数理解析研究所講究録 1366 巻 2004 年 129-140

FINITE COMPLETION OF COMMA-FREE CODES. Part II

NGUYEN HUONG LAM*

Hanoi Institute of Mathematics
P.O.Box 631, Bo Ho, 10 000 Hanoi, Vietnam

Abstract. This paper is a sequel to an earlier paper of the present author, in which it was proved that every finite comma-free code is embedded into a so-called (finite) canonical comma-free code. In this paper, it is proved that every (finite) canonical comma-free code is embedded into a finite maximal comma-free code, which thus achieves the conclusion that every finite comma-free code has finite completions.

Keywords. Comma-free Code, Completion, Finite Maximal Comma-free Code.

§1. Introduction. This paper continues the previous one of the present author [L]. Taken as a whole, they represent a solution to the problem of finite completion of comma-free codes.

The problem of completing a code of some class within this class is among problems in general theory of codes [BP] that have some attention of researchers in recent years. For (finite) prefix codes the problem is easy (positive answer), but for finite codes in general, the answer is negative and the argument is more sophisticated (see Restivo [R] or Berstel and Perrin [BP]).The situation is same for finite bifix codes: there exist finite bifix codes which are not included in any finite maximal bifix code [BP]. More on the positive side we can mention finite infix codes [IJST] and we can also prove that every finite *outfix* code is included in a finite maximal outfix code (a set X is an outfix code provided uv, $uxv \in X$ implies $x = 1$ for any words u, v, x).

As for comma-free code, in [L] we proved that every finite comma-free code is included in a so-called (finite) canonical comma-free code and in this paper we shall prove further that every finite canonical code is included in a finite maximal comma-free code. Thus we add one more class of codes having a positive ansewr to the finite completion problem.

This paper is organized as follows: In the next two sections we review some background and prove several simple technical statements which are almost folklore and will be used in later constructions. After that we prove an instrumental proposition, which enable us to make a ramification respective to the set of so-called ilr-words. If this set is finite (in §4) the completion is straightforward. Else, if infinite, this set contains a "short" ilr-word with rich properties and starting from this word we construct finite maximal comma-free codes, more or less explicit, that all contain the original comma-free code (in §5).

§2. Notions and Notation. We briefly specify our standard vocabulary and state some prerequisites.

Let A be a finite alphabet. Then A^* denotes the set of words on A including the empty word 1 and as usual A^+ denotes the set of non-empty words. For subsets of words we use interchangeably the plus and minus signs to denote the union and difference of them, besides the ordinary notation.

The set of words is equipped with the concatenation as product with the empty

* E-mail: nhlam@thevinh.ncst.ac.vn

word 1 as the unit. For subsets X and X' of A^* we denote

$$\begin{aligned} XX' &= \{xx' : x \in X, x' \in X'\} \\ X^0 &= \{1\} \\ X^{i+1} &= X^i X, \quad i = 0, 1, 2, \ldots \\ X^* &= \cup_{i \geq 0} X^i. \end{aligned}$$

Our subject-matter is comma-free codes which are defined as follows [S].

DEFINITION 2.1. *A subset $X \subseteq A^+$ is said to be a* comma-free *code if $X^2 \cap A^+ X A^+ = \emptyset$.*

A comma-free code is called *maximal* if it is not a proper subset of any other comma-free code. A completion of a comma-free is a maximal comma-free code containing it. In view of Zorn's lemma, every comma-free code always has completions.

EXAMPLE 2.2. Every primitive word constitutes a comma-free code. This means that for a primitive word p, $p^2 = up^2v$ implies $u = 1$ or $v = 1$.

We shall use frequently the following result (Fine and Wilf): If $u\{u, v\}^*$ and $\{u, v\}^*$ have a common left factor of length at least $|u| + |v|$, in particular, if $uv = vu$, then u and v are copowers.

Comma-free codes are closely connected to the notion of overlap. We say that two words u and v, not necessarily distinct, *overlap* if

$$u = tw, \quad v = ws$$

for some non-empty words $s, t \in A^+$ and $w \in A^+$, or equivalently,

$$us = tv$$

for some non-empty words s, t such that $|t| < |u|$ and $|s| < |v|$. We call w an *overlap*, s a *right border* and t a *left border* of the two *overlapping* words u, v. We say also that u *self-overlaps* if u and u overlap, that is, u overlaps itself. A right (left) border of a set X is a right (left, resp.) border of any two overlapping words of X. We denote the sets of right and left borders of X by $R(X)$ and $L(X)$, respectively.

With each comma-free code X we associate the following set, which plays a central role in our treatment

$$E(X) = A^+ - R(X)A^* - A^*L(X) - A^*XA^*.$$

We recall the principal object of this paper, which has been defined in the previous paper [L]. Let N be a positive integer.

DEFINITION 2.3. *A comma-free code X is called* N-canonical *if for an arbitrary word $w \in E(X)$ and an arbitrary factorization $w = xuy$ with $x, y, u \in A^*$ and $|u| \geq N$, there exist factorizations $u = pp' = ss'$ such that $xp \in E(X)$ and $s'y \in E(X)$, or just the same, $xp \notin A^*L(X)$ and $s'y \notin R(X)A^*$. A comma-free code is* canonical *if it is N-canonical for some N.*

Equivalently, a comma-free code X is N-canonical if and only if for any word $w \in E(X)$ and for any integer n, $0 < n \leq |w|$, there is a left factor p and a right factor s of w such that $n \leq |p|, |s| < n + N$ and $p, s \in E(X)$, or just the same, $p \notin A^*L(X)$ and $s \notin R(X)A^*$.

Our aim now is to prove that we can complete every finite N-canonical comma-free code to a finite maximal comma-free code.

Surely, we have to make a completion out of those words u outside X, for which $X + u$ is still a comma-free code. We term such words *good* words for X. Explicitely,

1. $u \notin A^*XA^*$.
2. $u \notin F(X^2)$.
3. $u \notin A^*L(X) + R(X)A^*$.
4. $u \in I(X) = \{u : u^2 \notin A^*XA^*\}$.
5. $A^+u \cap uP(X) = \emptyset$ and $uA^+ \cap S(X)u = \emptyset$.
6. u is primitive: $u \in Q$.

Let u be an arbitrary word. Consider the following conditions concerning u:

(r) u avoids X (i.e. u has no factors in X), u has no left factor in $R(X)$:

$$u \in A^+ - R(X)A^* - A^*XA^*.$$

(l) u avoids X, u has no right factor in $L(X)$:

$$u \in A^+ - A^*L(X) - A^*XA^*.$$

We call the words satisfying the conditions (r), (l), both (l) and (r) *r-words, l-words, lr-words* respectively. We call an lr-word *ilr-word* if, in addition, it satifies the condition 4 in the definition of a good word above. Notice that the set $E(X)$ mentioned in the definition of canonical comma-free codes is nothing but the set of lr-words. We also denote the set of l-words and r-words by $E_l(X)$ and $E_r(X)$, respectively.

The good word u is called *R-good* if uv avoids X for all r-words v. Similarly, u is *L-good* provided vu avoids X for all l-words v.

We say that the word u is an Lr-lr-word if it is an lr-word and for all r-words v, vu avoids X. Similarly, we say that u is an Rl-lr-word if it is an lr-word and for all l-words v, uv avoids X.

§3. Auxiliary Technical Results. We present several preliminary lemmas here in one section for easy reference in the sequel. First we discuss the notion of sesquipower, which is closely connected to the notion of self-overlap.

Let k be a positive real number, the word w is called a *k-sesquipower* if it is a left factor of u^+ for some word u of length less or equal to k, $|u| \leq k$, or equivalently, w is is a right factor of v^+ for some word v of length $|v| \leq k$. We have the following assertion, which is a folklore, relating sesquipowers to self-overlapping words.

PROPOSITION 3.1. *For any words x, y and u the following assertions are equivalent:*

(i) $xu = uy$,

(ii) u *is a left factor of* x^+,

(iii) u *is a right factor of* y^+,

(iv) $x = pq, u = (pq)^s q = p(qp)^s, y = qp$ *for some words* p, q.

It is straightforward to see that the if $|w| > k$, w is k-sesquipower if and only if w self-overlaps with borders no longer than k. So in the sequel if we want to prove some word not to self-overlap with borders which are left or right factors of X we just show that it is not a k-sesquipower for $k \geq \max\{|x| : x \in X\}$.

In the following simple statement we show that we can pick out of three special words, not self-overlapping with short borders, a primitive one. Let N be a possitive integer.

PROPOSITION 3.2. *Let u, v_1, v_2 be non-empty words such that $|u| \geq 3N$, $|v_1| \leq N$, $|v_2| \leq N$. Suppose that u, uv_1 and uv_1v_2 do not self-overlap with borders shorter or equal to N. Then at least one of them is primitive.*

The proof of the proposition requires two lemmas below.

LEMMA 3.3. *Let u and v be words such that $|u| \geq 3N, 0 < |v| \leq N$, $u = \lambda^m, uv = \mu^n$ with primitive words λ, μ and integers $m \geq 2, n \geq 2$. If not both of u and uv self-overlap with borders of length shorter than or equal to N then $m = n = 2$.*

LEMMA 3.4. *Let u and v be non-empty words such that $|u| > |v|$ and $u = \lambda^2, uv = \mu^2$ for some primitive words λ, μ. Then $\mu = \lambda\bar{\lambda}^n$ for some positive integer n and some primitive word $\bar{\lambda}$ such that λ is a left factor of $\bar{\lambda}^+$ and $|\bar{\lambda}| < \frac{|v|}{2}$.*

The following lemma will be used later in the proof of the existence of short ilr-words.

LEMMA 3.5. *Let w not be a k-sesquipower and let u be the longest proper left factor of w, $w = uv$ and $v \neq 1$, which is a k-sesquipower, that is, $u = u_1^s u_2$ with u_2 a proper left factor of u_1, u_1 primitive and $|u_1| \leq k$. Then for every integer t such that $|u_1^t u_2| \geq \min(|u_1^t u_2|, 2k)$ the word $u_1^t u_2 v$ is not a k-sesquipower.*

The following fact is left as an easy exercise.

LEMMA 3.6. *Let p not be a factor of q and $|q| \geq 2|p|$. Then qp^n is primitive for all integers $n > 0$.*

§4. **Short ilr-words.** Let X be a finite N-canonical word with $m = \max\{|x| : x \in X\}$. Suppose that h is a primitive ilr-word for X of length greater than m. We put $K = \max(N, m)$ and $f = h^k$, where $k \geq \frac{6K+6N}{|h|}$. We have the following key statement.

THEOREM 4.1. *f^2 contains a factor of length greater than $3K$ and less than or equal to $3K + 3N$ which is either an R-good or an L-good word.*

Proof. We first prove that f has a factorization $f = f'f''$ such that either f' is an Rl-lr-word or f'' is an Lr-lr-word and

$$\left\lfloor \frac{|f|}{2} \right\rfloor - m < |f'|, |f''| < \left\lfloor \frac{|f|}{2} \right\rfloor + m.$$

Let $f = f_1 f_2$ be a factorization such that

$$\left\lfloor \frac{|f|}{2} \right\rfloor + 1 \geq |f_1|, |f_2| \geq \left\lfloor \frac{|f|}{2} \right\rfloor.$$

Note that f_1 is an l-word and f_2 is an r-word. Suppose that there exists a word u_1 of X or $E_l(X)$ such that f_1u_1 contains a factor, not a right one, in X

$$f_1u_1 \in A^*XA^+.$$

Since f_1 avoids X and u contains no proper factor in X, we see that f_1 has a right factor x_1 which is a proper left factor of X:

$$f_1 = f_1' x_1.$$

Consider now the word $x_1 f_2$. If there is some word u_2 in $X + E_r(X)$ such that $u_2 x_1 f_2$ contains a factor, not a left one, $x \in X$, that is

$$u_2 x_1 f_2 = wxv$$

for some words $w \in A^+$ and $v \in A^*$. Since $x_1 f_2$, being a factor of f, avoids X and u_2 does not contains x if $u_2 \in E_r(X)$ and does not contain properly x if $u_2 \in X$, we have

$$|w| < |u_2| < |wx|.$$

Thus, $x_1 f_2$ has a left factor x_2 which is a right factor (of x) in X and we can write

$$x_1 f_2 = x_2 f_2'.$$

We have then a factorization

$$f = f_1 x_2 f_2'$$

with $|f_1 x_2| < \left\lfloor \frac{|f|}{2} \right\rfloor + m$ and $|f_1 x_2| = |f_2| - |x_2| > \left\lfloor \frac{|f|}{2} \right\rfloor - m$, because $0 < |x_2| < m$. Note that

$$|x_1| < |x_2|.$$

Now we proceed similarly with the latter factorization and with $f_1 x_2$ playing the role of f_1 in the former factorization for f to obtain a left factor x_3 of x and some factorization of f with the relevant relations, such that

$$|x_1| < |x_2| < |x_3|$$

and so on. However we cannot iterate the argument infinitely, as the length of factors of X are bounded by m. So we stop in some step, no later than the $m-1$-th one, to obtain a factorization

$$f = f' f''$$

with the claimed properties regarding as on which step we get stuck, even or odd.

Suppose for definiteness that f'' is an Lr-lr-word. Recall that $|f''| > \left\lfloor \frac{|f|}{2} \right\rfloor - m$. Let u be the longest left factor which is an m-sesquipower of $f'' f' f''$. We write

$$u = u_1^s u_2$$

for $s \geq 0$ and u_2 is a proper left factor of u_1. Since f is a power of a primitive word, h, of length longer than m and $|u_1| \leq m$ by Fine and Wilf we have

$$|u| < |f| + m$$

otherwise $u_1 \in h^+$, hence $|u_1| > m$, a contradiction.

Put $u_0 = u$ if $|u| < 2m$ and $u_0 = u_1^t u_2$, where t is the smallest integer such that $|u_1^t u_2| \geq 2m$, otherwise. In any case, we have

$$\min(|u|, 2m) \leq |u_0| < 3m.$$

Note that u_0 is a right, and left, factor of u. Now let u_3 be the left factor of $f''f'f''$ of length $3K$, we see that u_0 is a proper left factor of u_3. We have the following relations for some words $l \in A^*$, $r, v \in A^+$

$$f''f'f'' = lu_0rv = lu_3v,$$

where

$$lu_0 = u, \quad u_0r = u_3.$$

If $u = u_0$, that is if $l = 1$, then

$$|v| = |f''f'f''| - |u_0| > |f''f'| - 3m > \left\lfloor \frac{|f|}{2} \right\rfloor - m + |f| - 3m$$
$$\geq \frac{3}{2}|f| - 4m \geq 9K + 9N - 4m > 3N.$$

If $|u| \geq 2K$ then $|u_0| \geq 2m$ and $|l| = |u| - |u_0| < |f| + m - 2m = |f| - m$, hence

$$|v| = |f''f'f''| - |l| - |u_0r| > |f''| + |f| - (|f| - m) - |u_3| = |f''| + |m| - 3K$$
$$> \left\lfloor \frac{|f|}{2} \right\rfloor - m + m - 3K \geq \frac{|f|}{2} - 3K = 3N.$$

Now we use the hypothesis. Since X is N-canonical and $f^2 \in E(X)$, for the factorization

$$f^2 = f'lu_3v$$

with respect to the factor v of length $|v| \geq 3N$, there exist three words v_1, v_2, v_3 such that $0 < |v_1|, |v_2|, |v_3| \leq N$ and $v_1, v_1v_2, v_1v_2v_3$ all are left factors of v and

$$f'lu_3v_1, \quad f'lu_3v_1v_2, \quad f'lu_3v_1v_2v_3 \notin A^*L(X)$$

which means

$$u_3v_1, \quad u_3v_1v_2, \quad u_3v_1v_2v_3 \notin A^*L(X).$$

because of the large length ($> m$) of the latter words.

If $|u| \leq 2K$ then $u_0 = u$ and u is a proper left factor of all three u_3v_1, $u_3v_1v_2$, $u_3v_1v_2v_3$, hence all of them cannot be m-sesquipowers in view of the maximality of $|u|$. If $|u| \geq 2K$ then by Lemma 3.5 all of them cannot be m-sesquipowers either. So in any case

$$u_3v_1, \quad u_3v_1v_2, \quad u_3v_1v_2v_3$$

are not m-sesquipowers.

Moreover, by Proposition 3.2, as $|u_3| = 3K \geq 3N$, one of the three words, say, $u_3v_1v_2v_3$, should be primitive.

Now it is routine to verify that $g = u_3v_1v_2v_3$ is a good word, and more than that, an L-good one. Let us, for instance, check the points (3), (4) and (5) in the definition of a good word.

Clearly $g \notin A^*L(X)$. The fact that $g \notin R(X)A^*$ follows from the fact that u_3 is a left factor of f'' if $u_0 = u$ and u_3 has u_0 as a left factor, which is a left factor of u of length at least $2K \geq 2m > m$, hence of f'' if $|u| \geq 2K$. This shows also that g is Lr-lr-word, as f'' is so. Finally (5) holds, otherwise, g is an m-sesquipower, a contradiction. Certainly

$$3K < |g| = |u_3| + |v_1| + |v_2| + |v_3| \leq 3K + 3N$$

what is desired to prove.

In virtue of Theorem 4.1, we have the following dichotomy. First, there are no primitive ilr-words of length longer than m. Because good words are primitive ilr-words, all of them have length shorter or equal to m. In order to complete X, then, all we have to do is to search for appropriate goods words among the words of length not exceeding m. Second, there is a primitive ilr-word longer than m. This implies the existence of an L or R-good word, claimed in Theorem 4.1.

Nevertheless, how could we know in which branch of the dichotomy we are? The answer is an easy consequence of the following results by Ito, Katsura, Shyr and Yu [IKSY]:

PROPOSITION 4.2. *Let R be a regular set accepted by a deterministic automaton consisting of $n > 1$ states. Then*

(i) R contains a primitive word if and only if it contain a primitive word of length not exceeding $3n - 3$

(ii) R contains infinitely many primitive words if and only if it contains a primitive word of length in the range $[n, 3n - 3]$

PROPOSITION 4.3. *If R contains only a finite number of primitive words then all of them have length less than n.*

The next section is devoted to the completing X, starting from an L- or R-good word.

§5. Short Good Words. We may now suppose that we dispose of, say, an L-good word g satisfying

$$3K < |g| \leq 3K + 3N$$

in view of the discussion in the preceeding section. In order to complete X. We follow the steps below:

(a) If for almost all (i.e. all but finitely many) primitive ilr-words v, v contains a factor in $X+g$, or, vg contains a factor in X or an occurences of g different from the last one (this issue we can effectively test in view of Proposition 4.2) then the set of good words for $X + g$ is finite (the maximum length is effectively computable by Proposition 4.3) and we are finished. Otherwise

(b) We can effectively pick out a primitive ilr-word v such that

$$|v| > 2|g|$$

and vg contains no occurrence of any word in $X + g$, except the last one (of g). We state that vg is both an L- and an R-good word for X. Indeed,

1. vg is both an Lr- and an Rl-lr-word, because of the current assumption on g and on the set of ilr-words.

2. vg is not in $F(X^2)$, as $|vg| > |g| > 3m > 2m$, too long to be a factor of X^2.

3. vg is primitive, in view of Lemma 3.6.

4. vg is not a $6K$-sesquipower (hence not a m-sesquipower). Because from any equality for the overlapping

$$xvg = vgy$$

where $x, y \in A^+$, $|x| = |y| < |vg|$, it follows $|x| > |y|$ for g does not contain v and vg does not contain any occurences of g different from the last one. Thus the borders are longer than $|v| > 2|g| > 6K \geq 6m$.

(c) Put $p = vg$. So p is both an L- and an R-good word and $|p| > 3|g| > 9K \geq 9m$. It may self-overlap only with borders longer than $6m$.

If for almost all l-words $w \in E_l(X)$, either wp contains a factor in X or w contains p then we are done, the comma-free code $X + p$ has only a finite number of good words (of course, the hypotheses can be effectively tested), we can complete it at least by trial. Otherwise we can choose (again, effectively) an l-word $q \in E_l(X)$ with $|q| \geq 2|p|$ such that qp avoids $X + p$ and q does not contain any occurrence of p other than the last one.

By Lemma 3.6 qp^i is primitive for all positive integers i. We choose a positive integer n satisfying

$$(n-1)|p| > |q| + 6N.$$

Note that $n > 2$. We have first

REMARK 5.1. It is routine to check that qp^{n+1} is a good word for X.

Let G_i, for every $i = 0, 1, \ldots, n-1$, be the set consisting of words of the form

$$up^i qp^n$$

satisfying the following conditions.

(i) $|u| \geq |p|$

(i) u is an l-word and up avoids X: $u \in E_l(X)$, $up \notin A^*XA^*$

(iii) p is not a right or left factor of u

(iiii) $up^i qp^n$ is primitive: $up^i qp^n \in Q$.

We have a few preliminary remarks.

REMARK 5.2. Since $|p| > 9m > m$ and p is primitive, $|q| \geq 2|p|$ and p is not a factor of q, all words of G_i are not m-sesquipowers.

REMARK 5.3. All words of G_i avoid X and are not factors of X^2.

REMARK 5.4. All words of G_i are ilr-words, $G_i \subseteq E(X)$, because u is an l-word and p is an R-good word.

REMARK 5.5. If $up^i qp^n$ has another occurrence of p^n, apart from the last one, then it must occur in up if $i > 0$ and in uq if $i = 0$. This is because $|q| \geq 2|p|$, q does not contain p, $n > 2$ and p is primitive.

These remarks give rise to the following assertion.

PROPOSITION 5.6. *(g) Every word of G_i is a good word for X.*
(gg) All words of G_i are not factors of $p^n qp^n$.

Next, we define the set H as follows: H consists of the words of the form vp^n satisfying

(j) $|v| \geq |q|$ $(\geq 2|p| > |p|)$

(jj) v is l-word and vp avoids X, in other words, vp is l-word: $vp \in E_l(X)$.

(jjj) p is not a right or left factor of v, q is not a right factor of v.

(jjjj) vp^n is primitive: $vp^n \in Q$.

It is routine to verify that the counterparts of Remarks 5.2 – 5.4 and Proposition 5.6 are also valid for H (instead of G_i). Also, by the similar reasons, we have

REMARK 5.7. If vp^n has another occurrence of p^n different from the last one, then it must be one in vp.

Set

$$\bar{G}_i = G_i - A^+G_i$$
$$\bar{H} = H - A^+H$$

as the sets of "minimal" words of G_i and H. The following proposition says that the "minimal" words are of bounded length, hence $\bar{G}_i$ and $\bar{H}$ are finite.

PROPOSITION 5.8. *(i) If wp^iqp^n is a lr-word with $n > i \geq 0$, $|w| \geq 6N + |p|$ and if p is not a right factor of w then wp^iqp^n has a right factor in G_i, hence in $\bar{G}_i$.*
(ii) If wp^n is an lr-word with $|w| \geq 6N + |q|$ and if both p, q are not right factors of w then wp^n has a right factors in H, hence in $\bar{H}$.

Proof. (i) Since $|w| \geq 6N + |p|$ and X is N-canonical, we can write

$$w = w'w_6w_5w_4w_3w_2w_1w_0$$

where $w' \in A^*, |w_0| = |p|, |w_j| \leq N$ and

$$w_j \dots w_1w_0p^iqp^n$$

is an l-word (hence a lr-word) for $j = 1, \dots, 6$. In view of Proposition 3.2, there exist two different integers

$$1 \leq s \leq 3 < t \leq 6$$

such that

$$w_s \dots w_1w_0p^iqp^n$$

and

$$w_t \dots w_1w_0p^iqp^n$$

both are primitive, for, first $|p^iqp^n| > 3N$, and, second all $w_j \dots w_1w_0p^iqp^n$, $j = 1, \dots, 6$ are not N-sesquipowers, as $|p| > 9K \geq 9N$, $n > 2$ and q has no factor p. Moreover, at least one of them has no left factor p, otherwise, p is self-overlaps with borders shorter than $(s - t)N < 6N \leq 6K$, which contradicts a property of p which says that p is not a $6K$-sesquipower. Say

$$w_s \dots w_1w_0p^iqp^n$$

has no left factor p. Finally,

$$w_s \dots w_1w_0p^iqp^n$$

as a factor of an lr-word, avoids X. All together, the facts above mean that

$$w_s \dots w_1w_0p^iqp^n \in G_i.$$

(ii) is handled analogously. The proposition is proved.

The following statement is an immediate consequence of the preceding proposition.

THEOREM 5.9. *Every word of $\bar{G}_i$ is no longer than $6N + (n + i + 1)|p| + |q| \leq 6N + 2n|p| + |q|$ for $i = 0, 1, \dots, n - 1$ and every word of $\bar{H}$ is no longer than $6N + n|p| + |q|$.*

We need a simple fact about the words of $\bar{G}_i$ and $\bar{H}$.

COROLLARY 5.10. *Every word of $\bar{G}_i$ and $\bar{H}$ has a unique occurence of p^n.*

PROPOSITION 5.11. *(h) No word of $\bar{H}$ is a factor of $\bar{G}_i$, for all $i = 0, 1, \ldots, n-1$, and vice versa.*
(hh) No word of $\bar{H}$ or $\bar{G}_i$ is a factor of qp^{n+1} and vice versa, qp^{n+1} is not a factor of $\bar{H}$ or $\bar{G}_i$, for all $i = 0, 1, \ldots, n-1$.
(hhh) No word of $\bar{G}_i$ is a proper factor factor of $\bar{G}_j$, $0 \leq i \leq j < n$.
(hhhh) No word of $\bar{H}$ is a proper factor of another word in $\bar{H}$.

Put now

$$\bar{X} = qp^{n+1} + \bigcup_{i=0}^{n-1} \bar{G}_i + \bar{H}.$$

Recall that every word of $\bar{X}$ is a good word for X. How long are the borders of $\bar{X}$? By a mild argument we can show that they are much longer than m which is helpful in proving the comma-freeness of $X + \bar{X}$.

As we might expect, all the constructions we have done so far aim at the following

THEOREM 5.12. *$X + \bar{X}$ is a comma-free code.*

Proof. Suppose the contrary that $X + \bar{X}$ is not comma-free. Then, in virtue of Proposition 5.11, we can assume that there exists some words, not necessarily distinct, $x_1, x_2, x_3 \in X + \bar{X}$ and $r, l \in A^*$ such that

$$x_1 x_2 = l x_3 r$$

and $|l| < |x_1|, |r| < |x_2|$.

All x_1, x_2, x_3 should be in $\bar{X}$ due to the following reasons: p is both an Lr- and an Rl-word, every word of $\bar{X}$ is a good word (a little more: product of any two words of $\bar{X}$ avoids X), the borders of $\bar{X}$ is larger than m and X is comma-free. But x_3 has an occurrence of p^n and every word of $\bar{X}$, different from qp^{n+1}, has only one occurrence of p^n, so the foregoing occurrence of p^n in x_3 must overlap x_1 and x_2, if $x_2 \neq qp^{n+1}$. However this possibility is ruled out since p is primitive, $n > 2$ and every word in $\bar{X}$ has no left factor p but has a right factor p^n. So we have $x_2 = qp^{n+1}$. Note that qp^{n+1} has exactly two occurrences of p^n, hence x_3 is a right factor of $x_1 qp^n$. If $x_3 = qp^{n+1}$ then p is a right factor of q, contradiction. Otherwise $x_3 \in \bar{G}_i$ or $x_3 \in \bar{H}$ then $p^n qp^n$ is a (right) factor of x_3, again contradiction and thus the proof is completed.

We present our ultimate statement, the completion theorem.

THEOREM 5.9. *The finite comma-free code $X + \bar{X}$ is maximal.*

Proof. It suffices to prove that good words for X are no longer good ones for $X + \bar{X}$. It can be done as follows.

Let f be an arbitrary good word for X. Consider the word f^l with l arbitrarily large but fixed integer.

1. If f is a factor of qp^{n+1} then obviously f is not a good word for $X + \bar{X}$. Now suppose that f is not a factor of qp^{n+1}. If p^i is a factor of f^l then

$$i|p| < |f| + |p|$$

otherwise, by Fine and Wilf and primitivity of f, f is a conjugate of p, hence a factor of p^2 and all the more a factor of qp^{n+1}, despite the assumption. So we get

$$i < \frac{|f|}{|p|} + 1$$

which simply means that i is bounded.

2. Suppose that f^l contains an occurrence of p^{n+1}:

$$f^l = rp^{n+1}s$$

for some words r, s with r sufficiently long and p not being a right factor of r. If, however, q is a right factor of r then f^l contains qp^{n+1} and f is not good for $X + \bar{X}$. If q is not a right factor of r then rp^{n+1} is an (sufficiently long) lr-word for X, as f is so. Therefore rp^{n+1} contains a right factor in $\bar{H}$ in virtue of Proposition 5.7 (ii), that is, in $\bar{X}$, and we are done for this alternative.

3. Now suppose that f^l contains no occurrence of p^{n+1}. Consider the word

$$f^l qp^{n+1}.$$

If it has a factor in X, clearly, it cannot be a good word for $X + \bar{X}$. Else, consider the word

$$f^l qp^n.$$

Denote w the longest right factor of $f^l qp^n$ which is in $(qp^n)^*$. Certainly $|w| \geq |qp^n|$. On the other hand, by Fine and Wilf

$$|w| \leq |qp^n| + |f| + |qp^n|,$$

because in the opposite case, $f = qp^n$ in view of primitivity of both f and qp^n. Contradiction (or f is not good for $X + \bar{X}$).

Let write $w = (qp^n)^{d+1}$, $d \geq 0$, and

$$f^l qp^n = rw = r(qp^n)(qp^n)^d.$$

Let further p^i be the longest right factor of r in p^*. Since f^l is free from any occurrence of p^{n+1}, we have $i \leq n$. We write

$$r = tp^i$$

for some words t such that p is not a right factor of t.

If $i = n$, by maximality of $|w|$, q is not a right factor of t. This implies that $r = tp^n$ has a (right) factor in $\bar{H}$, as r, therefore t, is chosen arbitrarily large at the onset. Thus

$$f^l qp^n = rw$$

contains a factor in $\bar{H} \subseteq \bar{X}$ and f is not a good word for $X + \bar{X}$.

Last possibility, if $0 \leq i < n$ then

$$tp^i qp^n$$

has a (right) factor in $\bar{G}_i$ and the word

$$f^l q p^n = t p^i w$$

has a factor in $\bar{X}$: f is not a good word for $X + \bar{X}$ either, which thus concludes the proof.

References

[BP] J. Berstel, D. Perrin, "Theory of Codes", Academic Press, Orlando, 1985.

[GGW] S. W. Golomb, B. Gordon, L. R. Welch, *Comma-free Codes*, Canad. J. Math. **10**(1958)202–209.

[GVD] S. W. Golomb, L. R. Welch, M. Delbrück, *Construction and Properties of Comma-free Codes*, Biol. Medd. Dan. Vid. Selsk. **23**(1958), 3–34.

[IKSY] M. Ito, M. Katsura, H. J. Shyr, S. S. Yu, *Automata Accepting Primitive Words*, Semigroup Forum **37**(1988), 45–52.

[J] B. H. Jiggs, *Recent Results in Comma-free Codes*, Canad. J. Math. **15**(1963), 178–187.

[L] N. H. Lam, *Finite Completion of Comma-Free Codes. Part 1*, to appear in the Proceedings of DLT2002, Lecture Notes in Computer Science, Springer.

[IJST] M. Ito, H. Jürgensen, H. J. Shyr, G. Thierrin, *Outfix and Infix Codes and Related Classes of Languages*, Journal of Computer and System SCiences **43**(1991), 484–508.

[R] A. Restivo, *On Codes Having No Finite Completions*, Discreet Mathematics **17** (1977), 306-316.

[S] H. J. Shyr, "Free Monoids and Languages", Lecture Notes, Hon Min Book Company, Taichung, 2001.