RAIRO-Inf. Theor. Appl. **38** (2004) 91-115
DOI: 10.1051/ita:2004006

FINITE COMPLETION OF COMMA-FREE CODES PART 1

Nguyen Huong Lam[1]

Abstract. This paper is the first step in the solution of the problem of finite completion of comma-free codes. We show that every finite comma-free code is included in a finite comma-free code of particular kind, which we called, for lack of a better term, canonical comma-free code. Certainly, finite maximal comma-free codes are always canonical. The final step of the solution which consists in proving further that every canonical comma-free code is completed to a finite maximal comma-free code, is intended to be published in a forthcoming paper.

Mathematics Subject Classification. 68R15, 68S05.

1. Introduction

In this paper we consider (finite) comma-free codes and their completion problem. We sketch a few lines on their origin, history and development. Comma-free codes were defined rigorously in 1958, as mathematical objects, in Golomb, Gordon and Welch [3], although a rudimentary notion had been suggested by Crick, Griffith and Orgel earlier in 1957 [2], in connection with the famous discovery of DNA structure [14]; for more details on the biological origin of the problem, see [4], or [1], annotation to Chapter VII.

During the late 1950s and the 1960s there had been quite extensive investigation on comma-free codes of constant length (block codes), free from biological considerations. Then the main line of study was concerned with the maximal size of comma-free codes of a given length and on alphabets with a fixed number of letters. The most impressive achievement was a proof by Eastman [5] of a conjecture of Golomb, Gordon and Welch [3] on the maximal number of words in a

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

comma-free code on the alphabet comprising an odd number of letters; see [1], Chapter VII, for alternative proofs.

The research on variable-length comma-free codes was initiated a bit later, in 1969 [12], and it still goes on as of 1998 [6]. In the later paper some language-theoretic aspects of maximal comma-free codes are discussed. We would like to remark that we actually lack a satisfactory, elegant characterization of maximal comma-free codes.

Several problems often encountered in theory of codes are of the following kind: given code belongs to some special family of codes, is it possible to complete it within this family? For instance, it is trivial to see that every finite prefix code can be embedded into a finite maximal prefix code. It also represents little difficulty to prove that every finite infix code can be completed to a finite maximal infix code [7]. The case of general codes is more difficult. It was proved by Markov [10] and subsequently by Restivo [11] that there exist finite codes that are not included in any finite maximal codes. More on the negative side we can mention the case of bifix codes [1] (Chap. III).

In this paper, we manage to prove that every finite comma-free code has a finite completion, or in other words, is included in a finite maximal comma-free code. Thus we give the affirmative answer to the completion problem for another class of codes.

Our solution is complex, so we divide it into two parts. First, we attempt to complete every finite comma-free code to a (finite) comma-free code of a special kind, which we call *canonical* comma-free codes. Canonical comma-free codes are defined by a neat condition which warrants that, if not maximal, they can be made into maximal comma-free codes by adding "not very" long words. This part is the content of the present paper. Next, we prove, by a concrete and explicit construction, that every canonical comma-free code is completed to a finite maximal comma-free code. This final part we hope to publish in another paper.

The present paper is structured as follows: following this introduction, Section 2 introduces standard notation and basic notions. In Section 3, canonical words, the "bricks" with which a completion is built, are introduced and studied. Section 4, the most heavy, describes the construction, a repeated process, which is entitled to produce desired finite completions. Of course, the number of steps to be repeated should be determined by the code at the starting point. This is proved by means of the notions of index and type of right borders. This section prepares the ground for the definition of canonical comma-free codes and the completing procedure itself in the last Section 5.

2. Notions and notation

Our terminology is standard and the notions are few. Let A be a finite alphabet. Then A^* denotes the set of words on A, including the empty word 1, and as usual $A^+ = A^* - \{1\}$. (We often use the plus and minus signs to denote the union and

difference of two sets.) The set A^* is equipped with the concatenation as product. For $w \in A^*$ we denote by $|w|$ the length of the word w. For subsets X, X' of A^*, we denote

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

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 of v is proper if it is not the empty word or the whole word v. A subset of words is an infix code if no word of it is a factor of another, hence an infix code is a prefix, suffix and bifix code, all at a time. For a subset X of A^*, we denote by $P(X)$ and $S(X)$ the set of prefixes and suffixes of X, respectively.

We are in a position to define the principal objects of this paper [13].

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

A comma-free code is *maximal* if it is not included in any other comma-free codes. A completion of a comma-free code is a maximal comma-free code containing it. Every comma-free always has completions, in view of Zorn's lemma. In this paper we deal exclusively with finite comma-free codes and finite completions.

Example 2.2 [8]**.** Let $A^2 = X + Y$ be a partition of the set of words of length 2 and let $k > 0$. Then

$$C = \{a_1 \dots a_k b_1 \dots b_k : a_1 b_1 \in X, a_i b_i \in Y, i > 1\}$$

is a comma-free code.

Comma-free codes are closely connected with the notion of overlapping. We say that the two words u and v *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 the equalities above *overlapping equalities* for a pair of words u, v and we call s a *right border* and t a *left border* of the pair of overlapping words u, v. A right (left) border of a subset X is a right (left, resp.) border of any pair of overlapping words in X. We denote the sets of left borders and right borders of X by $L(X)$ and $R(X)$, respectively.

Throughout this paper, comma-free codes are assumed not to be subsets of the alphabet, just because for those comma-free codes the completion problem is trivial: the alphabet!

3. Precanonical and canonical words

In this section we introduce the concept of precanonical words and canonical words and study their properties in detail. One of the important conclusions is that the existence of canonical words means that the comma-free code under consideration is not maximal.

3.1. Precanonical words

Let X be a finite comma-free code, t be a positive integer.

Definition 3.1.1. The word w is said to be a left X-precanonical word of tag t, if every right factor of w of length longer or equal to t has a left factor in $R(X)$.

We also define the symmetrical version.

Definition 3.1.2. The word w is said to be a right X-precanonical word of tag t, if every left factor of w longer equal to t has a right factor in $L(X)$.

Clearly, a left X-precanonical word w of tag t has the form

$$w = us, \quad u \in A^*, \quad s \in A^+, \quad |s| = t$$

and for every right factor u', including 1, of u

$$u's \in R(X)A^*.$$

Similarly, a right X-precanonical word w of tag t has the form

$$w = su, \quad u \in A^*, \quad s \in A^+, \quad |s| = t$$

and for every left factor u', including 1, of u

$$su' \in A^*L(X).$$

We often use the concept of precanonical word under the following variation. Let m be a non-negative integer.

Definition 3.1.3. The word w is said to be a left X, m-precanonical word if and only if it is a left X-precanonical word of tag $|w| - m$.

Definition 3.1.4. The word w is said to be a right X, m-precanonical word if and only if it is a right X-precanonical word of tag $|w| - m$.

Clearly, if w is a left X, m-precanonical word then we can write w in the form

$$w = us, \quad s \in A^+, \quad u \in A^*, \quad |u| = m$$

and for every right factor u' of u

$$u's \in R(X)A^*.$$

Similarly, if w is a right X, m-precanonical word then we can write w in the form

$$w = su, \quad s \in A^+, \quad u \in A^*, \quad |u| = m$$

and for every left factor u' of u

$$su' \in A^*L(X).$$

Here we make no emphasis on the tag.

It follows immediately from the definition that the concept of precanonical word bears a dual (symmetric) character. A left precanonical word of a comma-free code is a right precanonical word for the mirror image of this code. Therefore it is natural to see that almost all assertions in this paper are valid, and formulated, for both left and right precanonical words, but in their proofs, we are silent about the treatment for one or another case. The reason for this is evident: the proofs for both parts are just identical, only in another way around.

Definition 3.1.5. A left (right) X, m-precanonical word is called minimal left (right, resp.) X, m-precanonical, $m \geq 0$, if it has no proper left (right, resp.) factor which is a left (right, resp.) X, m-precanonical word.

The set of X, m-precanonical words is not finite, for any fixed m, but the set of minimal ones is finite because of the following proposition.

Proposition 3.1.6. *If $w = us$ ($w = pu$), $|u| = m$, is a minimal left (right, resp.) X, m-precanonical word then there is a right (left, resp.) factor u' of u such that $u's \in R(X)$ ($pu' \in L(X)$, resp.). As a consequence, $|s| < \max X$ ($|p| < \max X$, resp.).*

Proof. Suppose that there is no such right factor. Then for every factorization $u = u_2u_1$ let $r(u_1)$ be the longest word of $R(X)$ which is a left factor of u_1s. By assumption $|u_1s| > |r(u_1)|$, hence

$$|us| = |u_2u_1s| > |u_2r(u_1)|.$$

Let $u = \bar{u}_2\bar{u}_1$ be a factorization for which $|\bar{u}_2r(\bar{u}_1)|$ attains the maximum. Note that $|\bar{u}_2r(\bar{u}_1)| > |u|$, since when $u_1 = 1$, we get at least $|ur(1)| > |u|$. Now it is straightforward to see that $\bar{u}_2r(\bar{u}_1)$ is also a left X, m-precanonical word, but it is a proper left factor of u. A contradiction. The proposition is proved. □

The following observation follows directly from the definition.

Remark 3.1.7. Every right (left) factor of length longer or equal to t of a left (right, resp.) X-precanonical word of tag t is also a left (right, resp.) X-precanonical word of tag t. More concretely, every right (left) factor of length $t+k$ of a left (right, resp.) X-precanonical word of tag t is a left (right, resp.) X,k-precanonical word of tag t.

Every left (right) factor of length longer or equal to t of a left (right, resp.) X-precanonical word of tag t is also a left (right, resp.) X-precanonical word of tag t. (It is vacuously true but useless, by definition, that *any* word of length less than t is left (and right) X-precanonical of tag t!)

Every left (right, resp.) X-precanonical word of tag t is left (right, resp.) X-precanonical word of every tag t', $t' \geq t$, consequently, every left (right) X,m-precanonical word is left (right, resp.) X,m'-precanonical word for every integer m', $0 \leq m' \leq m$.

In what follows when the reference to the code X is clear, or is irrelevant, from the context, we often omit it saying simply "X-precanonical", "m-precanonical", etc. or just "precanonical". The following remark describes the situation we shall encounter in a later proof.

Remark 3.1.8. Let us be a minimal left $X,|u|$-precanonical word. There exists then a right factor u_1 of u such that $u_1 s \in R(X)$; we can assume that u_1 is the shortest right factor satisfying this property. If $u_1 \neq 1$, then for every proper right factor u_2 (including 1) of u_1, on one hand $u_2 s \notin R(X)$ and on the other hand $u_2 s \in R(X)A^*$ as $u_2 s$ is a left $X,|u_2|$-precanonical word. Therefore $u_2 s \in R(X)A^+$ which shows that $|s| > 1$ as $R(X)$ does not contain 1 and we can choose $u_2 = 1$; then we can write $s = s_1 a$ for $s_1 \in A^+$ and $a \in A$. Consequently, $u_2 s_1 \in R(X)A^*$ for all proper right factors u_2, including 1, of u_1, which in turn means that $u_2 s_1$ *is still a left* $X,|u_2|$*-precanonical word, for every proper right factor* u_2 *of* u_1*, including* 1.

We present some technical statements which will be in use later.

Proposition 3.1.9. *Let X be a finite comma-free code, Y be a subcode of X and let $a_1 \dots a_k \dots a_{k+t}$ ($a_{k+t} \dots a_k \dots a_1$) be a left (right, resp.) Y,k-precanonical word (of tag t) for a non-negative integer k, a positive integer t and letters $a_1, \dots, a_k$, $a_{k+1} \dots, a_{k+t}$. Then for any integer d such that $d > 0$ and $d \leq k+1$ there exists an integer i such that $i \leq k+t$ and $d \leq i < d + \max Y - 1$ for which $a_1 \dots a_i$ ($a_i \dots a_1$, resp.) has no right (left, resp.) factor in $L(X)$ ($R(X)$, resp.) of length less than or equal to d.*

Proof. We suppose on the contrary that for every i, $i \leq k+1$ and $d \leq i < d + \max Y - 1$, the word $a_1 \dots a_i$ has a right factor in $L(X)$ of length less than or equal to d.

Set $i(0) = d$. We have

$$a_{i(1)} \dots a_{i(0)} \in L(X)$$

for some $i(1)$, $1 \leq i(1) \leq i(0)$ and $1 + i(1) - i(0) \leq d$. Since $i(1) \leq d \leq k+1$ and

$$|a_{i(1)} \dots a_k a_{k+1} \dots a_{k+t}| = t + k + 1 - i(1) \geq t,$$

by definition, $a_{i(1)} \dots a_k a_{k+1} \dots a_{k+t}$ has a left factor in $R(Y)$, that is

$$a_{i(1)} \dots a_{i(2)} \in R(Y)$$

for some $i(2)$, $i(1) \leq i(2) \leq k+t$. We must have

$$i(0) < i(2)$$

in virtue of the comma-freeness of X. Note that

$$\begin{aligned} i(2) &= i(1) + i(2) - i(1) \\ &\leq i(0) + \max Y - 2 \\ &= d + \max Y - 2 < d + \max Y - 1. \end{aligned}$$

Now we repeat the argument with $i(2)$ playing the role of $i(0)$ and so on; the assumption by contradiction allows us to obtain the chain

$$1 \leq \dots < i(3) < i(1) \leq i(0) < i(2) < \dots \leq k+t$$

of integers. But this is impossible, as k and t are fixed. □

Corollary 3.1.10. *Let Y, Z be subcodes of a finite comma-free code X. If, for some integers $k > 0$, $t > 0$, the word $a_1 \dots a_k a_{k+1} \dots a_{k+t}$ is simultaneously a left Y, k-precanonical word of tag t and a right Z, t-precanonical word (of tag k) with letters $a_1, \dots, a_k, a_{k+1} \dots, a_{k+t}$ then $k < \max Z - 1$.*

Proof. Suppose that $k \geq \max Z - 1$. By the previous proposition (with $d = k$) there is an integer i such that $k + t \geq i \geq k$ and

$$a_1 \dots a_i$$

has no right factor in $L(X)$ of length less than or equal to k. In particular, it has no right factor in $L(Z)$, as $L(Z) \subseteq L(X)$ and $\max L(Z) \leq \max Z - 1 \leq k$. However, as $k \leq i \leq k+t$, and being a right Z, t-precanonical word, $a_1 \dots a_i$ should have a right factor in $L(Z)$: contradiction! □

3.2. Canonical words

Let X be a finite comma-free code and m be an integer greater than or equal to $\max X$: $m \geq \max X$. Consider the following set of words of the form aus satisfying

the following conditions:

1. a is a letter: $a \in A$, u is a word of length m: $u \in A^*, |u| = m$, s is a nonempty word: $s \in A^+$;
2. us is a left X, m-precanonical word;
3. aus has no left factor in $R(X)$: $aus \notin R(X)A^*$;
4. aus has no factor in X: $aus \notin A^*XA^*$.

These words are the main objects of this subsection.

Definition 3.2.1. A word of the form aus, $a \in A$, $u \in A^*$, $s \in A$, is said to be left X, m-canonical word if it satisfies the conditions 1, 2, 3 and 4 above.

Similarly, we have the symmetric version:

Definition 3.2.2. A word of the form pvb is said to be right X, m-canonical word if it satisfies the following conditions:

1. $b \in A$, $|v| = m$, $p \in A^+$;
2. pv is a right X, m-precanonical word;
3. $pvb \notin A^*L(X)$;
4. $pvb \notin A^*XA^*$.

One direct consequence of the definitions is that all left, as well as all right, X, m-canonical words are not factors of X^2. Indeed, say $yausz = x_1x_2$ for a left X, m-canonical word aus and $y, z \in A^*$, $x_1, x_2 \in X$. First, we notice that

$$|x_2| \leq m < m + |sz| = |usz|.$$

Next, as $|x_1| \leq m$, we have

$$|x_2| = |yausz| - |x_1| = |y| + m + 1 + |sz| - |x_1| > m + |sz| - m = |sz|.$$

Two equalities above show that x_2 is a right factor of usz of length greater than $|sz|$, which implies that x_2 has a left factor in $R(X)$, but this is impossible because of the comma-freeness of X.

Definition 3.2.3. A left (right) X, m-canonical word is called minimal left (right, resp.) X, m-canonical if it does not contain properly any left (right, resp.) X, m-canonical words as left (right, resp.) factors.

We denote the sets of minimal left and minimal right X, m-canonical words by $T(X, m)$ and $U(X, m)$, respectively. We give the characterization of the minimal canonical words, similar to the case of minimal precanonical words.

Proposition 3.2.4. *If aus (pvb) is a minimal left (right, resp.) X, m-canonical word then us (pv, resp.) is a minimal left (right, resp.) X, m-precanonical word. Consequently, $|s| < \max X \leq m$ ($|p| < \max X \leq m$, resp.) and $T(X, m)$ ($U(X, m)$, resp.) is finite.*

Proof. Let aus be a minimal X, m-canonical word. If us is not minimal, then for some proper left factor s_1 of s, us_1 is a left X, m-precanonical word. Hence aus_1

is also a left X, m-canonical word and aus is no more minimal left X, m-canonical word, contradiction. Therefore us is minimal precanonical word. The last claim follows from the Proposition 3.1.6. □

Given a finite comma-free code X, when left (right, resp.) X, m-canonical words exist? For their existence, first of all, there must exist left (right, resp) X, m-precanonical words which form a subset of the following set. Let $E(X)$ denote the set of all words that avoid X (*i.e.* they have no factors in X) and that have no left factors in $R(X)$ and no right factors in $L(X)$. In symbols

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

It is not difficult to see that $E(X)$ contains all words which are both left and right X, m-canonical words for all $m \geq 0$. We have the following criterion for non-emptiness of T and U.

Theorem 3.2.5. *The set $T(X, m)$ ($U(X, m)$) of minimal left (right, resp.) X, m-canonical words is not empty if and only if $E(X)$ contains a left (right, resp.) X, m-precanonical factor. Specifically, if a word of $E(X)$ contains a left (right, resp.) X, m-precanonical factor, it contains a word in $T(X, m)$ ($U(X, m)$, resp.).*

Proof. The "if" part is evident by the remark preceding the theorem. For the opposite direction, the "only if" part, let suppose that $E(X)$ contains a word e which has a left X, m-precanonical factor us, *i.e.*

$$e = zusy$$

for some $z, y \in A^*$. First, $z \in A^+$ because $e \notin R(X)A^*$. Further, we can assume that z is chosen as of minimal length among all possible factorizations like one above, *i.e.* us is chosen as the left-most occurrence of any left X, m-precanonical factors. This means that if we write $z = z'a$, for $a \in A$ and $z' \in A^*$, we have $aus \notin R(X)A^*$, which shows that aus is a left X, m-canonical which, indeed, has a minimal left X, m-canonical left factor. This proves the "only if" direction and the theorem follows. □

Now we reveal one of the main purposes which the concept of canonical word is introduced for.

Theorem 3.2.6. *The sets $X + T(X, m)$ and $X + U(X, m)$ both are comma-free codes for all $m \geq \max X$.*

Proof. We prove the assertion for $X + T(X, m)$, the other case is handled in the same way.

If $X+T(X, m)$ is not a comma-free code then we have seven cases of "incidence" to consider all of which we shall show lead to contradictions.

1. $x_1y = lx_2r$ for some $x_1, x_2 \in X$, $y \in T(X, m)$ and $l, r \in A^+$. We write $y = aus$ for $a \in A$, $u \in A^+, |u| = m$, $s \in A^+$.

If $|l| < |x_1|$ then $y \in R(X)A^*$ which is against the definition of $T(X,m)$. If, otherwise, $|x_1| \leq l$ then y contains a factor x_2 in X which is against also the definition of $T(X,m)$.

2. $yx_1 = ausx_1 = lx_2r$, where $x_1, x_2 \in X, y \in T(X,m)$; l, r, a, u, s have the same meaning as in the previous case.

If $0 < |r| < |x_2r| \leq |x_1|$ then x_2 is a (proper) factor of x_1 which is impossible because of the comma-freeness of X. If $0 < |r| < |x_1| < |x_2r| \leq |sx_1|$ then $s \in A^*L(X)$; at the same time, because of the minimality of the precanonical word us, we have $s \in S(X)$: these two facts are in a contradiction with the comma-freeness of X. Next, the case $|r| < |x_1| < |sx_1| \leq |x_2r|$ $(\leq |usx_1|)$ implies, as us is left precanonical, that $x_2 \in R(X)A^*$, which is impossible by comma-freeness of X. Finally, the case $|r| \geq |x_1|$ shows that x_2 is a factor of y violating the definition of $T(X,m)$.

3. $x_1x_2 = lyr$ for some $x_2, x_2 \in X$, $y \in T(X,m)$, $l, r \in A^+$. We see at one that this case is impossible because all words of $T(X,m)$ are not factors of X^2.

4. $y_1x = ly_2r$, where $y_1, y_2 \in T(X,m)$, $x \in X$, $l, r \in A^+$. We write $y_1 = a_1u_1s_1$, $y_2 = a_2u_2s_2$ with the usual meaning for $a_1, a_2, u_1, u_2, s_1, s_2$. So we have $a_1u_1s_1x = la_2u_2s_2r$.

First, note that $1 = |a_1| \leq |l|$. If $|l| \leq |a_1u_1|$ then $a_2u_2s_2 \in R(X)A^*$ because u_1s_1 is left X-precanonical: a contradiction with the comma-freeness of X. If $|l| > |a_1u_1|$ then $|s_2r| < |x| < |u_2s_2r|$ (note that $|u_2| = m > |s_1|$) and $x \in R(X)A^*$, because u_2s_2 is left X-precanonical, which is again a contradiction with the comma-freeness of X.

5. $xy_1 = ly_2r = xa_1u_1s_1 = la_2u_2s_2$, where $x \in X$, $y_1, y_2 \in T(X,m)$ and $a_1, a_2, u_1, u_2, s_1, s_2$, l, r have the usual meaning as in the preceding case.

The case $|l| < |x|$ $(< |a_2u_2|$, certainly, as $|x| \leq m = |u_2|)$ implies that $a_1u_1s_1 \in R(X)A^*$ as u_2s_2 is left precanonical. This is impossible since $y_1 \in T(X,m)$. The case $|x| = |l|$ implies that y_1 is not minimal left canonical word as $r \in A^+$, a contradiction. The case $|x| < |l| \leq |xa_1u_1|$ implies $y_2 \in R(X)A^*$ as u_1s_1 is left precanonical, which is also impossible since $y_2 \in T(X,m)$. Finally, $|l| > |xa_1u_1|$ is plainly impossible because $|s_1| < m = |u_2|$.

6. $y_1y_2 = lxr = a_1u_1s_1a_2u_2s_2$, where $y_1, y_2 \in T(X,m)$, $x \in X$ and a_1, a_2, u_1, u_2, s_1, s_2, l, r have the usual meaning as above.

The case $1 = |a_1| \leq |l| \leq |a_1u_1|$ shows that $x \in R(X)A^*$, which violates the comma-freeness of X. The case $|a_1u_1| < |l| < |a_1u_1s_1|$ shows that $y_2 \in R(X)A^*$, as $s_1 \in S(X), x \in X$, this is against the definition of $y_2 \in T(X,m)$. The case $|a_1u_1s_1| \leq |l|$ shows that x is a factor of y_2: impossible. Finally, the last alternative:

7. $y_1y_2 = ly_3r$ for $y_1, y_2, y_3 \in T(X,m)$ and $l, r \in A^+$. We have $y_1 = a_1u_1s_1$, $y_2 = a_2u_2s_2$ and $y_3 = a_3u_3s_3$ with a_1, a_2, a_3, u_1, u_2, u_3, s_1, s_2, s_3 having the same meaning as above.

First, note that $1 = |a_1| \leq |l|$. We have to consider the following cases that are all impossible, the reasons for this we indicate in parentheses:

- $|l| \leq |a_1u_1|$ $(y_3 \notin R(X)A^*)$;
- $|a_1u_1| < |l| < |a_1u_1s_1|$ $(y_2 \notin R(X)A^*)$;

- $|l| = |a_1u_1s_1|$ (y_2 is minimal left canonical word);
- $|a_1u_1s_1| < |l| \leq |a_1u_1s_1a_2u_2|$ ($y_3 \notin R(X)A^*$);
- $|a_1u_1s_1a_2u_2| < |l|$ ($|s_2| < m$).

The proof is completed. □

We state a basic property of $T(X,m)$ and $U(X,m)$, which shows that, when applied the constructions T and U bring no new left and right precanonical words, resp. (They could apparently only reduce the sets of canonical words.)

Proposition 3.2.7. *Let $Y = X + T(X,m)$ ($Y = X + U(X,m)$). Then every right (left, resp.) border in $R(Y)$ ($L(Y)$, resp.) has a (non-empty) left (right, resp.) factor which is a right (left, resp.) border in $R(X)$ ($L(X)$, resp.).*

Proof. Let $r \in R(Y)$ be a right border of the pair of overlapping words y_1, y_2 of Y:

$$ly_2 = y_1r.$$

We consider all cases that can happen.

1. $y_1 \in X$, $y_2 \in X$. Certainly, $r \in R(X)$, we have nothing to prove.

2. $y_1 \in X$, $y_2 \in T(X,m)$. We write $y_2 = aus$, $a \in A$, $u \in A^+, |u| = m$, $s \in A^+$. Since $|u| = m \geq |y_1|$, we have $|la| \leq |y_1| < |lau|$ which implies that $|us| > |r| > |s|$ and $r \in R(X)A^*$ as y_2 is left X-canonical.

3. $y_1 \in T(X,m)$, $y_2 \in X$. We have then $ly_2 = ausr$ with $y_1 = aus$ and a, u, s, l having the usual meaning. Since $|l| \geq 1$ and $|r| < |y_2|$ we get $|r| < |y_2| \leq |usr|$. If $|y_2| \leq |sr|$ then $r \in R(X)A^*$ as $s \in S(X)$; otherwise $|y_2| > |sr|$ then we get also $r \in R(X)A^*$ because us is left X-precanonical. Finally

4. $y_1, y_2 \in T(X,m)$, so $y_1 = a_1u_1s_1$, $y_2 = a_2u_2s_2$, where $a_1, a_2 \in A$, $u_1, u_2 \in A^+$, $|u_1| = |u_2| = m$, $s_1, s_2 \in A^+$. We have

$$la_2u_2s_2 = a_1u_1s_1r$$

with $|l| < |y_1|$. First, $|a_1u_1| < |l|$, otherwise $a_2u_2s_2 \in R(X)A^*$: a contradiction: y_2 is a left canonical word. Hence, we get $|s_2| < |r| \leq |u_2s_2|$, as $|l| < |y_1|$ and $|u| = m > |s_1|$, which implies $r \in R(X)A^*$ because u_2s_2 is left precanonical. The theorem is proved. □

We derive an immediate consequence.

Corollary 3.2.8. *Let $Y = X + T(X,m)$ ($Y = X + U(X,m)$). Then every left (right, resp.) Y,m-precanonical word is also a left (right, resp.) X,m-precanonical word.*

Proof. Straightforward by definition. Observe that $R(X) \subseteq R(Y)$ and $A^* - R(Y) \subseteq A^* - R(X)A^*$. □

The following theorem, which is also an easy corollary of Proposition 3.2.7, describes the effect of elimination of "long" precanonical factors under T and U.

Theorem 3.2.9. *Let $Y = X + T(X, m)$ ($Y = X + U(X, m)$). Then $E(Y)$ does not contain left (right, resp.) Y, m-precanonical factors.*

Proof. Suppose that a certain word e of $E(Y)$ contains a left Y, m-precanonical factor. By the preceding corollary this factor is also a left X, m-precanonical word. Since $e \in E(Y) \subseteq E(X)$, by Theorem 3.2.5, e contains a minimal left X, m-canonical factor, *i.e.* a factor in $T(X, m) \subseteq Y$. But this is a contradiction, as $E(Y)$ avoids Y, which proves the theorem. □

The next section sets a departure point for the completing process, in which the constructions T and U are repeated alternatively to eliminate long precanonical factors, both left and right, in the sets E. The main objective is to show that we can succeed in doing this in a finite number of steps that only depends on the original (finite) comma-free code X.

4. Constructions T and U and their right borders

The present section contains rather an extensive treatment, so we divide it into several subsections for easier presentation. We begin with the main construction which is an iterated process.

4.1. Constructions T and U

Let X_0 be a finite comma-free code, k and λ positive integers, λ is sufficiently large. We define the sequence of finite comma-free codes

$$X_0, Y_1, \ldots, X_{2k}, Y_{2k+1}$$

as follows.

$Y_1 = X_0 + U(X_0, \lambda n_0)$
$X_2 = Y_1 + T(Y_1, n_1)$
. . . . . .
$X_{2k} = Y_{2k-1} + T(Y_{2k-1}, n_{2k-1})$
$Y_{2k+1} = X_{2k} + U(X_{2k}, \lambda n_{2k})$

where $n_0 = \max X_0$, $n_2 = \max X_2, \ldots$, $n_{2k} = \max X_{2k}$ and $n_1 = (\lambda + 1)n_0$, $n_3 = (\lambda+1)n_0 + (\lambda+1)n_2 + n_1, \ldots, n_{2k+1} = (\lambda+1)n_0 + \ldots + (\lambda+1)n_{2k} + n_{2k-1}$.

We get then the ascending chain

$$X_0 \subseteq Y_1 \subseteq \cdots \subseteq X_{2k} \subseteq Y_{2k+1}.$$

We are going to establish some inequalities relating the n_i's and the maximal length of Y_i's. Note that

$$\max Y_{2i+1} \leq (\lambda + 1)n_{2i}$$

for $i = 0, \ldots k$ by the relations $Y_{2i+1} = X_{2i} + U(X_{2i}, \lambda n_{2i})$ and $n_{2i} = \max X_{2i}$.

Further,

$$n_{2i-1} \leq n_{2i} \leq \max Y_{2i-1} + n_{2i-1}$$

for $i = 1, \ldots, k$ by the relations $X_{2i} = Y_{2i-1} + T(Y_{2i-1}, n_{2i-1})$. Hence

$$n_{2i-1} \leq n_{2i} < 2n_{2i-1}$$

for $i = 1, \ldots, k$. On the other hand, by definition, we have

$$(\lambda + 1)n_{2i-2} \leq n_{2i-1}$$

for $i = 1, \ldots, k$. We highlight the following, and what is more important in the sequel,

$$(\lambda + 1)n_{2i-2} \geq \max U(X_{2i-2}, \lambda n_{2i-2})$$

for every $i = 1, \ldots, k$ and, consequently,

$$n_{2i} \geq n_{2i-1} > \max U(X_{2i-2}, \lambda n_{2i-2}).$$

We should remark that λ is arbitrarily large but fixed integer throughout the construction. The question how large λ is depends on the later applications and we shall specify it there.

Setting for short
$U_0 = U(X_0, \lambda n_0)$
$T_1 = T(Y_1, n_1)$
. . . . . .
$T_{2k-1} = T(Y_{2k-1}, n_{2k-1})$
$U_{2k} = U(X_{2k}, \lambda n_{2k})$,
we have

$$Y_{2i+1} = U_{2i} + T_{2i-1} + \ldots + T_1 + U_0 + X_0$$

for $i = 0, 1, \ldots, k$ and

$$T_{2i-1} \subseteq X_{2i} = T_{2i-1} + U_{2i-2} + \ldots + T_1 + U_0 + X_0$$

for $i = 1, 2, \ldots k$. Remark that $X_0, U_0, T_1, \ldots, T_{2k-1}, U_{2k}$ are all disjoint.

4.2. Right borders and simple right borders: Index and type

In this subsection we investigate right borders in detail, more exactly, simple right borders, which are defined as follows. We retain the notation of the previous subsection.

Definition 4.2.1. A right border s of Y_{2k+1} is called simple right border if it does not contain any other right border of Y_{2k+1} as a proper left factor, *i.e.,* $s \in R(Y_{2k+1})$ but $s \notin R(Y_{2k+1})A^+$.

We distinguish the right borders by means of their index, defined as follows.

Definition 4.2.2. Index of a right border s is the least integer i, such that s is a right border of the overlapping equality $y_2 s = l y_1$, where $|l| < |y_2|$, $y_1, y_2 \in Y_{2k+1}$ and $y_1 \in T_i$ or $y_1 \in U_i$ respective to i odd or even, or -1 if $y_1 \in X_0$.

The following technical feature of simple right borders will only be used later in estimating their length.

Proposition 4.2.3. *Let s be a simple right border of index i. If s is a right factor of a word $t = a_j u_j s_j \in T_j$ with $j \leq i$ then s is a proper right factor of s_j.*

Proof. Clearly $s \neq t$, so s is right factor of $u_j s_j$. Suppose that s_j is a right factor of s. Since $u_j s_j$ is a left Y_j, n_j-precanonical word, $|u_j| = n_j$, we see that s has then a left factor $s' \in R(Y_j) \subseteq R(Y_{2k+1})$. By simplicity of s, $s' = s$, therefore, $s \in R(Y_j)$ which shows that the index of s is at most $j-1$ as $Y_j = U_{j-1}+T_{j-2}+\ldots+T_0+U_0+X_0$, which is against the assumption. This contradiction proves the proposition. □

We comes now to the main task of this section, to define the concept of type. Our purpose is to assign, in any way, not necessarily unique, to each simple right border s of Y_{2k+1} an integral value $t(s)$ such that

$$-1 \leq t(s) \leq k$$

called *type* of s; then we shall show, and what is the main service of the concept, that

1. if $t(s) = -1$ then $s \in S(X_0)$, in particular, $|s| < \max X_0$;
2. if $t(s) \geq 0$ then $s \in S(U_{2t(s)})$ and $|s| > (\lambda - 2k + 1) n_{2t(s)}$.

To begin with, suppose that s has index j. We have then an equality

$$y_2 s = l y_1$$

with $y_2 \in Y_{2k+1}$, $y_1 \in T_j$ or $y_1 \in U_j$ and $|l| < |y_2|$ (or equivalently $|s| < |y_1|$).

If $j = -1$, that is, $y_1 \in X_0$, we define at once $t(s) = -1$. Clearly, in this case $s \in S(X_0)$. Otherwise, $j \geq 0$, we distinguish the following cases:

1. j even, $j = 2j(1)$, $0 \leq j(1) \leq k$ and $y_1 \in U_{2j(1)}$. We define $t(s) = j(1)$ and we write by definition

$$y_1 = p_{2j(1)} v_{2j(1)} a_{2j(1)}$$

where $a_{2j(1)} \in A$, $v_{2j(1)} \in A^+, |v_{2j(1)}| = \lambda n_{2j(1)}$, $p_{2j(1)} \in A^+$ and $|p_{2j(1)}| < n_{2j(1)}$. Clearly $|s| < |y_1| \leq \max U_{2j(1)} = \max U_{2t(s)}$. We have two issues concerning s.

(a) $|s| \geq |v_{2j(1)} a_{2j(1)}|$. In this case $|s| > \lambda n_{2j(1)} = \lambda n_{2t(s)}$;

(b) $0 < |s| < |v_{2j(1)} a_{2j(1)}|$. We can write then

$$v_{2j(1)} a_{2j(1)} = \bar{p} s$$

for some word $\bar{p} \in A^+$, which is a right factor of $y_2 \in Y_{2k+1}$. Since $v_{2j(1)}$ is a right $X_{2j(1)}$-precanonical word of tag $|p_{2j(1)}|$ which is less than $n_{2j(1)}$ (as a right factor of $p_{2j(1)} v_{2j(1)}$, see Rem. 3.1.7), we should have $|\bar{p}| < n_{2j(1)} - 1$, otherwise,

$\bar{p}$ has a right factor in $L(X_{2j(1)})$, which contradicts the comma-freeness of Y_{2k+1}. Therefore

$$|s| = |v_{2j(1)}a_{2j(1)}| - |\bar{p}| > \lambda n_{2j(1)} + 1 - (n_{2j(1)} - 1) = (\lambda - 1)n_{2j(1)} + 2.$$

Hence

$$|s| > (\lambda - 1)n_{2t(s)};$$

2. j odd, $j = 2j(1) + 1$, $j(1) < k$ and $y_1 \in T_{2j(1)+1}$. We write

$$y_1 = a_{2j(1)+1}u_{2j(1)+1}s_{2j(1)+1}$$

where $a_{2j(1)+1} \in A$, $u_{2j(1)+1} \in A^+$, $|u_{2j(1)+1}| = n_{2j(1)+1}$, $s_{2j(1)+1} \in A^+$, $|s_{2j(1)+1}| < n_{2j(1)+1}$. In view of Proposition 4.2.3 $|s| < |s_{2j(1)+1}|$, hence

$$s_{2j(1)+1} = \bar{s}'_1 s$$

with $\bar{s}'_1 \in S(Y_{2k+1})$. As y_1 is a minimal right $Y_{2j(1)+1}$-canonical word, there exists a right factor $\bar{p}'_2$ of $u_{2j(1)+1}$ such that

$$\bar{p}'_2 s_{2j(1)+1} \in R(Y_{2j(1)+1}).$$

We choose $\bar{p}'_2$ as the shortest possible right factor satisfying this property. We have

$$y_4\bar{p}'_2 s_{2j(1)+1} = l_3 y_3$$

with some $y_3, y_4 \in Y_{2j(1)+1}$, $|\bar{p}'_2 s_{2j(1)+1}| < |y_3|$.

Further, if again $y_3 \in T_{2j(3)+1}$ for some $j(3) < j(1)$ then we write

$$y_3 = a_{2j(3)+1}u_{2j(3)+1}s_{2j(3)+1}$$

with $a_{2j(3)+1} \in A$, $u_{2j(3)+1} \in A^+, |u_{2j(3)+1}| = n_{2j(3)+1}$, $s_{2j(3)+1} \in A^+$ and $|s_{2j(3)+1}| < n_{2j(3)+1}$ for which

$$y_4\bar{p}'_2 s_{2j(1)+1} = l_3 a_{2j(3)+1}u_{2j(3)+1}s_{2j(3)+1}.$$

Note that we have again by Proposition 4.2.3 that s is a proper right factor of $s_{2j(3)+1}$. Now, depending on the following alternatives

(i) $|s| < |s_{2j(3)+1}| \leq |s_{2j(1)+1}|$;

(ii) $|s_{2j(1)+1}| < |s_{2j(3)+1}| \leq |\bar{p}'_2 s_{2j(1)+1}|$;

(iii) $|\bar{p}'_2 s_{2j(1)+1}| < |s_{2j(3)+1}|$;

we have respectively

(i') $s_{2j(3)+1} = \bar{s}_1 s$ with $\bar{s}_1$ a right factor of $\bar{s}'_1$;

(ii') $s_{2j(3)+1} = \bar{p}_2\bar{s}'_1 s$ with $\bar{p}_2$ a right factor of $\bar{p}'_2$. Note that then $\bar{p}_2\bar{s}'_1 s = \bar{p}_2 s_{2j(1)+1}$ is a left $Y_{2j(1)+1}, |\bar{p}_2|$-precanonical and, because of the minimality of

$\bar{p}'_2$, it is not a minimal left $Y_{2j(1)+1}, |\bar{p}_2|$-precanonical word if $|\bar{p}_2| < |\bar{p}'_2|$ (see Rem. 3.1.8);

(iii') $s_{2j(3)+1} = \bar{s}_3\bar{p}'_2 s_{2j(1)+1}$ with $\bar{s}_3$ a right factor of $y_4 \in Y_{2j(1)+1} \subseteq Y_{2k+1}$. Thus, generally, we can write for convenience

$$s_{2j(3)+1} = \bar{s}_3\bar{p}_2\bar{s}_1 s$$

where $\bar{s}_1$ and $\bar{s}_3$, which may be empty, belong to $S(Y_{2k+1})$ and $\bar{p}_2$ is a right factor of $\bar{p}'_2$, which may also be empty, and if $|\bar{p}_2| > 0$ then $\bar{p}_2\bar{s}_1 s$ is a left $Y_{2j(1)+1}, |\bar{p}_2|$-precanonical word and, moreover, if $0 < |\bar{p}_2| < |\bar{p}'_2|$, it is not a minimal left $Y_{2j(1)+1}, |\bar{p}_2|$-precanonical word.

Suppose that we have repeated the argument for t times, $t > 1$, to obtain the sequence

$$1 \leq j(2t-1) < \cdots < j(3) < j(1) < k,$$

the words

$$\begin{gathered} y_1 = a_{2j(1)+1}u_{2j(1)+1}s_{2j(1)+1} \in T_{2j(1)+1} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_{2t-1} = a_{2j(2t-1)+1}u_{2j(2t-1)+1}s_{2j(2t-1)+1} \in T_{j(2t-1)+1} \end{gathered}$$

with the usual notation, and the words

$$\begin{gathered} s_{2j(1)+1} = \bar{s}_1 s \\ \cdots\cdots\cdots \\ s_{2j(2t-1)+1} = \bar{s}_{2t-1}\bar{p}_{2t-1}\ldots\bar{p}_2\bar{s}_1 s \end{gathered}$$

where $\bar{s}_1, \ldots, \bar{s}_{2t-1} \in S(Y_{2k+1})$, $\bar{p}_2, \ldots, \bar{p}_{2t-2}$ are right factors of $\bar{p}'_2, \ldots, \bar{p}'_{2t-2}$, respectively, $\bar{p}_{2i}\ldots\bar{s}_1 s$ is left $Y_{2j(i)+1}, |\bar{p}_{2i}|$-precanonical if $|\bar{p}_{2i}| > 0$, and it is so but not minimal if $0 < |\bar{p}_{2i}| < |\bar{p}'_{2i}|$ for $i = 1, \ldots, t-1$.

Now we are at the step $t+1$. Reasoning as in all steps before, let $\bar{p}'_{2t}$ be a right factor of $u_{2j(2t-1)+1}$, as short as possible, such that

$$\bar{p}'_{2t}s_{2j(2t-1)+1} \in R(Y_{2j(2t-1)+1}).$$

Observe that, as before, $\bar{p}'_{2t}s_{2j(2t-1)+1}$ is a left $Y_{2j(2t-1)+1}, |\bar{p}_{2t}|$-precanonical word.

We choose now an overlapping relation for the right border $\bar{p}'_{2t}s_{2j(2t-1)+1}$

$$y_{2t+2}\bar{p}'_{2t}s_{2j(2t-1)+1} = l_{2t+1}y_{2t+1}$$

with $|\bar{p}'_{2t}s_{2j(2t-1)}| < |y_{2t+1}|$ for some words $y_{2t+1}, y_{2t+2} \in Y_{2j(2t-1)+1}$.

We suppose further that at this moment we do not have $y_{2t+1} \in T_i$ for any $i \leq 2k+1$ as before, but instead

(a) $y_{2t+1} \in X_0$. In this case we assign the type -1 to s. We have certainly $s \in S(X_0)$ and $|s| < \max X_0$;

(b) $y_{2t+1} \in U_{2i}$ for some $i \geq 0$. In this case, we assign the type i to s. We describe some properties of s. We have as usual $y_{2t+1} = p_{2i}v_{2i}a_{2i}$ with $a_{2i} \in A$, $v_{2i} \in A^+, |v_{2i}| = \lambda n_{2i}$, $p_{2i} \in A^+, |p_{2i}| < n_{2i}$; and

$$y_{2t+2}\bar{p}'_{2t}s_{2j(2t-1)+1} = l_{2t+1}p_{2i}v_{2i}a_{2i}$$

or

$$y_{2t+2}\bar{p}'_{2t}\bar{s}_{2t-1}\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s = l_{2t+1}p_{2i}v_{2i}a_{2i}.$$

We estimate the length of s. On the one hand, obviously

$$|s| < |y_{2t+1}| \leq \max U_{2i} = \max U_{2t(s)}.$$

On the other hand, to bound $|s|$ from below, we observe that

$$|\bar{p}'_{2t}\bar{s}_{2t-1}\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s| > |v_{2i}a_{2i}|$$

for, otherwise, in view of the fact that $p_{2i}v_{2i}a_{2i}$ is right X_{2i}-precanonical of tag $|p_{2i}|$, y_{2t+2} has a right factor in $L(X_{2i})$ that is in contradiction with the comma-freeness of Y_{2k+1}.

If $|v_{2i}a_{2i}| \leq |s|$ then we get immediately

$$|s| \geq |v_{2i}| + 1 = \lambda n_{2i} + 1.$$

If, otherwise,

$$|s| < |v_{2i}a_{2i}| < |\bar{p}'_{2t}\bar{s}_{2t-1}\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s| < |p_{2i}v_{2i}a_{2i}| = |y_{2t+1}|$$

then we show that

$$|\bar{p}'_{2t}| \leq n_{2i} - 1, |\bar{p}_{2t-2}| \leq n_{2i} - 1, \dots, |\bar{p}_2| \leq n_{2i} - 1$$

and

$$|\bar{s}_{2t-1}| < n_{2i} - 1, \dots, |\bar{s}_1| < n_{2i} - 1.$$

Consider first the case of $\bar{p}'_{2t}$. If

$$|v_{2i}a_{2i}| \leq |\bar{s}_{2t-1}\bar{p}_{2t-2}\dots, \bar{p}_2\bar{s}_1 s|$$

which implies

$$n_{2i} - 1 \geq |p_{2i}| > |\bar{p}'_{2t}|$$

so we are done. For the other case

$$|v_{2i}a_{2i}| > |\bar{s}_{2t-1}\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s|$$

we have

$$|y_{2t+2}\bar{p}'_{2t}| > |l_{2t+1}p_{2i}|.$$

If $\bar{p}'_{2t} = 1$, we have nothing to prove, $|\bar{p}'_{2t}| = 0 < n_{2i} - 1$; otherwise, in case

$$\bar{s}_{2t-1}\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s$$

is a left $Y_{2j(2t-1)+1}, |\bar{p}'_{2t}|$-precanonical word and for the right factor $\bar{p}''_{2t}$ of length $|\bar{p}'_{2t}| - 1$ of $\bar{p}'_{2t}$,

$$\bar{p}''_{2t}\bar{s}_{2t-1}\dots\bar{p}_2\bar{s}_1 s$$

is a left $Y_{2j(2t-1)+1}, |\bar{p}''_{2t}|$-precanonical, but not minimal, word because of the assumption on the (minimal) length of $\bar{p}'_{2t}$. Consequently,

$$\bar{p}''_{2t}\bar{s}_{2t-1}\dots\bar{p}_2\bar{s}_1 s',$$

where $s'a_{2i} = s$, is also a left $Y_{2j(2t-1)+1}, |\bar{p}''_{2t}|$-precanonical word.

Suppose that $|\bar{p}''_{2t}| \geq n_{2i} - 1$. Then by Remark 3.1.7

$$\bar{p}''_{2t}\bar{s}_{2t-1}\dots\bar{p}_2\bar{s}_1 s'$$

being a right factor of $p_{2i}u_{2i}$ is also a right $X_{2i}, |\bar{p}''_{2t}|$-precanonical word. But this is in contradiction with the Corollary 3.1.10 which says that $|\bar{p}''_{2t}| < n_{2i} - 1$. So we must have $|\bar{p}''_{2t}| < n_{2i} - 1$, and $|\bar{p}'_{2t}| \leq n_{2i} - 1$.

The proofs for the cases $|\bar{p}_{2t-2}| \leq n_{2i} - 1$,..., $|\bar{p}_2| \leq n_{2i} - 1$ are just the same, only note that $\bar{p}_{2t-2}, \dots, \bar{p}_2$ are right factors of $\bar{p}'_{2t-2}, \dots, \bar{p}'_2$, respectively, on which the assumption of minimal length is made.

As of the cases of $\bar{s}_{2t-1}, \dots, \bar{s}_1$ the proofs are in the same vein but simpler. We prove, for example,

$$|\bar{s}_{2t-1}| < n_{2i} - 1,$$

the remaining cases are just identical. If

$$|\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s| \geq |v_{2i}a_{2i}|$$

then

$$|\bar{s}_{2t-1}| < |p_{2i}| \leq n_{2i} - 1.$$

If, otherwise,

$$|\bar{p}_{2t-2}\dots\bar{p}_2\bar{s}_1 s| < |v_{2i}a_{2i}|$$

then

$$|y_{2t+2}\bar{p}'_{2t}\bar{s}_{2t-1}| \geq |l_{2t+1}p_{2i}|.$$

Again, as $p_{2i}v_{2i}a_{2i}$ is right X_{2i}, n_{2i}-precanonical word (of tag $|p_{2i}|$), it follows that

$$y_{2t+2}\bar{p}'_{2t}\bar{s}_{2t-1} \in L(X_{2i})A^*.$$

Therefore, we should have $|\bar{s}_{2t-1}| < n_{2i} - 1$, since in the opposite case, the inequality $|\bar{s}_{2t-1}| \geq n_{2i} - 1 \geq \max L(X_{2i})$ leads to $\bar{s}_{2t-1} \in L(X_{2i})A^*$, which contradicts the comma-freeness of Y_{2k+1} as $\bar{s}_{2t-1} \in S(Y_{2k+1})$.

Now we are in a position to bound $|s|$ from below. Since $|v_{2i}| = \lambda n_{2i}$ and each of the values $|\bar{s}_1|, \ldots, |\bar{s}_{2t-1}|$ is less than $n_{2i} - 1$ and each of the values $|p'_{2t}|, |p'_{2t-2}|, \ldots, |p'_2|$ is no more than $n_{2i} - 1$, we get

$$\begin{aligned} |s| &> |v_{2i}a_{2i}| - |p'_{2t}\bar{s}_{2t-1}\ldots\bar{p}_2\bar{s}_1| \\ &> \lambda n_{2i} + 1 - t(n_{2i} - 1) - t(n_{2i} - 2) = (\lambda - 2t)n_{2i} + 3t + 1 \\ &> (\lambda - 2t)n_{2i}. \end{aligned}$$

From the chain

$$1 \leq j(2t-1) < \cdots < j(3) < j(1) < k$$

we deduce $t < k$ and finally we get

$$|s| > (\lambda - 2(k-1))n_{2i}.$$

Summarizing, we have so far established

1. every simple right border s of Y_{2k+1} has type $t(s)$, which lies between -1 and k;
2. if $t(s) = -1$ then s is a right factor of X_0; in particular $|s| < \max X_0$;
3. if $t(s) \geq 0$ then s is a right factor of $U_{2t(s)}$; in particular $|s| < \max U_{2t(s)}$ and $|s| > \min(\lambda n_{2t(s)}, (\lambda-1)n_{2t(s)}, (\lambda-2(k-1))n_{2t(s)}) > (\lambda-2k+1)n_{2t(s)}$.

We are just one step away from the following.

Theorem 4.2.4. *For a simple right border s of Y_{2k+1}, if $t(s) = -1$ then s is a right factor of X_0, in particular $|s| < \max X_0$. If $t(s) \geq 0$ then s is a right factor of $U_{2t(s)}$ and $(\lambda - 2k + 1)n_{2t(s)} < |s| < \max U_{2t(s)}$, moreover, s is a right $X_{2t(s)}, m_{2t(s)}$-canonical word with tag less than $n_{2t(s)}$ and $m_{2t(s)} > (\lambda - 2k)n_{2t(s)}$.*

Proof. We have to show only the last claim. Suppose that $t(s) \geq 0$. Since s is a right factor of some word of $U_{2t(s)}$, which is a (minimal) right $X_{2t(s)}, \lambda n_{2t(s)}$-canonical word of tag less than $n_{2t(s)}$ then s is a right $X_{2t(s)}, m_{2t(s)}$-canonical word of the same tag and for some $m_{2t(s)}$. From the fact that $|s| > (\lambda-2(k-1))n_{2t(s)} > (\lambda - 2k + 1)n_{2t(s)}$, we see at once that

$$m_{2t(s)} > (\lambda - 2k)n_{2t(s)}$$

which concludes the proof. □

5. Canonical comma-free codes

In this section, as a first step in completing finite comma-free code to a finite maximal one, we introduce the concept of canonical comma-free code and we prove that every finite comma-free code is included in a finite canonical comma-free code. We continue using the notation of the previous section.

5.1. Basic theorems. Canonical comma-free code

Let $\lambda \geq 2k+2$ and $g \geq n_{2k+1}$. We have the following instrumental result.

Theorem 5.1.1. *If $E(Y_{2k+1})$ has a left Y_{2k+1}, g-precanonical factor then there exists a left border $s \in R(Y_{2k+1})$ such that $s \in S(X_0)$ but $s \notin R(X_{2k})$, hence $s \notin R(Y_{2k-1})$.*

Proof. Let

$$a_1 \dots a_g a_{g+1} \dots a_{g+t},$$

where $a_1, \dots, a_{g+t}$ are letters, be a left Y_{2k+1}, g-precanonical factor of $E(Y_{2k+1})$ (of tag t). We prove the following

Claim k. For every positive integer i,

$$i \leq g - (\lambda+1)n_{2k}$$

the word

$$a_i \dots a_g a_{g+1} \dots a_{g+t}$$

has no left factors which are right borders of type k.

In order to prove this, suppose by contradiction that $a_i \dots a_g a_{g+1} \dots a_{g+t}$ has a left factor

$$s = a_i \dots a_h a_{h+1}$$

which is a simple right border of type k. By Theorem 4.2.4, $|s| < \max U_{2k}$, which is less than $(\lambda+1)n_{2k}$, hence $h+1 \leq g$ as $g - i \geq (\lambda+1)n_{2k}$, and s is a right X_{2k}, m_{2k}-canonical word of tag less than n_{2k} and with $m_{2k} \geq (\lambda-2k)n_{2k} \geq 2n_{2k}$. To be more precise, we write

$$s = a_i \dots a_f a_{f+1} \dots a_h a_{h+1},$$

where $|a_i \dots a_f| < n_{2k}$, $h - f = m_{2k}$ and

$$s' = a_i \dots a_f a_{f+1} \dots a_h$$

is a right X_{2k}-precanonical word of tag $f - i + 1$. Put $d = h - n_{2k}$, $e = h - 2n_{2k}$. Certainly,

$$i \leq f \leq e < d < h < g.$$

We show that for every integer j in the range

$$e < j \leq d$$

the word

$$a_j \dots a_g a_{g+1} \dots a_{g+t}$$

has no left factor which is a simple right border of Y_{2k+1} of type k.

In fact, if it has such a right border

$$a_j \dots a_{j+h'}$$

as a left factor, which we present, as a right Y_{2k+1}, m_{2k}-canonical word (as $X_{2k} \subseteq Y_{2k+1}$), as follows:

$$a_j \dots a_{j+f'-1} a_{j+f'} \dots a_{j+h'-1} a_{j+h'}$$

where

$$n_{2k} > f' > 0, \quad h' - f' = m_{2k}.$$

Now observe that

$$j + f' < j + n_{2k} \leq d + n_{2k} = h$$

and

$$h = e + 2n_{2k} < j + 2n_{2k} < j + f' + 2n_{2k} < j + f' + m_{2k} = j + h',$$

hence

$$h + 1 < j + h'.$$

These two inequalities show that, with respect to the right Y_{2k+1}, m_{2k}-canonical word $a_j \dots a_{j+h'}$ of tag f', the word $a_j \dots a_{h+1}$ and, therefore, the word $a_i \dots a_h a_{h+1}$ have a right factor which is in $L(Y_{2k+1})$. But it is impossible, because $a_i \dots a_h a_{h+1}$ is a right Y_{2k+1}-canonical word. Therefore for every j, $e < j \leq d$,

$$a_j \dots a_g a_{g+1} \dots a_{g+t}$$

must have a left a factor which is a right border of type less than k and which is of length less than $\max U_{2(k-1)} < n_{2k-1} \leq n_{2k}$. But this contradicts Proposition 3.1.9 with respect to the right X_{2k}-precanonical word s' (herewith, $X_{2k} = Y$) which says that there should be an integer j such that

$$e = d - n_{2k-1} < d - \max U_{2(k-1)} < j \leq d,$$

for which $a_j \dots a_g a_{g+1} \dots a_{g+t}$ has no such a factor! This total contradiction proves Claim k.

Thus $a_i \dots a_g a_{g+1} \dots a_{g+t}$ has left factors that are simple right borders only of type less than k, for every $i \leq g - (\lambda+1)n_{2k}$. By an argument entirely identical to the one above, we can further establish for every $l \leq 0$.

Claim l. For every $i \leq g - (\lambda+1)n_{2k} - \dots - (\lambda+1)n_{2l}$ the word

$$a_i \dots a_g a_{g+1} \dots a_{g+t}$$

has no left factors which are simple right borders of type greater than or equal to l.

When $l = 0$, we finally get that for every integer i satisfying

$$i \leq g - (\lambda+1)n_{2k} - \dots - (\lambda+1)n_0 = (g - n_{2k+1}) + n_{2k-1}$$

the word

$$a_i \dots a_g a_{g+1} \dots a_{g+t}$$

has only right simple borders of Y_{2k+1} of type -1 as left factors.

Now that $E(Y_{2k+1}) \subseteq E(X_{2k})$ and $E(X_{2k})$ has no factors that are left X_{2k}, n_{2k-1}-precanonical words, it immediately follows that there exists a (simple) right border of Y_{2k+1} of type -1, meaning that it is a right factor of X_0, and it is not a right border of X_{2k}. This achieves the proof of the theorem. □

All we have done so far is to prepare the ground for the following statement.

Theorem 5.1.2. *Let $k = |S(X_0)|+1$ and $\lambda = 2k+2$ then $E(Y_{2k+1})$ does not contain any left Y_{2k+1}, n_{2k+1}-precanonical factors. Consequently, $E(Y_{2k+1})$ contains neither left nor right Y_{2k+1}, n_{2k+1}-precanonical factors.*

Proof. We first show that if two adjacent members of the sequence $Y_1, Y_3, \ldots, Y_{2k+1}$ equal then it stabilizes from that place on. For instance, we prove that

$$X_{2i} = Y_{2i+1}$$

implies

$$Y_{2i+1} = X_{2i+2}$$

for any i, $0 < i < k$. (The other case is handled analogously.) The equality

$$Y_{2i+1} = X_{2i+2}$$

is equivalent to

$$T(Y_{2i+1}, n_{2i+1}) = \emptyset$$

or the same

$$T(X_{2i}, n_{2i+1}) = \emptyset.$$

The equality above holds, since

$$X_{2i} = Y_{2i-1} + T(Y_{2i-1}, n_{2i-1}),$$

by Theorems 3.2.9 and 3.2.5, $T(X_{2i}, n_{2i-1}) = \emptyset$ and $n_{2i+1} > n_{2i-1}$.

We now turn to the proof of the theorem itself. If $Y_{2k+1} = X_{2k}$, we are done, since $E(X_{2k})$ has no left X_{2k}, n_{2k-1}-precanonical factors, while $n_{2k-1} < n_{2k+1}$. Otherwise, we see by the stability observation above that for $0 < i \leq k$

$$X_{2i} \neq Y_{2i-1},$$

or equivalently,

$$T(Y_{2i-1}, n_{2i-1}) \neq \emptyset$$

which means that $E(Y_{2i-1})$ contains left Y_{2i-1}, n_{2i-1}-precanonical factors.

By Theorem 5.1.1, as $\lambda = 2k+2 \geq 2i+2$, there exist a word $s_{i-1} \in S(X_0)$ such that $s_{i-1} \in R(Y_{2i-1})$ but $s_{i-1} \notin R(Y_{2i-3})$ for each $1 < i \leq k$. Now if in addition $E(Y_{2k+1})$ has a left Y_{2k+1}, n_{2k+1}-precanonical factor, then there exists $s_k \in S(X_0)$ such that $s_k \in R(Y_{2k+1})$ and $s_k \notin R(Y_{2k-1})$. This implies that $s_1, \ldots, s_k$ are all distinct, hence $k \leq |S(X_0)|$. But this is impossible because of the presumed value of k, which proves the first claim. The last claim is straightforward. □

The Theorem 5.1.2 motivates the central notion of this paper. Let N be a positive integer.

Definition 5.1.2. A comma-free code X is said to be N-canonical if $E(X)$ contains no left X, N-precanonical factor and no right X, N-precanonical factors. A comma-free code is said to be canonical if it is N-canonical for some positive integer N.

Evidently, this definition is equivalent to the next, which is more explicit.

Definition 5.1.3. 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 $s'y \notin R(X)A^*$.

In other words, 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 $p, s \in E(X)$ (or just the same $p \notin A^*L(X)$, $s \notin R(X)A^*$) and $n \leq |p|, |s| < n + N$. We can imagine that for an N-canonical comma-free code X and for every word w of $E(X)$ the left factors and the right factors of w which are also in $E(X)$ are distributed uniformly (no more than an N-distance apart).

Example 5.1.3. Let $A = \{a, b\}$; the set $\{bab\}$ is a 2-canonical comma-free code, for which $R = \{ab\}$; $L = \{ba\}$ and which is not maximal as $\{bab, baa\}$ is a comma-free code. Moreover, it can be proved that every one-word comma-free code is canonical.

5.2. Completing to canonical comma-free codes

Now we come to the culminating point of this section, the completion theorem.

Theorem 5.2.1. *Every finite comma-free code X can be completed to a finite N-canonical comma-free code Y with*

$$\max Y \leq 4(\lambda + 1)^2(\lambda + 2)\max X_0$$

if $k = 2$ or

$$\max Y \leq 2^{k-1}(\lambda + 1)^2(\lambda + 2)^{k-2}\max X_0$$

if $k > 2$ and

$$N = 2^{k-1}(\lambda + 1)(\lambda + 2)^{k-1}\max X_0$$

where $k = |S(X)| + 1$, $\lambda = 2k + 2$.

Proof. We see by Theorem 5.1.2 that $X = X_0$ is included in Y_{2k+1} which is a n_{2k+1}-canonical finite comma-free code. We estimate n_{2k+1} and $\max Y_{2k+1}$.

By the inequalities

$$\max Y_{2i+1} \leq (\lambda + 1)n_{2i}, \quad i = 0, 1, \ldots, k$$

and

$$n_{2i} < 2n_{2i-1}, \quad i = 1, \ldots, k$$

we have

$$\max Y_{2i+1} < 2(\lambda+1)n_{2i-1}, \quad i = 1, \ldots, k.$$

By definition, for $1 < i \leq k$,

$$\begin{aligned} n_{2i+1} - n_{2i-1} &= (\lambda+1)n_0 + \cdots + (\lambda+1)n_{2i} = n_{2i-1} - n_{2i-3} + (\lambda+1)n_{2i} \\ &< n_{2i-1} + 2(\lambda+1)n_{2i-1} = (2\lambda+3)n_{2i-1}. \end{aligned}$$

Hence

$$n_{2i+1} < (2\lambda+4)n_{2i-1}$$

for $i > 1$. Due to this relation and the fact that $k \geq 2$ and $n_3 < (2(\lambda+1)^2 + 2(\lambda+1))n_0$ a trite calculation yields

$$n_{2k+1} \leq 2^k(\lambda+1)(\lambda+2)^k \max X_0$$

and

$$\max Y_{2k+1} \leq 4(\lambda+1)^2(\lambda+2)\max X_0$$

if $k = 2$, or else

$$\max Y_{2k+1} \leq 2^{k-1}(\lambda+1)^2(\lambda+2)^{k-2}\max X_0$$

if $k > 2$, as desired to prove. □

6. Concluding remarks

As said before, Theorem 5.2.1 represents the first stage of the problem of finite completion of comma-free code by proving that there always exists a finite N-canonical comma-free code containing a given finite comma-free code X. We can figure out that the procedure is actually an effective one. To wit, starting from a finite comma-free code X, the sets $Y_1, Y_3, \ldots, Y_{2k+1}$ are effectively constructible, at least by the fact that the sets of left and right m-canonical words are always finite for all m, putting aside complexity considerations.

We would like to remark that the estimates do not tend to be best possible; rather, we present them in a more or less quantitative manner to give flavor of the method. The present work is a full exposition of [9], with a slight modification concerning the sequence $n_0, n_1, \ldots, n_{2k}, n_{2k+1}$. This is carried out to ensure that $E(Y_{2k+1})$ is free of n_{2k+1}-precanonical factors, however, at the expense of somewhat larger bounds on $\max Y_{2k+1}$.

We hope to publish the remaining part of the solution, namely, the proof that every finite N-canonical comma-free code, for all $N > 0$, has a finite completion, in another paper.

Acknowledgements. I am greatly indebted to the referee for the comments and suggestions of high expertise.

References

[1] J. Berstel and D. Perrin, *Theory of Codes*. Academic Press, Orlando (1985).
[2] F.H.C. Crick, J.S. Griffith and L.E. Orgel, Codes without Commas. *Proc. Natl. Acad. Sci. USA* **43** (1957) 416-421.
[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] W.L. Eastman, On the Construction of Comma-free Codes. *IEEE Trans. Inform. Theory* **IT-11** (1965) 263-267.
[6] C.M. Fan and H.J. Shyr, Some Properties of Maximal Comma-free Codes. *Tamkang J. Math.* **29** (1998) 121-135.
[7] 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.
[8] B.H. Jiggs, Recent Results in Comma-free Codes. *Canad. J. Math.* **15** (1963) 178-187.
[9] N.H. Lam, Finite Completion of Comma-free Codes. Part I, in *Proc. DLT*. Springer-Verlag, *Lect. Notes Comput. Sci.* **2450** (2002) 357-368.
[10] Al. A. Markov, An Example of an Idependent 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] R.A. Scholtz, Maximal and Variable Word-length Comma-free Codes. *IEEE Trans. Inform. Theory* **IT-15** (1969) 555-559.
[13] H.J. Shyr, *Free Monoids and Languages*. Lecture Notes, Hon Min Book Company, Taichung (1991).
[14] 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.