International Journal of Mathematics and
Computer Science, **10**(2015), no. 2, 175–184

Two Conjectures About Recency Rank Encoding

Chris Buhse[1], Peter Johnson[1], William Linz[2], Matthew Simpson[3]

[1]Department of Mathematics and Statistics
Auburn University
Auburn, AL 36849, USA

[2]Department of Mathematics
Texas A&M University
College Station, TX 77843-3368, USA

[3]Department of Mathematics
North Carolina State University
Raleigh, North Carolina 27695, USA

email: johnspd@auburn.edu, willdomath@tamu.edu, mcsimps2@ncsu.edu

(Received July 22, 2015, Accepted August 5, 2015)

Abstract

In recency rank encoding, a source letter s is replaced by a code word u_j, where j is the number of different source letters that have appeared since the last occurrence of s in the source text. Let m be the number of source letters and suppose that the source is perfectly zeroth order, meaning that the source letters enter the text independently with fixed probabilities $f_1, \ldots, f_m$. For $k \in \{0, \ldots, m-1\}$, let $g_k = g_k(f_1, \ldots, f_m)$ be the probability that, if a source letter is chosen independently from the source text, exactly k different letters other than that letter have appeared since the last appearance of that letter. The conjectures of the title are:

(1) $g_0 \geq \ldots \geq g_{m-1}$ with equality at any point ($g_j = g_{j+1}$) if and only if $f_1 = \ldots = f_m = \frac{1}{m}$;

Key words and phrases: Huffman's algorithm, replacement coding, source coding, source frequencies, zeroth order source.
AMS (MOS) Subject Classifications: 68P30, 94A29, 26D15, 60E15, 60C05, 05A20, 05E05
ISSN 1814-0432, 2015, http://ijmcs.future-in-tech.net
This research was partially supported by NSF (DMS) grant number 1262930

(2) if the code words $u_0, \dots, u_{m-1}$ for recency rank encoding are obtained by applying Huffman's algorithm to the probabilities $g_0, \dots, g_{m-1}$, and the code words $w_1, \dots, w_m$ are obtained by applying Huffman's algorithm to the probabilities $f_1, \dots, f_m$ then $\sum_{j=1}^{m} f_j \cdot \mathrm{lgth}(w_j) \leq \sum_{j=0}^{m-1} g_j \cdot \mathrm{lgth}(u_j)$.

1 Recency rank vs. simple replacement encoding

Let $S = \{s_1, \dots, s_m\}$ and $A = \{a_1, \dots, a_n\}$, $m, n \geq 2$, be, respectively, a source alphabet and a code alphabet. In the modern problem of encoding, we wish to represent words "over" S by words over A in such a way that every source word is recoverable from its code representative. Subordinate to this requirement of unique decodability are various optimization goals: efficiency of encoding, efficiency of decoding, compression, error detection and correction. See [2] and [3]. The goal of compression is to minimize the average number of code letters per source letter in the encoding; we might call this average the compression index.

In simple replacement encoding, each source letter s_j is assigned a code word w_j, and the encoding proceeds by replacing each occurrence of s_j in the source text by w_j. Thus, the code representative of source text $s_{i_1} s_{i_2} \dots s_{i_N}$ will be the concatenation $w_{i_1} w_{i_2} \dots w_{i_N}$. For unique and efficient decodability, the list $w_1, \dots, w_m$ of code words is prefix-free (see [2]). For compression, common sense whispers that the w_j should be as short as possible, given the requirement that the list of w_j be prefix-free. If nothing is known about the source text, we may as well make the w_j as nearly equal in length and as short as possible, so that $\mathrm{lgth}(w_j) \in \{\lfloor \log_n m \rfloor, \lceil \log_n m \rceil\}$ for each $j = 1, \dots, m$. If the relative frequencies in the source text of $s_1, \dots, s_m$ are known – let's call them $f_1, \dots, f_m$ (f_j is the probability that a letter plucked at random from the source text will be s_j) – then one can apply Huffman's algorithm [2] to the relative frequencies to obtain a prefix-free list $w_1, \dots, w_m$ for simple replacement encoding which minimizes the compression index $\sum_{j=1}^{m} f_j \cdot \mathrm{lgth}(w_j)$.

In recency rank encoding, introduced by Elias [1] in 1987 (although he credits others for independently having the idea), we use a prefix-free list $u_0, \ldots, u_{m-1}$ of code words; an occurrence of $s \in S$ is replaced by u_k when exactly k elements of $S \setminus \{s\}$ have appeared in the source text since the last occurrence of s. (In theory, the source text has no beginning. In practice, every actual chunk of source text has a beginning, so there will have to be some convention for getting started, in actual application of recency rank encoding.) Decoding is unique and not terribly hard, although it is a little more time-consuming than simply recognizing code words, as in simple replacement.

Recency rank encoding was touted as an effective "on-line" method, requiring no knowledge of the statistics of the source text, but only the number of source letters. However, it does require user agreement on the code words $u_0, \ldots, u_{m-1}$ and, if nothing is known about the source text, it seems sensible to make the u_j all of length around $\log_n m$. What, then, is the advantage of recency rank encoding over simple replacement of the source letters s_j with code words w_j, of the lengths also around $\log_n m$, especially in view of the fact that decoding of simple replacement code is noticeably easier than recency rank decoding?

Perhaps because of such considerations, recency rank coding has not survived as a practical method in digital communication. However, as a source of interesting mathematical questions, recency rank coding is a rich, hitherto untapped (so far as we know) resource. Our aim here is to pose a couple of mathematical questions about recency rank coding, in the form of conjectures, and to confirm part of one of these conjectures.

Let the source be perfectly zeroth order; this means that letters enter the source text independently, as though they were drawn with replacement from an urn, in which $s_1, \ldots, s_m$ occur in proportions $f_1, \ldots, f_m$, respectively. (But note that the f_j are not required to be rational.) For text from such a source, if a block of k consecutive letters is chosen from the source text at random, then for any $i_1, \ldots, i_k \in \{1, \ldots, m\}$, the probability that the block chosen will be $s_{i_1} \ldots s_{i_k}$ is the product $\prod_{j=1}^{k} f_{i_j}$.

Text from such a source is the same, statistically, whether read forward or backwards. Therefore, we can define the probabilities of interest in analyzing recency rank coding as follows. For $k \in \{0, 1, \ldots, m-1\}$, let $g_k = g_k(f_1, \ldots, f_m)$ be the probability that, if a letter s is selected at random from the source text, exactly k different letters of $S \setminus \{s\}$ will appear in the source text (reading forward) before the next occurrence of s. Thus,

$$g_0 = \sum_{i=1}^{m} f_i^2$$

$$g_1 = \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{t=1}^{\infty} f_i f_j^t f_i = \sum_{i=1}^{m} f_i^2 \sum_{\substack{j=1 \\ j \neq i}}^{m} \frac{f_j}{1 - f_j}$$

and the expressions for $g_2, \ldots, g_{m-1}$ as formulas in $f_1, \ldots, f_m$ are a bit more complicated. For instance:

$$\begin{aligned} g_2 &= \sum_{i=1}^{m} f_i \left[\sum_{\substack{1 \leq j_1 < j_2 \leq m \\ j_1 \neq i \neq j_2}} \left[\sum_{k=2}^{\infty} (f_{j_1} + f_{j_2})^k - \sum_{k=2}^{\infty} f_{j_1}^k - \sum_{k=2}^{\infty} f_{j_2}^k \right] \right] f_i \\ &= \sum_{i=1}^{m} \sum_{\substack{1 \leq j_1 < j_2 \leq m \\ j_1 \neq i \neq j_2}} f_i^2 \left[\frac{(f_{j_1} + f_{j_2})^2}{1 - f_{j_1} - f_{j_2}} - \frac{f_{j_1}^2}{1 - f_{j_1}} - \frac{f_{j_2}^2}{1 - f_{j_2}} \right] \end{aligned}$$

In general, for $2 \leq k \leq m-1$, the formula for g_k is

$$g_k = \sum_{i=1}^{m} f_i^2 \left[\sum_{\substack{Q \subseteq \{1,\ldots,m\} \setminus \{i\} \\ |Q| = k}} \left[\frac{\left(\sum_{j \in Q} f_j \right)^k}{1 - \sum_{j \in Q} f_j} + \sum_{t=1}^{k-1} (-1)^{k-t} \sum_{\substack{S \subseteq Q \\ |S| = t}} \frac{\left(\sum_{j \in S} f_j \right)^k}{1 - \sum_{j \in S} f_j} \right] \right]$$

Conjecture 1.1. If $m \geq 2$, $f_1, \ldots, f_m > 0$, and $\sum_{i=1}^{m} f_i = 1$, then for all $k \in \{1, \ldots, m-1\}$, $g_{k-1}(f_1, \ldots, f_m) \geq g_k(f_1, \ldots, f_m)$, with equality if and only if $f_1 = \ldots = f_m = \frac{1}{m}$

Conjecture 1.2. Suppose $f_1, \ldots, f_m$ are as in Conjecture 1 and that $g_k = g_k(f_1, \ldots, f_m)$, $k = 0, \ldots, m-1$. Suppose that the application of Huffman's algorithm to $f_1, \ldots, f_m$ results in code words $w_1, \ldots, w_m$, and an application of Huffman's algorithm to $g_0, \ldots, g_{m-1}$ results in code words $u_0, \ldots, u_{m-1}$. Then $\sum_{i=1}^{m} f_i \cdot \text{lgth}(w_i) \leq \sum_{k=0}^{m-1} g_k \cdot \text{lgth}(u_k)$

In Conjecture 1.2, "lgth" stands for length, meaning the number of code letters in the word. The conjecture is that for a zeroth order source, the best possible compression index achievable by recency rank encoding, by a shrewd choice of code words, is no better (smaller) than the best possible compression index achievable by simple replacement.

We do not have a good guess about conditions for equality in Conjecture 1.2, although, if Conjecture 1.1 holds, then $f_1 = \ldots = f_m = \frac{1}{m}$ implies $g_0 = \ldots = g_{m-1} = \frac{1}{m}$ (because the g_i must sum to 1), which implies equality in Conjecture 1.2.

It is straightforward to see that $g_0(\frac{1}{m}, \ldots, \frac{1}{m}) = \frac{1}{m}$. Therefore, if the inequalities $g_0 \geq \ldots \geq g_{m-1}$ hold for all $f_1, \ldots, f_m$, then $f_1 = \ldots = f_m = \frac{1}{m}$ implies that $g_0 = \ldots = g_{m-1} = \frac{1}{m}$.

We do not have strong reasons for these conjectures. They have withstood testing for small values of m. In section 3, we will prove that $g_0 \geq g_1$ with equality if and only if $f_1 = \ldots = f_m = \frac{1}{m}$. In section 2, we will develop the analysis to be used in the proof in section 3; some may find this analysis of interest in itself.

2 Useful inequalities

Lemma 2.1. *If $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, $\alpha \geq \beta$ and $\gamma \geq \delta$, then $\alpha\gamma + \beta\delta \geq \alpha\delta + \beta\gamma$ with equality if and only if either $\alpha = \beta$ or $\gamma = \delta$.*

Proof. The proposed inequality is equivalent to $(\alpha-\beta)(\gamma-\delta)\geq 0$, which obviously follows from the hypotheses. The sufficiency and necessity of the conditions for equality also follow from this equivalence. $\square$

Theorem 2.2. *Suppose that $m\geq 2$, $a_1\geq\ldots\geq a_m>0$, and $b_1\geq\ldots\geq b_m>0$. Then the function h defined by $h(x)=\left(\sum_{i=1}^m a_i^x\right)^{-1}\left(\sum_{j=1}^m a_j^x b_j\right)$ is strictly increasing on $[0,\infty)$, unless either $a_1=\ldots=a_m$ or $b_1=\ldots=b_m$. Clearly, if either $a_1=\ldots=a_m$ or $b_1=\ldots=b_m$, then h is constant.*

Proof. Suppose that $y>x\geq 0$. We aim to show that $h(y)\geq h(x)$ and that equality implies that either $a_1=\ldots=a_m$ or $b_1=\ldots=b_m$.

Each of the following after the first is clearly equivalent to the inequality preceding:

$$(1)\qquad h(y)\geq h(x);$$

$$(2)\quad \left(\sum_{i=1}^m a_i^x\right)\left(\sum_{j=1}^m a_j^y b_j\right)\geq\left(\sum_{i=1}^m a_i^y\right)\left(\sum_{j=1}^m a_j^x b_j\right);$$

$$(3)\quad \sum_{i=1}^m a_i^{x+y}b_i+\sum_{\substack{1\leq i,j\leq m\\ i\neq j}} a_i^x a_j^y b_j\geq\sum_{i=1}^m a_i^{x+y}b_i+\sum_{\substack{1\leq i,j\leq m\\ i\neq j}} a_i^y a_j^x b_j;$$

$$(4)\quad \sum_{1\leq i<j\leq m}\left(a_i^x a_j^y b_j+a_j^x a_i^y b_i\right)\geq\sum_{1\leq i<j\leq m}\left(a_i^y a_j^x b_j+a_j^y a_i^x b_i\right);$$

$$(5)\quad \sum_{1\leq i<j\leq m}\left(a_i a_j\right)^x\left[a_j^{y-x}b_j+a_i^{y-x}b_i\right]\geq\sum_{1\leq i<j\leq m}\left(a_i a_j\right)^x\left[a_i^{y-x}b_j+a_j^{y-x}b_i\right].$$

Now, $y-x>0$ and $a_i\geq a_j$ for $1\leq i<j\leq m$ implies that $a_i^{y-x}\geq a_j^{y-x}$. Therefore, by Lemma 2.1, with $\alpha=a_i^{y-x}$, $\beta=a_j^{y-x}$, $\gamma=b_i$, $\delta=b_j$, for each pair i,j such that $1\leq i<j\leq m$, $a_i^{y-x}b_i+a_j^{y-x}b_j\geq a_i^{y-x}b_j+a_j^{y-x}b_i$,

with equality if and only if either $a_i^{y-x} = a_j^{y-x}$ ($\iff$ $a_i = a_j$) or $b_i = b_j$. Therefore, inequality (5) holds because each term on the left is greater than or equal to the corresponding term (indexed by (i, j)) on the right. Therefore, equality holds if and only if it holds for each term; consequently, if equality holds, then for each pair (i, j), $1 \le i < j \le m$, either $a_i = a_j$ or $b_i = b_j$. This implies that, if equality holds in (1), and thus in (5), then either $a_1 = \ldots = a_m$ or $b_1 = \ldots = b_m$. For (assuming equality holds) if the a_i are not all equal, then $a_1 > a_m$, which implies $b_1 = b_m$, whence $b_1 = \ldots = b_m$. $\square$

Although we will make no use of it here, it would be churlish of us not to point out the following corollary.

Corollary 2.3. *Suppose that $k, m \ge 2$ and that $A = [a_{ij}]$ is a $k \times m$ matrix of positive real numbers such that each row is non-constant and non-increasing. Then the function $H : \left\{(x_1, \ldots, x_k) \in \mathbb{R}^k | x_i > 0, i = 1, \ldots, k\right\} \to (0, \infty)$ defined by*

$$H(x_1, \ldots, x_k) = \left[\prod_{i=1}^{k} \left(\sum_{j=1}^{m} a_{ij}^{x_i}\right)\right]^{-1} \left[\sum_{j=1}^{m} \left(\prod_{i=1}^{k} a_{ij}^{x_i}\right)\right]$$

is strictly increasing in each variable x_i, $i = 1, \ldots, n$.

Proof. If $k - 1$ of the variables – without loss of generality, say $x_2, \ldots, x_k$ – are fixed, then the resulting function of the remaining variable, x_1, is

$$p(x_1) = c \left(\sum_{j=1}^{m} a_{1j}^{x_1}\right)^{-1} \left(\sum_{j=1}^{m} a_{1j}^{x_1} b_j\right),$$

where $c = \prod_{i=2}^{k} \left(\sum_{j=1}^{m} a_{ij}^{x_i}\right)^{-1}$ and $b_j = \prod_{i=2}^{k} a_{ij}^{x_i}$, $j = 1, \ldots, m$. Since $a_{11} \ge \ldots \ge a_{1m}$ and $b_1 \ge \ldots \ge b_m$, and neither finite sequence is constant by assumptions about A, we have $p(x_1) = c \cdot h(x_1)$, with h being of the form given in Theorem 2.2, strictly increasing, and $c > 0$. $\square$

Both Theorem 2.2 and Corollary 2.3 have generalizations to arbitrary finite, positive measure spaces. We will give, without proof, the generalization of Theorem 2.2.

Theorem 2.4. *Suppose that (M, μ) is a positive measure space, $0 < \mu(M) < \infty$ and $a, b : M \to (0, \infty)$ are measurable functions such that for all $s, t \in M$, $a(s) \geq a(t)$ if and only if $b(s) \geq b(t)$. Then $h : [0, \infty) \to (0, \infty)$, defined by*

$$h(x) = \left(\int_M a(t)^x d\mu(t) \right)^{-1} \left(\int_M a(t)^x b(t) d\mu(t) \right)$$

is strictly increasing on $[0, \infty)$ unless one of a, b is essentially constant.

Theorem 2.2 is the special case of Theorem 2.4 in which $M = \{1, \dots, m\}$ and μ is the counting measure.

3 $g_0 \geq g_1$, and a necessary and sufficient condition for equality

Theorem 3.1. *Suppose that $m \geq 2$, $f_1, \dots, f_m > 0$, and $\sum_{j=1}^{m} f_j = 1$. Then $g_0(f_1, \dots, f_m) \geq g_1(f_1, \dots, f_m)$, with equality if and only if $f_1 = \ldots = f_m = \frac{1}{m}$.*

Proof. We may as well suppose that $1 > f_1 \geq \ldots \geq f_m > 0$, which implies that $\frac{1}{1-f_1} \geq \ldots \geq \frac{1}{1-f_m} > 0$. From Section 1, $g_0(f_1, \dots, f_m) = \sum_{i=1}^{m} f_i^2$ and

$$
\begin{aligned}
g_1(f_1, \ldots, f_m) &= \sum_{i=1}^{m} f_i^2 \left(\sum_{\substack{j=1 \\ j \neq i}}^{m} \frac{f_j}{1-f_j} \right) \\
&= \sum_{i=1}^{m} f_i^2 \left[\left(\sum_{j=1}^{m} \frac{f_j}{1-f_j} \right) - \frac{f_i}{1-f_i} \right] \\
&= \left(\sum_{i=1}^{m} f_i^2 \right) \sum_{j=1}^{m} \frac{f_j}{1-f_j} - \sum_{i=1}^{m} f_i^2 \frac{f_i}{1-f_i}.
\end{aligned}
$$

Therefore, the inequality $g_0 \geq g_1$ is equivalent to

$$
\sum_{i=1}^{m} f_i^2 \frac{1}{1-f_i} \geq \left(\sum_{i=1}^{m} f_i^2 \right) \sum_{j=1}^{m} f_j \frac{1}{1-f_j}
$$

or

$$
\left(\sum_{i=1}^{m} f_i^2 \right)^{-1} \sum_{j=1}^{m} f_j^2 \frac{1}{1-f_j} \geq \sum_{j=1}^{m} f_j \frac{1}{1-f_j} = \left(\sum_{i=1}^{m} f_i \right)^{-1} \sum_{j=1}^{m} f_j \frac{1}{1-f_j}, \text{ since } \sum_{i=1}^{m} f_i = 1.
$$

Setting $f_i = a_i$ and $\frac{1}{1-f_i} = b_i$ and referring to Theorem 2.2, the left-hand side of the inequality above is $h(2)$, and the right-hand side is $h(1)$. By Theorem 2.2, therefore, we have $g_0 \geq g_1$, with equality only if (and also if) either $a_1 = \ldots = a_m$ or $b_1 = \ldots = b_m$; each is equivalent to $f_1 = \ldots = f_m = \frac{1}{m}$. □

References

[1] Peter Elias, Interval and recency rank source coding, two on-line adaptive variable-length schemes, *IEEE Transactions in Information Theory,* **33**, no. 1, 1987, 3-10.

[2] Darrel Hankerson, Greg Harris, Peter Johnson, *Introduction to Information Theory and Data Compression,* 2nd edition, Chapman & Hall/CRC, Boca Raton, 2003.

[3] D. C. Hankerson, D. G. Hoffman, D. A. Leonard, Charles C. Lindner, K.T. Phelps, C.A. Rodger, J. R. Wall, *Coding Theory and Cryptography: The Essentials,* 2nd edition, Taylor & Francis, 2000.