Bifix codes and Sturmian words

Jean Berstel, Clelia de Felice, Dominique Perrin, Christophe Reutenauer, Giuseppina Rindone

Bifix codes and Sturmian words

Jean Berstel[1], Clelia De Felice[2], Dominique Perrin[1],
Christophe Reutenauer[3], Giuseppina Rindone[1]

[1]Université Paris Est, [2]Università degli Studi di Salerno,
[3] Université du Québec à Montréal

January 13, 2011 18 h 19

Abstract

We study bifix codes in factorial sets of words. We generalize most properties of ordinary maximal bifix codes to bifix codes maximal in a recurrent set F of words (F-maximal bifix codes). In the case of bifix codes contained in Sturmian sets of words, we obtain several new results. Let F be a Sturmian set of words, defined as the set of factors of a strict episturmian word. Our results express the fact that an F-maximal bifix code of degree d behaves just as the set of words of F of length d. An F-maximal bifix code of degree d in a Sturmian set of words on an alphabet with k letters has $(k-1)d+1$ elements. This generalizes the fact that a Sturmian set contains $(k-1)d+1$ words of length d. Moreover, given an infinite word x, if there is a finite maximal bifix code X of degree d such that x has at most d factors of length d in X, then x is ultimately periodic. Our main result states that any F-maximal bifix code of degree d on the alphabet A is the basis of a subgroup of index d of the free group on A.

Contents

1 Introduction **2**

2 Factorial sets **5**
2.1 Recurrent sets 5
2.2 Recurrent words 6
2.3 Episturmian words 7
2.4 Probability distributions 10

3 Prefix codes in factorial sets **13**
3.1 Prefix codes 14
3.2 Automata 15
3.3 Maximal prefix codes 16
3.4 Average length 18

4 Bifix codes in recurrent sets **19**
4.1 Parses 19
4.2 Maximal bifix codes 21
4.3 Derivation 28
4.4 Finite maximal bifix codes 33

5 Bifix codes in Sturmian sets **37**
5.1 Sturmian sets 37
5.2 Cardinality 39
5.3 Periodicity 44

6 Bases of subgroups **47**
6.1 Group automata 48
6.2 Main result 51
6.3 Incidence graph 53
6.4 Coset automaton 58
6.5 Return words 59
6.6 Proof of the main result 60

7 Syntactic groups **64**
7.1 Preliminaries 64
7.2 Group of a bifix code 68

1 Introduction

This paper studies a new relation between two objects previously unrelated: bifix codes and Sturmian words. We first give some elements on the background of both.

The study of bifix codes goes back to founding papers by Schützenberger [52] and by Gilbert and Moore [25]. These papers already contain significant results. The first systematic study is in the papers of Schützenberger [53, 54]. The general idea is that the submonoids generated by bifix codes are an adequate generalization of the subgroups of a group. This is illustrated by the striking fact that, under a mild restriction, the average length of a maximal bifix code with respect to a Bernoulli distribution on the alphabet is an integer. Thus, in some sense a maximal bifix code behaves as the uniform code formed of all the words of a given length. The theory of bifix codes was developed in a considerable way by Césari. He proved that all the finite maximal bifix codes may be obtained by internal transformations from uniform codes [9]. He also defined the notion of derived code which allows to build maximal bifix codes by increasing degrees [10].

Sturmian words are infinite words over a binary alphabet that have exactly $n + 1$ factors of length n for each $n \geq 0$. Their origin can be traced back to the astronomer J. Bernoulli III. Their first in-depth study is by Morse and Hedlund [41]. Many combinatorial properties were described in the paper by

Coven and Hedlund [15]. Note that, although Sturmian words appear first in the work of Morse and Hedlund, their finitary version, Christoffel and standard words, appear much before in the work of Christoffel [12] and, apparently independently, in the work of Markoff [38, 39]; the latter constructed the famous Markoff numbers by using them. The Markoff theory (which was designed to study minima's of quadratic forms) was revisited often by mathematicians, notably by Frobenius [24], Dickson [20], H. Cohn [13], Cusick and Flahive [17] and Bombieri [7]. There, the connection with the free group on two generators was established. Other connection of Christoffel words with the free group may be found in Osborne and Zieschang [42] and Kassel and Reutenauer [34]. Moreover, the Sturmian morphisms (substitutions that preserve Sturmian words) are the positive endomorphisms of the free group on two generators, see Wen and Wen [57], Mignosi and Séébold [40]. Thus Sturmian words are closely related to the free group. This connection is one of the main points of the present paper.

Sturmian words were generalized to arbitrary alphabets. Following an initial work by Arnoux and Rauzy [2] and developing ideas of De Luca [19], Droubay, Justin and Pirillo introduced in [22] the notion of episturmian words which generalizes Sturmian words to arbitrary finite alphabets.

In this paper, we consider the extension of the results known for bifix codes maximal in the free monoid to bifix codes maximal in more restricted sets of words, and in particular the sets of factors of episturmian words.

We extend most properties of ordinary maximal bifix codes to bifix codes that are maximal in a recurrent set F of words (F-maximal bifix codes). We show in particular that the average length of a finite F-maximal bifix code of degree d in a recurrent set F with respect to an invariant probability distribution on F is equal to d (Corollary 4.3.8).

Our main objective is the case of the set of factors of an episturmian word. We actually work with the set of factors of a strict episturmian word, called simply a Sturmian set. The number of factors of length d of a strict episturmian word over an alphabet of k letters is known to be $(k-1)d+1$. Our main result is that a maximal bifix code of degree d in a Sturmian set over an alphabet of k letters is always a basis of a subgroup of index d of the free group (Theorem 6.2.1). In particular, it has $(k-1)d+1$ elements (Theorem 5.2.1). Since the set of all words of length d is a maximal bifix code of degree d, this yields a strong generalization of the previous property. In particular, every finite maximal bifix code of degree d over a two letter alphabet contains exactly $d+1$ factors of any Sturmian word.

Finally, bifix codes X contained in restricted sets of words are used to study the groups in the syntactic monoid of the submonoid X^* (Theorem 7.2.3). This aspect was first considered by Schützenberger in [55]. He has studied the conditions under which parameters linked with the syntactic monoid M of a finitely generated submonoid X^* of a free monoid A^* can be bounded in terms of $\mathrm{Card}(X)$ only. One of his results is that, apart from a special case where the group is cyclic, the cardinality of a group contained in M is such a parameter. In [55], Schützenberger conjectured a refinement of his result which was subsequently proved by Césari. This study led to the Critical Factorization Theorem

that we will meet again here (Theorem 5.3.8).

The extension of the results concerning codes in free monoids to codes in a restricted set of words has already been considered by several authors. However, most of them have focused on general codes rather than on the particular class of bifix codes. In [49] the notion of codes of paths in a graph has been introduced. Such paths can also be viewed as words in a restricted set. The notion of a bifix code of paths has been studied in [18] where the internal transformation is generalized. In [47], the notion of code in a factorial set of words was introduced. The definition of a code X in a factorial set F requires that the set X^* of all concatenations of words in X is included in F. This approach was pushed further in [29]. A more general notion was considered in [4]. It only requires that $X \subset F$ and that no word of F has two distinct factorizations but not necessarily that $X^* \subset F$. The connection with unambiguous automata was considered later in [5]. Codes in Sturmian sets have been studied before in [8]. Finally, prefix codes X contained in restricted sets of words are used in [45] to study the groups in the syntactic monoid of the submonoid X^*.

Our paper is organized as follows.

In a first section (Section 2), we recall some definitions concerning prefix-closed, factorial, recurrent and uniformly recurrent sets, in relation with infinite words. We also introduce probability distributions on these sets.

In Section 3, we introduce prefix codes in factorial sets, especially maximal ones. We introduce some basic notions on automata. We define the average length with respect to a probability distribution on the factorial set.

In Section 4, we develop the theory of maximal bifix codes in recurrent sets. We generalize most of the properties known in the classical case. In particular, we show that the notion of degree and that of derived code can be defined (Theorem 4.3.1). We show that, for a uniformly recurrent set F, any F-thin bifix code contained in F is finite (Theorem 4.4.3). In the case of Sturmian sets, we prove our main results. First, a bifix code of degree d maximal in a Sturmian set on a k-letter alphabet has $(k-1)d+1$ elements (Theorem 5.2.1). Next, given an infinite word x, if there is a finite maximal bifix code X of degree d such that x has at most d factors of length d in X, then x is ultimately periodic (Corollary 5.3.3). The proof uses the Critical Factorization Theorem (see e.g. [35, 16]).

Section 6 presents our results concerning free groups. Our main result (Theorem 6.2.1) in this area states that for a Sturmian set F, a bifix code $X \subset F$ is a finite and F-maximal bifix code of F-degree d if and only if it is a basis of a subgroup of index d of the free group on A. We finally present in Section 7 a consequence of Theorem 6.2.1 concerning syntactic groups. We show that any transitive permutation group of degree d which can be generated by k elements is a syntactic group of a bifix code with $(k-1)d+1$ elements (Theorem 7.2.3).

Many results of this paper are extensions or generalizations of results contained in [6]. We always give the reference of the corresponding result in [6]. The proofs sometimes consist in the verification that the proof of the book still holds in the more general setting, and sometimes require new and more involved developments. In order to make the paper self contained, and to avoid repetitive

references to the book, we have tried to always give complete proofs.

2 Factorial sets

In this section, we introduce the basic notions of prefix-closed, factorial, recurrent and uniformly recurrent sets. These form a descending hierarchy. These notions are closely related with the analogous notions for infinite words which are defined in Section 2.2. In Section 2.4, we introduce probability distributions on factorial sets.

We use the standard terminology and notation on words, in particular concerning prefixes, suffixes and factors (see [35] for example). Let A be a finite alphabet. All words considered below are supposed to be on the alphabet A. We denote by 1 the empty word. We denote by A^* the set of all words on A and by A^+ the set of nonempty words.

The *reversal* of a word $w = a_1a_2\cdots a_n$, where $a_1, a_2, \ldots, a_n$ are letters, is the word $\widetilde{w} = a_n\cdots a_2a_1$. In particular, the reversal of the empty word is the empty word. A set X of words is *closed under reversal* if it contains the reversals of its elements.

Given a set X of words, we define, for a word u, the set $u^{-1}X$ by

$$u^{-1}X = \{y \in A^* \mid uy \in X\}\,.$$

Next, we say that a word is a *prefix of* X if it is a prefix of a word of X.

A nonempty set $F \subset A^*$ of words is said to be *prefix-closed* if it contains the prefixes of all its elements. Symmetrically, it is said to be *suffix-closed* if it contains the suffixes of all its elements. It is said to be *factorial* if it contains the factors of all its elements.

The *right* (resp. *left*) *order* of a word w with respect to F is the number of letters a such that $wa \in F$ (resp. $aw \in F$).

A set F is said to be *right essential* if it is prefix-closed and if any $w \in F$ has right order at least 1. If F is right essential, then for any $u \in F$ and any integer $n \geq 1$, there is a word v of length n such that $uv \in F$. Symmetrically, a set F is said to be *left essential* if it is suffix-closed and if any $w \in F$ has left order at least 1.

2.1 Recurrent sets

A set F of words is said to be *recurrent* if it is factorial and if for every $u, w \in F$ there is a $v \in F$ such that $uvw \in F$. A recurrent set $F \neq \{1\}$ is right and left essential.

Example 2.1.1 The set $F = A^*$ is recurrent.

Example 2.1.2 Let $A = \{a, b\}$. Let F be the set of words on A without factor bb. Thus $F = A^* \setminus A^*bbA^*$. The set F is recurrent. Indeed, if $u, w \in F$, then $uaw \in F$.

A set F is said to be *uniformly recurrent* if it is factorial and right essential and if, for any word $u \in F$, there exists an integer $n \geq 1$ such that u is a factor of every word in $F \cap A^n$.

Proposition 2.1.3 *A uniformly recurrent set is recurrent.*

Proof. Let $u, w \in F$. Let n be such that w is a factor of any word in $F \cap A^n$. Since F is right essential, there is a word v of length n such that $uv \in F$. Since w is a factor of v, we have $v = rws$ for some words r, s. Thus $urw \in F$. ∎

The converse of Proposition 2.1.3 is not true as shown in the example below.

Example 2.1.4 The set $F = A^*$ on $A = \{a, b\}$ is recurrent but not uniformly recurrent since $b \in F$ but b is not a factor of $a^n \in F$ for any $n \geq 1$.

2.2 Recurrent words

We denote by $F(x)$ the set of factors of an infinite word $x \in A^{\mathbb{N}}$. The set $F(x)$ is factorial and right essential.

An infinite word $x \in A^{\mathbb{N}}$ is said to be *recurrent* if for any word $u \in F(x)$ there is a $v \in F(x)$ such that $uvu \in F(x)$. Equivalently, each factor of a recurrent word x has an infinite number of occurrences in x.

Proposition 2.2.1 *For any recurrent set F there is an infinite word x such that $F(x) = F$.*

Proof. Set $F = \{u_1, u_2, \ldots\}$. Since F is recurrent and $u_1, u_2 \in F$, there is a word v_1 such that $u_1 v_1 u_2 \in F$. Further, since $u_1 v_1 u_2, u_3 \in F$ there is a word v_2 such that $u_1 v_1 u_2 v_2 u_3 \in F$. In this way, we obtain an infinite word $x = u_1 v_1 u_2 v_2 \cdots$ such that $F(x) = F$. ∎

Proposition 2.2.2 *For any infinite word x, the set $F(x)$ is recurrent if and only if x is recurrent.*

Proof. Set $F = F(x)$. Suppose first that F is recurrent. For any u in F, there is a $v \in F$ such that $uvu \in F$. Thus x is recurrent. Conversely, assume that x is recurrent. Let u, v be in F. Then there is a factorization $x = puy$ with $p \in F$ and $y \in A^{\mathbb{N}}$. Since x is recurrent, the word v is a factor of y. Set $y = qvz$ with $q \in F$ and $z \in A^{\mathbb{N}}$. Then uqv is in F. Thus F is recurrent. ∎

An infinite word $x \in A^{\mathbb{N}}$ is said to be *uniformly recurrent* if the set $F(x)$ is uniformly recurrent. There exist recurrent infinite words which are not uniformly recurrent, as shown in the following example.

Example 2.2.3 Let x be the infinite word obtained by concatenating all binary words in radix order: by increasing length, and for each length in lexicographic order. Thus, x starts as follows.

$$x = ab\,aaabbabb\,aaaaababaabbbaababbbabbb\cdots$$

The infinite word x is recurrent since every factor occurs infinitely often. However, x is not uniformly recurrent since each a^n, for $n > 1$, is a factor of x, thus two consecutive occurrences of say the letter b may be arbitrarily far one from each other. The word x is closely related to the Champernowne word [11].

We use indifferently the terms of *morphism* or *substitution* for a monoid morphism from A^* into itself. Let $f : A^* \to A^*$ be a morphism and assume there is a letter $a \in A$ such that $f(a) \in aA^+$. The words $f^n(a)$ for $n \geq 1$ are prefixes of one another. If $|f^n(a)| \to \infty$ with n, then we denote by $f^\omega(a)$ the infinite word which has all $f^n(a)$ as prefixes. It is called a *fix-point* of f.

Example 2.2.4 Set $A = \{a, b\}$. The *Thue–Morse morphism* is the substitution $f : A^* \to A^*$ defined by $f(a) = ab$ and $f(b) = ba$. The *Thue–Morse word* $x = abbabaab\cdots$ is the fix-point $f^\omega(a)$ of f. It is uniformly recurrent (see [36] Example 1.5.10). We call *Thue–Morse set* the set of factors of the Thue–Morse word.

An infinite word $x \in A^{\mathbb{N}}$ *avoids* a set X of words if $F(x) \cap X = \emptyset$. We denote by S_X the set of infinite words avoiding a set $X \subset A^*$. A (one sided) *shift space* is a set S of infinite words of the form S_X for some $X \subset A^*$.

A shift space $S \subset A^{\mathbb{N}}$ is *minimal* if for any shift space $T \subset S$, one has $T = \emptyset$ or $T = S$.

For any infinite word $x \in A^{\mathbb{N}}$, we denote by $S(x)$ the set of infinite words $y \in A^{\mathbb{N}}$ such that $F(y) \subset F(x)$. The set $S(x)$ is a shift space. Indeed, we have $y \in S(x)$ if and only if $F(y) \subset F(x)$ or equivalently $F(y) \cap X = \emptyset$ for $X = A^* \setminus F(x)$.

The following property is standard (see for example [36] Theorem 1.5.9).

Proposition 2.2.5 *An infinite word $x \in A^{\mathbb{N}}$ is uniformly recurrent if and only if $S(x)$ is minimal.*

2.3 Episturmian words

A *Sturmian word* is an infinite word x on a binary alphabet A such that the set $F(x) \cap A^n$ has $n + 1$ elements for any $n \geq 0$.

Example 2.3.1 Set $A = \{a, b\}$. The *Fibonacci morphism* is the substitution $f : A^* \to A^*$ defined by $f(a) = ab$ and $f(b) = a$. The *Fibonacci word*

$$x = abaababaabaababaababaabaababaabaab\cdots$$

is the fix-point $f^\omega(a)$ of f. It is a Sturmian word (see [36] Example 2.1.1). We call *Fibonacci set* the set of factors of the Fibonacci word.

Episturmian words are an extension of Sturmian words to arbitrary finite alphabets.

Recall that, given a set F of words over an alphabet A, the right (resp. left) order of a word u in F is the number of letters a such that $ua \in F$ (resp. $au \in F$). A word u is *right-special* (resp. *left-special*) if its right order (resp. left order) is at least 2. A right-special (resp. left-special) word is *strict* if its right (resp. left) order is equal to $\mathrm{Card}(A)$. In the case of a 2-letter alphabet, all special words are strict.

By definition, an infinite word x is *episturmian* if $F(x)$ is closed under reversal and if $F(x)$ contains, for each $n \geq 1$, at most one word of length n which is right-special.

Since $F(x)$ is closed under reversal, the reversal of a right-special factor of length n is left-special, and it is the only left-special factor of length n of x. A suffix of a right-special factor is again right-special. Symmetrically, a prefix of a left-special factor is again left-special.

As a particular case, a *strict* episturmian word is an episturmian word x with the two following properties: x has exactly one right-special factor of each length and moreover each right-special factor u of x is strict, that is satisfies the inclusion $uA \subset F(x)$ (see [22]).

It is easy to see that for a strict episturmian word x on an alphabet A with k letters, the set $F(x) \cap A^n$ has $(k-1)n+1$ elements for each n. Thus, for a binary alphabet, the strict episturmian words are just the Sturmian words, since a Sturmian word has one right-special factor for each length and its set of factors is closed under reversal.

An episturmian word s is called *standard* if all its left-special factors are prefixes of s. For any episturmian word s, there is a standard one t such that $F(s) = F(t)$. This is a rephrasing of Theorem 5 in [22].

Example 2.3.2 Consider the following generalization of the Fibonacci word to the ternary alphabet $A = \{a, b, c\}$. Consider the morphism $f : A^* \to A^*$ defined by $f(a) = ab$, $f(b) = ac$ and $f(c) = a$. The fix-point

$$f^\omega(a) = abacabaabacababacabaabacabacabaabacab\cdots$$

is the *Tribonacci word*. It is a strict standard episturmian word (see [32]).

The following is, in the case of Sturmian words, Proposition 2.1.25 in [36]. The general case results from Theorems 2 and 5 in [22].

Proposition 2.3.3 *An episturmian word x is uniformly recurrent and $S(x)$ is minimal.*

The converse is false as shown by the following example.

Example 2.3.4 The Thue–Morse word of Example 2.2.4 is not Sturmian. Indeed, it has four factors of length 2.

We recall now some notions and properties concerning episturmian words. A detailed exposition with proofs is given in [32, 22, 30, 31]. See also the survey paper [26]. For $a \in A$, denote by ψ_a the morphism of A^* into itself, called *elementary morphism*, defined by

$$\psi_a(b) = \begin{cases} ab & \text{if } b \neq a \\ a & \text{otherwise} \end{cases}$$

Let $\psi : A^* \to \mathrm{End}(A^*)$ be the morphism from A^* into the monoid of endomorphisms of A^* which maps each $a \in A$ to ψ_a. For $u \in A^*$, we denote by ψ_u the image of u by the morphism ψ. Thus, for three words u, v, w, we have $\psi_{uv}(w) = \psi_u(\psi_v(w))$.

A *palindrome* is a word w which is equal to its reversal. Given a word w, we denote by $w^{(+)}$ the *palindromic closure* of w. It is, by definition, the shortest palindrome which has w as a prefix.

The *iterated palindromic closure* of a word w is the word $\mathrm{Pal}(w)$ defined recursively as follows. One has $\mathrm{Pal}(1) = 1$ and for $u \in A^*$ and $a \in A$, one has $\mathrm{Pal}(ua) = (\mathrm{Pal}(u)a)^{(+)}$. Since $\mathrm{Pal}(u)$ is a proper prefix of $\mathrm{Pal}(ua)$, it makes sense to define the iterated palindromic closure of an infinite word x as the infinite word which is the limit of the iterated palindromic closure of the prefixes of x.

Justin's Formula is the following. For every words u and v, one has

$$\mathrm{Pal}(uv) = \psi_u(\mathrm{Pal}(v))\,\mathrm{Pal}(u)\,.$$

This formula extends to infinite words: if u is a word and v is an infinite word, then

$$\mathrm{Pal}(uv) = \psi_u(\mathrm{Pal}(v))\,. \tag{2.1}$$

There is a precise combinatorial description of standard episturmian words (see e.g. [32, 26]).

Theorem 2.3.5 *An infinite word s is a standard episturmian word if and only if there exists an infinite word $\Delta = a_0 a_1 \cdots$, where the a_n are letters, such that*

$$s = \lim_{n\to\infty} u_n\,,$$

where the sequence (u_n) is defined by $u_n = \mathrm{Pal}(a_0 a_1 \cdots a_{n-1})$. Moreover, the word s is episturmian strict if and only if every letter appears infinitely often in Δ.

The infinite word Δ is called the *directive word* of the standard word s. The description of the infinite word s can be rephrased by the equation

$$s = \mathrm{Pal}(\Delta)\,.$$

As a particular case of Justin's Formula, one has

$$u_{n+1} = \psi_{a_0\cdots a_{n-1}}(a_n)u_n\,. \tag{2.2}$$

The words u_n are the only prefixes of s which are palindromes.

Example 2.3.6 The Fibonacci word x of Example 2.3.1 is a standard episturmian word. It has the directive word $(ab)^\omega$, that is $x = \mathrm{Pal}((ab)^\omega)$ [26]. The Tribonacci word of Example 2.3.2 has the directive word $\Delta = (abc)^\omega$ [32]. The corresponding sequence (u_n) starts with $u_1 = a$, $u_2 = aba$, $u_3 = abacaba$. Observe that $\psi_{ab}(c) = abac$, so that indeed $u_3 = abacu_2$, as claimed in (2.2).

Example 2.3.7 Let $A = \{a, b, c\}$ and $\Delta = c(ab)^\omega$. Then, we have $u_1 = c$, $u_2 = cac$, $u_3 = cacbcac$, $u_4 = cacbcacacbcac$. By Justin's Formula 2.1, the limit is the word $x = \psi_c(y)$, where $y = \mathrm{Pal}((ab)^\omega)$ is the Fibonacci word on $\{a, b\}$. This means that x is obtained from y by inserting a letter c before every letter of y. The word x is not strict. Indeed, the letters a and b are not right-special and the letter c is not strict right special since cc is not a factor.

Example 2.3.8 Let $A = \{a, b, c\}$ and $\Delta = abc^\omega$. It is easily checked that $\mathrm{Pal}(\Delta)$ is the periodic word $(abac)^\omega$. The only right-special factors of this word are 1 and a ([26]).

2.4 Probability distributions

Let $F \subset A^*$ be a prefix-closed set of words. For $w \in F$, denote by $S(w)$ the set $S(w) = \{a \in A \mid wa \in F\}$. A *right probability distribution* on F is a map $\delta : F \to [0, 1]$ such that

(i) $\delta(1) = 1$,

(ii) $\sum_{a \in S(w)} \delta(wa) = \delta(w)$, for any $w \in F$.

For a right probability distribution δ on F and a set $X \subset F$, we denote $\delta(X) = \sum_{x \in X} \delta(x)$. See [6] for the elementary properties of right probability distributions. Note in particular that for any $u \in F$ and $n \geq 0$, one has, as a consequence of condition (ii),

$$\delta(uA^n \cap F) = \delta(u). \tag{2.3}$$

In particular, if δ is a right probability distribution on F, then $\delta(F \cap A^n) = 1$ for all $n \geq 0$.

The distribution is said to be *positive* on F if $\delta(x) > 0$ for any $x \in F$.

Symmetrically, for a suffix-closed set F, a *left probability distribution* is a map $\delta : F \to [0, 1]$ satisfying condition (i) above and

(iii) $\sum_{a \in P(w)} \delta(aw) = \delta(w)$, for any $w \in F$,

with $P(w) = \{a \in A \mid aw \in F\}$.

When F is factorial, an *invariant probability distribution* is both a left and a right probability distribution.

Proposition 2.4.1 *For any right essential set F of words, there exists a positive right probability distribution δ on F.*

Proof. Consider the map $\delta : F \to [0,1]$ defined for $w = a_1 a_2 \cdots a_n$ by

$$\delta(w) = \frac{1}{d_0 d_1 \cdots d_{n-1}}$$

where $d_i = \mathrm{Card}(S(a_1 \cdots a_i))$ for $0 \leq i < n$. Since F is right essential, $d_i \neq 0$ for $0 \leq i < n$. By convention, $\delta(1) = 1$.

Let us verify that δ is a right probability distribution on F. Indeed, let $w = a_1 a_2 \cdots a_n$. The set $S(w)$ is nonempty. Let $a \in S(w)$, we have $\delta(wa) = 1/d_0 d_1 \cdots d_n$. Since $\mathrm{Card}(S(w)) = d_n$, we obtain that δ satisfies condition (ii) and thus it is a right probability distribution. It is clearly positive. ∎

We will now turn to the existence of positive invariant probability distributions.

A *topological dynamical system* is a pair (S, σ) of a compact metric space S and a continuous map σ from S into S. Any shift space S becomes a topological dynamical system when it is equipped with the *shift* map defined by $\sigma(x_0 x_1 \cdots) = x_1 x_2 \cdots$. Indeed, we consider $A^{\mathbb{N}}$ as a metric space for the distance defined for $x = x_0 x_1 \cdots$ and $y = y_0 y_1 \cdots$ by $d(x,y) = 0$ if $x = y$ and $d(x,y) = 2^{-n}$ where n is the least integer such that $x_n \neq y_n$ otherwise.

A subset T of a topological dynamical system (S, σ) is said to be *stable under* σ or *stable* for short if $\sigma(T) \subset T$. A stable subset is also called (topologically) *invariant*.

The following property is well-known (although usually stated for two sided-infinite words, see for example Proposition 1.5.1 in [36]).

Proposition 2.4.2 *The shift spaces are the stable and closed subsets of* $(A^{\mathbb{N}}, \sigma)$.

Proof. It is clear that a shift space is both closed and stable. Conversely, let $S \subset A^{\mathbb{N}}$ be closed and stable under the shift. Let X be the set of words which are not factors of words of S. Then $S = S_X$. Indeed, if $y \in S$, then $F(y) \cap X = \emptyset$ and thus $y \in S_X$. Conversely, let $y \in S_X$. Let w_n be the prefix of length n of y. Since $w_n \notin X$ there is an infinite word $y^{(n)} \in S$ such that $w_n \in F(y^{(n)})$. Since S is stable under the shift, we may assume that w_n is a prefix of $y^{(n)}$. The sequence $y^{(n)}$ converges to y. Since S is closed, this forces $y \in S$. ∎

Let S be a metric space. The family of *Borel subsets* of S is the smallest family $\mathcal{F}$ of subsets of S containing the open sets and closed under complement and countable union. A function μ from $\mathcal{F}$ to $\mathbb{R}$ is said to be *countably additive* if $\mu(\bigcup_{n \geq 0} X_n) = \sum_{n \geq 0} \mu(X_n)$ for any sequence (X_n) of pairwise disjoint Borel subsets of S. A *Borel probability measure* on S is a function μ from $\mathcal{F}$ into $[0,1]$ which is countably additive and such that $\mu(S) = 1$.

Let (S, σ) be a topological dynamical system. A Borel probability measure on S is said to be *invariant* if $\mu(\sigma^{-1}(B)) = \mu(B)$ for any $B \in \mathcal{F}$. Note that since σ is continuous, $\sigma^{-1}(B) \in \mathcal{F}$ and thus $\mu(\sigma^{-1}(B))$ is well defined.

The following result is from [46, Theorem 4.2].

Theorem 2.4.3 *For any topological dynamical system, there exist invariant Borel probability measures.*

A dynamical system (S, σ) is said to be *minimal* if the only closed stable subsets of S are S and $\emptyset$. Note that, by Proposition 2.4.2, this definition is consistent with the definition of a minimal shift space. A Borel probability measure μ on S is *positive* if $\mu(U) > 0$ for every nonempty open set $U \subset S$.

Proposition 2.4.4 *Any invariant Borel probability measure on a minimal topological dynamical system is positive.*

Proof. Let μ be an invariant Borel probability measure on the topological dynamical system (S, σ). Let $U \subset S$ be a nonempty open set. Let $Y = \cup_{n \geq 0} \sigma^{-n}(U)$ and $Z = S \setminus Y$. Since U is open and σ is continuous, each $\sigma^{-n}(U)$ is open. Thus Y is open and Z is closed. The set Z is also stable. Indeed, if for $z \in Z$ we had $\sigma(z) \notin Z$, then there would be an integer $n \geq 0$ such that $\sigma(z) \in \sigma^{-n}(U)$. Thus $z \in \sigma^{-n-1}(U) \subset Y$, a contradiction. Thus $\sigma(Z) \subset Z$. Since (S, σ) is minimal, this implies that $Z = \emptyset$ or $Z = S$. Since U is nonempty, we have $Z = \emptyset$ and thus $Y = S$. Since μ is invariant, we have $\mu(\sigma^{-1}(U)) = \mu(U)$ and thus $\mu(\sigma^{-n}(U)) = \mu(U)$ for all $n \geq 0$. Hence we cannot have $\mu(U) = 0$ since it would imply $\mu(S) \leq \sum_{n \geq 0} \mu(\sigma^{-n}(U)) = 0$, a contradiction since $\mu(S) = 1$. ∎

Corollary 2.4.5 *For any recurrent set F there exists an invariant probability distribution on F. When F is uniformly recurrent, such a distribution is positive.*

Proof. Let F be a recurrent set. By Proposition 2.2.1 there is a recurrent infinite word x such that $F(x) = F$, and if F is uniformly recurrent, then x is uniformly recurrent.

By Theorem 2.4.3 there is an invariant Borel probability measure μ on $S = S(x)$.

Let δ be the map from F to $[0, 1]$ defined by $\delta(w) = \mu(wA^{\mathbb{N}} \cap S)$. Let us verify that δ is an invariant probability distribution. Indeed, one has $\delta(1) = \mu(S) = 1$. Next, for $w \in F$

$$\sum_{a \in S(w)} \delta(wa) = \sum_{a \in S(w)} \mu(waA^{\mathbb{N}} \cap S) = \mu(wA^{\mathbb{N}} \cap S) = \delta(w).$$

In the same way

$$\sum_{a \in P(w)} \delta(aw) = \sum_{a \in P(w)} \mu(awA^{\mathbb{N}} \cap S) = \mu(\sigma^{-1}(wA^{\mathbb{N}} \cap S)) = \mu(wA^{\mathbb{N}} \cap S) = \delta(w).$$

If x is uniformly recurrent, by Proposition 2.2.5, the shift space $S = S(x)$ is minimal. By Proposition 2.4.4, the measure μ is positive. Since $wA^{\mathbb{N}} \cap S$ is a nonempty open set for any $w \in F$, we have $\delta(w) = \mu(wA^{\mathbb{N}} \cap S) > 0$ and thus δ is positive. ∎

In some cases, there exists a unique invariant probability distribution on the set F. A morphism $f : A^* \to A^*$ is *primitive* if there exists an integer k such that, for all $a, b \in A$, the letter b appears in $f^k(a)$. If f is a primitive morphism and if $f(a)$ starts with the letter a for some $a \in A$, then $x = f^\omega(a)$ is a fix-point of f and there is a unique invariant probability distribution δ_F on the set $F(x)$ ([46, Theorem 5.6]). Moreover, this distribution is positive. We illustrate this result by the following examples.

Example 2.4.6 Let F be the Fibonacci set (see Example 2.3.1). Since the morphism f defined by $f(a) = ab$ and $f(b) = a$ is primitive, there is a unique invariant probability distribution on F. Its values on the words of length at most 4 are shown on Figure 2.1 with $\lambda = (\sqrt{5} - 1)/2$. The values of δ_F can be

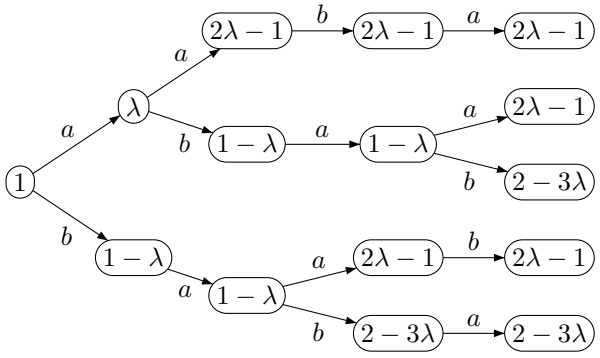

Figure 2.1: The invariant probability distribution on the Fibonacci set.

obtained as follows (see [46]). The vector $v = \begin{bmatrix}\delta_F(a) & \delta_F(b)\end{bmatrix}$ is an eigenvector for the eigenvalue $1/\lambda$ of the $A \times A$-matrix M defined by $M_{ab} = |f(a)|_b$. Here, we have

$$M = \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}$$

This implies $v = \begin{bmatrix}\lambda & 1-\lambda\end{bmatrix}$. The other values can be computed using conditions (ii) and (iii) of the definition of an invariant probability distribution.

Example 2.4.7 Let F be the Thue–Morse set (see Example 2.2.4). Since the Thue-Morse morphism is primitive, there is a unique invariant probability distribution on F. Its values on the words of length at most 4 are shown on Figure 2.2.

3 Prefix codes in factorial sets

In this section, we study prefix codes in a factorial set. We will see that most properties known in the usual case are also true in this more general situation.

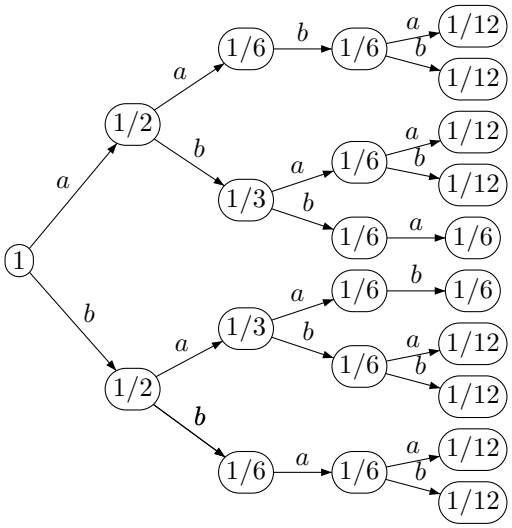

Figure 2.2: The invariant probability distribution on the Thue–Morse set.

Some of them are even true in the more general case of a prefix-closed set instead of a factorial set. In particular, this holds for the link between prefix codes and probability distributions (Proposition 3.3.4).

Recall that a set $X \subset A^+$ of nonempty words over an alphabet A is a *code* if the relation

$$x_1 \cdots x_n = y_1 \cdots y_m$$

with $n, m \geq 1$ and $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$. For the general theory of codes, see [6].

3.1 Prefix codes

The *prefix order* is defined, for $u, v \in A^*$, by $u \leq v$ if u is a prefix of v. Two words u, v are *prefix-comparable* if one is a prefix of the other. Thus u and v are prefix-comparable if and only if there are words x, y such that $ux = vy$ or, equivalently, if and only if $uA^* \cap vA^* \neq \emptyset$. The *suffix order*, and the notion of suffix-comparable words, are defined symmetrically.

A set $X \subset A^+$ of nonempty words is a *prefix code* if any two distinct elements of X are incomparable for the prefix order. A prefix code is a code.

The dual notion of a *suffix* code is defined symmetrically with respect to the suffix order.

The submonoid M generated by a prefix code satisfies the following property: if $u, uv \in M$ then $v \in M$. Such a submonoid of A^* is said to be *right unitary*. One can show that conversely, any right unitary submonoid of A^* is generated by a prefix code (see [6]). The symmetric notion of a *left unitary* submonoid is defined by the condition $v, uv \in M$ implies $u \in M$.

We denote by $\underline{X}$ the *characteristic series* of a set $X \subset A^*$. By definition,

for any $x \in A^*$,

$$(\underline{X}\,, x) = \begin{cases} 1 & \text{if } x \in X \\ 0 & \text{otherwise} \end{cases}$$

The following is Proposition 3.1.6 in [6].

Proposition 3.1.1 *Let X be a prefix code and let $U = A^* \setminus XA^*$. Then*

$$\underline{A^*} = \underline{X}^*\underline{U} \quad \textit{and} \quad \underline{X} - 1 = \underline{U}(\underline{A} - 1). \tag{3.1}$$

3.2 Automata

We recall the basic results on deterministic automata and prefix codes (see [6] for a more detailed exposition).

We denote $\mathcal{A} = (Q, i, T)$ a deterministic automaton with Q as set of states, $i \in Q$ as initial state and $T \subset Q$ as set of terminal states. For $p \in Q$ and $w \in A^*$, we denote $p \cdot w = q$ if there is a path labeled w from p to the state q and $p \cdot w = \emptyset$ otherwise.

The set *recognized* by the automaton is the set of words $w \in A^*$ such that $i \cdot w \in T$. A set of words is *rational* if is recognized by a finite automaton.

All automata considered in this paper are deterministic and we call them simply automata.

The automaton $\mathcal{A}$ is *trim* if for any $q \in Q$, there is a path from i to q and a path from q to some $t \in T$.

An automaton is called *simple* if it is trim and if it has a unique terminal state which coincides with the initial state.

An automaton $\mathcal{A} = (Q, i, T)$ is *complete* if for any state $p \in Q$ and any letter $a \in A$, one has $p \cdot a \neq \emptyset$.

For a set $X \subset A^*$, we denote by $\mathcal{A}(X)$ the *minimal automaton* of X. The states of $\mathcal{A}(X)$ are the nonempty sets $u^{-1}X = \{v \in A^* \mid uv \in X\}$ for $u \in A^*$. The initial state is the set X and the terminal states are the sets $u^{-1}X$ for $u \in X$.

Let $X \subset A^*$ be a prefix code. Then there is a simple automaton $\mathcal{A} = (Q, 1, 1)$ that recognizes X^*. Moreover, the minimal automaton of X^* is simple.

Let X be a prefix code and let P be the set of proper prefixes of X. The *literal automaton* of X^* is the simple automaton $\mathcal{A} = (P, 1, 1)$ with transitions defined for $p \in P$ and $a \in A$ by

$$p \cdot a = \begin{cases} pa & \text{if } pa \in P\,, \\ 1 & \text{if } pa \in X\,, \\ \emptyset & \text{otherwise.} \end{cases}$$

One verifies that this automaton recognizes X^*.

Let $\mathcal{A} = (Q, i, T)$ be an automaton. For $w \in A^*$, we denote $\varphi_{\mathcal{A}}(w)$ the partial map from Q to Q defined by $p\varphi_{\mathcal{A}}(w) = q$ if $p \cdot w = q$. The *transition monoid* of $\mathcal{A}$ is the monoid of partial maps from Q to Q of the form $\varphi_{\mathcal{A}}(w)$ for $w \in A^*$.

3.3 Maximal prefix codes

Let F be a subset of A^*. A set $X \subset A^*$ is *right dense* in $F \subset A^*$, or right F-dense, if any $u \in F$ is a prefix of X.

A set $X \subset F$ is *right complete* in F, or right F-complete, if X^* is right dense in F, that is if every word in F is a prefix of X^*.

A prefix code $X \subset F$ is *maximal* in F, or F-maximal, if it is not properly contained in any other prefix code $Y \subset F$. The notion of an F-maximal suffix code is symmetrical.

The following propositions are extensions of Propositions 3.3.1 and 3.3.2, and of Theorem 3.3.5 in [6].

Proposition 3.3.1 *Let F be a subset of A^*. For any prefix code $X \subset F$, the following conditions are equivalent.*

(i) *Every element of F is prefix-comparable with some element of X,*
(ii) *X is an F-maximal prefix code.*

Proof. (i) implies (ii). Any word $u \in F$ is prefix-comparable with some word of X. This implies that if $u \notin X$, then $X \cup u$ is no longer a prefix code. Thus X is an F-maximal prefix code.

(ii) implies (i). Assume that $u \in F$ is not prefix-comparable to any word in X. Then $X \cup u$ is prefix, and X is not an F-maximal prefix code. ∎

Proposition 3.3.2 *Let F be a factorial subset of A^*. For any set $X \subset F$ of nonempty words, the following conditions are equivalent.*

(i) *Every element of F is prefix-comparable with some element of X,*
(ii) *XA^* is right F-dense,*
(iii) *X is right F-complete.*

Proof. (i) implies (ii). Let $u \in F$. Let $x \in X$ be prefix-comparable with u. Then there exist v, w such that $uv = xw$. Thus XA^* is right F-dense.

(ii) implies (iii). Consider a word $u \in F$. Let us show that u is a prefix of X^*. Since XA^* is right F-dense, one has $uw = xw'$ for some word $x \in X$ and $w, w' \in A^*$. If u is a prefix of X, there is nothing to prove. Otherwise, x is a proper prefix of u. Thus $u = xu'$ for some $x \in X$ and $u' \in A^*$. Since u is in F and since F is factorial, we have $u' \in F$. Since $x \neq 1$, we have $|u'| < |u|$. Arguing by induction, the word u' is a prefix of X^*. Thus u is a prefix of X^*.

(iii) implies (i). Let $u \in F$. Then u is a prefix of X^*, and consequently u is prefix-comparable with a word in X.

∎

The propositions have a dual formulation, replacing prefix by suffix, and right by left.

Example 3.3.3 The set $X = \{a, ba\}$ is a maximal prefix code in the Fibonacci set F since XA^* is right F-dense.

The following is a generalization of Propositions 3.7.1 and 3.7.2 in [6].

Proposition 3.3.4 *Let F be a right essential set. Let δ be a positive right probability distribution on F. Any prefix code $X \subset F$ satisfies $\delta(X) \leq 1$. If X is finite, it is F-maximal if and only if $\delta(X) = 1$.*

Proof. Assume first that X is finite. Let n be the maximal length of the words in X. We have

$$\bigcup_{x \in X} xA^{n-|x|} \cap F \subset A^n \cap F \tag{3.2}$$

and the terms of the union are pairwise disjoint. Thus, using Equation (2.3)

$$\delta(X) = \sum_{x \in X} \delta(xA^{n-|x|} \cap F) \leq \delta(A^n \cap F) = 1\,. \tag{3.3}$$

If X is maximal in F, any word in $F \cap A^n$ has a prefix in X. Thus we have equality in (3.2) and thus also in (3.3). This shows that $\delta(X) = 1$. The converse is clear since δ is positive on F.

If X is infinite, then $\delta(Y) \leq 1$ for any finite subset Y of X. Thus $\delta(X) \leq 1$. ∎

The statement has a dual for a suffix code included in a factorial set F with a positive left probability distribution on F.

Example 3.3.5 Let F be the Fibonacci set. The set $X = \{a, ba\}$ is a maximal prefix code (Example 3.3.3). One has $\delta_F(X) = 1$ where δ_F is defined in Example 2.4.6.

We will use the following result in the proof of Proposition 4.4.5.

Proposition 3.3.6 *Let F be a right essential subset of A^*, and let $G \subset F$ be a right essential subset of F. For any finite F-maximal prefix code $X \subset F$, the set $X \cap G$ is a finite G-maximal prefix code.*

Proof. Set $Y = X \cap G$. The set Y is clearly a finite prefix code. We show that every $u \in G$ is prefix-comparable with some word in Y. This will imply that Y is G-maximal by Proposition 3.3.1. Let $u \in G$. Since G is right essential, there are arbitrary long words w such that $uw \in G$. Choose the length of uw larger than the maximal length of the words of X. Since X is an F-maximal prefix code, uw has a prefix x in X. This prefix x is in Y since $uw \in G$. Thus u is prefix-comparable to $x \in Y$. ∎

The following example shows that Proposition 3.3.6 is false for infinite prefix codes.

Example 3.3.7 Let $F \subset A^*$ be a right essential set with $F \neq A^*$, and let x be a word which is not in F. Let $X = A^*x \setminus A^*xA^+$ be the prefix code of words in A^* ending with x and having no other occurrence of x. X is a maximal prefix code, and $X \cap F = \emptyset$ is not F-maximal.

We will use later the following result on transformations of prefix codes. It is adapted from Proposition 3.4.9 in [6].

Proposition 3.3.8 *Let F be a factorial set and let $X \subset F$ be an F-maximal prefix code. Let w be a nonempty prefix of X and set $D = w^{-1}X$. The set $Y = (X \setminus wD) \cup w$ is an F-maximal prefix code.*

Proof. It is clear that Y is a prefix code. To show that it is F-maximal, we apply Proposition 3.3.1 and prove that every word $u \in F$ is prefix-comparable with a word of Y. So consider a word $u \in F$. Since X is F-maximal, u is prefix-comparable with a word of X. Thus u is prefix-comparable with a word of $X \setminus wD$ or it is prefix-comparable with a word of wD. In the second case, either u is a prefix of a word wd with $d \in D$ or u has wd as a prefix. Consequently, u is prefix-comparable with w. This proves that u is prefix-comparable with a word of Y. ∎

Proposition 3.3.8 has a dual formulation for suffix codes.

3.4 Average length

Let F be a right essential set and let δ be a right probability distribution on F. Let $X \subset F$ be a prefix code such that $\delta(X) = 1$. The *average length* of X with respect to δ is the sum

$$\lambda(X) = \sum_{x \in X} |x|\delta(x)\,.$$

Proposition 3.4.1 *Let F be a right essential set and let δ be a positive right probability distribution on F. Let $X \subset F$ be a finite F-maximal prefix code and let P be the set of proper prefixes of X. Then $\delta(X) = 1$ and $\lambda(X) = \delta(P)$.*

Proof. We already know that $\delta(X) = 1$ by Proposition 3.3.4. Let us show that for any $p \in P$,

$$\delta(p) = \sum_{x \in pA^+ \cap X} \delta(x)\,. \tag{3.4}$$

Let indeed n be an integer larger than the lengths of the words of X. Then by Equation (2.3), $\delta(p) = \delta(pA^n \cap F)$. Since X is an F-maximal prefix code, each word of $pA^n \cap F$ has a prefix in X, and conversely, each word in X which has p as a prefix is itself a prefix of $pA^n \cap F$. Thus

$$pA^n \cap F = \bigcup_{x \in pA^+ \cap X} xA^{n+|p|-|x|} \cap F\,.$$

Since $\delta(xA^{n+|p|-|x|} \cap F) = \delta(x)$, this proves Equation (3.4).

By Equation (3.4), one gets

$$\sum_{p \in P} \delta(p) = \sum_{p \in P} \sum_{x \in pA^+ \cap X} \delta(x) = \sum_{x \in X} \sum_{p \in P : x \in pA^+} \delta(x) = \sum_{x \in X} |x|\delta(x)\,.$$

Thus

$$\delta(P) = \sum_{p \in P} \delta(p) = \sum_{x \in X} |x| \delta(x) = \lambda(X)\,.$$

■

A dual statement of Proposition 3.4.1 holds for a suffix code and its set of proper suffixes, for a positive left probability distribution.

Example 3.4.2 Let F be the Fibonacci set and let $X = \{a, ba\}$. We have already seen in Example 3.3.5 that X is an F-maximal prefix code and that $\delta_F(X) = 1$ where δ_F is the unique invariant probability distribution on F defined in Example 2.4.6. We have $\lambda(X) = \lambda + 2(1-\lambda) = 2 - \lambda$. On the other hand the set of proper prefixes of X is $P = \{1, b\}$ and thus $\delta_F(P) = 1 + (1 - \lambda) = 2 - \lambda$.

4 Bifix codes in recurrent sets

In this section, we study bifix codes contained in a recurrent set. Since A^* itself is a recurrent set, it is a generalization of the usual situation. We will see that all results on maximal bifix codes can be generalized in this way. In particular, the notions of degree, of kernel and of derived code can be defined in this more general framework.

4.1 Parses

Recall that a set X of nonempty words is a *bifix code* if any two distinct elements of X are incomparable for the prefix order and for the suffix order.

A *parse* of a word w with respect to a set X is a triple (v, x, u) such that $w = vxu$ with $v \in A^* \setminus A^*X$, $x \in X^*$ and $u \in A^* \setminus XA^*$.

Proposition 4.1.1 *Let F be a factorial set and let $X \subset F$ be a set. For any factorization $w = uv$ of $w \in F$, there is a parse (s, yz, p) of w with $y, z \in X^*$, $sy = u$ and $v = zp$.*

Proof. Since $v \in F$, there exist, by Proposition 3.1.1, words $z \in X^*$ and $p \in A^* \setminus XA^*$ such that $v = zp$. Symmetrically, there exist $y \in X^*$ and $s \in A^* \setminus A^*X$ such that $u = sy$. Then (s, yz, p) is a parse of w which satisfies the conditions of the statement. ■

The number of parses of a word w with respect to X is denoted by $\pi_X(w)$. The function $\pi_X : A^* \to \mathbb{N}$ is the *parse enumerator* with respect to X.

The *indicator* of a set X is the series L_X defined for $w \in A^*$ by $(L_X, w) = \pi_X(w)$.

Example 4.1.2 Let $X = \emptyset$. Then $\pi_X(w) = |w| + 1$.

The following is a reformulation of Proposition 6.1.6 in [6].

Proposition 4.1.3 *Let F be a factorial set and let $X \subset F$ be a prefix code. For every word $w \in F$, the number $\pi_X(w)$ is equal to the number of prefixes of w which have no suffix in X.*

Proof. For every prefix v of w which is in $A^* \setminus A^*X$, there is a unique parse of w of the form (v, x, u). Since any parse is obtained in this way, the statement is proved. ∎

Proposition 4.1.3 has a dual statement for suffix codes.

Note that, as a consequence of Proposition 4.1.3, we have for two prefix codes X, Y, and for all words w,

$$X \subset Y \Rightarrow \pi_Y(w) \leq \pi_X(w). \tag{4.1}$$

Indeed, a word without suffix in Y is also a word without suffix in X.

Proposition 4.1.4 *Let X be a prefix code and let $V = A^* \setminus A^*X$. Then*

$$\underline{V} = L_X(1 - \underline{A})\,. \tag{4.2}$$

If X is bifix, one has

$$1 - \underline{X} = (1 - \underline{A})L_X(1 - \underline{A})\,. \tag{4.3}$$

Proof. Set $L = L_X$. Let $U = A^* \setminus XA^*$. By definition of the indicator, we have $L = \underline{V}\ \underline{X}^*\underline{U}$. Since X is prefix, we have by Proposition 3.1.1, the equality $\underline{A^*} = \underline{X}^*\underline{U}$. Thus we obtain $L = \underline{V}\,\underline{A^*}$ (note that this is actually equivalent to Proposition 4.1.3). Multiplying both sides on the right by $(1 - \underline{A})$, we obtain Equation (4.2).

If X is suffix, we have by the dual of Proposition 3.1.1, the equality $1 - \underline{X} = (1 - \underline{A})\underline{V}$. This gives Equation (4.3) by multiplying both sides of Equation (4.2) on the left by $1 - \underline{A}$. ∎

The following is Proposition 6.1.11 in [6].

Proposition 4.1.5 *A function $\pi : A^* \to \mathbb{N}$ is the parse enumerator of some bifix code if and only if it satisfies the following conditions.*

(i) *For any $a \in A$ and $w \in A^*$*

$$0 \leq \pi(aw) - \pi(w) \leq 1\,. \tag{4.4}$$

(ii) *For any $w \in A^*$ and $a \in A$*

$$0 \leq \pi(wa) - \pi(w) \leq 1\,. \tag{4.5}$$

(iii) *For any $a, b \in A$ and $w \in A^*$*

$$\pi(aw) + \pi(wb) \geq \pi(w) + \pi(awb)\,. \tag{4.6}$$

(iv) $\pi(1) = 1$.

The following is a reformulation of Proposition 6.1.12 in [6].

Proposition 4.1.6 *Let X be a prefix code. For any $u \in A^*$ and $a \in A$, one has*

$$\pi_X(ua) = \begin{cases} \pi_X(u) & \text{if } ua \in A^*X \\ \pi_X(u) + 1 & \text{otherwise} \end{cases} \tag{4.7}$$

Proof. This follows directly from Proposition 4.1.3. ∎

Proposition 4.1.6 has a dual for suffix codes expressing $\pi_X(au)$ in terms of $\pi_X(u)$.

Recall also that by Proposition 6.1.8 in [6], for a bifix code X and for all $u, v, w \in F$ such that $uvw \in F$, one has

$$\pi_X(v) \leq \pi_X(uvw). \tag{4.8}$$

Moreover, if $uvw \in X$ and $u, w \in A^+$ then the inequality is strict, that is,

$$\pi_X(v) < \pi_X(uvw). \tag{4.9}$$

4.2 Maximal bifix codes

Let F be set of words. A set $X \subset F$ is said to be *thin* in F, or F-thin, if there exists a word of F which is not a factor of a word in X.

The following example shows that there exist a uniformly recurrent set F, and a bifix code $X \subset F$ which is not F-thin.

Example 4.2.1 Let F be the Thue–Morse set, which is the set of factors of a fix-point of the substitution f defined by $f(a) = ab$, $f(b) = ba$ (see Example 2.2.4). Set $x_n = f^n(a)$ for $n \geq 1$. Note that $x_{n+1} = x_n\bar{x}_n$ where $u \to \bar{u}$ is the substitution defined by $\bar{a} = b$ and $\bar{b} = a$. Note also that $u \in F$ if and only if $\bar{u} \in F$. Consider the set $X = \{x_nx_n \mid n \geq 1\}$. We have $X \subset F$. Indeed, for $n \geq 1$, $x_{n+2} = x_{n+1}\bar{x}_{n+1} = x_n\bar{x}_n\bar{x}_nx_n$ implies that $\bar{x}_n\bar{x}_n \in F$ and thus $x_nx_n \in F$. Next X is a bifix code. Indeed, for $n < m$, x_m begins with $x_n\bar{x}_n$, and thus cannot have x_nx_n as a prefix. Similarly, since x_m ends with $\bar{x}_nx_n$ or with $x_n\bar{x}_n$, it cannot have x_nx_n as a suffix. Finally any element of F is a factor of a word in X. Indeed, any element u of F is a factor of some x_n, and thus of $x_nx_n \in X$.

A simpler proof uses Theorem 4.4.3 proved later.

An *internal factor* of a word x is a word v such that $x = uvw$ with u, w nonempty. Let $F \subset A^*$ be a factorial set and let $X \subset F$ be a set. Denote by

$$I(X) = \{w \in A^* \mid A^+wA^+ \cap X \neq \emptyset\}$$

the set of internal factors of words in X.[1]

When F is right essential and left essential, then X is F-thin if and only if $F \setminus I(X) \neq \emptyset$. Indeed, the condition is necessary. Conversely, if w is in $F \setminus I(X)$, let $a, b \in A$ be such that $awb \in F$. Since awb cannot be a factor of a word in X, it follows that X is F-thin.

We say that a bifix code $X \subset F$ is *maximal* in F, or F-maximal, if it is not properly contained in any other bifix code $Y \subset F$.

The following is a generalisation of Proposition 6.2.1 in [6].

Theorem 4.2.2 *Let F be a recurrent set and let $X \subset F$ be an F-thin set. The following conditions are equivalent.*

(i) *X is an F-maximal bifix code.*
(ii) *X is a left F-complete prefix code.*
(ii') *X is a right F-complete suffix code.*
(iii) *X is an F-maximal prefix code and an F-maximal suffix code.*

As a preparation for the proof of Theorem 4.2.2, we introduce the following notation. Let F be a recurrent set and let $X \subset F$.

A *factorization* of a word u is a pair (p, s) of words such that $u = ps$. We denote by $\mathrm{Fact}(u)$ the set of factorizations of u.

Let $C(X, F)$ be the set of pairs (u, v) of words such that $uvu \in F$, $v \neq 1$ and u is not an internal factor of X. We define for each pair $(u, v) \in C(X, F)$ a relation $\varphi_{u,v}$ on the set $\mathrm{Fact}(u)$ as follows. For $\pi = (p, s), \rho = (q, t) \in \mathrm{Fact}(u)$, one has $(\pi, \rho) \in \varphi_{u,v}$ if and only if the pair (π, ρ) satisfies one of the following conditions (see Figure 4.1).

(i) $px = q$ for some $x \in X$,
(ii) $svq = x_1 \cdots x_n$ with $n \geq 1$ and $x_i \in X$ for $1 \leq i \leq n$, s is a proper prefix of x_1 and q is a proper suffix of x_n.

Since $ps = qt$, the condition (i) is equivalent to $s = xt$. This means that both conditions are symmetric for reading from left to right or from right to left.

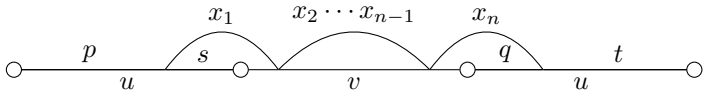

Figure 4.1: The relation $\varphi_{u,v}$ (case (ii)).

We prove a series of lemmas concerning the relations $\varphi_{u,v}$ (see Exercise 6.2.1 in [6]).

Lemma 4.2.3 *Let F be a recurrent set and let $X \subset F$ be an F-thin set. If X is a prefix code, then for all pairs $(u, v) \in C(X, F)$, the relation $\varphi_{u,v}$ is a partial function from* $\mathrm{Fact}(u)$ *into itself, that is*

$$(\pi, \rho), (\pi, \rho') \in \varphi_{u,v} \quad \Rightarrow \quad \rho = \rho' \,. \tag{4.10}$$

[1] The set $I(X)$ is denoted by $H(X)$ in [6].

Conversely, if X is an F-maximal suffix code, and if (4.10) *holds for all pairs $(u, v) \in C(X, F)$, then X is a prefix code.*

Define the *transpose* $\varphi'_{u,v}$ of the relation $\varphi_{u,v}$ by the condition $(\rho, \pi) \in \varphi'_{u,v}$ if $(\pi, \rho) \in \varphi_{u,v}$. Then (4.10) expresses the fact that the transpose $\varphi'_{u,v}$ is injective.

Proof. Assume first that X is a prefix code. For $(u, v) \in C(X, F)$, let $\pi = (p, s), \rho = (q, t), \rho' = (q', t')$ be three factorizations of u such that $(\pi, \rho), (\pi, \rho') \in \varphi_{u,v}$. We prove that $\rho = \rho'$. By definition, the following cases may occur for $(\pi, \rho), (\pi, \rho')$.

(1) $px = q$ and $px' = q'$, with $x, x' \in X$,
(2) $px = q$ with $x \in X$, and $svq' = x'_1 \cdots x'_m$, with $m \geq 1$ and $x'_1, \ldots, x'_m \in X$, and moreover s is a proper prefix of x'_1 and q' is a proper suffix of x'_m,
(3) $px' = q'$ with $x' \in X$ and $svq = x_1 \cdots x_n$, with $n \geq 1$ and $x_1, \ldots, x_n \in X$, and moreover s is a proper prefix of x_1 and q is a proper suffix of x_n,
(4) $svq = x_1 \cdots x_n$ and $svq' = x'_1 \cdots x'_m$, with $n \geq 1$, $m \geq 1$, $x_1, \ldots, x_n$, $x'_1, \ldots, x'_m \in X$, and moreover s is a proper prefix both of x_1 and of x'_1, q is a proper suffix of x_n and q' is a proper suffix of x'_m.

(1) Assume that $px = q$, $px' = q'$, with $x, x' \in X$. Since q and q' are prefixes of u, they are prefix-comparable. Thus x and x' are also prefix-comparable. Since X is a prefix code, it follows that $x = x'$, whence $q = q'$ and $\rho = \rho'$.

(2) We show that this case is impossible. Indeed, x is a prefix of s (by $ps = qt = pxt$) and s is a proper prefix of x'_1, thus x is a proper prefix of x'_1, and this is impossible because X is a prefix code. The same argument holds in the symmetric case (3).

(4) Since $u = qt = q't'$, the words q and q' are prefix-comparable. We may suppose that $q = q'w$ (see Figure 4.2). Since svq, svq' are in X^* and X is a prefix code, we have $w \in X^*$. Since X is a code, the decompositions $svq = x_1 \cdots x_n = svq'w = x'_1 \cdots x'_m w$ coincide. Consequently, $w = x_{m+1} \cdots x_n$. By hypothesis, $q = q'w = q'x_{m+1} \cdots x_n$ is a proper suffix of x_n. This forces $n = m$, $w = 1$ and $q = q'$, hence $\rho = \rho'$.

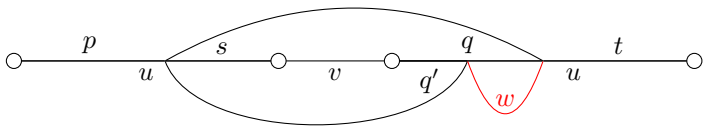

Figure 4.2: The factorizations (p, s), (q, t) and (q', t') with $t' = wt$.

Conversely, assume that X is an F-maximal suffix code and that it is not a prefix code. Let x', x'' be distinct words in X such that x' is a prefix of x''. Set $x'' = x'r'$ with $r' \neq 1$.

Since X is F-thin, there is a word $w \in F \setminus I(X)$. Since F is recurrent, there is a word r'' such that $x''r''w \in F$. Let $u = r'r''w$. Then $x''r''w = x'u \in F$. Let t be a word such that $utx'u \in F$. Set $v = tx'$. Thus $(u, v) \in C(X, F)$ (see

Figure 4.3). By the dual of Equation (3.1), there exist $p \in A^* \setminus A^*X$ and $z \in X^*$ such that $ut = pz$.

Since X is left F-complete, p is a proper suffix of a word in X. Since $u \notin I(X)$, p is a prefix of u. Thus $z = 1 = t$ or $z \in X^+$. In the latter case, set $z = z_1 \cdots z_n$ with $z_i \in X$. Since $ut = pz$, one of the following two cases holds:

(1) $u = pz'$, with $z', t \in X^*$,
(2) there is an i with $1 \leq i \leq n$ such that $z_i = rs$ with $u = pz_1 \cdots z_{i-1}r$, $t = sz_{i+1} \cdots z_n$, and $r \neq 1$, $s \neq 1$.

In case (1), consider the three factorizations $\pi = (u, 1)$, $\rho = (1, u)$, $\rho' = (r', r''w)$ of u. Since $r' \neq 1$, we have $\rho \neq \rho'$. We have $v = tx' \in X^+$, and thus $(\pi, \rho) \in \varphi_{u,v}$ (this is case (ii) of the definition with $s = q = 1$). Next, $vr' = tx'r' = tx'' \in X^+$, with $t \in X^*$ and where r' is a proper suffix of x''. Hence $(\pi, \rho') \in \varphi_{u,v}$. Thus, $\varphi_{u,v}$ is not a partial function.

In case (2), let $\pi = (pz_1 \cdots z_{i-1}, r)$ and let ρ, ρ' be as above. We have $rv = rtx' = rsz_{i+1} \cdots z_n x' \in X^+$, whence $(\pi, \rho) \in \varphi_{u,v}$. Next, $rvr' = rsz_{i+1} \cdots z_n x'r' = rsz_{i+1} \cdots z_n x'' \in X^+$, and r' is a proper suffix of x''. Thus $(\pi, \rho') \in \varphi_{u,v}$. Since $\rho \neq \rho'$, $\varphi_{u,v}$ is not a partial function. ∎

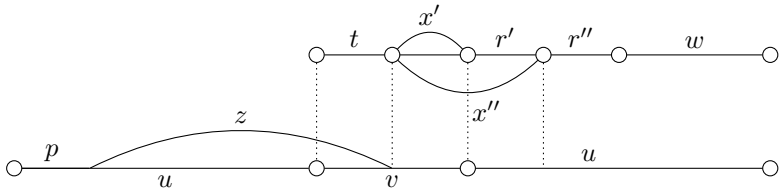

Figure 4.3: $\varphi_{u,v}$ is not a partial function.

Lemma 4.2.3 has a dual formulation for suffix codes: if X is a suffix code, then for all pairs $(u, v) \in C(X, F)$, the relation $\varphi_{u,v}$ is injective: if $(\pi, \rho), (\pi', \rho) \in \varphi_{u,v}$, then $\pi = \pi'$. Conversely, if X is an F-maximal prefix code, and if this implication holds for all pairs $(u, v) \in C(X, F)$, then X is a suffix code.

Recall that a set $X \subset F$ is right F-complete if any word of F is a prefix of X^*.

Lemma 4.2.4 *Let F be a recurrent set and let $X \subset F$ be an F-thin set. The set X is right F-complete if and only if, for all pairs $(u, v) \in C(X, F)$, the relation $\varphi_{u,v}$ contains a total function from* $\mathrm{Fact}(u)$ *into itself, that is for every $\pi \in \mathrm{Fact}(u)$, there exists $\rho \in \mathrm{Fact}(u)$ such that $(\pi, \rho) \in \varphi_{u,v}$.*

Proof. Assume first that X is right F-complete. Let $u, v \in F$ be such that $(u, v) \in C(X, F)$. Let $\pi = (p, s) \in \mathrm{Fact}(u)$. Suppose first that s has a prefix x in X. Let $s = xt$, with $x \in X$. Thus $u = ps = pxt$. Let $q = px$ and $\rho = (q, t)$. Then $(\pi, \rho) \in \varphi_{u,v}$. Suppose next that s has no prefix in X. Since X is right F-complete, there exists a word w such that $svuw = x_1 \cdots x_m$, with $x_1, \ldots, x_m \in X$.

Let n be the smallest integer such that sv is a prefix of $x_1 \cdots x_n$, $1 \leq n \leq m$. Let q be the prefix of uw such that $svq = x_1 \cdots x_n$. Since $sv \neq 1$, q is a proper suffix of x_n. The word q is a prefix of u since u is not an internal factor of X. Define the factorization $\rho = (q, t)$ of u by $svq = x_1 \cdots x_n$. Since s has no prefix in X, the word s is a proper prefix of x_1. Therefore, $(\pi, \rho) \in \varphi_{u,v}$. This shows that $\varphi_{u,v}$ contains a total function.

Conversely, assume that for all $(u, v) \in C(X, F)$, the relation $\varphi_{u,v}$ contains a total function from $\mathrm{Fact}(u)$ into itself. We show that any $u \in F$ is prefix-comparable with a word of X. By Proposition 3.3.2, this implies that X is right F-complete.

Let $u \in F$. Since X is F-thin, the set $F \setminus I(X)$ is nonempty. Let $w \in F \setminus I(X)$ and let v be such that $uvw \in F$. Set $r = uvw$. Note that $r \in F \setminus I(X)$. Let $z \neq 1$ be such that $rzr \in F$. Then $(r, z) \in C(X, F)$. Set $\pi = (1, r)$. Since $\varphi_{r,z}$ contains a total function, there is a factorization $\rho = (q, t)$ of r such that $(\pi, \rho) \in \varphi_{r,z}$. If $q \in X$, then r has the prefix q in X, the word u is prefix-comparable with q, and we obtain the conclusion. Otherwise, we have $uvwzq = x_1 \cdots x_n$ with $x_i \in X$ and uvw is a prefix of x_1, whence our conclusion again. ∎

Lemma 4.2.4 has a dual formulation for left F-complete sets: the set X is left F-complete if and only if, for all pairs $(u, v) \in C(X, F)$, the transpose of the relation $\varphi_{u,v}$ contains a total function from $\mathrm{Fact}(u)$ into itself.

Proposition 4.2.5 *Let F be a recurrent set and let $X \subset F$ be an F-thin and F-maximal prefix code. Then X is a suffix code if and only if it is left F-complete.*

Proof. Since X is an F-maximal prefix code, by Lemmas 4.2.3 and 4.2.4, for any pair $(u, v) \in C(X, F)$, the relation $\varphi_{u,v}$ is a total function from $\mathrm{Fact}(u)$ into itself.

Assume first that X is a suffix code. Then, by the dual of Lemma 4.2.3, for any pair $(u, v) \in C(X, F)$, the function $\varphi_{u,v}$ from $\mathrm{Fact}(u)$ into itself is injective. Since $\mathrm{Fact}(u)$ is a finite set, $\varphi_{u,v}$ is also surjective for any pair $(u, v) \in C(X, F)$. This implies by the dual of Lemma 4.2.4 that X is left F-complete.

Assume conversely that X is left F-complete. By the dual of Lemma 4.2.4, the function $\varphi_{u,v}$ maps $\mathrm{Fact}(u)$ onto itself for every pair $(u, v) \in C(X, F)$. This implies as above that it is also injective. By the dual of Lemma 4.2.3, and since X is an F-maximal prefix code, X is a suffix code. ∎

Proposition 4.2.5 has a dual formulation for an F-maximal suffix code.

Proof of Theorem 4.2.2. We first show that (i) implies (ii). If X is an F-maximal suffix code, then X is left F-complete and thus condition (ii) is true. Assume next that X is an F-maximal prefix code. Since X is suffix, by Proposition 4.2.5, it is left F-complete and thus (ii) holds. Finally assume that X is neither an F-maximal prefix code nor an F-maximal suffix code. Let $y, z \in F$ be such that $X \cup y$ is prefix and $X \cup z$ is suffix. Since F is recurrent, there is a word u such that $yuz \in F$. Then $X \cup yuz$ is bifix and thus we get a contradiction.

The proof that (i) implies (ii') is similar.

(ii) implies (iii). Consider the set $Y = X \setminus A^+X$. It is a suffix code by definition. It is prefix since it is contained in X. It is left F-complete. Indeed, one has $A^*X = A^*Y$ and thus A^*Y is left F-dense by the dual of Proposition 3.3.2. Hence Y is an F-maximal suffix code. By the dual of Proposition 4.2.5, the set Y is right F-complete. Thus Y is an F-maximal prefix code. This implies that $X = Y$ and thus that X is an F-maximal prefix code and an F-maximal suffix code.

The proof that (ii') implies (iii) is similar. It is clear that (iii) implies (i).

■

Example 4.2.6 Let $A = \{a, b\}$ and let F be the set of words without factor bb (Example 2.1.2). The set $X = \{aaa, aaba, ab, baa, baba\}$ is a finite F-maximal bifix code. As an example of computation of the relation $\varphi_{u,v}$, note that for $u = aaa$ and $v = b$, we have $\mathrm{Fact}(u) = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ with $\pi_1 = (1, aaa)$, $\pi_2 = (a, aa)$, $\pi_3 = (aa, a)$, $\pi_4 = (aaa, 1)$. The function $\varphi_{u,v}$ is the cycle $(\pi_1\pi_4\pi_3)$ and fixes π_2.

The following example shows that Theorem 4.2.2 is false if F is not recurrent.

Example 4.2.7 Let $F = a^*b^*$. Then $X = \{aa, ab, b\}$ is an F-maximal prefix code. It is not a suffix code but it is left F-complete as it can be easily verified.

Let $F \subset A^*$ be a factorial set. The *F-degree*, denoted $d_F(X)$, of a set $X \subset A^*$ is the maximal number of parses of words of F with respect to X, that is

$$d_F(X) = \max_{w \in F} \pi_X(w)\,.$$

The F-degree of a set X is finite or infinite. The A^*-degree is called the *degree*, and is denoted $d(X)$. Observe that $d_F(X) = d_F(X \cap F)$, and that $d_F(X) \leq d(X)$.

The following is a generalization of Theorem 6.3.1 in [6].

Theorem 4.2.8 *Let F be a recurrent set and let $X \subset F$ be a bifix code. Then X is an F-thin and F-maximal bifix code if and only if its F-degree $d_F(X)$ is finite. In this case,*

$$I(X) = \{w \in F \mid \pi_X(w) < d_F(X)\}\,. \tag{4.11}$$

Proof. Assume first that X is an F-thin and F-maximal bifix code. Since X is F-thin, $F \setminus I(X)$ is not empty. Let $u \in F \setminus I(X)$ and $w \in F$. Since F is recurrent, there is a word $v \in F$ such that $uvw \in F$. Since X is prefix, by Proposition 4.1.3, the number of parses of u is equal to the number of prefixes of u which have no suffix in X. Since X is left F-complete, the set of words in F which have no suffix in X coincides with the set S of words which are proper suffixes of words in X. Since u is not an internal factor of a word in X, any prefix of uvw which is in S is a prefix of u. Thus $\pi_X(uvw) = (\underline{S}\,\underline{A}^*, uvw) = (\underline{S}\,\underline{A}^*, u) = \pi_X(u)$. Since by Equation (4.8), $\pi_X(w) \leq \pi_X(uvw)$, we get $\pi_X(w) \leq \pi_X(u)$. This shows that

π_X is bounded, and thus that the F-degree of X is finite. Moreover, this shows that $F \setminus I(X)$ is contained in the set of words of F with maximal value of π_X. Conversely, consider $w \in I(X)$. Then there exists $w' \in X$ and $p, s \in A^+$ such that $w' = pws$. Then by Equation (4.9) $\pi_X(w') > \pi_X(w)$, and thus $\pi_X(w)$ is not maximal in F. This proves Equation (4.11).

Conversely, let $w \in F$ be a word with $\pi_X(w) = d_F(X)$. For any nonempty word $u \in F$ such that $uw \in F$ we have $uw \in XA^*$. Indeed, set $u = au'$ with $a \in A$ and $u' \in F$. Then $\pi_X(au'w) \geq \pi_X(u'w) \geq \pi_X(w)$ by Equation (4.8). This implies $\pi_X(au'w) = \pi_X(u'w) = \pi_X(w)$. By the dual of Equation (4.7) we obtain that $uw \in XA^*$.

This implies first that X is F-thin and next that XA^* is right F-dense. Indeed suppose that w is an internal factor of a word in X. Let $p, s \in F \setminus 1$ be such that $pws \in X$. Since $pw \in F$, the previous argument shows that $pw \in XA^*$, a contradiction. Thus $w \in F \setminus I(X)$. This shows that X is F-thin.

Next, and since F is recurrent, for any $v \in F$, there is a word $u \in F$ such that $vuw \in F$. Then $vuw \in XA^*$ by using again the above argument. Thus XA^* is right F-dense and X is an F-maximal bifix code by Theorem 4.2.2. ∎

Example 4.2.9 Let F be the Fibonacci set. The set $X = \{a, bab, baab\}$ is a finite bifix code. Since it is finite, it is F-thin. It is an F-maximal prefix code as one may check on Figure 2.1. Thus it is, by Theorem 4.2.2, an F-thin and F-maximal bifix code. The parses of the word bab are $(1, bab, 1)$ and (b, a, b). Since bab is not in $I(X)$, one has $d_F(X) = 2$.

Example 4.2.10 Let F be the Fibonacci set. The set $X = \{aaba, ab, baa, baba\}$ is a bifix code. It is F-maximal since it is right F-complete (see Figure 2.1). It has F-degree 3. Indeed, the word $aaba$ has three parses $(1, aaba, 1)$, (a, ab, a) and $(aa, 1, ba)$ and it is in $F \setminus I(X)$.

The following result establishes the link between maximal bifix codes and F-maximal ones.

Theorem 4.2.11 *Let F be a recurrent set. For any thin maximal bifix code $X \subseteq A^+$ of degree d, the set $Y = X \cap F$ is an F-thin and F-maximal bifix code. One has $d_F(Y) \leq d$ with equality when X is finite.*

Proof. Recall that $d_F(Y) = d_F(X \cap F) = d_F(X) \leq d$. Thus $d_F(Y)$ is finite and by Theorem 4.2.8, Y is an F-thin and F-maximal bifix code. If X is finite, then each word which is in F and is longer than the longest words in X has d parses. Thus $d_F(X) = d$, whence $d_F(Y) = d$. ∎

Example 4.2.12 The set $X = a \cup ba^*b$ is a maximal bifix code of degree 2. Let F be the Fibonacci set. Then $X \cap F = \{a, baab, bab\}$ (see Figure 2.1).

As another example, let $Z = \{a^3, a^2ba, a^2b^2, ab, ba^2, baba, bab^2, b^2a, b^3\}$. The set Z is a finite maximal bifix code of degree 3 (see [6]). Then $Z \cap F = \{a^2ba, ab, ba^2, baba\}$ (see Figure 2.1).

Example 4.2.13 Let F be the Thue–Morse set. Consider again $X = a \cup ba^*b$. Then $X \cap F = \{a, baab, bab, bb\}$ is a finite F-maximal bifix code of F-degree 2 (see Figure 2.2).

The following examples show that a strict inequality can hold in Theorem 4.2.11. The second example shows that this may happen even if all letters occur in the words of F.

Example 4.2.14 Let $A = \{a, b\}$ and let $X = a \cup ba^*b$. The set X is a maximal bifix code of degree 2. Let $F = a^*$. Then F is a recurrent set. We have $Y = X \cap F = a$. The F-degree of Y is 1.

Example 4.2.15 Let $A = \{a, b\}$ and let $X \subset A^+$ be the maximal bifix code of degree 3 with kernel $K = \{aa, ab, ba\}$. Let F be the Fibonacci set. Since $K = A^2 \cap F$, K is an F-maximal bifix code. Since $K \subset X \cap F$ and K is F-maximal, one has $X \cap F = K$. Next $K = A^2 \cap F$ and Theorem 4.2.11 imply that $d_F(K) = 2$. Thus $d(X) = 3$ and $d_F(X \cap F) = 2$.

4.3 Derivation

We first show that the notion of derived code can be extended to F-maximal bifix codes. The following result generalizes Proposition 6.4.4 in [6].

The *kernel* of a set of words X is the set of words in X which are internal factors of words in X. We denote by $K(X)$ the kernel of X. Note that $K(X) = I(X) \cap X$.

Theorem 4.3.1 *Let F be a recurrent set. Let $X \subset F$ be a bifix code of finite F-degree $d \geq 2$. Set $I = I(X)$ and $K = K(X)$. Let $G = (IA \cap F) \setminus I$ and $D = (AI \cap F) \setminus I$. Then the set $X' = K \cup (G \cap D)$ is a bifix code of F-degree $d - 1$.*

The code X' is called the *derived* code of X with respect to F or F-derived code.

The proof uses two lemmas. Let P be the set of proper prefixes of X and let S be the set of proper suffixes of X.

Lemma 4.3.2 *One has $G \subset S$ and $D \subset P$.*

Proof. By Theorem 4.2.8, the parse enumerator of X is bounded on F and $F \setminus I(X) = F \setminus I$ is the set of words in F with maximal value $d_F(X)$. Let $y = ha$ be in G with $h \in I$ and $a \in A$. Since $y \notin I$, we have $\pi_X(ha) > \pi_X(h)$. Thus, by Proposition 4.1.6, $y = ha$ does not have a suffix in X. Since A^*X is left F-dense, this implies that y is a proper suffix of a word in X. Thus y is in S. The proof that $D \subset P$ is symmetrical. ∎

Lemma 4.3.3 *For any $x \in X \setminus K$, the shortest prefix of x which is not in I is in X'.*

Proof. Since $x \notin K$, we have $x \notin I$. Let x' be the shortest prefix of x which is not in I or, equivalently such that $\pi_X(x') = d_F(X)$. Let us show that $x' \in X'$. First, x' is a proper prefix of x. Set indeed $x = pa$ with $p \in A^*$ and $a \in A$. Since $x \in X$, we have by Equation (4.7), $\pi_X(x) = \pi_X(p)$. Thus $p \notin I$ and x' is a prefix of p.

Since $1 \in I$, we have $x' \neq 1$. Set $x' = p'a'$ with $p' \in A^*$ and $a' \in A$. By definition of x' we have $p' \in I$. Thus $x' \in G = (IA \cap F) \setminus I$.

Next, set $x' = a''s$ with $a'' \in A$ and $s \in A^*$. Since $x' \notin XA^*$, we have by the dual of Equation (4.7), $\pi_X(s) < \pi_X(x')$. Thus s is in I. This shows that $x' \in D$. Thus we conclude that $x' \in G \cap D \subset X'$. ∎

There is a dual of Lemma 4.3.3 concerning the shortest suffix of a word in $X \setminus K$.

Proof of Theorem 4.3.1.

We first prove that X' is a prefix code. Suppose first that $k \in K$ is a prefix of a word z in $G \cap D$. By Lemma 4.3.2, a word in D is a proper prefix of X. Thus $k \in X$ would be a proper prefix of X, which is impossible since X is prefix.

Suppose next that a word u of $G \cap D$ is a prefix of a word k in K. Since k is in I, it follows that u is in I, a contradiction.

Finally, no word $y \in G \cap D$ can be a proper prefix of another word y' in $G \cap D$, otherwise $y' = yz$, with $z \in A^+$. Therefore, since $G \subset S$ by Lemma 4.3.2, there is $t \in A^+$ such that $ty' = tyz \in X$. Consequently, $y \in G \cap I$, a contradiction.

Thus X' is a prefix code. To show that it is F-maximal, it is enough to show that any word in X has a prefix in X'.

Consider indeed $x \in X$. If x is in K then $x \in X'$. Otherwise, let x' be the shortest prefix of x which is not in I. By Lemma 4.3.3, we have $x' \in X'$.

Thus X' is an F-maximal prefix code.

A symmetric argument shows that X' is an F-maximal suffix code.

Let us show that $d_F(X') = d_F(X) - 1$. We first note that $G \cap D \neq \emptyset$. Indeed, let $x \in X$ be such that $\pi_X(x)$ is maximal on X. If x were an internal factor of a word $y \in X$, then by Equation (4.9) $\pi_X(x) < \pi_X(y)$ which contradicts our assumption. Thus $x \notin K$. This shows that K is not an F-maximal bifix code and thus that $X' \setminus K = G \cap D \neq \emptyset$. Consider $x' \in G \cap D$. Since $(G \cap D) \cap I(X)$ is empty, and since $I(X') \subset I(X)$, x' cannot be in $I(X')$. Thus the number of parses of x' with respect to X' is $d_F(X')$.

Let P' be the set of proper prefixes of X'. We show that x' has $d_F(X) - 1$ suffixes which are in P'. This will show that $d_F(X') = d_F(X) - 1$ by the dual of Proposition 4.1.3.

Since $x' \in F \setminus I$, we have $\pi_X(x') = d_F(X)$. Thus x' has $d_F(X)$ suffixes in P. One of them is x' itself since $x' \in D \subset P$. Let p be a proper suffix of x' which is in P. Let us show that p does not have a prefix in X'. Indeed, arguing by contradiction, assume that $x'' \in X'$ is a prefix of p. We cannot have $x'' \in K$ since p is a proper prefix of a word in X. We cannot have either $x'' \in G \cap D$. Indeed, since x' is in AI, p is in I and thus also $x'' \in I$. Thus p cannot have a prefix in X'. Since X' is an F-maximal prefix code, this implies that p is a

proper prefix of X'. Thus, the $d_F(X) - 1$ proper suffixes of x' which are in P are in P'. ∎

Example 4.3.4 Let F be the Fibonacci set. Let $X = \{a, bab, baab\}$. The set X is an F-thin and F-maximal bifix code of F-degree 2 (see Example 4.2.9). We have $K = \{a\}$, $I = \{1, a, aa\}$, $G = \{b, ab, aab\}$ and $D = \{b, ba, baa\}$. Thus $X' = \{a, b\}$.

The following is a generalization of Proposition 6.3.14 in [6].

Proposition 4.3.5 *Let F be a recurrent set. Let $X \subset F$ be a bifix code of F-degree $d \geq 2$. Let S be the set of proper suffixes of X and set $I = I(X)$. The set $S \setminus I$ is an F-maximal prefix code and the set $S \cap I$ is the set of proper suffixes of the derived code X'.*

The proof uses the following lemma.

Lemma 4.3.6 *Let F be a recurrent set. Let $X \subset F$ be an F-thin and F-maximal bifix code. Let S be the set of proper suffixes of X and set $I = I(X)$. For any $w \in F \setminus I$ the longest prefix of w which is in S is not in I.*

Proof. Let s be the longest prefix of w which is in S. Set $w = st$. Let us show that for any prefix t' of t, we have $\pi_X(st') = \pi_X(s)$. It is true for $t' = 1$. Assume that it is true for t' and let $a \in A$ be the letter such that $t'a$ is a prefix of t. Since $st'a \notin S$, we have $st'a \in A^*X$. Thus by Equation (4.7), this implies $\pi_X(st'a) = \pi_X(st')$. Thus $\pi_X(st'a) = \pi_X(s)$. We conclude that $\pi_X(st) = \pi_X(s)$. Since $w = st$ is in $F \setminus I$, and since $F \setminus I$ is the set of words in F with maximal value of π_X, this implies that $s \in F \setminus I$. ∎

This lemma has a dual statement for the longest suffix of a word in $w \in F \setminus I$ which is in P.

Proof of Proposition 4.3.5. Set $Y = S \setminus I$. Let us first show that Y is prefix. Assume that $u, uv \in Y$. Since $uv \in S$ there is a nonempty word p such that $puv \in X$. Since $u \notin I$, this forces $v = 1$. Thus Y is prefix.

We show next that YA^* is right F-dense. Consider $u \in F$ and let $w \in F \setminus I$. Since F is recurrent, there exists $v \in F$ such that $uvw \in F$. Let s be the longest word of S which is a prefix of uvw. By Lemma 4.3.6, we have $s \in F \setminus I$. Thus $s \in S \setminus I = Y$ and $uvw \in YA^*$. This shows that YA^* is right F-dense.

Let us now show that the set S' of proper suffixes of the words of X' is $S \cap I$. Let s be a proper suffix of a word $x' \in X'$. If $x' \in K$, then s is in $S \cap I$. Suppose next that $x' \in G \cap D$. Since $G \subset S$ by Lemma 4.3.2, we have $s \in S$. Furthermore, since $D \subset AI$, we have $s \in I$. This shows that $s \in S \cap I$.

Conversely, let s be in $S \cap I$. Let $x \in X$ be such that s is a proper suffix of x. If x is in K then x is in X' and thus s is in S'. Otherwise, let y be the shortest suffix of x which is in not in I. By the dual of Lemma 4.3.3, the word y is in X'. Then s is a proper suffix of y (since $s \in I$ and $y \notin I$) and therefore s is in S'. ∎

There is a dual version of Proposition 4.3.5 concerning the set of proper prefixes of an F-thin and F-maximal bifix code $X \subset F$.

The following property generalizes Theorem 6.3.15 in [6].

Theorem 4.3.7 *Let F be a recurrent set. Let X be a bifix code of finite F-degree d. The set of its nonempty proper suffixes is a disjoint union of $d-1$ F-maximal prefix codes.*

Proof. Let S be the set of proper suffixes of X. If $d = 1$, then $S \setminus 1$ is empty. If $d \geq 2$, by Proposition 4.3.5, the set $Y = S \setminus I$ is an F-maximal prefix code and the set $S \cap I$ is equal to the set S' of proper suffixes of the words of the derived code X'. Arguing by induction, the set $S' \setminus 1$ is a disjoint union of $d-2$ F-maximal prefix codes. Thus $S \setminus 1 = Y \cup (S' \setminus 1)$ is a disjoint union of $d-1$ F-maximal prefix codes. ∎

The following generalizes Corollary 6.3.16 in [6], with two restrictions. First, it applies only in the case of finite maximal bifix codes instead of thin bifix codes (in order to be able to use Proposition 3.3.4). Next, it applies only for recurrent sets such that there exists a positive invariant probability distribution (in order to be able to use Proposition 3.4.1).

Corollary 4.3.8 *Let F be a recurrent set such that there exists a positive invariant probability distribution δ on F. Let X be a finite bifix code of finite F-degree d. The average length of X with respect to δ is equal to d.*

Proof. Let δ be a positive invariant probability distribution on F. By the dual of Proposition 3.4.1, one has $\lambda(X) = \delta(S)$. In view of Theorem 4.3.7, we have $S \setminus 1 = Y_1 \cup \ldots \cup Y_{d-1}$ where each Y_i is a finite F-maximal prefix code. By Proposition 3.3.4, we have $\delta(Y_i) = 1$ for $1 \leq i \leq d-1$. Thus $\lambda(X) = d$. ∎

Example 4.3.9 Let F be the Fibonacci set and let $X = \{a, bab, baab\}$ (Example 4.3.4). The set X is an F-maximal bifix code of F-degree 2. With respect to the unique invariant probability distribution of F (Example 2.4.6), we have $\lambda(X) = \lambda + 3(2 - 3\lambda) + 4(2\lambda - 1) = 2$.

Now we show that an F-thin and F-maximal bifix code is determined by its F-degree and its kernel. We first prove the following generalization of Proposition 6.4.1 from [6].

Proposition 4.3.10 *Let F be a recurrent set. Let $X \subset F$ be a bifix code of finite F-degree d and let K be the kernel of X. Let Y be a set such that $K \subset Y \subset X$. Then for all $w \in I(X) \cup Y$,*

$$\pi_Y(w) = \pi_X(w). \tag{4.12}$$

For all $w \in F$,

$$\pi_X(w) = \min\{d, \pi_Y(w)\}. \tag{4.13}$$

Proof. Denote by $F(w)$ the set of factors of the word w. Notice that Equation (4.3) is equivalent to $L_X = \underline{A^*}(1 - \underline{X})\underline{A^*}$. Thus, to prove (4.12), we have to show that for any $w \in I(X) \cup Y$ one has $F(w) \cap X = F(w) \cap Y$. The inclusion $F(w) \cap Y \subset F(w) \cap X$ is clear. Conversely, if w is in $I(X)$, then $F(w) \cap X \subset K$ and thus $F(w) \cap X \subset F(w) \cap Y$. Next, assume that w is in Y. The words in $F(w) \cap X$ other than w are all in K. Thus we have again $F(w) \cap X \subset F(w) \cap Y$.

To show Equation (4.13), assume first that $w \in I(X)$. Then $\pi_X(w) < d$ by Theorem 4.2.8. Moreover, $\pi_X(w) = \pi_Y(w)$ by Equation (4.12). Thus Equation (4.13) holds. Next, suppose that $w \in F \setminus I(X)$. Then $\pi_X(w) = d$. Since $Y \subset X$, we have $\pi_X(w) \leq \pi_Y(w)$ by Equation (4.1). This proves (4.13). ∎

Proposition 4.3.10 will be used to prove the following generalization of Theorem 6.4.2 in [6].

Theorem 4.3.11 *Let F be a recurrent set and let $X \subset F$ be a bifix code of finite F-degree d. For any $w \in F$, one has*

$$\pi_X(w) = \min\{d, \pi_{K(X)}(w)\}\,.$$

In particular X is determined by its F-degree and its kernel.

Proof. Take $Y = K(X)$ in Proposition 4.3.10. Then the formula follows from Equation (4.13). Next X is determined by L_X, and so by π_X, through Equation (4.3). ∎

We now state the following generalization of Theorem 6.4.3 in [6].

Theorem 4.3.12 *Let F be a recurrent set. A bifix code $Y \subset F$ is the kernel of some bifix code of finite F-degree d if and only if*

(i) *Y is not an F-maximal bifix code,*
(ii) $\max\{\pi_Y(y) \mid y \in Y\} \leq d - 1$.

Proof. Let X be an F-thin and F-maximal bifix code of F-degree d and let $Y = K(X)$ be its kernel. Condition (i) is satisfied because $X = Y$ implies that X is equal to its derived code which has F-degree $d - 1$. Moreover, for every $y \in Y$ one has $\pi_X(y) \leq d - 1$. Since $\pi_X(y) = \pi_Y(y)$ by Equation (4.12), condition (ii) is also satisfied.

Conversely, let $Y \subset F$ be a bifix code satisfying conditions (i) and (ii). Let $\pi : A^* \to \mathbb{N}$ be the function defined by

$$\pi(w) = \min\{d, \pi_Y(w)\}.$$

It can be verified that the function π satisfies the four conditions of Proposition 4.1.5. Thus π is the parse enumerator of a bifix code Z. Let $X = Z \cap F$. Then π_X and π have the same restriction to F. Since π is bounded on F, the same holds for π_X. This implies that the code X is an F-thin and F-maximal bifix code by Theorem 4.2.8. Since the code Y is not an F-maximal bifix code, the

F–parse enumerator π_Y is not bounded. Consequently $\max\{\pi(w) \mid w \in F\} = d$, showing that X has F-degree d. Let us prove finally the Y is the kernel of X. Since, by condition (ii), $\max\{\pi_Y(y) \mid y \in Y\} \leq d-1$, we have $Y \subset I(X)$.

Moreover, for $w \in I(X)$ we have $\pi_X(w) = \pi_Y(w)$. Let L (resp. L_Y) be the indicator of X (resp. of Y). Since $1 - \underline{X} = (1 - \underline{A})L(1 - \underline{A})$ and $1 - \underline{Y} = (1 - \underline{A})L_Y(1 - \underline{A})$ by Equation (4.3), we conclude that for $w \in I(X)$, we have $(\underline{X}, w) = (\underline{Y}, w)$. This implies that if $w \in I(X)$, then w is in X if and only if w is in Y. Thus $K(X) = I(X) \cap X = I(X) \cap Y = Y$ and Y is the kernel of X. ∎

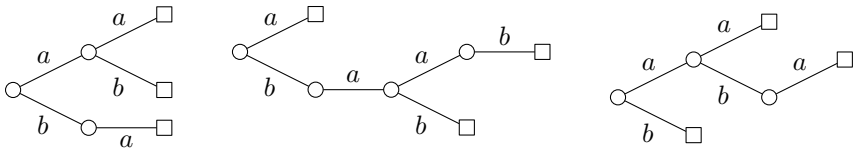

Figure 4.4: The three F-maximal bifix codes of F-degree 2 in the Fibonacci set F.

Example 4.3.13 Let $A = \{a, b\}$ and let $F \subset A^*$ be the Fibonacci set. There are three maximal bifix codes of F-degree 2 in F represented on Figure 4.4. Indeed, by Theorem 4.3.12, the possible kernels are $\emptyset$, $\{a\}$ and $\{b\}$.

4.4 Finite maximal bifix codes

The following generalizes Theorem 6.5.2 of [6].

Theorem 4.4.1 *For any recurrent set F and any integer $d \geq 1$ there is a finite number of finite F-maximal bifix codes $X \subset F$ of F-degree d.*

Proof. The only F-maximal bifix code of F-degree 1 is $F \cap A$. Arguing by induction on d, assume that there are only finitely many finite F-maximal bifix codes of F-degree d. Each finite F-maximal bifix code $X \subset F$ of F-degree $d+1$ is determined by its kernel which is a subset of its derived code X'. Since X' is a finite F-maximal bifix code of F-degree d, there are only a finite number of kernels and we are done. ∎

Example 4.4.2 Let $A = \{a, b\}$ and let F be the set of words without factor bb. There are two finite F-maximal bifix codes of F-degree 2, namely the code $\{aa, ab, ba\}$ with empty kernel and the code $\{aa, aba, b\}$ with kernel b. The code of F-degree 2 with kernel a is $a \cup ba^+b$, and thus is infinite.

The following result shows that the case of a uniformly recurrent set contrasts with the case $F = A^*$ since in A^*, as soon as $\mathrm{Card}(A) \geq 2$, there exist infinite maximal bifix codes of degree 2 and thus of all degrees $d \geq 2$; see e.g. [6, Example 6.4.7] for degree 2 and [6, Theorem 6.4.6] for the general case.

Theorem 4.4.3 *Let F be a uniformly recurrent set. Any F-thin bifix code $X \subset F$ is finite. Any finite bifix code is contained in a finite F-maximal bifix code.*

Proof. Let $X \subset F$ be an F-thin bifix code. Since X is F-thin, there exists a word $w \in F \setminus I(X)$. Since F is uniformly recurrent there is an integer r such that w is factor of every word in $F_r = F \cap A^r$. Assume $F_k \cap X \neq \emptyset$ for some $k \geq r+2$, and let $x \in F_k \cap X$. Set $x = pqs$, with $q \in F_r \cap I(X)$, and p, s nonempty. Then w is factor of q, hence w is in $I(X)$, a contradiction. We deduce that each x in X has length at most $r+1$. Thus X is finite.

Let $X \subset F$ be a finite bifix code which is not F-maximal. Let $d = \max\{\pi_X(x) \mid x \in X\}$. By Theorem 4.3.12, X is the kernel of an F-thin and F-maximal bifix code $Z \subset F$ of F-degree $d+1$. By the previous argument, Z is finite. ∎

By Theorem 6.6.1 of [6], any rational bifix code is contained in a maximal rational bifix code. We have seen that the situation is simpler for bifix codes in uniformly recurrent sets.

Example 4.4.4 Let F be the Fibonacci set. Let $X = \{a, bab\}$. Then X is contained in the bifix code $\{a, bab, baab\}$ which has F-degree 2 (see Figure 4.4). It is also the kernel of $\{a, baabaab, baababaab, bab\}$ which is a bifix code of F-degree 3 (see Table 5.1).

The following is a generalization of Proposition 6.2.10 in [6]. The equality $d(Y) = d(X)$ is stated as a comment following Proposition 6.3.9 in [6, page 243], in a more general framework.

Proposition 4.4.5 *Let F be a recurrent set, let $X \subset F$ be a finite F-maximal bifix code and let w be a nonempty word in F. Let $G = Xw^{-1}$, and $D = w^{-1}X$. If*

$$G \neq \emptyset\,, \quad D \neq \emptyset\,, \text{ and } \quad Gw \cap wD = \emptyset\,,$$

then the set

$$Y = (X \cup w \cup (GwD \cap F)) \setminus (Gw \cup wD) \tag{4.14}$$

is a finite F-maximal bifix code with the same F-degree as X.

We use in the proof the following proposition which is an extension of Corollary 3.4.7 of [6].

Proposition 4.4.6 *Let F be a recurrent set and let $X \subset F$ be a F-maximal prefix code. Let $X = X_1 \cup X_2$ be a partition of X into two prefix codes and let Y be a finite prefix code such that $Y \cap x^{-1}F$ is $x^{-1}F$-maximal for all $x \in X_2$. Then the set $Z = X_1 \cup (X_2Y \cap F)$ is an F-maximal prefix code.*

Proof. We first prove that Z is a prefix code. Let z and z' be distinct words in Z. We show that they are not prefix-comparable. Since X_1 is a prefix code, this holds if both words are in X_1. Assume next that $z \in X_2Y$. Then $z = xy$ with $x \in X_2$ and $y \in Y$.

If z' is in X_1, then z' and x are not prefix-comparable because they are distinct since $z' \in X_1$ and $x \in X_2$, and so z and z' are not prefix-comparable.

If $z' \in X_2Y$, set $z' = x'y'$ with $x' \in X_2$ and $y' \in Y$. Either x and x' are not prefix-comparable, and then so are z and z', or $x = x'$. In the latter case, y and y' are not prefix-comparable because Y is a prefix code, and again z and z' are not prefix-comparable. Thus Z is a prefix code.

Let us show that Z is F-maximal. Let $u \in F$. Since X is an F-maximal prefix code, there is an $x \in X$ which is prefix-comparable with u. If x is in X_1, then $x \in Z$ and thus u is prefix-comparable with a word of Z. Otherwise, we have $x \in X_2$.

Suppose first that u is a prefix of x. Since Y is a finite $x^{-1}F$-maximal prefix code, it is not empty and u is a prefix of xv for every $v \in Y \cap x^{-1}F$.

Suppose next that $u = xv$ for some word v. Since v is in $x^{-1}F$ and since $Y \cap x^{-1}F$ is an $x^{-1}F$-maximal prefix code, the word v is prefix-comparable with some $y \in Y \cap x^{-1}F$. Thus u is prefix-comparable with $xy \in Z$. ∎

Proof of Proposition 4.4.5. The condition $G \neq \emptyset$ (resp. $D \neq \emptyset$) means that w is a suffix of X (resp. a prefix of X). The condition $Gw \cap wD = \emptyset$ implies that w is not in X.

By Theorem 4.2.2, the set X is an F-maximal prefix code. By Proposition 3.3.8, the set $Y_1 = (X \cup w) \setminus wD$ is an F-maximal prefix code. Next, we have

$$Y = (Y_1 \setminus Gw) \cup (GwD \cap F).$$

We show that Y is an F-maximal prefix code, by applying Proposition 4.4.6. Indeed, consider the partition $Y_1 = X_1 \cup X_2$ with $X_1 = Y_1 \setminus Gw$ and $X_2 = Gw$. Then $Y = X_1 \cup (X_2D \cap F)$. Clearly D is a finite $w^{-1}F$-maximal prefix code. Since $(gw)^{-1}F$ is a subset of $w^{-1}F$ for all $g \in G$, the set $D \cap (gw)^{-1}F$ is a finite $(gw)^{-1}F$-maximal prefix code for all $g \in G$ by Proposition 3.3.6, So the claim follows from by Proposition 4.4.6. This proves that Y is an F-maximal prefix code. Since Y it is also a suffix code, it follows that Y is an F-maximal bifix code by Theorem 4.2.2.

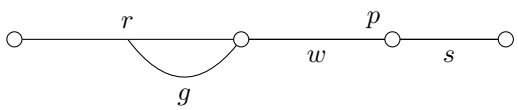

Figure 4.5: Construction of $\varphi(p)$ (second case).

To show that X and Y have the same degree, consider a word $u \in F$ which is not an internal factor of X nor Y. Such a word exists since X and Y are finite. Let P (resp. Q) be the set of proper prefixes of the words of X (resp. Y). We define a bijection φ between the set $P(u)$ of suffixes of u which are in P and the

set $Q(u)$ of suffixes of u which are in Q. This will imply that $d_F(X) = d_F(Y)$ by the dual of Proposition 4.1.3.

Let $p \in P$ be a suffix of u and set $u = rp$. If w is not a prefix of p, then p is in Q. Otherwise, set $p = ws$. Since the words in P starting with w are all prefixes of wD, the word s is a proper prefix of D. Since G is an Fw^{-1}-maximal suffix code, r is suffix-comparable with a word of G. If r is a proper suffix of G, then $urws$ is an internal factor of GwD, a contradiction. Thus r has a suffix $g \in G$. This suffix is unique because G is a suffix code. Since $gp = gws$, the word gp is a proper prefix of GwD, and thus a proper prefix of Y. Thus $gp \in Q(u)$. We set (see Figure 4.5)

$$\varphi(p) = \begin{cases} p & \text{if } p \notin wA^*, \\ gp & \text{if } p \in wA^* \text{ and } g \in G \text{ is the suffix of } r \text{ in } G. \end{cases}$$

Thus φ maps $P(u)$ into $Q(u)$. We show that it is injective. Suppose that $\varphi(p) = \varphi(p')$ for some $p, p' \in P(u)$. Assume that $\varphi(p) = gp$ and $\varphi(p') = g'p'$ with $g, g' \in G$. Since p and p' start with w, the word $gp = g'p'$ starts with the words gw and $g'w$ which are in X. This shows that $g = g'$ and thus $p = p'$. Assume next that $p = g'p'$ with $g' \in G$ and $p' \in wA^*$. But then $g'w$ is a prefix of p, with p in P and $g'w$ in X, a contradiction.

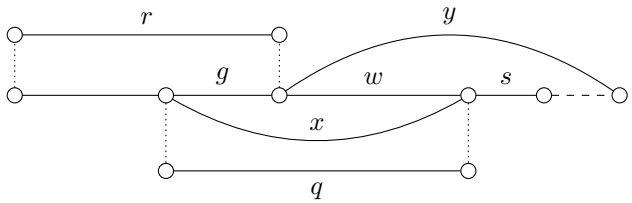

Figure 4.6: Reconstruction of the factorization.

To show that φ is surjective, consider $q \in Q(u)$. Assume first that q has a prefix x in X (see Figure 4.6). By Equation (4.14), one has $x = gw$ and $q = gws$ for some $g \in G$ and s a proper prefix of the word d in D. Thus ws is a proper prefix of $wD \subset X$, and consequently ws is a proper prefix of X. Since ws is a suffix of q, it is a suffix of u. Thus $ws \in P(u)$. Set $u = rws$. Then g is a suffix of r. Moreover $ws \in wA^*$. Consequently $\varphi(ws) = q$.

Finally, if q has no prefix in X, then q is a proper prefix of X. Moreover, since q is a prefix of Y, either q is a proper prefix of w or q is not a prefix of wD. In both cases, w is not a prefix of q and therefore $\varphi(q) = q$. Thus φ is surjective. ∎

The set Y defined by Equation (4.14), is said to be obtained from X by *internal transformation* (with respect to w).

Example 4.4.7 Let F be the Fibonacci set. The set $X = \{aa, ab, ba\}$ is an F-maximal bifix code of F-degree 2. Then $Y = \{aa, aba, b\}$ is a bifix code of F-degree 2 which is obtained from X by internal transformation with respect to $w = b$. Indeed, here $G = D = \{a\}$, $Gw = \{ab\}$, $wD = \{ba\}$ and $GwD = \{aba\}$.

The following theorem is due to Césari. It is Theorem 6.5.4 in [6].

Theorem 4.4.8 *For any finite maximal bifix code X over A of degree d, there is a sequence of internal transformations which, starting from the code A^d, gives the code X.*

Theorem 4.4.8 has been generalized to finite F-maximal bifix codes when F is the set of paths in a strongly connected graph (see [18]). It is not true in any recurrent, or even uniformly recurrent set, as shown by the following example.

Example 4.4.9 Let F be the Fibonacci set. The set $X = \{a, bab, baab\}$ is a finite bifix code of F-degree 2. It cannot be obtained by a sequence of internal transformations from the code $A^2 \cap F = \{aa, ab, ba\}$. Indeed, the only internal transformation which can be realized is with respect to $w = b$. The result is $\{aa, aba, b\}$ by Example 4.4.7. Next, no internal transformation can be realized from this code. See also Figure 4.4.

A more general form of internal transformation is described in [6] in Proposition 6.2.8. We do not know whether its adaptation to finite F-maximal bifix codes allows one to obtain all finite F-maximal bifix codes of F-degree d starting with the code $A^d \cap F$.

5 Bifix codes in Sturmian sets

In this section, we study bifix codes in Sturmian sets. This time, the situation is completely specific. First of all, as we have already seen, any F-thin bifix code included in a uniformly recurrent set F is finite (Theorem 4.4.3). Next, in a Sturmian set F, any bifix code of finite F-degree d on k letters has $(k-1)d+1$ elements (Theorem 5.2.1). Since A^d is a bifix code of degree d, this generalizes the fact that $\mathrm{Card}(F \cap A^d) = (k-1)d + 1$ for all $d \geq 1$.

Additionally, if an infinite word x is X-stable, that is if , for some thin maximal bifix code X, one has $d_{F(y)}(X) = d_{F(x)}(X)$ for all suffixes y of x, then the inequality $\mathrm{Card}(X \cap F(x)) \leq d_{F(x)}(X)$ implies that x is ultimately periodic (Theorem 5.3.2).

5.1 Sturmian sets

Let F be a factorial set on the alphabet A. Recall that a word w is strict right-special if $wA \subset F$. It is strict left-special if $Aw \subset F$. A suffix of a (strict) right-special word is (strict) right-special, a prefix of a (strict) left-special word is (strict) left-special.

A set of words F is called *Sturmian* if it is the set of factors of a strict episturmian word. By Proposition 2.3.3 a Sturmian set F is uniformly recurrent. Moreover, every right-special (left-special) word in F is strict.

The following statement gives a direct definition of Sturmian sets.

Proposition 5.1.1 *A set F is Sturmian if and only if it is uniformly recurrent and*

(i) *it is closed under reversal,*
(ii) *for each n, there is exactly one right-special word in F of length n, and this right-special word is strict.*

Proof. If $F = F(x)$ for some strict episturmian word, then the conclusions of the proposition hold.

Conversely, assume that F has the required properties. For each n, the reversal of the strict right-special word of length n is a strict left-special word. Since all these left-special words are prefixes one of the other, there is an infinite word x that such that all its prefixes are these strict left-special words. Clearly, $F(x) \subset F$. To show that x is strict episturmian, we verify that $F(x)$ is closed under reversal. Let $u \in F(x)$. Then $u \in F$. Since F is uniformly recurrent, there is an integer m such that u is a factor of the right-special word w of length m. Consequently the reversal $\widetilde{u}$ of u is a factor of the left-special word $\widetilde{w}$ of length m, and therefore is in $F(x)$.

To prove that $F \subset F(x)$, let $u \in F$. Since F is uniformly recurrent, there is an integer m such that u is a factor of the left-special word w of length m. Since w is a prefix of x, this shows that $u \in F(x)$. ∎

The following statement is a direct consequence of the previous proof.

Proposition 5.1.2 *Let F be a Sturmian set of words. There is a unique strict standard episturmian infinite word s such that $F = F(s)$.*

As a consequence of Proposition 5.1.2, for every left-special word w of a Sturmian set F, exactly one of the words wa, for $a \in A$, is left-special in F. Symmetrically, for every right-special word w in F, exactly one of the words aw for $a \in A$ is right-special in F. More generally, for every $n \geq 1$ there is exactly one word u of length n such that uw is a right-special word in F.

Proposition 5.1.3 *Any word in a Sturmian set F is a prefix of some right-special word in F.*

Proof. Let indeed $u \in F$. Since F is uniformly recurrent, there is an integer n such that u is a factor of any word in F of length n. Let w be the right-special word of length n. Then u is a factor of w, thus $w = pus$ for some words p, s. Since w is right-special, its suffix us is also right-special. Thus u is a prefix of a right-special word. ∎

The following example shows that for a Sturmian set F, there exists bifix codes $X \subset F$ which are not F-thin (we have seen such an example for a uniformly recurrent but not Sturmian set in Example 4.2.1).

Example 5.1.4 Let F be a Sturmian set. Consider the following sequence $(x_n)_{n\geq 1}$ of words of F. Set $x_1 = a$, for some $a \in A$.

Suppose inductively that $x_1, \ldots, x_n$ have been defined in such a way that $X_n = \{x_1, x_2, \ldots, x_n\}$ is bifix and not F-maximal bifix. Define x_{n+1} as follows. By Theorem 4.2.2, X_n is not right F-complete, thus there is a word u in F which is incomparable for the prefix order with the words of X_n. By Proposition 5.1.3, the word u is a prefix of a right special word v in F. Symmetrically, since X_n is not an F-maximal bifix code, there is a word $w \in F$ which is incomparable with the words of X_n for the suffix order. Since F is recurrent, there is a word t such that $vatw \in F$. Then we choose $x_{n+1} = vatw$.

The set $X_{n+1} = X_n \cup x_{n+1}$ is a bifix code since x_{n+1} is incomparable with the words of X_n for the prefix and for the suffix order. It is not an F-maximal prefix code since vb, for all letters $b \neq a$, is incomparable for the prefix order with the words of X_{n+1}: indeed, its prefix u is incomparable for the prefix order with all words in X_n and vb is incomparable with x_{n+1}. Since it is finite, it is not an F-maximal bifix code by Theorem 4.2.2. The infinite set $X = \{x_1, x_2, \ldots\}$ is a bifix code included in F and it is not F-thin by Theorem 4.4.3.

Proposition 5.1.5 *Let F be a Sturmian set and let $X \subset F$ be a prefix code. Then X contains at most one left-special word. If X is a finite F-maximal prefix code, it contains exactly one left-special word.*

Proof. Assume on the contrary that $x, y \in X$ are two left-special words. We may assume that $|x| < |y|$. Let x' be the prefix of y of length $|x|$. Then x' is left-special and thus x, x' are two left-special words of the same length. This implies that $x = x'$. Thus x is a prefix of y. Since X is prefix, this implies $x = y$.

Assume now that X is a finite F-maximal prefix code. Let n be the maximal length of the words in X. Let $u \in F$ be the left-special word of length n. Since XA^* is right F-dense, there is a prefix x of u which is in X. Thus x is a left-special element of X. It is unique by the previous statement. ∎

A dual of Proposition 5.1.5 holds for suffix codes and right-special words.

5.2 Cardinality

The following result shows that Theorem 4.4.3 can be made much more precise for Sturmian sets.

Theorem 5.2.1 *Let F be a Sturmian set on an alphabet with k letters. For any finite F-maximal bifix code $X \subset F$, one has* $\mathrm{Card}(X) = (k-1)d_F(X) + 1$.

The following corollary is strong generalization of a result related to Sturmian words.

Corollary 5.2.2 *Let x be a Sturmian word over $A = \{a, b\}$, and let $X \subset A^+$ be a finite maximal bifix code of degree d. Then* $\mathrm{Card}(X \cap F(x)) = d + 1$.

Indeed, since A^d is a finite maximal bifix code of degree d, this corollary (re)proves that any Sturmian word x has $d+1$ factors of length d, and it extends this to arbitrary finite maximal bifix code of degree d. A similar extension holds for strict episturmian words.

Proof of Corollary 5.2.2. Set $F = F(x)$. In view of Theorem 4.2.11, one has $d = d_F(X \cap F)$. Consequently, by the formula of Theorem 5.2.1, $\mathrm{Card}(X \cap F) = d_F(X) + 1 = d + 1$. ∎

The proof of Theorem 5.2.1 uses two lemmas.

Lemma 5.2.3 *Let F be a Sturmian set. Let $X \subset F$ be a finite bifix code of finite F-degree d and let P be the set of proper prefixes of X. There exists a right-special word $u \in F$ such that $\pi_X(u) = d$. The d suffixes of u which are in P are the right-special words contained in P.*

Proof. Let $n \geq 1$ be larger than the length of the words of X. By definition, there is a right-special word u of length n. Then u is not a factor of a word of X. By Theorem 4.2.8 it implies that $\pi_X(u) = d_F(X)$.

By the dual of Proposition 4.1.3, the word u has $d_F(X)$ suffixes which are in P. They are all right-special words. Furthermore, any right-special word p contained in P is a suffix of u. Indeed, the suffix of u of the same length than p is the unique right-special word of this length. ∎

The next lemma is a well-known property of trees translated into the language of prefix codes. Let X be a prefix code or the set $\{1\}$ and let P be the set of proper prefixes of X. For $p \in P$, let $d(p) = \mathrm{Card}\{a \in A \mid pa \in P \cup X\}$.

Lemma 5.2.4 *Let A be an alphabet with k letters. Let $X \subset A^*$ be a finite prefix code or the set $\{1\}$ and let P be the set of proper prefixes of the words of X. Assume that for all $p \in P$, $d(p) = k$ or 1. Let $Q_X = \{p \in P \mid d(p) = k\}$. Then,* $\mathrm{Card}(X) = (k-1)\,\mathrm{Card}(Q_X) + 1$.

Proof. Let us prove the property by induction on the maximal length n of the words in X. The property is true for $n = 0$ since in this case $X = \{1\}$ and $P = Q_X = \emptyset$. Assume $n \geq 1$. If $1 \notin Q_X$, then all words of X begin with the same letter a. We have then $X = aY$, Y is a prefix code or the set $\{1\}$ and $\mathrm{Card}(Q_Y) = \mathrm{Card}(Q_X)$. Hence, by induction hypothesis $\mathrm{Card}(X) = \mathrm{Card}(Y) = (k-1)\,\mathrm{Card}(Q_Y) + 1 = (k-1)\,\mathrm{Card}(Q_X) + 1$. Otherwise, $X = \cup_{a \in A} aX_a$. Set $t_a = \mathrm{Card}(Q_{X_a})$. We have $\sum_{a \in A} t_a = \mathrm{Card}(Q_X) - 1$. By induction hypothesis, $\mathrm{Card}(X_a) = (k-1)t_a + 1$. Therefore, $\mathrm{Card}(X) = \sum_{a \in A} \mathrm{Card}(X_a) = \sum_{a \in A} (k-1)t_a + k = (k-1)\,\mathrm{Card}(Q_X) + 1$. ∎

Proof of Theorem 5.2.1. Let P be the set of proper prefixes of X. An element p of P satisfies $pA \subset P \cup X$ if and only it is right-special. Thus the conclusion follows directly by Lemmas 5.2.3 and 5.2.4. ∎

code	kernel	derived code
aab, aba, baa, bab	$\emptyset$	aa, ab, ba
$aa, aba, baab, bab$	aa	
$aaba, ab, baa, baba$	ab	
$aab, abaa, abab, ba$	ba	
$aa, ab, baaba, baba$	aa, ab	
$aa, abaab, abab, ba$	aa, ba	
$aabaa, aababaa, ab, ba$	ab, ba	
$a, baabaab, baabab, babaab$	a	$a, baab, bab$
$a, baab, babaabaabab, babaabab$	$a, baab$	
$a, baabaab, baababaab, bab$	a, bab	
$aaba, abaa, ababa, b$	b	aa, aba, b
$aa, abaaba, ababa, b$	aa, b	
$aabaa, aababa, aba, b$	aba, b	

Table 5.1: The 13 F-maximal bifix codes of F-degree 3 in the Fibonacci set F.

Example 5.2.5 Let F be the Fibonacci set. We have seen in Example 4.3.13 that there are 3 F-maximal bifix codes of F-degree 2. It appears that there are 13 F-maximal bifix codes of degree 3 listed in Table 5.1. These codes are determined by their derived F-maximal bifix codes of F-degree 2, and by the choice of the kernel. The construction of the code can be done by Theorem 4.3.11. By Theorem 5.2.1, all these codes have 4 elements.

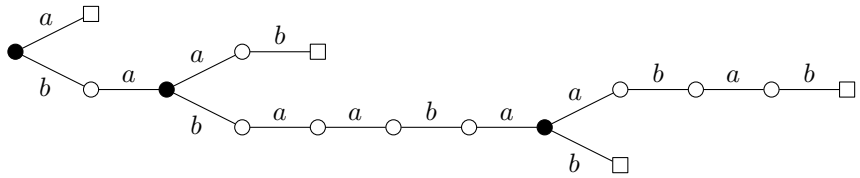

Figure 5.1: The F-maximal bifix code of F-degree 3 with kernel $\{a, baab\}$.

Example 5.2.6 We may illustrate the proof of Theorem 5.2.1 on the code $X = \{a, baab, babaabaabab, babaabab\}$ (see Table 5.1). According to Lemma 5.2.3, the right-special word $ababaaba$ (which is the reversal of the prefix $abaababa$ of the Fibonacci word) has exactly three suffixes which are proper prefixes of words of X, namely 1, ba and $babaaba$ (these are the "fork nodes", that is the nodes with two childred, indicated in black on Figure 5.1). This implies, by Lemma 5.2.4, that X has four elements.

The following example shows that Theorem 5.2.1 is not true for the set of factors of an episturmian word which is not strict.

Example 5.2.7 Set $A = \{a, b, c\}$. Let y be the Fibonacci word and let $x = \psi_c(y)$ be the infinite word of Example 2.3.7. It is an episturmian word which is not strict. Set $F = F(x)$. Let $\psi : A^* \to G$ be the morphism from A^* onto the group $G = (\mathbb{Z}/2\mathbb{Z})^3$ defined by $\psi(a) = (1, 0, 0)$, $\psi(b) = (0, 1, 0)$ and $\psi(c) = (0, 0, 1)$. Let Z be the group code such that $Z^* = \psi^{-1}(0, 0, 0)$. Since G has 8 elements, the degree of Z is 8 (see Proposition 6.1.5 below). The bifix code $X = Z \cap F$ has 10 elements obtained by inserting c in two possible ways in the 5 words of the bifix code $Z \cap F(y)$. The latter has degree 4 by Theorem 5.2.1. The bifix code $X = Z \cap F$ is given in Figure 5.2. The numbering of the nodes is for later use, in Example 7.2.7.

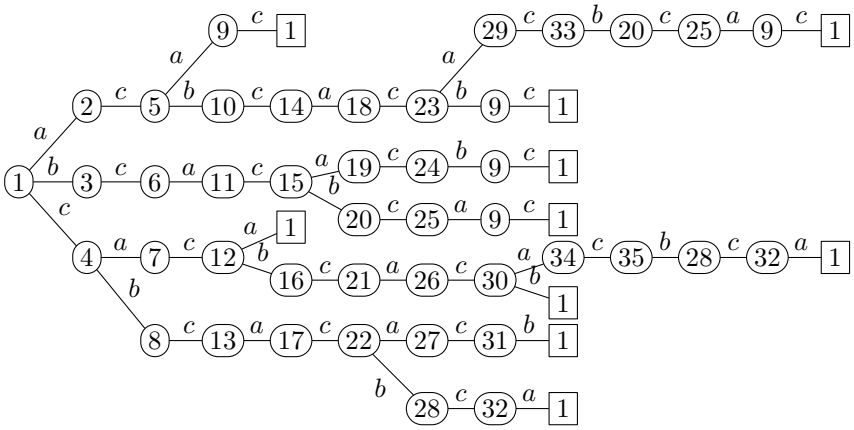

Figure 5.2: An F-maximal bifix code with 10 elements. The numbers in the vertices are for later use.

By Theorem 4.2.11, X is an F-maximal bifix code. Its F-degree is 8. Indeed, the word *acbcacbc* has 8 parses. Thus Theorem 5.2.1 is not true in this case.

As a consequence of Theorem 5.2.1, an internal transformation does not change the cardinality of a finite F-maximal bifix code for a Sturmian set F. Indeed, by Proposition 4.4.5, an internal transformation preserves the F-degree.

Actually, if Y is obtained from X by internal transformation with respect to w, we have

$$Y = (X \cup w \cup (GwD \cap F)) \setminus (Gw \cup wD) \tag{5.1}$$

and

$$\mathrm{Card}(Y) = \mathrm{Card}(X) + 1 + \mathrm{Card}(GwD \cap F) - \mathrm{Card}(G) - \mathrm{Card}(D).$$

The fact that internal transformations preserve the cardinality can be proved directly by the following statement. This statement applies to the internal transformation (5.1) because $Gw \cup wD$ is a bifix code, which implies property (i) and $DA^* = w^{-1}XA^*$ (resp. $A^*G = A^*Xw^{-1}$) which implies property (ii) (resp. (iii)).

Proposition 5.2.8 *Let F be Sturmian set, let $w \in F$ be a nonempty word and let D, G be finite sets such that*

(i) *any word u has at most one factorization $u = gwd$ with $g \in G$ and $d \in D$,*
(ii) *wD is a prefix code contained in F and DA^* is right $w^{-1}F$-dense,*
(iii) *Gw is a suffix code contained in F and A^*G is left Fw^{-1}-dense.*

Then $\mathrm{Card}(GwD \cap F) = \mathrm{Card}(G) + \mathrm{Card}(D) - 1$.

Proof. Let $V = (1 \otimes G) \cup (D \otimes 1)$ be a set made of copies of G and D. The tensor product notation is used to emphasize that the copies of G and D are disjoint. Let $H = (V, E)$ be the undirected graph having V as set of vertices and as edges the pairs $\{1 \otimes g, d \otimes 1\}$ such that $gwd \in F$ (this graph is close to, but slightly different from the incidence graph for GwD as it will be defined in Section 6.3). We have $\mathrm{Card}(V) = \mathrm{Card}(G) + \mathrm{Card}(D)$ and, by condition (i), $\mathrm{Card}(E) = \mathrm{Card}(GwD \cap F)$. We show that the graph H is a tree. This implies our conclusion since, in a tree, one has $\mathrm{Card}(E) = \mathrm{Card}(V) - 1$.

Let us prove that the graph H is a tree by induction on the sum of the lengths of the words of D, assuming that the pair G, D satisfies conditions (ii) and (iii). Assume first that $D = \{1\}$. Since $Gw \subset F$, one has $GwD \subset F$. Consequently, $\{1 \otimes g, 1 \otimes 1\} \in E$ for any $g \in G$. Thus H is a tree.

Next, assume that $D \neq \{1\}$. Let $d \in D$ be of maximal length. Set $d = d'a$ with $a \in A$.

Suppose first that $d'A \cap D = \{d\}$. Let $D' = (D \cup d') \setminus d$. Since DA^* is $w^{-1}F$-dense, the word wd' is not right-special. Thus for each $g \in G$, we have $gwd' \in F$ if and only if $gwd \in F$. This shows that the graph H is isomorphic to the graph H' corresponding to the pair (G, D'). The set D' satisfies condition (ii). By induction H' is a tree. Consequently H is a tree.

Suppose next that $d'A \cap D$ has more than one element. Then d' is right-special and $d'A \cap D = d'A$. Let $D' = (D \cup d') \setminus d'A$. Then D' satisfies condition (ii). Let H' be the graph corresponding to the pair (G, D'). By induction hypothesis, the graph H' is tree. Since $wD \subset F$, wd' is right-special. Let uwd' be a right-special word such that u is longer than any word of G. Since A^*G is Fw^{-1}-dense, and since $u \in Fw^{-1}$, u has a suffix g in G. Thus gwd' is right-special. We have $\{1 \otimes g, d'a \otimes 1\} \in E$ for all $a \in A$. For any other element $g' \in G$ such that $g'wd' \in F$, since $g'wd'$ is not right-special, there is exactly one $a' \in A$ such that $g'wd'a' \in F$. There is a path between g and every $g' \neq g$, since $\{1 \otimes g', 1 \otimes d'a'\} \in E$ for some a' and $\{1 \otimes g, 1 \otimes d'a'\} \in E$ for all a' (see Figure 5.3). Thus the graph H is connected and acyclic, and therefore is a tree. ∎

The following example shows that condition (i) is necessary.

Example 5.2.9 Let F be the Fibonacci set. Let $G = \{ab, aba\}$, $w = a$ and $D = \{ab, b\}$. Then conditions (ii) and (iii) are satisfied but not condition (i). We have $GwD = \{abaab, abab\}$ and thus the conclusion of Proposition 5.2.8 is false.

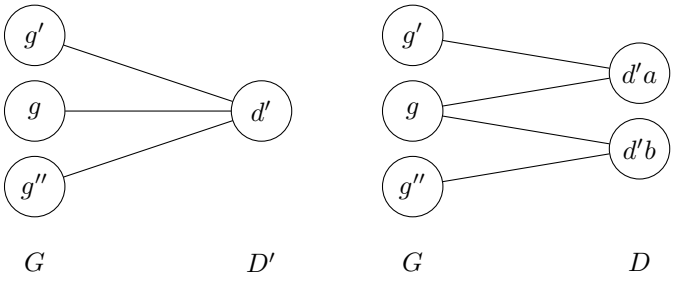

Figure 5.3: The graphs H' and H.

Example 5.2.10 Let F be the Fibonacci set and let $X = \{aaba, abaa, abab, baab, baba\}$ be the set of words of F of length 4. The internal transformation from X relative to the word $w = aba$ gives $Y = \{aabaa, aabab, aba, baab, babaa\}$. We have $G = D = \{a, b\}$. The codes X, Y and the graph H of the proof of Proposition 5.2.8 are represented on Figure 5.4.

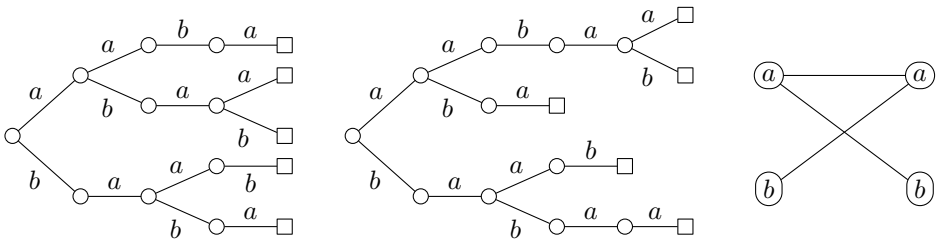

Figure 5.4: The codes X, Y and the graph H.

5.3 Periodicity

Let $x = a_0a_1 \cdots$, with $a_i \in A$, be an infinite word. It is *periodic* if there is an integer $n \geq 1$ such that $a_{i+n} = a_i$ for all $i \geq 0$. It is *ultimately periodic* if the equalities hold for all i large enough. Thus, x is ultimately periodic if there is a word u and a periodic infinite word y such that $x = uy$. The following result, due to Coven and Hedlund, is well-known (see [36], Theorem 1.3.13).

Theorem 5.3.1 *Let $x \in A^{\mathbb{N}}$ be an infinite word on an alphabet with k letters. If there exists an integer $d \geq 1$ such that x has at most $d + k - 2$ factors of length d then x is ultimately periodic.*

We will prove a generalization of this result.

Let x be an infinite word and let X be a thin maximal bifix code. Let u be a prefix of x and set $x = uy$. Since $F(y) \subset F(x)$, one has $d_{F(y)}(X) \leq d_{F(x)}(X)$. The word x is called X-*stable* if $d_{F(y)}(X) = d_{F(x)}(X)$ for all suffixes y of x. Let u be a prefix of x such that $d_{F(y)}(X)$ is minimal. Then the infinite word y is X-stable.

For example, if $x = ba^\omega$ and $X = a \cup ba^*b$, then an X-stable suffix of x is a^ω.

Theorem 5.3.2 *Let X be a thin maximal bifix code and let $x \in A^{\mathbb{N}}$ be an X-stable infinite word. If $\mathrm{Card}(X \cap F(x)) \leq d_{F(x)}(X)$, then x is ultimately periodic.*

Corollary 5.3.3 *Let $x \in A^{\mathbb{N}}$ be an infinite word. If there exists a finite maximal bifix code X of degree d such that $\mathrm{Card}(X \cap F(x)) \leq d$, then x is ultimately periodic.*

Proof. Since any long enough word has d parses, $d_{F(x)}(X) = d$ and x is X-stable. Since $\mathrm{Card}(X \cap F(x)) \leq d$, the conclusion follows by Theorem 5.3.2. ∎

Corollary 5.3.3 implies Theorem 5.3.1 in the case $k = 2$ since A^d is a maximal bifix code of degree d.

Example 5.3.4 Let us consider again the finite maximal bifix code X of degree 3 defined by $X = \{a^3, a^2ba, a^2b^2, ab, ba^2, baba, bab^2, b^2a, b^3\}$ (see Example 4.2.12). Assume that $X \cap F = \{a^2ba, ab, baba\}$, where $F = F(x)$ and $x \in A^{\mathbb{N}}$. Since ba is a factor of x, there exist a word u and an infinite word y such that $x = ubay$. Next, the first letter of y is b (otherwise, $ba^2 \in X \cap F$) and the second letter of y is a (otherwise, $bab^2 \in X \cap F$). This argument shows that whenever uba is a prefix of x then $ubaba$ is also a prefix of x, i.e., $x = u(ba)^\omega$, with $u \in A^*$.

Example 5.3.5 The set $X = a \cup ba^*b$ is a maximal bifix code of degree 2. An argument similar to the previous one shows that any infinite word $x \in A^{\mathbb{N}}$ such that $X \cap F(x) = \{a, bab\}$ belongs to the set $a^*(ba)^\omega$. Thus it is ultimately periodic.

Corollary 5.3.6 *Let $x \in A^{\mathbb{N}}$ be an infinite word and let X be a thin maximal bifix code. Let y be an X-stable suffix of x and let $F = F(y)$. If $\mathrm{Card}(X \cap F) \leq d_F(X)$, then x is ultimately periodic.*

Proof. By Theorem 5.3.2, the word y is ultimately periodic, and so is x. ∎

The following example shows that Corollary 5.3.6 may become false if we replace $F = F(y)$ by $F = F(x)$ in the statement.

Example 5.3.7 Let X be the maximal bifix code of degree 4 on the alphabet $A = \{a, b, c\}$ with kernel $K = \{a, b\}^2$.

Let $x = ccay$ where y is an infinite word without any occurrence of c. Then cca has no factor in X. Indeed, a word of X of length at most 3 is in the kernel of X and thus is not a factor of cca. Thus cca has 4 parses with respect to X, namely $(1, 1, cca)$, $(c, 1, ca)$, $(cc, 1, a)$ and $(cca, 1, 1)$. Thus we have $d_{F(x)}(X) = 4$. On the other hand $X \cap F(x) \subset \{a, b\}^2$ and thus $\mathrm{Card}(X \cap F(x)) \leq d_{F(x)}(x)$ although x need not be ultimately periodic. This shows that we cannot replace $F(y)$ by $F(x)$ in the statement of Corollary 5.3.6.

The proof uses the Critical Factorization Theorem (see [35, 16]) that we recall below. For a pair of words $(p, s) \neq (1, 1)$, consider the set of nonempty words r such that

$$A^*p \cap A^*r \neq \emptyset, \quad sA^* \cap rA^* \neq \emptyset\,.$$

This is the set of nonempty words r which are prefix-comparable with s and suffix-comparable with p. This set is nonempty since it contains $r = sp$. The *repetition* $\mathrm{rep}(p, s)$ is the minimal length of such a nonempty word r.

Let $w = a_1a_2 \cdots a_m$ be a word with $a_i \in A$. An integer $n \geq 1$ is a *period* of w if for $1 \leq i \leq j \leq m$, $j - i = n$ implies $a_i = a_j$. Recall that a *factorization* of a word $w \in A^*$ is a pair (p, s) of words such that $w = ps$.

Theorem 5.3.8 (Critical Factorization Theorem) *For any word $w \in A^+$, the maximal value of $\mathrm{rep}(p, s)$ for all factorizations (p, s) of w is the least period of w.*

We will also use the following lemma.

Lemma 5.3.9 *Let x be an infinite word and $n \geq 1$ be an integer such that the least period of an infinite number of prefixes of x is at most n. Then x is periodic.*

Proof. Since the least period of an infinite number of prefixes of x is at most n, an infinity of them have the same least period. Let p be such that an infinite number of prefixes of x have least period p. Set $x = a_0a_1 \cdots$ with $a_i \in A$. For each $i \geq 0$, there is a prefix of x of length larger than $i + p$ with least period p. Thus $a_i = a_{i+p}$. This shows that x is periodic. ∎

Proof of Theorem 5.3.2.

Let $S = A^* \setminus A^*X$ and $P = A^* \setminus XA^*$. Set $F = F(x)$ and $d = d_F(X)$. Since $\mathrm{Card}(X \cap F) \leq d_F(X) \leq d(X)$, the set $X \cap F$ is finite. Since x is X-stable, there are an infinite number of factors and therefore also of prefixes of x which have d parses with respect to X. Indeed, for any factorization $x = uy$, we have $d_{F(y)}(X) = d$ and thus y has a factor which has d parses, so it has a prefix w with d parses, and finally uw is a prefix of x with d parses.

Let n be the maximal length of the words in $X \cap F$. Let u be a prefix of x of length larger than n which has d parses and set $x = uy$. Let w be a nonempty prefix of y and set $y = wz$. Let v be a prefix of z of length larger than n which has d parses.

Let (p, s) be a factorization of w. We show that $\mathrm{rep}(p, s) \leq n$.

Since up has d parses with respect to X, there are d suffixes $p_1, p_2, \ldots, p_d$ of up which are in P. We may assume that $p_1 = 1$. Similarly, there are d prefixes $s_1, s_2, \ldots, s_d$ of sv which are in S. We may assume that $s_1 = 1$.

Since $upsv$ has d parses, for each p_i with $2 \leq i \leq d$ there is exactly one s_j with $2 \leq j \leq d$ such that $p_is_j \in X$. Indeed, there is a prefix s' of sv such that $p_is' \in X$. Since s' must be one of the s_j, the conclusion follows.

We may renumber the s_i in such a way that $p_i s_i \in X$ for $2 \leq i \leq d$. Set $x_i = p_i s_i$. Since $up \notin S$, we have $up \in A^*X$. Let x_0 be the word of X which is a suffix of up. Similarly, let x_1 be the word of X which is a prefix of sv (see Figure 5.5).

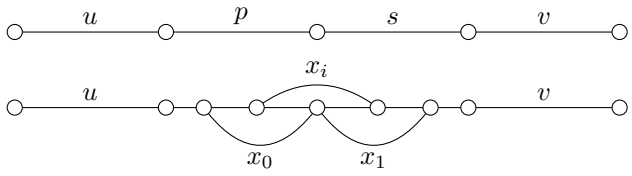

Figure 5.5: The $d+1$ words $x_0, x_1, \ldots, x_d$.

Since $\mathrm{Card}(X \cap F) \leq d$, two of the $d+1$ words $x_0, x_1, \ldots, x_d$ are equal.

If $x_0 = x_1$, then $\mathrm{rep}(p, s) \leq n$.

If $x_0 = x_i$ for an index i with $2 \leq i \leq d$, then s_i is a suffix of up (since it is a suffix of x_0) and a prefix of sv (by definition of s_i). Furthermore $|s_i| \leq n$ (since n is the maximal length of the words of $X \cap F$). Thus $\mathrm{rep}(p, s) \leq |s_i| \leq n$.

The case where $x_i = x_1$ for an index i with $2 \leq i \leq d$ is similar.

Assume finally that $x_i = x_j$ for some indices i, j such that $2 \leq i < j \leq d$. We may assume that $|p_i| < |p_j|$. Thus $p_j = p_i t$, $ts_j = s_i$. As a consequence, t is both a suffix of up (since it is a suffix of p_j) and a prefix of sv (since it is a prefix of s_i). Thus again, $\mathrm{rep}(p, s) \leq |t| \leq n$.

By the Critical Factorization Theorem, this implies that the least period of w is at most equal to n. Thus an infinite number of prefixes of y have least period at most n. By Lemma 5.3.9, it implies that y is periodic.

■

6 Bases of subgroups

In this section, we push further the study of bifix codes in Sturmian sets. The main result of Section 6.2 is Theorem 6.2.1. It states that a F-maximal bifix code $X \subset F$ of F-degree d is a basis of a subgroup of index d of the free group on A. The proof uses two sets of preliminary results. The first part concerns bases of subgroups composed of words over A, already considered in [48]. The second one uses the first return words, which were introduced independently in [23], [28], and which we use in the framework of [32] and [56], up to a left-right symmetry (see also [1]).

We denote by A° the free group generated by A. The *rank* of A° is $\mathrm{Card}(A)$. Note that all sets generating a free group of rank k have at least k elements. A basis is a minimal generating set. All bases have exactly k elements (see e.g. [37]).

Let H be a subgroup of rank n and of index d of a free group of rank k. Then

$$n = d(k-1) + 1\,. \tag{6.1}$$

Formula (6.1) is called *Schreier's Formula*.

The free monoid A^* is viewed as embedded in A°. An element of the free group is represented by its unique reduced word on the alphabet $A \cup A^{-1}$. The elements of the free monoid A^* are themselves reduced words since they do not contain any letter in A^{-1}. Thus A^* is a submonoid of A°. The subgroup of A° generated by a subset X of A° is denoted $\langle X \rangle$.

In any group G, the *right cosets* of a subgroup H are the sets of the form Hg for $g \in G$. Two right cosets of the same subgroup are disjoint or equal. The *index* of a subgroup is the number of its distinct right cosets. If K is a subgroup of the subgroup H, then the index of K in G is the product of the index of K in H and of the index of H in G. If H, K are two subgroups of index d of a group G, then $H \subset K$ implies $H = K$.

Assume now that G is a group of permutations over a set Q. For any q in Q, the set of elements of G that fixes q is a subgroup of G.

The group G is *transitive* if, for all $p, q \in Q$, there is an element $g \in G$ such that $pg = q$. In this case, the subgroup H of permutations fixing a given element p of Q has index $\mathrm{Card}(Q)$. Indeed, for each $q \in Q$ let g_q be an element of G such that $pg_q = q$. If $g \in G$ is such that $pg = q$, then $pgg_q^{-1} = p$ and consequently $gg_q^{-1} \in H$, whence $g \in Hg_q$. Thus each $g \in G$ is in one of the right cosets Hg_q, for $q \in Q$. Since these right cosets are pairwise disjoint, the index of H is $\mathrm{Card}(Q)$.

6.1 Group automata

A simple automaton $\mathcal{A} = (Q, 1, 1)$ is said to be *reversible* if for any $a \in A$, the partial map $\varphi_{\mathcal{A}}(a) : p \mapsto p \cdot a$ is injective. This condition allows to construct the *reversal* of the automaton as follows: whenever $q \cdot a = p$ in $\mathcal{A}$, then $p \cdot a = q$ in the reversal automaton. The state 1 is the initial and the unique terminal state of this automaton. Any reversible automaton is minimal [48]. The set recognized by a reversible automaton is a left and right unitary submonoid. Thus it is generated by a bifix code.

An automaton $\mathcal{A} = (Q, 1, 1)$ is a *group automaton* if for any $a \in A$ the map $\varphi_{\mathcal{A}}(a) : p \mapsto p \cdot a$ is a permutation of Q. When Q is finite, a group automaton is a reversible automaton which is complete.

The following result is from [48] (see also Exercise 6.1.2 in [6]).

Proposition 6.1.1 *Let $X \subset A^+$ be a bifix code. The following conditions are equivalent.*

(i) $X^* = \langle X \rangle \cap A^*$;

(ii) *the minimal automaton of X^* is reversible.*

Let $\mathcal{A} = (Q, i, T)$ be a deterministic automaton. A *generalized path* is a sequence $(p_0, a_1, p_1, a_2, \ldots, p_{n-1}, a_n, p_n)$ with $a_i \in A \cup A^{-1}$ and $p_i \in Q$, such that for $1 \leq i \leq n$, one has $p_{i-1} \cdot a_i = p_i$ if $a_i \in A$ and $p_i \cdot a_i^{-1} = p_{i-1}$ if $a_i \in A^{-1}$. The *label* of the generalized path is the element $a_1 a_2 \cdots a_n$ of the free group A°.

Note that if $\mathcal{A} = (Q, 1, 1)$, the set of labels of generalized paths from 1 to 1 in $\mathcal{A}$ is a subgroup of A°. It is called the *subgroup described* by $\mathcal{A}$.

A path in an automaton is a particular case of a generalized path. In the case where $\mathcal{A}$ has a unique terminal state which is equal to the initial state, the submonoid of A^* recognized by $\mathcal{A}$ is contained in the subgroup of A° described by $\mathcal{A}$.

Example 6.1.2 Let $\mathcal{A} = (Q, 1, 1)$ be the automaton defined by $Q = \{1, 2\}$, $1 \cdot a = 1 \cdot b = 2$ and $2 \cdot a = 2 \cdot b = \emptyset$. The submonoid recognized by $\mathcal{A}$ is $\{1\}$. The subgroup described by $\mathcal{A}$ is the cyclic group generated by ab^{-1}.

For any subgroup H of A°, the submonoid $H \cap A^*$ is right and left unitary. Thus $H \cap A^*$ is generated by a bifix code.

Proposition 6.1.3 *Let $\mathcal{A}$ be a simple automaton and let X be the prefix code generating the submonoid recognized by $\mathcal{A}$. The subgroup described by $\mathcal{A}$ is generated by X.*

Proof. Set $H = \langle X \rangle$. Let K be the subgroup described by $\mathcal{A}$. Let us show that $K = H$. First, $X \subset K$ implies $\langle X \rangle = H \subset K$. To prove the converse inclusion, let $h = a_1 a_2 \cdots a_n \in K$ with $a_i \in A \cup A^{-1}$. Let r be the number of indices i such that $a_i \in A^{-1}$. We show by induction on r that $h \in H$. This holds clearly if $r = 0$. Assume that it is true for $r - 1$. Let i be the least index such that $a_i \in A^{-1}$. Set $u = a_1 \cdots a_{i-1}$, $a = a_i^{-1}$, $v = a_{i+1} \cdots a_n$ in such a way that $h = ua^{-1}v$. Set also $p = 1 \cdot u$ and define $q \in Q$ by $q \cdot a = p$. Since $\mathcal{A}$ is trim there exist words $w, t \in A^*$ such that $p \cdot t = 1$ and $1 \cdot w = q$. Since $1 \cdot ut = 1 \cdot wat = 1 \cdot wv$ (see Figure 6.1), we have $ut, wat \in X^*$. By induction hypothesis, we have $wv \in H$. Then $ua^{-1}v = utt^{-1}a^{-1}w^{-1}wv = ut(wat)^{-1}wv$ is in H and thus $h \in H$. Thus $K \subset H$ and this completes the proof that $K = H$. ∎

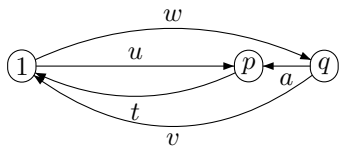

Figure 6.1: Paths in the automaton $\mathcal{A}$

A subgroup H of A° is *positively generated* if there is a set $X \subset A^*$ which generates H. In this case, the set $H \cap A^*$ generates the subgroup H. Let X be the bifix code which generates the submonoid $H \cap A^*$. Then X generates the subgroup H. This shows that, for a positively generated subgroup H, there is a bifix code which generates H.

Proposition 6.1.4 *For any positively generated subgroup H of A°, there is a unique reversible automaton $\mathcal{A}$ such that H is the subgroup described by $\mathcal{A}$.*

Proof. Let X be the bifix code generating the submonoid $H \cap A^*$, so that $X^* = H \cap A^*$. Since H is positively generated, the subgroup generated by X is equal to H, that is $\langle X \rangle = H$. Thus $X^* = \langle X \rangle \cap A^*$. In view of Proposition 6.1.1, the minimal automaton $\mathcal{A}$ of X^* is reversible.

If $\mathcal{B}$ is another reversible automaton such that H is the subgroup described by $\mathcal{B}$, then $\mathcal{B}$ recognizes the set $H \cap A^*$. Since $\mathcal{B}$ is minimal and since minimal automata are unique, the uniqueness follows.

The submonoid recognized by $\mathcal{A}$ is $H \cap A^*$. Thus the last assertion follows. ∎

The reversible automaton $\mathcal{A}$ such that H is the subgroup described by $\mathcal{A}$ is called the *Stallings automaton* of the subgroup H. It can also be defined for a subgroup which is not positively generated (see [3] or [33]).

Proposition 6.1.5 *The following conditions are equivalent for a submonoid M of A^*.*

(i) *M is recognized by a group automaton with d states.*
(ii) *$M = \varphi^{-1}(K)$, where K is a subgroup of index d of a group G and φ is a surjective morphism from A^* onto G.*
(iii) *$M = H \cap A^*$, where H is a subgroup of index d of A°.*

If one of these conditions holds, the minimal generating set of M is a maximal bifix code of degree d.

Proof. (i) implies (ii). Let $\mathcal{A} = (Q, 1, 1)$ be a group automaton with d states and let M be the set recognized by $\mathcal{A}$. Since a composition of permutations is a permutation, the monoid $G = \varphi_{\mathcal{A}}(A^*)$ is a permutation group. Since $\mathcal{A}$ is trim, there is a path from every state q to any state q' in $\mathcal{A}$. Consequently, G is transitive. Let K be the subgroup of G formed of the permutations fixing 1. As we have seen earlier, K has index d. Then $M = \varphi_{\mathcal{A}}^{-1}(K)$.
(ii) implies (iii). Let ψ be the morphism from A° onto G extending φ. Then $H = \psi^{-1}(K)$ is a subgroup of index d of A° and $M = H \cap A^*$.
(iii) implies (i). Let Q be the set of right cosets of H with 1 denoting the right coset H. The representation of A° by permutations on Q defines a group automaton $\mathcal{A}$ with d states and M is recognized by $\mathcal{A}$.

Finally, let X be the minimal generating set of a submonoid M satisfying one of these conditions. It is clearly a bifix code. Let P be the set of proper prefixes of X. The number of suffixes of a word which are in P is at most equal to d. Indeed, let $\mathcal{A} = (Q, 1, 1)$ be a group automaton recognizing X^*. If s, t are distinct suffixes of a word w which are in P, then $1 \cdot s \neq 1 \cdot t$. Indeed, otherwise, since s and t are suffix-comparable, we may assume that $s = ut$. Let $p = 1 \cdot u$. Then $p \cdot t = 1 \cdot ut = 1 \cdot s = 1 \cdot t$ and thus $p = 1$ since $\mathcal{A}$ is reversible. Thus $s = t$. Let w be a word with the maximal number of suffixes in P. Then w cannot be an internal factor of X. Moreover the number of suffixes of w in P is equal to d. Indeed, since $\mathcal{A}$ is a group code, for any $q \in Q$, there is a state q' such that $q' \cdot w = q$. Since w is not an internal factor of X, there is a factorization

$w = sxp$ such that $q' \cdot s = 1$, $1 \cdot x = 1$ and $1 \cdot p = q$, and such that (s, x, p) is a parse of w. Thus X has degree d. ∎

A bifix code Z such that Z^* satisfies one of the equivalent conditions of the proposition 6.1.5 is called a *group code*.

The following proposition shows in particular that a subgroup of finite index is positively generated.

Proposition 6.1.6 *Let H be a subgroup of finite index of A°. The minimal automaton $\mathcal{A}$ of $H \cap A^*$ is a group automaton which describes the subgroup H. Let X be the group code such that $\mathcal{A}$ recognizes X^*. The subgroup generated by X is H.*

Proof. By Proposition 6.1.5, the monoid $H \cap A^*$ is recognized by a group automaton $\mathcal{A} = (Q, 1, 1)$, which is the minimal automaton of $H \cap A^*$. The morphism $\varphi_{\mathcal{A}}$ from A^* onto the group $G = \varphi_{\mathcal{A}}(A^*)$ of Proposition 6.1.5 extends to a morphism ψ from A° onto G. The subgroup K is composed of the permutations that fix 1, and the subgroup H is formed of the elements $w \in A^\circ$ such that the permutation $\psi(w)$ fixes 1. There is a generalized path in $\mathcal{A}$ from p to q labeled w if and only if $p\psi(w) = q$. Thus $\psi(w)$ fixes 1 if and only there is a generalized path from 1 to 1 labeled w, that is if $w \in H$. Thus the subgroup described by $\mathcal{A}$ is H. By Proposition 6.1.3, the subgroup H is generated by X. ∎

Example 6.1.7 The set A^d is a group code by Proposition 6.1.5(ii). Thus it is a maximal bifix code of degree d. The intersection of the subgroup generated by A^d with A^* is the submonoid generated by A^d (Proposition 6.1.6). It is composed of the words with length a multiple of d.

6.2 Main result

We will prove the following result.

Theorem 6.2.1 *Let F be a Sturmian set and let $d \geq 1$ be an integer. A bifix code $X \subset F$ is a basis of a subgroup of index d of A° if and only if it is a finite F-maximal bifix code of F-degree d.*

Note that Theorem 5.2.1 is contained in Theorem 6.2.1 (we will use Theorem 5.2.1 in the proof of Theorem 6.2.1). Indeed, let X be an F-maximal bifix code of F-degree d. By Theorem 4.4.3, X is finite. By Theorem 6.2.1, the subgroup $\langle X \rangle$ has rank $\mathrm{Card}(X)$ and index d in the free group A°. By Schreier's Formula (6.1), one get $\mathrm{Card}(X) = (\mathrm{Card}(A) - 1)d + 1$.

Before proving Theorem 6.2.1, we list some corollaries.

Corollary 6.2.2 *Let F be a Sturmian set. For any $d \geq 1$, the set of words in F of length d is a basis of the subgroup of A° generated by A^d.*

Proof. The set A^d is a group code (see Example 6.1.7), and therefore is a maximal bifix code. The set $A^d \cap F$ is a finite bifix code. By Theorem 4.2.11, it is an F-maximal bifix code and has F-degree d. The corollary follows from Theorem 6.2.1. Indeed $\langle A^d \cap F\rangle = \langle A^d\rangle$ since both are subgroups of A° of index d. ∎

The following is also a complement to Theorem 4.2.11.

Corollary 6.2.3 *Let F be a Sturmian set. The map which associates to $X \subset F$ the subgroup $\langle X\rangle$ of A° generated by X is a bijection between F-maximal bifix codes of F-degree d and subgroups of A° of index d. Such a bifix code X is a basis of $\langle X\rangle$. The reciprocal bijection associates, to a subgroup H of A°, the set $Z \cap F$ where Z is the group code which is the minimal generating set of the submonoid $H \cap A^*$ of A^*.*

Proof. Let first X be a finite F-maximal bifix code of F-degree d. Then $\langle X\rangle$ is a subgroup of index d by Theorem 6.2.1.

Conversely, let H be a subgroup of index d of A° and let Z be the group code such that $Z^* = H \cap A^*$. By Theorem 4.2.11, the set $X = Z \cap F$ is an F-maximal bifix code of F-degree $e \leq d$. By Theorem 4.4.3, X is finite. By Theorem 6.2.1, the subgroup $\langle X\rangle$ has index e. Since $\langle X\rangle$ is a subgroup of H, e is a multiple of d. Thus $d = e$ and $\langle X\rangle = H$.

Finally, let X be an F-maximal bifix code of F-degree d. Then $H = \langle X\rangle$ is a subgroup of index d of A°. Let Z be the group code such that $Z^* = H \cap A^*$ and let $Y = Z \cap F$. Then $X \subset Y$ and thus $X = Y$ since X is an F-maximal bifix code. This shows that the two maps are mutually inverse.

∎

A set W of words of $\{a, b\}^*$ is *balanced* if for all $w, w' \in W$, $|w| = |w'|$ implies $||w|_a - |w'|_a| \leq 1$. It is a classical property that the set of factors of a Sturmian word is balanced (Theorem 2.1.5 in [36]). Thus any Sturmian set on two letters is balanced.

Following Richomme and Séébold [50], we say that a subset X of $\{a, b\}^*$ is *factorially balanced* if the set of factors of words of X is balanced. They show that a finite set $X \subset \{a, b\}^*$ is contained in some Sturmian set if and only if it is factorially balanced. Thus, we have the following consequence of Theorem 6.2.1.

Corollary 6.2.4 *Let $X \subset \{a, b\}^*$ be a bifix code. The following conditions are equivalent.*

(i) *There exists a Sturmian set $F \subset \{a, b\}^*$ such that $X \subset F$ and X is a finite F-maximal bifix code.*

(ii) *X is a factorially balanced basis of a subgroup of finite index of $\{a, b\}^\circ$.*

As a further consequence of Theorem 6.2.1, we have the following result.

Corollary 6.2.5 *Let F be a Sturmian set on an alphabet with k letters. The number $N_{d,k}$ of finite F-maximal bifix codes $X \subset F$ of F-degree d satisfies $N_{1,k} = 1$ and*

$$N_{d,k} = d(d!)^{k-1} - \sum_{i=1}^{d-1} [(d-i)!]^{k-1} N_{i,k}.$$

The formula results directly from the formula, due to Hall [27], for the number of subgroups of index d in a free group of rank k.

The values for $k = 2$ are given by the recurrence

$$N_{d,2} = d\ d! - \sum_{i=1}^{d-1} (d-i)! N_{i,2}.$$

The first values are

d	1	2	3	4	5	6	7	8	9	10
$N_{d,2}$	1	3	13	71	461	3447	29093	273343	2829325	31998903

The values for $d = 2, 3$ are consistent with Examples 4.3.13 and 5.2.6.

The formula is known to enumerate also the indecomposable permutations on $d+1$ elements (see [21], [43] and [14]).

6.3 Incidence graph

Let X be a set, let P be the set of its proper prefixes and S be the set of its proper suffixes. Set $P' = P \setminus 1$ and $S' = S \setminus 1$. The *incidence graph* of X is the undirected graph G defined as follows. The set of vertices is $V = 1 \otimes P' \cup S' \otimes 1$. Similarly to the proof of Proposition 5.2.8, the tensor product notation is used to emphasize that V is made of two disjoint copies of P' and S'. The edges of G are the pairs $\{1 \otimes p, s \otimes 1\}$, for $p \in P'$ and $s \in S'$, such that $ps \in X$.

Let C be a connected component of G, that is a maximal set of vertices connected by paths. The *trace* of C on P' is the set of $p \in P'$ such that $1 \otimes p \in C$. Similarly, the trace of C on S' is the set of $s \in S'$ such that $s \otimes 1 \in C$.

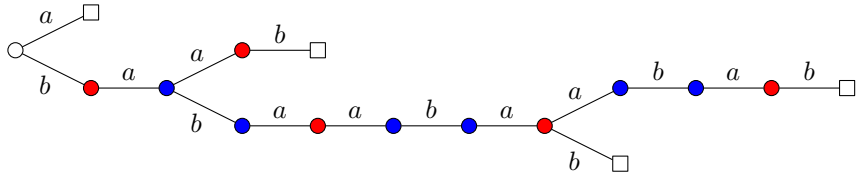

Figure 6.2: The F-maximal bifix code of F-degree 3 with kernel $\{a, baab\}$.

Example 6.3.1 Consider the F-maximal bifix code of F-degree 3 in the Fibonacci set F given in Figure 6.2. It is a colored copy of Figure 5.1. The

incidence graph of X is given in Figure 6.3. It has two connected components colored red and blue. The vertices on the left side are the $1 \otimes p$ (written simply p for convenience). The vertices on the right side are the $s \otimes 1$ with the same convention.

The color on the node in Figure 6.2 corresponds to the color of the corresponding prefix in Figure 6.3.

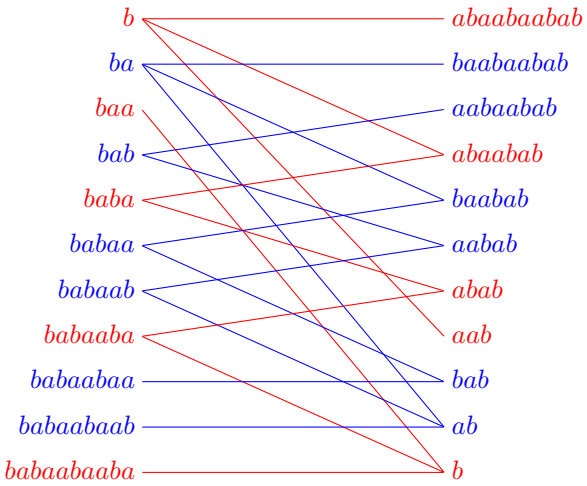

Figure 6.3: The incidence graph of X.

The following lemma uses an argument similar to Lemma 5.2.4.

Lemma 6.3.2 *Let $v_1, v_2, \ldots, v_{n+1}$ be words such that v_i, v_{i+1} are not prefix-comparable for $1 \leq i \leq n$. Let p_i be the longest common prefix of v_i, v_{i+1}, for $1 \leq i \leq n$. If two of the v_i are prefix-comparable, then two of the p_i are equal.*

Proof. Let $V = \{v_1, \ldots, v_{n+1}\}$, let P be the set of proper prefixes of V and let $W = V \setminus P$. The set W is the set of words of V which have no proper prefix in V. The set W is a prefix code. If two distinct words in V are prefix-comparable, then $\mathrm{Card}(W) < \mathrm{Card}(V) \leq n + 1$.

Let m be the number of distinct p_i. Since v_i, v_{i+1} are not prefix-comparable for $1 \leq i \leq n$, for each p_i there are at least two distinct letters a, b such that $p_i a, p_i b \in P \cup W$. This implies $m < \mathrm{Card}(W)$. Indeed, the set W can be seen as the set of leaves in a tree, and each p_i is a fork node (i. e. a node with at least two children) in this tree. It is well-known that the number of fork nodes is strictly less than the number of leaves. If two of the v_i are prefix-comparable, the inequality $\mathrm{Card}(W) < n + 1$ implies $m < \mathrm{Card}(W) \leq n$, and consequently two of the p_i are equal. ∎

Lemma 6.3.3 *Let F be a Sturmian set and let $X \subset F$ be a bifix code. Let P' (resp. S') be the set of nonempty proper prefixes (resp. suffixes) of X and let G be the incidence graph of X.*

(i) *The graph G is acyclic, that is a union of trees.*

(ii) *The trace on P' (resp. on S') of a connected component C of G is a suffix (resp. prefix) code.*

(iii) *In particular, this trace on P' (resp. on S') contains at most one right-special (resp. left-special) word.*

Proof. The last assertion follows from the second by Proposition 5.1.5. We call a path *reduced* if it does not use equal consecutive edges.

We prove by induction on $n \geq 1$ that if $s \otimes 1$ and $t \otimes 1$ (resp. $1 \otimes p$ and $1 \otimes q$) are connected by a reduced path of length $2n$ in G, then s, t are not prefix-comparable (resp. p, q are not suffix-comparable). This shows that G is acyclic. Indeed, if there were a cycle from s to $t = s$ in G, then s and t would be prefix-comparable. This shows also that two words in the same trace on P' (resp. on S') are not suffix-comparable (resp. are not prefix-comparable).

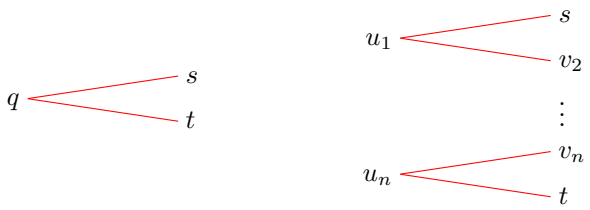

Figure 6.4: A path $(s \otimes 1, 1 \otimes q, t \otimes 1)$ on the left, and a path of length $2n$ on the right.

The property holds for $n = 1$. Indeed, a reduced path of length 2 from $s \otimes 1$ to $t \otimes 1$ is of the form $(s \otimes 1, 1 \otimes q, t \otimes 1)$ with $qs, qt \in X$. Since the path is reduced, $s \neq t$, and since X is prefix, s and t are not prefix-comparable, see Figure 6.4. The proof for prefixes is similar.

Let $n \geq 2$. A path of length $2n$ from $s \otimes 1$ to $t \otimes 1$ is a sequence $(v_1 \otimes 1, 1 \otimes u_1, v_2 \otimes 1, \ldots, 1 \otimes u_n, v_{n+1} \otimes 1)$ with $s = v_1$ and $t = v_{n+1}$ such that the $2n$ words defined for $1 \leq i \leq n$ by

$$x_{2i-1} = u_i v_i, \quad x_{2i} = u_i v_{i+1}.$$

are in X. Moreover, since the path is reduced, one has $x_j \neq x_{j+1}$ for $1 \leq j < 2n$.

For $1 \leq i \leq n$, let p_i be the longest common prefix of v_i, v_{i+1}. Since $x_{2i-1} \neq x_{2i}$ and since the code X is prefix, the words v_i and v_{i+1} are not prefix-comparable.

Arguing by contradiction, assume that v_1 and v_{n+1} are prefix-comparable. By Lemma 6.3.2, we have $p_i = p_j$ for some indices i, j with $1 \leq i < j \leq n$.

Set $v_i = p_i v_i'$ and $v_{i+1} = p_i v_i''$. Since v_i, v_{i+1} are not prefix-comparable, the words v_i', v_i'' are nonempty. Since their longest common prefix is empty, their

initial letters are distinct. Thus u_ip_i is right-special. Similarly u_jp_j is right-special. Thus u_ip_i and u_jp_j are suffix-comparable. Since $p_i = p_j$, u_i and u_j are suffix-comparable.

But $1 \otimes u_i$ and $1 \otimes u_j$ are connected by the path $(1 \otimes u_i, v_{i+1} \otimes 1, \ldots, v_j \otimes 1, 1 \otimes u_j)$ of length $2(j-i) \leq 2(n-1)$. By the induction hypothesis, u_i and u_j are not suffix-comparable, a contradiction.

The proof that if $1 \otimes p$ and $1 \otimes q$ are connected by a path of length $2n$ in G, then p, q are not suffix-comparable is similar. ∎

Let X be a bifix code and let P be the set of proper prefixes of X. Consider the equivalence θ_X on P which is the transitive closure of the relation formed by the pairs $p, q \in P$ such that $ps, qs \in X$ for some $s \in A^+$. Such a pair corresponds, when $p, q \neq 1$, to a path $(1 \otimes p, s \otimes 1, 1 \otimes q)$ in the incidence graph of X. Thus a class of θ_X is either reduced to the empty word or it is the trace on $P \setminus 1$ of a connected component of the incidence graph of X.

Example 6.3.4 Consider the code X of Example 6.3.1 above. The three classes of θ_X are the class $\{1\}$ of the empty word, and the two suffix codes which are the traces of connected components of the incidence graph on the set of nonempty proper prefixes of X. These codes are $\{babaabaaba, babaaba, baba, baa, b\}$ and $\{babaabaab, babaabaa, babaab, babaa, bab, ba\}$. They are shown in Figure 6.5.

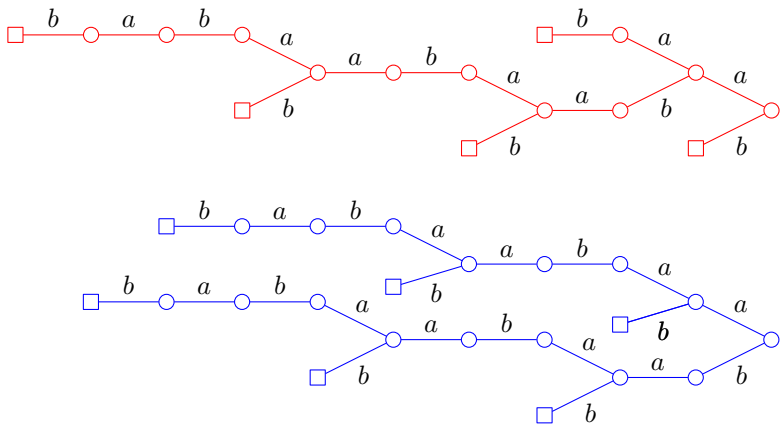

Figure 6.5: The two suffix codes which are classes of the equivalence θ_X.

The following property relates the equivalence θ_X with the right cosets of $H = \langle X \rangle$.

Proposition 6.3.5 *Let X be a bifix code, let P be the set of proper prefixes of X and let H be the subgroup generated by X. For any $p, q \in P$, $p \equiv q \bmod \theta_X$ implies $Hp = Hq$.*

Proof. Let $s \in A^+$ be such that $ps, qs \in X$. Then $p, q \in Hs^{-1}$ and thus $Hp = Hq$. ∎

Let $\mathcal{A} = (P, 1, 1)$ be the literal automaton of X^* (see Section 3.2). We show that the equivalence θ_X is compatible with the transitions of the automaton $\mathcal{A}$ in the following sense.

Lemma 6.3.6 *Let F be a Sturmian set. Let $X \subset F$ be a bifix code and let P be the set of proper prefixes of X. Let $p, q \in P$ and $a \in A$. If $p \equiv q \bmod \theta_X$ and if $p \cdot a, q \cdot a \neq \emptyset$ in the literal automaton of X^*, then $p \cdot a \equiv q \cdot a \bmod \theta_X$.*

Proof. Let G be the incidence graph of X.

Let $p, q \in P$ and $a \in A$ be such that $p \equiv q \bmod \theta_X$ and $p \cdot a, q \cdot a \neq \emptyset$. If $p = 1$, then $q = 1$ and the conclusion holds. Thus we may assume that $p \neq 1, q \neq 1$, and that $p \neq q$. Let $(1 \otimes u_0, v_1 \otimes 1, 1 \otimes u_1, \ldots, v_n \otimes 1, 1 \otimes u_n)$ be a path in G with $p = u_0$, $u_n = q$. The corresponding words in X are $u_0v_1, u_1v_1, u_2v_1, \ldots, u_nv_n$. We may assume that the words u_i are pairwise distinct, and that the v_i are pairwise distinct. Moreover, since $p \cdot a, q \cdot a \neq \emptyset$ there exist words v, w such that $pav, qaw \in X$.

The proof is in two steps. In the first step, we assume that v_1 and v_n both start with a. In the second step, we show that this condition is always fulfilled.

Assume that v_1 and v_n begin with a. There are two cases.

Case 1: Assume first that $pa, qa \in P$. Then $p \cdot a = pa$ and $q \cdot a = qa$. If all words v_i begin with a, then clearly the equivalence $pa \equiv qa \bmod \theta_X$ holds. Thus assume the contrary, and let $i > 1$ be minimal such that v_i begins with a letter distinct of a and let $i \leq j < n$ be maximal such that v_j begins with a letter distinct of a. Then both words u_{i-1} and u_j are right-special (since $u_{i-1}v_{i-1}$ starts with $u_{i-1}a$ and $u_{i-1}v_i$ starts with $u_{i-1}b$ for some letter $b \neq a$ and similarly for u_j). But since u_{i-1} and u_j are in the same trace on P' of a connected component of G, Lemma 6.3.3 implies that $u_{i-1} = u_j$, that is $i - 1 = j$. But this contradicts the inequality $i \leq j$.

Case 2: Suppose now that $pa \in X$. This implies that $v_1 = a$, since $pv_1 = u_0v_1$ is in X and begins with pa. If $v_n = aw$, then since v_1 and v_n , if they are distinct, are not prefix-comparable by Lemma 6.3.3, one $n = 1$ and $w = 1$. If $v_n \neq aw$, then $(v_1 \otimes 1, 1 \otimes u_1, \ldots, v_n \otimes 1, 1 \otimes u_n, aw \otimes 1)$ is a path from $v_1 \otimes 1$ to $aw \otimes 1$ (recall that $u_naw = qaw \in X$). Lemma 6.3.3 implies that $v_1 = a$ and aw, if they are distinct, are not prefix-comparable. Thus, one has again $w = 1$. In both cases, $qa \in X$ and therefore $p \cdot a = 1 = q \cdot a$.

We now show that the assumption that v_1 begins with a letter distinct of a leads to a contradiction (the case where v_n starts with a letter distinct from a is handled symmetrically). In this case since u_0v_1 is in X and $u_0av = pav \in X$, the word u_0 is right-special. Let i be the largest integer such that v_i begins with a letter distinct of a for $1 \leq i \leq n$. If $i < n$, then u_i is right-special. This contradicts Lemma 6.3.3(iii), since u_0 and u_i are distinct (because $i \geq 1$) elements of the trace on P' of a connected component of G. If $i = n$, then u_0

and u_n are right-special since $u_n v_n \in X$ and $u_n aw = qaw \in X$. We obtain again a contradiction since u_0 and u_n are distinct. ∎

6.4 Coset automaton

Let F be a Sturmian set and let $X \subset F$ be a bifix code. We introduce a new automaton denoted $\mathcal{B}_X$ or $\mathcal{B}$ for short, and called the *coset automaton* of X. Let R be the set of classes of θ_X with the class of 1 still denoted 1. The coset automaton of X is the automaton $\mathcal{B}_X = (R, 1, 1)$ with set of states R and transitions induced by the transitions of the literal automaton $\mathcal{A} = (P, 1, 1)$ of X^*. Formally, for $r, s \in R$ and $a \in A$, one has $r \cdot a = s$ in the automaton $\mathcal{B}$ if there exist p in the class r and q in the class s such that $p \cdot a = q$ in the automaton $\mathcal{A}$.

Observe first that the definition is consistent since, by Lemma 6.3.6, if $p \cdot a$ and $p' \cdot a$ are nonempty and p, p' are in the same class r, then $p \cdot a$ and $p' \cdot a$ are in the same class. Since the class $p \cdot a$ is uniquely defined, the automaton is indeed deterministic.

Observe next that if there is a path from p to p' in the automaton $\mathcal{A}$ labeled w , then there is a path from the class r of p to the class r' of p' labeled w in $\mathcal{B}_X$.

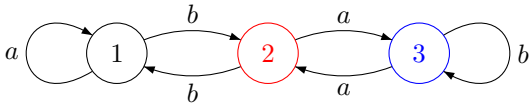

Figure 6.6: The automaton $\mathcal{B}_X$.

Example 6.4.1 For the code X of Example 6.3.1, the automaton $\mathcal{B}_X$ has three states. State 2 is the red class, that is the class containing b, and state 3 is the blue class containing ba. The code Z of this automaton is $a \cup b(ab^*a)^*b$. Observe that the word bb is in Z^* but it is not in X^*.

The following result shows that the coset automaton of X is the Stallings automaton of the subgroup generated by X.

Lemma 6.4.2 *Let F be a Sturmian set, and let $X \subset F$ be a bifix code. The coset automaton $\mathcal{B}_X$ is reversible and describes the subgroup generated by X. Moreover $X \subset Z$, where Z is the bifix code generating the submonoid recognized by $\mathcal{B}_X$.*

Proof. Set $\mathcal{B}_X = (R, 1, 1)$. Let $r, s \in R$ and $a \in A$ be such that $r \cdot a = s \cdot a$ is nonempty. Let $p, q \in P$ be elements of the classes r and s respectively, such that $p \cdot a, q \cdot a$ are nonempty. Then $pa, qa \in P \cup X$. To show that $\mathcal{B}_X$ is reversible, it is enough to show that $p \equiv q \bmod \theta_X$.

Suppose first that $pa \in X$. Then $r \cdot a = s \cdot a = 1$ and thus $qa \in X$ since 1 is isolated mod θ_X. Thus $p \equiv q \bmod \theta_X$.

Suppose next that $pa, qa \in P$. Then there is a path $(1 \otimes u_0, v_1 \otimes 1, \ldots, v_n \otimes 1, 1 \otimes u_n)$ in the incidence graph G of X, with $pa = u_0$ and $qa = u_n$. We may assume that the nodes of the path are pairwise distinct, except for a possible equality $u_0 = u_n$.

If all the words u_i end with a, then $p \equiv q \bmod \theta_X$.

Otherwise, let i be minimal such that u_i ends with a letter distinct of a and j, with $1 \leq i \leq j < n$ be maximal such that u_j ends with a letter distinct of a. Then v_i and v_{j+1} are left-special and they are distinct since $j + 1 > i$. This contradicts Lemma 6.3.3(iii) since v_i and v_{j+1} are distinct elements of the same trace on S'.

Thus the coset automaton is reversible.

Let Z be the bifix code generating the submonoid recognized by $\mathcal{B}_X$. To show the inclusion $X \subset Z$, consider a word $x \in X$. There is a path from 1 to 1 labeled x in $\mathcal{A}$, hence also in $\mathcal{B}_X$. Since the class of 1 modulo θ_X is reduced to 1, this path in $\mathcal{B}$ does not pass by 1 except at its ends. Thus x is in Z.

Let us finally show that the coset automaton describes the group $H = \langle X \rangle$. By Proposition 6.1.3, the subgroup described by $\mathcal{B}_X$ is equal to $\langle Z \rangle$. Set $K = \langle Z \rangle$. Since $X \subset Z$, we have $H \subset K$. To show the converse inclusion, let us show by induction on the length of $w \in A^*$ that, for $p, q \in Q$, there if there is a path from the class of p to the class of q in $\mathcal{B}_X$ with label w then $Hpw = Hq$. It is true for $w = 1$. Next, assume that it is true for w and consider wa with $a \in A$. Assume that there is a path $p \xrightarrow{w} q \xrightarrow{a} r$ in $\mathcal{B}_X$. By induction hypothesis, we have $Hpw = Hq$. Next, by definition of $\mathcal{B}_X$, there is an $s \equiv q \bmod \theta_X$ such that $sa \equiv r \mod \theta_X$. By Proposition 6.3.5, we have $Hs = Hq$ and $Hsa = Hr$. Thus $Hpwa = Hqa = Hsa = Hr$. This property shows that if $z \in Z$, then $Hz = Z$, that is $z \in H$. Thus $Z \subset H$ and finally $H = K$. ∎

6.5 Return words

Let F be a factorial set. For $u \in F$, define

$$\Gamma_F(u) = \{z \in F \mid uz \in A^+u \cap F\}\,, \qquad \Gamma'_F(u) = \{z \in F \mid zu \in uA^+ \cap F\}$$

and

$$R_F(u) = \Gamma_F(u) \setminus \Gamma_F(u)A^+\,, \qquad R'_F(u) = \Gamma'_F(u) \setminus A^+\Gamma'_F(u)\,.$$

When $F = F(x)$ for an infinite word x, the sets $\Gamma_F(u)$ and $R_F(u)$ are respectively the set of *right return words* to u and *first right return words* to u in x, and $\Gamma'_F(u)$ and $R'_F(u)$ are respectively the set of *left return words* to u and *first left return words* to u in x. The relation between $R_F(u)$ and $R'_F(u)$ is simply

$$uR_F(u) = R'_F(u)u\,. \tag{6.2}$$

Words in the set $uR_F(u) = R'_F(u)u$ are called *complete return words* in [32]. When there is no ambiguity, we will call the (first) right return words simply the (first) return words, omitting the 'right' specification.

Example 6.5.1 Let F be the Fibonacci set. The sets $R_F(u)$ and $R'_F(u)$ are given below for the first small words of F.

u	1	a	b	aa	ab	ba	aab	aba	baa	bab
$R_F(u)$	a	a	ab	baa	ab	ba	aab	ba	baa	$aabab$
	b	ba	aab	$babaa$	aab	aba	$abaab$	aba	$babaa$	$aabaabab$
$R'_F(u)$	a	a	ba	aab	ab	ba	aab	ab	baa	$babaa$
	b	ab	baa	$aabab$	aba	baa	$aabab$	aba	$baaba$	$babaabaa$

Vuillon has shown in [56] that x is a Sturmian word if and only if $R'_F(u)$ has exactly two elements for every factor u of x. Another proof of this result is given by Justin and Vuillon in [32].

In fact, they show in [32] the following theorem. Since this result is not exactly formulated in [32] as stated here, we show how it follows easily from their article.

Theorem 6.5.2 *Let F be a Sturmian set. For any word $u \in F$, the set $R_F(u)$ (and the set $R'_F(u)$) is a basis of the free group A°.*

By Equation (6.2), the sets $R_F(u)$ and $R'_F(u)$ are conjugate in the free group. Conjugacy by an element u is an automorphism of the free group. It follows that $R_F(u)$ is a basis if and only if $R'_F(u)$ is a basis. Thus, it suffices to prove the claim for $R'_F(u)$. We quote the following result of [32, Theorem 4.4, Corollaries 4.1 and 4.5], with the notations of Section 2.3.

Proposition 6.5.3 *Let s be a standard strict episturmian word over A, let $\Delta = a_0a_1\cdots$ be its directive word, and let (u_n) be its sequence of palindrome prefixes.*

(i) *The first left return words to u_n are the words $\psi_{a_0\cdots a_{n-1}}(a)$ for $a \in A$.*

(ii) *For each factor u of s, there exist a word z and an integer n such that the first left return words to u are the words zyz^{-1}, where y ranges over the first left return words to u_n.*

Proof of Theorem 6.5.2. We may assume that $F = F(s)$ for some standard and strict episturmian word s. By Proposition 6.5.3(i), the set of first left return words to u_n is the image of the alphabet by the endomorphism $\psi_{a_0\cdots a_{n-1}}$. It is easily seen that these endomorphisms define automorphisms of the free group. We deduce that the set of first left return words to u_n is a basis of the free group on A. By Proposition 6.5.3(ii), the set of first left return words to u is a basis, too. This ends the proof. ∎

6.6 Proof of the main result

Some preliminary results are needed for the proof of Theorem 6.2.1.

Proposition 6.6.1 *Let F be a Sturmian set and let $X \subset F$ be a finite F-maximal bifix code. Then $\langle X\rangle \cap F = X^* \cap F$.*

*Proof.*We have $X^* \cap F \subset \langle X \rangle \cap F$. To show the converse inclusion, consider the bifix code Z generating the submonoid recognized by the coset automaton $\mathcal{B}_X$ associated to X.

Let us show that $Z \cap F = X$. By Lemma 6.4.2, we have $X \subset Z$ and thus $X \subset Z \cap F$. Since X is an F-maximal bifix code, this implies that $X = Z \cap F$.

Since any reversible automaton is minimal and since the automaton $\mathcal{B}_X$ is reversible by Lemma 6.4.2, it is equal to the minimal automaton of Z^*. Let K be the subgroup generated by Z. By Proposition 6.1.1, we have $K \cap A^* = Z^*$.

This shows that

$$\langle X \rangle \cap F \subset K \cap F = K \cap A^* \cap F = Z^* \cap F = X^* \cap F\,.$$

The first inclusion holds because $X \subset Z$ implies $\langle X \rangle \subset K$. The last equality follows from the fact that if $z_1 \cdots z_n \in F$ with $z_1, \ldots, z_n \in Z$, then each z_i is in F hence in $Z \cap F = X$. Thus $\langle X \rangle \cap F \subset X^* \cap F$, which was to be proved. ∎

Note the following consequence of Proposition 6.6.1.

Corollary 6.6.2 *Let F be a Sturmian set and let $X \subset F$ be a finite F-maximal bifix code. Each right coset of the subgroup $\langle X \rangle$ generated by X contains at most one right-special proper prefix of X.*

Proof. Set $H = \langle X \rangle$. Let Q be the set of those proper prefixes of the words of X which are right-special.

Let us show that if $p, q \in Q$ belong to the same right coset, then $p = q$. We may assume that $p = uq$. Since $Hp = Hq$, one has $Huq = Hq$. Consequently, $Hu = H$ and thus $u \in H$. By Proposition 6.6.1, since $u \in F$, this implies that $u \in X^*$ and thus $u = 1$ since p is a proper prefix of X. ∎

Proof of Theorem 6.2.1.

Assume first that X is an F-maximal bifix code of F-degree d. Let P be the set of proper prefixes of X. Let Q be the set of words in P which are right-special. Let H be the subgroup generated by X.

By Lemma 5.2.3 there is a right-special word u such that $\pi_X(u) = d$. The d suffixes of u which are in P are the elements of Q. By Theorem 4.2.8, the word u is not an internal factor of X.

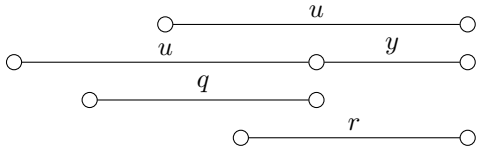

Figure 6.7: A word $y \in R_F(u)$.

Let

$$V = \{v \in A^\circ \mid Qv \subset HQ\}\,.$$

Any $v \in V$ defines a permutation of Q. Indeed, suppose that for $p, q \in Q$, one has $pv, qv \in Hr$ for some $r \in Q$. Then rv^{-1} is in $Hp \cap Hq$. This forces $Hp = Hq$ and thus $p = q$ by Corollary 6.6.2.

The set V is a subgroup of A°. Indeed, $1 \in V$. Next, let $v \in V$. Then for any $q \in Q$, since v defines a permutation of Q, there is a $p \in Q$ such that $pv \in Hq$. Then $qv^{-1} \in Hp$. This shows that $v^{-1} \in V$. Next, if $v, w \in V$, then $Qvw \subset HQw \subset HQ$ and thus $vw \in V$.

We show that the set $R_F(u)$ is contained in V. Indeed, let $q \in Q$ and $y \in R_F(u)$. Since q is a suffix of u, qy is a suffix of uy, and since uy is in F (by definition of $\Gamma_F(u)$), also qy is in F. The fact that X is F-maximal implies that there is a word $r \in P$ such that $qy \in X^*r$. We verify that the word r is a suffix of u. Since $y \in R_F(u)$, there is a word y' such that $uy = y'u$. Consequently, r is a suffix of $y'u$, and in fact the word r is a suffix of u. Indeed, one has $|r| \leq |u|$ since otherwise u is in $I(X)$ and this is not the case. Thus we have $r \in Q$ (see Figure 6.7). Since $X^* \subset H$ and $r \in Q$, we have $qy \in HQ$. Thus $y \in V$.

By Theorem 6.5.2, the group generated by $R_F(u)$ is A°. Since $R_F(u) \subset V$, and since V is a subgroup of A°, we have $V = A^\circ$. Thus $Qw \subset HQ$ for any $w \in A^\circ$. Since $1 \in Q$, we have in particular $w \in HQ$. Thus $A^\circ = HQ$. Since $\mathrm{Card}(Q) = d$, and since the right cosets Hq for $q \in Q$ are pairwise disjoint, this shows that H is a subgroup of index d. By Theorem 5.2.1 and in view of Schreier's Formula, X is a basis of H.

Assume conversely that the bifix code $X \subset F$ is a basis of the group $H = \langle X \rangle$ and that $\langle X \rangle$ has index d. Since X is a basis, by Schreier's Formula, we have $\mathrm{Card}(X) = (k-1)d+1$, where $k = \mathrm{Card}(A)$. The case $k = 1$ is straightforward; thus we assume $k \geq 2$. By Theorem 4.4.3, there is a finite F-maximal bifix code Y containing X. Let e be the F-degree of Y. By the first part of the proof, Y is a basis of a subgroup K of index e of A°. In particular, it has $(k-1)e+1$ elements. Since $X \subset Y$, we have $(k-1)d+1 \leq (k-1)e+1$ and thus $d \leq e$. On the other hand, since H is included in K, d is a multiple of e and thus $e \leq d$. We conclude that $d = e$ and thus that $X = Y$. ∎

Example 6.6.3 Let F be the Fibonacci set. Let $X \subset F$ be the bifix code shown on Figure 6.8. The right-special proper prefixes of the words of X are the four suffixes of aba and are indicated in black on the figure. The states of the coset

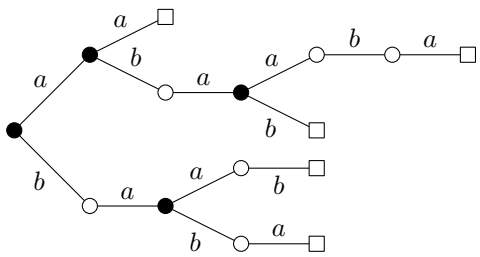

Figure 6.8: An F-maximal bifix code of F-degree 4.

automaton are the sets $\{1\}$, $\{a, bab, abaab\}$, $\{aba, b, baa\}$ and $\{ba, ab, abaa\}$. The code X has F-degree 4. Each state is represented by its right-special factor in Figure 6.9.

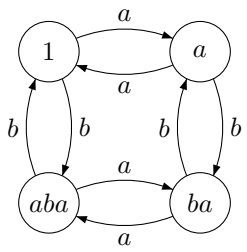

Figure 6.9: The associated coset automaton.

We end this section with a combinatorial consequence of Theorem 6.2.1.

Proposition 6.6.4 *Let F be a Sturmian set on an alphabet with k letters and let $X \subset F$ be a finite F-maximal bifix code of F-degree d. Let P (resp. S) be the set of proper prefixes (resp. suffixes) of X. Then*

$$\sum_{x \in X} |x| = \mathrm{Card}(P) + \mathrm{Card}(S) + (k-2)d.$$

We will use the following proposition, of independent interest.

Proposition 6.6.5 *Let F be a Sturmian set and let $X \subset F$ be a finite F-maximal bifix code of F-degree d. The coset automaton $\mathcal{B}_X$ is a group automaton with d states. Each state of $\mathcal{B}_X$ other than 1 is an F-maximal suffix code.*

Proof. Let $\mathcal{A} = (P, 1, 1)$ be the literal automaton of X^* and let $\mathcal{B} = (R, 1, 1)$ be the coset automaton of X^*. By Lemma 6.4.2, the automaton $\mathcal{B}$ is reversible and describes the subgroup H generated by X.

By Theorem 6.2.1, the subgroup H has index d in A°. Since $\mathcal{B}$ is reversible, it is minimal. Thus Proposition 6.1.6 applies, showing that $\mathcal{B}$ is a group automaton. Its number of states is d since a group automaton with d states describes a subgroup of index d.

Finally, consider $r \in R$ and let $X_r = \{p \in P \mid 1 \cdot p = r\}$. Let us show that any $w \in F$ is suffix-comparable with an element of X_r. We may assume that w is longer than any word of X. Since $\mathcal{B}_X$ is a group automaton, there is an $u \in R$ such that $u \cdot w = r$. Since w is longer than any word of X, the path from u to r labeled w passes through state 1. Thus w has a parse (s, x, p) such that $1 \cdot p = r$ and thus w has a suffix in X_r. This shows that X_r is an F-maximal suffix code. ∎

Note that the fact that the set P of nonempty proper prefixes of X is a disjoint union of $d-1$ F-maximal suffix codes is also a consequence of Theorem 4.3.7.

Proof of Proposition 6.6.4. Let H be the subgroup generated by X. By Theorem 6.2.1, the set X is a basis of H and the index of H is equal to $d = d_F(X)$. Let G be the incidence graph of X. Let E be the set of edges of G. One has

$$\mathrm{Card}(E) = \sum_{x\in X}(|x| - 1) = \sum_{x\in X}|x| - \mathrm{Card}(X) = \sum_{x\in X}|x| - (k-1)d - 1\,.$$

By Proposition 6.6.5, the classes of θ_X are the set $\{1\}$ and $d-1$ F-maximal suffix codes denoted P_i, for $i = 1, \ldots, d-1$. Each of the latter is the trace on $P \setminus 1$ of a connected component C_i of G. Let G_i be the subgraph of G induced by its connected component C_i. By Lemma 6.3.3, G_i is a tree.

Similarly, let S_i be the trace on $S \setminus 1$ of the connected component C_i. Let E_i be the set of edges of G_i. Since G_i is a tree, we have $\mathrm{Card}(E_i) = \mathrm{Card}(P_i) + \mathrm{Card}(S_i) - 1$ for $i = 1, \ldots, d-1$. Finally

$$\begin{aligned}\mathrm{Card}(E) &= \sum_{i=1}^{d-1}\mathrm{Card}(E_i) = \sum_{i=1}^{d-1}(\mathrm{Card}(P_i) + \mathrm{Card}(S_i) - 1)\\ &= \mathrm{Card}(P\setminus 1) + \mathrm{Card}(S\setminus 1) - (d-1)\,,\end{aligned}$$

whence the result. ∎

7 Syntactic groups

In this section, we introduce the notion of F-group of a bifix code $X \subset F$ of finite F-degree. It is a permutation group of degree $d_F(X)$. We investigate the relation between this group and the notion of group of a maximal bifix code (Theorem 7.2.5). We use Theorem 6.2.1 to prove a new result on the syntactic groups of bifix codes: any transitive permutation group G of degree d and with k generators is a syntactic group of a bifix code with $(k-1)d + 1$ elements (Theorem 7.2.3).

7.1 Preliminaries

We first recall the basic terminology on groups in monoids (see [6] for a more detailed exposition). We are mainly concerned with monoids of maps from a set into itself. The maps considered in this section are partial maps.

Let M be a monoid. A *group in* M is a subsemigroup of M which is isomorphic to a group. Note that the neutral element of a group contained in M needs not be equal to the neutral element of M.

A group in M is *maximal* if it not included in another group in M.

Proposition 7.1.1 *Let G be a group in a monoid M of partial maps from a set Q into itself. All elements of G have the same image I. The restriction of the elements of G to I is a faithful representation of G as a permutation group on I.*

Proof. Two elements $g, h \in G$ have the same image. Indeed, let k be the inverse of g in G. Then $h = hkg$ and thus the image of h is contained in the image of g. The converse inclusion is shown analogously. Then G is a permutation group on the common image I of its elements. Indeed, let e be the neutral element of G. Then for any $p \in I$, let $q \in Q$ be such that $qe = p$. Then $pe = qe^2 = qe = p$. This shows that e is the identity on I. Next, for any $g \in G$ the inverse k of g is such that $gk = kg = e$. Thus g is a permutation on I.

Let $g, g' \in G$ be such that they have the same restriction to I. Then for each $p \in Q$, $p(eg) = (pe)g = (pe)g' = p(eg')$ since $pe \in I$. Since $eg = g$ and $eg' = g'$, we obtain $g = g'$. This shows that the representation of G by permutations on I is faithful. ∎

Let G be a group in a monoid of maps from Q into itself as above. The *canonical* representation of G by permutations is the restriction of the maps in G to their common image.

A *syntactic group* of a prefix code X is the canonical representation by permutations of a maximal group in the monoid of transitions of the minimal automaton $\mathcal{A}(X^*)$ of X^*.

Let X be a prefix code and let $\mathcal{A} = \mathcal{A}(X^*)$. A syntactic group G of X is called *special* if $\varphi_{\mathcal{A}}^{-1}(G)$ is a cyclic submonoid of A^*. In particular a special syntactic group is cyclic.

The *degree* of a permutation group G on a set R is the cardinality of R. Recall that the group G is *transitive* if for any $r, s \in R$ there is some $g \in G$ such that $rg = s$.

A permutation group G on a set R and a permutation group H on a set S are *equivalent* if there exists a bijection $\beta : R \to S$ and an isomorphism $\sigma : G \to H$ such that, for all $g \in G$ and $r \in R$, one has

$$\beta(rg) = \beta(r)\sigma(g)\,,$$

in other terms, if the diagram of Figure 7.1 is commutative for all $g \in G$.

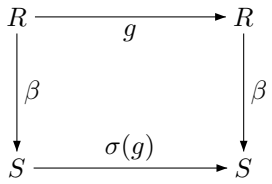

Figure 7.1: Equivalent permutation groups.

Let us recall the notation concerning Green relations in a monoid M (see [6]). We denote by $\mathcal{R}$ the equivalence in M defined by $m\mathcal{R}n$ if m, n generate the same right ideal, i.e. if $mM = nM$. We denote by $R(m)$ the $\mathcal{R}$-class of m.

Symmetrically, we denote by $\mathcal{L}$ the equivalence defined by $m\mathcal{L}n$ if m, n generate the same left ideal, i.e. if $Mm = Mn$. We denote by $L(m)$ the $\mathcal{L}$-class of m.

It is well known that the equivalences $\mathcal{L}$ and $\mathcal{R}$ commute. We denote by $\mathcal{D}$ the equivalence $\mathcal{LR} = \mathcal{RL}$. Finally, we denote by $\mathcal{H}$ the equivalence $\mathcal{L} \cap \mathcal{R}$.

A $\mathcal{D}$-class D is *regular* if it contains an idempotent. In this case, there is at least an idempotent in each $\mathcal{L}$-class of D and in each $\mathcal{R}$-class of D. The following statement is known as *Clifford and Miller's Lemma*. For $m, n \in M$, one has $mn \in R(m) \cap L(n)$ if and only if $R(n) \cap L(m)$ contains an idempotent.

Assume that M is a monoid of maps from a finite set Q into itself.

If $m, n \in M$ are $\mathcal{L}$-equivalent, then they have the same image. If they are $\mathcal{R}$-equivalent, then they have the same nuclear equivalence (the nuclear equivalence of a partial map m from Q into itself is the partial equivalence, for which $p, q \in Q$ are equivalent if m is defined on p and q and $pm = qm$).

If $m, n \in M$ are $\mathcal{H}$-equivalent, they have the same image and the same nuclear equivalence. The converse is not true but it holds in the following important particular case.

Proposition 7.1.2 *Let M is a monoid of maps from a finite set Q into itself. Let $e \in M$ be an idempotent. An element m of M is in the $\mathcal{H}$-class of e if and only if it has the same nuclear equivalence and the same image as e.*

Proof. If m and e are $\mathcal{H}$-equivalent, they have the same nuclear equivalence ρ and the same image I. Conversely, we have $me = m$ since e is the identity on its image I. For any $p \in Q$, $pe^2 = pe$ implies that p and pe are in the same class of ρ. This implies that $pem = pm$. Thus $em = m$.

Finally, the restriction of m to I is a permutation. Indeed, $pm = qm$ for $p, q \in I$ implies $pe = qe$ which forces $p = q$. Let $k > 0$ be such that the restriction of m^k to I is the identity. Then m^k and e are two idempotents with the same nuclear equivalence and the same image. This implies that they are equal. Thus m and e are in the same $\mathcal{H}$-class. ∎

Let F be a recurrent set and let $X \subset F$ be a bifix code of finite F-degree d. Let $\mathcal{A} = (Q, 1, 1)$ be a simple automaton recognizing X^*. We set $\varphi = \varphi_{\mathcal{A}}$ and we denote by M the transition monoid $\varphi(A^*)$ of $\mathcal{A}$.

For a word w, we denote by $\mathrm{Im}(w)$ the *image* of w with respect to $\mathcal{A}$, that is the set $\mathrm{Im}(w) = \{p \cdot w \mid p \in Q\}$. The *rank* of w (with respect to the automaton $\mathcal{A}$) is the number $\mathrm{rank}(w) = \mathrm{Card}(\mathrm{Im}(w))$. Then $\mathrm{Im}(w)$ is also the image of the map $\varphi(w)$ (recall that the action of M is on the right of the elements of Q), and the rank of w is also the rank of $\varphi(w)$. Clearly $\mathrm{rank}(uwv) \leq \mathrm{rank}(w)$ for all u, w, v.

Proposition 7.1.3 *The set of elements of $\varphi(F)$ of rank d is included in a regular $\mathcal{D}$-class of M.*

We use the following lemmas.

Lemma 7.1.4 *A word $w \in F$ which has d parses with respect to X has rank d with respect to $\mathcal{A}$. Moreover, $\mathrm{Im}(w)$ is the set of states $1 \cdot p$ for all p such that*

there is a parse (s, x, p) of w. For all $q \in \mathrm{Im}(w)$, there is a unique proper prefix p of P which is a suffix of w, and such that $q = 1 \cdot p$.

Proof. Consider first two states $q, r \in Q$ and suppose that $q \cdot w = r$. Since $\mathcal{A}$ is simple, it is trim. Consequently there exist two words u, v such that $1 \cdot u = q$ and $r \cdot v = 1$. It follows that $uwv \in X^*$. Since w has d parses, by Theorem 4.2.8 it is not an internal factor of a word in X. Thus there is a parse (s, x, p) of w such that $us, pv \in X^*$. Then $r = 1 \cdot p$. The relation $r \to (s, x, p)$ is a function. Indeed, let us show that if (s, x, p) and (s', x', p') are two distinct parses of w, then $1 \cdot p \neq 1 \cdot p'$. Assume the contrary. Then we have $pv, p'v \in X^*$ for the same word v. Since p, p' are suffixes of w, they are suffix-comparable and thus $p = p'$ since X is bifix. This is impossible if the parses are distinct. Of course, the function $r \mapsto (s, x, p)$ is injective since $\mathcal{A}$ is deterministic.

Conversely, let (s, x, p) be a parse of w. Since X is an F-maximal bifix code, there exist by Theorem 4.2.2 words u, v such that $us, pv \in X^*$. Thus we have $1 \cdot us = 1 \cdot x = 1 \cdot pv = 1$. Consequently $(1 \cdot u) \cdot w = 1 \cdot usxp = 1 \cdot xp = 1 \cdot p$. This shows that $1 \cdot p \in \mathrm{Im}(w)$.

∎

Lemma 7.1.5 *Let $u \in F$ be a word. If $\mathrm{rank}(u) = d$, then $\mathrm{rank}(uv) = d$ for all v such that $uv \in F$.*

Proof. Since X is F-thin, there exists $w \in F$ which is not a factor of a word in X. This word w has d parses. Assume $uv \in F$. Since F is recurrent, there exists a word t such that $uvtw \in F$. Then $uvtw$ also has d parses. By Lemma 7.1.4, this implies that the rank of $uvtw$ is d. Since $d = \mathrm{rank}(uvtw) \leq \mathrm{rank}(uv) \leq \mathrm{rank}(u) = d$, one has $\mathrm{rank}(uv) = d$. ∎

Proof of Proposition 7.1.3. Let $u, v \in F$ be two words of rank d. Set $m = \varphi(u)$ and $n = \varphi(v)$. Let w be such that $uwv \in F$. We show first that $m\mathcal{R}\varphi(uwv)$ and $n\mathcal{L}\varphi(uwv)$.

For this, let t be such that $uwvtu \in F$. Set $z = wvtu$. By Lemma 7.1.5, the rank of uz is d. Since $\mathrm{Im}(uz) \subset \mathrm{Im}(z) \subset \mathrm{Im}(u)$, this implies that the images are equal. Consequently, the restriction of $\varphi(z)$ to $\mathrm{Im}(u)$ is a permutation. Since $\mathrm{Im}(u)$ is finite, there is an integer $\ell \geq 1$ such that $\varphi(z)^\ell$ is the identity on $\mathrm{Im}(u)$. Set $e = \varphi(z)^\ell$ and $s = tuz^{\ell-1}$. Then, since e is the identity on $\mathrm{Im}(u)$, one has $m = me$. Thus $m = \varphi(uwv)\varphi(s)$, and since $\varphi(uwv) = m\varphi(wv)$, it follows that m and $\varphi(uwv)$ are $\mathcal{R}$-equivalent.

Similarly n and $\varphi(uwv)$ are $\mathcal{L}$-equivalent. Indeed, set $z' = tuwv$. Then $\mathrm{Im}(vz') \subset \mathrm{Im}(z') \subset \mathrm{Im}(v)$. Since vz' is a factor of z^2 and z has rank d, it follows that $d = \mathrm{rank}(z^2) \leq \mathrm{rank}(vz') \leq \mathrm{rank}(v) = d$. Therefore, vz' has rank d and consequently the images $\mathrm{Im}(vz')$, $\mathrm{Im}(z')$ and $\mathrm{Im}(v)$ are equal. There is an integer $\ell' \geq 1$ such that $\varphi(z')^{\ell'}$ is the identity on $\mathrm{Im}(v)$. Set $e' = \varphi(z')^{\ell'}$. Then $n = ne' = n\varphi(z')^{\ell'-1}\varphi(tuwv) = nq\varphi(uwv)$, with $q = \varphi(z')^{\ell'-1}\varphi(t)$. Since $\varphi(uwv) = \varphi(uw)n$, one has $n\mathcal{L}\varphi(uwv)$. Thus m, n are $\mathcal{D}$-equivalent, and $\varphi(uwv) \in R(m) \cap L(n)$.

Set $p = \varphi(wv)$. Then $p = \varphi(w)n$ and, with the previous notation, $n = ne' = nq\varphi(u)p$, so $L(n) = L(p)$. Thus $mp = \varphi(uwv) \in R(m) \cap L(p)$, and by Clifford and Miller's Lemma, $R(p) \cap L(m)$ contains an idempotent. Thus the $\mathcal{D}$-class of m, p and n is regular. ∎

7.2 Group of a bifix code

Let M be a monoid. The $\mathcal{H}$-class of an idempotent e is denoted $H(e)$. It is the maximal group contained in M and containing e.

All groups $H(e)$ for e idempotent in a regular $\mathcal{D}$-class D are isomorphic. The *structure group* (or Schützenberger group) of D is any one of them. When M is a monoid of maps from a set Q into itself, the canonical representations of the groups $H(e)$ are equivalent permutation groups. See [6, Proposition 9.1.9]. We then also consider the structure group as a permutation group.

Let F be a recurrent set and let $X \subset F$ be a bifix code of finite F-degree d. Let $\mathcal{A} = (Q, 1, 1)$ be a simple automaton recognizing X^*. Set $\varphi = \varphi_A$. The structure group of the $\mathcal{D}$-class of elements of rank d of $\varphi(F)$ is a permutation group of degree d. By Proposition 9.5.1 in [6], this group does not depend on the choice of the simple automaton $\mathcal{A}$ recognizing X^*. It is called the *F-group* of the code X and denoted $G_F(X)$.

When $F = A^*$, the group $G_F(X)$ is the group $G(X)$ of the code X defined in [6]. Indeed, in this case, the $\mathcal{D}$-class of elements of rank d coincides with the minimal ideal of the monoid $\varphi(A^*)$.

The following example shows that the F-group of an F-maximal bifix code is not always transitive.

Example 7.2.1 Let $X = \{ab, ba\}$ and let $F = F((ab)^*)$. Then X is an F-maximal bifix code of F-degree 2. It can be verified easily that the syntactic monoid of X^* contains only trivial subgroups (see also Exercises 7.1.1, 7.2.1 in [6]). Thus $G_F(X)$ is reduced to the identity.

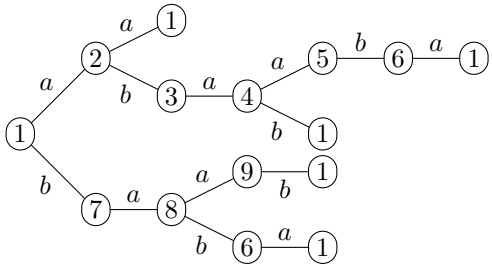

Figure 7.2: An F-maximal bifix code of F-degree 4.

Example 7.2.2 We consider again the code of Example 6.6.3. The minimal automaton of X^* is represented on Figure 7.2.

	$1,2,4,8$	$1,2,5,9$	$1,3,6,7$
$1\|2,6\|3\|7$		a^2	*
$1\|2\|4,9\|5,8$	* ba	*	b
$1\|2,6\|3,8\|4,7$	* aba		* ab

Figure 7.3: The $\mathcal{D}$-class of rank 4.

We have represented on Figure 7.3 the $\mathcal{D}$-class of elements of rank 4 meeting $\varphi(F)$. It is composed of three $\mathcal{L}$-classes and three $\mathcal{R}$-classes. Each $\mathcal{L}$-class is represented by a column and each $\mathcal{R}$-class by a row. On top of each column, we have indicated the common image of the its elements. On the left of each row, we have indicated the common nuclear equivalence of its elements (recall that two elements are equivalent for the nuclear equivalence if and only if they have the same image). The $\mathcal{H}$-classes containing an idempotent are indicated by a star. Each $\mathcal{H}$-class has four elements, and five of them are groups (this happens when the image is a system of representatives of the nuclear equivalence). For instance, the five classes in the nuclear equivalence of $\varphi(ba)$ are $\{1\}$, $\{2\}$, $\{4,9\}$, $\{5\}$ and $\{8\}$, and the $\mathcal{H}$-class of (the image of) ba is composed of the following elements:

Word	Permutation
ba	$(18)(24)$
$baaba$	$(12)(48)$
$baba$	(1)
$babaaba$	$(14)(28)$

The structure group of this $\mathcal{D}$-class is the Abelian group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. It is the F-group of the code.

The aim of this section is to prove the following theorem.

Theorem 7.2.3 *Any transitive permutation group of degree d which can be generated by k elements is a syntactic group of a bifix code with $(k-1)d+1$ elements.*

Theorem 7.2.3 was known before in particular cases. In [44] it is shown that any transitive permutation group is a syntactic group of a finite bifix code. The bound $d+1$ on the cardinality of the bifix code is proved for the case of a group generated by a d-cycle and another permutation. In [51], it is proved that for an Abelian group of rank 2 and order d there exists a bifix code X such that $\mathrm{Card}(X) - 1 = d$. The proof is based on the fact that the Cayley graph of an Abelian group contains a Hamiltonian cycle.

Let us call *minimal rank* of a group G the minimal cardinality of a generating set for G. Theorem 7.2.3 is related to the following conjecture [44].

Let X be a finite bifix code and let G be a transitive permutation group of degree d and minimal rank k. If G is a syntactic group of X, then $\mathrm{Card}(X) \geq (k-1)d + 1$.

Theorem 7.2.3 shows that the lower bound is sharp.

The following result, which is from [45], shows that the conjecture holds for $k = 2$.

Theorem 7.2.4 *Let G be a permutation group of degree d. If G is a nonspecial syntactic group of a finite prefix code X, then* $\mathrm{Card}(X) \geq d + 1$.

Theorem 7.2.4 is clearly not true for special syntactic groups since $\mathbb{Z}/n\mathbb{Z}$ is a syntactic group of $X = a^n$ for any $n \geq 1$.

Theorem 7.2.3 is a consequence of the following theorem which can be viewed as a complement to Theorem 4.2.11. The proof itself makes use of Theorem 6.2.1.

Theorem 7.2.5 *Let $Z \subset A^*$ be a group code of degree d. Let F be a Sturmian set. The set $X = Z \cap F$ is an F-maximal bifix code of F-degree d and $G_F(X) = G(Z)$.*

Proof. The fact that X is an F-maximal bifix code of F-degree d results from Corollary 6.2.3.

Let us show that $G_F(X) = G(Z)$. Let $\mathcal{B} = (R, 1, 1)$ be the minimal automaton of Z^*. Set $\psi = \varphi_{\mathcal{B}}$ and $G = \psi(A^*)$. Thus G is a permutation group equivalent to $G(Z)$.

Let $\mathcal{A} = (Q, 1, 1)$ be the minimal automaton of X^*. Set $\varphi = \varphi_{\mathcal{A}}$. Denote by $\mathrm{Im}(w)$ the image of $\varphi(w)$ with respect to $\mathcal{A}$. Thus $\mathrm{Im}(w) = \{t \in Q \mid s \cdot w = t \text{ for some } s \in Q\}$.

Let $u \in F$ be a word with d parses with respect to X. Let $I = \mathrm{Im}(u)$. By Lemma 7.1.4, the word u has rank d and thus $\mathrm{Card}(I) = d$.

Let $Y = R_F(u)$ be the set of first return words to u. By Theorem 6.5.2, the set Y is a basis of the free group A°. For any $y \in Y$, the restriction of $\varphi(y)$ to I is a permutation of I. Indeed, $uy \in A^+u$ implies $\mathrm{Im}(uy) \subset I$. Since $uy \in F$, the set $\mathrm{Im}(uy)$ has d elements by Lemma 7.1.5. Thus $\mathrm{Im}(uy) = I$. Since $\mathrm{Im}(u) = I$, this proves the claim.

Let e be an idempotent in $\varphi(Y^+)$. The restriction of e to I is the identity. Any long enough element of $\varphi^{-1}(e) \cap Y^*$ has u as a suffix. Thus the image of e is I. Moreover, since $\varphi(u)e = \varphi(u)$ and $e \in \varphi(A^*u)$, e and $\varphi(u)$ belong to the same $\mathcal{L}$-class and thus to the $\mathcal{D}$-class. Thus e belongs to the $\mathcal{D}$-class of $\varphi(A^*)$ which contains the elements of rank d in $\varphi(F)$.

Let G' be the maximal group contained in $\varphi(A^*)$ which contains e. It is a permutation group on I which is equivalent to $G_F(X)$.

For $y \in Y^*$, let $\chi(y)$ be the restriction of $\varphi(y)$ to the set I.

For any $y \in Y^*$, $e\varphi(y)e$ has the same nuclear equivalence and the same image as e. By Proposition 7.1.2 it implies that they are in the same $\mathcal{H}$-class. Thus $e\varphi(y)e$ is in G'.

Since $e\varphi(y)e$ and $\varphi(y)$ have the same restriction to I and since $e\varphi(y)e$ belongs to the $\mathcal{H}$-class of e, χ is a morphism from Y^* into the permutation group G'. Since Y generates A°, this morphism is surjective. Indeed, if $\varphi(w) \in G'$, let $y_1, \ldots, y_n \in Y$ be such that $w = y_1^{\varepsilon_1} \cdots y_n^{\varepsilon_n}$ with $\varepsilon_i = \pm 1$. Then $\chi(w) = \chi(y_1)^{\varepsilon_1} \cdots \chi(y_n)^{\varepsilon_n}$. Since G' is a finite group $\chi(y)^{-1} \in \chi(Y^*)$. Thus $\chi(w) \in \chi(Y^*)$.

Let us show that G and G' are equivalent as permutation groups.

For this, let us define a bijection $\beta : I \to R$ as follows. Let P be the set of proper prefixes of the words of X and let S be the set of elements of P which are suffixes of u. For $i \in I$, there is a unique $q \in S$ such that $i = 1 \cdot q$ by Lemma 7.1.4. Set $\beta(i) = 1\psi(q)$. We show that β is injective. Let $q, t \in S$ be such that $1\psi(q) = 1\psi(t)$. Assume that $|q| \leq |t|$. Since q, t are suffix-comparable, we have $t = vq$. Since $1\psi(t) = 1\psi(v)\psi(q) = 1\psi(q)$ and since $\psi(q)$ is a permutation, we have $1\psi(v) = 1$ and thus $v \in Z^*$. Since v is in F and since $Z^* \cap F \subset X^*$, this implies $v \in X^*$ and thus $v = 1$. This shows that $q = t$ and thus that β is injective. Since $\mathrm{Card}(R) = \mathrm{Card}(I) = d$, we have shown that β is a bijection.

Let us verify that for any $i, j \in I$ and $y \in Y^*$, we have

$$i\varphi(y) = j \quad \Longleftrightarrow \quad \beta(i)\psi(y) = \beta(j). \tag{7.1}$$

Let us first prove (7.1) for $y \in Y$. For this, let $q, t \in S$ be such that $i = 1 \cdot q$, $j = 1 \cdot t$. The states q, t exist by Lemma 7.1.4. Then

$$i\varphi(y) = j \iff 1\varphi(qy) = 1\varphi(t) \iff qy \in X^*t.$$

The last equivalence holds because $1 \cdot qy = 1 \cdot v$ for the word $v \in P$ such that $qy \in X^*v$. But since $uy \in A^*u$, v is a suffix of u and thus $v \in S$. This forces $t = v$.

Since $qy \in F$, we have

$$qy \in X^*t \iff qy \in Z^*t$$

and thus, we obtain

$$i\varphi(y) = j \iff qy \in Z^*t \iff \beta(i)\psi(y) = \beta(j).$$

This proves (7.1) for $y \in Y$. Next, let us show that if $y, z \in Y^*$ satisfy (7.1) for all $i, j \in I$, the same is true for yz. Assume first that for $i, j \in I$, one has $i\varphi(yz) = j$. Since the restrictions of $\varphi(y), \varphi(z)$ to I are permutations, there is a unique $k \in I$ such that $i\varphi(y) = k$ and $k\varphi(z) = j$. Then, since y, z satisfy (7.1), we have $\beta(i)\psi(y) = \beta(k)$ and $\beta(k)\psi(z) = \beta(j)$. Thus $\beta(i)\psi(yz) = \beta(j)$. Conversely, assume that $\beta(i)\psi(yz) = \beta(j)$. Since β is a bijection from I onto R, there is a unique $k \in I$ such that $\beta(k) = \beta(i)\psi(y)$. Then $\beta(k)\psi(z) = \beta(j)$. By (7.1), we have $i\varphi(y) = k$ and $k\varphi(z) = j$ whence $i\varphi(yz) = j$. This proves that yz satisfies (7.1).

Equation (7.1) shows that we may define a morphism α from G' to G by $\alpha(g) = \psi(y)$ for $y \in Y^*$ such that $\chi(y) = g$. This map is injective. Indeed, if $\alpha(g) = \alpha(g')$, let $y, y' \in Y^*$ be such that $\chi(y) = g$ and $\chi(y') = g'$. Then,

$\alpha(g) = \psi(y)$ and $\alpha(g') = \psi(y')$ imply that $\psi(y) = \psi(y')$. By (7.1), $\psi(y) = \psi(y')$ implies that $\chi(y) = \chi(y')$ and thus $g = g'$. Since Y generates the free group A°, the map is surjective. Indeed, for any $a \in A$ we have $a = y_1^{\epsilon_1} \cdots y_n^{\epsilon_n}$ with $y_i \in Y$ and $\epsilon_i = \pm 1$. Thus $\psi(a) = \psi(y_1)^{\epsilon_1} \cdots \psi(y_n)^{\epsilon_n} = \alpha(g_1^{\epsilon_1} \cdots g_n^{\epsilon_n})$ with $\chi(y_i) = g_i$.

Finally, the commutative diagrams of Figure 7.4 show that the pair (α, β) is an equivalence of permutation groups.

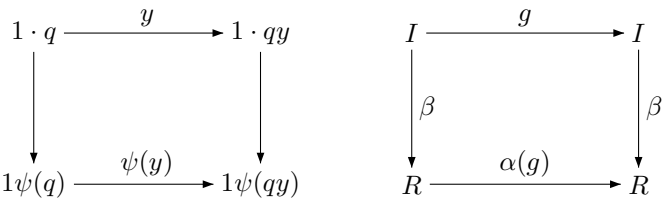

Figure 7.4: The equivalence of G and G'.

■

Example 7.2.6 We illustrate the proof of Theorem 7.2.5. Let Z be the group code of degree 4 recognized by the automaton of Figure 7.5. It is the automaton of Figure 6.9 with more convenient labels for the states. It is clear the $G(Z)$ is

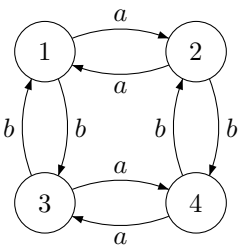

Figure 7.5: A group automaton.

$\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Let F be the Fibonacci set. The code $X = Z \cap F$ is the code of Example 6.6.3. The minimal automaton of X^* is represented on Figure 7.2. Let us chose $u = aba$. It has rank 4, and $\mathrm{Im}(u) = \{1, 2, 4, 8\}$. One gets $Y = \{ba, aba\}$. Next $\chi(ba) = (18)(24)$ and $\chi(aba) = (14)(28)$. The function β maps $1, 2, 4, 8$ to $1, 2, 3, 4$ respectively.

The following example shows that Theorem 7.2.5 does not hold for the set of factors of an episturmian word which is not strict.

Example 7.2.7 Let F and X be as in Example 5.2.7. The bifix code X has F-degree 8. Let $\mathcal{A}$ be the minimal automaton of X^* represented on Figure 5.2. The image of $\varphi(bc)$ is the set $I = \{1, 4, 6, 13, 14, 21, 25, 32\}$. The submonoid $U = \{u \in A^* \mid I \cdot u = I\}$ is generated by $acbc$ and $acacbc$. The restrictions to I

of $\varphi(acbc)$ and $\varphi(acacbc)$ are

$$(1\ 14)(25\ 6)(4\ 21)(32\ 13), \quad (1\ 6)(14\ 25)(4\ 13)(21\ 32).$$

These permutations generate a group which has two orbits: $\{1, 6, 14, 25\}$ and $\{4, 13, 21, 32\}$. The restriction to each orbit is isomorphic to $(\mathbb{Z}/2\mathbb{Z})^2$. Thus the F-group of X is $(\mathbb{Z}/2\mathbb{Z})^2$. However $X = Z \cap F$ where Z is a group code such that $G(Z) = (\mathbb{Z}/2\mathbb{Z})^3$.

Proof of Theorem 7.2.3. Let G be a transitive permutation group of degree d and let Z be a group code on an alphabet A with k letters such that $G(Z) = G$. Let F be a Sturmian set on the alphabet A and let $X = Z \cap F$. Then, by Theorem 7.2.5, $G_F(X) = G$ and, by Theorem 5.2.1, X has $(k-1)d+1$ elements. ∎

Acknowledgments We wish to thank Mike Boyle, Aldo De Luca, Thierry Monteil, Patrice Séébold, Martine Queffélec and Gwénaël Richomme for their help in the preparation of this manuscript.

References

[1] Isabel M. Araújo and Véronique Bruyère. Words derivated from Sturmian words. *Theoret. Comput. Sci.*, 340(2):204–219, 2005.

[2] Pierre Arnoux and Gérard Rauzy. Représentation géométrique de suites de complexité $2n + 1$. *Bull. Soc. Math. France*, 119(2):199–215, 1991.

[3] L. Bartholdi and Pedro Silva. Rational subsets of groups. European science Foundation, 2011.

[4] Marie-Pierre Béal and Dominique Perrin. Codes and sofic constraints. *Theoret. Comput. Sci.*, 340(2):381–393, 2005.

[5] Marie-Pierre Béal and Dominique Perrin. Completing codes in a sofic shift. *Theoret. Comput. Sci.*, 410(43):4423–4431, 2009.

[6] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Cambridge University Press, 2009.

[7] Enrico Bombieri. Continued fractions and the Markoff tree. *Expo. Math.*, 25(3):187–213, 2007.

[8] Arturo Carpi and Aldo de Luca. Codes of central Sturmian words. *Theoret. Comput. Sci.*, 340(2):220–239, 2005.

[9] Yves Césari. Sur un algorithme donnant les codes bipréfixes finis. *Math. Systems Theory*, 6:221–225, 1972.

[10] Yves Césari. Propriétes combinatoires des codes bipréfixes. In D. Perrin, editor, *Théorie des Codes (actes de la septième École de Printemps d'Informatique Théorique)*, pages 20–46. LITP, 1979.

[11] David Gawen Champernowne. The construction of decimals normal in the scale of ten. *J. London Math. Soc.*, 8:254–260, 1933.

[12] Elwin B. Christoffel. Observatio arithmetica. *Annali di Matematica*, 6:145–152, 1875.

[13] Harvey Cohn. Markoff forms and primitive words. *Math. Ann.*, 196:8–22, 1972.
[14] Robert Cori. Indecomposable permutations, hypermaps and labeled Dyck paths. *J. Combin. Theory Ser. A*, 116(8):1326–1343, 2009.
[15] Ethan M. Coven and G. A. Hedlund. Sequences with minimal block growth. *Math. Systems Theory*, 7:138–153, 1973.
[16] Maxime Crochemore and Dominique Perrin. Two-way string-matching. *J. Assoc. Comput. Mach.*, 38(3):651–675, 1991.
[17] Thomas W. Cusick and Mary E. Flahive. *The Markoff and Lagrange spectra*, volume 30 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1989.
[18] Clelia De Felice. Finite biprefix sets of paths in a graph. *Theoret. Comput. Sci.*, 58(1-3):103–128, 1988.
[19] Aldo de Luca. Sturmian words: structure, combinatorics, and their arithmetics. *Theoret. Comput. Sci.*, 183(1):45–82, 1997.
[20] Leonard E. Dickson. *Studies in the theory of numbers*. U. Chicago Press, 1930. Reprinted by Chelsea Publications Co. in 1957.
[21] Andreas W. M. Dress and R. Franz. Parametrizing the subgroups of finite index in a free group and related topics. *Bayreuth. Math. Schr.*, (20):1–8, 1985.
[22] Xavier Droubay, Jacques Justin, and Giuseppe Pirillo. Episturmian words and some constructions of de Luca and Rauzy. *Theoret. Comput. Sci.*, 255(1-2):539–553, 2001.
[23] Fabien Durand. A characterization of substitutive sequences using return words. *Discrete Mathematics*, 179(1-3):89–101, 1998.
[24] Ferdinand Georg Frobenius. *Über die Markoffschen Zahlen*. Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, Berlin, 1913. Also in: Frobenius, F. G., *Gesammelte Abhandlungen*, Springer-Verlag, 1968, vol. 3, 598–627, ed. J.-P. Serre.
[25] Edgar N. Gilbert and Edward F. Moore. Variable length binary encodings. *Bell System Tech. J.*, 38:933–967, 1959.
[26] Amy Glen and Jacques Justin. Episturmian words: a survey. *Theor. Inform. Appl.*, 43:403–442, 2009.
[27] Marshall Hall, Jr. Subgroups of finite index in free groups. *Canadian J. Math.*, 1:187–190, 1949.
[28] Charles Holton and Luca Q. Zamboni. Descendants of primitive substitutions. *Theory Comput. Syst.*, 32(2):133–157, 1999.
[29] Soonjo Hong and Sujin Shin. Cyclic renewal systems. *Theoret. Comput. Sci.*, 410(27-29):2675–2684, 2009.
[30] Jacques Justin and Giuseppe Pirillo. Episturmian words and episturmian morphisms. *Theoret. Comput. Sci.*, 276(1-2):281–313, 2002.
[31] Jacques Justin and Giuseppe Pirillo. Episturmian words: shifts, morphisms and numeration systems. *Intern. J. Found. Comput. Sci.*, 15(2):329–348, 2004.
[32] Jacques Justin and Laurent Vuillon. Return words in Sturmian and episturmian words. *Theor. Inform. Appl.*, 34(5):343–356, 2000.

[33] Ilya Kapovich and Alexei Myasnikov. Stallings foldings and subgroups of free groups. *J. Algebra*, 248(2):608–668, 2002.

[34] Christian Kassel and Christophe Reutenauer. Sturmian morphisms, the braid group B_4, Christoffel words and bases of F_2. *Ann. Mat. Pura Appl. (4)*, 186(2):317–339, 2007.

[35] M. Lothaire. *Combinatorics on Words*. Cambridge University Press, second edition, 1997. (First edition 1983).

[36] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002.

[37] Roger C. Lyndon and Paul E. Schupp. *Combinatorial group theory*. Classics in Mathematics. Springer-Verlag, 2001. Reprint of the 1977 edition.

[38] André A. Markoff. Sur les formes quadratiques binaires indéfinies. *Math. Ann.*, 15(3):381–406, 1879.

[39] André A. Markoff. Sur les formes quadratiques binaires indéfinies. *Math. Ann.*, 17(3):379–399, 1880. (second memoire).

[40] Filippo Mignosi and Patrice Séébold. Morphismes Sturmiens et règles de Rauzy. *J. Théor. Nombres Bordeaux*, 5(2):221–233, 1993.

[41] Marston Morse and Gustav A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *Amer. J. Math.*, 62:1–42, 1940.

[42] Richard P. Osborne and Heiner Zieschang. Primitives in the free group on two generators. *Invent. Math.*, 63(1):17–24, 1981.

[43] Patrice Ossona De Mendez and Pierre Rosenstiehl. Transitivity and connectivity of permutations. *Combinatorica*, 24(3):487–501, 2004.

[44] Dominique Perrin. Sur les groupes dans les monoïdes finis. In *Noncommutative Structures in Algebra and Geometric Combinatorics (Naples 1978)*, volume 109 of *Quaderni de "La Ricerca Scientifica"*, pages 27–36. CNR, 1981.

[45] Dominique Perrin and Giuseppina Rindone. On syntactic groups. *Bull. Belg. Math. Soc. Simon Stevin*, 10(suppl.):749–759, 2003.

[46] Martine Queffélec. *Substitution dynamical systems—spectral analysis*, volume 1294 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, second edition, 2010.

[47] Antonio Restivo. Codes and local constraints. *Theoret. Comput. Sci.*, 72(1):55–64, 1990.

[48] Christophe Reutenauer. Une topologie du monoïde libre. *Semigroup Forum*, 18(1):33–49, 1979.

[49] Christophe Reutenauer. Ensembles libres de chemins dans un graphe. *Bull. Soc. Math. France*, 114(2):135–152, 1986.

[50] Gwénaël Richomme and Patrice Séébold. On factorially balanced sets of words,. Technical report, LIRMM, 2010.

[51] Giuseppina Rindone. Construction d'une famille de codes associés à certains groupes finis. *Theoret. Comput. Sci.*, 54(2-3):165–179, 1987.

[52] Marcel-Paul Schützenberger. On an application of semigroup methods to some problems in coding. *IRE Trans. Inform. Theory*, IT-2:47–60, 1956.

[53] Marcel-Paul Schützenberger. On a family of submonoids. *Publ. Math. Inst. Hungar. Acad. Sci. Ser. A*, VI:381–391, 1961.

[54] Marcel-Paul Schützenberger. On a special class of recurrent events. *Ann. Math. Statist.*, 32:1201– 1213, 1961.

[55] Marcel-Paul Schützenberger. A property of finitely generated submonoids of free monoids. In *Algebraic theory of semigroups (Proc. Sixth Algebraic Conf., Szeged, 1976)*, volume 20 of *Colloq. Math. Soc. János Bolyai*, pages 545–576. North-Holland, Amsterdam, 1979.

[56] Laurent Vuillon. A characterization of Sturmian words by return words. *European J. Combin.*, 22(2):263–275, 2001.

[57] Zhi Xiong Wen and Zhi Ying Wen. Local isomorphisms of invertible substitutions. *C. R. Acad. Sci. Paris Sér. I Math.*, 318(4):299–304, 1994.