Non-Overlapping Codes

Simon R. Blackburn
Department of Mathematics
Royal Holloway, University of London
Egham, Surrey TW20 0EX
United Kingdom

February 6, 2014

Abstract

We say that a $q$-ary length $n$ code is non-overlapping if the set of non-trivial prefixes of codewords and the set of non-trivial suffixes of codewords are disjoint. (This property is often called cross-bifix-free.)

Bajić and Stojanović were the first to consider non-overlapping codes, motivated by applications requiring fast and reliable frame synchronisation. Chee, Kiah, Purkayastha and Wang showed that a $q$-ary length $n$ non-overlapping code contains at most $q^n/(2^n-1)$ codewords, and gave a construction of a family of non-overlapping codes that are within a constant factor of this upper bound when $q$ is fixed.

We provide a simple combinatorial proof of the upper bound of Chee et al. (in fact, the upper bound is improved slightly). The construction of Chee et al. is good for small alphabet sizes, but performs much less well when the alphabet size $q$ is large compared to the length $n$ of the code. We provide a construction that generalises the construction of Chee et al. which performs well for all parameter sizes. More precisely, we show that our construction provides non-overlapping codes whose cardinality is within an (absolute) constant factor of the upper bound, for all parameters $n$ and $q$.

We also consider codes of short length, showing that the upper bound of $q^n/(2^n-1)$ is not always asymptotically tight. We provide a conjecture on the leading term of the cardinality of an optimal non-overlapping code when $n$ is fixed and $q \to \infty$. 
1 Introduction

Let \( u \) and \( v \) be two words (not necessarily distinct) of length \( n \), over a finite alphabet \( F \) of cardinality \( q \). We say that \( u \) and \( v \) are overlapping if a non-empty proper prefix of \( u \) is equal to a non-empty proper suffix of \( v \), or if a non-empty proper prefix of \( v \) is equal to a non-empty proper suffix of \( u \).

So, for example, the binary words 00000 and 01111 are overlapping; so are the words 10001 and 11110. However, the words 11111 and 01110 are non-overlapping.

We say that a code \( C \subseteq F^n \) is non-overlapping if for all (not necessarily distinct) \( u, v \in C \), the words \( u \) and \( v \) are non-overlapping. The following is an example of a non-overlapping binary code of length 6 containing 3 codewords:

\[
C = \{001101, 001011, 001111\}.
\]

We write \( C(n,q) \) for the maximum number of codewords in a \( q \)-ary non-overlapping code of length \( n \). It is easy to see that \( C(1,q) = q \). From now on, to avoid trivialities, we always assume that \( n \geq 2 \).

Inspired by the use of distributed sequences in frame synchronisation applications by van Wijngaarden and Willink [7], Bajić and Stojanović [2] were the first to study non-overlapping codes. (Bajić and Stojanović used the term cross-bifix-free sets rather than non-overlapping codes.) See also [1, 3, 4, 5, 6, 7] for various aspects of non-overlapping (cross-bifix-free) codes and their applications to synchronisation.

Chee, Kiah, Purkayastha and Wang [6] showed that

\[
C(n,q) \leq \frac{q^n}{2^n - 1},
\]

and provided a construction of a class of non-overlapping codes whose cardinality is within a constant factor of the bound (1) when the alphabet size \( q \) is fixed and the length \( n \) tends to infinity.

Chee et al. established the bound (1) by appealing to the application in synchronisation (deriving the bound from the fact that a certain variance must be positive). In Section 2 below, we provide a direct combinatorial proof of this bound. (Indeed, the combinatorial derivation allows us to improve the bound slightly to a strict inequality.)

The construction of Chee et al. becomes poor for large alphabet sizes (in the sense that the ratio between the number of codewords in the construction
and the upper bound tends to 0 as \( q \) increases). In Section 3 we provide a simple generalisation of their construction which performs well when \( q \) is large. Indeed, in Section 5 we show that this generalised construction produces non-overlapping codes whose cardinality is within a constant factor of the upper bound even when the alphabet size \( q \) is allowed to grow.

In Section 4 we provide exact values for \( C(2, q) \) and \( C(3, q) \); these values show that the bound is not always asymptotically tight. We also state a conjecture on the asymptotics of \( C(n, q) \) when \( n \) is fixed and \( q \) tends to infinity.

2 An upper bound

We provide a direct combinatorial proof of the following theorem, that slightly strengthens the bound due to Chee et al. [6].

**Theorem 1.** Let \( n \) and \( q \) be integers with \( n \geq 2 \) and \( q \geq 2 \). Let \( C(n, q) \) be the number of codewords in the largest non-overlapping \( q \)-ary code of length \( n \). Then

\[
C(n, q) < \frac{q^n}{2n - 1}.
\]

**Proof.** Let \( C \) be a non-overlapping code of length \( n \) over an alphabet \( F \) with \( |F| = q \). Consider the set \( X \) of pairs \((w, i)\) where \( w \in F^{2n-1} \), \( i \in \{1, 2, \ldots, 2n-1\} \) and the (cyclic) subword of \( w \) starting at position \( i \) lies in \( C \). So, for example, if \( C \) is the code in the introduction then \((01111110011, 8)\) \( \in X \).

We see that \( |X| = (2n - 1)|C| q^{n-1} \), since there are \( 2n - 1 \) choices for \( i \), then \( |C| \) choices for the codeword starting in the \( i \)th position of \( w \), then \( q^{n-1} \) choices for the remaining positions in \( w \).

Since \( C \) is non-overlapping, two codewords cannot appear as distinct cyclic subwords of any word \( w \) of length \( 2n - 1 \). Thus, for any \( w \in F^{2n-1} \) there is at most one choice for an integer \( i \) such that \((w, i) \in X \). Moreover, no subword of any of the \( q \) constant words \( w \) of length \( 2n - 1 \) can appear as a codeword in a non-overlapping code. So \( |X| \leq q^{2n-1} - q < q^{2n-1} \).

The theorem now follows from the inequality

\[
(2n - 1)|C| q^{n-1} \leq |X| < q^{2n-1}.
\]
3 Constructions of non-overlapping codes

Let $F = \{0, 1, \ldots, q-1\}$. Chee et al. provide the following construction of a non-overlapping code of length $n$ over $F$.

**Construction 1** (Chee et al. [6]). Let $k$ be an integer such that $1 \leq k \leq n-1$. Let $C$ be the set of all words $c \in F^n$ such that:

- $c_i = 0$ for $1 \leq i \leq k$ (so all codewords start with $k$ zeroes);
- $c_{k+1} \neq 0$, and $c_n \neq 0$;
- the sequence $c_{k+2}, c_{k+3}, \ldots, c_{n-1}$ does not contain $k$ consecutive zeroes.

Then $C$ is a non-overlapping code.

It is not hard to see that the construction above is indeed a non-overlapping code. Chee et al. show that the construction is already good for small parameters. Indeed, they show that for binary codes, Construction 1 (with the best choice of $k$) achieves the best possible code size whenever $n \leq 14$ and $n \neq 9$.

It less clear how to choose $k$ in general so that $C$ is as large as possible, and what the resulting asymptotic size of the code is. Much of the paper of Chee et al. sets out to answer these questions. Indeed, the authors argue that when $q$ is fixed, and $k$ is chosen appropriately (as a function of $n$), we have that

$$\liminf_{n \to \infty} \frac{|C|}{(q^n/n)} \geq \frac{q-1}{qe},$$

where $e$ is the base of the natural logarithm. This shows that Theorem [1] is tight to within a constant factor when $q$ is fixed. Their result uses a delicate argument using techniques from algebraic combinatorics. In fact, the following much simpler argument gives a similar, though weaker, result.

**Lemma 2.** Let $q$ be a fixed integer, $q \geq 2$. Then the codes in Construction 1 show that

$$\liminf_{n \to \infty} C(n,q)/(q^n/n) \geq \frac{(q-1)^2(2q-1)}{4q^4}.$$
To see this, note that any sequence that fails the condition of containing no $k$ consecutive sequences of zeroes must contain $k$ consecutive zeros starting at some position $i$, where $1 \leq i \leq n - k - 2 - (k - 1)$. Since there are $n - 2k - 1$ possibilities for $i$, and $q^{n-2k-2}$ sequences containing $k$ zeros starting at position $i$, our claim follows. Thus, if $C$ is the non-overlapping code in Construction \( \text{[II]} \)

\[ |C| \geq (q - 1)^2(q^{n-k-2} - nq^{n-2k-2}) = \left(\frac{q - 1}{q}\right)^2 q^n(q^k - nq^{-2k}). \]

The function $q^{-k} - nq^{-2k}$ is maximised when $k = \log_2(q)(2n) + \delta$, where $\delta$ is chosen so that $|\delta| < 1$ and $k$ is an integer. In this case, the value of $q^{-k} - nq^{-2k}$ is bounded below by $(2q - 1)/(4n^2q^2)$ (this can be shown by always taking $\delta$ to be non-negative). Thus

\[ |C| \geq \left(\frac{(q - 1)^2(2q - 1)}{4n^4}\right) q^n. \]

When the alphabet size $q$ is much larger than the length $n$, Construction \( \text{[I]} \) produces codes that are much smaller than the upper bound in Theorem \( \text{[I]} \).

The following generalisation of Construction \( \text{[II]} \) does not have this drawback; we discuss this issue further in Sections \( \text{[I]} \) and \( \text{[II]} \) below.

Let $S \subseteq F^k$. We say that a word $x_1x_2\cdots x_r \in F^r$ is $S$-free if $r < k$, or if $r \geq k$ and $x_i x_{i+1} \cdots x_{i+k-1} \notin S$ for all $i \in \{1, 2, \ldots, r - k + 1\}$.

**Construction 2.** Let $k$ and $\ell$ be such that $1 \leq k \leq n - 1$ and $1 \leq \ell \leq q - 1$. Let $F = I \cup J$ be a partition of a set $F$ of cardinality $q$ into two parts $I$ and $J$ of cardinalities $\ell$ and $q - \ell$ respectively. Let $S \subseteq I^k \subseteq F^k$. Let $C$ be the set of all words $c \in F^n$ such that:

- $c_1c_2\cdots c_k \in S$;
- $c_{k+1} \in J$, and $c_n \in J$;
- the word $c_{k+2}, c_{k+3}, \ldots, c_{n-1}$ is $S$-free.

Then $C$ is a non-overlapping code.

It is easy to see that Construction \( \text{[II]} \) is the special case of Construction \( \text{[II]} \) with $\ell = 1$, $I = \{0\}$ and $S = \{0^k\}$.  

4 Non-overlapping codes of small length

This section considers non-overlapping codes of fixed length \( n \), when the alphabet size \( q \) becomes large. In this situation, Construction 1 produces codes that are much smaller than the upper bound in Theorem 1. To see this, note that there are at most \( q^n - k \) codewords in a code \( C \) from Construction 1, since the first \( k \) components of any codeword are fixed. So, since \( k \) is positive, \(|C| \leq q^n - 1 \) and therefore \(|C| / (q^n/n) \leq n/q \).

The proof of the following theorem shows that the codes given by Construction 2 are within a constant factor of the bound in Theorem 1 whenever \( n \) is fixed and \( q \to \infty \).

**Theorem 3.** Let \( n \) be a fixed positive integer, \( n \geq 2 \). Then

\[
\liminf_{q \to \infty} C(n, q) / (q^n/n) \geq \left( \frac{n-1}{n} \right)^{n-1}.
\]

**Proof.** We use Construction 2 in the special case when \( k = n - 1 \) and \( S = I^k \). In this case (in the notation of Construction 2) \( C \) is the set of words whose first \( n - 1 \) components lie in \( I \), and whose final component lies in \( J \). So here \(|C| = \ell^{n-1}(q - \ell)\).

Let \( \ell = \lceil((n-1)/n)q \rceil \). Since \( q - \ell \geq (1/n)q - 1 \), we find that

\[
|C| = \frac{1}{n} \left( \frac{n-1}{n} \right)^{n-1} q^n - O(q^{n-1}),
\]

and so the theorem follows. \( \square \)

The following theorem shows (in particular) that the bound \((|C| < q^2/3)\) of Theorem 1 is not asymptotically tight when \( n = 2 \) and \( q \to \infty \).

**Theorem 4.** A largest \( q \)-ary length 2 non-overlapping code has \( C(2, q) \) codewords, where \( C(2, q) = \lfloor q/2 \rfloor \lceil q/2 \rceil \). In particular,

\[
\lim_{q \to \infty} C(2, q) / (q^2/2) = \frac{1}{2}.
\]

**Proof.** Construction 2 in the case \( n = 2 \), \( k = 1 \), \( \ell = \lfloor q/2 \rfloor \) and \( S = I^k \) provides the lower bound on \( C(2, q) \) we require.

Let \( C \) be a \( q \)-ary non-overlapping code of length \( q \). Let \( I \) be the set of symbols which occur in the first position of a codeword in \( C \), and let \( J \) be
the set of symbols that occur in the final position of a codeword in $C$. Since $C$ is non-overlapping, $I$ and $J$ are disjoint. Thus

$$|C| \leq |I||J| \leq |I|(q - |I|) \leq \lfloor q/2 \rfloor \lceil q/2 \rceil.$$ 

\[ \Box \]

In the following theorem, $[x]$ denotes the nearest integer to the real number $x$.

**Theorem 5.** A largest $q$-ary length 3 non-overlapping code has $C(3, q)$ codewords, where $C(3, q) = 2q/3(q - \lfloor 2q/3 \rfloor)$. In particular,

$$\lim_{q \to \infty} C(3, q)/(q^2/3) = \frac{4}{9}.$$ 

**Proof.** Construction 2 in the case $n = 3$, $k = 2$, $\ell = \lfloor 2q/3 \rfloor$ and $S = I^k$ provides the lower bound on $C(2, q)$ we require.

Let $C$ be a $q$-ary non-overlapping code of length $q$ of maximal size. Let $F$ be the underlying alphabet of $C$, so $|F| = q$.

Let $I$ be the set of symbols which occur in the first position of a codeword in $C$. Let $J$ be the complement of $I$ in $F$, so $|J| = q - |I|$. Since $C$ is non-overlapping, the symbols that occur in the final component of any codeword lie in $J$. So we may write $C$ as a disjoint union $C = C_1 \cup C_2$, where $C_1 \subseteq I \times I \times J$ and $C_2 \subseteq I \times J \times J$.

Let $X$ be the set of all pairs $(b, c) \in I \times J$ such that $abc \in C$ for some $a \in I$. Define

$$\overline{C_1} = \{abc \mid a \in I \text{ and } (b, c) \in X\},$$

$$\overline{C_2} = \{bcd \mid (b, c) \in (I \times J) \setminus X \text{ and } d \in J\}.$$ 

Clearly $C_1 \subseteq \overline{C_1}$. Moreover, $C_2 \subseteq \overline{C_2}$, since whenever $bcd \in C$ is a codeword, the fact that $C$ is non-overlapping implies that $(b, c) \notin X$. But $\overline{C} = \overline{C_1} \cup \overline{C_2}$ is a non-overlapping code, and so $C = \overline{C}$ as $C$ is maximal.

We have

$$|C| = |\overline{C}| = |X||I| + (|J||J| - |X|)|J| = |X|(|I| - |J|) + |I||J|^2.$$ 

If $|I| \leq |J|$, then the maximum value of $|C|$ is achieved when $|X| = 0$, at max$_{i \in \{1, 2, \ldots, q/2\}} i^2(q - i)$. If $|I| > |J|$, the maximum value of $|C|$ is achieved when $|X| = |I||J|$, at max$_{i \in \{q/2, \ldots, q/2 + 1, \ldots, q - 1\}} i^2(q - i)$. Thus

$$|C| \leq \max_{i \in \{1, 2, \ldots, q - 1\}} i^2(q - i) = [2q/3]^2(q - \lfloor 2q/3 \rfloor).$$
and so the theorem follows. \[ \square \]

It would be interesting to determine the asymptotic behaviour of \( C(n, q) \) when \( q \to \infty \) for a general fixed length \( n \). I believe the following two conjectures are true.

**Conjecture 1.** Let \( n \) be an integer such that \( n \geq 2 \). Then

\[
\lim_{q \to \infty} \frac{C(n, q)}{q^n} = \frac{1}{n} \left( \frac{n - 1}{n} \right)^{n-1}.
\]

The following conjecture implies Conjecture 1.

**Conjecture 2.** Let \( n \) be an integer such that \( n \geq 2 \). For all sufficiently large integers \( q \), a largest \( q \)-ary non-overlapping code of length \( n \) is given by Construction \(^2\) in the case \( k = n - 1 \) (and some value of \( \ell \)).

## 5 Good constructions for general parameters

This section shows that Construction \(^2\) is always good, in the sense that it produces non-overlapping codes of cardinality within a constant factor of the upper bound given by Theorem \(^1\) for all parameters. This is implied by the proof of the following theorem.

**Theorem 6.** There exist absolute constants \( c_1 \) and \( c_2 \) such that

\[
c_1(q^n/n) \leq C(n, q) \leq c_2(q^n/n)
\]

for all integers \( n \) and \( q \) with \( n \geq 2 \) and \( q \geq 2 \).

**Proof.** The existence of \( c_2 \) follows by the upper bound on \( C(n, q) \) given by Theorem \(^1\). We prove the lower bound by showing that there exists a constant \( c_1 \) such that for all choices of \( n \) and \( q \), one of the constructions given by Construction \(^2\) contains at least \( c_1(q^n/n) \) codewords.

Let \((n_1, q_1), (n_2, q_2), \ldots\) be an infinite sequence of pairs of integers where \( n_i \geq 2 \) and \( q_i \geq 2 \). It suffices to show that \( C(n_i, q_i)/(q_i^{n_i}/n_i) \) is always bounded below by some positive constant as \( i \to \infty \). Suppose, for a contradiction, that this is not the case. By passing to a suitable subsequence if necessary, we may assume that \( C(n_i, q_i)/(q_i^{n_i}/n_i) \to 0 \) as \( i \to \infty \). If the
integers \( q_i \) are bounded, then Lemma 2 gives a contradiction. If the integers \( n_i \) are bounded, we again have a contradiction, by Theorem 3. So we may assume, without loss of generality, that the integer sequences \( (n_i) \) and \( (q_i) \) are unbounded. By passing to a suitable subsequence if necessary, we may therefore assume that \( (n_i) \) and \( (q_i) \) are strictly increasing sequences (and that \( n_i \) and \( q_i \) are sufficiently large for our purposes below). In particular, we may assume that \( n_i \to \infty \) and \( q_i \to \infty \) as \( i \to \infty \).

Let \( k_i = \lceil \log_2 2n_i \rceil \), and set \( s_i = \lfloor q_i^{k_i}/(2n_i) \rfloor \). Let \( F_i \) be a set of size \( q_i \). Let \( I_i \subseteq F_i \) have cardinality \( \ell_i \), where \( \ell_i = \lceil s_i^{1/k_i} \rceil \). Let \( J_i \) be the complement of \( I_i \) in \( F_i \). Let \( S_i \) be a subset of \( I_i \) of cardinality \( s_i \). Note that such a set \( S_i \) exists, by our choice of \( \ell_i \).

Let \( C_i \) be the \( q_i \)-ary non-overlapping code of length \( n_i \) given by Construction 2 in the case \( k = k_i \), \( \ell = \ell_i \), \( I = I_i \), \( J = J_i \) and \( S = S_i \). Then

\[
|C_i| = |S|(q_i - \ell_i)^2f_i
\]  

(2)

where \( f_i \) is the number of \( S \)-free sequences of length \( n_i - k_i - 2 \). We now aim to find a lower bound on \( |C_i| \).

Since \( q_i \to \infty \) as \( i \to \infty \), we see that

\[
q_i^{k_i}/(2n_i) \geq q_i^{\log_2(2n_i)}/(2n_i) = 2^{(\log_2(q_i) - 1)(\log_2(2n_i))} \to \infty.
\]

Hence

\[
|S| \sim q_i^{k_i}/(2n_i)
\]  

(3)

as \( i \to \infty \).

Note that

\[
(2n_i)^{(1/k_i)} \geq 2^{\log_2(2n_i)/2\log_2(2n_i)} = 2^{1/2},
\]

and hence

\[
s_i^{1/k_i} \leq \left( \frac{q_i^{k_i}}{2n_i} \right)^{1/k_i} \leq 2^{-1/2} q_i.
\]

Since \((1 - 2^{-1/2})^2 > (1/12)\), we see that

\[
(q_i - \ell_i)^2 > (1/12)q_i^2
\]  

(4)

for all sufficiently large \( i \).

The number of \( S \)-free \( q \)-ary sequences of length \( r \) is at least \( q^r - (r - k + 1)|S|q^{r-k} \), since every word that is not \( S \)-free must contain an element of \( S \).
somewhere as a subword. So the number of $S$-free $q$-ary sequences of length $r$ is at least $q^r - r|S|q^{r-k} = q^r(1 - r|S|q^{-k})$. Thus

$$f_i \geq q_i^{n_i-k_i-2}(1 - (n_i - k_i - 2)|S_i|q_i^{-k_i})$$

$$\geq \frac{1}{2}q_i^{n_i-k_i-2}(2 - 2n_i|S_i|q_i^{-k_i})$$

$$\sim \frac{1}{2}q_i^{n_i-k_i-2},$$

the last step following from (3).

Now (3), (4) and (5) combine with (2) to show that $|C_i| > (1/50)(q_i^{n_i}/n_i)$ for all sufficiently large $i$. This contradiction completes the proof of the theorem.

References

[1] Dragana Bajić, ‘On construction of cross-bifix-free kernel sets’, presented at 2nd COST 2100 MCM, Lisbon, Portugal, February 2007.

[2] Dragana Bajić and Jakov Stojanović, ‘Distributed sequences and search process’, Proc. IEEE Internat. Conf. on Communications 2004, 514–518.

[3] Dragana Bajić, Jakov Stojanović and Juergen Lindner, ‘Multiple window-sliding search’, Proc. Int. Symp. Inform. Theory, 2003, 249.

[4] Stefano Bilotta, Elisabetta Grazzini, Elisa Pergola and Renzo Pinzani, ‘Avoiding cross-bifix-free binary words’, Acta Informatica, to appear.

[5] Stefano Bilotta, Elisa Pergola and Renzo Pinzani, ‘A new approach to cross-bifix-free sets’, IEEE Trans. Inform. Theory, 58 (2012), 4058–4063.

[6] Yeow Meng Chee, Han Mao Kiah, Punarbasu Purkayastha and Chengmin Wang, ‘Cross-bifix-free codes within a constant factor of optimality’, IEEE Trans. Inform. Theory, to appear. Preprint dated 3 September 2012 available at http://arxiv.org/abs/1209.0236.

[7] Adriaan J. de Lind van Wijngaarden and Tricia J. Willink, ‘Frame synchronization using distributed sequences’, IEEE Trans. Inform. Theory 48 (2000), 2127–2138.