[3] R. M. Capocelli, A. De Santis, and G. Persiano, "Binary prefix codes ending in a '1'," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1296–1302, July 1994.
[4] S. Chan and M. Golin, "A dynamic programming algorithm for constructing optimal '1'-ended binary prefix-free codes," in *Proc. IEEE Int. Symp. Information Theory*, Boston, MA, 1998, p. 45.
[5] D. A. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE*, vol. 40, pp. 1098–1101, 1952.
[6] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic, 1979.
[7] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1997, ISBN 0-691-01586-4.

A Simple Upper Bound on the Redundancy of Huffman Codes

Chunxuan Ye and Raymond W. Yeung, *Senior Member, IEEE*

***Abstract*—Upper bounds on the redundancy of Huffman codes have been extensively studied in the literature. Almost all of these bounds are in terms of the probability of either the most likely or the least likely source symbol. In this correspondence, we prove a simple upper bound in terms of the probability of any source symbol.**

***Index Terms*—Huffman code, prefix code, redundancy.**

I. Introduction

Let $\{p_1, \ldots, p_n\}$ be the probability distribution of a source, and let C be a prefix code for the source. The redundancy of a code C is defined as the difference between the average codeword length L of the code and $H = -\sum_i p_i \log p_i$, the entropy of the source. (In this correspondence, all logarithms are of base 2.) We define R as the redundancy of a Huffman code. It is well known that $0 \leq R < 1$. These bounds on R are the tightest possible when nothing is known about the source distribution. However, when partial knowledge about the source distribution is available, these bounds can be improved. Gallager [5] proved that if p_1, the probability of the most likely source symbol, is given, the upper bound on R can be improved. His result is summarized as

$$R \leq \begin{cases} p_1 + \sigma, & \text{if } p_1 < \frac{1}{2} \\ 2 - p_1 - H_b(p_1), & \text{if } p_1 \geq \frac{1}{2} \end{cases} \tag{1}$$

where

$$H_b(x) = -x \log_2 x - (1-x) \log_2 (1-x)$$

and

$$\sigma = 1 - \log_2 e + \log_2(\log_2 e) \approx 0.086.$$

This upper bound is tight for $p_1 \geq \frac{1}{2}$, but it is not tight for $p_1 < \frac{1}{2}$. Subsequently, a number of improvements on the upper bound on R in terms of p_1 with $p_1 < \frac{1}{2}$ have been obtained [1]–[4], [6]–[9]. On the other hand, upper bounds on R in terms of p_n, the probability of the least likely source symbol, have also been obtained [1], [4], [10], [12].

Johnsen [6] obtained a tight lower bound on R in terms of p_1. His result is that for $p_1 \geq 0.4$, $R \geq 1 - H_b(p_1)$. However, Johnsen's work does not give lower bounds on R in terms of p_1 with $p_1 < 0.4$. Such lower bounds were subsequently obtained by Montgomery and Abrahams in [8]. The lower bounds on R in [6] and [8] are for binary Huffman codes, and generalizations of these bounds to arbitrary code alphabets have been obtained in [3] and [7]. Furthermore, lower bounds on R in terms of p_n and p_{n-1} have also been obtained in [4] and [12].

In some cases, the probability of a source symbol is given, but the "rank" of this source symbol in the source distribution is not known. In other words, there is no explicit information on how large the probability of this symbol is compared with those of other symbols. For example, if we are given that $p_1 = 0.2$, we not only know that there is a source symbol whose probability of occurrence is 0.2, but we also know that the probability of any other source symbol is at most 0.2. However, if we are given that the probability of a source symbol is 0.2, we can only deduce that the probability of any other source symbol is less than 0.8 (one minus the probability of the given source symbol).

The main result in this correspondence is an upper bound on R in terms of the probability of any source symbol. This result is proved in Section III after the preliminaries are presented in Section II.

II. Preliminaries

Much work [1]–[10], [12] has been devoted to the study of better bounds on the redundancy when some partial knowledge about the source is available. In particular, it is obtained in [7] that $R \leq f(p_1)$, where

$$f(p_1) = \begin{cases} 2 - p_1 - H_b(p_1), & \text{if } 0.5 \leq p_1 < 1 \\ 3 - 5p_1 - H_b(2p_1), & \text{if } 0.4505 \leq p_1 < 0.5 \\ 1 + 0.5(1-p_1) - H_b(p_1), & \text{if } \frac{1}{3} \leq p_1 < 0.4505 \\ 3 - 7.7548p_1 - H_b(3p_1), & \text{if } 0.3138 \leq p_1 < \frac{1}{3} \\ 2 - 1.25(1-p_1) - H_b(p_1), & \text{if } 0.2 \leq p_1 < 0.3138 \\ 4 - 18.6096p_1 - H_b(5p_1), & \text{if } 0.1971 \leq p_1 < 0.2 \\ 2 - 1.3219(1-p_1) - H_b(p_1), & \text{if } \frac{1}{6} \leq p_1 < 0.1971 \\ p_1 + 0.086, & \text{if } p_1 < \frac{1}{6}. \end{cases} \tag{2}$$

This upper bound is tight when $p_1 \geq \frac{1}{6}$. We note that upper bounds tighter than the one given in (2) for $p_1 < \frac{1}{6}$ have been reported in [7], but we do not need to invoke this result in the current work. Fig. 1 shows the upper bound given by (2). We can immediately obtain the following facts from this figure.

Fact 1: The upper bound on R for a source approaches 1 if and only if the probability of the most likely source symbol, p_1, tends to 1.

If p_1 tends to 1, the entropy of the source tends to 0 and the average codeword length of a Huffman code for this source tends to 1. Hence, R tends to 1. On the other hand, if p_1 does not tend to 1, from Fig. 1, the upper bound on R for this source is bounded away from 1.

For a source with alphabet size n, Fact 1 asserts that

$$\{1 - \epsilon_2 - \cdots - \epsilon_n, \epsilon_2, \epsilon_3, \ldots, \epsilon_n\}$$

with $\epsilon_2, \ldots, \epsilon_n$ tending to zero, is the unique form for a sequence of source distributions for which the redundancy of a Huffman code approaches 1.

Manuscript received May 30, 2000; revised May 4, 2001.

C. Ye was with the Department of Information Engineering, The Chinese University of Hong Kong. He is now with the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mail: cxye@Glue.umd.edu).

R. W. Yeung is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong (e-mail: whyeung@ie.cuhk.edu.hk).

Communicated by M. Weinberger, Associate Editor for Source Coding.
Publisher Item Identifier S 0018-9448(02)05164-7.

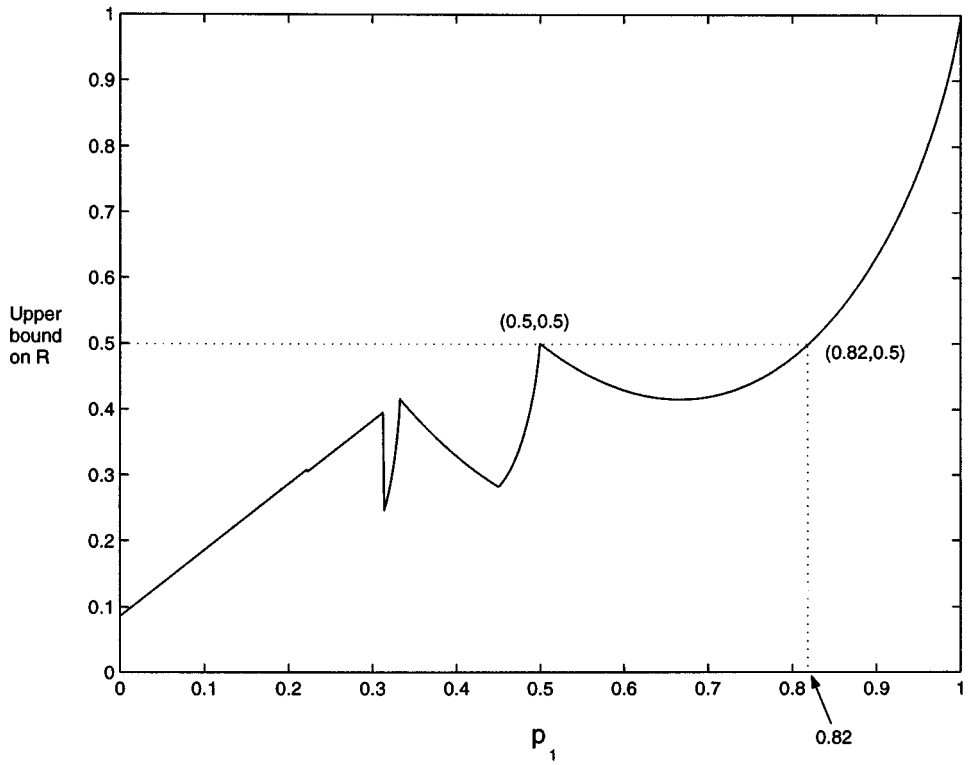

Fig. 1. Upper bound on R when p_1 is known.

Let A be a constant. A simple analysis of Fig. 1 gives the following two facts.

Fact 2: If $p_1 \leq A$ and $A \leq 0.82$, then $R \leq 0.5$, with equality if and only if $p_1 = 0.5$ or 0.82.

Fact 3: If $p_1 \leq A$ and $0.82 < A \leq 1$, then $R \leq 2 - A - H_b(A)$, with equality if and only if $p_1 = A$.

Now we can easily obtain the following theorem.

Theorem 1: Let q be the probability of any source symbol. Then

$$R \leq \begin{cases} 2 - q - H_b(q), & \text{if } 0.5 \leq q < 1 \\ 0.5, & \text{if } 0.18 \leq q < 0.5 \\ 1 + q - H_b(q), & \text{if } q < 0.18. \end{cases} \tag{3}$$

Proof: If $q \geq 0.5$, then q must be equal to p_1. It is obvious from (1) that

$$R \leq 2 - q - H_b(q).$$

If $q < 0.5$, then $q \leq p_1 \leq 1 - q$, which means that p_1 can be any value between q and $1 - q$. We can then obtain an upper bound on R by (2) for each possible p_1. Since no further information about p_1 is available, we take the maximum of all these upper bounds. Hence,

$$R \leq \max_{q \leq p_1 \leq 1-q} f(p_1).$$

Furthermore, if $0.18 \leq q < 0.5$, then $p_1 \leq 1-q \leq 0.82$. By regarding $1 - q$ as the constant A in Fact 2, we see that $R \leq 0.5$. The equality is satisfied if p_1 is equal to 0.5 or 0.82. If $0 \leq q < 0.18$, then $p_1 \leq 1-q$, where $0.82 < 1 - q \leq 1$. Hence, from Fact 3, we have

$$\begin{aligned} R &\leq 2 - (1 - q) - H_b(1 - q) \\ &= 1 + q - H_b(q). \end{aligned}$$

This equality is satisfied if and only if $p_1 = 1 - q$. Therefore, the theorem is proved. □

Fig. 2 shows this upper bound on R as a function of q. It is obviously tight when $q \geq 0.5$. For any given q with $q < 0.18$, consider the source distribution $\{q, 1 - q - \epsilon, \epsilon\}$, where ϵ tends to zero. The redundancy of a Huffman code for this source tends to $1 + q - H_b(q)$. Hence, the upper bound on R in (3) with $q < 0.18$ is also tight. In next section, we will obtain a tighter upper bound on R for $0.18 \leq q < 0.5$.

III. Main Result

Lemma 1: For a source with at least two symbols, let q be the probability of any source symbol. Then

$$R < 1 - H_b(q) - (1-q)\log\left(1 - 2^{-\lceil -\log q \rceil}\right) - q(1 - \lceil -\log q \rceil). \tag{4}$$

Proof: We prove this lemma by showing a particular prefix code C whose redundancy R_c is upper-bounded by the right-hand side (RHS) of (4). This will establish the lemma since the redundancy of a Huffman code must be less than that of any prefix code, in particular that of C.

Let $q = p_j$ for some j, where $1 \leq j \leq n$. Define the length of the code C as

$$l_i = \begin{cases} \lceil -\log p_j \rceil, & \text{if } i = j \\ \left\lceil -\log\left(p_i \dfrac{1 - 2^{-\lceil -\log p_j \rceil}}{1 - p_j}\right) \right\rceil, & \text{if } 1 \leq i \leq n \text{ and } i \neq j. \end{cases} \tag{5}$$

We obtain the above definition of l_i by using the Lagrange multipliers in order to minimize the upper bound on R. The details are given in the Appendix. It is obvious that $l_j \geq 1$. For $1 \leq i \leq n$ and $i \neq j$, since

$$\begin{aligned} p_i \frac{1 - 2^{-\lceil -\log p_j \rceil}}{1 - p_j} &\leq 1 - 2^{-\lceil -\log p_j \rceil} \\ &< 1 \end{aligned}$$

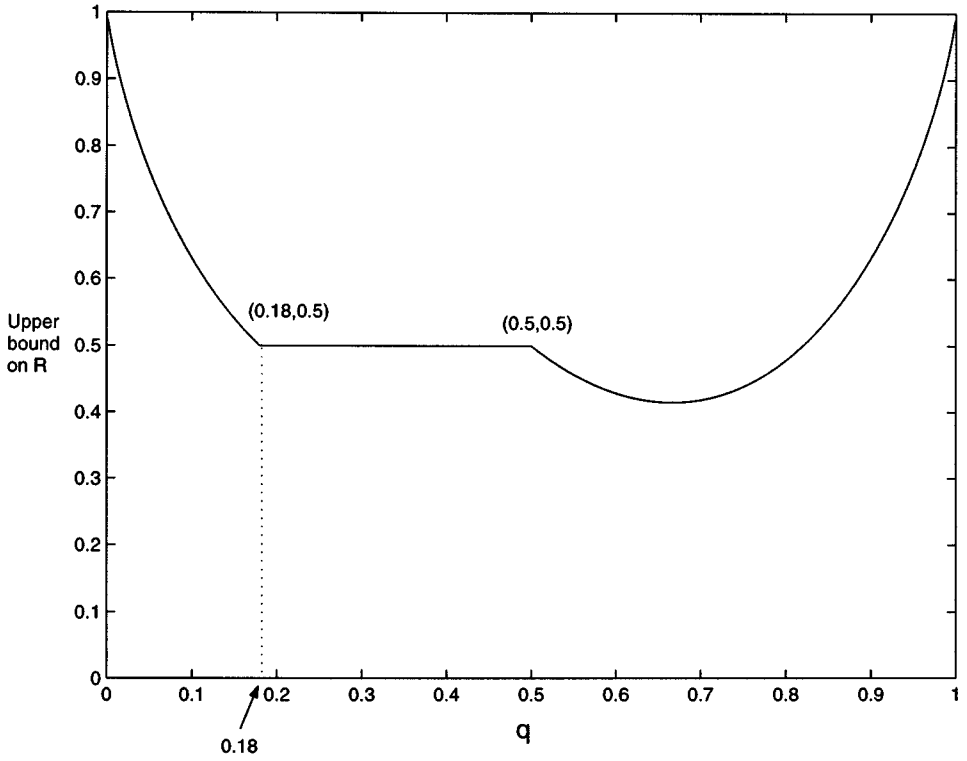

Fig. 2. Upper bound on R when the probability of a symbol is known.

we see that $l_i \geq 1$. Hence, $l_i \geq 1$ for $1 \leq i \leq n$, so they are valid lengths for codewords. From (5), we have

$$\begin{aligned}\sum_{1\leq i\leq n} 2^{-l_i} &= 2^{-l_j} + \sum_{\substack{1\leq i\leq n\\ i\neq j}} 2^{-l_i}\\ &\leq 2^{-\lceil -\log p_j\rceil} + \sum_{\substack{1\leq i\leq n\\ i\neq j}} \left(p_i \frac{1-2^{-\lceil -\log p_j\rceil}}{1-p_j}\right)\\ &= 2^{-\lceil -\log p_j\rceil} + 1 - 2^{-\lceil -\log p_j\rceil}\\ &= 1.\end{aligned}$$

By the Kraft inequality, there exists a prefix code with these codeword lengths. Then the redundancy of this code is upper-bounded as

$$\begin{aligned}R_c &= L - H\\ &= p_j(l_j + \log p_j) + \sum_{\substack{1\leq i\leq n\\ i\neq j}} p_i(l_i + \log p_i)\\ &< p_j(\lceil -\log p_j\rceil + \log p_j)\\ &\quad + \sum_{\substack{1\leq i\leq n\\ i\neq j}} p_i\left[-\log\left(p_i\frac{1-2^{-\lceil -\log p_j\rceil}}{1-p_j}\right) + 1 + \log p_i\right]\\ &= p_j(\lceil -\log p_j\rceil + \log p_j)\\ &\quad + \sum_{\substack{1\leq i\leq n\\ i\neq j}} p_i\left[-\log p_i - \log\left(\frac{1-2^{-\lceil -\log p_j\rceil}}{1-p_j}\right) + 1 + \log p_i\right]\\ &= p_j(\lceil -\log p_j\rceil + \log p_j)\\ &\quad + \sum_{\substack{1\leq i\leq n\\ i\neq j}} p_i\left[-\log\left(\frac{1-2^{-\lceil -\log p_j\rceil}}{1-p_j}\right) + 1\right]\\ &= 1 - (1-p_j)\log\left(\frac{1-2^{-\lceil -\log p_j\rceil}}{1-p_j}\right)\\ &\quad - p_j(1 - \log p_j - \lceil -\log p_j\rceil)\\ &= 1 - H_b(p_j) - (1-p_j)\log\left(1-2^{-\lceil -\log p_j\rceil}\right)\\ &\quad - p_j(1 - \lceil -\log p_j\rceil)\\ &= 1 - H_b(q) - (1-q)\log\left(1-2^{-\lceil -\log q\rceil}\right) - q(1-\lceil -\log q\rceil).\end{aligned}$$

Hence,

$$R < 1 - H_b(q) - (1-q)\log\left(1-2^{-\lceil -\log q\rceil}\right) - q(1-\lceil -\log q\rceil),$$

and the lemma is proved. □

Note that unlike most other upper bounds on R which depends on partial knowledge of the source distribution, this upper bound does not result from exploitation of the structure of Huffman codes. The code defined in (5) can be regarded as a modified Shannon code.

Fig. 3 shows this upper bound on R as a function of q. It is not difficult to prove that the RHS of (4) is not less than the RHS of (3), which implies the upper bound in Lemma 1 is looser than that in Theorem 1. This is not unexpected because the upper bound in (4) depends on less knowledge about the source distribution than the upper bound in (3). Although the upper bound in Lemma 1 is not the tightest possible, it can readily be generalized to the case when the probabilities of any k source symbols, $k \geq 2$, are known. In the next lemma, we illustrate how this can be done for $k = 2$.

Lemma 2: For a source with at least three symbols, let q_1 and q_2 be the probabilities of any two source symbols. Then $R < g(q_1, q_2)$, where

$$\begin{aligned}g(q_1, q_2) &= 1 - (1-q_1-q_2)\log\frac{1-2^{-\lceil -\log q_1\rceil} - 2^{-\lceil -\log q_2\rceil}}{1-q_1-q_2}\\ &\quad - q_1(1 - \log q_1 - \lceil -\log q_1\rceil)\\ &\quad - q_2(1 - \log q_2 - \lceil -\log q_2\rceil).\end{aligned} \tag{6}$$

Proof: Let $q_1 = p_j$ and $q_2 = p_k$ for some j and k, where $1 \leq j, k \leq n$, and $j \neq k$. Define (7) as shown at the bottom of the next page.

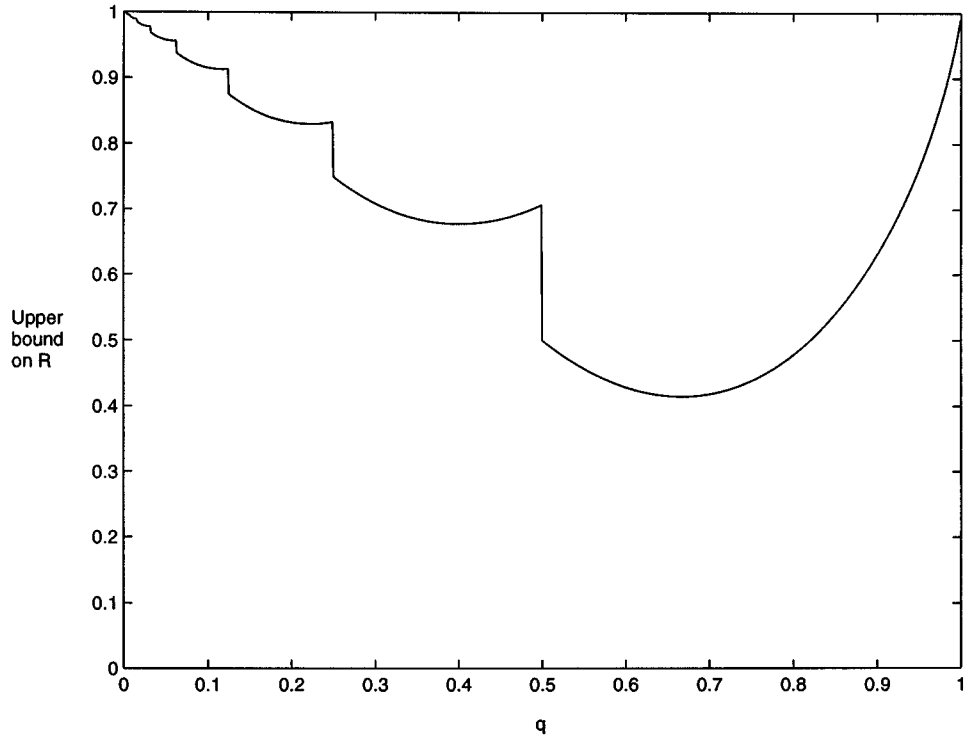

Fig. 3. Upper bound on R when the probability of a symbol is known.

Following arguments, similar to those in Lemma 1, we see that there exists a prefix code with codeword lengths defined in (7). Now the redundancy of this code is upper-bounded as

$$\begin{aligned} L - H &= p_j(l_j + \log p_j) + p_k(l_k + \log p_k) + \sum_{\substack{1 \le i \le n \\ i \ne j,\, k}} p_i(l_i + \log p_i) \\ &< p_j(\lceil -\log p_j \rceil + \log p_j) + p_k(\lceil -\log p_k \rceil + \log p_k) \\ &\quad + \sum_{\substack{1 \le i \le n \\ i \ne j,\, k}} p_i \left(-\log \frac{1 - 2^{-\lceil -\log p_j \rceil} - 2^{-\lceil -\log p_k \rceil}}{1 - p_j - p_k} + 1 \right) \\ &= 1 - (1 - p_j - p_k) \log \frac{1 - 2^{-\lceil -\log p_j \rceil} - 2^{-\lceil -\log p_k \rceil}}{1 - p_j - p_k} \\ &\quad - p_j(1 - \log p_j - \lceil -\log p_j \rceil) - p_k(1 - \log p_k - \lceil -\log p_k \rceil) \\ &= 1 - (1 - q_1 - q_2) \log \frac{1 - 2^{-\lceil -\log q_1 \rceil} - 2^{-\lceil -\log q_2 \rceil}}{1 - q_1 - q_2} \\ &\quad - q_1(1 - \log q_1 - \lceil -\log q_1 \rceil) - q_2(1 - \log q_2 - \lceil -\log q_2 \rceil). \end{aligned}$$

Hence,

$$\begin{aligned} R &< 1 - (1 - q_1 - q_2) \log \frac{1 - 2^{-\lceil -\log q_1 \rceil} - 2^{-\lceil -\log q_2 \rceil}}{1 - q_1 - q_2} \\ &\quad - q_1(1 - \log q_1 - \lceil -\log q_1 \rceil) - q_2(1 - \log q_2 - \lceil -\log q_2 \rceil) \end{aligned}$$

and the lemma is proved. □

In the next lemma, we prove an alternative upper bound on R in terms of q_1 and q_2.

Lemma 3: For a source with at least three symbols, let q_1 and q_2 be the probabilities of any two source symbols such that $q_1 \ge q_2$. Then $R < g'(q_1, q_2)$, where

$$\begin{aligned} g'(q_1, q_2) &= 3 + (1 - q_1 - q_2) \log(1 - q_1 - q_2) \\ &\quad + q_1(\log q_1 - 2) + q_2(\log q_2 - 1). \end{aligned} \tag{8}$$

Proof: Let $q_1 = p_j$ and $q_2 = p_k$ for some j and k, where $1 \le j, k \le n$ and $j \ne k$. Define

$$l_i = \begin{cases} 1, & \text{if } i = j \\ 2, & \text{if } i = k \\ \left\lceil -\log \left(p_i \frac{0.25}{1 - p_j - p_k} \right) \right\rceil, & \text{if } 1 \le i \le n \\ & \text{and } i \ne j,\, k. \end{cases} \tag{9}$$

Following arguments similar to those in Lemma 1, we see that there exists a prefix code with codeword lengths defined in (9). Now the redundancy of this code is upper-bounded as

$$\begin{aligned} L - H &= p_j(l_j + \log p_j) + p_k(l_k + \log p_k) \\ &\quad + \sum_{\substack{1 \le i \le n \\ i \ne j,\, k}} p_i(l_i + \log p_i) \\ &< p_j(1 + \log p_j) + p_k(2 + \log p_k) \\ &\quad + \sum_{\substack{1 \le i \le n \\ i \ne j,\, k}} p_i \left[-\log \left(\frac{0.25}{1 - p_j - p_k} \right) + 1 \right] \\ &= 1 + (1 - q_1 - q_2)[2 + \log(1 - q_1 - q_2)] \end{aligned}$$

$$l_i = \begin{cases} \lceil -\log p_j \rceil, & \text{if } i = j \\ \lceil -\log p_k \rceil, & \text{if } i = k \\ \left\lceil -\log \left(p_i \dfrac{1 - 2^{-\lceil -\log p_j \rceil} - 2^{-\lceil -\log p_k \rceil}}{1 - p_j - p_k} \right) \right\rceil, & \text{if } 1 \le i \le n \text{ and } i \ne j,\, k. \end{cases} \tag{7}$$

$$
\begin{aligned}
&+ q_1 \log q_1 + q_2(1+\log q_2) \\
&= 3 + (1-q_1-q_2)\log(1-q_1-q_2) \\
&\quad + q_1(\log q_1 - 2) + q_2(\log q_2 - 1).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
R < 3 + (1-q_1-q_2)&\log(1-q_1-q_2) \\
&+ q_1(\log q_1 - 2) + q_2(\log q_2 - 1)
\end{aligned}
$$

and the lemma is proved. □

We now compare $g(q_1, q_2)$ in (6) and $g'(q_1, q_2)$ in (8). For some values of q_1 and q_2, $g(q_1, q_2)$ is greater than $g'(q_1, q_2)$, while for other values of q_1 and q_2, $g'(q_1, q_2)$ is greater than $g(q_1, q_2)$. This can be seen from two examples. For $q_1 = 0.5$, $q_2 = 0.2$, $g(q_1, q_2) = 0.339$, and $g'(q_1, q_2) = 0.3145$. Hence, $g(q_1, q_2) > g'(q_1, q_2)$. For $q_1 = 0.3$, $q_2 = 0.2$, $g(q_1, q_2) = 0.5536$ and $g'(q_1, q_2) = 0.7145$. Hence, $g(q_1, q_2) < g'(q_1, q_2)$.

We now prove an enhancement of Theorem 1, which is the main result in this correspondence.

Theorem 2: Let q be the probability of any source symbol. If $0.18 \le q < 0.5$, then

$$
R(q) \le \max_{q<p_1<1-q} \min[f(p_1), g(p_1, q), g'(p_1, q)] \tag{10}
$$

where the functions f, g, and g' are defined in (2), (6), and (8), respectively.

Proof: We distinguish three mutually exclusive cases for which one of them must be true, but we do not know which one it is.

Case 1: If the source has only two symbols, then R is unambiguously determined as $1 - H_b(q)$.

Case 2: If the source has at least three symbols and the given q is the probability of the most likely source symbol, then R is upper-bounded by (2), where $p_1 = q$.

Case 3: If the source has at least three symbols and the given q is not the probability of the most likely source symbol, then we let p_1 be this probability, where $p_1 > q$. Now the redundancy of this code is upper-bounded by (6) and (8), where $q_1 = p_1$ and $q_2 = q$, and it is also upper-bounded by (2). Thus, it is upper-bounded by the minimum of these three upper bounds. However, p_1 can take any value between q and $1 - q$. Since we have no further information about p_1, we take the maximum of these upper bounds over all $q < p_1 < 1 - q$.

Finally, since we do not know which of the three cases is true, we take the maximum of the upper bounds given in all three cases. Hence,

$$
\begin{aligned}
R(q) \le \max\Big\{&1 - H_b(q),\ f(q), \\
&\max_{q<p_1<1-q} \min[f(p_1), g(p_1, q), g'(p_1, q)]\Big\}.
\end{aligned} \tag{11}
$$

The remaining task is to simplify the RHS of (11), i.e., we will show that $1 - H_b(q)$ and $f(q)$ do not contribute to the final upper bound on $R(q)$.

For any given q between 0.18 and 0.5, we take $p_1 = 1-q-\epsilon$, where ϵ tends to zero. Note that p_1 is within the open set $(q, 1-q)$. Hence,

$$
\begin{aligned}
\max_{q<p_1<1-q} &\min[f(p_1), g(p_1, q), g'(p_1, q)] \\
&\ge \min[f(1-q-\epsilon), g(1-q-\epsilon, q), g'(1-q-\epsilon, q)].
\end{aligned}
$$

In the following, we will calculate $f(1-q-\epsilon)$, $g(1-q-\epsilon, q)$, and $g'(1-q-\epsilon, q)$, respectively. First

$$
\begin{aligned}
f(1-q-\epsilon) &= 2 - (1-q-\epsilon) - H_b(1-q-\epsilon) \\
&\approx 1 + q - H_b(q).
\end{aligned} \tag{12}
$$

Since $0.18 \le q < 0.5$, $p_1 = 1-q-\epsilon > 0.5$ and $\lceil -\log p_1 \rceil = 1$. From (6), we have

$$
\begin{aligned}
g(1-q-\epsilon, q) &= 1 - \epsilon \log\left(\frac{\frac{1}{2} - 2^{-\lceil -\log q \rceil}}{\epsilon}\right) \\
&\quad + (1-q-\epsilon)\log(1-q-\epsilon) \\
&\quad - q(1 - \log q - \lceil -\log q \rceil) \\
&= 1 - \epsilon \log\left(\tfrac{1}{2} - 2^{-\lceil -\log q \rceil}\right) - q(1 - \lceil -\log q \rceil) \\
&\quad - H(q, 1-q-\epsilon, \epsilon)
\end{aligned}
$$

where $H(q, 1-q-\epsilon, \epsilon)$ is the entropy of a source with distribution $\{q, 1-q-\epsilon, \epsilon\}$. It is obvious that $H(q, 1-q-\epsilon, \epsilon) \approx H_b(q)$. Also considering

$$
\lceil -\log q \rceil = \begin{cases} 2, & \text{if } 0.25 \le q < 0.5 \\ 3, & \text{if } 0.18 \le q < 0.25 \end{cases}
$$

we have

$$
\begin{aligned}
&g(1-q-\epsilon, q) \\
&\quad \approx 1 - \epsilon \log\left(\frac{1}{2} - 2^{-\lceil -\log q \rceil}\right) - q(1 - \lceil -\log q \rceil) - H_b(q) \\
&\quad = \begin{cases} 1 + 2\epsilon + q - H_b(q), & \text{if } 0.25 \le q < 0.5 \\ 1 + \epsilon \log \frac{8}{3} + 2q - H_b(q), & \text{if } 0.18 \le q < 0.25 \end{cases} \\
&\quad \approx \begin{cases} 1 + q - H_b(q), & \text{if } 0.25 \le q < 0.5 \\ 1 + 2q - H_b(q), & \text{if } 0.18 \le q < 0.25. \end{cases}
\end{aligned} \tag{13}
$$

Similarly, from (8), we have

$$
\begin{aligned}
g'(1-q-\epsilon, q) &= 3 + \epsilon \log \epsilon + (1-q-\epsilon)\log(1-q-\epsilon) \\
&\quad - 2(1-q-\epsilon) + q \log q - q \\
&= 1 + q + 2\epsilon + \epsilon \log \epsilon + q \log q \\
&\quad + (1-q-\epsilon)\log(1-q-\epsilon) \\
&= 1 + q + 2\epsilon - H(q, 1-q-\epsilon, \epsilon) \\
&\approx 1 + q - H_b(q).
\end{aligned} \tag{14}
$$

Finally, from (12)–(14), we obtain

$$
\max_{q<p_1<1-q} \min[f(p_1), g(p_1, q), g'(p_1, q)] \ge 1 + q - H_b(q).
$$

Now $1 + q - H_b(q)$ is always greater than $1 - H_b(q)$ for positive value q. Also for $0.18 \le q < 0.5$, it is not difficult to prove that $1+q-H_b(q)$ is greater than $f(q)$, where the function f is given by (2).

Therefore, the terms $1 - H_b(q)$ and $f(q)$ on the RHS of (11) do not contribute to the final upper bound on $R(q)$, and thus the theorem is proved. □

Using Theorem 2, we can improve the upper bound in Theorem 1 for $0.18 \le q < 0.5$. The resulting upper bound is shown in Fig. 4. From this figure, we can obtain the following fact which is readily seen to be equivalent to Fact 1.

Fact 4*:* The upper bound on R for a source approaches 1 if and only if the probability of every source symbol tends to 1 or 0.

From Fig. 4, it looks as though the upper bound on R in terms of q is symmetric about $q = 0.5$. Hence, we conjecture that the upper bound on R in terms of q with $0.18 \le q < 0.5$, is also $1 + q - H_b(q)$. This conjecture can be verified if we can prove that

$$
\max_{q<p_1<1-q} \min[f(p_1), g(p_1, q), g'(p_1, q)] \le 1 + q - H_b(q).
$$

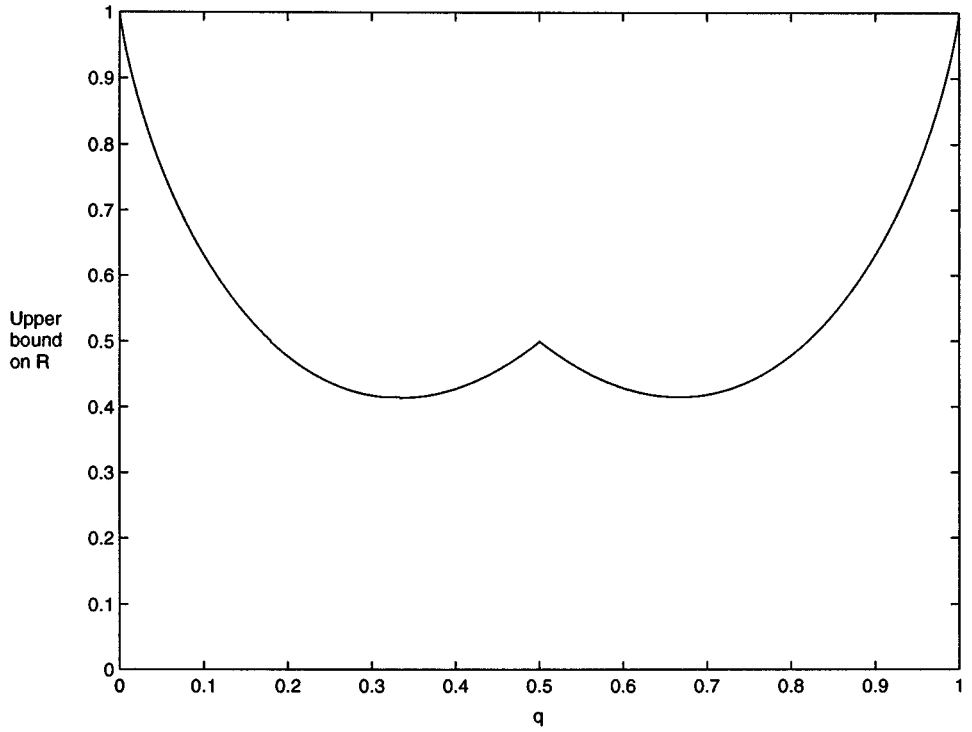

Fig. 4. Enhanced upper bound on R when the probability of a symbol is known.

IV. Conclusion

In this correspondence, we have obtained an upper bound on R in terms of the probability of any given source symbol. An upper bound on R in terms of the probabilities of any two given source symbols is also obtained. To obtain a lower bound on R in terms of the probability of any given symbol would be an interesting problem for further research.

We mention that the technique used in the proof of Lemma 1 has been modified in [11] to obtain upper bounds on the redundancy of an optimal fix-free code.

Appendix
Optimization of an Upper Bound on R for a Given p_j

In this appendix, we show how to optimize an upper bound on R for a given p_j. First, we let $l_j = \lceil -\log p_j \rceil$. Now suppose

$$l_i = \lceil -\log(p_i + x_i) \rceil, \qquad \text{for } i \neq j$$

where $0 \leq x_i < 1 - p_i$. Here $x_i < 1 - p_i$ guarantees that $l_i \geq 1$, so that it is a valid length for a codeword. Then

$$\begin{aligned} L - H &= p_j(l_j + \log p_j) + \sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i(l_i + \log p_i) \\ &< p_j(\lceil -\log p_j \rceil + \log p_j) \\ &\quad + \sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i \left[-\log \left(p_i \cdot \left(1 + \frac{x_i}{p_i} \right) \right) + 1 + \log p_i \right] \\ &= p_j(\lceil -\log p_j \rceil + \log p_j) \\ &\quad + \sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i \left[-\log \left(1 + \frac{x_i}{p_i} \right) + 1 \right] \\ &= 1 - p_j(1 - \lceil -\log p_j \rceil - \log p_j) \\ &\quad - \sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i \log \left(1 + \frac{x_i}{p_i} \right). \end{aligned} \tag{15}$$

Since p_j is fixed, to minimize the upper bound on the redundancy of this code is to maximize $\sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i \log(1 + \frac{x_i}{p_i})$. On the other hand

$$\begin{aligned} \sum_{1 \leq i \leq n} 2^{-l_i} &= 2^{-l_j} + \sum_{\substack{1 \leq i \leq n \\ i \neq j}} 2^{-l_i} \\ &\leq 2^{-\lceil -\log p_j \rceil} + \sum_{\substack{1 \leq i \leq n \\ i \neq j}} (p_i + x_i) \\ &= 2^{-\lceil -\log p_j \rceil} + 1 - p_j + \sum_{\substack{1 \leq i \leq n \\ i \neq j}} x_i. \end{aligned} \tag{16}$$

By the Kraft inequality, the RHS of (16) must be less than 1 in order to guarantee the existence of a prefix code. Hence,

$$\sum_{\substack{1 \leq i \leq n \\ i \neq j}} x_i \leq p_j - 2^{-\lceil -\log p_j \rceil}. \tag{17}$$

Let us, for the time being, ignore the constraints $0 \leq x_i < 1 - p_i$ for $i \neq j$. To maximize $\sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i \log(1 + \frac{x_i}{p_i})$ under the constraint in (17), we define

$$J = \sum_{\substack{1 \leq i \leq n \\ i \neq j}} p_i \log \left(1 + \frac{x_i}{p_i} \right) + \lambda \left(\sum_{\substack{1 \leq i \leq n \\ i \neq j}} x_i \right).$$

From $\frac{\partial J}{\partial x_i} = 0$, we have

$$x_i = -p_i \cdot \left(1 + \frac{1}{\lambda'} \right) \tag{18}$$

where $\lambda' = \lambda \cdot (\ln 2)$. Putting (18) into (17), we obtain

$$-\left(1 + \frac{1}{\lambda'} \right) \leq \frac{p_j - 2^{-\lceil -\log p_j \rceil}}{1 - p_j}.$$

Then from (18), for $1 \leq i \leq n$, $i \neq j$

$$x_i \leq p_i \frac{p_j - 2^{-\lceil -\log p_j \rceil}}{1 - p_j}.$$

Since the upper bound in (15) is a decreasing function of x_i for each i, in order to minimize this upper bound, we take the upper bound

$$x_i = p_i \frac{p_j - 2^{-\lceil -\log p_j \rceil}}{1 - p_j}. \tag{19}$$

With this choice of x_i, since $p_j \geq 2^{-\lceil -\log p_j \rceil}$, $x_i \geq 0$. On the other hand,

$$p_i + x_i = \frac{p_i}{1 - p_j} \left(1 - 2^{-\lceil -\log p_j \rceil}\right) < 1.$$

Hence we conclude that $0 \leq x_i < 1 - p_i$, as required.

Finally, for $1 \leq i \leq n$ and $i \neq j$, $x_i \geq 0$ implies

$$\begin{aligned} l_i &= \lceil -\log(p_i + x_i) \rceil \\ &\leq \lceil -\log p_i \rceil. \end{aligned}$$

This explains why the code defined by $\{l_i,\ 1 \leq i \leq n\}$ as above gives a tighter upper bound on R than the Shannon code.

REFERENCES

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

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

[3] ——, "A note on D-ary Huffman codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 174–179, Jan. 1991.

[4] ——, "New bounds on the redundancy of Huffman codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1095–1104, July 1991.

[5] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 668–674, Nov. 1978.

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

[7] D. Manstetten, "Tight bounds on the redundancy of Huffman codes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 144–151, Jan. 1992.

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

[9] 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, pp. 293–296, Mar. 1987.

[10] R. De Prisco and A. De Santis, "On the redundancy achieved by Huffman codes," *Inform. Sci.*, vol. 88, pp. 131–148, Jan. 1996.

[11] C. Ye and R. W. Yeung, "Some basic properties of fix-free codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 72–87, Jan. 2001.

[12] R. W. Yeung, "Local redundancy and progressive bounds on the redundancy of a Huffman code," *IEEE Trans. Inform. Theory*, vol. 37, pp. 687–691, May 1991.

An Axiomatization of Partition Entropy

Dan A. Simovici, *Member, IEEE*, and Szymon Jaroszewicz

***Abstract*—The aim of this correspondence is to present an axiomatization of a generalization of Shannon's entropy starting from partitions of finite sets. The proposed axiomatization defines a family of entropies depending on a real positive parameter that contains as a special case the Havrda–Charvat entropy, and thus, provides axiomatizations for the Shannon entropy, the Gini index, and for other types of entropy used in classification and data mining.**

***Index Terms*—Gini index, Havrda–Charvat entropy, non-Shannon entropy, Shannon entropy.**

I. INTRODUCTION AND BASIC NOTATIONS

The notion of partition of a finite set is naturally linked to the notion of probability distribution. Namely, if A is a finite set and $\pi = \{B_1, \ldots, B_n\}$ is a partition of A, then the probability distribution attached to π is $(p_1, \ldots, p_n)$, where $p_i = \frac{|B_i|}{|A|}$ for $1 \leq i \leq n$. Thus, it is natural to consider the notion of entropy of a partition via the entropy of the corresponding probability distribution. Axiomatizations for entropy and entropy-like characteristics of probability distributions represent a problem with a rich history in information theory. Previous relevant work includes the results of Khinchin [12], Faddeev [6], Ingarden and Urbanik [9], Rényi [15] who investigated various axiomatizations of entropy, and Daróczy who presented in [5] an unified treatment of entropy-like characteristics of probability distributions using the notion of information function.

In our previous work (see [16], [10]), we introduced an axiomatization for the notion of functional entropy. This numerical characteristic of functions is related to the complexity of circuits that realize functions (cf. [1]) and serves as an estimate for power dissipation of a circuit realizing a function (cf. [8]) and is linked to the notion of entropy for partitions, since every function $f\colon A \longrightarrow B$ between the finite sets A, B defines a partition on its definition domain A whose blocks are $\{f^{-1}(b) \mid b \in \mathrm{Ran}(f)\}$, where $\mathrm{Ran}(f)$ is the range of the function f.

Information measures, especially conditional entropy of a logic function and its variables, have been used for minimization of logic functions (see [13] and [2]).

In a different direction, starting from the notion of impurity of a set relative to a partition, we found a common generalization of Shannon entropy and of Gini index and we used this generalization in clustering of noncategorial data (see [17]). Devijer used the Gini index in pattern recognition in [4].

Our main result is an axiomatization of this generalization that illuminates the common nature of several known ways of evaluating concentrations of values of functions.

The set of reals, the set of nonnegative reals, the set of rational numbers, the set of natural numbers, and the set of positive natural numbers are denoted by $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, $\mathbb{Q}$, $\mathbb{N}$, $\mathbb{N}_1$, respectively. All other sets considered in the following discussion are nonempty and finite.

Manuscript received June 27, 2001; revised January 18. 2002. A preliminary form of this paper was published in the *Proceedings of the 31 International Symposium on Multiple-Valued Logic*, Warsaw, Poland, 2001, pp. 259–268.

The authors are with the Department of Computer Science, University of Massachusetts, Boston, MA 02125 USA (e-mail: dsim@cs.umb.edu; sj@cs.umb.edu).

Communicated by G. Battail, Associate Editor At Large.

Publisher Item Identifier S 0018-9448(02)05145-3.