the average codeword length, entropy and redundancy of finite and infinite AUH codes in terms of the alphabet size of the source. These bounds are tighter than similar bounds. Also a tight upper bound on the entropy of AUH codes is presented in terms of the average cost of the code. The Fibonacci distribution is introduced, which plays a fundamental role in AUH codes. It is shown that such distributions maximise the average length and the entropy of the code for a given alphabet size. The authors also show that the minimum average cost of a code is achieved by an AUH codes in a highly unbalanced cost regime.

1 Introduction

Consider a discrete source with finite-size alphabet $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and associated ordered probability distribution $\mathcal{P} = (p_1, p_n, \ldots, p_n)$, where $p_1 \geq p_2 \geq \cdots \geq p_n$. It is well known that the Huffman encoding algorithm [1] provides an optimal prefix-free (A prefix-free code is a code, typically a variable-length code, with the 'prefix property': there is no valid codeword in the code that is a prefix (start) of any other valid codeword in the set.) code for this source. A binary Huffman code is usually represented using a binary tree $\mathcal{T}$, whose leaves correspond to the source symbols; the two edges emanating from each intermediate node of $\mathcal{T}$ are labelled either 0 and 1, and the codeword corresponding to a symbol is the string of labels on the path from the root to the corresponding leaf. Huffman's algorithm is a recursive bottom-up construction of $\mathcal{T}$, where at each time the two smallest probabilities are merged into a new unit, and henceforth represented by an intermediate node in the tree.

We denote by ℓ_i the length of the codeword associated with the symbol s_i that is the number of edges on the path between the root and the node s_i on the Huffman tree. Then, the expected length of the Huffman code is defined by

$$L(\mathcal{P}) = \sum_{i=1}^{n} p_i \ell_i \tag{1}$$

Similarly, the entropy of the source is defined as

$$H(\mathcal{P}) = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

where all the logarithms in this paper are in base 2. The redundancy $R(\mathcal{P})$ of the code is defined as the difference between the average codeword length $L(\mathcal{P})$, and the entropy $H(\mathcal{P})$ of the source. It is well known that the redundancy of the Huffman code is always non-negative and never exceeds 1.

Consider a binary storage system where there is a constant cost of restoring for each bit. A similar scenario can be considered in a transmission system (e.g. an optic system) where the cost of sending bits '0' and '1' are different. The average cost of the code is an objective function that has to be minimised in such situations. Let c_0 and c_1 be the positive associated cost for the bits 0 and 1, receptively. The average cost of a code is defined by

$$C(\mathcal{P}, \mathcal{C}) = \sum_{i=1}^{n} p_i (n_0(i)c_0 + n_1(i)c_1) \tag{3}$$

where by $n_0(i)$ and $n_1(i)$ we denote the number of 0's and 1's in the codeword corresponding to the source symbol s_i.

The Huffman encoding algorithm is optimal in the sense that no other uniquely decipherable code for distribution $\mathcal{P}$ can have a smaller expected length than $L(\mathcal{P})$. Also this algorithm is optimal to find codes with minimum average cost where $c_0 = c_1$, that is the balanced cost situation.

In contrast with the uniform Huffman code wherein $|\ell_i - \ell_j| \leq 1$, a finite code (source) is called anti-uniform Huffman (AUH) [2, 3] if $\ell_i = i$ for $i = 1, \ldots, n-1$ and $\ell_n = n - 1$. Fig. 1 shows the structure of an AUH tree. Such sources can be generated by several probability distributions. It has been shown in [4] that the normalised tail of the Poisson distribution satisfies AUH structure. These kinds of distributions are also considered by Kato *et al.* [5] and in particular, it is shown that the geometric

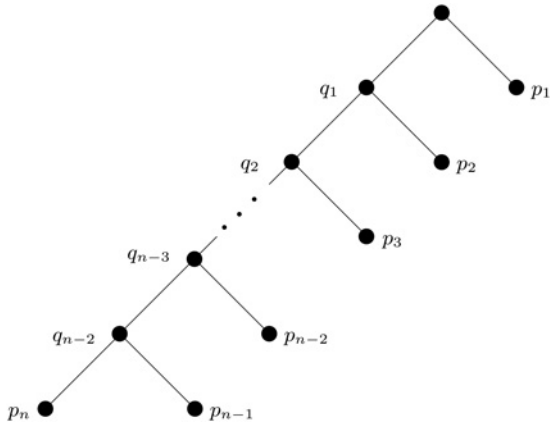

Fig. 1 *AUH tree for a source with n symbols*

distribution with success probability greater than some critical value satisfies AUH condition.

As mentioned above, these sources are represented by a special type of Huffman trees. Apart from this interesting graphical structure, we should note that these sources appear in a wide variety of situations in the real word. It is shown that the geometric, negative binomial, quasi-geometric, Poisson and exponential distributions lie in the class of AUH sources for some regimes of their parameters [2, 3, 4, 6]. Furthermore, the class of AUH sources are also known for their property of achieving minimum redundancy in different situations. It has been shown in [7] 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. A similar result by Capocelli and Santis [8] show that AUH structure achieves the minimum redundancy of a Huffman code when p_n, the probability of the least likely symbol, is known. Moreover, AUH codes are efficient codes with minimal average cost in highly unbalanced cost regime among all prefix-free codes.

In this paper we consider the AUH structure and derive tight bounds on the average codeword length, entropy and redundancy of such codes in terms of n, the alphabet size of the sources. A summary of the obtained tight bounds is provided in Table 1. The rest of the paper is organised as follows. We will start with preliminaries in the next section. A characterisation of probability distributions lead to an AUH code is presented in Section 3. Then we state and prove our bound on the average length, entropy and redundancy of AUH codes in Sections 4–6, respectively. We also show that AUH codes are optimal to achieve the minimum average cost in highly unbalanced regime, where $c_1 \gg c_0$ (or $c_0 \gg c_1$). A connection between the average cost and this class of codes is presented in Section 7. We finally conclude the paper in Section 8.

2 Preliminaries

One can simply define the probability (This is in fact the weight of the intermediate node, since unlike a leaf, an intermediate node is not corresponding to a source symbol, and therefore no probability mass is associated. However, with slight abuse we may call it probability.) of an intermediate node in the Huffman tree as the sum of the probabilities of the leaves lying under it. In an AUH tree of a source with n symbols (see Fig. 1), there are $n-2$ intermediate nodes which are labelled as $q_1, \ldots, q_{n-2}$. We denote the part of a Huffman tree lying under any intermediate node u, by Δ_u (see Fig. 2). It is clear that Δ_u is a subtree that satisfies the Huffman structure, unless the probability of the root, which is not 1. So, by normalising the probabilities of all the leaves by u, the probability of the intermediate node, we obtain a new Huffman tree that is denoted by $u^{-1*}\Delta_u$.

On the other hand, we can merge all the leaves lying in a subtree Δ_u in $\mathcal{T}$ and obtain a new Huffman tree that is denoted by Γ_u.

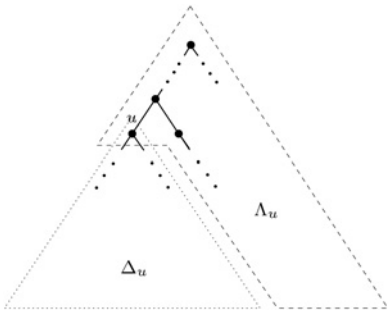

Fig. 2 *Decomposition of a Huffman tree around an intermediate node u*

Table 1 Summary of the obtained tight bounds

Parameter	Bounds	Achieving distribution
average length	$L_n^{\min} = 1$	$\left(1-\varepsilon, \frac{\varepsilon}{2}, \frac{\varepsilon}{2^2}, \ldots, \frac{\varepsilon}{2^{n-3}}, \frac{\varepsilon}{2^{n-2}}, \frac{\varepsilon}{2^{n-2}}\right)$
	$L_n^{\max} = \frac{f_{n+3}-3}{f_{n+1}}$	$\left(\frac{f_{n-1}}{f_{n+1}}, \frac{f_{n-2}}{f_{n+1}}, \ldots, \frac{f_3}{f_{n+1}}, \frac{f_2}{f_{n+1}}, \frac{f_1}{f_{n+1}}, \frac{f_2}{f_{n+1}}\right)$
entropy	$H_n^{\min} = 0$	$\left(1-\varepsilon, \frac{\varepsilon}{2}, \frac{\varepsilon}{2^2}, \ldots, \frac{\varepsilon}{2^{n-3}}, \frac{\varepsilon}{2^{n-2}}, \frac{\varepsilon}{2^{n-2}}\right)$
	$H_n^{\max} = \log f_{n+1} - \sum_{i=1}^{n-1} \frac{f_i}{f_{n+1}} \log f_i$	$\left(\frac{f_{n-1}}{f_{n+1}}, \frac{f_{n-2}}{f_{n+1}}, \ldots, \frac{f_3}{f_{n+1}}, \frac{f_2}{f_{n+1}}, \frac{f_1}{f_{n+1}}, \frac{f_2}{f_{n+1}}\right)$
redundancy	$R_n^{\min} = 0$	$\left(\frac{1}{2}, \frac{1}{2^2}, \ldots, \frac{1}{2^{n-3}}, \frac{1}{2^{n-2}}, \frac{1}{2^{n-1}}, \frac{1}{2^{n-1}}\right)$
	$R_n^{\max} = 1$	$\left(1-\varepsilon, \frac{\varepsilon}{2}, \frac{\varepsilon}{2^2}, \ldots, \frac{\varepsilon}{2^{n-3}}, \frac{\varepsilon}{2^{n-2}}, \frac{\varepsilon}{2^{n-2}}\right)$

doi: 10.1049/iet-com.2010.0049

The following lemma [7] relates the parameters of a source and the its corresponding tree to the parameters of its subtrees.

Lemma 1 ([7]): For any intermediate node with probability u

$$H(\mathcal{T}) = H(\Lambda_u) + uH(u^{-1} * \Delta_u) \tag{4}$$

The same equation holds for the average length and redundancy.

3 Construction

In this section we present a criterion on the probability distribution $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ of a source $\mathcal{S}$ such that employing Huffman coding on this source results in an AUH code.

Theorem 1 ([2, 3]): 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 anti-uniform is

$$p_{i+2} + p_{i+3} + \cdots + p_n \leq p_i, \quad \text{for } 1 \leq i \leq n-3 \tag{5}$$

A generalisation of Theorem 1 to arbitrary source is straightforward. In words, a (not necessarily finite) source $\mathcal{S}$ with ordered probability distribution $\mathcal{P} = (p_1, p_2, \ldots)$ induces an (finite or infinite) AUH code if and only if

$$\sum_{k \geq i+2} p_k \leq p_i, \quad \text{for } i \geq 1 \tag{6}$$

A probability distribution $\mathcal{P} = (p_1, p_2, \ldots)$ that satisfies the above property is called anti-uniform distribution [2, 3].

There are many well-known distributions lie in the class of anti-uniform distributions. It is known that Poisson distributions [4] with parameter $\lambda \leq 1$ and geometric distributions [6] $p_i = \varpi^k(1-\varpi)$, $k \geq 0$, with $0 < \varpi \leq (\sqrt{5} - 1/2)$ are among the class of infinite alphabet anti-uniform sources.

In the following we show that the discrete form of the exponential distribution is also anti-uniform for some regime of its parameter, λ. Consider

$$f(x; \lambda) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{7}$$

as the probability density function of the exponential distribution. Define $F(i) = P(X \leq i) = 1 - \mathrm{e}^{-\lambda i}$, and $p_i = F(i) - F(i-1) = \mathrm{e}^{-\lambda i}(\mathrm{e}^{\lambda} - 1)$ for $i \geq 1$. We have

$$\begin{aligned} p_i + p_{i+1} + \cdots &= \mathrm{e}^{-\lambda i}(\mathrm{e}^{\lambda} - 1)(1 + \mathrm{e}^{-\lambda} + \mathrm{e}^{-2\lambda} + \cdots) \\ &= \mathrm{e}^{-\lambda i} \frac{\mathrm{e}^{\lambda} - 1}{1 - \mathrm{e}^{-\lambda}} \end{aligned}$$

Now, since (6) is necessary and sufficient for characterising anti-uniform sources, then it is enough to have

$$\mathrm{e}^{-\lambda i} \frac{\mathrm{e}^{\lambda} - 1}{1 - \mathrm{e}^{-\lambda}} = \sum_{k \geq i} p_k \leq p_{i-2} = \mathrm{e}^{-\lambda(i-2)}(\mathrm{e}^{\lambda} - 1) \tag{8}$$

Inequality (8) holds if and only if $\mathrm{e}^{2\lambda} - \mathrm{e}^{\lambda} - 1 \geq 0$, which is true for $\lambda \geq \ln(1 + \sqrt{5}/2) \simeq 0.48$. Consequently, the exponential distribution is anti-uniform if and only if $\lambda \geq 0.48$.

4 Average length

The average length of any non-trivial code is lower bounded by (1). Using Lemma 1, it can be shown that for any arbitrary n the average length of the AUH source with distribution

$$\mathcal{P}_{n,\varepsilon} = \left(1 - \varepsilon, \frac{\varepsilon}{2}, \frac{\varepsilon}{4}, \ldots, \frac{\varepsilon}{2^{n-3}}, \frac{\varepsilon}{2^{n-2}}, \frac{\varepsilon}{2^{n-2}}\right)$$

tends to 1 as $\varepsilon \to 0$. Therefore average length of an AUH code is tightly lowerbounded by (1), that is, $L_n^{\min} = 1$ for every $n \geq 2$.

In the following, we will state a tight upperbound on the average length of AUH codes in terms of the alphabet size of the source. A similar result is also shown independently in [9].

Theorem 2: Let $\mathcal{P}$ be an ordered distribution over a discrete source of alphabet size n. Then $\mathrm{L}(\mathcal{P})$ is upperbounded by

$$L_n^{\max} = \frac{f_{n+3} - 3}{f_{n+1}} \tag{9}$$

where f_n is the nth Fibonacci number defined as $f_1 = f_2 = 1$ and

$$f_n = f_{n-1} + f_{n-2} \quad n \geq 3 \tag{10}$$

Furthermore, this bound is tight and can be achieved by the Fibonacci distribution

$$\mathcal{P}_n^{(F)} = \left(\frac{f_{n-1}}{f_{n+1}}, \frac{f_{n-2}}{f_{n+1}}, \ldots, \frac{f_3}{f_{n+1}}, \frac{f_2}{f_{n+1}}, \frac{f_1}{f_{n+1}}, \frac{f_2}{f_{n+1}}\right)$$

Before stating the proof, we show two simple lemmas that simplify the proof.

Lemma 2: In any probability distribution $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ which maximises the average length of the Huffman code, the probability of any arbitrary leaf is not greater than the probability of the intermediate node in the same level, that is

$$p_i \leq q_i \quad i = 1, \ldots, n-2$$

where $q_i = \Sigma_{j>i} p_j$.

Lemma 3: Any probability distribution $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ with maximum average length, satisfies

$$p_1 = q_2 = \sum_{i>2} p_i \tag{11}$$

The proof of these lemmas are presented in Appendix. Now, we are ready to prove the main theorem of this section.

Proof of Theorem 2: We prove the theorem using induction over the alphabet size, n. It is clear that $L_2^{\max} = 1 = (f_5 - 3)/f_3$. For $n = 3$, one can argue that $\mathcal{P} = (1/3, 1/3, 1/3)$ has the maximum average length $L_3^{\max} = (f_6 - 3)/f_4 = 5/3$. Let the theorem be true for any $k < n$, and $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ achieve the maximum average length of an AUH for n symbols. We consider two cases as follows:

1. $p_1 \geq (f_{n-1}/f_n + 1)$: We denote the subtree lying under $q_1 = 1 - p_1$ by Δ_{1-p_1} as before. Note that

$(1-p_1)^{-1} * \Delta_{1-p_1}$ is an AUH tree over $n-1$ symbols, and its average length is upper bounded by $L_{n-1}^{\max}$ according to the assumption of the induction for $k = n-1$. Therefore using Lemma 1, we have

$$\begin{aligned} L(\mathcal{P}) &= 1 + (1-p_1)L((1-p_1)^{-1} * \Delta_{1-p_1}) \\ &\le 1 + \left(1 - \frac{f_{n-1}}{f_{n+1}}\right)\frac{f_{n+2}-3}{f_n} \\ &= \frac{f_{n+3}-3}{f_{n+1}} \end{aligned}$$

2. $p_1 \ge (f_{n-1}/f_n + 1)$: Using Lemma 3 we have $q_2 = p_1$. By expanding $L(\mathcal{P})$ with respect to q_2 and using Lemma 1 we can write

$$\begin{aligned} L(\mathcal{P}) &= L(\Lambda_{q_2}) + q_2 L(q_2^{-1} * \Delta_{q_2}) \\ &= (2-p_1) + p_1 L(q_2^{-1} * \Delta_{q_2}) \\ &\le 2 + \frac{f_{n-1}}{f_{n+1}}\left(\frac{f_{n+1}-3}{f_{n-1}} - 1\right) \\ &= \frac{f_{n+3}-3}{f_{n+1}} \end{aligned}$$

where the inequality is the assumption of the induction for $k = n-2$. □

Remark 1: One can simply show that $\mathcal{P}_n^{(F)}$ meets the upper bound for any alphabet size. Although some other distributions such as $\mathcal{P}_4 = (0.35, 0.30, 0.20, 0.15)$ can meet the bound, it can be shown that the maximal distribution is unique for $n > 4$.

Remark 2: Note that the Fibonacci probability distribution tends to

$$Q_n = (t^2, t^3, \ldots, t^{n-1}, t^n, t^{n-1})$$

as n grows, where $t = (\sqrt{5}-1/2)$ is the positive root of $x^2 + x - 1 = 0$, that is $\lim \Sigma_{n\to\infty} \Sigma_{i=1}^{n} |p_i - q_i| = 0$.

Furthermore, it is easy to see that $\{L_n^{\max}\}_{n=1}^{\infty}$ is an increasing sequence and tends to $t^{-2} = (3+\sqrt{5}/2) \simeq 2.618$ in the asymptotic case.

5 Entropy

Since only very particular sources satisfy the AUH structure, the range of the entropy of such sources is not so wide. It is easy to check that the minimum entropy of such sources can be arbitrary close to zero for any alphabet size n. In order to see this, one may compute the entropy of

$$P_{n,\varepsilon} = \left(1-\varepsilon, \frac{\varepsilon}{2}, \frac{\varepsilon}{4}, \ldots, \frac{\varepsilon}{2^{n-3}}, \frac{\varepsilon}{2^{n-2}}, \frac{\varepsilon}{2^{n-2}}\right)$$

for $\varepsilon \le 2/3$ and show that $H(P_{n,\varepsilon}) \to 0$ as $\varepsilon \to 0$.

In spite of that, upperbounding the entropy of AUH codes is not trivial. It has been shown in [3] that the entropy of an infinite length AUH codes with a given average length L is upper bounded by

$$\hat{H}_{\infty}^{\max}(L) = L\log L - (L-1)\log(L-1) \tag{12}$$

This bound is only valid for infinite sources. It can be shown that for any dyadic source with n symbols

$$\mathcal{P}_D = \left(\frac{1}{2}, \frac{1}{2^2}, \ldots, \frac{1}{2^{n-3}}, \frac{1}{2^{n-2}}, \frac{1}{2^{n-1}}, \frac{1}{2^{n-1}}\right) \tag{13}$$

we have $H(\mathcal{P}_D) = L(\mathcal{P}_D) = 2 - 2^{2-n}$, which is greater that $\hat{H}_{\infty}^{\max}(L)$. The following theorem states a tight upperbound on the entropy of AUH sources with n symbols.

Theorem 3: The entropy of a finite anti-uniform source with n symbols is upperbounded by

$$H_n^{\max} = H(\mathcal{P}_n^{(F)}) = \log f_{n+1} - \frac{1}{f_{n+1}}\sum_{i=1}^{n-1} f_i \log f_i \tag{14}$$

Remark 3: Note that $\{H_n^{\max}\}_{n=1}^{\infty}$ is also increasing with n, and tends to $((1+t^2)/t^2)\log(1/t)$. Recall that $L_{\infty}^{\max} = t^{-2}$, and therefore, $\hat{H}_{\infty}^{\max}(L_{\infty}^{\max}) = \hat{H}_{\infty}^{\max}(t^{-2}) = ((1+t^2)/t^2)\log(1/t)$. This shows that our result is consistent with that of [3], for infinite alphabet sources.

The proof of this theorem is fairly similar to that of Theorem 2. The following lemmas show some basic properties on distributions that achieve maximum entropy. We present the proofs of the lemmas in Appendix.

Lemma 4: Let $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ be a distribution over n symbols with the maximum possible entropy. Then $p_i \le q_i$ for any $i = 1, \ldots, n-2$.

Lemma 5: For any distribution $\mathcal{P} = (p_1, p_2, \ldots, p_n)$ the achieves maximum entropy, $p_1 = q_2$.

Now we can present the proof of Theorem 3.

Proof of Theorem 3: Similar to the proof of Theorem 2, we prove this theorem by induction over the alphabet size of the source. Since the uniform distribution satisfies the AUH constraints for $n = 3$ and $n = 2$, we have $L_2^{\max} = 1$ and $L_3^{\max} = \log 3$ that coincide with (14). Now, let $n \ge 4$ and the bound is valid for $k < n$. We consider two cases.

1. $p_1 \ge (f_{n-1}/f_{n+1})$: Using Lemma 1 and by expansion the entropy with respect to q_1, we have

$$\begin{aligned} H(\mathcal{P}) &= h_b(p_1) + q_1 H(q_1^{-1} * \Delta_{q_1}) \\ &\le h_b(p_1) + (1-p_1)H_{n-1}^{\max} \\ &\le h_b\left(\frac{f_{n-1}}{f_{n+1}}\right) + \left(1 - \frac{f_{n-1}}{f_{n+1}}\right)H_{n-1}^{\max} \\ &= \log f_{n+1} - \frac{1}{f_{n+1}}\sum_{i=1}^{n-1} f_i \log f_i \end{aligned} \tag{15}$$

where $h_b(x) = -x\ \log x - (1-x)\ \log(1-x)$ is the binary entropy function. Note that the first inequality follows form the assumption of induction for $k = n-1$ and the second inequality follows is owing to the fact that the function $\alpha(x) = h(x) + (1-x)H_{n-1}^{\max}$ is non-increasing for $x \ge f_{n-1}/f_{n+1}$.

doi: 10.1049/iet-com.2010.0049

It can be shown by taking the derivative of $\alpha(x)$ as

$$\begin{aligned}\frac{d\alpha}{dx} &= \log\frac{1-p}{p} - H_{n-1}^{\max} \\ &\le \log\frac{1-f_{n-1}/f_{n+1}}{f_{n-1}/f_{n+1}} - H_{n-1}^{\max} \\ &= \log\frac{f_n}{f_{n-1}} - \log f_n + \frac{1}{f_n}\sum_{i=1}^{n-2} f_i \log f_i \\ &= \frac{1}{f_n}\left[\sum_{i=1}^{n-2} f_i \log\frac{f_i}{f_{n-1}} - \log f_{n-1}\right] \le 0\end{aligned}$$

2. $p_1 \le (f_{n-1}/f_{n+1})$: Using Lemma 3, we can only focus on the distributions for which $q_2 = p_1$. Now we use Lemma 1 to expand the entropy with respect to q_2.

$$\begin{aligned}H(\mathcal{P}) &= h(\Lambda_{q_2}) + q_2 H(q_2^{-1} * \Delta_{q_2}) \\ &= -2p\log p - (1-2p)\log(1-2p) + pH(q_2^{-1} * \Delta_{q_2}) \\ &\le -2p\log p - (1-2p)\log(1-2p) + pH_{n-2}^{\max} \\ &\le -2\frac{f_{n-1}}{f_{n+1}}\log\frac{f_{n-1}}{f_{n+1}} - \left(1-2\frac{f_{n-1}}{f_{n+1}}\right)\log\left(1-2\frac{f_{n-1}}{f_{n+1}}\right) \\ &\quad + \frac{f_{n-1}}{f_{n+1}} H_{n-2}^{\max} \\ &= \log f_{n+1} - \frac{1}{f_{n+1}}\sum_{i=1}^{n-1} f_i \log f_i \end{aligned} \tag{16}$$

where again the assumption of induction for $k = n-2$ implies the first inequality and the second one follows from the fact that $\beta(x) = -2x\log x - (1-2x)\log(1-2x) + xH_{n-2}^{\max}$ is an increasing function for $x > f_{n-1}/f_{n+1}$. □

Corollary: Note that the maximum achievable entropy for an infinite-size alphabet is

$$\lim_{n\to\infty} H_n^{\max} = -\sum_{i=2}^{\infty} t^i \log t^i = \left(1+\frac{1}{t^2}\right)\log\frac{1}{t} \tag{17}$$

which is obtained for $\mathcal{P}_\infty^{(F)}$ and coincides with

$$\begin{aligned}&L_\infty^{\max}\log L_\infty^{\max} - (L_\infty^{\max}-1)\log(L_\infty^{\max}-1) \\ &= \frac{1}{t^2}\log\frac{1}{t^2} - \left(\frac{1}{t^2}-1\right)\log\left(\frac{1}{t^{2x}}-1\right) \\ &= \left(1+\frac{1}{t^2}\right)\log\frac{1}{t}\end{aligned}$$

This simply proves (12).

6 Redundancy

It is known that the Huffman code associated with any dyadic source has zero redundancy. Since such distributions exist for any arbitrary alphabet size and satisfying the AUH constraints, the redundancy of AUH code is tightly lowerbounded by zero. On the other extreme, it can be shown that the redundancy of AUH codes can be arbitrary close to 1.

Theorem 4: For any alphabet size n and $\delta > 0$, there exist AUH sources on n symbols for which $R(\mathcal{P}) > 1-\delta$.

Proof: Take an arbitrary AUH distribution $\mathcal{Q} = (q_1, \ldots, q_{n-1})$ over $n-1$ symbols and $\varepsilon > 0$ small enough such that $1-\varepsilon \ge \max\{q_1\varepsilon, (1-q_1)\varepsilon\}$. Therefore $\mathcal{P} = (1-\varepsilon, q_1, \varepsilon, \ldots, q_{n-1}\varepsilon)$ is a distribution with AUH code. Using Lemma 1 we have

$$\begin{aligned}R(\mathcal{P}) &= 1 - h_b(\varepsilon) + \varepsilon R(\varepsilon^{-1} * \Delta_\varepsilon) \\ &= 1 - h_b(\varepsilon) + \varepsilon R(\mathcal{Q})\end{aligned}$$

Note that $R(\mathcal{Q})$ is bounded by $L(\mathcal{Q}) \le L_{n-1}^{\max} < t^{-2}$, and $R(\mathcal{P})$ tends to 1 as $\varepsilon \to 0$. □

7 Average cost

In this section we consider the problem of designing minimum average cost codes. In particular, we only focus on the highly unbalanced costs regime, where $c_1 \gg c_0$ (or similarly $c_0 \gg c_1$). The following theorem shows that the class of AUH codes are optimal for this regime.

Theorem 5: Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be a finite alphabet with associated equal probability distribution $\mathcal{P} = (p_1, p_2, \ldots, p_n)$, $p_1 = p_2 = \cdots = p_n = (1/n)$. If $c_1 \gg c_0 > 0$ (or $c_0 \gg c_1 > 0$), then a prefix-free code for $\mathcal{S}$ with minimum average cost has an AUH structure.

Proof: It is clear that since $c_1 \gg c_0$, one can decrease the average cost by choosing codewords with minimum number of 1's. On the other hand, since the code is prefix-free, we can have at most one all zero codeword. Furthermore, the position of the 1's in the non-zero codewords should be different. These lead to an AUH structure, with codeword set $\{1, 01, 001, \ldots, 0^{n-3}1, 0^{n-2}1, 0^{n-1}\}$, which minimises the average cost of the code. Similarly, we can show that if $c_0 \gg c_1$ then the code with minimal average cost is $\{0, 10, 100, \ldots, 1^{n-3}0, 1^{n-2}0, 1^{n-1}\}$. □

Having the structure of the cost minimising codes, we can also give a tight upper bound on the entropy of an AUH source, with given average cost.

Theorem 6: Let $\mathcal{S}$ be an infinite size alphabet where $c_0 \gg c_1$, and $\mathcal{C}$ be an AUH code with average cost C associated to $\mathcal{S}$. Then the entropy of the source is upper bounded by

$$H_\infty^{\max}(C) = \frac{C}{c_0}\log\left(1+\frac{c_0}{C-c_1}\right) + \log\left(\frac{C-c_1}{c_0}\right)$$

Proof: We are going to use Lagrange multipliers method to prove the theorem. The goal is to minimise $H(\mathcal{P})$ subject to two constraints $g_1(\mathcal{P}) \triangleq \Sigma_{i\ge1}(p_i((i-1)c_0 + c_1)) - C = 0$ and $g_2(\mathcal{P}) \triangleq \Sigma_i p_i - 1 = 0$. Define

$$\begin{aligned}\Gamma(\mathcal{P}) &\triangleq H(\mathcal{P}) + \alpha g_1(\mathcal{P}) + \beta g_2(\mathcal{P}) \\ &= -\sum_i p_i \log p_i + \alpha\left(\sum_i p_i((i-1)c_0 + c_1) - C\right) \\ &\quad + \beta\left(\sum_i p_i - 1\right)\end{aligned}$$

By taking the derivative of $\Gamma(\mathcal{P})$ with respect to p_i and setting it to be zero, we have

$$\frac{\partial \Gamma(\mathcal{P})}{\partial p_i} = c_0\alpha(i-1) + \alpha c_1 + \beta - \log p_i - \frac{1}{\ln 2} = 0$$

for $i \geq 1$. Therefore

$$\ln p_i = c_0\alpha'(i-1) + \alpha' c_1 + \beta' - 1 \tag{18}$$

where $\alpha' = \alpha \ln 2$ and $\beta' = \beta \ln 2$ have to be chosen such that the constraints $g_1(\mathcal{P}) = 0$ and $g_2(\mathcal{P}) = 0$ are satisfied. This gives us

$$\alpha' = \frac{1}{c_0} \ln \frac{C - c_1}{C - c_1 + c_0}$$

$$\beta' = 1 + \ln \frac{c_0}{C - c_1}$$

The resulting entropy of such distribution can be computed as

$$\begin{aligned} H(\mathcal{P}) &= -\sum_i p_i \log p_i \\ &= \frac{-1}{\ln 2}\left(\sum_i p_i(c_0\alpha'(i-1) + \alpha' c_1 + \beta' - 1)\right) \\ &= \frac{-1}{\ln 2}(\alpha' C + \beta' - 1) \end{aligned} \tag{19}$$

Replacing α' and β' by the corresponding obtained values we get the bound. □

8 Conclusion

In this paper we obtained tight upper and lower bounds on the average length, entropy and redundancy of finite and infinite AUH codes, with these bounds we have tighten the similar bounds presented in [2, 3]. Furthermore, we have shown that for a given alphabet size, Fibonacci distribution maximises the average length and entropy. We showed that the minimum average cost of a code is achieved by an AUH code in highly unbalanced cost regime. Moreover, we presented a tight upper bound on the entropy of an AUH source with given average length or cost.

9 References

1 Huffman, D.A.: 'A method for the construction of minimum-redundancy codes', *Proc. IRE*, 1952, **40**, (2), pp. 1098–1101
2 Esmaeili, M., Kakhbod, A.: 'On antiuniform and partially antiuniform sources'. Proc. IEEE ICC, June 2006, pp. 1611–1615
3 Esmaeili, M., Kakhbod, A.: 'On information theory parameters of infinite anti-uniform sources', *IET Commun.*, 2007, **1**, pp. 1039–1041
4 Humblet, P.: 'Optimal source coding for a class of integer alphabets', *IEEE Trans. Inf. Theory*, 1978, **24**, (1), pp. 110–112
5 Kato, A., Han, T., Nagaoka, H.: 'Huffman coding with an infinite alphabet', *IEEE Trans. Inf. Theory*, 1996, **42**, (3), pp. 977–984
6 Gallager, R., Van Voorhis, D.: 'Optimal source coding for geometrically distributed integer alphabets', *IEEE Trans. Inf. Theory*, 1975, **21**, (2), pp. 228–230
7 Mohajer, S., Pakzad, P., Kakhbod, A.: 'Tight bounds on the redundancy of huffman codes'. Proc. IEEE ITW, March 2006, pp. 131–135
8 Capocelli, R.M., Santis, A.D.: 'New bounds on the redundancy of huffman codes', *IEEE Trans. Inf. Theory*, 1991, **37**, (4), pp. 1095–1104
9 Esmaeili, M.: 'On the weakly superincreasing distributions and the Fibonacci–Hessenberg matrices', *ARS Comb.*, 2007, **84**, pp. 217–224

10 Appendix

Proof of Lemma 2: Let $p_i > q_i$ for some i. This implies $\varepsilon = (p_i - q_i)/2$ is positive. Defining $\varepsilon_i = -\varepsilon$ and $\varepsilon_k = p_k\varepsilon/q_i$ for $k > i$, we can show the distribution

$$\mathcal{P}' = (p_1, \ldots, p_{i-1}, p_i + \varepsilon_i, p_{i+1} + \varepsilon_{i+1}, \ldots, p_n + \varepsilon_n)$$

satisfies the AUH constraint in (5). Moreover

$$\begin{aligned} L(\mathcal{P}') - L(\mathcal{P}) &= \sum_{k=i}^{n-1} \varepsilon_k k + \varepsilon_n(n-1) \\ &= \sum_{k=i+1}^{n-1} \varepsilon_k(k-i) + \varepsilon_n(n-1-i) > 0 \end{aligned}$$

which is in contradiction with the maximality of $\mathcal{P}$. □

Proof of Lemma 3: The structure of the Huffman tree implies $p_1 \geq q_2$ owing to the location of the corresponding nodes on the tree. Assume the inequality is strict and so $\varepsilon = (p_1 - q_2)/2 > 0$. Defining $\varepsilon_1 = -\varepsilon$, and $\varepsilon_k = p_k\varepsilon/q_1$ for $k > 1$, it is easy to show that

$$\mathcal{P}' = (p_1 + \varepsilon_1, p_2 + \varepsilon_2, \ldots, p_n + \varepsilon_n)$$

satisfies the AUH structure, and we have

$$\begin{aligned} L(\mathcal{P}') - L(\mathcal{P}) &= \sum_{k=1}^{n-1} \varepsilon_k k + \varepsilon_n(n-1) \\ &= \sum_{k=2}^{n-1} \varepsilon_k(k-1) + \varepsilon_n(n-2) > 0 \end{aligned}$$

This refuses the assumption of maximality of $\mathcal{P}$. □

Proof of Lemma 4: Similar to the proof of Lemma 2, we assume that the condition does not hold for some i and make a contradiction. Assume $\varepsilon = (p_i - q_i)/2 > 0$, and define the modified distribution

$$\begin{aligned} \mathcal{P}' = (p'_1, \ldots, p'_n) = (p_1, p_2, \ldots, p_{i-1}, p_i + \varepsilon_i, p_{i+1} \\ + \varepsilon_{i+1}, \ldots, p_n + \varepsilon_n) \end{aligned} \tag{20}$$

where $\varepsilon_i = -\varepsilon$ and $\varepsilon_k = p_k\varepsilon/q_i$ for $k > i$. We can write

$$\begin{aligned} H(\mathcal{P}') - H(\mathcal{P}) &= \sum_k p_k \log p_k - \sum_k p'_k \log p'_k \\ &= \sum_k p_k \log p_k - \sum_{k<i} p_k \log p'_k \\ &\quad - \sum_{k \geq i} (p_k + \varepsilon_k) \log p'_k \\ &= D(\mathcal{P}||\mathcal{P}') + \sum_{k>i} \varepsilon_k \log \frac{p'_i}{p'_k} > 0 \end{aligned}$$

where $D(\cdot\|\cdot)$ is the Kullback–Leibler divergence and the last inequality follows from the facts that $\Sigma_{k>i}\varepsilon_k = -\varepsilon_i$

doi: 10.1049/iet-com.2010.0049

and $\mathcal{P}'$ is an decreasing sequence. This inequality is in contradiction with the assumption, which implies the desired result. □

Proof of Lemma 5: Assume $\mathcal{P}$ is an entropy maximal distribution. The structure of AUH tree implies $p_1 \geq q_2$. Let the inequality be strict and define $\varepsilon = (p_1 - q_2)/2 > 0$. Consider the distribution

$$\mathcal{P}' = (p'_1, \ldots, p'_n) = (p_1 + \varepsilon_1, p_2 + \varepsilon_2, \ldots, p_n + \varepsilon_n)$$

where $\varepsilon_1 = -\varepsilon$ and $\varepsilon_k = p_k\varepsilon/q_1$ for $k > 1$. We have

$$\begin{aligned} H(\mathcal{P}') - H(\mathcal{P}) &= \sum_k p_k \log p_k - \sum_k p'_k \log p'_k \\ &= \sum_k p_k \log p_k - \sum_k p'_k \log (p_k + \varepsilon_k) \\ &= D(\mathcal{P}||\mathcal{P}') + \sum_{k>1} \varepsilon_k \log \frac{p'_k}{p'_i} > 0 \end{aligned}$$

which refuses the assumption we made in the beginning. □