INFORMATIQUE THÉORIQUE ET APPLICATIONS

YANNICK GUESNET

On codes with finite interpreting delay : a defect theorem

Informatique théorique et applications, tome 34, nº 1 (2000), p. 47-59

<http://www.numdam.org/item?id=ITA_2000__34_1_47_0>

Theoretical Informatics and Applications
Theoret. Informatics Appl. **34** (2000) 47–59

ON CODES WITH FINITE INTERPRETING DELAY: A DEFECT THEOREM

YANNICK GUESNET[1]

Abstract. We introduce two new classes of codes, namely adjacent codes and codes with finite interpreting delay. For each class, we establish an extension of the defect theorem.

AMS Subject Classification. 94A45.

INTRODUCTION

In theoretical computer science, the questions connected to coding play a prominent part, by their mathematical specificity as well as their potentiality of practical applications. From this point of view, the aim of the theory of codes consists in studying the properties concerning factorization of words. Remarkable results illustrate the relevance and the difficulty of this issue. Some famous special classes of codes, like bifix codes [2] and codes with finite deciphering delay [4], are directly concerned. In fact, when reading sequentially a word, these types of sets allow an efficient deciphering of the corresponding message.

A generalization of the notion of factorization of a word is the concept of interpretation. An interpretation of a word w with respect to a code X is a triplet (s, d, p) such that $s.d.p = w$ where $d \in X^*$ and s (resp. p) is a proper suffix (resp. proper prefix) of a word in X. This notion is natural when considering the different configurations linked to the factors of w in X^*. Clearly a word may have several interpretations, but of course, in view of an easy deciphering, a minimal number of interpretations is required. From this point of view, some famous classes of codes have been introduced, namely codes with finite synchronization delay [5, 7] and circular codes [2]. However, the unicity of interpretation is not reached.

In this paper, we introduce a new class of codes, namely *codes with finite interpreting delay* (f.i.d. codes for short), of which main specificity consists in avoiding several interpretations. Unformally, if X is a code with finite interpreting delay,

[1] LIFAR, Université de Rouen, Place Émile Blondel, 76821 Mont-Saint-Aignan, France; e-mail: Yannick.Guesnet@dir.univ-rouen.fr

then any "long enough" word w in X^* has a unique interpretation. In fact, we define the *interpreting delay* of X as the smallest integer n such that $\beta X^* \alpha \cap X^n = \emptyset$, for all pair of words (α, β), with α prefix of a word of X, β suffix of a word of X and such that at least $\alpha \notin X^*$ or $\beta \notin X^*$. In the point of view of the deciphering of a message $w_1 x w_2 \in X^*$, when an error of transmission occurs in $x \in X$, the f.i.d. codes allows the word w_2 to be recognized. Note that with uniformly synchronous codes with delay n only the word w_3, where $w_2 = w' w_3$ with $w' \in X^n$ and $w_3 \in X^n.X^*$, can be recognized: we lose the information included in w'.

Clearly, apart form its powerful applications connected to deciphering, this class of codes must satisfied strict theoretical criteria: investigating these properties is the aim of our paper. In this matter, we prove a first remarkable result: any finite intersection of submonoids generated by f.i.d. codes is itself generated by a f.i.d. code.

Moreover, it is well-known that each of the preceding classes of codes, namely prefix codes, codes with finite deciphering delay, circular codes, satisfies an extension of the defect theorem [2,3,6]. In our paper, we establish that f.i.d. codes also satisfy a version of this theorem:

Theorem 1. *For any finite subset $X \subset \Sigma^*$, there exists a smallest code Y with finite interpreting delay satisfying $X \subset Y^*$ and $|Y| \leqslant |X|$.*

The proof relies on the characterization of the class of codes with finite interpreting delay as the intersection of the classical class of circular codes with the "adjacent codes" which we define as follow: a non-empty set X is an adjacent code iff $X \cap (S(X) \setminus \{\varepsilon\}).X^+ = \emptyset$ and $X \cap X^+.(P(X) \setminus \{\varepsilon\}) = \emptyset$. Where $S(X)$ ($P(X)$) stands for the suffixes (prefixes) of the words of X. In fact we establish an another extension of the defect theorem for these codes. Our method leads to an algorithm for computing the preceding code Y itself [9].

We now describe the contents of our paper. First section deals with classical elementary notions from the free monoid theory.

In Section 2, we introduce codes with finite interpreting delay. Some basic results are established. In particular we show that any code with finite interpreting delay is circular [2].

In the last section, we show that codes with finite interpreting delay admit an extension of the defect theorem.

1. Preliminaries

1.1. Definitions and notations

In all this paper, we denote by Σ a finite alphabet, by Σ^* the free monoid it generates and by ε the empty word.

For all subsets X and Y of Σ^*, we denote by $X^{-1}Y$ (XY^{-1}) the set $\{v \in \Sigma^* / \exists u \in X, uv \in Y\}$ ($\{u \in \Sigma^* / \exists v \in Y, uv \in X\}$) and by $X.Y$ their concatenation product.

For any subset X of Σ^*, we denote by X^* the submonoid generated by X and by X^+ the set $X^* \setminus \{\varepsilon\}$.

Given a word $w \in \Sigma^*$, the set of all factors (prefixes, suffixes) of w is denoted by $F(w)$ $(P(w), S(w))$. A set X is *prefix* (*suffix*) if no element of X is prefix (suffix) of another one. If $w \in X^+$, we say that $u \in F(X)$ is an *X-factor* of w iff $w \in X^*uX^*$. For any subset X, we denote by $\overline{P(X)}$ $(\overline{S(X)})$ the set $P(X) \setminus \{\varepsilon\}$ $(S(X) \setminus \{\varepsilon\})$.

Two words w and w' are *conjugated* if there exist two words u and v such that $w = uv$ and $w' = vu$; if $u, v \in X^*$ we say that w and w' are *X-conjugated.*

Given a word $w \in \Sigma^*$, we denote by $|w|$ the length of the word.

Let $X \subset \Sigma^*$ and let $w \in \Sigma^*$. An interpretation of w with respect to X is a triplet (s, d, p) such that $s.d.p = w$ where $d \in X^*$, $p \in P(X) \setminus X$ and $s \in S(X) \setminus X$. Two X-interpretations (s, d, p) and (s', d', p') of the word w are *adjacent* if there exist $d_1, d_2, d_1', d_2' \in X^*$ such that

$$d = d_1 d_2, \quad d' = d_1' d_2', \quad s d_1 = s' d_1' \quad \text{and} \quad d_2 p = d_2' p'.$$

The interpretation $(\varepsilon, w, \varepsilon)$ is the *trivial interpretation* of w and an interpretation of w is *proper* if it is not adjacent to the trivial one.

Let $w \in \Sigma^*$ and X be a code, we denote by $\delta_X(w)$ the maximal number of pairwise disjoint X-interpretations of w.

1.2. Some definitions on codes

(*i*) A code X has a *finite deciphering delay* if there exists $d \geqslant 0$ such that

$$\forall x, x' \in X,\ \forall y \in X^d,\ \forall u \in \Sigma^*, \qquad xyu \in x'X^* \Rightarrow x = x'.$$

(*ii*) A code X is *circular* if for all $n, m \geqslant 1$, $x_1, \ldots, x_n \in X$, $y_1, \ldots, y_m \in X$, $p \in \Sigma^*$ and $s \in \Sigma^+$ the equalities

$$s x_2 \ldots x_n p = y_1 \ldots y_m, \qquad x_1 = ps$$

imply

$$n = m, \quad p = \varepsilon \quad \text{and} \quad x_i = y_i \quad \text{for} \quad i = 1, \ldots, n.$$

(*iii*) Let $p, q \geqslant 0$. A submonoid M of Σ^* satisfies *condition* $C(p,q)$ if for any sequence $u_0, u_1, \ldots, u_{p+q}$ of words of Σ^*, the condition $u_{i-1}u_i \in M$ $(1 \leqslant i \leqslant p+q)$ implies $u_p \in M$ (Fig. 1). A code X is *(p,q)-limited* if X^* satisfies $C(p,q)$.

A code X is *limited* if there exist $p, q \geqslant 0$ such that X is (p,q)-limited.

Proposition 1.1 ([2], p. 330). *Any limited code is a circular code.*

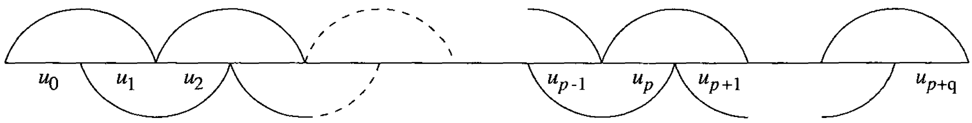

FIGURE 1. Condition $C(p, q)$.

1.3. THE DEFECT THEOREM AND AN EXTENSION TO CIRCULAR CODES

In this section, we are interested by the following problem. Let X be a subset of Σ^*. Assume that X is not a code. How can we compute a "convenient" code generating the elements of X.

One way to answer this question consists in constructing the smallest free submonoid containing X. This is justified by the fact that the intersection of all free submonoids containing X is still a free submonoid. In fact, we say that this free submonoid is the *free hull* of X, whose main property is the famous defect theorem (see *e.g.* [3], pp. 48-50):

Theorem 1.2 (Defect theorem). *Let $X \subset \Sigma^*$ and let Y be the base of the free hull of X. If X is not a code, then*

$$|Y| \leqslant |X| - 1.$$

Several extensions of this result have been established for special well-known classes of codes, due to the fact that the corresponding submonoids are closed under finite intersections. Let's mention prefix codes, bifix codes and codes with finite deciphering delay [3].

Actually, we are interested by the class of circular codes:

The *circular hull* of a set $X \subset \Sigma^*$ is the smallest submonoid generated by a circular code containing X. We have the following result:

Theorem 1.3 (see *e.g.* [6]). *Let $X \subset \Sigma^*$ and let Y be the base of its circular hull. Then*

$$|Y| \leqslant |X|.$$

Note that in this case we have not a strict inequality between the cardinality of Y and the cardinality of X.

2. CODES WITH FINITE INTERPRETING DELAY

In this section, we introduce the notion of codes with finite interpreting delay (f.i.d. codes for short) and we compare these codes with the preceding circular codes.

2.1. The basic properties

Definition 2.1. Let X be a code. X has a finite interpreting delay if there exists $m \geqslant 1$ such that for all $\alpha \in P(X)$, $\beta \in S(X)$, $(\alpha, \beta) \notin X^* \times X^*$, we have:

$$\beta X^* \alpha \cap X^m = \emptyset. \tag{1}$$

The *interpreting delay* is the smallest integer m satisfying condition (1). In other words, if m is the interpreting delay of a code X, for any X-interpretation (s, d, p) of a word $w \in X^m$ we have $s, p \in X^*$; thus there exists no proper interpretation of w, *i.e.*

$$\forall w \in X^m, \qquad \delta_X(w) = 1.$$

Although our notion of interpreting delay seems close to the concept of deciphering delay (*cf. e.g.* [4]), it is different, as attested by the following example. The regular code $a + b + ab^+c$ has an interpreting delay 1 but has no finite deciphering delay. Other examples will be presented in Section 2.2.

Lemma 2.2. *If a code has an interpreting delay n, it satisfies the condition (1) of Definition 2.1 for all $m \geqslant n$.*

Proof. Given a code X with interpreting delay n, assume that there exist an integer $m > n$ and $(\alpha, \beta) \in P(X) \times S(X)$, $(\alpha, \beta) \notin X^* \times X^*$ such that $\beta X^* \alpha \cap X^m \neq \emptyset$. More precisely, let $k \in \mathbb{N}$ and $x_1, x_2, \ldots, x_k \in X$, $y_1, y_2, \ldots, y_m \in X$ such that

$$\beta x_1 x_2 \ldots x_k \alpha = y_1 y_2 \ldots y_m.$$

Without loss of generality, we assume that $\beta \notin X^*$ (the case $\alpha \notin X^*$ being examined in a symmetrical way). Let i be the smallest integer such that $\sum_{h=1}^{n} |y_{h+i}| > |\beta|$, let $\beta' = (y_1 \ldots y_i)^{-1}\beta$ and let j be the greatest integer which satisfies $|\beta x_1 x_2 \ldots x_j| < |y_1 y_2 \ldots y_{n+i}|$.

If $j = k$, then we set

$$\alpha' = \alpha(y_{n+i+1} \cdots y_m)^{-1},$$

otherwise we set

$$\alpha' = (\beta x_1 \ldots x_j)^{-1} y_1 \ldots y_{n+i}.$$

We hold $(\alpha', \beta') \in P(X) \times S(X)$ and $\beta' \notin X^*$ (indeed we have $\beta = y_1 \ldots y_i \beta' \notin X^*$). Moreover we have $\beta' x_1 \ldots x_j \alpha' = y_{i+1} \ldots y_{i+n}$ (Fig. 2). This contradicts the fact that X has finite interpreting delay n.

As a consequence, X satisfies the condition (1) of Definition 2.1 for all $m \geqslant n$. □

We say that a submonoid $M \subset \Sigma^*$ generated by a f.i.d. code is a *f.i.d. submonoid.*

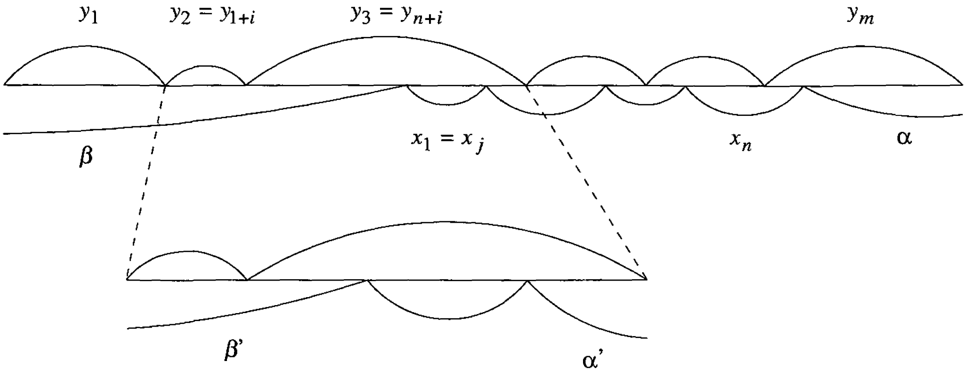

FIGURE 2. m greater than the interpreting delay (here 2).

First, we shall establish basic properties of f.i.d. codes.

Proposition 2.3. *Given a f.i.d. code X, any non-empty subset of X is a f.i.d. code. Moreover if X has interpreting delay n, then these subsets have an interpreting delay lower than or equal to n.*

Proof. Let X' be a non-empty subset of X. Trivially, X' is a code. We shall prove that X' has a finite interpreting delay.

We denote by n the interpreting delay of X. Let $\alpha \in P(X')$, $\beta \in S(X')$. Assume that

$$\beta X'^* \alpha \cap X'^n \neq \emptyset.$$

We shall establish that $(\alpha, \beta) \in X'^* \times X'^*$. Indeed, since $X' \subset X$, there exist $\alpha \in P(X)$, $\beta \in S(X)$ and a word w such that $w \in \beta X^* \alpha \cap X^n$. But X has delay n hence, by definition, we have $(\alpha, \beta) \in X^* \times X^*$. This implies that the word w may be factorized upon X^* (as $\beta.w'.\alpha$ with $w' \in X^*$) and X'^* ($w \in X'^n$). Since $X' \subset X$, these factorizations must correspond. Thus we obtain $\beta, \alpha \in X'^*$.

We have proved that X' has an interpreting delay lower than or equal to n. □

Remark 2.4. The classical algorithm of Sardinas and Patterson [8] can be modified in order to decide whether a finite code has a finite interpreting delay. Indeed, let (U_n), (V_n) be the sequences defined as indicated in the following:

$$U_1 = \overline{S(X)}^{-1}.X \setminus \{\varepsilon\}, \quad U_{n+1} = X^{-1}U_n \cup U_n^{-1}X \quad \text{for } n \geqslant 1,$$
$$V_1 = X.\overline{P(X)}^{-1} \setminus \{\varepsilon\}, \quad V_{n+1} = XV_n^{-1} \cup V_n X^{-1} \quad \text{for } n \geqslant 1.$$

X has a finite interpreting delay iff there exists $n \geqslant 1$ such that $U_n = V_n = \emptyset$ and $\varepsilon \notin \bigcup_{1 \leqslant i < n}(U_i \cup V_i)$. Indeed the condition $\varepsilon \in U_i$, $i > 0$ ($\varepsilon \in V_i$) corresponds to the existence of an interpretation of type (α, u, ε) $((\varepsilon, u, \beta))$.

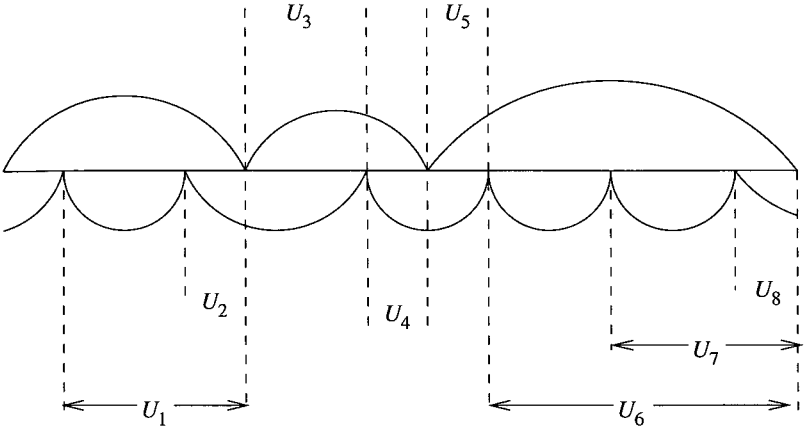

FIGURE 3. Some words in U_n.

2.2. COMPARISON WITH CLASSICAL CLASSES OF CODES

In fact, as established by the following proposition, any f.i.d. code is circular. More precisely, we establish that f.i.d. codes are limited.

Proposition 2.5. *Any f.i.d. code is circular.*

Proof. Let X be a code with interpreting delay n. We shall prove that the code X is $(1, 2n)$-limited and $(2n, 1)$-limited.

Let $u_0, u_1, \ldots, u_{2n+1}$ be words of Σ^* such that $u_{i-1}u_i \in X^*, 1 \leqslant i \leqslant 2n+1$.

- First, we assume that $u_0, u_1, \ldots, u_{2n+1} \in \Sigma^+$. With such a condition, there exists $m \geqslant n$ such that $u_1 \ldots u_{2n} \in X^m$ (indeed, we have $u_1u_2, u_3u_4, \ldots, u_{2n-1}u_{2n} \in X^+$). Since we have $u_0u_1 \in X^*$, we obtain $u_1 \in S(X).X^*$. In a similar way, we have $u_{2n} \in X^*.P(X)$. But the code X has an interpreting delay n, therefore, by Lemma 2.2 it satisfies condition (1) of Definition 2.1 for $m \geqslant n$, thus we have $u_1, u_{2n} \in X^*$.
- Now, we assume that there exists a smallest integer $i \in [0, 2n+1]$ such that $u_i = \varepsilon$.
  - If $i = 0$, since we have $u_0u_1 \in X^*$, we have in fact $u_1 \in X^*$.
  - If $i = 1$, we have $u_1 = \varepsilon$ thus $u_1 \in X^*$.
  - Assume that we have $i > 1$.

  In fact, we have $u_{i-1} \in X^+$, $u_{i-2}u_{i-1}.X^* \in X^+$ and $u_{i-2} \in \overline{S(X^*)}$. But $u_{i-1}.X^* \in X^+$, therefore $u_{i-2} \in X^+$ (indeed X has a finite interpreting delay). By decreasing induction we obtain $u_1 \in X^+$.

A similar argument with the greater j such that $u_j = \varepsilon$, leads to the conclusion $u_{2n} \in X^*$.

We have established that the code X is $(1, 2n)$ and $(2n, 1)$-limited. According to Proposition 1.1, X is a circular code. □

Remark 2.6. Let $w, w' \in X^*$, as a direct consequence of Proposition 2.5, if w and w' are conjugated, then they are in fact X-conjugated.

We have seen that the two notions of f.i.d. codes and codes with finite deciphering delay are different. However, in the case of finite sets, we have the following inclusion:

Corollary 2.7. *Any finite f.i.d. code has a finite deciphering delay.*

Proof. The proof is trivial: any finite circular code has a finite deciphering delay (*e.g.* see [2]). □

If no restriction is imposed, the following examples show that the deciphering delay and the interpreting delay are non-equivalent notions:

Example 2.8. The code $\{a, abc, b\}$ has an interpreting delay 1 and has a deciphering delay 2.

Example 2.9. The prefix code $\{abcd, bc, dc, ba\}$ (deciphering delay 0) has interpreting delay 2.

Remark 2.10. The converse of Proposition 2.5 does not hold in general. In fact, the class of f.i.d. codes is strictly included in the class of limited codes, as shown by the following example:

Example 2.11. The code $X = \{ba, bad, db\}$ is limited, since it is circular and finite ([2], p. 333), but any word belonging to $X^*.\{bad\}$ has an X-interpretation of the form $(\varepsilon, w.ba, d)$ with $w \in X^*$.

2.3. A characterization of finite f.i.d. codes

It is convenient to introduce the following notation: given a set X, we set $lg(X) = \sum_{x \in X}(|x| - 1)$.

Following proposition gives a property that circular codes must satisfy to be f.i.d. codes.

Proposition 2.12. *Let X be a finite circular code. The two following conditions are equivalent:*

1. *X is a f.i.d. code.*
2. *$X \cap \overline{S(X)}.X^+ = \emptyset$ and $X \cap X^+.\overline{P(X)} = \emptyset$.*

As we shall see below, condition 2 plays an important part in the proof of our main result (Sect. 3). This result states that the class of f.i.d. codes is the intersection between two suitable classes of codes. Before to prove Proposition 2.12, it is convenient to introduce the following definition:

Definition 2.13. $X \subset \Sigma^*$ is an adjacent set if $X \cap \overline{S(X)}.X^+ = \emptyset$ and $X \cap X^+.\overline{P(X)} = \emptyset$.

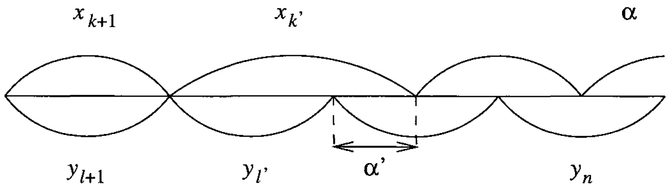

FIGURE 4. Definition of α'.

Clearly, if ε does not belong to X, the set X is a code. We say that X is an *adjacent code.*

The proof of Proposition 2.12 applies the following proposition [6]:

Proposition 2.14. *Let X be a finite circular code and $w \in X^*$ a word such that $|w|_X \geqslant lg(X)$. Then any X-interpretation of w is adjacent to $(\varepsilon, w, \varepsilon)$.*

Proof of Proposition 2.12.

- Assume that condition 1 holds. Since X is a f.i.d. code:
  - If $X \cap \overline{S(X)}.X^+ \neq \emptyset$ then we have $X \cap X.X^+ \neq \emptyset$ (otherwise, we have $X \cap (S(X) \setminus X^*).X^+ \neq \emptyset$, which contradicts the fact that X is a f.i.d. code).
  - In a similar way, if $X \cap X^+.\overline{P(X)} \neq \emptyset$ then we obtain $X \cap X^+.X \neq \emptyset$.

  Since X is a code, we have $X \cap \overline{S(X)}.X^+ = \emptyset$ and $X \cap X^+.\overline{P(X)} = \emptyset$.

- We show that condition 2 implies condition 1.

Assume that X is a finite circular code satisfying condition 2. If X has no finite interpreting delay, there exist $n \geqslant lg(X)$, $m \geqslant 0$, $x_1, \ldots, x_m \in X$, $y_1, \ldots, y_n \in X$ and $(\alpha, \beta) \in P(X) \times S(X)$, with $(\alpha, \beta) \notin X^* \times X^*$, such that $\beta x_1 \ldots x_m \alpha = y_1 \cdots y_n$.

According to Proposition 2.14 $(\beta, x_1 \ldots x_m, \alpha)$ is an X-interpretation of the word $y_1 \ldots y_n$ which is adjacent to $(\varepsilon, y_1 \ldots y_n, \varepsilon)$. Let k, l be integers such that $\beta x_1 \ldots x_k = y_1 \ldots y_l$ (we have $x_{k+1} \ldots x_m \alpha = y_{l+1} \cdots y_n$).

If $\alpha \notin X^*$ then there exists a unique word $\alpha' \in X^* P(X)$ such that $x_{k'}\alpha' = y_{l'}$ or $x_{k'} = y_{l'}\alpha'$ with $k' > k$ and $l' > l$ (see Fig. 4). Clearly this contradicts $X \cap X^+.\overline{P(X)} = \emptyset$.

In a similar way, if $\beta \notin X^*$ then we get a contradiction with $X \cap \overline{S(X)}.X^+ = \emptyset$.

As a consequence, if X satisfies condition 2, it satisfies 1.

Consequently conditions 1 and 2 are equivalent. This completes the proof of Proposition 2.12. □

Example 2.15. In [1] the authors exhibit a circular code. This 20 elements code has been introduced for representing trinucleotids coding proteins:

AAC	AAT	ACC	ATC	ATT
CAG	CTC	CTG	GAA	GAC
GAG	GAT	GCC	GGC	GGT
GTA	GTC	GTT	TAC	TTC

Since all the words of this code have a length 3, this code satisfies condition 2. Hence it has a finite interpreting delay – in fact the interpreting delay is 4.

3. The defect theorem for f.i.d. codes

3.1. Adjacent codes

Given $w \in \Sigma^*$ and given a code X, if X is an adjacent code then all the X-interpretations of w are pairwise disjoint.

According to Proposition 2.5 and Proposition 2.12, we obtain a characterization of finite f.i.d. codes:

Theorem 3.1. *Let X be a finite subset of Σ^*. Then X is a f.i.d. code iff X is both circular and adjacent.*

The following example shows that the finiteness of the sets is required:

Example 3.2. The rational code $a + bc + ce + da^*b$ is circular and adjacent. However it is not f.i.d. since for any $n \in \mathbb{N}$, $(a^n b, \varepsilon, c)$ is an interpretation of the word $a^n bc \in X^{n+1}$.

Remark 3.3. In fact we can prove that a rational code X is a f.i.d. code iff it is a both circular and adjacent and if it exists an integer n such that for any $(s,p) \in \overline{S(X)} \times \overline{P(X)}$ satisfying $sX^*p \cap X \neq \emptyset$ we have $X^n s \cap S(X) = pX^n \cap P(X) = \emptyset$. The idea of the proof is that any X-interpretation of a long enough word w over X induces a circular one for an X-factor of w.

Our main result consists in a defect theorem for f.i.d. monoids. Before to proceed to its proof, we need to establish the following lemma, which will be of a common use.

Lemma 3.4. *Let $X \subset \Sigma^*$. For all $n, m \geqslant 1$, $x_1, \ldots, x_n \in X$, $y_1, \ldots, y_m \in X$ and $(\alpha, \beta) \in P(X) \times S(X)$, the two following conditions hold:*

(*i*) *If $X \cap X^+.\overline{P(X)} = \emptyset$ then the equality*

$$x_1 \ldots x_n \alpha = y_1 \ldots y_m$$

implies

$$m \geqslant n+1,\ x_i = y_i\ (1 \leqslant i \leqslant n),\ \alpha = y_{n+1} \cdots y_m.$$

(ii) If $X \cap \overline{S(X)}.X^+ = \emptyset$ then the equality

$$\beta x_1 \dots x_n = y_1 \dots y_m$$

implies

$$m \geqslant n+1,\ x_{i-m+n} = y_i\ (m-n+1 \leqslant i \leqslant m),\ \beta = y_1 \dots y_{m-n}.$$

Proof. By considering the reversed words, property (i) of Lemma 3.4 and property (ii) of Lemma 3.4 are equivalent. We shall establish property (i) of Lemma 3.4.

Assume that there exists a (smallest) integer i such that $x_i \neq y_i$.

If $|y_i| > |x_i|$, we have $y_i \in x_i.\overline{P(X^+)}$ thus $y_i \in X.\overline{P(X^+)}$, which contradicts $X \cap X^+.\overline{P(X)} = \emptyset$. If $|x_i| > |y_i|$, we have $x_i \in X.\overline{P(X^+)}$ and we hold a similar contradiction. As a consequence, we have $x_i = y_i$ for $1 \leqslant i \leqslant n$ and $m \geqslant n+1$. It follows that $\alpha = y_{n+1} \dots y_m$. Thus we have proved property (i) of Lemma 3.4. □

In fact, we first establish the defect theorem for adjacent codes. In a classical way, in order to prove that adjacent codes satisfy the defect theorem, we first prove that they are stable by intersections:

Proposition 3.5. *The intersection of an arbitrary family of submonoids generated by adjacent codes is generated by an adjacent code.*

Proof. Let $(X_i)_{i\in I}$ be a family of adjacent codes. We denote by Y the base of the free submonoid $\bigcap_{i\in I} X_i^*$. We shall prove that the code Y is adjacent.

Let us assume that $Y \cap Y^+.\overline{P(Y)} \neq \emptyset$. Let $y \in Y$ such that $y = y_1 \dots y_n \alpha$ with $n > 0$, $y_1, \dots, y_n \in Y$, $\alpha \in \overline{P(Y)}$.

Let $i > 0$. By definition of Y, we have $y, y_1, \dots, y_n \in X_i^+$ and $\alpha \in P(X_i^*)$. Hence:

$$y = x_{i,1} \dots x_{i,m_i},\ y_1 \dots y_n = x'_{i,1} \dots x'_{i,k_i},\ \alpha = \alpha_{i,1} \dots \alpha_{i,h_i} \alpha_i,$$

with $m_i, k_i > 0$, $h_i \geqslant 0$, $x_{i,1}, \dots, x_{i,m_i}, x'_{i,1}, \dots x'_{i,k_i} \in X_i$, $\alpha_{i,1}, \dots, \alpha_{i,h_i} \in X_i$ and $\alpha_i \in P(X_i) \setminus X_i^+$.

From $y = y_1 \dots y_n \alpha$, it follows that

$$x_{i,1} \dots x_{i,m_i} = x'_{i,1} \dots x'_{i,k_i} \alpha_{i,1} \dots \alpha_{i,h_i} \alpha_i.$$

Since the code X_i is adjacent, it satisfies the condition (i) of Lemma 3.4 therefore $\alpha_i = \varepsilon$, consequently $\alpha \in X_i^*$.

For all i, we have $\alpha \in X_i^*$, hence $\alpha \in Y^*$. But Y is a code, hence the equation $y = y_1 \dots y_n \alpha$ implies $n = 1$ and $\alpha = \varepsilon$, which contradicts $\alpha \in \overline{P(Y)}$. We have $Y \cap Y^+.\overline{P(Y)} = \emptyset$.

In a symmetrical way, we have $Y \cap \overline{S(Y)}.Y^+ = \emptyset$.

As a consequence, the intersection of $(X_i^*)_{i\in I}$ is generated by an adjacent code. □

We say that a submonoid generated by an adjacent code is an *adjacent submonoid.* Proposition 3.5 leads to introduce the notion of adjacent hull:

Definition 3.6. The adjacent hull of a subset $X \in \Sigma^*$ is the smallest adjacent submonoid containing X.

We prove the defect theorem for adjacent codes:

Theorem 3.7. *The base Y of the adjacent hull of a finite set X satisfies* $|Y| \leqslant |X|$.

Proof. As in [2] (p. 49) we prove that each word in Y appears as a suffix of some word in X.

Assume that there exists a word $y \in Y$ which is not in $(A^*)^{-1}.X$. Let $Z = y^*(Y - y)$.

We shall prove that Z is an adjacent set.

Assume first that $Z \cap Z^+.\overline{P(Z)} \neq \emptyset$. There exist $n \geqslant 1$, $z, z_1, \ldots, z_n \in Z$, $z' \in \overline{P(Z)}$ such that $z = z_1 \ldots z_n z'$.

By definition of Z, there exist $i, i_1, \ldots, i_n, i' \in \mathbb{N}$, $y', y_1, \ldots, y_n \in Y - y$, $y'' \in P(Y)$ such that $z = y^i y'$, $z_j = y^{i_j} y_j$ for $1 \leqslant j \leqslant n$ and $z' = y^{i'} y''$. Hence we hold

$$y^i y' = y^{i_1} y_1 \ldots y^{i_n} y_n y^{i'} y''. \tag{2}$$

By Lemma 3.4, we hold $i > i_1$ and $y_1 = y$, which is in contradiction with the definition of y_1. Hence $Z \cap Z^+.\overline{P(Z)} = \emptyset$.

In a similar way $Z \cap \overline{S(Z)}.Z^+ = \emptyset$.

Consequently Z is an adjacent set and we have $X \subset Z^* \subsetneq Y^*$ which yields a contradiction.

Therefore each word in Y appears as the suffix of some word in X, hence $|Y| \leqslant |X|$. □

Example 3.8. The base of the adjacent hull of the suffix code $\{abc, ab, cd\}$ is $\{ab, cd, c\}$.

3.2. Defect theorem and f.i.d. codes

We shall prove that Theorem 3.1 allows a version of the defect theorem for f.i.d. codes to be established:

We prove first, as in [3], that for any finite subset $X \subset \Sigma^*$, there exists a smallest f.i.d. code whose star contains X.

Indeed, given $X \subset \Sigma^*$, let $\mathcal{I}$ be the set of f.i.d. submonoids $Y^* \subset \Sigma^*$ such that $X \subset Y^*$ and $Y \subset F(X)$. Clearly, $\mathcal{I}$ is non-empty ($\Sigma^* \in \mathcal{I}$).

By Theorem 3.1, any $Y \in \mathcal{I}$ is circular and adjacent. Let Z be the base of the intersection of all these Y^*. Then by Proposition 3.5, Z is adjacent. Moreover since the intersection of an arbitrary family of submonoids generated by circular codes is generated by a circular code ([6], p. 145), the set Z is circular. As Z is included in $F(X)$, Z is finite and thus f.i.d.

As a consequence, there exists a smallest element Y^* in $\mathcal{I}$. For all f.i.d. submonoid Q containing X, we have $Q \cap F(X)^* \in \mathcal{I}$, hence $Y^* \subset Q$. Thus, Y^* is the smallest f.i.d. monoid containing X.

Hence, we can define the f.i.d. hull of a set:

Definition 3.9. The f.i.d. hull of a subset $X \in \Sigma^*$ is the smallest f.i.d. submonoid containing X.

Theorem 3.10. *The base Y of the f.i.d. hull of a finite set X satisfies $|Y| \leqslant |X|$.*

Proof. As in [3], we define the sequence (Z_n) by:

$$Z_0 = X,$$

$$Z_{n+1} = \begin{cases} \text{the base of the circular hull of } Z_n \text{ if } n \text{ is even,} \\ \text{the base of the adjacent hull of } Z_n \text{ if } n \text{ is odd.} \end{cases}$$

Since each Z_i is included in $F(X)$, the preceding sequence is stable from some index. According to Proposition 2.12, if $Z_n = Z_{n+1}$, then Z_n itself has a finite interpreting delay (because it is a circular and adjacent finite code). Moreover Z_n^* is the smallest submonoid which contains X (indeed each Z_i must be included in Y^*, which is circular and adjacent). □

References

[1] D. Arques and C.J. Michel, A possible code in the genetic code, edited by E.W. Mayr and C. Puech, *12th Annual Symposium on Theoretical Aspects of Computer Science.* Springer, *Lectures Notes in Comput. Sci.* **900** (1995) 640–651.

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

[3] J. Berstel, D. Perrin, J.F. Perrot and A. Restivo, Sur le théorème du défaut. *J. Algebra* **60** (1979) 169–180.

[4] V. Bruyère, Maximal codes with bounded deciphering delay. *Theoret. Comput. Sci.* **84** (1991) 53–76.

[5] S.W. Golomb and B. Gordon, Codes with bounded synchronization delay. *Inform. and Control* **8** (1965) 355–372.

[6] M. Leconte, *Codes sans répétition.* Ph.D. Thesis, Université Paris VII (1985).

[7] A. Restivo, A combinatorial property of codes having finite synchronization delay. *Theoret. Comput. Sci.* **1** (1975) 95–101.

[8] A.A. Sardinas and C. Patterson, A necessary and sufficient condition for the unique decomposition of coded messages. *IRE Internat. Conv. Rec.* **8** (1953) 104–108.

[9] J.C. Spehner, Quelques constructions et algorithmes relatifs aux sous-monoïdes d'un monoïde libre. *Semigroup Forum* **9** (1975) 334–353.

Communicated by Ch. Choffrut.
Received March, 1999. Accepted October, 1999.