International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 453–460

Monotonicity of Code Symbol Frequencies in Recency Rank Encoding

Matthew Ellison[1], **Joseph Johnson**[2]

[1]Department of Mathematics
Dartmouth College
Hanover, New Hampshire 03755, USA

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

email: matthew.ellison.gr@dartmouth.edu, jwjohns5@ncsu.edu

(Received September 3, 2021, Accepted September 28, 2021)

Abstract

In *recency rank encoding*, the encoder reads along the source text, and, as each symbol s is read, the code symbol c_j is added to the code text, where j is the number of distinct symbols which have appeared since the previous occurrence of s. We consider an idealized source text model which is two-way infinite ($\mathbb{Z}$-indexed), and where at each position, independently, one of m source symbols $s_1, ..., s_m$ is chosen at random according to fixed global frequencies $f_1, ..., f_m > 0$. Applying recency rank encoding to this source text yields a two-way infinite code text on code symbols $c_0, ..., c_{m-1}$, and we let $g_i = g_i(f_1, ..., f_m)$ denote the probability that the symbol at a given source index is encoded by c_i. We prove that $g_0 \geq \ldots \geq g_{m-1}$, with equality at any point if and only if $f_1 = \ldots = f_m = \frac{1}{m}$, confirming a conjecture of Buhse et al. [1].

1 Introduction

Consider a source text written with symbols $s_1, \ldots, s_m$. When applying recency rank encoding, each source symbol s is encoded by code symbol c_j,

Key words and phrases: Source coding, recency rank, inequalities.
AMS (MOS) Subject Classifications: 68P30, 05A20, 94A29, 26D15, 60E15, 60C05.
This work was supported in part by NSF grant 1560257.
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

where j is the number of distinct source symbols appearing since the previous occurrence of s. For example, in the section of source text $\ldots s_3 s_1 s_4 s_4 s_2 s_1 \ldots$, the final s_1 would be encoded by c_2, since two distinct source symbols (s_2 and s_4) have appeared since the previous occurrence of s_1. Together with a way to get the encoding 'off the ground' for the first appearance of each source symbol[1], recency rank encoding becomes a well-defined encoding method with straightforward decoding when the code symbols are prefix-free words over some code alphabet.

Recency rank encoding was introduced by Elias [2], though an equivalent scheme—now known as the BSTW Algorithm—was introduced by Bentley et al. [3]. One benefit of recency rank encoding is that it is naturally implemented as an online algorithm, requiring only a single pass over the data (Huffman coding requires two passes). Another benefit is that if occurrences of source symbols tend to be clumped together in the source text, recency rank encoding will take advantage of this if c_i has few bits for small i.

Though it must be admitted that recency rank encoding has not found wide use, it certainly leads to interesting mathematical questions. In this paper we investigate the action of recency rank encoding on an idealized source text which is two-way infinite ($\mathbb{Z}$-indexed), and where at each position, independently, one of m source symbols $s_1, ..., s_m$ is chosen at random according to fixed global frequencies $f_1, ..., f_m > 0$. Applying recency rank encoding to this source text yields a two-way infinite code text on code symbols $c_0, ..., c_{m-1}$, and we let $g_i = g_i(f_1, ..., f_m)$ denote the probability that the symbol at a given source index is encoded by c_i [2]. Note that since the statistics of the source text are the same under translation, each g_i is independent of the source index chosen.

A variety of questions can be asked about the code frequencies $g_0, ..., g_{m-1}$. How are the g_i related to the f_i and each other? Could we expect to improve on a compression scheme by applying it not to the original source text but to its recency rank encoding?

In Buhse et al. [1], the authors make two conjectures in this vein:

1. $g_0 \geq ... \geq g_{m-1}$, with equality at any point if and only if $f_1 = ... = f_m = 1/m$.

2. Applying Huffman coding (see, for example, [4]) to the recency rank

[1] For example by preceding the source message with the string $s_1 s_2 \ldots s_m$

[2] The zero-probability event that there is no prior occurrence of a symbol is not relevant for our results. If the reader wished, they could encode this situation by any of the c_i (or some new symbol) without altering the g_i of interest.

encoding yields compression no better than applying Huffman coding directly to source text. More precisely, if we let $\{u_i\}_{i=1}^{m}, \{v_i\}_{i=0}^{m-1}$ be the respective Huffman code words for frequencies $f_1, ..., f_m$ and $g_0, ..., g_{m-1}$ then

$$\sum_{i=1}^{m} f_i \cdot \mathrm{lgth}(u_i) \leq \sum_{i=0}^{m-1} g_i \cdot \mathrm{lgth}(v_i).$$

Buhse et al. proved part of the first conjecture, establishing that $g_0 \geq g_1$ with equality exactly when $f_1 = ... = f_m = \frac{1}{m}$. In this paper we establish the full first conjecture.

2 Preliminaries

It will be convenient, in introducing notation, to fix an index in the ($\mathbb{Z}$-indexed) source text for consideration. We choose index 0 for this purpose and for the remainder of the paper take g_i to be the probability that the source symbol at index 0 is encoded as c_i. As noted in the introduction, the translational invariance of the source text assures that each g_i is independent of this choice.

With this convention in place, we now define the event, denoted by $[s_j W^*(S')s_j]$, which forms the base of our analysis.

Definition 2.1. *Let S' be a proper subset of source symbol set $S = \{s_1, ..., s_m\}$, and let s_j be a source symbol in $S - S'$. We let $[s_j W^*(S')s_j]$ denote the event on the source text where the following conditions are all met:*

1. *The source symbol at index 0 is s_j.*

2. *The s_j at index 0 is immediately preceded by a segment of arbitrary length using only the symbols of S', which must contain all the symbols in S' (repetitions are allowed).*

3. *The segment of the above condition is immediately preceded by the symbol s_j.*

The notation $[s_j W^*(S')s_j]$ is meant to describe these conditions pictorially: the rightmost s_j representing the s_j at index 0, the $W^*(S')$ representing the preceding 'word' on all the symbols of S', and the leftmost s_j representing the s_j immediately preceding that word. Note that this definition allows the possibility that S' is empty, in which case the event translates to the symbols at index 0 and -1 both being s_j.

The key point is that if $|S'| = k$, and the event $[s_j W^*(S') s_j]$ occurs, the source symbol at index 0 will be encoded by c_k. A bit of thought shows that whenever the symbol at index 0 is encoded as c_k, it is due to exactly one event of this type occurring, yielding the following useful expression for g_i.

Lemma 2.2. *For $i \in \{0, ..., m-1\}$,*

$$g_i = \sum_{s_j \in S} \sum_{S' \in \binom{S - \{s_j\}}{i}} P[s_j W^*(S') s_j]$$

□

Here, we have used the notation $\binom{S-\{s_j\}}{i}$ to denote the set of all subsets of $S - \{s_j\}$ of size i, and have written $P[s_j W^*(S') s_j]$ to denote the probability of the event $[s_j W^*(S') s_j]$.

We end this preliminary section with a simple piece of notation which we will use frequently.

Definition 2.3. *Let S' be a subset of $S = \{s_1, ..., s_m\}$. Then we define*

$$P(S') = \sum_{s_i \in S'} f_i$$

In other words, we let $P(S')$ denote the sum of the frequencies of the source symbols in S'.

3 Main results

Recall that our aim is to establish, under the described source text model, that $g_0 \geq g_1 \geq ... \geq g_{m-1}$. The additional result that equality holds at any point exactly when all f_i are equal will follow easily from the proof.

Our approach is to show that $g_{i-1} - g_i \geq 0$ for all $i \in \{1, ..., m-1\}$. Lemmas 3.2 and 3.3 yield convenient expressions for g_i and g_{i-1}. Lemma 3.1 establishes a recursive identity, key for the proof of Lemma 3.2.

Lemma 3.1. *For S' a non-empty proper subset of $S = \{s_1, ..., s_m\}$, $s_j \in S - S'$,*

$$P[s_j W^*(S') s_j] = \sum_{s_k \in S'} \frac{f_k P[s_j W^*(S' - \{s_k\}) s_j]}{1 - P(S' - \{s_k\}) - f_k}.$$

Proof. The idea is to break into cases based on the source symbol at index -1.

$$\begin{aligned} P[s_jW^*(S')s_j] = & \sum_{s_k \in S'} P[s_jW^*(S')s_ks_j] + P[s_jW^*(S' - \{s_k\})s_ks_j] \\ =& P[s_jW^*(S')s_j] \sum_{s_k \in S'} f_k + \sum_{s_k \in S'} f_k P[s_jW^*(S' - \{s_k\})s_j] \end{aligned}$$

Combining terms and dividing gives

$$P[s_jW^*(S')s_j] = \sum_{s_k \in S'} \frac{f_k P[s_jW^*(S' - \{s_k\})s_j]}{1 - P(S')}$$

Substituting $P(S') = P(S' - \{s_k\}) + f_k$ then gives the result. □

Lemma 3.2. *For $i \in \{1, ..., m-1\}$,*

$$g_i = \sum_{\{s_j, s_k\} \in \binom{S}{2}} \sum_{S' \in \binom{S - \{s_j, s_k\}}{i-1}} \frac{f_k P[s_jW^*(S^{'})s_j]}{1 - P(S') - f_k} + \frac{f_j P[s_kW^*(S^{'})s_k]}{1 - P(S') - f_j}.$$

Proof. As noted in Section 2,

$$g_i = \sum_{s_j \in S} \sum_{S' \in \binom{S - \{s_j\}}{i}} P[s_jW^*(S^{'})s_j]$$

By the previous lemma, we then have

$$g_i = \sum_{s_j \in S} \sum_{S' \in \binom{S - \{s_j\}}{i}} \sum_{s_k \in S'} \frac{f_k P[s_jW^*(S' - \{s_k\})s_j]}{1 - P(S' - \{s_k\}) - f_k}$$

Now we reindex. With a little thought, the reader will see that the sets of ordered triples

$$\{(s_j, s_k, S');\ s_k \in S' \in \binom{S - \{s_j\}}{i}\} \text{ and}$$

$$\{(s_j, s_k, S');\ s_j \neq s_k,\ S' \in \binom{S - \{s_j, s_k\}}{i-1}\}$$

are in bijection. Thus we have

$$g_i = \sum_{\substack{(s_j,s_k)\\ s_j\neq s_k}} \sum_{S'\in\binom{S-\{s_j,s_k\}}{i-1}} \frac{f_k P[s_j W^*(S')s_j]}{1-P(S')-f_k}$$

Pairing indices of the form $((s_j, s_k), S')$ and $((s_k, s_j), S')$ then gives the result. □

Lemma 3.3. *For $i \in \{1, ..., m-1\}$,*

$$g_{i-1} = \sum_{\{s_j,s_k\}\in\binom{S}{2}} \sum_{S'\in\binom{S-\{s_j,s_k\}}{i-1}} \frac{f_k P[s_j W^*(S^{'})s_j]}{1-P(S')-f_j} + \frac{f_j P[s_k W^*(S^{'})s_k]}{1-P(S')-f_k}.$$

Proof. As in the previous Lemma, we have

$$g_{i-1} = \sum_{s_j\in S} \sum_{S'\in\binom{S-\{s_j\}}{i-1}} P[s_j W^*(S^{'})s_j]$$

To achieve a form compatible with the previous result, we introduce a variable s_k to our indexing by noting that for any $s_j \in S - S'$

$$1 = \sum_{s_k\in S-S'-\{s_j\}} \frac{f_k}{P(S-S'-\{s_j\})} = \sum_{s_k\in S-S'-\{s_j\}} \frac{f_k}{1-P(S')-f_j}$$

Thus we have

$$g_{i-1} = \sum_{s_j\in S} \sum_{S'\in\binom{S-\{s_j\}}{i-1}} \sum_{s_k\in S-S'-\{s_j\}} \frac{f_k P[s_j W^*(S^{'})s_j]}{1-P(S')-f_j}$$

Performing a reindexing similar to the previous Lemma and pairing terms yields the result. □

Theorem 3.4. *For positive source frequencies $f_1, ...f_m$, we have*

$$g_0 \geq g_1 \geq ... \geq g_{m-1}.$$

At each point, equality is attained if and only if $f_1 = ... = f_m = \frac{1}{m}$ (in which case we have equality at every point).

Proof. We show that we have $g_{i-1} - g_i \geq 0$ for $i \in \{1, ..., m-1\}$. Before applying the previous two lemmas we note that

$$\frac{P[s_j W^*(S')s_j]}{f_j^2} = \frac{P[s_k W^*(S')s_k]}{f_k^2}.$$

This identity can be seen by comparing corresponding terms in an expansion of each side over all values of $W^*(S^{'})$. Letting λ equal either side of this identity, we apply the preceding two lemmas to yield

$$\begin{aligned} g_{i-i} - g_i &= \sum_{\{s_j,s_k\}\in\binom{S}{2}} \sum_{S'\in\binom{S-\{s_j,s_k\}}{i-1}} \frac{f_j f_k^2\lambda - f_k f_j^2\lambda}{1-P(S')-f_k} + \frac{f_k f_j^2\lambda - f_j f_k^2\lambda}{1-P(S')-f_j} \\ &= \sum_{\{s_j,s_k\}\in\binom{S}{2}} \sum_{S'\in\binom{S-\{s_j,s_k\}}{i-1}} f_j f_k \lambda (f_k - f_j) \left(\frac{1}{1-P(S')-f_k} - \frac{1}{1-P(S')-f_j} \right) \\ &= \sum_{\{s_j,s_k\}\in\binom{S}{2}} \sum_{S'\in\binom{S-\{s_j,s_k\}}{i-1}} f_k f_j \lambda (f_k - f_j)^2 \left(\frac{1}{(1-P(S')-f_k)(1-P(S')-f_j)} \right) \end{aligned}$$

Each term of this sum is non-negative, and, since we have assumed each $f_i > 0$, any term in which $f_k \neq f_j$ is strictly positive. Thus we have that $g_{i-1} - g_i \geq 0$, with equality if and only if $f_1 = ... = f_m$, as desired. □

4 Further Questions

One interesting area of future research would be to make progress on the second conjecture of Buhse et al. (stated in the introduction), on the interplay between Huffman coding and the recency rank transform. Another interesting line of inquiry would be to investigate correlations between nearby code symbols. For example, one could consider two source locations k indices apart and look into the mutual information between the corresponding code symbols. Computer simulation with frequencies $f_1 = f_2 = f_3 = f_4 = \frac{1}{4}$ yield that the mutual information between adjacent code symbols (k=1) is $\approx .0001$. With $f_1 = \frac{7}{10}, f_2 = f_3 = f_4 = \frac{1}{10}$, on the other hand, the mutual information between adjacent code symbols was $\approx .08$. Both simulations were carried out using the Python Random library and data was collected with 100,000 randomly generated source texts. Perhaps the mutual information between two code symbols is zero exactly when all f_i are equal.

Another potentially interesting idea is to consider repeated application of recency rank encoding, treating the produced code text as a source text and so on. It seems likely that on repeated application the code frequencies would tend toward the uniform distribution, and perhaps bounds could be obtained on the rate of convergence.

A final line of inquiry might be to investigate the extent to which the $g_0 \geq g_1 \geq \ldots \geq g_{m-1}$ result holds under different source text models. It certainly doesn't hold generally (consider the sequence $...s_1s_2s_3s_1s_2s_3...$), but perhaps it holds for some broader class of source text models than considered here.

5 Acknowledgments

We would like to give special thanks to Peter Johnson, Alumni Professor at Auburn University. He introduced us to this problem at the 2017 Auburn University REU in Algebra and Discrete Mathematics, and, but for his saint-like patience, words of encouragement, and generous help editing, it seems unlikely this work would have seen the light of day.

References

[1] C. Buhse, P. Johnson, W. Linz, M. Simpson, Two conjectures about recency rank encoding, *International Journal of Mathematics and Computer Science*, **10**, no. 2, (2015), 175–184.

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

[3] J. Bentley, D. Sleator, R. Tarjan, V. Wei, A locally adaptive compression scheme, *Programming Techniques and Data Structures*, **29**, no. 4, (1986), 320–330.

[4] D. Hankerson, G. Harris, P. Johnson, *Introduction to Information Theory and Data Compression, 2nd Edition.* Chapman & Hall CRC, 2003.