RAIRO-Inf. Theor. Appl. **38** (2004) 117-136
DOI: 10.1051/ita:2004007

FINITE COMPLETION OF COMMA-FREE CODES PART 2

Nguyen Huong Lam[1]

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.

Mathematics Subject Classification. 68R15, 68S05.

1. Introduction

This paper continues the previous one of the present author [8]. 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 [1] 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, he answer is negative and the argument is more sophisticated (see Markov [10] or Restivo [11] or Berstel and Perrin [1]). The situation is same for finite bifix codes: there exist finite bifix codes which are not included in any finite maximal bifix code [1]. More on the positive side we can mention finite infix codes [6] 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 codes, in [8] we proved that every finite comma-free code is included in a so-called (finite) canonical comma-free code and in this paper we

Keywords and phrases. Comma-free code, completion, finite maximal comma-free code.

[1] Hanoi Institute of Mathematics, 18 Hoang Quoc Viet Road, 10 307 Hanoi, Vietnam; e-mail: nhlam@math.ac.vn

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 answer 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 i-words. If this set is finite (in Sect. 4) the completion is straightforward. Else, if infinite, this set contains a "short" special word with rich properties and starting from this word we construct maximal comma-free codes, more or less explicit, that all contain the original comma-free code (in Sect. 5).

2. Notions and notation

We briefly specify our vocabulary which is standard 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, beside the ordinary notation.

The set of words is equipped with the concatenation as product: the product of two words u and v is the concatenation uv and $u1 = 1u = u$ for all words u. For subsets X and X' of A^* we denote

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

For $w \in A^*$ we denote by $|w|$ the length of the word w. Note that $|uv| = |u| + |v|$ for every $u, v \in A^*$ and $|1| = 0$.

Let u and v be two words of A^*. The word u is a factor of v if $v = xuy$, a right factor if $v = xu$ and left factor if $v = uy$ for some words $x, y \in A^*$. A factor u is proper if it is not 1 or the whole word v. We denote by $F(X)$ the set of factors of the words in X.

A subset of words is an infix code (prefix code, suffix code, bifix code) if no word of it is a factor (left factor, right factor, both left and right factor, resp.) of another. Of course an infix code is a prefix code, suffix code and bifix code, simultaneously. Our subject-matter is comma-free codes which are defined as follows [12].

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

A comma-free code is of course a code in the general sense of [1], moreover, it is an infix code, hence, a prefix, suffix and bifix code, which is easily verified by the definition. One useful criterion for testing comma-freeness is that X is comma-free iff $(S \cap P)P \cap X = \emptyset$ iff $S(S \cap P) \cap X = \emptyset$, where P and S are the sets of proper left and right factors, respectively. A comma-free code is called *maximal* if it is 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.

Two words u and v are conjugates, or, in other words, u is a conjugate of v, and *vice versa*, if $u = xy$ and $v = yx$ for some words $x, y \in A^*$. A non-empty word is called primitive if it is not a power of another word, namely, the equality $u = v^n$ for any word v and integer $n > 0$ implies $v = u$, or, $n = 1$. A conjugate of a primitive word is also a primitive word. Every non-empty word is a power of a unique primitive word, which we called the primitive root of it. Primitive words have a "synchronizing" property expressed in the following

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

We shall use frequently the following result of Fine and Wilf [2]: let u and v be two words such that $u\{u, v\}^*$ and $v\{u, v\}^*$ have a common left factor of length longer than or equal to $|u| + |v| - \gcd(|u|, |v|)$ then u and v are powers of a same word (copowers). As a matter of fact, we use a weaker, but equally effective in practice, form of this result: if $u\{u, v\}^*$ and $v\{u, v\}^*$ 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 with 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

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

which consists of the words not containing any left factor in $R(X)$, any right factor in $L(X)$ and any factor in X.

We recall the principal object of this paper, which has been defined in the previous paper [8]. Let N be a positive integer.

Definition 2.3. A comma-free code X is called N-canonical if for any word $w \in E(X)$ and any 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 $s \notin R(X)A^*$.

Example 2.4. Let $A = \{a, b\}$. The set $\{ababb, babbabb\}$ is a comma-free code, for which $R = \{abb, abbabb\}$, $L = \{a, abab, bab\}$, but not maximal since $a^+b^+ - abb^+ \subseteq E$.

Example 2.5. The set $\{a^3b, a^2b^2, ab^3\}$ is a 4-canonical comma-free code with $R = \{b, b^2\}$, $L = \{a, a^2\}$.

In the previous paper [8], it is proved that every finite comma-free code is included in a finite N-canonical comma-free code, for some N. Our aim now is to prove further 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. Explicitly, which words are good ones? First, it contains no factors in X. Second, it is not a factor of X^2. Third, there are no two words x and y of X, for which $\{x, y, u\}$ is not a comma-free code by the equality $ux = vy$ or $xu = yv$ with $v \in A^*$ and $|v| < |u|$. Fourth, there is no word x of X, for which $\{x, u\}$, is not a comma-free set by the equality $uu = vxw$ for some $v, w \in A^*$ with $|w| < |u|$ and $|v| < |u|$. Fifth, there is no word x of X for which is not a comma-free code by the equality $vuw = ux$ with $0 < |v| < |u|$ or $vuw = xu$ with $0 < |w| < |u|$ for some $v, w \in A^*$. Finally, sixth, it is a primitive word (we denote the set of primitive words by Q). These wordy description obviously corresponds to the following formal conditions:

(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) $u \in Q$.

Let u be an arbitrary word. We call u an e-word if $u \in E(X)$; we call an e-word u *i-word* if, in addition, $u^2 \in E(X)$. It is easy to see that the e-word u is an i-word if $u \in I(X) - F(X)$. Let u be an e-word. Consider the following conditions concerning u:

(r) uv avoids X (*i.e.* u has no factors in X) for every e-word v:

$$uE(X) \cap A^*XA^* = \emptyset;$$

(l) vu avoids X for every e-word v:

$$E(X)u \cap A^*XA^* = \emptyset.$$

We call the words satisfying the conditions (r) and (l) *r-words* and *l-words* respectively.

The good word u is called *r-good* if it is a good word and an r-word at the same time. Similarly, u is *l-good* provided it is also an l-word.

We mention here a detail upon which we shall come later in the proof of Theorem 4.1. Let f' be a word in $A^* - R(X)A^*$ then it is straightforward to see that f' is an r-word iff $f'v$ has no factor in X for every $v \in E(X)$ and has no occurrence of X other than the last one if $v \in X$. Symmetrically, let f'' be a word in $A^* - A^*L(X)$ then f'' is an l-word iff vf'' has no factor in X for every $v \in E(X)$ and has no occurrence of X other than the first one if $v \in X$.

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 integer, the word w is called a *k-sesquipower* if it is a left factor of u^+ for some word u of length less than or equal to k, $|u| \leq k$, or which amounts to the same, $w = u^s u'$ for some left factor u' of u and non-negative integer s and $|u| \leq k$. Obviously, we have an equivalent statement: w is a k-sesquipower if and only if it is a right factor of v^+ or just the same, $w = v'v^t$ for some integer $t \geq 0$ where v 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^+, u is a right factor of y^+ and $|x| = |y|$;*
(iii) *$x = pq, u = (pq)^s p = p(qp)^s, y = qp$ for some words p, q.*

It is straightforward to see 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 a certain $k \geq \max\{|x| : x \in X\}$.

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

Lemma 3.2. *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 $m = n = 2$.*

Proof. We have

$$|\lambda| = \frac{|u|}{m}$$
$$|\mu| = \frac{|u|}{n} + \frac{|v|}{n} \leq \frac{|u|}{n} + \frac{N}{n}\cdot$$

If m or $n \geq 3$ we get

$$|\lambda| + |\mu| \leq \frac{|u|}{3} + \frac{|u|}{2} + \frac{N}{2} \leq \frac{5|u|}{6} + \frac{N}{2} \leq |u|$$

as $|u| \geq 3N$. By Fine and Wilf, $\lambda = \mu$, therefore $v = \lambda^{n-m}$, which implies $|\lambda| \leq |v| \leq N$, so both of u and uv are N-sesquipowers which contradicts the assumption. Hence $m = n = 2$. □

Lemma 3.3. *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 $|\bar{\lambda}| < \frac{|v|}{2}\cdot$*

Proof. Clearly, $|\mu| > |\lambda|$, so we can write

$$\mu = \lambda\lambda_1$$

where $|\lambda_1| = |\mu| - |\lambda| = \frac{|u|+|v|}{2} - \frac{|u|}{2} = \frac{|v|}{2}\cdot$ Let $\bar{\lambda}$ be the primitive root of λ_1, $\lambda_1 = \bar{\lambda}^n$, $n > 0$, then we have

$$\mu = \lambda(\bar{\lambda})^n$$

and

$$|\bar{\lambda}| \leq |\lambda_1| = \frac{|v|}{2}\cdot$$

From the equality

$$uv = (\lambda)^2 v = \lambda\lambda v = \mu^2 = \lambda\lambda_1\lambda\lambda_1$$

it follows $\lambda v = \lambda_1\lambda\lambda_1$. Since $|\lambda_1| = \frac{|v|}{2} < \frac{|u|}{2} = |\lambda|$, we see that λ self-overlaps with (left) border λ_1. Therefore λ is a left factor of $\bar{\lambda}^+$. □

Proposition 3.4. *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.*

Proof. Assume the contrary that all of u, uv_1, uv_1v_2 are not primitive

$$\begin{aligned} u &= \lambda^m, & m \geq 2 \quad (1)\\ uv_1 &= \mu^n, & n \geq 2 \quad (2)\\ uv_1v_2 &= \eta^p, & p \geq 2. \quad (3) \end{aligned}$$

By Lemma 3.2 $m = n = p = 2$. We show that this is impossible.

Apply Lemma 3.3 to (1) and (2), we obtain

$$\mu = \lambda\bar{\lambda}^r, \qquad r > 0, \tag{4}$$

where $\bar{\lambda}$ is primitive, λ is a left factor of $\bar{\lambda}^+$ and $|\bar{\lambda}| \leq \frac{|v_1|}{2} \leq \frac{N}{2}$.

Similarly, apply Lemma 3.3. to (2) and (3), we get that μ is a left factor of $\bar{\mu}^+$ for some primitive word $\bar{\mu}$, with $|\bar{\mu}| \leq \frac{|v_2|}{2} \leq \frac{N}{2}$.

Since λ is a left factor of μ, $\bar{\mu}^+$ and $\bar{\lambda}^+$ have a common left factor λ for which

$$|\lambda| = \frac{|u|}{2} \geq \frac{3N}{2} > N = \frac{N}{2} + \frac{N}{2} \geq |\bar{\mu}| + |\bar{\lambda}|.$$

Therefore, $\bar{\mu} = \bar{\lambda}$ and in view of (4) and the "synchronizing " property of primitive words we get $\mu \in \bar{\lambda}^+$ despite (4) and the primitivity of μ. □

The meaning of the following lemma is that any non-sesquipower can be "pumped" up to more non-sesquipowers.

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, $u = u_1^s u_2$, $s \geq 0$, with u_2 a proper left factor of u_1, u_1 primitive and $|u_1| \leq k$. Then for sufficiently large integers t, namely, for all t such that $|u_1^t u_2| \geq 2k$, the words $u_1^t u_2 v$ are not k-sesquipowers.*

Proof. Note that $u \neq 1$. We show actually that for every t such that $|u_1^t u_2| \geq \min(|u_1^s u_2|, 2k)$ the word $u_1^t u_2 v$ is not a k-sesquipower. Suppose that $u_1^t u_2 v$ is a k-sesquipower then it is a left factor of a power of a certain word of length no longer than k.

If $\min(|u_1^s u_2|, 2k) = |u_1^s u_2|$ then $t \geq s$ and $w = u_1^s u_2 v$ is a factor of $u_1^{t-s} u_1^s u_2 v = u_1^t u_2 v$. Thus w is a factor of a power of some word of length $\leq k$, or just the same, w is a left factor of a power of some word of length $\leq k$. This contradicts the assumption.

Else, if $\min(|u_1^s u_2|, 2k) = 2k$ and, being a k-sesquipower, $u_1^t u_2 v = u_3^n u_4$, where u_3 is primitive word, $|u_3| \leq k$ and u_4 is a proper left factor of u_3. By Fine and Wilf we have $u_1 = u_3$, consequently, $n \geq t$ and $u_2 v = u_3^{n-t} u_4 = u_1^{n-t} u_4$. Therefore $w = u_1^s u_2 v = u_1^s u_1^{n-t} u_4 = u_1^{n-t+s} u_4$ is a left factor of u_1^+, hence is a k-sesquipower, contradiction. The lemma is proved. □

The next lemma 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 $n > 0$.*

Now we start up properly for our task by the next section.

4. Short i-words

Let X be a finite N-canonical comma-free code with $m = \max\{|x| : x \in X\}$. Suppose that h is a primitive i-word for X of length greater than m. We put $K = \max(N, m)$ and $f = h^k$, where $k \geq \frac{3K+3N}{|h|}$. 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 r-word or f'' is an l-word and

$$\frac{|f|}{2} - m < |f'|, |f''| < \frac{|f|}{2} + m.$$

Let $f = f_0'f_0''$ be a factorization such that

$$\frac{|f|}{2} + 1 \geq |f_0'|, |f_0''| \geq \frac{|f|}{2}.$$

Note that $f_0' \in A^* - R(X)A^*$ and $f_0'' \in A^* - A^*L(X)$. The following repeated argument will lead to the desired r-word f' or l-word f''.

Suppose that f_0' is not r-word (otherwise we are done). Then there exists a word $u_1 \in X + E(X)$ such that $f_0'u_1$ contains a factor y_1, not a right one in case $u_1 \in X$, in X.

$$f_0'u_1 = v_1y_1w_1.$$

Since f_0' avoids X and u_1 contains no factor in X when $u_1 \in E(X)$ and no proper factor in X when $u_1 \in X$ we see that y_1 must overlap both f_0' and u_1. This means that f_0' has right factor x_1 which is a non-empty proper left factor of y_1:

$$f_0' = f_1'x_1$$

for $f_1' \in A^*$. At this moment we get the factorization

$$f = f_1'f_1''$$

where $f_1'' = x_1f_0''$. Note that $|f_0'| - |f_1'| = |x_1| < m$, we have

$$\frac{|f|}{2} - m \leq |f_0'| - m < |f_1'| < |f_0'| \leq \frac{|f|}{2} + m.$$

Therefore, as $|f_1'| + |f_1''| = |f|$, for $|f_1''|$ the same inequalities obtain

$$\frac{|f|}{2} - m < |f_1''| < \frac{|f|}{2} + m.$$

Consider now the symmetrical situation. Suppose that f_1'' is not an l-word, there is then some word $u_2 \in X + E(X)$ such that $u_2 f_1''$ contains a factor $y_2 \in X$, not a left one in case $u_2 \in X$. Reasoning as above: since f_1'', being a factor of f, avoids X and u_2 avoids X if $u_2 \in E(X)$ and does properly X if $u_2 \in X$, we again see that y_2 must overlap both u_2 and f_1'' that means that f_1'' has a left factor x_2 which is a non-empty proper right factor of y_2. We specify this situation in more detail by

$$u_2 = v_2 z_2, \quad y_2 = z_2 x_2, \quad f_1'' = x_2 f_2''$$

for $z_2 \in A^+$, $f_2'' \in A^*$. Now it is crucial to notice that, as $f_1'' = x_1 f_0''$ and x_1 is a left factor of X, x_2 must be longer than x_1, otherwise $z_2 \in L(X)$ despite $u_2 \in X + E(X) \subseteq A^* - A^* L(X)$. Hence x_1 is a proper left factor of x_2. Put $f_2' = f_1' x_2$. We have now the factorization

$$f = f_2' f_2''.$$

Further, it is directly to see that x_2 overlaps both f_0' and f_0'' as x_1 is a proper left factor of x_2 and x_1 is a right factor of f_0'. This implies that $|f_0''| - |f_2''| < |x_2| < m$ and, consequently,

$$\frac{|f|}{2} - m \le |f_0''| - m < |f_2''| < \frac{|f|}{2} + m,$$

and

$$\frac{|f|}{2} - m < |f_2'| < \frac{|f|}{2} + m.$$

Now we proceed similarly with the latter factorization and with f_2' playing the role of f_0' in the initial factorization of f to obtain a left factor x_3 of x for which x_2 is a proper right factor and the ensuing factorization $f = f_3' f_3''$. Of course, x_3 overlaps both f_0', f_0'' as x_2 does, from which follow the relevant inequalies for the length of f_3', f_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 $m-1$ ones, to obtain a factorization

$$f = f' f''$$

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

Suppose for instance that f'' is an l-word. Let u be the longest left factor of $f'' f' f'' f'$ which is an m-sesquipower. 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. If we write

$$f''f'f''f' = uy$$

for $y \in A^*$ then $|y| = 2|f| - |u| > 2|f| - |f| - m = |f| - m \geq 3K + 3N - m > 3K + 3N - 2m$. On the other hand $2|f| - |u| > |f| - |u| \geq 3K + 3N - |u|$. Therefore $|y| > \max(3K + 3N - 2m, 3K + 3N - |u|) = 3K + 3N - \min(|u|, 2m)$.

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, in particular, we can write

$$lu_0 = u$$

for $l \in A^*$. Note also that $|u_0 y| > |u_0| + 3K + 3N - \min(|u|, 2m) \geq \min(|u|, 2m) + 3K - \min(|u|, 2m) = 3K$. Now let u_3 be the left factor of $u_0 y$ of length $3K$, that is

$$u_0 y = u_3 v$$

for $v \in A^+$ and $|u_3| = 3K$. We see that u_0 is a proper left factor of u_3 and we write

$$u_0 r = u_3$$

for $r \in A^+$. We have the following relations

$$f''f'f''f' = uy = lu_0 y = lu_3 v = lu_0 rv.$$

Observe that because ur, which is a left factor of $f''f'f''f'$, is not an m-sesquipower, in any case (*i.e.*, if $u = u_0$ or $|u_0| \geq 2m$ by Lem. 3.5) $u_0 r$, that is u_3, is not an m-sesquipower either.

In order to employ the canonicity, we estimate the length of $|v|$. If $u = u_0$, that is, if $l = 1$, then

$$|v| = |f''f'f''f'| - |u_0| > 2|f''f'| - 3m = 2|f| - 3m \geq 2(3K + 3N) - 3m > 3N.$$

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

$$\begin{aligned} |v| &= |f''f'f''f'| - |l| - |u_3| > 2|f| - (|f| - m) - 3K \\ &= |f| + |m| - 3K > 3K + 3N - 3K \\ &= 3N. \end{aligned}$$

Now we use the hypothesis. Since X is N-canonical and $f^2 \in E(X)$ hence $f^3 \in E(X)$, for the factorization

$$f^3 = f'lu_3vf''$$

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 (larger than m) of the latter words.

All of the three words u_3v_1, $u_3v_1v_2$, $u_3v_1v_2v_3$ cannot be m-sesquipowers because u_3 is not so. Moreover, by Lemma 3.4, as $|u_3| = 3K \geq 3N$, one of them, 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 verify the the definition of a good word:

(1) g avoids X because g is a factor of f^3, which does so as $f^3 \in E(X)$;
(2) g is not a factor of X^2: g is too long for that $|g| > 3K > 2m$;
(3) clearly $g \notin A^*L(X)$. Also, if $u = u_0$ then u_3 is a common left factor of g and f''; if $|u_0| \geq 2m$ then u_0 is a common left factor of u and u_3, hence of f'' and g. That means in all cases g and f'' have a common factor of length at least $\min(3K, 2m) = 2m > m$ which implies that $g \in A^* - R(X)A^*$, for $f'' \in A^* - R(X)A^*$ as well, since the two inclusions concern only the left factors of length less than m. More than that, by the same reason and by (1), we have
(3') $E(X)g \cap A^*XA^* = \emptyset$;
(4) in view of (1) and (3) we have $g \in E(X)$ and in view of (3') $g^2 \in E(X)$;
(5) holds because g is not an m-sesquipower;
(6) g is primitive by choice.

So (1)–(6) shows that g is a good word, and in addition to them, (3') shows that g is an l-good word. Certainly,

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

what is the desired estimate. Moreover, as g is a factor of f^3 of length not exceeding $|f|$, it must be a factor of f^2. This concludes the proof. □

In virtue of Theorem 4.1, we have the following dichotomy. First, there are no primitive i-words of length longer than m. This means that they are finite in number. Because good words are primitive i-words, the good words are also finite in number and 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.

The following argument may help. For any two i-words u, v, not factors of X, with different primitive roots if uv and vu avoid X then u^nv^n avoids X for all positive integers n. They all are primitive i-words by [9], moreover, of arbitrarily large length, contradiction with the finitude assumption. So at least of uv, vu contains a factor in X and u, v cannot be both in any completion of X. Therefore, any completion of X differs from X by at most one word (of length no greater than m).

The next possibility: there are infinitely many primitive i-words. In this case there is always a relatively "short" l- or r-good word g, no longer than $3K + 3N$, provided by Theorem 4.1. The proof of theorem provides also an explicit, effective means to determine g.

Nevertheless, how could we know in which branch of the dichotomy we are: all of the primitive i-words are of length not exceeding m or some of them are longer than m? The answer is an instance of the following results by Ito, Katsura, Shyr and Yu [5]:

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 contains 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.*

In view of these propositions, we check, if there is a primitive word p with $n \leq |p| \leq 3n-3$. If yes, R contains an infinity of primitive words; if no R contains only finitely many (may be, none) primitive words and the number of them can be bounded by an effectively computable constant.

It is clear that the sets $R(X)$, $L(X)$ are finite, $E(X)$ is regular, as X is finite. Also, the set of i-words is easily seen to be regular. Our problem is to test the set of i-words minus $A^*A^{m+1}A^*$ for a primitive word in it.

The next section is devoted to the completion of X, starting from an l- or r-good word.

5. Short good words

In view of the discussion in the preceding section, for diversity of treatment, we may now suppose that we dispose of an r-good word g satisfying

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

We recall that "g is an r-good word" implies that gv is free of factors in X for every word v of $E(X)$ beside the *a priori* property of a good word (of an r-word, more exactly) that gv has only one occurrence in X, that of v itself, if $v \in X$. In order to complete X, we follow the steps below:

(a) if for almost all (*i.e.* all but finitely many) primitive i-words v, v contains g as a factor or vg contains a factor in X or an occurrence of g different from the last one (this issue we can effectively test in view of Prop. 4.2, to wit we test

$$\{v : vg \in A^*XA^*\} \cap \{v : vg \in A^*gA^+\} \cap \{v : v \in E(X), v^2 \in E(X)\}$$

for the finitude of primitive words) then the set of good words for $X + g$ is finite (the maximum length is effectively computable by Prop. 4.3) and we are finished. Otherwise

(b) we can effectively pick out a primitive i-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 r-good word for X. Indeed,

1. vg is an r-word, because of the current assumption on g and v;
2. vg is not in $F(X^2)$, as $|vg| > 3|g| > (K > 2m$, 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 an m-sesquipower). Because from any equality for the overlapping
$$xvg = vgy$$

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

(c) put $p = vg$. So p is an r-good word and $|p| > 3|g| > 9K$. It may self-overlap only with the borders longer than $6K$.

If for almost all e-words $w \in E(X)$, either wp contains a factor in X or an occurrence of p different from the last one 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 e-word q, $q \in E(X)$, with $|q| \geq 2|p|$ such that qp does not contain any factor in X and 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

$$(n - 2)|p| > |q| + 6N.$$

Certainly, $n > 2$. 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 the words of the form

$$up^iqp^n$$

satisfying the following conditions:

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

(ii) up (if $i > 0$) and uq (if $i = 0$) are e-words;
(iii) p is not a right or left factor of u;
(iiii) up^iqp^n is primitive.

We have a few further remarks. For every $i = 0, 1 \ldots, n - 1$:

Remark 5.2. All words of G_i are not m-sesquipowers since p is not an m-sesquipower.

Remark 5.3. All words of G_i avoid X and are not factors of X^2. First, by (ii) up and uq (for $i = 0$) avoid X; qp avoids X by definition; pq avoids X since p is r-good and q is an e-word. The next claim is obvious.

Remark 5.4. All words of G_i are i-words. The fact that up (for $i > 0$), uq (for $i = 0$) and p are e-words together with Remark 5.3 yield $G_i \subseteq E(X)$. Next, p is r-good and up is an e-word shows that pup avoid X hence so does pu which implies $G_i \subseteq I(X)$.

Remark 5.5. If up^iqp^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^nqp^n.

Proof. (g) follows from Remarks 5.2, 5.3, 5.4 and (iiii).
(gg) holds because p^nqp^n has only two occurrences of p^n and because of (i) and (iii). □

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

(j) $|v| \geq |q|$;
(jj) vp is an e-word;
(jjj) p is not a right or left factor of v, q is not a right factor of v;
(jjjj) vp^n is primitive.

It is routine to verify that the counterparts of Remarks 5.2–5.4 and Proposition 5.6 are also valid for H. 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 an e-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 e-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 $w' \in A^*, |w_0| = |p|, 0 < |w_j| \leq N$ and

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

is an e-word for $j = 1, \dots, 6$. In view of Proposition 3.4, 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 $w_j \dots w_1w_0p^iqp^n$ for $j = 1, \dots, 6$ are not N-sesquipowers, as $|p| > 9K > N$ and $n > 2$ 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 the 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 e-word, avoids X which also shows that $w_s \dots w_1w_0p$ for $i > 0$ and $w_s \dots w_1w_0q$ for $i = 0$ atre e-words. 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 $6N + n|p| + |q|$.*

The restriction put on the length of p, q and n is to ensure the following property, which will be used to establish the comma-freeness of the would-be completions of X.

Corollary 5.10. *For every word of up^iqp^n of $\bar{G}_i$, $|u| \leq (n-3)|p|$ and for every word vp^n of $\bar{H}$, $|v| \leq (n-2)|p|$. Moreover, the words of $\bar{G}_i$ and $\bar{H}$ have a unique occurrence of p^n.*

Proof. The upper bound for the length of u, v follows directly from the assumption on n and from Theorem 5.9. For the next claim, consider an arbitrary word up^iqp^n of $\bar{G}_i$, $0 \leq i < n$. There is already one occurrence of p^n, the terminal one. Another occurrence, if exists, should be a factor of up (in case $i > 0$) or uq (in case $i = 0$) by Remark 5.5, hence

$$|p^n| \leq \max(|uq|, |up|) = |uq|.$$

On the other hand, by Theorem 5.9

$$|u| \leq 6N + |p|.$$

This implies

$$n|p| \leq |u| + |q| \leq 6N + |p| + |q|.$$

But this contradicts the assumption that

$$(n-2)|p| > |q| + 6N.$$

It is handled similarly for the case of $\bar{H}$, where if some word vp^n of $\bar{H}$ has two occurrences of p^n then by Remark 5.7

$$|p^n| \leq |vp|.$$

This is again a contradiction, since by Proposition 5.9, $|v| \leq 6N + |q|$ despite the assumption

$$(n-2)|p| > |q| + 6N$$

which completes the proof. □

We intend to make up a completion of X, which includes $G_i's$ and H; for this purpose first we fix the following

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 $\bar{H}$.*

Proof. (h) because every word of $\bar{H}$ and $\bar{G}_i$ has a unique (terminal) occurrence of p^n and because of (jjj) q is not a right factor of v.

(hh), (hhh) and (hhhh): analogously handled. □

Put now

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

Note that $\bar{X}$ avoids X and the Proposition 5.11 says nothing but that $X + \bar{X}$ is an infix code. Recall that every word of $\bar{X}$ is a good word for X. How long are the borders of $\bar{X}$? We shall show that they are much longer than m which is helpful in proving the comma-freeness of $X + \bar{X}$.

If in some pair of overlapping words of $\bar{X}$ the overlap is shorter than $|p^n|$ then the resulting borders are of length at least $\min(|u|, |v|, |q|)$ which is indeed larger than m. If, otherwise, the overlap is longer than $|p^n|$, then it must contains an occurrence of p^n, being a right factor of one of the overlapping word, and p^n is a right factor of all words of $\bar{X}$. Hence one of the two overlapping words has at least two occurrences of p^n, so it is qp^{n+1} by Corollary 5.10 and thus the right border is p and the overlap is qp^n. Since q is not a right factor of v, the other of the overlapping words is $up^iqp^n \in \bar{G}_i$, $0 \leq i < n$. Hence the left border is up^i. In any case the borders are at least $\min(|u|, |v|, |q|, |p|)$ long. In sum, the borders are always longer than m.

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

$$x_1x_2 = lx_3r$$

and $0 < |l| < |x_1|, 0 < |r| < |x_2|$.

A little observation first: product of two words of $\bar{X}$ avoids X, since pu, pv (u, v are the words in (i), (j) of the definition of the words of G_i and H, respectively) avoid X (recall that p is an r-good word and up, vp are in $E(X)$). All x_1, x_2, x_3 should be in $\bar{X}$ due to the following reasons: the observation above, p is an r-word, every word of $\bar{X}$ is a good word, the borders of $\bar{X}$ is larger than m and X is comma-free. To wit, suppose that $x_3 \in X$. Since x_3 overlaps both x_1 and x_2 by the observation above, x_1 and x_2 cannot be both in $\bar{X}$. However, both of them are not in X because X is comma-free; if $x_1 \in X, x_2 \in \bar{X}$ or $x_1 \in \bar{X}, x_2 \in X$ then $x_2 \in R(X)A^*$ or $x_1 \in A^*L(X)$ correspondingly, contradictions. Therefore $x_3 \in \bar{X}$. Since x_3 is a good word, x_1 and x_2 cannot be both in X; if only one of them is not in X then the borders are shorter than m, a contradiction. So all three x_1, x_2, x_3 are in $\bar{X}$.

Further, 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, and only in case $x_2 \neq qp^{n+1}$. However this possibility is ruled out by the fact that $n > 2$, p is primitive and every word in $\bar{X}$ has no left factor p but has a right factor p^n and by the following reasons. First, if $x_2 = up^iqp^n \in \bar{G}_i$ then *that* occurrence must be in pup^iq and overlap the initial occurrence of p and, maybe in addition, p^i, or q (if $i = 0$), as q does not contains p and p is primitive, which means that $|u| > (n-2)|p|$, a contradiction by Corollary 5.10. Next, if $x_2 = vp^n \in \bar{H}$ then $|v| > (n-2)|p|$, again a contradiction by Corollary 5.10. 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_1qp^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 x_3 is a (right) factor of p^nqp^n by Proposition 5.6 (gg), again contradiction by Corollary 5.10 and thus the proof is completed. □

We present our ultimate statement, the completion theorem.

Theorem 5.13. *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 5.8(ii), that is, in $\bar{X}$, and we are done for this alternative.

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

$$f^lqp^{n+1}.$$

If it has a factor in X, clearly, f 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 word 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

$$f^l qp^n = tp^i w$$

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

Acknowledgements. I am greatly grateful to the referee for keen and penetrating remarks and comments that help improve the paper.

References

[1] J. Berstel and D. Perrin, *Theory of Codes*. Academic Press, Orlando (1985).

[2] N.J. Fine and H.S. Wilf, Uniqueness Theorem for Periodic Functions. *Proc. Amer. Math. Soc.* **16** (1965) 109-114.

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

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

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

[6] M. Ito, H. Jürgensen, H.J. Shyr and G. Thierrin, Outfix and Infix Codes and Related Classes of Languages. *J. Comput. Syst. Sci.* **43** (1991) 484-508.

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

[8] N.H. Lam, Finite Completion of Comma-Free Codes. Part 1, in *Proc. of DLT 2002*. Springer-Verlag, *Lect. Notes Comput. Sci.* **2450** 357-368.

[9] R.C. Lyndon and M.-P. Shützenberger, The Equation $a^M = b^N c^P$ in a Free Group. *Michigan Math. J.* **9** (1962) 289-298.

[10] Al.A. Markov, An Example of an Independent System of Words Which Cannot Be Included in a Finite Complete System. *Mat. Zametki* **1** (1967) 87-90.

[11] A. Restivo, On Codes Having No Finite Completions. *Discret Math.* **17** (1977) 306-316.

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

[13] J.D. Watson and F.C.H. Crick, A Structure for Deoxyribose Nucleic Acid. *Nature* **171** (1953) 737.

Communicated by J. Berstel.
Received April 22, 2003. Accepted February 10, 2004.