LABORATORY FOR COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MIT/LCS/TM-536

On the redundancy achieved by Huffman codes

Roberto De Prisco and Alfredo De Santis

September, 1995

545 TECHNOLOGY SQUARE; CAMBRIDGE, MASSACHUSETTS 02139 (617) 253-5851

On the redundancy achieved by Huffman codes

Roberto De Prisco* Alfredo De Santis†

Abstract

It has been recently proved that the redundancy r of any discrete memoryless source satisfies $r \leq 1 - \mathcal{H}(p_N)$, where p_N is the least likely source letter probability. This bound is achieved only by sources consisting of two letters. We prove a sharper bound if the number of source letters is greater than two. Also provided is a new upper bound on r, as function of the two least likely source letter probabilities which improve on previous results.

*MIT Laboratory for Computer Science, 545 NE Technology Square, Cambridge MA 02139, USA. This work has been done while the author was at the Dipartimento di Informatica ed Applicazioni, Università di Salerno, 84081 Baronissi (SA), Italy. E-mail: `robdep@theory.lcs.mit.edu`

†Dipartimento di Informatica ed Applicazioni, Università di Salerno, 84081 Baronissi (SA), Italy. E-mail: `ads@dia.unisa.it`

1 Introduction

Let $S = \{a_1, a_2, ..., a_N\}$ be a discrete source with N letters and let p_k denote the probability of letter a_k, $1 \leq k \leq N$. We assume, without loss of generality, that $p_1 \geq p_2 \geq \ldots \geq p_N$. Let $C = \{x_1, x_2, \ldots, x_N\}$ be a code for source S and let $n_1 \leq n_2 \leq \ldots \leq n_N$ be the codeword lengths. Codeword x_i encodes the letter a_i, for $i = 1, 2, \ldots, N$. The *length vector* of a code C is the vector $(n_1, n_2, \ldots, n_N)$. The *Huffman* encoding algorithm [6] provides an *optimal prefix code* C for the source S. The encoding is optimal in the sense that codeword lengths minimize the *redundancy* r, defined as the difference between the *average codeword length* L and the *entropy* $H(p_1, p_2, \ldots, p_N)$ of the source:

$$r = L - H(p_1, p_2, \ldots, p_N) = \sum_{i=1}^{N} p_i n_i + \sum_{i=1}^{N} p_i \log p_i$$

where log denotes the logarithm to base 2. It is well known that the redundancy of an optimal code satisfies $0 \leq r < 1$. These bounds can be improved if one knows partial information on the source. Gallager [4], Johnsen [7], Capocelli *et al.* [3], Capocelli and De Santis [1], and Manstetten [10] considered the problem of upper bounding r when p_1 is known.

Reza [9] and Horibe [5] considered the problem of upper bounding the redundancy when the least likely source letter probability p_N is known. Their bound has been recently improved by Capocelli and De Santis [2] who proved that, as function of p_N, the redundancy r of Huffman codes is upper bounded by

$$r \leq 1 - \mathcal{H}(p_N), \tag{1}$$

where $\mathcal{H}$ is the binary entropy function $\mathcal{H}(p) = -p \log p - (1-p) \log(1-p)$.

Recently, Yeung [13] introduced the notion of local redundancy and, by exploiting it, he derived the following bound as function of two least likely source letter probabilities:

$$r \leq 1 - \mathcal{H}(p_{N-1}) + (p_{N-1} + p_N) \left[1 - \mathcal{H}\left(\frac{p_N}{p_{N-1} + p_N}\right)\right]. \tag{2}$$

Bound (1) is clearly the best possible upper bound on r as function of only p_N. Indeed, any binary source $(1 - p_N, p_N)$ satisfies (1) with equality. If the source is not binary, the bound (1) is no more the best possible.

In this paper we derive upper bounds on r as function of p_N when additional information on the total number of source letters is known.

We prove that for $N \geq 3$ the following bound, as function of the least likely source letter probability, holds:

$$r \leq \begin{cases} 1 - \mathcal{H}(2p_{_N}), & \text{if } 0 < p_{_N} \leq \delta \\ 0.5 + 1.5p_{_N} - \mathcal{H}(p_{_N}), & \text{if } \delta < p_{_N} \leq 1/3 \end{cases} \tag{3}$$

where $\delta \simeq 0.1525$. This bound is the best possible expressed only as function of $p_{_N}$ for every $p_{_N} > 0$ and $N \geq 3$.

We also present three tight upper bounds as function of the least likely source letter probability, when it is known that there are exactly three, four and five source letters.

Finally, we use (3) to obtain an upper bound as function of the two least likely source letter probabilities which improves Yeung's bound (2).

This paper is organized as follows. In Section 2 we present some definitions and results. In Section 3 we provide tight upper bounds as function of $p_{_N}$ for the special cases $N = 3, 4, 5$. By using these bounds, in Section 4 we prove (3). Finally, in Section 5 we prove the upper bound on r, as function of $p_{_{N-1}}$ and $p_{_N}$.

2 Preliminaries

In this section we present some definitions and results, that will be useful in the rest of the paper.

An Huffman code can be represented as a rooted tree in which each source letter corresponds to a leaf on the tree and where the associated codeword is the sequence of labels on the path from root to leaf. The code tree is generated by the Huffman algorithm. Each internal node in the code tree has a probability defined as the sum of the probabilities of his two children. The external nodes have the probability of the corresponding source letter. Gallager [4] proved that an Huffman code tree has the sibling property, i.e. each node, except the root, has a sibling and all nodes can be listed in order of decreasing probability with each node being adjacent to its sibling. We number all the nodes, except the root, using the sibling property, so that for each k, $k = 1, \ldots, N-1$, nodes $2k-1$ and $2k$ are siblings, and the probability of a node i is greater than the probability of node $i+1$, for each $i = 1, \ldots, 2N-3$. Let q_i be the probability of node i, $i = 1, \ldots, 2N-2$. The list $(q_1, \ldots, q_{2N-3}, q_{2N-2})$ is called the *sibling list* [4].

Let u be a node at level l in the Huffman code tree, and v a node at level $l+1$. Then

$$q_u \geq q_v. \tag{4}$$

Lemma 1 *The redundancy r of a source, whose most and least likely source letter probabilities are respectively p_1 and p_N, is upper bounded by:*

$$r \leq p_1 + 0.086 - p_N \quad \text{for } 0 < p_1 \leq 1/6 \tag{5}$$

$$r \leq 2 - 1.3219(1 - p_1) - \mathcal{H}(p_1) - p_N \quad \text{for } 1/6 < p_1 \leq 0.1971 \tag{6}$$

$$r \leq 4 - 18.6096 p_1 - \mathcal{H}(5p_1) - p_N \quad \text{for } 0.1971 < p_1 \leq 0.2 \tag{7}$$

$$r \leq 2 - 1.25(1 - p_1) - \mathcal{H}(p_1) - p_N \quad \text{for } 0.2 < p_1 \leq 0.3138 \tag{8}$$

$$r \leq 3 - (3 + 3\log 3)p_1 - \mathcal{H}(3p_1) - p_N \quad \text{for } 0.3138 < p_1 \leq 1/3 \tag{9}$$

$$r \leq 1 + 0.5(1 - p_1) - \mathcal{H}(p_1) - 2p_N \quad \text{for } 1/3 < p_1 \leq 0.4505, \text{ and } N \geq 6 \tag{10}$$

Proof. Bound (5) is provided by Capocelli *et al.* [3].
Bounds (6) and (7) follow from Manstetten's work [10]. Manstetten proved that for $p_1 \in [1/(j+1), 1/j]$ the redundancy can be written as

$$r = s + \sum_{k=v}^{m} q'_k - H(q'_1, \ldots, q'_m) + \sum_{k=m}^{N-1} (q_{2k-1} + q_{2k})[1 - \mathcal{H}(q_{2k}/(q_{2k-1} + q_{2k}))],$$

where s is the integer such that $2^s - 1 \leq j \leq 2^{s+1} - 1$, and $q'_1, \ldots, q'_m$ are the entries $q_j, \ldots, q_{2j}$, in the sibling list so that $m = j + 1$, and $q'_i = q_{i+j-1}$ for $i = 1, \ldots, m$, and, finally, $v = 2^{s+1} - j$. The above expression of r can be upper bounded by

$$r \leq s + \sum_{k=v}^{m} q'_k - H(q'_1, \ldots, q'_m) + q'_m - p_N.$$

Manstetten provided upper bounds on $s + \sum_{k=v}^{m} q'_k - H(q'_1, \ldots, q'_m) + q'_m$, as function of p_1. From these bounds we can obtain bounds on r as function of p_1 and p_N. In particular for $j = 5, 6$, one obtains (6), (7) and (8).
Bound (9) are due to Capocelli and De Santis [1]. Bound (10) has been obtained by Capocelli *et al.* (see Theorem 5 and subsequent remark in [3]). □

A useful result due to Montgomery and Kumar [12] is the following:

Lemma 2 *Let S be a discrete memoryless source whose most likely source letter probability is p_1. If for some integer m*

$$\frac{2}{2^{m+1} + 1} < p_1 < \frac{1}{2^m - 1},$$

then an optimal code for S must have the minimum codeword length $n_1 = m$. Furthermore, if

$$\frac{1}{2^{m+1} - 1} \leq p_1 < \frac{2}{2^{m+1} + 1},$$

then an optimal code for S must have the minimum codeword length either $n_1 = m$ or $n_1 = m + 1$.

3 Upper bounds on r as function of $p_{_N}$, for fixed N

In this section we provide three upper bounds on r, as function of $p_{_N}$ when we know that the source has exactly $3, 4, 5$ letters. More precisely, in subsections 2.1, 2.2, 2.3 we consider the cases $N = 3, 4, 5$, respectively.

These three upper bounds will be useful to determine the general upper bound (3) on r as function of $p_{_N}$, when $N \geq 3$.

3.1 An upper bound when $N = 3$

In this section we derive the upper bound as function of the least likely source letter probability, for ternary sources.

Theorem 1 *Let $S = (p_1, p_2, p_3)$ be a discrete source and $p_3 = p_{_N}$ be its least likely source letter probability. The redundancy of the corresponding Huffman code is upper bounded by:*

$$r \leq \begin{cases} 1 - \mathcal{H}(2p_{_N}), & \text{if } 0 < p_{_N} \leq \delta \\ 0.5 + 1.5p_{_N} - \mathcal{H}(p_{_N}), & \text{if } \delta < p_{_N} \leq 1/3 \end{cases} \tag{11}$$

where $\delta \simeq 0.1525$ is the unique zero of the function $0.5 - 1.5x - \mathcal{H}(2x) + \mathcal{H}(x)$ in the interval $[0, 0.3]$. The bound is tight.

Proof. The Huffman code for S has length vector $(1, 2, 2)$, and its redundancy is

$$r = 1 + p_2 + p_3 - H(p_1, p_2, p_3).$$

Maximizing over all possible values for p_1 and p_2, given that $p_3 = p_{_N}$, it follows

$$r \leq 1 + p_{_N} + \max_{(y_1, y_2, p_{_N}) \varepsilon Q} \{y_2 - H(y_1, y_2, p_{_N})\}$$

where Q is the set of all source of three letters whose least likely source letter probability is $p_3 = p_{_N}$, i.e. $Q = \{(y_1, y_2, p_{_N}) | y_1 \geq y_2 \geq p_{_N}, y_1 + y_2 + p_{_N} = 1\}$. Alternatively, the set Q can be written as $Q = \{(1 - x - p_{_N}, x, p_{_N}) | p_{_N} \leq x \leq (1 - p_{_N})/2\}$. Since the function $x - H(1 - x - p_{_N}, x, p_{_N})$ is a convex $\cup$ function of x the maximum value must occur at an extreme point of x's interval of variation. The extreme points of this interval are $p_{_N}$ and $(1-p_{_N})/2$. If $x = p_{_N}$ then $r \leq 1 - \mathcal{H}(2p_{_N})$ whereas if $x = (1-p_{_N})/2$ then $r \leq 0.5 + 1.5p_{_N} - \mathcal{H}(p_{_N})$. Hence, we have

$$r \leq \max\{1 - \mathcal{H}(2p_{_N}), 0.5 + 1.5p_{_N} - \mathcal{H}(p_{_N})\}$$

that leads us to (11).

The bound is reached by the source $(1 - 2p_{_N}, p_{_N}, p_{_N})$, if $0 < p_{_N} \leq \delta$, and by the source $(\frac{1-p_N}{2}, \frac{1-p_N}{2}, p_{_N})$, if $\delta < p_{_N} \leq 1/3$. $\square$

Define the function $\alpha(p_{_N})$ as follows

$$\alpha(p_{_N}) = \begin{cases} 1 - \mathcal{H}(2x) & \text{if } 0 < p_{_N} \leq \delta \\ 0.5 + 1.5 - \mathcal{H}(x) & \text{if } \delta < p_{_N} \leq 1/3 \end{cases}$$

that is, the function α represent the bound on r given by (11).

The function $1 - \mathcal{H}(2x)$ is a decreasing function of x, for $x \in [0, \delta]$. The function $0.5 + 1.5x - \mathcal{H}(x)$ is a convex $\cup$ function of x with a minimum at $\chi \simeq 0.26$, the unique zero of the first derivate of $0.5 + 1.5 - \mathcal{H}(x)$ in $[\delta, 1/3]$. Thus, the bound (11) is a decreasing function of $p_{_N}$ for $p_{_N} \leq \chi$ and it is an increasing function of $p_{_N}$ for $\chi < p_{_N} \leq 1/3$. Hence the following lemma holds.

Lemma 3 *The function $\alpha(p_{_N})$ is a decreasing function of $p_{_N}$, for $p_{_N} \in [0, \chi]$, $\chi \simeq 0.26$, whereas for $p_{_N} \in [\chi, 1/3]$, $\alpha(p_{_N})$ is an increasing function of $p_{_N}$.*

Observe that, since $\alpha(\delta) > \alpha(1/3)$, Lemma 3 implies that if $p \in]0, \delta]$ one has $\alpha(p) > \alpha(x)$ for all $x \in]p, 1/3]$.

Bound (11) is depicted in Figure 1.

3.2 An upper bound when N=4

In this section we derive an upper bound as function of $p_{_N}$, for all sources consisting of four source letters.

Theorem 2 *Let $S = (p_1, p_2, p_3, p_4)$ be a discrete source and $p_4 = p_{_N}$ be its least likely source letter probability. The redundancy of the corresponding Huffman code is upper bounded by:*

$$r \leq \begin{cases} 1 + 5p_{_N} - H(1 - 3p_{_N}, p_{_N}, p_{_N}, p_{_N}) & \textit{if } 0 < p_{_N} \leq 1/9 \\ 2 - H(1/3, 1/3, 1/3 - p_{_N}, p_{_N}) & \textit{if } 1/9 < p_{_N} \leq 1/6 \\ 2 - H(2p_{_N}, 1 - 4p_{_N}, p_{_N}, p_{_N}) & \textit{if } 1/6 < p_{_N} \leq \delta_1 \\ 2 - H(\frac{1+p_{_N}}{3}, \frac{1-2p_{_N}}{3}, \frac{1-2p_{_N}}{3}, p_{_N}) & \textit{if } \delta_1 < p_{_N} \leq 1/5 \\ 2 - H(1 - 3p_{_N}, p_{_N}, p_{_N}, p_{_N}) & \textit{if } 1/5 < p_{_N} \leq 1/4 \end{cases} \tag{12}$$

where $\delta_1 \simeq 0.1708$ is the unique point in the interval $[1/6, 1/5[$ for which the function $H(2x, 1 - 4x, x, x)$ is equal to the function $H(\frac{1+x}{3}, \frac{1-2x}{3}, \frac{1-2x}{3}, x)$. The bound is tight.

Proof. The Huffman code for S has length vector (2,2,2,2) or (1,2,3,3). It is easy to see that if the length vector is (2,2,2,2) one has $p_1 \leq p_3 + p_4$, otherwise $p_1 \geq p_3 + p_4$. If $p_1 = p_3 + p_4$ then there are two Huffman codes with the two length vectors. Now we distinguish between the two cases:

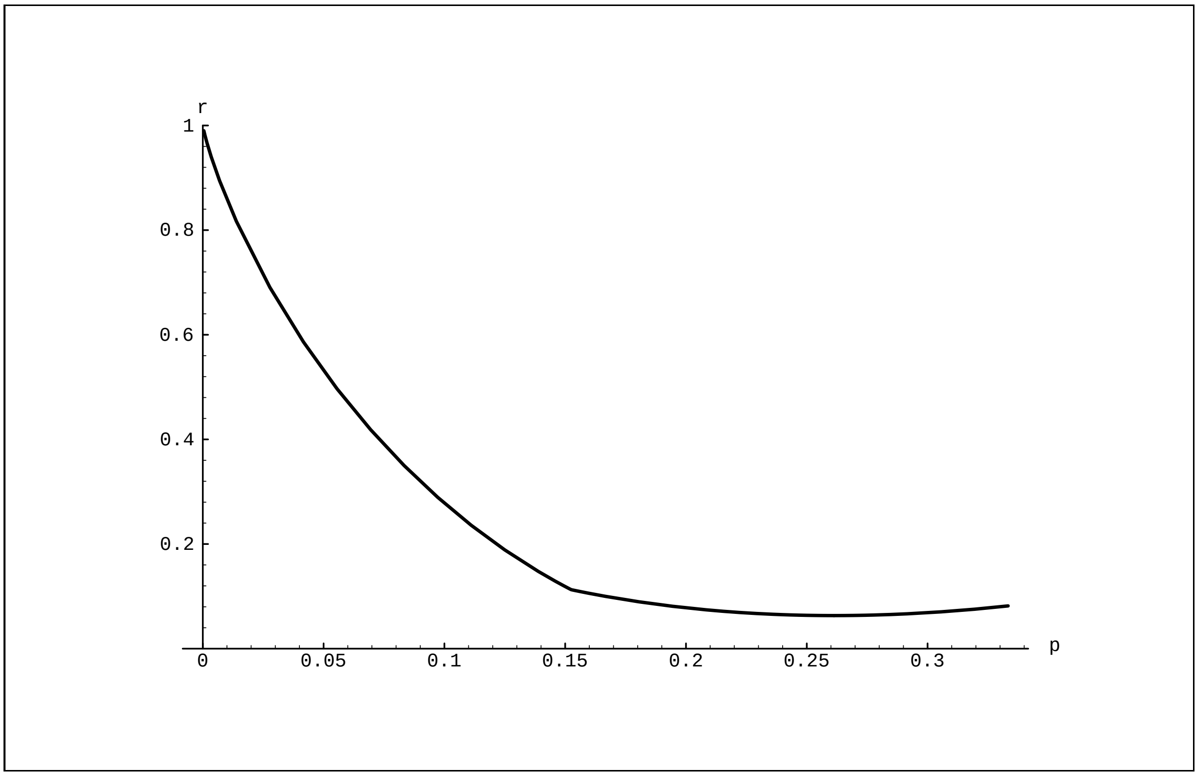

Figure 1: The upper bound on the redundancy r as function of p_N for sources with $N = 3$ letters

A) $p_1 < p_3 + p_N$
B) $p_1 \geq p_3 + p_N$.

<u>CASE A</u>: $p_1 < p_3 + p_N$. The expected codeword length is 2 and the redundancy is $r = 2 - H(p_1, p_2, p_3, p_4)$. For a fixed value of $p_4 = p_N$ and considering all possible values for p_1, p_2, p_3 we obtain that

$$r \leq 2 - \min_{(y_1, y_2, y_3, p_N) \varepsilon Q} H(y_1, y_2, y_3, p_N)$$

where $Q = \{(y_1, y_2, y_3, p_N) | y_1 \geq y_2 \geq y_3 \geq p_N, y_1 + y_2 + y_3 + p_N = 1, y_1 < y_3 + p_N\}$ is the set of sources with four letters whose least likely source letter has probability p_N and that satisfies $y_1 < y_3 + p_N$. Now we compute the minimum value of the entropy. The computation is based on the convexity of the function H. First note that the minimum over Q satisfies $y_2 = y_3$. Indeed, since $y_1 < y_3 + p_N$ if $y_2 > y_3$ then the point $y'_1 = y_1 + \epsilon, y'_2 = y_2 - \epsilon, y'_3 = y_3$ would satisfy $H(y'_1, y'_2, y'_3, p_N) < H(y_1, y_2, y_3, p_N)$, for small $\epsilon > 0$. Hence $y_2 = y_3$.
It is easy to see that the function to minimize, $H(1 - 2y_3 - p_N, y_3, y_3, p_N)$, is an

increasing function of y_3. Thus it assume the minimum value in the leftmost point in its interval of variation. From the conditions $y_3 \geq p_N$ and $y_1 < y_3 + p_N$ one has $\max\{p_N, (1-2p_N)/3\} \leq y_3$. The maximum between p_N and $(1-2p_N)/3$ is p_N if $p_N > 1/5$ and $(1-2p_N)/3$ otherwise. Hence

$$r \leq \begin{cases} 2 - H(1-3p_N, p_N, p_N, p_N) & \text{if } 1/5 < p_N \leq 1/4 \\ 2 - H(\frac{1+p_N}{3}, \frac{1-2p_N}{3}, \frac{1-2p_N}{3}, p_N) & \text{if } 0 < p_N \leq 1/5. \end{cases} \tag{13}$$

<u>CASE B</u>: $p_1 \geq p_3 + p_N$. Notice that $p_N \leq 1/5$. Indeed if $p_N > 1/5$ then $p_1 < p_3 + p_N$. The expected codeword length is $1 + p_2 + 2p_3 + 2p_4$. For a fixed value of $p_4 = p_N$ and considering all possible values for p_1, p_2, p_3 we obtain that

$$r \leq 1 + 2p_N + \max_{(y_1,y_2,y_3,p_N)\varepsilon Q}\{y_2 + 2y_3 - H(y_1, y_2, y_3, p_N)\}$$

where $Q = \{(y_1, y_2, y_3, p_N) | y_1 \geq y_2 \geq y_3 \geq p_N, y_1 + y_2 + y_3 + p_N = 1, y_1 \geq y_3 + p_N\}$ is the set of sources with four letters whose least likely source letter has probability p_N and that satisfies $y_1 > y_3 + p_N$. The function to maximize, $y_2 + 2y_3 - H(1 - y_2 - y_3 - p_N, y_2, y_3, p_N)$, is a convex $\cup$ function of y_2. Therefore it assumes the maximum at an extreme point of the variation interval of y_2. From conditions $y_1 \geq y_2 \geq y_3$ and $y_1 \geq y_3 + p_N$ we get $y_3 \leq y_2 \leq \min\{1 - 2y_3 - 2p_N, (1 - y_3 - p_N)/2\}$. It is easy to see that $1 - 2y_3 - 2p_N \leq (1 - y_3 - p_N)/2$ iff $y_3 \geq 1/3 - p_N$. We distinguish the three cases
B.1) $y_2 = 1 - 2y_3 - 2p_N$ and $y_3 \geq 1/3 - p_N$
B.2) $y_2 = (1 - y_3 - p_N)/2$ and $y_3 \leq 1/3 - p_N$
B.3) $y_2 = y_3$.

<u>CASE B.1</u>: $y_1 \geq y_3 + p_N, y_2 = 1 - 2y_3 - 2p_N$ and $y_3 \geq 1/3 - p_N$. The function to maximize is $1 - 2p_N - H(y_3 + p_N, 1 - 2p_N - 2y_3, y_3, p_N)$. This function is a convex $\cup$ function of y_3 and thus it assumes the maximum value at an extreme point of its variation interval. From conditions $y_1 \geq y_2 \geq y_3 \geq p_N$ one has $\max\{p_N, 1/3 - p_N\} \leq y_3 \leq (1 - 2p_N)/3$. The maximum between p_N and $1/3 - p_N$ is p_N iff $p_N \geq 1/6$.
If y_3 assume the value of the leftmost point of its variation interval we get:

$$r \leq \begin{cases} 2 - H(2p_N, 1 - 4p_N, p_N, p_N) & \text{if } 1/6 < p_N \leq 1/5 \\ 2 - H(1/3, 1/3, 1/3 - p_N, p_N) & \text{if } 0 < p_N \leq 1/6. \end{cases} \tag{14}$$

If $y_3 = (1 - 2p_N)/3$ then (13) holds.

<u>CASE B.2</u>: $y_1 \geq y_3 + p_N, y_2 = (1 - y_3 - p_N)/2$ and $y_3 \leq 1/3 - p_N$. Observe that $1/3 - p_N \geq y_3 \geq p_N$ implies that $p_N \leq 1/6$. The function to maximize, $(1 - y_3 - p_N)/2 + 2y_3 - H((1 - y_3 - p_N)/2, (1 - y_3 - p_N)/2, y_3, p_N)$, is a convex $\cup$ function of y_3, and thus it assumes the maximum at an extreme point of its variation interval. From

conditions $y_2 \geq y_3 \geq p_N$ and $y_1 \geq y_3 + p_N$ we get $p_N \leq y_3 \leq (1 - 3p_N)/3$.
If $y_3 = p_N$ we get

$$r \leq 1.5 + 3p_N - H((1 - 2p_N)/2, (1 - 2p_N)/2, p_N, p_N) \quad \text{if } 0 < p_N \leq 1/6. \tag{15}$$

If $y_3 = (1 - 3p_N)/3$ then (14) holds.
<u>CASE B.3</u>: $y_1 \geq y_3 + p_N$ and $y_2 = y_3$. The function to maximize, $y_3 + 2y_3 - H(1 - 2y_3 - p_N, y_3, y_3, p_N)$, is a convex $\cup$ function of y_3, and thus it assumes the maximum at an extreme point of its variation interval. From conditions $y_3 \geq p_N$ and $y_1 \geq y_3 + p_N$, one has $p_N \leq y_3 \leq (1 - 2p_N)/3$.
If $y_3 = (1 - 2p_N)/3$ then (15) holds, otherwise

$$r \leq 1 + 5p_N - H(1 - 3p_N, p_N, p_N, p_N) \quad \text{if } 0 < p_N \leq 1/5. \tag{16}$$

Now, comparing (13)-(16) and taking the maximum we get (12).
The bound is reached by the following sources:

$$\begin{cases} (1 - 3p_N, p_N, p_N, p_N) & \text{if } 0 < p_N \leq 1/9 \\ (1/3, 1/3, 1/3 - p_N, p_N) & \text{if } 1/9 < p_N \leq 1/6 \\ (2p_N, 1 - 4p_N, p_N, p_N) & \text{if } 1/6 < p_N \leq \delta_1 \\ (\frac{1+p_N}{3}, \frac{1-2p_N}{3}, \frac{1-2p_N}{3}, p_N) & \text{if } \delta_1 < p_N \leq 1/5 \\ (1 - 3p_N, p_N, p_N, p_N) & \text{if } 1/5 < p_N \leq 1/4. \end{cases}$$

This concludes the proof. □

Bound (12) is depicted in Figure 2.

3.3 An upper bound when N=5

In this section we consider sources consisting of five source letters and derive an upper bound as function of the least likely source letter probability.

Theorem 3 *Let $S = (p_1, p_2, p_3, p_4, p_5)$ be a discrete source and $p_5 = p_N$ be its least likely source letter probability. The redundancy of the corresponding Huffman code is upper bounded by:*

$$r \leq \begin{cases} 1 + 8p_N - H(1 - 4p_N, p_N, p_N, p_N, p_N) & \text{if } 0 < p_N \leq \delta_2 \\ 13/6 + p_N/2 - H(\frac{1}{3}, \frac{1}{3}, \frac{1-3p_N}{6}, \frac{1-3p_N}{6}, p_N) & \text{if } \delta_2 < p_N \leq 1/9 \\ 2 + 2p_N - H(3p_N, 1 - 6p_N, p_N, p_N, p_N) & \text{if } 1/9 < p_N \leq 1/8 \\ 2 + 2p_N - H(\frac{1-2p_N}{2}, \frac{1-4p_N}{2}, p_N, p_N, p_N) & \text{if } 1/8 < p_N \leq \delta_3 \\ 11/5 + 4p_N/5 - H(\frac{2(1-p_N)}{5}, \frac{1-p_N}{5}, \frac{1-p_N}{5}, \frac{1-p_N}{5}, p_N) & \text{if } \delta_3 < p_N \leq 1/6 \\ 2 + 2p_N - H(1 - 4p_N, p_N, p_N, p_N, p_N) & \text{if } 1/6 < p_N \leq \delta_4 \\ 9/4 + 3p_N/4 - H(\frac{1-p_N}{4}, \frac{1-p_N}{4}, \frac{1-p_N}{4}, \frac{1-p_N}{4}, p_N) & \text{if } \delta_4 < p_N \leq 1/5 \end{cases} \tag{17}$$

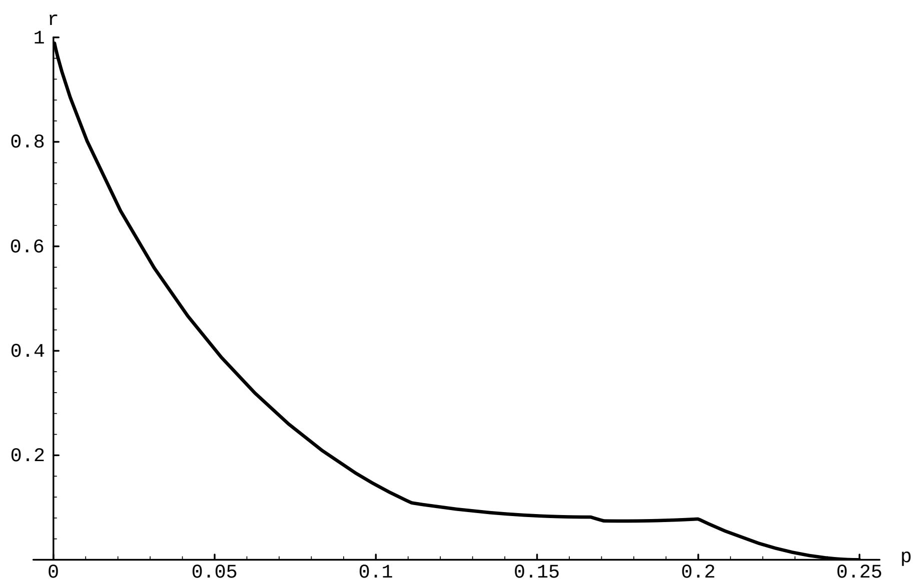

Figure 2: The upper bound on the redundancy r as function of p_N for sources with $N = 4$ letters

*where $\delta_2 \simeq 0.078184$ is the unique point in $]0, 1/9[$ for which the function $1 + 8x - H(1-4x, x, x, x, x)$ is equal to the function $13/6 + x/2 - H(\frac{1}{3}, \frac{1}{3}, \frac{1-3x}{6}, \frac{1-3x}{6}, x)$;
$\delta_3 \simeq 0.143815$ is the unique point in $]1/8, 1/6[$ for which the function $2 + 2x - H(\frac{1-2x}{2}, \frac{1-4x}{2}, x, x, x)$ is equal to the function $6/5 + 4x/5 - H(\frac{2(1-x)}{5}, \frac{1-x}{5}, \frac{1-x}{5}, \frac{1-x}{5}, x)$;
$\delta_4 \simeq 0.179669$ is the unique point in $]1/6, 1/5[$ for which the function $2 + 2x - H(1-4x, x, x, x, x)$ is equal to the function $5/4 + 3x/4 - H(\frac{1-x}{4}, \frac{1-x}{4}, \frac{1-x}{4}, \frac{1-x}{4}, x)$.
The bound is tight.*

Proof. We distinguish between four possible cases:

A) $p_2 \leq p_4 + p_5$ and $p_1 < p_2 + p_3$
B) $p_2 \leq p_4 + p_5$ and $p_1 \geq p_2 + p_3$
C) $p_2 > p_4 + p_5$ and $p_1 < p_3 + p_4 + p_5$
D) $p_2 > p_4 + p_5$ and $p_1 \geq p_3 + p_4 + p_5$.

CASE A: $p_2 \leq p_4 + p_5$ and $p_1 < p_2 + p_3$. It is easy to see that the Huffman code has length vector (2,2,2,3,3) and the expected codeword length is $2 + p_4 + p_5$. Maximizing the redundancy over all possible values of p_1, p_2, p_3, p_4, given that $p_5 = p_N$, we get

$$r \leq 2 + p_N + \max_{(y_1, y_2, y_3, y_4, p_N) \varepsilon Q} \{y_4 - H(y_1, y_2, y_3, y_4, p_N)\}$$

where $Q = \{(y_1, y_2, y_3, y_4, p_N) | y_1 \geq y_2 \geq y_3 \geq y_4 \geq p_N, y_1 + y_2 + y_3 + y_4 + p_N = 1, y_2 \leq y_4 + p_N, y_1 < y_2 + y_3\}$ is the set of sources with five letters whose probabilities satisfy the constraints of case A.
Observe that the maximum satisfies $y_2 = y_3$. Indeed, since $y_1 < y_2 + y_3$, if $y_2 > y_3$ the point $y'_2 = y_2 + \epsilon, y'_3 = y_3 - \epsilon, y'_i = y_i,\ i = 1, 4, 5$ would satisfy $H(y'_1, y'_2, y'_3, y_4, p_N) < H(y_1, y_2, y_3, y_4, p_N)$, for small $\epsilon > 0$. Following a similar reasoning we have that $y_3 = y_4$. Hence, the function to maximize is $y_4 - H(1 - 3y_4 - p_N, y_4, y_4, y_4, p_N)$. This is a convex $\cup$ function of y_4, and thus it assumes the maximum at an extreme point of y_4's variation interval. The conditions $y_1 \geq y_2, y_4 \geq p_N$ and $y_1 < y_2 + y_3$ imply that $\max\{p_N, (1 - p_N)/5\} \leq y_4 \leq (1 - p_N)/4$. The maximum between p_N and $(1 - p_N)/5$ is p_N if $p_N \geq 1/6$ and $(1 - p_N)/5$ otherwise. Hence

$$r \leq \begin{cases} 11/5 + 4p_N/5 - H(2\frac{1-p_N}{5}, \frac{1-p_N}{5}, \frac{1-p_N}{5}, \frac{1-p_N}{5}, p_N) & \text{if } 0 < p_N \leq 1/6 \\ 2 + 2p_N - H(1 - 4p_N, p_N, p_N, p_N, p_N) & \text{if } 1/6 < p_N \leq 1/5. \end{cases} \tag{18}$$

If $y_4 = (1 - p_N)/4$ we get

$$r \leq 9/4 + 3p_N/4 - H\left(\frac{1 - p_N}{4}, \frac{1 - p_N}{4}, \frac{1 - p_N}{4}, \frac{1 - p_N}{4}, p_N\right) \text{ if } 0 < p_N \leq 1/5 \tag{19}$$

<u>CASE B</u>: $p_2 \leq p_4 + p_5$ and $p_1 \geq p_2 + p_3$. Notice that $p_N \leq 1/6$. Indeed, if $p_N > 1/6$ one would has $\sum p_i > 1$. The Huffman code has length vector (1,3,3,3,3) and expected codeword length $3 - 2p_1$. Fixed $p_5 = p_N$, maximizing the redundancy over all possible values of $p_i,\ i = 1, 2, 3, 4$ one has:

$$r \leq 3 - \min_{(y_1, y_2, y_3, y_4, p_N) \varepsilon Q} \{2y_1 + H(y_1, y_2, y_3, y_4, p_N)\}$$

where $Q = \{(y_1, y_2, y_3, y_4, p_N) | y_1 \geq y_2 \geq y_3 \geq y_4 \geq p_N, y_1 + y_2 + y_3 + y_4 + p_N = 1, y_2 \leq y_4 + p_N, y_1 \geq y_2 + y_3\}$ is the set of sources with five letters whose probabilities satisfies the constraints of case B.
The minimum satisfies $y_4 = p_N$ or $y_2 = y_3$. Indeed, if $y_4 > p_N$ and $y_2 > y_3$ the point $y'_3 = y_3 + \epsilon, y'_4 = y_4 - \epsilon, y'_i = y_i,\ i = 1, 2, 5$ would have a smaller entropy. We distinguish between the two cases
B.1) $y_4 = p_N$
B.2) $y_2 = y_3$.

<u>CASE B.1</u> (The following inequalities hold: $y_2 \leq y_4 + y_5$, $y_1 \geq y_2 + y_3$ and $y_4 = p_N$). A point of minimum satisfies $y_3 = y_4$ or $y_2 = 2p_N$. Indeed, if $y_3 > y_4$ and $y_2 > 2p_N$

the point $y_3' = y_3 + \epsilon, y_4' = y_4 - \epsilon, y_i' = y_i,\ i = 1, 2, 5$ would have a smaller entropy. We distinguish between the cases

B.1.1) $y_3 = y_4$

B.1.2) $y_2 = 2p_N$

CASE B.1.1 (The following inequalities hold: $y_2 \le y_4 + y_5$, $y_1 \ge y_2 + y_3$, $y_4 = p_N$, $y_3 = y_4$). The function to minimize, $2(1 - y_2 - 3p_N) + H(1 - y_2 - 3p_N, y_2, p_N, p_N, p_N)$, is a convex $\cap$ function of y_2 and it assumes the maximum at an extreme point of its variation interval. From the conditions $y_2 \ge y_3$, $y_2 \le y_4 + p_N$ and $y_1 \ge y_2 + y_3$ one has $p_N \le y_2 \le \min\{2p_N, (1 - 4p_N)/2\}$. The minimum between $2p_N$ and $(1 - 4p_N)/2$ is $2p_N$ iff $p_N \le 1/8$. Hence

$$r \le \begin{cases} 1 + 10p_N - H(1 - 5p_N, 2p_N, p_N, p_N, p_N) & \text{if } 0 < p_N \le 1/8 \\ 2 + 2p_N - H(\frac{1-2p_N}{2}, \frac{1-4p_N}{2}, p_N, p_N, p_N) & \text{if } 1/8 < p_N \le 1/6. \end{cases} \tag{20}$$

If $y_4 = p_N$ we get

$$r \le 1 + 8p_N - H(1 - 4p_N, p_N, p_N, p_N, p_N) \quad \text{if } 0 < p_N \le 1/6. \tag{21}$$

CASE B.1.2 (The following inequalities hold: $y_2 \le y_4 + y_5$, $y_1 \ge y_2 + y_3$, $y_4 = p_N$, $y_2 = 2p_N$). Notice that $p_N \le 1/8$. Indeed if p_N 1/8 one would have $\sum p_i > 1$. The function to minimize, $2(1 - 4p_N - y_3) + H(1 - 4p_N - y_3, 2p_N, y_3, p_N, p_N)$, is a convex $\cap$ function of y_3 and it assumes the maximum at an extreme point of y_3's variation interval. From the conditions $y_1 \ge y_2 + y_3, y_2 \ge y_3 \ge y_4$ we get $p_N \le y_3 \le \min\{2p_N, (1 - 6p_N)/2\}$. The minimum between $2p_N$ and $(1 - 6p_N)/2$ is $2p_N$ iff $p_N \le 1/10$. Hence

$$r \le \begin{cases} 1 + 6p_N - H(1 - 6p_N, 2p_N, 2p_N, p_N, p_N) & \text{if } 0 < p_N \le 1/10 \\ 2 + 2p_N - H(\frac{1-2p_N}{2}, 2p_N, \frac{1-6p_N}{2}, p_N, p_N) & \text{if } 1/10 < p_N \le 1/8. \end{cases} \tag{22}$$

If $y_3 = p_N$ (20) holds.

CASE B.2 (The following inequalities hold: $y_2 \le y_4 + y_5$, $y_1 \ge y_2 + y_3$ and $y_2 = y_3$). The function to minimize, $2y_1 + H(y_1, \frac{1-y_1-y_4-p_N}{2}, \frac{1-y_1-y_4-p_N}{2}, y_4, p_N)$, is a convex $\cap$ function of y_4. Thus, it assumes the maximum value at an extreme point of y_4's variation interval. From conditions $y_1 \ge y_2, y_3 \ge y_4 \ge p_N$, and $y_2 \le y_4 + p_N$ one has $\max\{p_N, (1 - y_1 - 3p_N)/3\} \le y_4 \le 1 - 2y_1 - p_N$. The maximum between p_N and $(1 - y_1 - 3p_N)/3$ is p_N iff $p_N \ge (1 - y_1)/6$. We distinguish between the three cases

B.2.1) $y_4 = 1 - 2y_1 - p_N$

B.2.2) $y_4 = p_N$ and $p_N \ge (1 - y_1)/6$

B.2.3) $y_4 = (1 - y_1 - 3p_N)/3$ and $p_N < (1 - y_1)/6$.

CASE B.2.1 (The following inequalities hold: $y_2 \le y_4 + y_5$, $y_1 \ge y_2 + y_3$, $y_2 = y_3$, $y_4 = 1 - 2y_1 - p_N$). The function to minimize, $2y_1 + H(y_1, y_1/2, y_1/2, 1 - 2y_1 - p_N, p_N)$, is a convex $\cap$ function of y_1 and, thus, it assumes the maximum value at an extreme point of y_1's variation interval. From condition $y_1 \geq y_2 + y_3, y_3 \geq y_4 \geq p_N$ and $y_2 \leq y_4 + y_5$ one has $2(1-p_N)/5 \leq y_1 \leq \min\{(1-2p_N)/2, 2/5\}$. The minimum between $(1-2p_N)/2$ and $2/5$ is $2/5$ iff $p_N \leq 1/10$. Hence

$$r \leq \begin{cases} 11/5 - H(2/5, 1/5, 1/5, 1/5 - p_N, p_N) & \text{if } 0 < p_N \leq 1/10 \\ 2 + 2p_N - H(\frac{1-2p_N}{2}, \frac{1-2p_N}{4}, \frac{1-2p_N}{4}, p_N, p_N) & \text{if } 1/10 < p_N \leq 1/6. \end{cases} \tag{23}$$

If $y_1 = \frac{2(1-p_N)}{5}$ then (18) holds.

<u>CASE B.2.2</u> (The following inequalities hold: $y_2 \leq y_4 + y_5$, $y_1 \geq y_2 + y_3$, $y_2 = y_3$, $y_4 = p_N$ and $p_N \geq (1-y_1)/6$). The function to minimize $2(1-2y_3-2p_N)+H(1-2y_3-2p_N, y_3, y_3, p_N, p_N)$ is a convex $\cap$ function of y_3 and thus, assumes the maximum at an extreme point of its variation interval. From condition $y_1 \geq y_2 + y_3, y_3 \geq y_4, y_2 \leq y_4 + p_N$ and $p_N \geq (1-y_1)/6$ one gets $p_N \leq y_3 \leq \min\{2p_N, (1-2p_N)/4\}$. One has $2p_N \leq (1-2p_N)/4$ iff $p_N \leq 1/10$. If $y_3 = 2p_N$ we get

$$r \leq 1 + 12p_N - H(1-6p_N, 2p_N, 2p_N, p_N, p_N) \quad \text{if } 0 < p_N \leq 1/10. \tag{24}$$

If $y_3 = p_N$ or $y_3 = (1-2p_N)/4$ then respectively (21),(23) hold.

<u>CASE B.2.3</u> (The following inequalities hold: $y_2 \leq y_4 + y_5$, $y_1 \geq y_2 + y_3$, $y_2 = y_3$, $y_4 = (1-y_1-3p_N)/3$ and $p_N < (1-y_1)/6$). The function to minimize, $2y_1 + H(y_1, (1-y_1)/3, (1-y_1)/3, (1-y_1-3p_N)/3, p_N)$, is a convex $\cap$ function of y_1. From condition $y_1 \geq y_2 + y_3, y_4 \geq p_N$ and $p_N < (1-y_1)/6$ one gets $2/5 \leq y_1 \leq 1 - 6p_N$. In these extreme points (23) and (24) hold.

<u>CASE C</u>: $p_2 > p_4 + p_5$ and $p_1 < p_3 + p_4 + p_5$. Notice that $p_N \leq 1/7$. Indeed, if $p_N > 1/7$ one would have $\sum p_i > 1$. The Huffman code has length vector (2,2,2,3,3) and the expected codeword length is $2 + p_4 + p_5$. Like previous cases we fix the least probability $p_5 = p_N$, and consider all possible values of $p_i, i = 1, 2, 3, 4$, obtaining an upper bound for the redundancy.

$$r \leq 2 + p_N + \max_{(y_1, y_2, y_3, y_4, p_N) \varepsilon Q} \{y_4 - H(y_1, y_2, y_3, y_4, p_N)\}$$

where $Q = \{(y_1, y_2, y_3, y_4, p_N) | y_1 \geq y_2 \geq y_3 \geq y_4 \geq p_N, y_1 + y_2 + y_3 + y_4 + p_N = 1, y_2 > y_4 + p_N, y_1 < y_3 + y_4 + p_N\}$ is the set of sources with five letters whose probabilities satisfies the constraints of this case.

Observe that a point of maximum satisfies $y_1 = y_2$ or $y_3 = y_4$. Indeed, if $y_1 > y_2$ and $y_3 > y_4$ then point $y_2' = y_2 + \epsilon, y_3' = y_3 - \epsilon, y_i' = y_i,\ i = 1, 4, 5$ would satisfy $H(y_1', y_2', y_3', y_4', p_N) < H(y_1, y_2, y_3, y_4, p_N)$, for small $\epsilon > 0$. We distinguish between the

two cases
C.1) $y_1 = y_2$
C.2) $y_3 = y_4$.

<u>CASE C.1</u> (The following inequalities hold: $y_2 > y_4 + p_N$, $y_1 < y_3 + y_4 + p_N$ and $y_1 = y_2$). The function to maximize, $y_4 - H(\frac{1-y_3-y_4-p_N}{2}, \frac{1-y_3-y_4-p_N}{2}, y_3, y_4, p_N)$, is a decreasing function of y_3, and assumes the maximum value at leftmost point of y_3's variation interval. From conditions $y_3 \geq y_4$ and $y_1 < y_3 + y_4 + p_N$ one has $\max\{y_4, 1/3 - y_4 - p_N\} \leq y_3$. It is easy to see that $y_4 \geq 1/3 - y_4 - p_N$ iff $y_4 \geq (1-3p_N)/6$. We distinguish between the two cases
C.1.1) $y_3 = y_4$ and $y_4 \geq (1-3p_N)/6$
C.1.2) $y_3 = 1/3 - y_4 - p_N$ and $y_4 < (1-3p_N)/6$.

<u>CASE C.1.1</u> (The following inequalities hold: $y_2 > y_4 + p_N$, $y_1 < y_3 + y_4 + p_N$, $y_1 = y_2$, $y_3 = y_4$ and $y_4 \geq (1-3p_N)/6$). The function to maximize, $y_4 - H(\frac{1-2y_4-p_N}{2}, \frac{1-2y_4-p_N}{2}, y_4, y_4, p_N)$, is a convex $\cup$ function of y_4, thus assumes the maximum at an extreme point of its interval of variation. From conditions $y_4 \geq p_N, y_2 > y_4 + p_N, y_1 < y_3 + y_4 + p_N$ and $y_4 \geq (1-3p_N)/6$ one gets $\max\{p_N, (1-3p_N)/6\} \leq y_4 \leq (1-3p_N)/4$. p_N is greater than $(1-3p_N)/6$ iff $p_N \leq 1/9$. Hence

$$r \leq \begin{cases} 13/6 + p_N/2 - H(\frac{1}{3}, \frac{1}{3}, \frac{1-3p_N}{6}, \frac{1-3p_N}{6}, p_N) & \text{if } 0 < p_N \leq 1/9 \\ 2 + 2p_N - H(\frac{1-3p_N}{2}, \frac{1-3p_N}{2}, p_N, p_N, p_N) & \text{if } 1/9 < p_N \leq 1/7. \end{cases} \tag{25}$$

If $y_4 = (1-3p_N)/4$ then

$$r \leq 9/4 + p_N/4 - H(\frac{1+p_N}{4}, \frac{1+p_N}{4}, \frac{1-3p_N}{4}, \frac{1-3p_N}{4}, p_N) \quad \text{if } 0 < p_N \leq 1/7. \tag{26}$$

<u>CASE C.1.2</u> (The following inequalities hold: $y_4 + p_N$, $y_1 < y_3 + y_4 + p_N$, $y_1 = y_2$, $y_3 = 1/3 - y_4 - p_N$ and $y_4 \leq (1-3p_N)/6$). Observe that $p_N \leq y_4 \leq (1-3p_N)/6$ implies $p_N \leq 1/9$. The function to maximize, $y_4 - H(1/3, 1/3, 1/3 - y_4 - p_N, y_4, p_N)$, is a convex $\cup$ function of y_4, and assumes the greatest value at an extreme point of its variation interval. From conditions $y_3 \geq y_4 \geq p_N$ one gets $p_N \leq y_4 \leq (1-3p_N)/6$. Hence

$$r \leq 2 + 2p_N - H(1/3, 1/3, (1-6p)/3, p_N, p_N) \quad \text{if } 0 < p_N \leq 1/9. \tag{27}$$

If $y_4 = (1-3p_N)/6$ then (25) holds.

<u>CASE C.2</u> (The following inequalities hold: $y_2 > y_4 + p_N$, $y_1 < y_3 + y_4 + p_N$ and $y_3 = y_4$). Since $y_1 < y_3 + y_4 + p_N$ and $y_2 > y_4 + p_N$ a point in this case can not be of maximum. Indeed the point $y_1' = y_1 + \epsilon, y_2' = y_2 - \epsilon, y_i' = y_i, i = 3, 4$, would have smaller entropy.

CASE D: $p_2 > p_4 + p_5$ and $p_1 \geq p_3 + p_4 + p_N$. Notice that $p_N \leq 1/8$. Indeed if $p_N > 1/8$ one would has $\sum p_i > 1$. The Huffman code has length vector (1,2,3,4,4). Like the previous cases we upper bound the redundancy by

$$r \leq 1 + 3p_N + \max_{(y_1,y_2,y_3,y_4,p_N)\varepsilon Q}\{y_2 + 2y_3 + 3y_4 - H(y_1, y_2, y_3, y_4, p_N)\}$$

where $Q = \{(y_1, y_2, y_3, y_4, p_N)|y_1 \geq y_2 \geq y_3 \geq y_4 \geq p_N, y_1 + y_2 + y_3 + y_4 + p_N = 1, y_2 > y_4 + p_N, y_1 \geq y_3 + y_4 + p_N\}$ is the set of sources with five letters whose probabilities satisfies the constraints of case D.
The function to maximize, $y_2 + 2y_3 + 3y_4 - H(1 - y_2 - y_3 - y_4 - p_N, y_2, y_3, y_4, p_N)$, is a convex $\cup$ function of y_2 and thus assumes the maximum at an extreme point of y_2's variation interval. From conditions $y_1 \geq y_2 > y_2 + p_N$, $y_1 \geq y_3 + y_4 + p_N$ we get $y_4 + p_N \leq y_2 \leq \min\{\frac{1-y_3-y_4-p_N}{2}, 1 - 2y_3 - 2y_4 - 2p_N\}$. Moreover $\frac{1-y_3-y_4-p_N}{2} \leq 1 - 2y_3 - 2y_4 - 2p_N$ iff $y_3 + y_4 + p_N \leq 1/3$. We distinguish between the three cases
D.1) $y_2 = y_4 + p_N$
D.2) $y_2 = \frac{1-y_3-y_4-p_N}{2}$ and $y_3 + y_4 + p_N \leq 1/3$
D.3) $y_2 = 1 - 2y_3 - 2y_4 - 2p_N$ and $y_3 + y_4 + p_N > 1/3$.

CASE D.1 (The following inequalities hold: $y_2 > y_4+y_5$, $y_1 \geq y_3+y_4+p_N$, $y_2 = y_4+p_N$). The function to maximize, $(y_4+p_N)+2y_3+3y_4-H(1-y_3-2y_4-2p_N, y_4+p_N, y_3, y_4, p_N)$, is a convex $\cup$ function of y_3 and assumes the greatest value at an extreme point of y_3's variation interval. From conditions $y_1 \geq y_3 + y_4 + p_N$, $y_2 \geq y_3 \geq y_4$ one gets $y_4 \leq y_3 \leq \min\{y_4+p_N, \frac{1-3y_4-3p_N}{2}\}$. Furthermore $y_4+p_N \leq \frac{1-3y_4-3p_N}{2}$ iff $y_4+p_N \leq 1/5$. We distinguish the cases
D.1.1) $y_3 = y_4$
D.1.2) $y_3 = y_4 + p_N \leq 1/5$
D.1.3) $y_3 = \frac{1-3y_4-3p_N}{2}$ and $y_4 + p_N > 1/5$.

CASE D.1.1 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = y_4+p_N$, $y_3 = y_4$). The function to maximize, $y_4+p_N+2y_4+3y_4-H(1-3y_4-2p_N, y_4+p_N, y_4, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. By conditions $y_1 \geq y_2 > y_4 + p_N$ and $y_1 \geq y_3 + y_4 + p_N$ one gets $p_N \leq y_4 \leq (1 - 3p_N)/5$. Hence

$$r \leq 1 + 10p_N - H(1 - 5p_N, 2p_N, p_N, p_N, p_N) \text{ if } 0 < p_N \leq 1/8 \tag{28}$$

$$r \leq 11/5 + 2p_N/5 - H(\tfrac{2-p_N}{5}, \tfrac{1+2p_N}{5}, \tfrac{1-3p_N}{5}, \tfrac{1-3p_N}{5}, p_N) \text{ if } 0 < p_N \leq 1/8. \tag{29}$$

CASE D.1.2 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = y_4 + p_N$, $y_3 = y_4 + p_N \leq 1/5$). The condition $y_4 + p_N \leq 1/5$ together $p_N \leq y_4$ implies that $p_N \leq 1/10$. The function to maximize, $y_4 + p_N + 2(y_4 + p_N) + 3y_4 - H(1 - 3y_4 -$

$3p_N, y_4+p_N, y_4+p_N, y_4, p_N)$, is a convex $\cup$ function of y_4 and assume the maximum at an extreme point of its variation interval. From conditions $p_N \leq y_4$ and $y_1 \geq y_3+y_4+p_N$ one gets $p_N \leq y_4 \leq (1-5p_N)/5$. If $y_4 = p_N$ then (24) holds. If $y_4 = (1-5p_N)/5$ then (23) holds.

CASE D.1.3 (The following inequalities hold: $y_2 > y_4+y_5$, $y_1 \geq y_3+y_4+p_N$, $y_2 = y_4+p_N$, $y_3 = \frac{1-3y_4-3p_N}{2}$ and $y_4+p_N > 1/5$). The function to maximize, $y_4+p_N+1-3y_4-3p_N+3p_N-H(\frac{1-y_4-p_N}{2}, y_4+p_N, \frac{1-3y_4-3p_N}{2}, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. By conditions $y_2 \geq y_3 \geq y_4 \geq p_N$ one gets $\max\{p_N, (1-5p_N)/5\} \leq y_4 \leq (1-3p_N)/5$. Moreover, p_N is greater than $(1-5p_N)/5$ iff $p_N \geq 1/10$. If $y_4 = (1-5p_N)/5$ then (23) holds, if $y_4 = (1-3p_N)/5$ then (29) holds, and if $y_4 = p_N$ then (22) holds.

CASE D.2 (The following inequalities hold: $y_2 > y_4+y_5$, $y_1 \geq y_3+y_4+p_N$, $y_2 = \frac{1-y_3-y_4-p_N}{2}$ and $y_3+y_4+p_N \leq 1/3$). Observe that, since $y_3+y_4+p_N \leq 1/3$ one has $p_N \leq 1/9$. The function to maximize, $\frac{1-y_3-y_4-p_N}{2}+2y_3+3y_4-H(\frac{1-y_3-y_4-p_N}{2}, \frac{1-y_3-y_4-p_N}{2}, y_3, y_4, p_N)$, is a convex $\cup$ function of y_3 and assumes the maximum at an extreme point of its variation interval. From conditions $y_3 \geq y_4$ and $y_1 \geq y_3+y_4+p_N$ one gets $y_4 \leq y_3 \leq 1/3-y_4-p_N$. We distinguish the two cases

D.2.1) $y_3 = y_4$

D.2.2) $y_3 = 1/3-y_4-p_N$.

CASE D.2.1 (The following inequalities hold: $y_2 > y_4+y_5$, $y_1 \geq y_3+y_4+p_N$, $y_2 = \frac{1-y_3-y_4-p_N}{2}$,$y_3+y_4+p_N \leq 1/3$ and $y_3 = y_4$). The function to maximize, $\frac{1-2y_4-p_N}{2}+2y_4+3y_4-H((1-2y_4-p_N)/2, (1-2y_4-p_N)/2, y_4, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. By conditions $y_4 \geq p_N$ and $y_1 \geq y_3+y_4+p_N$ we have $p_N \leq y_4 \leq (1-3p_N)/6$. Hence, if $y_4 = p_N$ then

$$r \leq 3/2+13p_N/2-H\left(\frac{1-3p_N}{2}, \frac{1-3p_N}{2}, p_N, p_N, p_N\right) \quad \text{if } 0 < p_N \leq 1/9. \tag{30}$$

If $y_4 = (1-3p_N)/6$ then

$$r \leq 2+2p_N-H(1/3, 1/3, (1-3p_N)/6, (1-3p_N)/6, p_N) \quad \text{if } 0 < p_N \leq 1/9. \tag{31}$$

CASE D.2.2 (The following inequalities hold: $y_2 > y_4+y_5$, $y_1 \geq y_3+y_4+p_N$, $y_2 = \frac{1-y_3-y_4-p_N}{2}$,$y_3+y_4+p_N \leq 1/3$ and $y_3 = 1/3-y_4-p_N$). The function to maximize, $1/3+\frac{2}{3}(1-3y_4-3p_N)+3y_4-H(1/3, 1/3, \frac{1-3y_4-3p_N}{3}, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. By conditions $y_3 \geq y_4 \geq p_N$ one gets $p_N \leq y_4 \leq (1-3p_N)/6$. Hence If $y_4 = (1-3p_N)/6$ then (31) holds and if $y_4 = p_N$ then

$$r \leq 2+2p_N-H(1/3, 1/3, (1-6p_N)/3, p_N, p_N) \quad \text{if } 0 < p_N \leq 1/9. \tag{32}$$

CASE D.3 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = 1 - 2y_3 - 2y_4 - 2p_N$ and $y_3 + y_4 + p_N > 1/3$). The function to maximize, $(1 - 2y_3 - 2y_4) + 2y_3 + 3y_4 - H(y_3 + y_4 + p_N, 1 - 2y_3 - 2y_4, y_3, y_4, p_N)$, is a convex $\cup$ function of y_3 and assumes the maximum at an extreme point of its variation interval. From conditions $y_1 \geq y_2 \geq y_3 \geq y_4$ and $y_2 > y_4 + p_N$ one has that $\max\{\frac{1-3y_4-3p_N}{3}, y_4\} \leq y_4 \leq \min\{\frac{1-2y_4-2p_N}{3}, \frac{1-3y_4-3p_N}{2}\}$. It easy to see that $\frac{1-3y_4-3p_N}{3} \geq y_4$ iff $y_4 \leq (1-3p_N)/6$ and $\frac{1-2y_4-2p_N}{3} \leq \frac{1-3y_4-3p_N}{2}$ iff $y_4 \leq (1-5p_N)/5$. We distinguish among the four cases
D.3.1) $y_3 = \frac{1-3y_4-3p_N}{3}$ and $y_4 \leq (1-3p_N)/6$
D.3.2) $y_3 = y_4$ and $y_4 > (1-3p_N)/6$
D.3.3) $y_3 = \frac{1-2y_4-2p_N}{3}$ and $y_4 \leq (1-5p_N)/5$
D.3.4) $y_3 = \frac{1-3y_4-3p_N}{2}$ and $y_4 > (1-5p_N)/5$.

CASE D.3.1 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = 1 - 2y_3 - 2y_4 - 2p_N$, $y_3 + y_4 + p_N > 1/3$, $y_3 = \frac{1-3y_4-3p_N}{3}$ and $y_4 \leq (1-3p_N)/6$). By condition $p_N \leq y_4 \leq (1-3p_N)/6$ one has $p_N \leq 1/9$. The function to maximize , $1/3 + \frac{2}{3}(1 - 3y_4 - 3p_N) + 3y_4 - H(1/3, 1/3, \frac{1-3y_4-3p_N}{3}, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. From conditions of this case we get $p_N \leq y_4 \leq (1-3p_N)/6$. In both extreme point a bound already obtained holds. Precisely (31) and (32) hold.

CASE D.3.2 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = 1-2y_3-2y_4-2p_N$, $y_3+y_4+p_N > 1/3$, $y_3 = y_4$ and $y_4 > (1-3p_N)/6$). The function to maximize, $(1-4y_4-2p_N)+2y_4+3y_4-H(2y_4+p_N, 1-4y_4-2p_N, y_4, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. From conditions of this case we get $\max\{p_N, (1-3p_N)/6\} \leq y_4 \leq (1-3p_N)/5$. The maximum between p_N and $(1-3p_N)/6$ is p_N iff $p_N \geq 1/9$. Hence

$$r \leq 2 + 2p - H(3p_N, 1 - 6p_N, p_N, p_N, p_N) \quad \text{if } 1/9 < p_N \leq 1/8. \tag{33}$$

Whereas if $y_4 = (1-3p_N)/6$ or $y_4 = (1-3p_N)/5$ then (31) and (29) respectively holds.

CASE D.3.3 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = 1 - 2y_3 - 2y_4 - 2p_N$, $y_3 + y_4 + p_N > 1/3$, $y_3 = \frac{1-2y_4-2p_N}{3}$ and $y_4 \leq (1-5p_N)/5$). The condition $p_N \leq y_4 \leq (1-5p_N)/5$ implies that $p_N \leq 1/10$. The function to maximize, $\frac{1-2y_4-2p_N}{3} + \frac{2(1-2y_4-2p_N)}{3} + 3y_4 - H(\frac{1+y_4+p_N}{3}, \frac{1-2y_4-2p_N}{3}, \frac{1-2y_4-2p_N}{3}, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. From conditions of this case we get $p_N \leq y_4 \leq (1-5p_N)/5$. Hence

$$r \leq 2 + 2p_N - H\left(\frac{1+2p_{N_N}}{3}, \frac{1-4p_N}{3}, \frac{1-4p_N}{3}, p_N, p_N\right) \quad \text{if } 1/9 < p_N \leq 1/10. \tag{34}$$

If $y_4 = (1-5p_N)/5$ then (23) holds.

CASE D.3.4 (The following inequalities hold: $y_2 > y_4 + y_5$, $y_1 \geq y_3 + y_4 + p_N$, $y_2 = 1 - 2y_3 - 2y_4 - 2p_N$, $y_3 + y_4 + p_N > 1/3$, $y_3 = \frac{1-3y_4-3p_N}{2}$ and $y_4 > (1-5p_N)/5$). The function to maximize , $y_4 + p_N + 1 - 3y_4 - 3p_N + 3y_4 - H(\frac{1-y_4-p_N}{2}, y_4 + p_N, \frac{1-3y_4-3p_N}{2}, y_4, p_N)$, is a convex $\cup$ function of y_4 and assumes the maximum at an extreme point of its variation interval. From conditions of this case we get $\max\{p_N, (1-5p_N)/5\} \leq y_4 \leq (1-3p_N)/5$. Moreover $p_N \geq (1-5p_N)/5$ iff $p_N \geq 1/10$. If $y_4 = p_N, (1-5p_N)/5, (1-3p_N)/5$ then respectively (22),(23),(31) holds.

Comparing (18)-(34) and taking the maximum we get (17).
The bound is reached by the following sources:

$$\begin{cases} (1-4p_N, p_N, p_N, p_N, p_N) & \text{if } 0 < p_N \leq \delta_1 \\ (1/3, 1/3, (1-3p_N)/6, (1-3p_N)/6, p_N) & \text{if } \delta_1 < p_N \leq 1/9 \\ (3p_N, 1-6p_N, p_N, p_N, p_N) & \text{if } 1/9 < p_N \leq 1/8 \\ ((1-2p_N)/2, (1-4p_N)/2, p_N, p_N, p_N) & \text{if } 1/8 < p_N \leq \delta_2 \\ (2(1-p_N)/5, (1-p_N)/5, (1-p_N)/5, (1-p_N)/5, p_N) & \text{if } \delta_2 < p_N \leq 1/6 \\ (1-4p_N, p_N, p_N, p_N, p_N) & \text{if } 1/6 < p_N \leq \delta_3 \\ ((1-p_N)/4, (1-p_N)/4, (1-p_N)/4, (1-p_N)/4, p_N) & \text{if } \delta_3 < p_N \leq 1/5. \end{cases}$$

This concludes the proof. $\square$

Bound (17) is depicted in Figure 3.

A simple but tedious calculus shows that the functions that represents the bound decreases as N increases, for $N \leq 5$ and for any fixed value of the least likely source letter probability. More formally, we have the following

Lemma 4 *Let $\alpha(p_N)$, $\beta(p_N)$ and $\gamma(p_N)$ be the functions that represent the bounds (11), (12), and (17) respectively, i.e. the bounds for $N = 3, 4, 5$. Then,*

$$\begin{array}{rcll} \alpha(p_N) & \geq & \beta(p_N) & for \quad 0 < p_N \leq 1/4 \\ \beta(p_N) & \geq & \gamma(p_N) & for \quad 0 < p_N \leq 1/5. \end{array}$$

Figure 4 shows bounds (1), (11), (12), and (17).

4 An upper bound when only p_N is known and $N \geq 3$

The bound $r \leq 1 - \mathcal{H}(p_N)$ is achieved by sources of two letters. In this section we prove that whenever $N \geq 3$ a sharper bound holds. More precisely, we prove that bound (3) holds whenever $N \geq 3$.

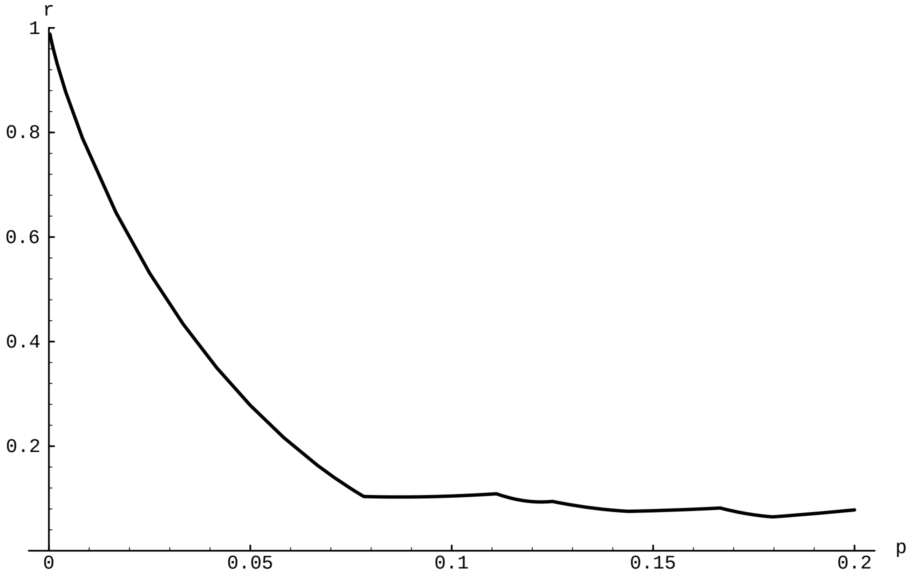

Figure 3: The upper bound on the redundancy r as function of p_N for sources with $N = 5$ letters

Theorem 4 *Let S be a source with $N \geq 3$ letters and p_N its least likely source letter. The redundancy of the corresponding Huffman code is upper bounded by:*

$$r \leq \begin{cases} 1 - \mathcal{H}(2p_N), & \text{if } 0 < p_N \leq \delta \\ 0.5 + 1.5p_N - \mathcal{H}(p_N), & \text{if } \delta < p_N \leq 1/3 \end{cases} \tag{35}$$

where $\delta \simeq 0.1525$ is the unique zero of the function $0.5 - 1.5x - \mathcal{H}(2x) + \mathcal{H}(x)$ in the interval $[0, 0.3]$. The bound is tight.

Proof. By Lemma 4 the redundancy satisfies (35), for $N = 3, 4, 5$. Hence we assume $N \geq 6$ and thus $p_N \leq 1/6$.
We distinguish between the two cases $p_N \leq 1/7$ and $1/7 < p_N \leq 1/6$.

CASE A: $p_N \leq 1/7$. We have to prove that $r \leq 1 - \mathcal{H}(2p_N)$. Now we distinguish among few possible cases according to the value of the most likely source letter probability p_1.

CASE A.1: $0 < p_1 \leq 1/6$. From (5) we have $r \leq p_1 + 0.0860 - p_N$ which is $\leq 1/6 + 0.0860 - p_N < 0.253 - p_N < 1 - \mathcal{H}(2p_N)$.

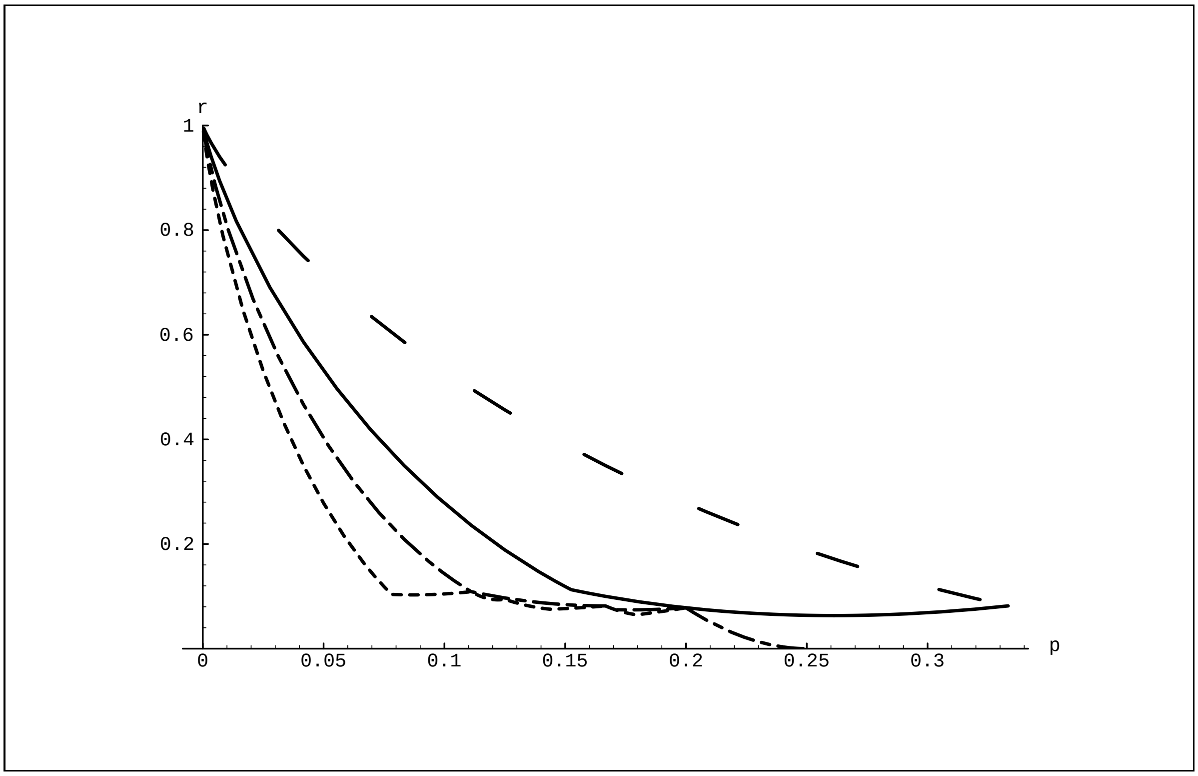

Figure 4: Upper bounds on the redundancy r as functions of p_N for sources with $N = 2, 3, 4, 5$ letters

<u>CASE A.2</u>: $1/6 < p_1 \le 0.1971$. From (6) we have that $r \le 2 - 1.3219(1 - p_1) - \mathcal{H}(p_1) - p_N$. The right-hand side is a decreasing function of p_1, and thus $r \le 2 - 1.3219(1 - 1/6) - \mathcal{H}(1/6) - p_N < 0.249 - p_N < 1 - \mathcal{H}(2p_N)$.

<u>CASE A.3</u>: $0.1971 < p_1 \le 0.2$. From (7) we have that $r \le 4 - 18.6096p_1 - \mathcal{H}(5p_1) - p_N$. The right-hand side is a decreasing function of p_1, and thus $r \le 4 - (18.6096)(0.2) - \mathcal{H}(1) - p_N < 0.278 - p_N < 1 - \mathcal{H}(2p_N)$.

<u>CASE A.4</u>: $0.2 < p_1 \le 0.3138$. From (8) we have that $r \le 2 - 1.25(1 - p_1) - \mathcal{H}(p_1) - p_N$. The right-hand side is an increasing function of p_1, and thus $r \le 2 - 1.25(1 - 0.2) - \mathcal{H}(0.2) - p_N < 0.278 - p_N < 1 - \mathcal{H}(2p_N)$.

<u>CASE A.5</u>: $0.3138 < p_1 < 1/3$. From (9) we have that $r \le 3 - 3(1 + \log 3)p_1 - \mathcal{H}(3p_1) - p_N$. The right-hand side is a decreasing function of p_1, and thus $r \le 3 - 3(1+\log 3)(1/3) - \mathcal{H}(1) - p_N < 0.415 - p_N$. It is easy to see that $0.415 - p_N < 1 - \mathcal{H}(2p_N)$ for $p_N \le 0.0884$. Now assume $p_N > 0.0884$. Let t be the number of leaves at level 2 in the Huffman code tree for the source $S = (p_1, \ldots, p_N)$. Since $N \ge 6$ then $t \le 3$.
As stated in [3] (see Theorem 3, equation (35)), the redundancy can be upper bounded by $r \le l - H(q''_1, \ldots, q''_{2^l}) + q_{2m-1} - p_N$, where l is a level for which the code tree is

complete, and m is the smallest integer such that node $2m-1$ with probability q_{2m-1} in the sibling list is at level $l+1$, and $q''_i, i=1,\ldots,2^l$ are the entries in the sibling list at level l. We consider the case $l=2$. For $t=1,2$,Capocelli and De Santis [2] proved that $2-H(q''_1,\ldots,q''_{2^l})+q_{2m-1}\le 0.75+1.25p_1-\mathcal{H}(p_1)$. Hence $r\le 0.75+1.25p_1-\mathcal{H}(p_1)-p_N<0.278-p_N<1-\mathcal{H}(2p_N)$ for $p_N\le 1/7$.
Suppose $t=3$. Since $N\ge 6$, $p_N>0.0884$ and $0.3138<p_1<1/3$, recalling (4), one has that an optimal code for S must have length vector (2,2,2,3,4,4). Let $S=(p_1,\ldots,p_6)$ be a source of 6 letters. There are two possible situations in which Huffman procedure generates a code whose length vector is (2,2,2,3,4,4):
(a) $p_3\ge p_5+p_6$, $p_2\ge p_4+p_5+p_6$, and $p_1\le p_3+p_4+p_5+p_6$
(b) $p_3\ge p_5+p_6$, $p_2\le p_4+p_5+p_6$, and $p_1\le p_2+p_3$.
Case (a) is not possible: indeed, since $p_6>0.0884$ and $p_1\ge 0.3138$, we would have $\sum p_i\ge p_1+(p_4+p_5+p_6)+(p_5+p_6)+p_4+p_5+p_6\ge p_1+8p_6>0.3138+8(0.0884)>1$. In case (b), nodes with probabilities p_2 and p_3 are merged at level 2 of the code tree. Observe that $p_2+p_3\ge 2p_3\ge 2(p_5+p_6)\ge 4p_6>1/3$, and since $p_1<1/3$ then $p_1<p_2+p_3$. The redundancy is $r=2+p_4+2p_5+2p_6-H(p_1,p_2,p_3,p_4,p_5,p_6)$. It is well known that $H(p_1+\epsilon,p_2-\epsilon,p_3,p_4,p_5,p_6)>H(p_1,p_2,p_3,p_4,p_5,p_6)$ for small ϵ (that is, for $\epsilon<\min\{p_2-p_3,1-p_1\}$). Since $p_1<p_2+p_3$, if $p_2>p_3$ then the source $(p_1+\epsilon,p_2-\epsilon,p_3,p_4,p_5,p_6)$ would have a redundancy greater than that of the source $(p_1,p_2,p_3,p_4,p_5,p_6)$. Thus the maximum redundancy, for all sources satisfying $p_3\ge p_5+p_6$, $p_2\le p_4+p_5+p_6$, and $p_1<p_2+p_3$, for a fixed value of p_6, is reached when $p_2=p_3$. On other hand, if $p_2=p_3$ it is easy to see that the redundancy r of S is equal to the redundancy r' of the source of five letters $(p_2+p_3,p_1,p_2,p_3,p_4)$. From Theorem 3 and Lemma 4 we have $r'\le 1-\mathcal{H}(2p_N)$.

<u>CASE A.6</u>: $1/3<p_1\le 0.4505$. From (10) we have that $r\le 1+0.5(1-p_1)-\mathcal{H}(p_1)-2p_N$. The right-hand side is a decreasing function of p_1, and thus $r\le 1+0.5(1-1/3)-\mathcal{H}(1/3)-2p_N<0.415-2p_N<1-\mathcal{H}(2p_N)$ for $p_N\le 1/7$.

<u>CASE A.7</u>: $0.4505<p_1$. Let B be the source obtained by deleting the first letter in S and normalizing the remaining probabilities, i.e. $B=(p_2/(1-p_1),\ldots,p_N/(1-p_1))$. Let L and L_B be the expected codeword lengths of sources S and B, and H and H_B their entropies. It is well known that

$$H=\mathcal{H}(p_1)+(1-p_1)H_B$$

and

$$L=1+(1-p_1)L_B.$$

Using bound (1), we have

$$\begin{aligned} r &= L-H \\ &= 1-\mathcal{H}(p_1)+(1-p_1)r_B \end{aligned}$$

$$
\begin{aligned}
&\leq 1-\mathcal{H}(p_1)+(1-p_1)\left[1-\mathcal{H}\left(\frac{p_N}{1-p_1}\right)\right] \\
&= 2-p_1-\mathcal{H}(p_1)-(1-p_1)\mathcal{H}\left(\frac{p_N}{1-p_1}\right) \\
&= 2-p_1-H(p_1,1-p_1-p_N,p_N)
\end{aligned}
$$

The function $2-x-H(x,1-x-p_N,p_N)$ is a convex $\cup$ function of x and thus it assumes the maximum value at an extreme point of x's variation interval. Since $N \geq 6$ one has $0.4505 \leq x \leq 1-5p_N \leq 1-2p_N$. If $x=1-2p_N$ we have that $2-(1-2p_N)-H(1-2p_N,p_N,p_N) \leq 1-\mathcal{H}(2p_N)$. If $x=0.4505$ we have

$$r \leq 2-0.4505-\mathcal{H}(0.4505)-(0.5495)\mathcal{H}\left(\frac{p_N}{0.5495}\right)$$

which by a simple calculus is less than $1-\mathcal{H}(2p_N)$ for $p_N \leq 1/7$.

CASE B: $1/7 < p_N \leq 1/6$. Since $p_N > 1/7$ and $N \geq 6$ then $N=6$. Thus $1/6 \leq p_1 \leq 1-5p_N = 0.28$. By Lemma 2, we have that, for $1/6 \leq p_1 \leq 0.28$, the minimum codeword length of the Huffman code satisfies either $n_1=2$ or $n_1=3$. If $n_1=3$ then N would be ≥ 8, therefore $n_1=2$. Recalling (4) we have that the Huffman code has length vector (2,2,3,3,3,3). Observe that if $p_1=p_2$ or $p_3=p_4$ then the redundancy of source S is equal to the redundancy of a source of five letters. From Theorem 3 and Lemma 4, (35) holds. Hence assume $p_1>p_2$ and $p_3>p_4$. The redundancy is equal to $r=2+p_3+p_4+p_5+p_6-H(p_1,p_2,p_3,p_4,p_5,p_6)$. It is well known that $H(p_1,p_2,p_3,p_4+\epsilon,p_5-\epsilon,p_6) > H(p_1,p_2,p_3,p_4,p_5,p_6)$, for small $\epsilon>0$. Since $p_3>p_4$, if $p_5>p_6$ then the source $(p_1,p_2,p_3,p_4+\epsilon,p_5-\epsilon,p_6)$ would have a redundancy r' greater than the redundancy r of the source S. Thus a source with maximum redundancy, among all sources satisfying the constraints of this case, for a fixed value of the least likely source letter probability p_6, has $p_5=p_6$. Hence the maximum redundancy is equal to the redundancy of a source consisting of five letters whose least likely source letter probability is p_4. From Lemma 4 the redundancy of a source of five letters whose least likely source letter probability is p_4, is less than the redundancy of a source of three letters whose least likely source letter probability is p_4. On the other hand $p_4 \leq (1-p_5-p_6)/4 \leq (1-2/7)/4 \leq 0.18$. Thus, recalling that $p_N \leq p_4$, from Lemma 3 we have that $\alpha(p_N) > \alpha(p_4) > r$.

$\square$

5 An upper bound as function of p_{N-1} and p_N

In this section, exploiting the bound provided in Theorem 4, we obtain an upper bound as function of p_{N-1} and p_N which improve Yeung's bound (2).

Theorem 4 tells us that $r \leq \alpha(p_N)$ where, recalling the definition of α,

$$\alpha(p_N) = \begin{cases} \phi(p_N) & \text{if } 0 < p_N \leq \delta \\ \psi(p_N) & \text{if } \delta < p_N \leq 1/3 \end{cases}$$

and $\phi(x) = 1 - \mathcal{H}(2x)$ and $\psi(x) = 0.5 + 1.5 - \mathcal{H}(x)$.
The two special cases $N = 2$ and $N = 3$ are trivial since we know all probabilities of the source, and we can compute the exact value of the redundancy. We have

$$r = 1 - \mathcal{H}(p_{N-1}) = 1 - \mathcal{H}(p_N) \text{ if } N = 2$$

$$r = 1 + p_{N-1} + p_N - H(1 - p_{N-1} - p_N, p_{N-1}, p_N) \text{ if } N = 3$$

Thus, we suppose that $N \geq 4$. Let $A = (p_1, \ldots, p_{N-1}, p_N)$ be a source with N letters and let B the source obtained merging p_{N-1} and p_N, i.e. with source letter probabilities $\{p_1, \ldots, p_{N-2}, p_{N-1} + p_N\}$. The redundancies r and r_B of sources A and B respectively, satisfy

$$r = r_B + l(p_{N-1}, p_N) \tag{36}$$

where $l(p_{N-1}, p_N) = (p_{N-1} + p_N) + \left[1 - \mathcal{H}\left(\frac{p_N}{p_N + p_{N-1}}\right)\right]$ is the contribute to the redundancy due to p_{N-1} and p_N. Since A has at least four letters, B has at least three letters and thus we can upper bound r_B using Theorem 4. We distinguish between the two cases $p_{N-1} + p_N \leq p_{N-2}$ and $p_{N-1} + p_N > p_{N-2}$.

Consider the case $p_{N-1} + p_N \leq p_{N-2}$. The least likely source letter probability of B is $p_{N-1} + p_N$. Since B has at least three letters, we have $p_{N-1} + p_N \leq 1/3$. By Theorem 4, r_B is upper bounded by

$$r_B \leq \begin{cases} \phi(p_N + p_{N-1}), & \text{if } 0 < p_N + p_{N-1} \leq \delta \\ \psi(p_N + p_{N-1}), & \text{if } \delta < p_N + p_{N-1} \leq 1/3 \end{cases} \tag{37}$$

Consider the case $p_{N-1} + p_N > p_{N-2}$. The least likely source letter probability of B is p_{N-2} and it satisfies

$$p_{N-1} \leq p_{N-2} \leq \min\{1/3, p_{N-1} + p_N\}.$$

By Theorem 4 we have that

$$r_B \leq \max_{p_{N-1} \leq x \leq \min\{1/3, p_{N-1} + p_N\}} \alpha(x).$$

By Lemma 3, if $p_{N-1} \leq \delta$ then $\alpha(p_{N-1})$ is the maximum of $\alpha(x)$ in $[p_{N-1}, \min\{1/3, p_{N-1} + p_N\}]$ and thus

$$r_B \leq \phi(p_{N-1}) \text{ if } p_{N-1} \leq \delta. \tag{38}$$

Assume $p_{N-1} > \delta$.We distinguish the two cases $p_{N-1} + p_N \leq 1/3$ and $p_{N-1} + p_N > 1/3$. In the case $p_{N-1} + p_N \leq 1/3$, by the convexity of ψ we have

$$r_B \leq \max\{\psi(p_N + p_{N-1}), \psi(p_{N-1})\} \text{ if } p_{N-1} > \delta \text{ and } p_{N-1} + p_N \leq 1/3 \tag{39}$$

whereas, in the case $p_{N-1} + p_N > 1/3$ we have

$$r_B \leq \max\{\psi(1/3), \psi(p_{N-1})\} \text{ if } p_{N-1} > \delta \text{ and } p_{N-1} + p_N > 1/3. \tag{40}$$

Observe that $p_{N-1} + p_N > 1/3$ implies $p_{N-1} > \delta$.

Comparing (37)-(40) and taking the maximum we obtain the following upper bound on r_B,

$$r_B \leq \begin{cases} \max\{\phi(p_N + p_{N-1}), \phi(p_{N-1})\} & \text{if } p_{N-1} + p_N \leq \delta \\ \max\{\psi(p_N + p_{N-1}), \phi(p_{N-1})\} & \text{if } p_{N-1} + p_N > \delta, p_{N-1} \leq \delta \\ \max\{\psi(p_N + p_{N-1}), \psi(p_{N-1})\} & \text{if } p_{N-1} > \delta, p_{N-1} + p_N \leq 1/3 \\ \max\{\psi(p_{N-1}), \phi(1/3)\} & \text{if } p_{N-1} + p_N > 1/3 \end{cases}$$

Substituting above bound in (36), and observing that $\max\{\phi(p_N + p_{N-1}), \phi(p_{N-1})\} = \phi(p_{N-1})$, we have

$$r \leq \omega(p_{N-1}, p_N) \tag{41}$$

where

$$\omega(p_{N-1}, p_N) = \begin{cases} \phi(p_{N-1}) + l(p_{N-1}, p_N) & \text{if } p_{N-1} + p_N \leq \delta \\ \max\{\psi(p_N + p_{N-1}), \phi(p_{N-1})\} + l(p_{N-1}, p_N) & \text{if } p_{N-1} + p_N > \delta, p_{N-1} \leq \delta \\ \max\{\psi(p_N + p_{N-1}), \psi(p_{N-1})\} + l(p_{N-1}, p_N) & \text{if } p_{N-1} > \delta, p_{N-1} + p_N \leq 1/3 \\ \max\{\psi(p_{N-1}), \phi(1/3)\} + l(p_{N-1}, p_N) & \text{if } p_{N-1} + p_N > 1/3 \end{cases}$$

The above results can be summarized in the following theorem

Theorem 5 *Let $S = (p_1, \ldots, p_{N-1}, p_N)$ be a discrete source. The redundancy of the corresponding Huffman code is upper bounded by:*

$$r \leq 1 - \mathcal{H}(p_N) \textit{ if } p_{N-1} + p_N = 1 \tag{42}$$

$$r \leq 1 + p_{N-1} + p_N - H(1 - p_{N-1} - p_N, p_{N-1}, p_N) \textit{ if } (1 - p_N)/4 < p_{N-1} \leq (1 - p_N)/3 \tag{43}$$

$$r \leq \max\{1 + p_{N-1} + p_N - H(1 - p_{N-1} - p_N, p_{N-1}, p_N), \omega(p_{N-1}, p_N)\} \textit{ if } p_{N-1} \leq (1 - p_N)/4 \tag{44}$$

Notice that if $p_{N-1} + p_N < 1$ then $p_{N-1} \leq (1 - p_N)/3$.
Figure 5 shows bound (44).

Following a similar reasoning but using $1 - \mathcal{H}(p_N)$ for upper bounding r_B instead of (3) we get the bound provided by Yeung. Therefore it is clear that the bound (41) is sharper than the bound of Yeung, since bound (3) is sharper than $1 - \mathcal{H}(p_N)$.

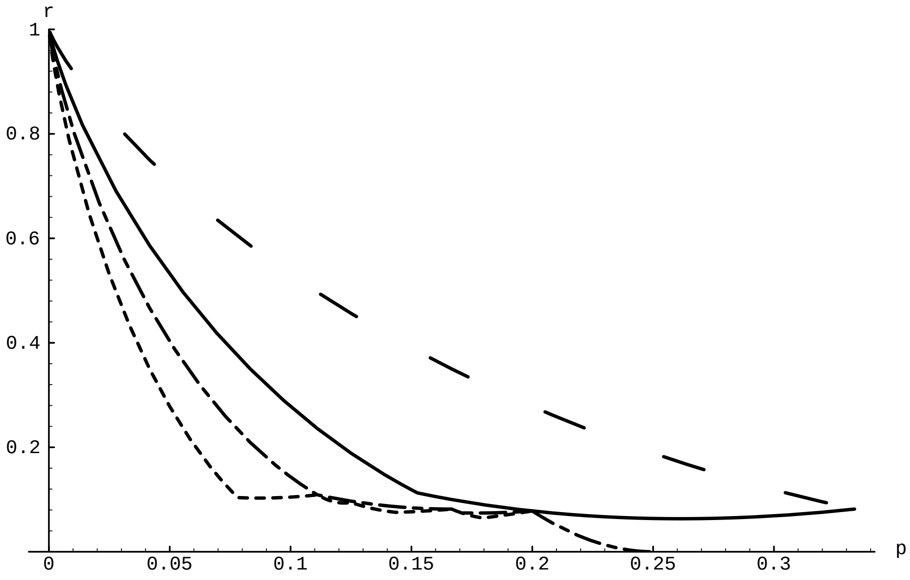

Figure 5: An Upper bounds on the redundancy r as functions of p_{N-1} and p_N

References

[1] R. M. Capocelli and A. De Santis, "Tight upper bounds on the redundancy of Huffman codes", *IEEE Trans. Inform. Theory*, vol. IT-35, n. 5, pp. 1084–1091, Sept 1989.

[2] R. M. Capocelli and A. De Santis, "New bounds on the redundancy of Huffman codes", *IEEE Trans. Inform. Theory*, vol. IT-37, n. 4, pp. 1095–1104, July 1991.

[3] R. M. Capocelli, R. Giancarlo, and I. J. Taneja, "Bounds on the redundancy of Huffman codes", *IEEE Trans. Inform. Theory*, vol. IT-32, n. 6, pp. 854–857, Dec 1986.

[4] R. G. Gallager, "Variation on a theme by Huffman", *IEEE Trans. Inform. Theory*, vol. IT-24, n. 6, pp. 668–674, Dec 1978.

[5] Y. Horibe, "An improved bound for weight-balanced trees", *Inform. Contr.*, vol. 24, n. 2, pp. 148–151, 1977.

[6] D. A. Huffman, "A Method for the construction of minimum redundancy codes", *Proc. IRE*, 40, n.2, pp. 1098–1101, 1952.

[7] O. Johnsen, "On the redundancy of Huffman codes", *IEEE Trans. Inform. Theory*, vol. IT-26, n. 2, pp. 220–222, Mar 1980.

[8] G.O.H. Katona and T.O.H. Nemetz, "Huffman codes and self-information", *IEEE Trans. Inform. Theory*, vol. IT-22, n. 3, May 1976.

[9] F. M. Reza, *Introduction to information theory.* New York: McGraw-Hill, 1951.

[10] D.Manstetten, "Tight bounds on the redundancy of Huffman codes", *IEEE Trans. Inform. Theory*, vol. IT-38, n. 1, Jan 1992.

[11] B.L.Montgomery and J.Abrahams, "On the redundancy of optimal prefix-condition codes for finite and infinite sources", *IEEE Trans. Inform. Theory*, vol.IT-33, n. 1, pp. 156–160, Jan 1987.

[12] B.L.Montgomery and B.V.K.V.Kumar, "On the average codeword length of optimal binary codes for extended sources", *IEEE Trans. Inform. Theory* , vol. IT-33, n. 2, pp. 293–296, Mar 1987.

[13] R. Yeung, "The redundancy theorem and new bounds of the expected length of the Huffman code", *IEEE Trans. Inform. Theory*, vol. IT-37, n. 3, pp. 687–691 May 1991.