A Family of Comma-Free Codes with Even Word-Length

L. J. Cummings, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
ljcummin@math.uwaterloo.ca

Abstract

For even codeword length $n = 2k, k > 1$ and alphabet size $\sigma > 1$ a family of comma-free codes is constructed with $\lfloor\frac{\sigma^2}{3}\rfloor^r(\sigma^2 - \lfloor\frac{\sigma^2}{3}\rfloor)^{k-r}$ codewords where $1 \leq r < k$. In particular, a new maximal comma-free code with $n = 4$ and $\sigma = 4$ is given by one of these codes.

In a noiseless channel a message stream generated by using a comma-free code avoids misframing errors since comma-free codes do not contain overlaps of codewords by definition. Comma-free codes can still prevent misframing errors even in a noisy channel if bit error correction is also used. Accuracy is then a function of the error correction and the channel used. Pioneering papers collected in the anthology [7] contain early studies of simultaneous correction of both types of errors. Comma-free codes were first discussed in a biology paper [3] and the first mathematical treatment was [6].

A block code C with codewords of fixed length $n > 1$ over a finite alphabet of σ letters is *comma-free* if for all codewords $\mathbf{x} = x_1 \dots x_n, \mathbf{y} = y_1 \dots y_n \in C$ *none* of the *overlaps*

$$x_{t+1} \cdots x_n y_1 \cdots y_t \quad t = 1, \dots, n-1 \tag{1}$$

are in C.

We will refer to (1) as the $t^{\underline{th}}$ overlap of $\mathbf{x}$ and $\mathbf{y}$.

The Witt bound on the maximum number of codewords in any comma-free code over an alphabet of size σ with codewords of length n is

$$\frac{1}{n}\sum_{d|n} \mu(n/d)\sigma^d, \tag{2}$$

where μ is the Möbius function of elementary number theory [6]. The function (2) is well known in other contexts [4].

Golomb, Gordon and Welch [6] conjectured that the upper bound (2) was tight for all odd n. Seven years later, Eastman [5] found a construction which resolved their conjecture affirmatively. An easily implemented algorithm was subsequently given by Scholtz [10] for the case when n is odd, but for even n the bound is not attained in general. For $n = 2$ Golomb, Gordon, and Welch [6] showed that $\lfloor \frac{\sigma^2}{3} \rfloor$ is a tight upper bound. All isomorphism classes for word length 2 were determined in [2]. For $n = 3$ all comma-free codes over finite alphabets were determined in [1].

Block comma-free codes over a binary alphabet attaining the Witt bound for $n = 2, 4, 6, 8$ have been know for some time [6, 9]. In 1973 Niho used a backtracking program to find a maximal binary comma-free code when $n = 10$ with 99 codewords which meets the Witt bound [8].

It was first shown by Golomb, Gordon, and Welch [6] that the Witt bound (2) is not attained for all even word lengths $n = 2k$ when $\sigma > 3^k$. Subsequently Jiggs [9] refined this by showing that the Witt bound is not attained if $\sigma > 2^k + k$. Currently the best result known is that the Witt bound is not attained for $\sigma > k^{\frac{\log k}{0.71}} + k$ and $n > 8$ [11].

Earlier Jiggs [9] had shown that the Witt bound is not attained for $n = \sigma = 4$ with an exhaustive backtracking program written by Lee Laxdal. He found a maximal comma-free code with 57 codewords but the Witt bound is 60 for these parameters.

Theorem 1 *Let Σ be a finite alphabet with σ letters. There exists a comma-free code $C_r = C_r(n = 2k, \sigma)$ over Σ with $\lfloor \frac{\sigma^2}{3} \rfloor^r (\sigma^2 - \lfloor \frac{\sigma^2}{3} \rfloor)^{k-r}$ codewords where $k > 1$ and $1 \leq r < k$.*

Proof: Let D denote a maximal comma-free code contained in the set of pairs Σ^2. It is well known that such a code has $\lfloor \frac{\sigma^2}{3} \rfloor$ codewords [6].

Let C_r denote the set of codewords:

$$a_1 \cdots a_r c_1 \cdots c_s b_1 \cdots b_r d_1 \cdots d_s \tag{3}$$

where $a_i b_i \in D, c_j d_j \in \Sigma^2 \backslash D, i = 1, \ldots r; j = 1, \ldots s$ and $r + s = k$. It is crucial to note that each pair of entries a_i and b_i are separated by exactly $k - 1$ entries in each codeword as are the pairs c_i and d_i.

To prove C_r is comma-free we argue by assuming that if a variable codeword $\mathbf{w}$ of the form (3) is the $t^{\underline{th}}$ overlap of codewords

$$\begin{aligned} \mathbf{w}' &= a'_1 \cdots a'_r c'_1 \cdots c'_s b'_1 \cdots b'_r d'_1 \cdots d'_s \\ \mathbf{w}'' &= a''_1 \cdots a''_r c''_1 \cdots c''_s b''_1 \cdots b''_r d''_1 \cdots d''_s. \end{aligned}$$

in C_r then $\mathbf{w}$ is not in C_r.

Case 1 If $1 \leq t < r$ the $t^{\underline{th}}$ overlap of the concatenation $\mathbf{w}'\mathbf{w}''$ is

$$a'_{t+1} \cdots a'_r c'_1 \cdots c'_s b'_1 \cdots b'_r d'_1 \cdots d'_s a''_1 \cdots a''_t \tag{4}$$

where $r + s = k$.

If (4) is also in C_r then it has the form (3). If $t \leq s$ then the length of the prefix $a'_{t+1} \cdots a'_r c'_1 \cdots c'_s$ is greater than the length of $a_1 \cdots a_r$ and $a_r = c'_t$ since the $t^{\underline{th}}$ overlap is being considered. Since a_r and b_r are separated by exactly $k-1$ entries as are c'_t and d'_t we conclude $b_r = d'_t$ and $a_r b_r = c'_t d'_t \in \Sigma^2 \backslash D$, a contradiction unless (4) is not in C_r. On the other hand, if $t > s$ then $a_r = b'_{t-s}$ and $b_r = a''_{t-s}$ as is seen by counting $k-1$ entries in (4). In this case $a_r b_r = b'_{t-s} a''_{t-s}$ which is an overlap of $a'_{t-s} b'_{t-s}$ and $a''_{t-s} b''_{t-s} \in D$. Since D is comma-free, $a_r b_r \notin D$, contrary to the definition of C_r. Therefore, (4) is not in C_r in this subcase as well.

<u>Case 2</u> If $r \leq t < k$ then the $t^{\underline{th}}$ overlap of $\mathbf{w}'\mathbf{w}''$ is

$$c'_{t-r+1} \cdots c'_s b'_1 \cdots b'_r d'_1 \cdots d'_s a''_1 \cdots a''_r c''_1 \cdots c''_{t-r}. \tag{5}$$

If (5) is in C_r then it has the form (3) and $a_1 = c'_{t-r+1}$. Arguing as before, $b_1 = d'_{t-r+1}$ and $a_1 b_1 = c'_{t-r+1} d'_{t-r+1} \in \Sigma^2 \backslash D$, contrary to the definition of C_r unless (5) is not in C_r.

<u>Case 3</u> If $k \leq t < k + r$ then the $t^{\underline{th}}$ overlap of $\mathbf{w}'\mathbf{w}''$ is

$$b'_{t-k+1} \cdots b'_r d'_1 \cdots d'_s a''_1 \cdots a''_r c''_1 \cdots c''_s b''_1 \cdots b''_{t-k}. \tag{6}$$

Here, $a_1 = b'_{t-k+1}$ and $b_1 = a''_{t-k+1}$ showing that $a_1 b_1 = b'_{t-k+1} a''_{t-k+1}$ is an overlap of $a'_{t-k+1} b'_{t-k+1}$ and $a''_{t-k+1} b''_{t-k+1}$ in D, a contradiction unless (6) is not in C_r.

<u>Case 4</u>

If $k + r \leq t < n$ then the $t^{\underline{th}}$ overlap of $\mathbf{w}'\mathbf{w}''$ is

$$d'_{t-k-r+1} \cdots d'_s a''_1 \cdots a''_r c''_1 \cdots c''_s b''_1 \cdots b''_r d''_1 \cdots d''_{t-k-r}. \tag{7}$$

The prefix $a_1 \cdots a_r c_1 \cdots c_s$ of $\mathbf{w}$ necessarily has length greater than the prefix $d'_{t-k-r+1} \cdots d'_s a''_1 \cdots a''_r$ of (7). Since $k + r \leq t$, $a''_r = c_{t-k-r+1}$ and, arguing as in previous cases, $b''_r = d_{t-k-r+1}$ yielding the contradiction $a''_r b''_r = c_{t-k-r+1} d_{t-k-r+1} \in \Sigma^2 \backslash D$, unless (7) $\notin C_r$.

This completes the proof.

The construction (3) for $r = 1$ first appeared in [6] in a proof of a theorem giving bounds on the asymptotic density of the number of words in a maximal comma-free code codewords of even length.

Consider the binary alphabet $\Sigma = \{0, 1\}$. If $n = 4$ then $k = 2$ and only $r = 1$ is possible in the construction (3). The only maximal comma-free code in Σ^2 is $D = \{01\}$ (or $\{10\}$) and $\Sigma^2 \backslash D = \{00, 10, 11\}$ ($\{00, 01, 11\}$). The resulting comma-free code is $\{0010, 0110, 0111\}$ if 01 is chosen and this meets the Witt bound of 3.

If $\Sigma = \{0,1\}$ and $n = 6$ then $k = 3$ then both $r = 1$ and $r = 2$ are possible in the construction (3). If $r = 1$ then again there is only $D = \{01\}$ (or $\{10\}$) and $\Sigma^2 \backslash D = \{00, 10, 11\}$ ($\{00, 01, 11\}$), but since the word length is 6 the resulting code has 9 codewords again meeting the Witt bound. If $r = 2$ then the only possibility is to repeat the the single pair 01 (or 10) in (3) so that there are still only 9 codewords provided by the construction.

More generally, if $n > 6$ and $\sigma = 2$ then $\lfloor \frac{\sigma^2}{3} \rfloor^r (\sigma^2 - \lfloor \frac{\sigma^2}{3} \rfloor)^{k-r} = 3^{\frac{n}{2}-r}$. Thus, the greatest number of codewords a binary code C_r can have occurs when $r = 1$.

Corollary 1 *If $\sigma = 2$ and $n = 2k, 1 < k$, then any comma-free code C_r can have at most $3^{\frac{n}{2}-1}$ codewords of length n.*

If $n = \sigma = 4$ then C_r has $5^r 11^{2-r}$ codewords by Theorem 1. The only feasible value for r is 1 and $C_1(4,4)$ has 55 codewords, two short of the 57 codewords in the code found by Jiggs in a computer search [9]. Nevertheless it is interesting because it is maximal.

Let $CF(n, \sigma)$ denote the class of comma-free codes with words of length n over an alphabet Σ with σ letters.

Theorem 2 *There exists a maximal $CF(4,4)$ code with 55 codewords.*

Proof: We list the 55 codewords of a $C_1(4,4)$ code obtained by the construction given by Theorem 1. We begin by choosing the maximal $CF(2,4)$ code given by $\{01, 02, 21, 31, 32\}$.

0010	0020	2010	3010	3020
0013	0023	2013	3013	3023
0110	0120	2110	3110	3120
0111	0121	2111	3111	3121
0112	0122	2112	3112	3122
0113	0123	2113	3113	3123
0210	0220	2210	3210	3220
0212	0222	2212	3212	3222
0213	0223	2213	3213	3223
0310	0320	2310	3310	3320
0313	0323	2313	3313	3323

There are 256 possible codewords. Of these 16 are periodic and so cannot appear in any comma-free code. Each codeword of $C_1(4,4)$ generates an equivalence class [w] of 4 codewords under cyclic permutation. These classes are necessarily disjoint since a comma-free code cannot contain two words from the same class. Removing periodic codewords and those in the equivalence classes of C leaves 20 codewords which form 5 classes under cyclic permutation:

$$[0003],\ [0033],\ [0131],\ [0232],\ [0333].$$

Routine checking shows that each codeword in each class creates an overlap with the codewords in C. Therefore, C is a maximal comma-free code.

References

[1] A.H. Ball, *The construction of comma-free codes with odd word length*, Ph.D thesis, Department of Combinatorics and Optimization, University of Waterloo, Canada, 1980.

[2] A.H. Ball and L.J. Cummings, *The comma-free codes with words of length two*, Bull. Austral. Math. Soc. **14**(1976), 249–258.

[3] F.H.C. Crick, J.S. Griffith, and L.E. Orgel, *Codes without commas*, Proceedings of the National Academy of Sciences, Washington, **43**(1957), 416–21.

[4] L.J. Cummings and M.E. Mays, On the parity of the Witt formula, *Congressus Numerantium* **80**(1991), 49–56.

[5] W.L. Eastman, *On the construction of comma-free codes*, IEEE Trans. Information Theory **11**(1965), 263–267.

[6] S.W. Golomb, B. Gordon, and L.R. Welch, *Comma-free codes*, Canadian J. Math. **10**(1958), 202–209.

[7] W.E. Hartnett, editor, Foundations of Coding Theory, D. Reidel, Boston, 1974.

[8] Y. Niho, *On maximal comma-free codes*, IEEE Trans. Inform. Theory (1973), 580–581.

[9] B.H. Jiggs, *Recent results in comma-free codes*, Canadian J. Math. **15**(1963), 178–187.

[10] R.A. Scholtz, *Maximal and variable word- length comma-free codes*, IEEE Trans. Inform. Theory **IT15**(1969), 300–306.

[11] B. Tang, S.W. Golomb, and R.L. Graham, *A New Result on Comma-Free Codes of Even Word Length*, Canadian J. Math. **39**(1987), 513–526.