Bifix codes and Sturmian words

Jean Berstel¹, Clelia De Felice², Dominique Perrin¹, Christophe Reutenauer³, Giuseppina Rindone¹

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

September 1, 2010

Contents

1 Introduction 2

2 Factorial sets 4
  2.1 Recurrent sets 4
  2.2 Recurrent words 5
  2.3 Probability distributions 8

3 Prefix codes in factorial sets 12
  3.1 Prefix codes 12
  3.2 Automata 13
  3.3 Maximal prefix codes 14
  3.4 Average length 16

4 Bifix codes in recurrent sets 17
  4.1 Indicator 17
  4.2 Maximal bifix codes 19
  4.3 Degree 25
  4.4 Kernel 28
  4.5 Finite maximal bifix codes 30

5 Bifix codes in Sturmian sets 31
  5.1 Sturmian sets 31
  5.2 Cardinality 33
  5.3 Periodicity 34

6 Basis of subgroups 37
  6.1 Main result 37
  6.2 Preliminaries 39
  6.3 Proof of the main result 43
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. We also prove that any $F$-maximal bifix code of degree $d$ is the basis of a subgroup of index $d$ of the free group on the alphabet.

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 [32] and by Gilbert and Moore [15]. These papers already contain significant results. The first systematic study is in the papers of Schützenberger [33],[34]. 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 [7]. He also defined the notion of derived code which allows to build maximal bifix codes by increasing degrees [8].

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 [22]. Many combinatorial properties were described in the paper by Coven and Hedlund [10]. A connexion between Sturmian words and free groups was discovered in the study of Sturmian morphisms by Wen and Wen [37]. This connexion is one of the main points of this paper.
generalized to arbitrary alphabets. Following an initial work by Arnoux and Rauzy [2] and developing ideas of De Luca [12], Droubay, Justin and Pirillo introduced in [14] 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 Sturmian words.

We extend most properties of ordinary maximal bifix codes to bifix codes 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 a Sturmian word. Such words are, by definition, infinite words on a two-letter alphabet which have for all $n \geq 0$, $n + 1$ factors of length $n$. We actually work with the set of factors of a strict episturmian word, called simply a Sturmian set. This allows us to work on an alphabet with $k$ letters. The number of factors of length $n$ of a strict episturmian word is $(k - 1)n + 1$. Our main result is that a maximal bifix code of degree $d$ in a Sturmian set is always a basis of a subgroup of index $d$ of the free group (Theorem 6.1.1). In particular, it has $(k - 1)d + 1$ elements (Theorem 5.2.1). 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 7.3.2). This aspect was first considered by Schützenberger in [35]. 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 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 [35], Schützenberger conjectured a refinement of his result which was subsequently proved by Césari. This study lead to the Critical Factorization Theorem that we will meet again here (Theorem 5.3.4).

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 [29] 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 [11] where the internal transformation is generalized. In [27], 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 [17]. A more general notion was considered in [3]. 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 connexion with unambiguous automata was considered later in [4]. Codes in Sturmian sets have been studied before in [6]. Finally, prefix codes $X$ contained in restricted sets of words are used in [25] 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 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 factorial 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 a bifix code thin and maximal in a uniformly recurrent set is finite (Proposition 4.5.3). 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 (Theorem 5.3.2). The proof uses the Critical Factorization Theorem (see [20]).

Section 6 presents our results concerning free groups. We first prove our main result (Theorem 6.1.1) which states that for a Sturmian set $F$, a bifix code $X \subset F$ is a finite and $F$-maximal bifix code 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 6.1.1 concerning syntactic groups.

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 2.3, we introduce probability distributions on factorial sets.

2.1 Recurrent sets

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. We use the standard terminology on words, in particular concerning prefixes, suffixes and factors (see [20] for example).

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

A set $F$ is said to be \textit{right essential} if it is prefix-closed and for any $w \in F$ there is a letter $a \in A$ such that $wa \in F$. 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 $uv \in F$. neighbour.
Symmetrically, a set $F$ is said to be left essential if it is suffix-closed and if for any $w \in F$ there is a letter $a \in A$ such that $aw \in F$.

A set $F$ 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 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 $uw \in F$.

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 $x \in A^\infty$. The set $F(x)$ is factorial and right essential.

An infinite word $x \in A^\infty$ is said to be recurrent if for any word $u \in F(x)$ there is a $v \in F(x)$ such that $uvu \in F(x)$. Since every factor of a recurrent word $x$ has a second occurrence, it has an infinite number of occurrences.

**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 $u_1v_1u_2 \in F$. Further, since $u_1v_1u_2, u_3 \in F$ there is a word $v_2$ such that $u_1v_1u_2v_2u_3 \in F$. In this way, we obtain an infinite word $x = u_1v_1u_2v_2 \cdots$ 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 $y \in A^N$. Since $x$ is recurrent, the word $v$ is a factor of $y$. Set $y = qvz$ with $q \in F$ and $z \in A^N$. Then $uvq$ is in $F$. Thus $F$ is recurrent.

An infinite word $x \in A^N$ 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.

$$x = ab aaabbabb aaaaababaabbbaabbaabbbabbabb \cdots$$

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, 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 sequence.

We use indifferently the terms of morphism or substitution for a monoid morphism from $A^*$ into itself.

**Example 2.2.4** Set $A = \{a, b\}$. The Thue-Morse morphism is the substitution $f : A^* \rightarrow A^*$ defined by $f(a) = ab$ and $f(b) = ba$. The Thue-Morse word $x = abbaababaababbaabbbabbbabbb \cdots$ is the fixpoint $f^\omega(a)$ of $f$. It is uniformly recurrent (see [21] Example 1.5.10).

An infinite word $x \in A^N$ 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 of infinite words of the form $S_X$ for some $X \subset A^*$.

For any infinite word $x \in A^N$, we denote by $S(x)$ the set of infinite words $y \in A^N$ 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 $X = A^* \setminus F(x)$.

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

The following property is classical (see for example [21] Theorem 1.5.9).

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

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


Example 2.2.6 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 $x = abaababa \cdots$ is the fixpoint $f^\omega(a)$ of $f$. It is a Sturmian word (see [21] Example 2.1.1).

Episturmian words are a generalization of Sturmian words to arbitrary finite alphabets. A factor $u$ of an infinite word $x$ over an alphabet $A$ is right-special if the set $uA \cap F(x)$ contains at least two elements. Symmetrically, a factor $u$ is left-special if the set $Au \cap F(x)$ contains at least two elements.

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 $u$ of length $n$ which is right-special. Since $F(x)$ is closed under reversal, the reversal of the 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 such that each right-special factor satisfies the inclusion $uA \subset F(x)$ (see [14]).

An episturmian word is called standard if all its prefixes are left-special. For any episturmian word $s$, there is a standard one $t$ such that $F(s) = F(t)$. This is the word that has as prefixes the left-special factors of $s$.

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.

Example 2.2.7 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 fixpoint $f^\omega(a) = abacaba \cdots$ is the Tribonacci word. It is an episturmian word (see [19]).

The following is, in the case of Sturmian words, Proposition 2.1.25 in [21]. The general case is Theorem 2 in [14].

Proposition 2.2.8 If $x$ is episturmian, then $S(x)$ is minimal and $x$ is uniformly recurrent.

The converse is false as shown by the following example.

Example 2.2.9 The Thue-Morse word of Example 2.2.4 is not Sturmian. Indeed, it has four factors of length 2.

For $a \in A$, denote by $\psi_a$ the elementary morphism of $A^*$ into itself defined by

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

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

A palindrome is a word 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 \( P(w) \) defined recursively as follows. One has \( P(1) = 1 \) and for \( u \in A^{*} \) and \( a \in A \), one has \( P(ua) = (P(u)a)^{(+)} \).

By Justin’s Formula, one has for every \( u, v \in A^{*} \)

\[
P(uv) = \psi_{u}(P(v))P(u).
\]

An arbitrary episturmian word can be obtained as follows. Let \( \Delta = a_{1}a_{2} \cdots \) be an infinite word. Set \( u_{0} = 1 \) and \( u_{n+1} = P(u_{n}a_{n}) \) for \( n \geq 0 \). Then the \( u_{n} \) form an increasing sequence for the prefix order. The limit is a standard episturmian word \( s \) and any standard episturmian word is obtained in the way. The infinite word \( \Delta \) is called the directive word of \( s \).

**Example 2.2.10** Let \( A = \{a, b, c\} \) and \( \Delta = (abc)^{\omega} \). Then, we have \( u_{1} = a, u_{2} = aba, u_{3} = abacaba \). The limit is the Tribonacci word of Example 2.2.7.

A standard episturmian word is strict if and only if its directive word is such that any letter in \( A \) occurs infinitely often in \( \Delta \).

### 2.3 Probability distributions

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

(i) \( \pi(1) = 1 \),

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

For a right probability distribution \( \pi \) on \( F \) and a set \( X \subset F \), we denote \( \pi(X) = \sum_{x \in X} \pi(x) \). See [5] 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)

\[
\pi(uA^{n} \cap F) = \pi(u).
\]

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

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

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

(iii) \( \sum_{a \in P(w)} \pi(aw) = \pi(w) \), 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 2.3.1** For any right essential set \( F \) of words, there exists a positive right probability distribution \( \pi \) on \( F \).

**Proof.** Consider the map \( \pi : F \to [0,1] \) defined for \( w = a_1a_2\cdots a_n \) by

\[
\pi(w) = \frac{1}{d_0d_1\cdots d_{n-1}}
\]

where \( d_i = \text{Card}(S(a_1\cdots a_i)) \) for \( 0 \leq i \leq n \). By convention, \( \pi(1) = 1 \).

Let us verify that \( \pi \) is a right probability distribution on \( F \). Indeed, let \( w = a_1a_2\cdots a_n \). Since \( F \) is right essential, the set \( S(w) \) is nonempty. Let \( a \in S(w) \), we have \( \pi(wa) = 1/d_0d_1\cdots d_n \). Since \( \text{Card}(S(w)) = d_n \), we obtain that \( \pi \) 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 \((S,\sigma)\) of a compact metric space \( S \) and a continuous map \( \sigma \) from \( S \) into \( S \). Any shift space \( S \) is a topological dynamical system with the transformation defined by the shift map defined by \( \sigma(x_0x_1\cdots) = x_1x_2\cdots \). Indeed, we consider \( A^\mathbb{N} \) as a metric space for the distance defined for \( x = x_0x_1\cdots \) and \( y = y_0y_1\cdots \) 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 \((S,\sigma)\) is said to be invariant if \( \sigma^{-1}(T) = T \).

The following property is well-known (although usually stated for two sided-infinite words, see for example Proposition 1.5.1 in [21]).

**Proposition 2.3.2** The shift spaces are the invariant and closed subsets of \((A^\mathbb{N},\sigma)\).

**Proof.** It is clear that a shift space is both closed and invariant. Conversely, let \( S \subset A^\mathbb{N} \) be closed and invariant 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 \( w_n \in F(y) \) there is an infinite word \( y^{(n)} \in S \) such that \( w_n \in F(y^{(n)}) \). Since \( S \) is invariant 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 \((S,\sigma)\) be a topological dynamical system. A probability measure \( \mu \) on the family \( \mathcal{F} \) of Borel subsets of \( S \) is invariant if \( \mu(\sigma^{-1}B) = \mu(B) \) for any \( B \in \mathcal{F} \).

The following result is from [26] (Krylov and Bogolioubov’s Theorem 4.2).
**Theorem 2.3.3** For any topological dynamical system, there exist invariant probability measures.

Let $F$ be a uniformly recurrent set. By Proposition 2.2.1 there is an infinite word $x$ such that $F(x) = F$. Such an infinite word is by definition uniformly recurrent. By Proposition 2.2.5, the shift space $S = S(x)$ is minimal.

By Theorem 2.3.3 there is an invariant probability measure $\mu$ on $S$. Since $S$ is minimal, every nonempty open set in $S$ has positive measure. Indeed, let $T$ be a nonempty open set with measure 0. Then the set $U = \bigcup_{n \in \mathbb{Z}} \sigma^n(T)$ is a nonempty open invariant set of measure 0. Its complement $V$ is a closed invariant subset of $S$ such that $V \neq \emptyset$ (since $\mu(V) = 1$) and $V \neq S$ (since $U \neq \emptyset$) a contradiction with the fact that $S$ is minimal. Since for any $w \in F$, the set $wA^N \cap S$ is open, we have shown in particular that $\mu(\sigma^{-1}(wA^N \cap S)) = 1$.

Thus we have proved the following result.

**Corollary 2.3.4** For any uniformly recurrent set $F \subset A^*$, there exists positive invariant probability distributions on $F$.

Corollary 2.3.4 is not true in general for a recurrent set.

In some cases, there is a unique invariant probability distribution on the set $F$. Indeed, let $f : A^* \to A^*$ be a morphism. We say that $f$ is primitive if there is a letter $a \in A$ such that $f(a) \in aA^+$ and if $\lim_{n \to \infty} |f^n(b)| = \infty$ for every $b \in A$. Let $x = \lim_{n \to \infty} f^n(a)$. Then, there is a unique invariant probability distribution $\pi_F$ on the set $F(x)$ ([26], Theorem 5.6). We illustrate this result by the following examples.

**Example 2.3.5** Let $x = ababaababaab\cdots$ be the Fibonacci word and let $F$ be the set of factors of $x$. 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 $\pi_F$ can be obtained as follows (see [26]). The vector $v = [\pi(a) \ \pi(b)]$ 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 \( v = [\lambda, 1 - \lambda] \). The other values can be computed using conditions (ii) and (iii) of the definition of an invariant probability distribution.

**Example 2.3.6** Let \( x = abbabaab \cdots \) be the Thue-Morse word and let \( F = F(x) \). 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. 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 3.3.5).

3.1 Prefix codes

The prefix order is defined, for \( u, v \in A^* \), by \( u \leq v \) if \( u \) is a prefix of \( v \). A set \( X \subseteq A^* \) of nonempty words is a prefix code if any two distinct elements of \( X \) are incomparable for the prefix order.

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

We will use formal series to express properties of prefix codes. Let \( F \) be a factorial set. An \( F \)-series is a map \( \sigma \) from \( F \) into \( \mathbb{Z} \). The value of \( \sigma \) on \( w \in F \) is denoted \((\sigma, w)\). We denote by \( \mathbb{Z}^F \) the set of \( F \)-series.

The set \( \mathbb{Z}^F \) is a semiring for the operations of sum and product defined by

\[
(\sigma + \tau, w) = (\sigma, w) + (\tau, w) \\
(\sigma \tau, w) = \sum_{uv = w} (\sigma, u)(\tau, v)
\]

If \( \sigma \) is an \( F \)-series such that \((\sigma, 1) = 0\), we denote \( \sigma^* = \sum_{n \geq 0} \sigma^n \). Then it can be verified that \( \sigma^* \) is the inverse of \( 1 - \sigma \).

For a set \( S \subseteq A^* \), we denote by \( \mathbb{Z}^F S \) the characteristic \( F \)-series of \( S \). By definition, for any \( w \in F \),

\[
(S, w) = \begin{cases} 
1 & \text{if } w \in S \cap F \\
0 & \text{otherwise}
\end{cases}
\]

Note that \( S \) need not be included in \( F \) but that \( \mathbb{Z}^F S = S \cap F \). In particular, \( \mathbb{Z}^F A^* = F \).

Note also that \( \mathbb{Z}^F A^* \) is the inverse of \( 1 - \mathbb{Z}^F A^* \). More generally, for any prefix code \( X \), the \( F \)-series \( X^* \) is the inverse of \( 1 - X \). The following is adapted from Proposition 3.1.6 in [5].

**Proposition 3.1.1** Let \( F \) be a factorial set, let \( X \subseteq F \setminus 1 \) and let \( U = A^* \setminus XA^* \).

Then

\[
F \subseteq X^*U. \tag{3.1}
\]

If \( X \) is a prefix code, then

\[
F = X^*U \quad \text{and} \quad X - 1 = U(A - 1) \tag{3.2}
\]
Proof. We prove Equation (3.1) by induction on the length of $w \in F$. It is true for $w = 1$ since $1 \in U$. Next, if $w \in F$ is nonempty, either $w \in U$ or $w \in XA^*$. In the first case, the conclusion $w \in X^*U$ holds. In the second case, set $w = xw'$ with $x \in X$. Since $F$ is factorial, we have $w' \in F$. By induction hypothesis, we have $w' = yu$ for $y \in X^*$ and $u \in U$. Thus $w = xyu$ is in $X^*U$.

Assume that $X$ is a prefix code. Then it is easy to see that $X^*U = X^*U$. Furthermore, by (3.1), we have $\overline{F} = X^*U \cap F = X^*U$. This shows the first equality in (3.2). The second one is a consequence of the first one using $X^*(1 - X) = 1$ and $\overline{F}(1 - A) = 1$.

3.2 Automata

We recall the basic results on deterministic automata and prefix codes (see [5] for a more detailed exposition).

We denote $A = (Q, i, T)$ 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.

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

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

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

The initial state is the set $X$ and the terminal states are the sets $u^{-1}X$ for $u \in X$.

For $p \in Q$, we denote $\text{Stab}_A(p) = \{ x \in A^* \mid p \cdot x = p \}$.

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

Let $X \subset A^*$ be a prefix code. Then there is a simple automaton $A = (Q, 1, 1)$ such that $X^* = \text{Stab}_A(1)$. In particular, the minimal automaton of $X^*$ is simple.

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

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

It is immediate that this automaton recognizes $X^*$.

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

13
3.3 Maximal prefix codes

We say that a set $E \subset A^*$ is right dense in $F$, or right $F$-dense, if any $u \in F$ is a prefix of an element of $E$.

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

(i) $XA^*$ is right $F$-dense,

(ii) every element of $F$ is comparable with some element of $X$ for the prefix order.

**Proof.** Assume that $XA^*$ is right $F$-dense. For any $u \in F$, there is a word $v$ such that $uv = xv$ with $x \in X$. Then $u$ and $x$ are comparable for the prefix order. Thus (ii) holds. Conversely, let $u \in F$. Let $x \in X$ be comparable with $u$ for the prefix order. Then there exist $v, w$ such that $uv = xw$. Thus $XA^*$ is right $F$-dense.

Let $F \subset A^*$. A set $X \subset F$ is right complete in $F$, or right $F$-complete, if $X^*$ is right dense in $F$.

The following is a generalization to subsets of a factorial set of Proposition 3.3.2 in [5].

**Proposition 3.3.2** Let $F$ be a factorial set and let $X \subset F$ be a set of nonempty words of $F$. The following conditions are equivalent.

(i) $X$ is right $F$-complete,

(ii) $XA^*$ is right $F$-dense,

**Proof.** (i) implies (ii). Let $u$ be a nonempty word in $F$. Since $X$ is right $F$-complete, there exists $v \in A^*$ such that $uv \in X^*$. Then $uv$ has a prefix in $X$ and thus $uv \in XA^*$.

(ii) implies (i). Consider a word $u \in F$. Let us show that $u$ is a prefix of a word in $X^*$. If $u$ is a prefix of a word of $X$, there is nothing to prove. Otherwise, $u$ has a proper prefix in $X$. 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 a word in $X^*$. Thus $u$ is a prefix of some word in $X^*$.

We say that 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 result is a generalization to subsets of a prefix-closed set of Theorem 3.3.5 in [5].

**Proposition 3.3.3** Let $F$ be a prefix-closed set and let $X \subset F$ be a prefix code. Then $X$ is $F$-maximal if and only if $XA^*$ is right $F$-dense.
Proof. Suppose first that $X$ is maximal in $F$. Assume that $u \in F$ is not a prefix of any word in $X A^*$. Then $X \cup u$ is prefix, a contradiction.

Conversely, suppose that $X A^*$ is right dense in $F$. Any word $u \in F$ is a prefix of word in $X A^*$. Thus $u$ is comparable for the prefix order with some word of $X$. This implies that $X$ is maximal in $F$.  

Example 3.3.4 The set $X = \{a, ba\}$ is a maximal prefix code in the set $F$ of factors of the Fibonacci word since $X A^*$ is right $F$-dense.

The following is a generalization of Propositions 3.7.1 and 3.7.2 in [5].

Proposition 3.3.5 Let $F$ be a prefix-closed set. Let $\pi$ be a positive right probability distribution on $F$. Any prefix code $X \subset F$ satisfies $\pi(X) \leq 1$. If $X$ is finite, it is $F$-maximal if and only if $\pi(X) = 1$.

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

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

and the terms of the union are pairwise disjoint. Thus, using Equation (2.1)

$$\pi(X) = \sum_{x \in X} \pi(x A^n - |x| \cap F) \leq \pi(A^n \cap F) = 1.$$  

If $X$ is maximal in $F$, any word in $F \cap A^n$ has a prefix in $X$. Thus we have equality in (3.3) and thus also in (3.4). This shows that $\pi(X) = 1$. The converse is clear since $\pi$ is positive on $F$.

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

The statement has a dual for a suffix code included in a suffix-closed set $F$ with a left probability distribution on $F$.

Example 3.3.6 Let $F$ be the set of factors of the Fibonacci word. The set $X = \{a, ba\}$ is a maximal prefix code (Example 3.3.4). One has $\pi_F(X) = 1$ where $\pi_F$ is defined in Example 2.3.5.

We will use the following result (see Theorem 4.2.11).

Proposition 3.3.7 Let $F$ be a right essential set. For any finite maximal prefix code $X \subset A^+$ the set $X \cap F$ is a finite $F$-maximal prefix code.

Proof. Set $Y = X \cap F$. The set $Y$ is clearly a finite prefix code. We show that $Y A^*$ is right $F$-dense. This will imply that $Y$ is $F$-maximal by Proposition 3.3.3. Let $u \in F$ and suppose that $u$ does not have a prefix in $Y$. Since $F$ is right essential, the word $u$ is a prefix of arbitrary long words $w \in F$. Choose the length of $w$ larger than the maximal length of the words of $X$. Since $X$ is a
maximal prefix code, $XA^*$ is right dense and thus $w$ has a prefix in $X$. This prefix is in $Y$ since $w \in F$. This implies that $u$ is a prefix of a word in $YA^*$. ■

The following example shows that Proposition 3.3.7 is false for infinite prefix codes.

**Example 3.3.8** Let $F = a^*$ and let $X = a^*b$. The set $X$ is a maximal prefix code on the alphabet $A = \{a, b\}$. However $X \cap F = \emptyset$ and thus $X \cap F$ is not $F$-maximal.

### 3.4 Average length

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

$$\lambda(X) = \sum_{x \in X} |x| \pi(x)$$

**Proposition 3.4.1** Let $F$ be a right-essential set and let $\pi$ 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 the words of $X$. Then $\pi(X) = 1$ and $\lambda(X) = \pi(P)$.

**Proof.** We already know that $\pi(X) = 1$ by Proposition 3.3.5. Let us show that for any $p \in P$,

$$\pi(p) = \sum_{x \in pA^* \cap X} \pi(x). \quad (3.5)$$

Let indeed $n$ be an integer larger than the lengths of the words of $X$. Then by Equation (2.1), $\pi(p) = \pi(pA^n \cap F)$. Since $X$ is an $F$-maximal prefix code, each word of $pA^n \cap F$ has prefix in $X$. Thus

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

Since $\pi(xA^{n+|p|-|x|} \cap F) = \pi(x)$, this proves Equation (3.5).

Thus,

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

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

**Example 3.4.2** Let $F$ be the set of factors of the Fibonacci word and let $X = \{a, ba\}$. We have already seen that $X$ is an $F$-maximal prefix code and that $\pi_F(X) = 1$ where $\pi_F$ is the unique invariant probability distribution on $F$. 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 $\pi_F(P) = 1 + (1 - \lambda) = 2 - \lambda$. 

16
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 Indicator

In this section, we generalize the notion of indicator of a bifix code as an $F$-series on a factorial set $F$. Contrary to the sections that follow, the results do not require the hypothesis that $F$ is recurrent.

Let $F$ be a factorial set of words. A set $X \subset F$ 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 $w \in F$ with respect to a set $X \subset F$ is a triple $(v,x,u)$ such that $w = vxu$ with $v \in A^* \setminus A^* X$, $x \in X^*$ and $u \in A^* \setminus X A^*$. We denote by $\Pi(w)$ the set of parses of $w$.

**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 $v = pz$.

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

The $F$-indicator of a set $X \subset F$ is the $F$-series denoted $L_{X,F}$ or $L_X$ when $F$ is understood or simply $L$ when $X$ is also understood such that for any $w \in F$, $(L,w)$ is the number of parses of $w$ with respect to $X$.

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

The following is a reformulation of Proposition 6.1.6 in [5].

**Proposition 4.1.3** Let $F$ be a factorial set and let $X \subset F$ be a prefix code. For every word $w \in F$, $(L,w)$ 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.

17
Note that, as a consequence of Proposition 4.1.3, we have for two prefix codes \( X, Y \),
\[ X \subset Y \Rightarrow L_Y \leq L_X. \] (4.1)
Indeed, a word without suffix in \( Y \) is also a word without suffix in \( X \).

**Proposition 4.1.4** Let \( F \) be factorial set. Let \( X \subset A^\ast \) be a prefix code and let \( V = A^\ast \setminus A^\ast X \). Then
\[ V = L(1 - A). \] (4.2)
If \( X \) is bifix, one has
\[ 1 - X = (1 - A)L(1 - A) \] (4.3)

**Proof.** Let \( U = A^\ast \setminus X A^\ast \). By definition of the \( F \)-indicator, we have \( L = V X^\ast U \).
Since \( X \) is prefix, we have by Proposition 3.1.1, the equality \( F = X^\ast U \). Thus we obtain \( L = V F \) (note that this is actually equivalent to Proposition 4.1.3).
Multiplying both sides on the right by \((1 - A)\), we obtain Equation (4.2).
If \( X \) is suffix, we have by the dual of Proposition 3.1.1, the equality \( 1 - X = (1 - A)V \), whence the result multiplying both sides of Equation (4.2) on the left by \((1 - A)\).

The following is a generalization of Proposition 6.1.11 in [5]. The proof is quite similar.

**Proposition 4.1.5** Let \( F \) be a factorial set. An \( F \)-series \( L \) is the indicator of a bifix code \( X \subset F \) if and only if it satisfies the following conditions.

(i) For any \( a \in A \) and \( w \in F \) such that \( aw \in F \)
\[ 0 \leq (L, aw) - (L, w) \leq 1 \] (4.4)

(ii) For any \( w \in F \) and \( a \in A \) such that \( wa \in F \)
\[ 0 \leq (L, wa) - (L, w) \leq 1 \] (4.5)

(iii) For any \( a, b \in A \) and \( w \in F \) such that \( awb \in F \)
\[ (L, aw) + (L, wb) \geq (L, w) + (L, awb) \] (4.6)

(iv) \((L, 1) = 1\)

The following is a reformulation of Proposition 6.1.12 in [5].

**Proposition 4.1.6** Let \( F \) be a factorial set and let \( X \subset F \) be a prefix code.
For any \( u \in F \) and \( a \in A \) such that \( ua \in F \), one has
\[ (L, ua) = \begin{cases} (L, u) & \text{if } ua \in A^\ast X \\ (L, u) + 1 & \text{otherwise} \end{cases} \] (4.7)
Proof. This follows directly from Proposition 4.1.4.

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

Recall also that by Proposition 6.1.8 in [5], for a bifix code \(X\) and for all \(u, v, w \in F\) such that \(uvw \in F\), one has

\[
(L, v) \leq (L, uvw).
\] (4.8)

4.2 Maximal bifix codes

Let \(F\) be factorial set. 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, for a uniformly recurrent set \(F\), there exist bifix codes \(X \subset F\) which are not \(F\)-thin.

Example 4.2.1 Let \(F\) be the set of factors of the Thue-Morse word, which is a fixpoint 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 \(u \rightarrow \bar{u}\) 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 \(X = \{x_{2n}x_{2n}^* \mid n \geq 1\}\). We have \(X \subset F\). Indeed, for \(n \geq 2\), \(x_{n+2} = x_{n+1} \bar{x}_{n+1} = x_n \bar{x}_n \bar{x}_n x_n\) implies that \(x_n \bar{x}_n \in F\) and thus \(x_n x_n \in F\). Next \(X\) is a bifix code. Indeed, for \(n < m\), \(x_{2m}\) begins with \(x_{2n} \bar{x}_{2n}\) and thus cannot have \(x_{2n}^2\) as a prefix. Similarly, since \(x_{2m}\) ends with \(\bar{x}_{2n} x_{2n}\), it cannot have \(x_{2n}^2\) 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 \(x_n\). If \(n\) is even, then \(u\) is a factor of \(x_{2n}\) as a prefix. Otherwise, it is a factor of \(x_{2n+1} = x_n \bar{x}_n \bar{x}_n \bar{x}_n\).

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

\[H(X) = \{w \in A^* \mid A^+wA^+ \cap X \neq \emptyset\}\]

the set of internal factors of words in \(X\).

When \(F\) is right essential and left essential, \(X\) is \(F\)-thin if and only if \(F \setminus H(X) \neq \emptyset\). Indeed, the condition is necessary. Conversely, if \(w\) is in \(F \setminus H(X)\), 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 set \(X \subset F\) is maximal in \(F\), or \(F\)-maximal, if it is not properly contained in any other bifix subset of \(F\).

The following is a generalisation Proposition 6.2.1 in [5].

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 $C(X, F)$ be the set of pairs $(u, v)$ of words such that $uvu \in F$ 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 $\Pi(u)$ of parses of $u$ as follows. Let $\pi = (s, x, p)$ and $\pi' = (s', x', p')$ be two parses of $u$. Then $(\pi, \pi') \in \varphi_{u,v}$ if and only if $pvs' \in X^*$ (see Figure 4.1). We prove a series of lemmas concerning the relations $\varphi_{u,v}$ (see Exercise 6.2.1 in [5]).

![Figure 4.1: The relation $\varphi_{u,v}$](image)

**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 $\Pi(u)$ into itself. The converse is true if $X$ is an $F$-maximal suffix code.

**Proof.** Assume first that $X$ is a prefix code. For $(u, v) \in C(X, F)$, let $\pi = (s, x, p)$, $\pi' = (s', x', p')$ and $\pi'' = (s'', x'', p'')$ be three parses of $u$ such that $(\pi, \pi')$ and $(\pi, \pi'')$ are in $\varphi_{u,v}$. We may suppose that $s' = s''$. Since $pvs', pvs''$ are in $X^*$, we have $w \in X^*$. Since $s' \notin A^*X$, this forces $s' = s'\prime$. Hence $x'p' = x''p''$. By Equation (3.2), every word in $F$ has exactly one factorization $wu$ with $u \in X^*$ and $v \in U$. Thus $x'p' = x''p''$ implies $x' = x''$, and $p' = p''$. Thus $\pi' = \pi''$.

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 $x'' = x'p'$. Since $X$ is a suffix code, we have $p' \in A^* \setminus A^*X$.

Since $X$ is $F$-thin, there is a word $w \in F \setminus H(X)$. Since $F$ is recurrent, there is a word $r''$ such that $x''r''w \in F$. Let $u = r''r''w$ and let $t$ be such that $utx'u \in F$. Set $v = tx'$. Thus $(u, v) \in C(X, F)$. By Equation (3.1) there exist $z', z'' \in X^*$ and $p', p'' \in A^* \setminus XA^*$ such that $u = z'p' = r'z''p''$ (see Figure 4.2). By the dual of Equation (3.1), there exist $s \in A^* \setminus A^*X$ and $z \in X^*$ such that $ut = sz$.

Since $X$ is left $F$-complete and $u \notin H(X)$, there is a parse $\pi = (s, x, p)$ such that $z = xpt$ with $pt \in X^*$. Then $\pi = (s, x, p)$, $\pi' = (1, z', p')$ and $\pi'' = (r', z'', p'')$ are three parses of $u$ such that $(\pi, \pi')$, $(\pi, \pi'') \in \varphi_{u,v}$ with $\pi' \neq \pi''$. Thus $\varphi_{u,v}$ is not a partial function.

$\blacksquare$
Lemma 4.2.3 has a dual formulation for suffix codes. Recall that a set $X \subset F$ is right $F$-complete if any word in $F$ is a prefix of a word in $X^*$.

**Lemma 4.2.4** Let $F$ be a recurrent set and let $X$ 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 $\Pi(u)$ into itself.

**Proof.** Assume first that $X$ is right $F$-complete. Let $u, v \in F$ be such that $(u, v) \in C(X, F)$. Let us show that for any $\pi \in \Pi(u)$, there is a parse $\pi' \in \Pi(u)$ such that $(\pi, \pi') \in \varphi_{u,v}$. Let $\pi = (s, x, p)$ be a parse of $u$. Then $puv \in F$. Since $X$ is right $F$-complete, there is word $w$ such that $pvu \in X^*$. Since $u \in F \setminus H(X)$, this implies that there is a parse $\pi' = (s', x', p')$ of $u$ such that $pvs', p'w \in X^*$. Thus $(\pi, \pi') \in \varphi_{u,v}$.

Conversely, assume that for all $(u, v) \in C(X, F)$, the relation $\varphi_{u,v}$ contains a total function from $\Pi(u)$ onto itself. Let $u \in F$. Let $w \in F \setminus H(X)$ and let $v$ be such that $uvw \in F$. Set $r = uvw$. Let $t$ be such that $rtr \in F$. Then $(r, t) \in C(X, F)$. Let $\pi = (s, x, p)$ be a parse of $r$ such that $s = 1$ (such a parse exists by (3.1)). By the hypothesis, there is a parse $\pi' = (s', x', p')$ of $u$ such that $(\pi, \pi') \in \varphi_{r,t}$. Then $pts' \in X^*$. Since $u$ is a prefix of $r$ which is a prefix of $xpts'$, we have shown that $u$ is a prefix of a word in $X^*$. Thus $X$ is right $F$-complete.

Lemma 4.2.4 has a dual formulation for left $F$-complete sets.

**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 $\Pi(u)$ 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 $\Pi(u)$ into itself is injective. Since $\Pi(u)$ is a finite set, it 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 $\Pi(u)$ 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 ∈ F be such that X ∪ y is
prefix and X ∪ z is suffix. Since F is uniformly recurrent, there is a word u such
that yuz ∈ F. Then X ∪ yuz is bifix and thus we reach a contradiction.

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

(ii) implies (iii). Consider the set
Y = X \ 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.

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.

The following is a generalization of Theorem 6.3.1 in [5].

Theorem 4.2.8 Let F be a recurrent set and let X ⊂ F be a bifix code. Then X
is an F-thin and F-maximal bifix code if and only if the F-indicator L = L_{X,F}
of X is bounded. In this case,

\[ H(X) = \{ w ∈ F \mid (L, w) < d_F \} \]  \hspace{1cm} (4.9)

where d_F is defined as d_F = max\{(L, w) \mid w ∈ F\}.

Proof. Assume first that X is an F-thin and F-maximal bifix code. 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 X is F-thin,
F \ H(X) is not empty. Let u ∈ F \ H(X) and w ∈ F. Since F is recurrent, there
is a word v ∈ F such that uvw ∈ 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 \( u \) is not an internal factor of a word in \( X \), any prefix of \( uw \) which is in \( S \) is a prefix of \( u \). Thus \( (L, uw) = (S \Delta^*, uw) = (S \Delta^*, u) = (L, u) \) by Equation (4.8). Since by Equation (4.8), \((L, w) \leq (L, uw)\), we get \((L, w) \leq (L, u)\). This shows that \( L \) is bounded and moreover that \( F \setminus H(X) \) is contained in the set of words of \( F \) with maximal value of \( L \). Conversely, consider \( w \in H(X, F) \). Then there exists \( w' \in X \) and \( p, s \in A^+ \) such that \( w' = pws \). Then \((L, w') \succ (L, w)\) and thus \((L, w)\) is not maximal in \( F \). This proves also Equation (4.9).

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

This implies that a) \( X \) is \( F \)-thin and b) \( XA^* \) 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 \( pws \in X \). Then \( pw \in F \) implies \( pw \in XA^* \), a contradiction. Thus \( w \in F \setminus H(X) \). Furthermore, for any \( v \in F \), since \( F \) is recurrent, there is a word \( u \in F \) such that \( vuw \in F \). Then \( vuw \in XA^* \) by the above argument.

Thus \( XA^* \) is right \( F \)-dense and \( X \) is an \( F \)-maximal bifix code by Theorem 4.2.2.

Let \( F \) be a recurrent set. The \( F \)-degree, denoted \( d_F(X) \), of a bifix code \( X \subseteq F \) is the maximal number of parses of words of \( F \) with respect to \( X \). By Theorem 4.2.8, a bifix code has finite degree if and only if it is \( F \)-thin and \( F \)-maximal.

**Example 4.2.9** Let \( F \) be the set of factors of the Fibonacci word. 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 \((b, a, b)\). Thus \( d_F(X) = 2 \).

**Example 4.2.10** Let \( F \) be the set of factors of the Fibonacci word. Then \( X = \{aaba, ab, baa, baba\} \) 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 \((aa, ba)\) and it is in \( F \setminus H(X) \).

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.

The proof uses the following lemma. In both, we denote \( L = L_{X, A^*} \).
Lemma 4.2.12 Let $X \subset A^+$ be a thin maximal bifix code. For any words $u \in A^+$ and $v \in A^*$ such that $(L, uvw) = (L, u)$, one has $uvw \in XA^*$.

Proof. Since $(L, uvw) = (L, u)$, there is a bijection between the parses of $uvw$ and $u$. Thus, for any parse $(s, z, p)$ of $uvw$, $s$ is a prefix of $u$ and $p$ is a suffix of $u$. Thus for any parse $(s, x, q) \cdot (t, y, p)$ of $u$ there is a unique parse $(t, y, p)$ of $u$ such that $(s, xqvt, y, p)$ is a parse of $uvw$. If $u$ is in $XA^*$, there is nothing to prove. Otherwise $(1, 1, u)$ is a parse of $u$. Let $(t, y, p)$ be a parse of $u$ such that $(1, uvty, p)$ is a parse of $uvw$. Then $uvty$ is in $X^*$ and in fact in $X^+$ since $u \in A^+$. Thus $uvw \in XA^*$.

Proof of Theorem 4.2.11. Since $X$ is thin, its indicator is bounded. Let $w \in F$ be such that $(L, w)$ is maximal among the values of $L$ on $F$. Then $w$ cannot be an internal factor of a word in $Y$. Thus $Y$ is $F$-thin.

We show that $XA^*$ is right $F$-dense. Since $XA^* \cap F = YA^* \cap F$, it will imply that $YA^*$ is right $F$-dense and thus that $Y$ is an $F$-maximal prefix code by Proposition 3.3.3.

Let $u \in F$. We prove that there is a word $v \in F$ such that $uwv \in XA^*$. Let $(w_n)_{n \geq 0}$ be the sequence of words of $F$ defined as follows. Set $w_0 = u$. For $n \geq 0$, define inductively $w_{n+1}$ from $w_n$ as follows. Since $F$ is recurrent and $w_n \in F$ there exists a word $v_n \in F$ such that $w_nv_nw_n \in F$. We define $w_{n+1} = w_nv_nw_n$. Since $(L, w_0) \leq (L, w_1) \leq \ldots$ there is an integer $n$ such that $(L, w_n) = (L, w_{n+1})$. By Lemma 4.2.12, this implies $w_{n+1} \in XA^*$. Since $u$ is a prefix of all $w_n$ this implies that $u$ is a prefix of a word in $XA^*$. Thus $XA^*$ is right $F$-dense.

The proof that $Y$ is an $F$-maximal suffix code is symmetrical.

Since a parse of a word in $F$ with respect to $Y$ is also a parse with respect to $X$, we have $d_F(Y) \leq d(X)$.

Assume that $X$ is finite. To show that $d_F(Y) = d(X)$, consider a word $y$ such that $y$ is not in $H(Y, F)$. Thus $(L_Y, y) = d_F(Y)$. If $(L, y) < d(X)$, then $y$ is in $H(X)$. Thus there exist $p, s \in A^+$ such that $pyv \in X$. Consider $V = \{v \in A^+ \mid pyv \in X\}$. The set $V$ is a finite maximal prefix code. Since $y^{-1}F$ is a right essential prefix-closed set, by Proposition 3.3.7, the set $V \cap y^{-1}F$ is an $F$-maximal prefix code. Thus it is nonempty. Let $v \in V \cap y^{-1}F$ and let $W = \{w \in A^+ \mid wyv \in X\}$. Since $X$ is a finite maximal suffix code, the set $W$ is a finite maximal suffix code. Consider the set $G = F(yv)^{-1}$. It is a suffix-closed set and since $yw \in F$, it is left essential. By the dual of Proposition 3.3.7, the set $W \cap G$ is a $G$-maximal suffix code. Thus it is nonempty. Let $w \in W \cap G$. Then $wyv \in Y$ and thus $y$ is in $H(Y, F)$, a contradiction.

Example 4.2.13 The set $X = a \cup ba^*b$ is a maximal bifix code of degree 2. Let $F$ be the set of factors of the Fibonacci word. Then $X \cap F = \{a, baab, bab\}$ (see Figure 2.1).

As another example, let $Z = \{a^3, a^2ba, a^2b^2, ab, ba^2, baba, b^2a, b^2a, b^3\}$. The set $Z$ is a finite maximal bifix code of degree 3 (see [5]). Then $Z \cap F = \{a^2ba, ab, ba^2, baba\}$ (see Figure 2.1).

24
Example 4.2.14 Let \( F \) be the set of factors of the Thue-Morse word. 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 example shows that a strict inequality can hold in Theorem 4.2.11.

Example 4.2.15 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.

We will see in Example 4.4.3 that a strict inequality can hold in Theorem 4.2.11 even if all letters occur in \( F \).

4.3 Degree

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 [5].

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 \( K(X) = H(X, F) \cap X \).

Theorem 4.3.1 Let \( F \) be a recurrent set. Let \( X \subset F \) be an \( F \)-thin and \( F \)-maximal bifix code of degree \( d_F(X) \geq 2 \). Set \( H = H(X, F) \) and \( K = K(X) \). Let \( G = (HA \cap F) \setminus H \) and \( D = (AH \cap F) \setminus H \). Then the set \( X' = K \cup (G \cap D) \) is an \( F \)-thin set which is an \( F \)-maximal bifix code of degree \( d_F(X) - 1 \).

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 words in \( X \) and let \( S \) be the set of proper suffixes of words in \( X \).

Lemma 4.3.2 One has \( G \subset S \) and \( D \subset P \).

Proof. By Theorem 4.2.8, the \( F \)-indicator of \( X \) is bounded and \( F \setminus H(X) = F \setminus H \) is the set of words in \( F \) with maximal value \( d_F(X) \). Let \( y = ha \) be in \( G \) with \( h \in H \) and \( a \in A \). Since \( y \notin H \), we have \( (L_X, ha) > (L_X, h) \). Thus, by Proposition 4.1.6, \( y = ha \) does not have a suffix in \( X \). Since \( A^+X \) is right 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. \( \blacksquare \)

Lemma 4.3.3 For any \( x \in X \setminus K \), the shortest prefix of \( x \) which is not in \( H \) is in \( X' \).

Proof. Since \( x \notin K \), we have \( x \notin H \). Let \( x' \) be the shortest prefix of \( x \) which is not in \( H \) or, equivalently such that \( (L_X, x') = d_F(X) \). 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), \( (L_X, x) = (L_X, p) \). Thus \( p \notin H \) and \( x' \) is a prefix of \( p \).
Set \( x' = p'a' \) with \( p' \in A^* \) and \( a' \in A \). By definition of \( x' \) we have \( p' \in H \).

Thus \( x' \in G = (HA \cap F) \setminus H \).

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), \((L_X, s) < (L_X, x')\). Thus \( s \) is in \( H \). 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 a word in \( X \). Thus \( k \in X \) would be a proper prefix of a word in \( X \), 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 \( H \), it follows that \( u \) is in \( H \), a contradiction.

Finally, no word \( y \in G \cap D \) can be a proper prefix of another word \( y' \) in \( G \cap D \), otherwise \( y = y'z \), with \( z \in A^+ \). Therefore, since \( G \subset S \) by Lemma 4.3.2, there is \( t \in A^+ \) such that \( ty = ty'z \in X \). Consequently, \( y' \in G \cap H \), 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 \( H \). 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 \((L_X, x)\) is maximal on \( X \). If \( x \) were an internal factor of a word \( y \in X \), then \((L_X, x) < (L_X, y)\) which is impossible. 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 \((G \cap D) \cap H(X)\) is empty, and since \( H(X') \subset H(X) \), \( x' \) cannot be in \( H(X') \). Thus the number of parses of \( x' \) with respect to \( X' \) is \( d_F(X') \).

Let \( P' \) be the set of proper prefixes of words in \( X' \). 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 Proposition 4.1.3.

Since \( x' \in F \setminus H \), we have \((L_X, x') = d_F(X)\). 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 \( AH \), \( p \) is in \( H \) and thus also \( x'' \in H \). 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 a word 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 set of factors of the Fibonacci word. Let \( X = \{a, bab, baab\} \). The set \( X \) is an \( F \)-thin and \( F \)-maximal bifix code of degree 2.
(see Example 4.2.9). We have \( K = \{a\} \), \( H = \{1, a, aa\} \), \( G = \{b, ab, aab\} \) and \( D = \{b, ba, baa\} \). Thus \( X' = \{a, b\} \).

The following is a generalization of Proposition 6.3.14 in [5].

**Proposition 4.3.5** Let \( F \) be a recurrent set. Let \( X \subseteq F \) be an \( F \)-thin and \( F \)-maximal bifix code of degree \( d_F(X) \geq 2 \). Let \( S \) be the set of proper suffixes of \( X \) and set \( H = H(X, F) \). The set \( S \setminus H \) is an \( F \)-maximal prefix code and the set \( S \cap H \) 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 \( X \subseteq F \) be an \( F \)-thin and \( F \)-maximal bifix code. Let \( S \) be the set of proper suffixes of \( X \) and set \( H = H(X, F) \). For any \( w \in F \setminus H \) the longest prefix of \( w \) which is in \( S \) is not in \( H \).

**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 \((L_X, st') = (L_X, s)\). 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 \( s t'a \notin S \), we have \( s t'a \in A^*X \). Thus by Equation (4.7), this implies \((L_X, st'a) = (L_X, st')\). Thus \((L_X, st'a) = (L_X, s)\). We conclude that \((L_X, st) = (L_X, s)\). Since \( w = st \) is in \( F \setminus H \), and since \( F \setminus H \) is the set of words in \( F \) with maximal value of \( L_X \), this implies that \( s \in F \setminus H \). \( \blacksquare \)

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

**Proof of Proposition 4.3.5.**

Set \( Y = S \setminus H \). 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 \( u \notin H \), this forces \( v = 1 \). Thus \( Y \) is prefix.

We show next that \( YA^* \) is right \( F \)-dense. Consider \( u \in F \) and let \( w \in F \setminus H \). Since \( F \) is recurrent, there exists \( v \in F \) such that \( uvw \in F \). Let \( s \) be the longest word of \( S \) which a prefix of \( uvw \). By Lemma 4.3.6, we have \( s \in F \setminus H \). Thus \( s \in S \setminus H = Y \) 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 \( S \cap H \). Let \( s \) be a proper suffix of a word \( x' \in X' \). If \( x' \in K \), then \( s \) is in \( S \cap H \). Suppose next that \( x' \in Y \cap D \). Since \( Y \subseteq S \) by Lemma 4.3.2, we have \( s \in S \). Furthermore, since \( D \subseteq AH \), we have \( s \in H \). This shows that \( s \in S \cap H \).

Conversely, let \( s \) be in \( S \cap H \). 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 \( H \). By the dual of Lemma 4.3.3, the word \( y \) is in \( X' \). Then \( s \) is a proper suffix of \( y \) (since \( s \in H \) and \( y \notin H \)) 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 \( X \subseteq F \).

The following property generalizes Theorem 6.3.15 in [5].
Theorem 4.3.7 Let $F$ be a recurrent set. Let $X$ be an $F$-thin and $F$-maximal bifix code of $F$-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 the words of $X$. If $d = 1$, then $S \setminus 1$ is empty. If $d \geq 2$, by Proposition 4.3.5, the set $Y = S \setminus H$ is an $F$-maximal prefix code and the set $S \cap H$ is equal to the set $S'$ of proper suffixes of the words of $X'$. 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 union of $d - 1$ $F$-maximal prefix codes. □

The following generalizes Corollary 6.3.16 in [5], 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 3.3.5). 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 $\pi$ on $F$. Let $X$ be a finite $F$-maximal bifix code of $F$-degree $d$. The average length of $X$ with respect to $\pi$ is equal to $d$.

Proof. Let $\pi$ be a positive invariant probability distribution on $F$. By the dual of Proposition 3.4.1, one has $\lambda(X) = \pi(S)$. 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 3.3.5, we have $\pi(Y_i) = 1$ for $1 \leq i \leq d - 1$. Thus $\lambda(X) = d$. □

Example 4.3.9 Let $F$ be the set of factors of the Fibonacci word and let $X = \{a, bab, baab\}$ (Example 4.3.4). The set $X$ is an $F$-maximal bifix code of degree 2. With respect to the unique invariant probability distribution of $F$ (Example 2.3.5), we have $\lambda(X) = \lambda + 3(2 - 3\lambda) + 4(2\lambda - 1) = 2$.

4.4 Kernel

In this section, 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 [5].

Proposition 4.4.1 Let $F$ be a recurrent set. Let $X \subset F$ be an $F$-thin and $F$-maximal bifix code of $F$-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 $w \in H(X,F) \cup Y$,

\[
(L_Y, w) = (L_X, w). \tag{4.10}
\]

For all $w \in F$,

\[
(L_X, w) = \min\{d, (L_Y, w)\}. \tag{4.11}
\]
Proof. Denote by $F(w)$ the set of factors of the word $w$. Notice that Equation (4.3) is equivalent to

$$L_X = E(X - 1)E.$$ 

Thus, to prove (4.10), we have to show that for any $w \in H(X, F) \cup Y$ 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 $H(X, F)$, then $F(w) \cap X \subset K$ and thus $F(w) \cap X \subset F(w) \cap Y$. Next, assume that $w$ is in $X$. 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 (4.11), assume first that $w \in H(X, F)$. Then $(L_X, w) < d$ by Theorem 4.2.8. Moreover, $(L_X, w) = (L_Y, w)$ by Equation (4.10). Thus Equation (4.11) holds. Next, suppose that $w \in \bar{H}(Y, F)$. Then $(L_X, w) = d$. Since $Y \subset X$, we have $(L_X, w) \leq (L_Y, w)$ by Equation (4.1). This proves (4.11).

Proposition 4.4.1 will be used to prove the following generalization of Theorem 6.4.2 in [5].

**Theorem 4.4.2** Let $F$ be a recurrent set and let $X \subset F$ be an $F$-thin and $F$-maximal bifix code. Set $K = K(X)$ and $d = d_F(X)$. Then for any $w \in F$

$$(L_X, w) = \min\{d, (L_K, w)\}.$$ 

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

**Proof.** Take $Y = K$ in Proposition 4.4.1. Then the formula follows from Equation (4.11). Next $X$ is determined by $L_X$ through Equation (4.3).

The next example shows that a strict inequality can hold in Theorem 4.2.11 even if all letters occur in the words of $F$.

**Example 4.4.3** Let $F$ be the set of factors of the Fibonacci word. Let $X$ be the maximal bifix code of degree 3 with kernel $K = \{aa, ab, ba\}$. Then $X \cap F = K$ since $K$ is an $F$-maximal bifix code. Thus $d(X) = 3$ but $d_F(X \cap F) = 2$.

We now state the following generalization of Theorem 6.4.3 in [5].

**Theorem 4.4.4** Let $F$ be a recurrent set. A bifix code $Y \subset F$ is the kernel of some $F$-thin and $F$-maximal bifix code of $F$-degree $d$ if and only if

(i) $Y$ is not an $F$-maximal bifix code,

(ii) $\max\{(L_Y, y) \mid y \in Y\} \leq d - 1$.

**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 degree $d - 1$. Moreover, for every $y \in Y$ one has $(L_X, y) \leq d - 1$. Since $(L_X, y) = (L_Y, y)$ by Equation (4.10), condition (ii) is also satisfied.
Conversely, let $Y \subset F$ be a bifix code satisfying conditions (i) and (ii). Let $L \in \mathbb{Z}^F$ be the $F$-series defined by

$$(L, w) = \min\{d, (L_Y, w)\}.$$  

It can be verified that $L$ satisfies the four conditions of Proposition 4.1.5. Thus $L$ is the $F$-indicator of a bifix code $X \subset F$. Since $L = L_X$ is bounded, 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 $F$-series $L_Y$ is not bounded. Consequently $\max\{(L, w) \mid w \in F\} = d$, showing that $X$ has $F$-degree $d$. Let us prove finally that $Y$ is the kernel of $X$. Since, by condition (ii), $\max\{(L_Y, y) \mid y \in Y\} \leq d - 1$, we have $Y \subset H(X, F)$.

Moreover, for $w \in H(X, F)$ we have $(L_X, w) = (L_Y, w)$. Since $1 - X = (1 - A)L(1 - A)$ and $1 - Y = (1 - A)L_Y(1 - A)$ by Equation (4.3), we conclude that for $w \in H(X, F)$, we have $(X, w) = (Y, w)$. This implies that if $w \in H(X, F)$, then $w$ is in $X$ if and only if $w$ is in $Y$. Thus $K(X) = H(X, F) \cap X = H(X, F) \cap Y = Y$ and $Y$ is the kernel of $X$.

**Example 4.4.5** Let $A = \{a, b\}$ and let $F \subset A^*$ be the set of factors of the Fibonacci word. There are three maximal bifix codes of degree 2 in $F$ represented on Figure 4.3.

![Figure 4.3: The three maximal bifix codes of degree 2 in the factors of the Fibonacci word](image)

**4.5 Finite maximal bifix codes**

The following generalizes Theorem 6.5.2 of [5].

**Theorem 4.5.1** 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 degree $d$.

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

![image](image)
Example 4.5.2 Let $A = \{a, b\}$ and let $F$ be the set of words without factor $bb$. There are two $F$-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 degree 2 with kernel $a$ 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 $\text{Card}(A) \geq 2$, there exist infinite maximal bifix codes of degree 2 and thus of all degrees $d \geq 2$.

Proposition 4.5.3 Let $F$ be a uniformly recurrent set. Any $F$-thin bifix code $X \subset F$ is finite.

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

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 (Proposition 4.5.3). 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). This generalizes the fact that $\text{Card}(F \cap A^n) = (k - 1)n + 1$ for all $n \geq 1$. Additionally, if an infinite word $x$ is such that $\text{Card}(F(x) \cap X) \leq d$ for some finite maximal bifix code $X$, then $x$ is ultimately periodic (Theorem 5.3.2).

5.1 Sturmian sets

A set of words is called Sturmian if it is the set of factors of a strict episturmian word.

By Proposition 2.2.8 a Sturmian set is uniformly recurrent.

Let $F$ be a Sturmian set on the alphabet $A$. A word $w$ is right-special if one has $wA \subset F$. It is said to be left-special if $Aw \subset F$.

Observe that a suffix of a right-special word is right-special and that a prefix of a left-special word is left-special.

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 recurrent and

(i) it is closed under reversal,

(ii) for each $n$, there is exactly one right-special word $w$ of length $n$. 

31
The following statement is a direct consequence of the definitions. A *left infinite* word is a sequence \( x = \cdots a_{-2} a_{-1} a_0 \) of letters indexed by the set of negative integers and 0.

**Proposition 5.1.2** Let \( F \) be a Sturmian set of words. There is a unique left infinite word such that all its suffixes are right-special.

A symmetric statement holds for left-special words.

As a consequence of Proposition 5.1.2, for every right-special word \( w \), exactly one of the words \( aw \) for \( a \in A \) is right-special. More generally, for every \( n \geq 1 \) there is exactly one word \( g \) of length \( n \) such that \( gw \) is right-special.

Similarly, for every left-special word, there is exactly one of the words \( wa \) for \( a \in A \) which is left-special.

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

*Proof.* Let indeed \( F \) be a Sturmian set and let \( u \in F \). Since \( F \) is uniformly recurrent, there is an integer \( n \) such that \( u \) is a factor of any word in \( F \cap A^n \). Let \( w \) be the right-special word of length \( n \). Then \( w = pus \). Since \( w \) is right-special, its suffix \( us \) is also right-special. Thus \( u \) is a prefix of a right-special word. \( \Box \)

The following example shows that for a Sturmian set \( F \), there exists bi-fix 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 the set of factors of the Fibonacci word. Consider the following sequence \( (x_n)_{n \geq 1} \) of words of \( F \). Set \( x_1 = a \). 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, not \( F \)-maximal and such that \( (L_{X_n}, x_n) \geq n \). Define \( x_{n+1} \) as follows. Since \( X_n \) is not \( F \)-maximal, \( L_{X_n} \) is not bounded. Let \( u \) be a word in \( F \) which is incomparable for the prefix order with the words of \( X_n \) and such that \( (L_{X_n}, u) \geq n - 1 \). By Proposition 5.1.3, the word \( u \) is a prefix of a right special word \( v \). Since \( X_n \) is not an \( F \)-maximal bifix code, there is a word \( w \) 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 the suffix order. It is not an \( F \)-maximal prefix code since \( vb \) is incomparable for the prefix order with the words of \( X_{n+1} \). Since it is finite, it is not an \( F \)-maximal bifix code. Moreover, since \( x_{n+1} \) has no suffix in \( X_n \), \( (L_{X_n}, x_{n+1}) \geq (L_{X_n}, u) + 1 \) by Equation (4.7). Thus

\[
(L_{X_{n+1}}, x_{n+1}) = (L_{X_n}, x_{n+1}) + 1 \geq (L_{X_n}, u) + 2 \geq n + 1
\]

by construction of \( x_{n+1} \). It is clear that \( X = \{ x_1, x_2, \ldots \} \) is a bifix code included in \( F \) which is not \( F \)-thin.
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. Let $x, y \in X$ be two left-special words. We may assume that $|x| \leq |y|$. 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 $X A^*$ 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 Proposition 4.5.3 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 $\text{Card}(X) = (k-1)d_F(X) + 1$.

The proof uses two lemmas.

Lemma 5.2.2 Let $F$ be a Sturmian set. Let $X \subset F$ be a finite $F$-maximal bifix code and let $P$ be the set of proper prefixes of the words of $X$. There exists a right-special word $u \in F$ such that $(L_X, u) = d_F(X)$. The $d_F(X)$ 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 $(L_X, u) = d_F(X)$.

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.

Since any right-special word contained in $P$ is a suffix of $u$, this proves the statement.

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

Lemma 5.2.3 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, $\text{Card}(X) = (k-1)\text{Card}(Q_X) + 1$. 


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 \( 1 \not\in Q_X \), 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 \( \text{Card}(Q_Y) = \text{Card}(Q_X) \). Hence, by induction hypothesis \( \text{Card}(X) = \text{Card}(Y) = (k-1) \text{Card}(Q_Y) + 1 = (k-1) \text{Card}(Q_X) + 1 \). Otherwise, \( X = \bigcup_{a \in \mathcal{A}} aX_a \). Set \( t_a = \text{Card}(Q_{X_a}) \). We have \( \sum_{a \in \mathcal{A}} t_a = \text{Card}(Q_X) \). By induction hypothesis, \( \text{Card}(X_a) = (k-1)t_a + 1 \). Therefore, \( \text{Card}(X) = \sum_{a \in \mathcal{A}} \text{Card}(X_a) = \sum_{a \in \mathcal{A}} (k-1)t_a + k = (k-1) \text{Card}(Q_X) + 1 \).

Proof of Theorem 5.2.1.

Let \( P \) be the set of proper prefixes of the words 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 5.2.2 and 5.2.3.

Example 5.2.4 Let \( F \) be the set of factors of the Fibonacci word. We have seen in Example 4.4.5 that there are 3 \( F \)-maximal bifix codes of \( F \)-degree 2. There are 13 \( F \)-maximal bifix codes of degree 3 listed below. We may illustrate the proof of Theorem 5.2.1 on this example. Consider the code \( X \) with kernel \( K = \{a, baab\} \) (see Table 5.1). There exactly three right-special words which are proper prefixes of words of \( X \), namely 1, \( ba \) and \( babaaba \) (indicated in black on Figure 5.1).

<table>
<thead>
<tr>
<th>code</th>
<th>kernel</th>
<th>derived code</th>
</tr>
</thead>
<tbody>
<tr>
<td>aab, aba, ba, bab</td>
<td>∅</td>
<td>aa, ab, ba</td>
</tr>
<tr>
<td>aa, aba, bab, baba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aaba, ab, baba</td>
<td>ab</td>
<td></td>
</tr>
<tr>
<td>aab, aba, abab, ba</td>
<td>ba</td>
<td></td>
</tr>
<tr>
<td>aa, ab, baaba, baba, bababab</td>
<td>aa, ab</td>
<td></td>
</tr>
<tr>
<td>aa, abab, abab, baba</td>
<td>aa, ba</td>
<td></td>
</tr>
<tr>
<td>aabaa, aababa, ab, ba</td>
<td>ab, ba</td>
<td></td>
</tr>
<tr>
<td>a, baabaab, baabab, bababab</td>
<td>a, baab, bab</td>
<td></td>
</tr>
<tr>
<td>a, baab, bababaabab, bababab</td>
<td>a, baab</td>
<td></td>
</tr>
<tr>
<td>a, baabaab, bababaab, babab</td>
<td>a, bab</td>
<td></td>
</tr>
<tr>
<td>aa, aba, abab, babababa</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>aa, ababa, ababa, bba, bababab</td>
<td>aa, ab, b</td>
<td></td>
</tr>
<tr>
<td>aabaa, aababa, aba, b</td>
<td>ab, b</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: The 13 maximal bifix codes of degree 3 in the factors of the Fibonacci word

5.3 Periodicity

Let \( x = a_0a_1 \cdots \), with \( a_i \in \mathcal{A} \), 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 there
Figure 5.1: The maximal bifix code of degree 3 with kernel \{a, baab\}

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 [21], Theorem 1.3.13).

**Theorem 5.3.1** Let \(x \in A^\mathbb{N}\) be an infinite word. If there exists an integer \(d \geq 1\) such that \(x\) has at most \(d\) factors of length \(d\) then \(x\) is ultimately periodic.

We prove the following generalization.

**Theorem 5.3.2** Let \(x \in A^\mathbb{N}\) be an infinite word and let \(F = F(x)\). If there exists a finite maximal bifix code \(X\) such that \(\text{Card}(X \cap F) \leq d(X)\), then \(x\) is ultimately periodic.

Theorem 5.3.2 implies Theorem 5.3.1 since \(A^d\) is a maximal bifix code of degree \(d\).

**Example 5.3.3** Let us consider again the finite maximal bifix code \(Z\) of degree 3 defined by \(Z = \{a^3, a^2ba, a^2b^2, ab, ba^2, bab, bab^2, b^2a, b^3\}\) (see Example 4.2.13). Assume that \(Z \cap F = \{ab, bab\}\), where \(F = F(x)\) and \(x \in A^\mathbb{N}\). Thus, \(x = ybabax'\), with \(y' \in A^*\). Therefore, the first letter in \(x'\) is \(b\) (otherwise, \(ba^2 \in Z \cap F\)) and the second letter in \(x'\) is \(a\) (otherwise, \(bab^2 \in Z \cap F\)). This argument shows that if \(yba\) is a prefix of \(x\) then \(ybab\) is also a prefix of \(x\), i.e., \(x = y(ba)^\omega\), with \(y \in A^*\).

The proof uses the Critical Factorization Theorem (see [20]) that we recall below. For a pair of words \((p, s)\), the set of words \(r\) such that

\[ A^* p \cap A^* r \neq \emptyset, \quad s A^* \cap r A^* \neq \emptyset. \]

is nonempty since it contains \(r = sp\). The repetition \(\text{rep}(p, s)\) is the minimal length of such a word \(r\).

Let \(w = a_1 a_2 \cdots a_m\) be a word with letters \(a_i \in A\). An integer \(n \geq 1\) is a period of \(w\) if for \(1 \leq i \leq j \leq m, j - i = n\) implies \(a_i = a_j\). A factorization of a word \(w \in A^*\) is a pair \((p, s)\) of words such that \(w = ps\).

**Theorem 5.3.4** (Critical Factorization Theorem) For any word \(w \in A^+\), the maximal value of \(\text{rep}(p, s)\) for all factorizations \((p, s)\) of \(w\) is the least period of \(w\).

We will also use the following lemma.
Lemma 5.3.5 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 \( i \geq 1 \), 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. \( \blacksquare \)

Proof of Theorem 5.3.2.
Let \( n \) be the maximal length of the words of \( X \). Let \( S = A^* \setminus A^*X \) and \( P = A^* \setminus XA^* \).

Let \( u \) be a prefix of \( x \) of length larger than \( n \) 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 \( n \).

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

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 \( p_i s_j \in X \). Indeed, there is a prefix \( s' \) of \( sv \) such that \( p_i s' \in X \). 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 5.2).

![Figure 5.2: The \( d+1 \) words \( x_0, x_1, \ldots, x_d \).](image)

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

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

If \( x_0 = x_i \) for an index \( i \) with \( 1 \leq i \leq d \), 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 \( X \)). Thus \( \text{rep}(q, r) \leq |s_i| \leq n \).

Set \( t = x_0 = x_i \) and \( x_0 = qp_t \). We have \( qt = ts_i \). Thus \( |s_i| = |q| \) is a period of the word \( qt \). Again \( \text{rep}(p, s) \leq |q| \leq n \). The case where \( x_i = x_1 \) for an index \( i \) with \( 1 \leq i \leq d \) 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 \( p_j = p_i t, t s_j = s_i \). 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 \(s_i\)).

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 5.3.5, it implies that \(y\) is periodic.

6 Basis of subgroups

In this Section, we push further the study of bifix codes in Sturmian sets. The main result of Section 6.1 is Theorem 6.1.1. It states that a finite \(F\)-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 positive bases of subgroups already considered in [28]. The second one uses the first return words which are considered in [36] and [19], up to a left-right symmetry (see also [1]).

6.1 Main result

We denote by \(A^\circ\) the free group generated by \(A\). The rank of \(A^\circ\) is \(\text{Card}(A)\).

We will prove the following result.

Theorem 6.1.1 Let \(F\) be a Sturmian set. A bifix code \(X \subset F\) is a basis of a subgroup of index \(d\) of \(A^\circ\) if and only if it is a finite \(F\)-maximal bifix code of \(F\)-degree \(d\).

Note that Theorem 6.1.1 implies Theorem 5.2.1. Indeed, by Schreier’s formula, if \(H\) is a subgroup of rank \(n\) and index \(d\) of a free group of rank \(k\), then

\[n - 1 = d(k - 1)\]  \hfill (6.1)

Let \(X\) be a finite \(F\)-maximal bifix code of degree \(d\). By Theorem 6.1.1, it is a basis of a subgroup of index \(d\) of the free group \(A^\circ\) which has rank 2. Thus \(\text{Card}(X) = (k - 1)d + 1\) by Schreier’s formula (6.1).

Before proving Theorem 6.1.1, we list some corollaries.

Recall that a group code of degree \(d\) is the minimal generating set of the submonoid \(\varphi^{-1}(H)\) where \(\varphi\) is a morphism from \(A^\circ\) onto a group \(G\) and \(H\) is a subgroup of index \(d\) of \(G\).

The following is a complement to Theorem 4.2.11.

Corollary 6.1.2 Let \(F\) be a Sturmian set and let \(Z\) be a group code of degree \(d\). Then \(Z \cap F\) is a finite \(F\)-maximal bifix code of \(F\)-degree \(d\).

Proof. Let \(K\) be the group generated by \(Z\). By Theorem 4.2.11, the set \(X = Z \cap F\) is an \(F\)-thin and \(F\)-maximal bifix code of \(F\)-degree \(e \leq d\). By Proposition 4.5.3, it is finite. By Theorem 6.1.1, it is a basis of a subgroup \(H\).
of index $e$. Since $X \subset Z$, we have $H \subset K$. Thus the index $d$ of $K$ divides the index $e$ of $H$. This implies $d = e$.

The next corollary implies the well-known fact that a subgroup of finite index of a free group has a positive basis.

A basis of a subgroup of $A^\circ$ contained in a Sturmian set $F$ is called an $F$-basis or a Sturmian basis.

**Corollary 6.1.3** Each subgroup of finite index of $A^\circ$ has, for each Sturmian set $F$, an $F$-basis. In particular, it has a Sturmian basis.

**Proof.** Let $K$ be a subgroup of index $d$ of $A^\circ$. Let $Z$ be the group code such that $Z^* = K \cap A^*$. Then $X = Z \cap F$ is a finite $F$-maximal bifix code of degree $d$. Thus $X$ is a basis of $K$. □

**Corollary 6.1.4** Let $F$ be a Sturmian set. The following conditions are equivalent for $X \subset F$.

(i) $X$ is a finite $F$-maximal bifix code of degree $d$.

(ii) $X$ is the intersection with $F$ of some group code of degree $d$.

As a further consequence of Theorem 6.1.1, we have the following result.

**Corollary 6.1.5** 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

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

The formula results directly from the formula, due to Hall [16], for the number of subgroups of index $d$ in a free group of rank $k$.

Observe that for $k = 2$, the first values are

<table>
<thead>
<tr>
<th>$d$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>13</td>
<td>71</td>
<td>461</td>
<td>3447</td>
<td>29093</td>
<td>273343</td>
</tr>
</tbody>
</table>

The values for $d = 2, 3$ are consistent with Examples 4.4.5 and 5.2.4. The values for $k = 2$ are given by the recurrence

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

It is known to enumerate also the indecomposable permutations on $d+1$ elements (see [13], [23] and [9]).
6.2 Preliminaries

Some preliminary results needed for the proof of Theorem 6.1.1.

**Proposition 6.2.1** Let $F$ be a Sturmian set and let $X \subset F$ be a finite $F$-maximal bifix code. Let $H$ be the subgroup generated by $X$. Then $H \cap F = X^* \cap F$.

The proof of Proposition 6.2.1 uses itself the following lemmas.

An automaton $A = (Q, 1, 1)$ is said to be bideterministic if it is deterministic and if for $p, q, r \in Q$ and $a \in A$, $r = p \cdot a = q \cdot a$ implies $p = q$ (the second condition expresses the fact that the reversal of the automaton is also deterministic). The following is from [28] (see also Exercise 6.1.2 in [5]).

**Lemma 6.2.2** Let $X \subset A^+$ be a bifix code and let $H$ be the subgroup of $A^\circ$ generated by $X$. The following conditions are equivalent.

(i) $X^* = H \cap A^*$

(ii) The minimal automaton of $X^*$ is bideterministic

The next lemma uses an argument similar to Lemma 5.2.3.

**Lemma 6.2.3** Let $v_1, v_2, \ldots, v_{n+1}$ be $n + 1$ words such that $v_i, v_{i+1}$ are not comparable for the prefix order for $1 \leq i \leq n$. Let $p_i$ be, for $1 \leq i \leq n$, the longest common prefix of $v_i, v_{i+1}$. If two of the $p_i$ are comparable for the prefix order, then two of the $p_i$ are equal.

**Proof.** Let $V = \{v_1, \ldots, v_{n+1}\}$ and let $W = V \setminus VA^+$. The set $W$ is a prefix code. Let $P$ be the set of proper prefixes of the words of $W$. Let $m$ be the number of distinct $p_i$. Since $v_i, v_{i+1}$ are not comparable for the prefix order for $1 \leq i \leq n$, for each $p_i$ there are at least two distinct letters $a, b$ such that $pa, pb \in P \cup W$. This implies $\text{Card}(W) \geq m + 1$. Thus $\text{Card}(W) < \text{Card}(V)$ implies $m \leq n - 1$.

**Lemma 6.2.4** Let $F$ be a Sturmian set and let $X \subset F$ be a bifix code. Let $n \geq 1$ and let $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_{n+1})$ be sequences of words of $F$ such that the $2n$ words $x_i$ defined by

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

for $1 \leq i \leq n$ are all in $X$ and such that $x_j \neq x_{j+1}$ for $1 \leq j < 2n$. Then $v_1$ and $v_{n+1}$ are incomparable for the prefix order.

**Proof.** We prove the property by induction on $n$. The property is true for $n = 1$ since $x_1, x_2$ are in $X$ and thus are not comparable for the prefix order.

Let $n \geq 2$. Let $p_i$ be, for $1 \leq i \leq n$, the longest common prefix of $v_i, v_{i+1}$. Since $x_{2i-1} \neq x_{2i}$, the words $v_i, v_{i+1}$ are incomparable for the prefix order. If
v_1 and v_n are comparable for the prefix order, by Lemma 6.2.3, 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 u_{i+1} = p_i v''_i. Since v_i, v_{i+1} are incomparable for the prefix order, the words v'_i, v''_i are nonempty. Since their longest common prefix is empty, their initial letters are distinct. Thus u_i p_i is right-special. Similarly u_j p_j is right-special. Thus u_i p_i and u_j p_j are comparable for the suffix order. Since p_i = p_j, u_i and u_j are comparable for the suffix order.

Since j - i \leq n - 1, we may apply the induction hypothesis to the reverse of X and the 2(j - i) words x_{2i-1}, ..., x_{2j-1} with the factorisations

\[ x_{2k} = v_{k+1} u_k, \quad x_{2k+1} = v_{k+1} u_{k+1}, \]

for i \leq k \leq j - 1. It implies that u_i and u_j are incomparable for the suffix order, a contradiction.

Lemma 6.2.4 has a dual formulation as follows. If (u_1, ..., u_{n+1}) and (v_1, ..., v_n) are sequences such that the 2n words

\[ x_{2i-1} = u_i v_i, \quad x_{2i} = u_{i+1} v_i, \]

for 1 \leq i \leq n are all in X and such that x_j \neq x_{j+1} for 1 \leq j < 2n, then u_1 and u_{n+1} are incomparable for the suffix order.

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. Consider the equivalence \( \theta_X \) 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^*.

One has p \equiv q \mod \theta_X if and only if there exist an integer n \geq 0, a sequence \((u_0, ..., u_n)\) with u_0 = p, u_n = q and a sequence \((v_1, ..., v_n)\) such that x_{2i-1} = u_{i-1} v_i and x_{2i} = u_i v_i are in X for 1 \leq i \leq n. Moreover, one may assume that x_j \neq x_{j+1} for 1 \leq j < 2n.

Indeed, we may assume that all u_i are distinct and that all v_i are distinct. But x_{2i-1} = x_{2i} implies u_{i-1} = u_i and x_{2i} = x_{2i+1} implies v_i = v_{i+1}.

We say that such sequences \((u_i)\) and \((v_i)\) are associated with the pair \((p, q)\).

Note that the class of 1 for the equivalence \( \theta_X \) is reduced to 1.

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

**Lemma 6.2.5** For p, q \in P and a \in A, if p \equiv q \mod \theta_X and p \cdot a, q \cdot a \neq \emptyset then p \cdot a \equiv q \cdot a \mod \theta_X.

**Proof.** Let p, q \in P and a \in A be such that p \equiv q \mod \theta_X and p \cdot a, q \cdot a \neq \emptyset.

Let \((u_0, ..., u_n)\) and \((v_1, ..., v_n)\) be sequences associated with the pair \((p, q)\).

Since p \cdot a, q \cdot a \neq \emptyset there exist v, w such that pav, qaw \in X.

We distinguish several cases.

Case 1: v_1 and v_0 begin with a.

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 $pa \equiv qa \mod \theta_X$. Otherwise, let $i$ be minimal such that $v_i$ begins with a letter distinct of $a$ and $j$ be maximal such that $v_j$ begins with a letter distinct of $a$. Then $u_{i-1}$ and $u_j$ are right-special. But the dual of Lemma 6.2.4, applied to the sequences $(u_{i-1}, \ldots, u_j)$ and $(v_1, \ldots, v_j)$ implies that $u_{i-1}$ and $u_j$ are incomparable for the suffix order, a contradiction.

Next, suppose that $pa \in X$ and thus that $v_1 = a$ (since $v_1$ begins with $a$ and $X$ is prefix). If $v_n \neq aw$, then Lemma 6.2.4, applied to the sequences $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n, aw)$, implies that $v_1 = a$ and $aw$ are not comparable for the prefix order, a contradiction. Thus $v_n = aw$. If $n \geq 2$, Lemma 6.2.4, applied to the sequences $(u_1, \ldots, u_{n-1})$ and $(v_1, \ldots, v_n)$ implies that $v_1 = a$ is not comparable for the prefix order with $v_n = aw$, a contradiction. Finally, if $v_n = aw$ and $n = 1$, then $v_1 = aw$, thus $w = 1$, $qa \in X$ and therefore $p \cdot a = 1 = q \cdot a$.

Case 2: $v_1$ begins with $b$. Then, since $u_0v_1$ is in $X$ and $u_0av = pav \in X$, $u_0$ 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. Thus $u_0$ and $u_i$ are comparable for the suffix order, a contradiction with the dual of Lemma 6.2.4, applied to the sequences $(u_0, \ldots, u_i)$ and $(v_1, \ldots, v_i)$. If $i = n$, then $u_0$ and $u_n$ are right-special since $u_nv_n \in X$ and $u_naw = qaw \in X$. We obtain a contradiction with the dual of Lemma 6.2.4 applied to the sequences $(u_0, \ldots, u_n)$ and $(av, v_1, \ldots, v_n, aw)$ (we may apply the lemma since $u_0av \neq u_0v_1$ because $v_1$ begins with $b$, and $u_n v_n \neq u_n aw$ for a similar reason). All this shows that Case 2 does not happen.

Let $R$ be the set of classes of $\theta_X$ with the class of 1 still denoted 1. Let $B = (R, 1, 1)$ be the automaton with set of states $R$ and transitions induced by the transitions of $A$.

**Lemma 6.2.6** The automaton $B$ is bideterministic.

**Proof.** Let $r, s \in R$ and $a \in A$ be such that $r \cdot a = s \cdot a$. Let $p, q \in P$ be representatives of the classes $r$ and $s$ respectively. It is enough to show that $p \equiv q \mod \theta_X$.

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

Suppose next that $pa, qa \in P$. Let $(u_0, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ be sequences associated with the pair $(pa, qa)$.

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

Otherwise, let $i$ be minimal such that $u_i$ ends with a letter distinct of $a$ and $j$ be maximal such that $u_j$ ends with a letter distinct of $a$. Then $v_i$ and $v_{j+1}$ are left-special and thus comparable for the prefix order. This contradicts Lemma 6.2.4 applied to the sequences $(u_1, \ldots, u_j)$ and $(v_1, \ldots, v_{j+1})$.

**Proof of Proposition 6.2.1.**

We have $X^* \cap F \subset H \cap F$.

To show the converse inclusion, consider the bifix code $Z$ such that $Z^* = \text{Stab}_B(1)$.
Let us show that $Z \cap F = X$. If $x$ is in $X$, then there is a path in the automaton $B$ from 1 to 1 labeled $x$. It does not pass by 1 except at its ends since the class of 1 modulo $\theta_X$ is reduced to 1. Thus $x$ is in $Z$. Conversely, since $X$ is an $F$-maximal prefix code, each $z \in Z \cap F$ is comparable for the prefix order with some $x \in X$. As we saw before $x$ is in $Z$ and thus $x = z$ because $Z$ is prefix.

Since the automaton $B$ is bideterministic by Lemma 6.2.6, it is equal to the minimal automaton of $Z^*$. Let $K$ be the subgroup generated by $Z$. By Lemma 6.2.2, we have $K \cap A^* = Z^*$. This shows that $H \cap F \subset K \cap F \subset Z^* \cap F \subset X^* \cap F$.

The first inclusion holds because $X \subset Z$ implies $H \subset K$. The last one 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 $H \cap F \subset X^* \cap F$, which was to be proved.

Let $F$ be a factorial set. For $u \in F$, define
$$\Gamma_F(u) = \{z \in F \mid uz \in A^* u \cap F\}$$
and
$$R_F(u) = \Gamma_F(u) \setminus \Gamma_F(u) A^+.$$ 
Thus $\Gamma_F(u) = R_F(u)^*$. When $F = F(x)$ for an infinite word $x$, the sets $\Gamma_F(u)$ and $R_F(u)$ are respectively the set of return words to $u$ and first return words to $u$ in $x$. Vuillon has shown in [36] 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 [19]. Note that their definition of return words is a left-right symmetrical of ours. This means that they are conjugated by a fixed word to ours.

In fact, they show much more in [19].

**Theorem 6.2.7** Let $F$ be a Sturmian set. For a word $u \in F$, the set $R_F(u)$ is a basis of the free group $A^\circ$.

Since this result is not clearly formulated as above in [19], we prove below how it may be easily deduced from their article.

**Proof of Theorem 6.2.7.** By Definition 7 of [14], we may assume that $F = F(s)$ for some standard and strict episturmian word $s$. Now, by Theorem 4.4 of [19], for the $n + 1$-th palindromic prefix $u_{n+1}$ of $s$, its set of first return words is the set of $\mu(x)$, for $x$ any $i$-th letter of the directive sequence of $s$, $i \geq n$, and where $\mu$ is a product of elementary morphisms $\psi_a$ (see page 6 in [19]). Note that such a $\mu$ is therefore an automorphism of the free group. Since $s$ is strict, any letter of the alphabet of $s$ appears infinitely often in its directive sequence (see Definition 2.3 in [18]). Hence $x$ above may be any letter of the alphabet $A$.

We deduce that the set of first returns of $u_{n+1}$ is a basis of the free group on $A$.

Suppose now that $w$ is any factor of $s$. Then by Corollary 4.1 of [19], there exist $n$ and a word $f$ such that the first return words to $w$ are of the form $fyf^{-1},$
with \( y \) any first return of \( u_n \). Thus the set of first returns of \( w \) is a basis, too. This ends the proof.

6.3 Proof of the main result

Proof of Theorem 6.1.1. Assume first that \( X \) is a finite \( F \)-maximal bifix code of \( F \)-degree \( d \). Let \( P \) be the set of proper prefixes of the words of \( X \). Let \( Q \) be the set of words in \( P \) which are right-special. Let \( H \) be the group generated by \( X \).

For \( p,q \in Q \), \( Hp = Hq \) implies \( p = q \). Suppose indeed that \( Hp = Hq \). Then \( Huq = Hq \) implies \( Hu = H \) and thus \( u \in H \). By Proposition 6.2.1, since \( u \in F \), this implies that \( u \in X^* \) and thus \( u = 1 \) since \( p \) is a proper prefix of a word of \( X \).

By Lemma 5.2.2 there is a right-special word \( u \) such that \((L_X, u) = d \). The \( d \) suffixes of \( u \) which are in \( P \) are the elements of \( Q \).

Let \( \mathcal{A} = (P, 1, 1) \) be the literal automaton of \( X^* \) (see Section 3.2). Recall from Section 3.2 that \( \varphi_A \) is the morphism from \( A^* \) onto the transition monoid of \( \mathcal{A} \). Set \( \varphi = \varphi_A \). For \( p,q \in P \), and \( w \in F \), one has \( pw \in X^*q \).

For \( y \in R_F(u) \), the restriction of \( \varphi(y) \) to \( Q \) is a permutation of \( Q \). Indeed, let \( q \in Q \). Since \( q \) is a suffix of \( u \), \( qy \) is a suffix of \( uy \) and thus \( qy \) is in \( F \). Thus there is a word \( r \in P \) such that \( qy \in X^*r \). Since \( y \in R_F(u) \), the word \( r \) is a suffix of \( u \) and thus we have \( r \in Q \) (see Figure 6.1). Each element of \( Q \) is obtained exactly once in this way. This shows that the restriction of \( \varphi(y) \) to \( Q \) is a permutation.

![Figure 6.1: A word \( y \in R_F(u) \).](image)

Since, by Theorem 6.2.7, \( R_F(u) \) is a basis of the free group \( A^\circ \), the group generated by \( R_F(u) \) is \( A^\circ \). Let

\[
V = \{ v \in A^\circ \mid Qv \subset HQ \}
\]

Any \( v \in V \) defines a permutation of \( Q \). Indeed, suppose that for \( p,q,r \in Q \), one has \( pv, qv \in Hr \). Then \( rv^{-1} \) is in \( Hp \cap Hq \). This forces \( p = q \) as we have seen above.

The set \( V \) is a subgroup of \( A^\circ \). 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 \( Qvw \subset Qw \subset Q \) and thus \( vw \in V \).

Since \( R_F(u) \subset V \), and since \( V \) is a subgroup of \( A^\circ \), 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 $\text{Card}(Q) = d$, and since the cosets $Hq$ for $q \in Q$ are distinct, this shows that $H$ is a subgroup of index $d$.

Assume finally that $X \subset F$ is a bifix code such that the group $H$ generated by $X$ has index $d$. Let $Y \subset A^+$ be the bifix code such that $H \cap A^* = Y^*$. Then $Y$ is a thin maximal bifix code of degree $d$. By Theorem 4.2.11, the set $X' = Y \cap F$ is an $F$-thin and $F$-maximal bifix code. By the preceding argument, the group $H'$ generated by $X'$ is a subgroup of finite index $d'$. By Theorem 4.2.11, we have $d' \leq d$. Since $X' \subset Y$, we have $H' \subset H$. Thus the index of $H$ divides the index of $H'$, which implies $d \leq d'$. Thus $d = d'$ and $H = H'$, which implies that $X = X'$.

**Example 6.3.1** Let $F$ be the set of factors of the Fibonacci word. The set that $R_F(u)$ is shown below for the first elements of $F$.

<table>
<thead>
<tr>
<th>$u$</th>
<th>$a$</th>
<th>$b$</th>
<th>$aa$</th>
<th>$ab$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_F(u)$</td>
<td>$a, ba$</td>
<td>$ab, aab$</td>
<td>$baa, babaa$</td>
<td>$ab, aab$</td>
</tr>
</tbody>
</table>

**Example 6.3.2** Let $F$ be the set of factors of the Fibonacci word. Let $X$ be the bifix code shown on Figure 6.2. The right-special proper prefixes of the words of $X$ are indicated in black. The representation of $A^\circ$ on the cosets of the group generated by $X$ is shown on Figure 6.3.

![Figure 6.2: An $F$-maximal bifix code of $F$-degree 4.](image)

Figure 6.2: An $F$-maximal bifix code of $F$-degree 4.

![Figure 6.3: The group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.](image)

Figure 6.3: The group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

We end this section with a combinatorial consequence of Theorem 6.1.1.
Proposition 6.3.3 Let F be a Sturmian set on an alphabet with k letters and let X ⊂ F be a finite F-maximal bifix code. Let P (resp. S) be the set of proper prefixes (resp. suffixes) of the words of X. Then
\[ \sum_{x \in X} |x| = \text{Card}(P) + \text{Card}(S) + (k - 2)d. \]

We will use the following lemma. Let X be a bifix code and let P be the set of proper prefixes of the words of X. We consider again the equivalence \( \theta_X \) on P which is the transitive closure of the relation on P formed by the pairs \((p, q)\) such that \( ps, qs \in X \) for some word s (this is the same equivalence which is used in the proof of Proposition 6.2.1).

Lemma 6.3.4 Let F be a Sturmian set and let X ⊂ F be a finite F-maximal bifix code. Let P be the set of proper prefixes of the words of X. If X generates a subgroup of index d of \( A^* \), then \( \theta_X \) has d classes.

Proof. Let H be the group generated by X. We show that \( p \equiv q \mod \theta_X \) if and only if \( Hp = Hq \). It is clear that \( p \equiv q \mod \theta_X \) implies \( Hp = Hq \).

Since the d right-special words which are in P are in distinct cosets of H, the number of classes of \( \theta \) is at least equal to d.

Let \( A = (P, 1, 1) \) be the literal automaton of \( X^* \) and let \( B = (R, 1, 1) \) be the automaton on the set of classes of \( \theta_X \) as in Lemma 6.2.5. By Lemma 6.2.6, the automaton B is bideterministic. Let \( C \) be a group automaton on R with transitions extending those of B. Let T be the bifix code such that \( T^* = \text{Stab}_C(1) \). Then T has degree \( \text{Card}(R) \). Let Z be the bifix code such that \( Z^* = H \cap A^* \). Then X ⊂ T implies that Z ⊂ T^* and thus that the degree of T divides the degree d of Z. In particular, we have \( d(T) \leq d \). This shows that \( \text{Card}(R) = d \).

Proof of Proposition 6.3.3. Let H be the group generated by X. By Theorem 6.1.1, the index of H is equal to \( d = d_F(X) \). Let \( P' = P \setminus 1 \) and \( S' = S \setminus 1 \). Let \( E = \{(p, s) \in P' \times S' \mid ps \in X\} \). One has
\[ \text{Card}(E) = \sum_{x \in X} (|x| - 1) = \sum_{x \in X} |x| - \text{Card}(X) = \sum_{x \in X} |x| - (k - 1)d - 1. \]

Let G be the nonoriented graph on the set of vertices \( P' \cup S' \) with E as set of edges. Let \( G = \bigcup_{i \in I} G_i \) be the partition of G in connected components. Let \( P = \bigcup_{i \in I} P_i \) and \( S = \bigcup_{i \in I} S_i \) be the partitions of P and S such that \( P_i \cup S_i \) is the set of vertices of \( G_i \).

Since X is a basis of H, the partition \( P = \bigcup_{i \in I} P_i \) coincides with the decomposition of P in classes of \( \theta_X \). Thus \( \text{Card}(I) = d - 1 \) by Lemma 6.3.4.

Each \( G_i \) is a tree. Indeed, assume that \((p_1, s_1, \ldots, p_n, s_n)\) is a simple cycle of \( G_i \). Then \( p_1s_1, p_2s_1, \ldots, p_ns_n, p_1s_n \) are in X. But \( p_1s_n = (p_1s_1)(p_2s_1)^{-1} \cdots (p_ns_n) \) in contradiction with the fact that X is a basis of the subgroup H, by Theorem 6.1.1.
Let \( E_i \) be the set of edges of \( G_i \). Since \( G_i \) is a tree, we have \( \text{Card}(E_i) = \text{Card}(P_i) + \text{Card}(S_i) - 1 \). Finally

\[
\text{Card}(E) = \sum_{i \in I} \text{Card}(E_i) = \sum_{i \in I} (\text{Card}(P_i) + \text{Card}(S_i) - 1) = \text{Card}(P) + \text{Card}(S) - d - 1.
\]

whence the result.

\[\square\]

7 Syntactic groups

In Section 7, we use Theorem 6.1.1 to prove a new result on the syntactic groups of prefix codes.

7.1 Preliminaries

We first recall the basic terminology on groups in monoids (see [5] for a more detailed exposition). We then prove our main result which states that any transitive permutation group \( G \) of degree \( d \) and rank 2 is a syntactic group of a bifix code with \( d + 1 \) elements (Theorem 7.3.2).

Let \( M \) be a monoid of maps from a set \( Q \) to itself. A group contained in \( M \) is a subsemigroup of \( M \) which is isomorphic to a group. Note that the neutral element of a group contained in \( M \) need not be equal to the neutral element of \( M \).

A group \( G \) contained in \( M \) is maximal if it not included in another group \( H \) contained in \( M \).

**Proposition 7.1.1** Let \( G \) be a group contained in a monoid \( M \) of maps from a set \( Q \) to itself. All elements of \( G \) have the same image \( I \). The restriction 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 \( I \). \[\square\]

We often identify the group \( G \) with its representation by permutations on \( I \).

A syntactic group of a prefix code \( X \) is a maximal group in the monoid of transitions of \( \mathcal{A}(X^*) \).

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

The following result is from [25].

**Theorem 7.1.2** Let $G$ be a permutation group of degree $d$. If $G$ is a nonspecial syntactic group of a prefix code $X$, then $\text{Card}(X) \geq d + 1$.

Theorem 7.1.2 was proved before in a weaker form ($\text{Card}(X) \geq d$) but with a more general hypothesis (with a set $X$ of words instead of a prefix code). The general idea is that some parameters in the transition monoid of the minimal automaton of $X^*$ can be bounded in terms of $\text{Card}(X)$ only, instead of the sum of the lengths of the words of $X$. The proof uses the Critical Factorization Theorem (see [25] for a bibliography on this problem).

Theorem 7.1.2 is clearly not true for special syntactic groups since $\mathbb{Z}/n\mathbb{Z}$ is a syntactic group of $X = a^n$ for any $n \geq 1$.

### 7.2 Group of a bifix code

Let us recall the notation concerning Green relations in a monoid $M$. (see [5]).

We denote by $R$ the equivalence in $M$ defined by $mRn$ if $m,n$ generate the same right ideal, i.e. if $mM = nM$. We denote by $R(m)$ the $R$-class of $m$.

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

It is a classical result that the equivalences $L$ and $R$ commute. We denote by $D$ the equivalence $LR = RL$. Finally, we denote by $H$ the equivalence $L \cap R$.

A $D$-class is regular if it contains an idempotent. In this case, there is at least an idempotent in each $L$ class and each $R$-class. By Green’s lemma, for $m, n \in M$, $R(m) \cap L(n) \neq \emptyset$ if and only if $R(n) \cap L(m)$ contains an idempotent.

The $H$ 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 $D$-class $D$ are equivalent. The structure group of $D$ is any one of them.

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

**Proposition 7.2.1** The set of elements of $\varphi(F)$ of rank $d$ is a regular $D$-class of $M$.

We use the following lemma. For a word $w$, the rank of $w$ is the rank of $\varphi(w)$. We denote $\text{Im}(w) = \{ q \in Q \mid p \cdot w = q \text{ for some } p \in Q \}$ the image of $w$.

**Lemma 7.2.2** Let $u \in F$ be of rank $d$. If $v \in F$ is such that $uv \in F$, then $uv$ has rank $d$. 

47
Proof. Since $X$ is $F$-thin, there exists $w \in F$ which is not a factor of a word in $X$. Then $w$ has $d$ interpretations. Since $F$ is recurrent, there exists a $t$ such that $uwv \in F$. Then $uwv$ has $d$ interpretations. This implies that the rank of $uwv$ is $d$. Thus the rank of $uwv$ cannot be less than $d$. ■

Proof of Proposition 7.2.1 Let $u, v \in F$ be two words of rank $d$. Let $w$ be such that $uwv \in F$ and let $t$ be such that $uwvtu \in F$. By Lemma 7.2.2, the rank of $uwvtu$ is $d$. Since $\text{Im}(uwvtu) \subset \text{Im}(u)$, this implies that the restriction of $\varphi(uwvtu)$ to $\text{Im}(u)$ is a permutation. Since $\text{Im}(u)$ is finite, there is an integer $n \geq 1$ such that $(uwvtu)^n$ is the identity on $\text{Im}(u)$. Set $e = \varphi(uwvtu)^n$ and $s = tu(uwvtu)^{n-1}$. Then, since $e$ is the identity on $\text{Im}(u)$, $\varphi(u) = \varphi(ue)$. Thus $\varphi(u) = \varphi(uw)v\varphi(s)$. This shows that $u$ and $uwv$ are $\mathcal{R}$-equivalent. Similarly, $v$ and $uwv$ are $\mathcal{L}$ equivalent. Thus $u, v$ are $\mathcal{D}$-equivalent. ■

The structure group of the $\mathcal{D}$-class of elements of rank $d$ in $\varphi(F)$ is a transitive group of degree $d$. By Proposition 9.5.1 in [5], this group does not depend on the choice of the simple automaton $A$ 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 of the code $X$ defined in [5]. Indeed, in this case, the $\mathcal{D}$-class of elements of rank $d$ coincides with the minimal ideal of the monoid $\varphi(A^*)$.

7.3 Main result

We will use Theorem 6.1.1 to prove the following result.

**Theorem 7.3.1** Let $Z \subset A^*$ be a thin maximal bifix code of degree $d$. Let $F$ be Sturmian set. The set $X = Z \cap F$ is an $F$-maximal bifix code of degree $d$ and $G_F(X) = G(Z)$ are identical.

**Proof.** Let $P, Q$ be the sets of proper prefixes of the words of $X$ and $Z$ respectively. Let $A = (P, 1, 1)$ and $B = (Q, 1, 1)$ be the literal automata of $X^*$ and $Z^*$ respectively. Let $u \in F$ be a word which is not a factor of a word in $X$ (and therefore not a factor of a word in $Z$ either). Then $u$ has $d$ interpretations with respect to $X$, which are also its $d$ interpretations with respect to $Z$. Thus $u$ has rank $d$ with respect to both automata. Let $I$ be the set of suffixes of $u$ which are in $P$ (and also in $Q$). Then any word $v$ such that $uv \in F$ defines an permutation on $I$. The set of these permutations is the group $G_F(X)$ and is included in $G(Z)$. Thus $G_F(X) \subset G(Z)$.

Let us show the converse inclusion. Let $Y = R_F(u)$ be the set of first returns to $u$. By Theorem 6.2.7, the set $Y$ is a basis of $A^*$. For any $y \in Y$, we have $uy \in F \cap A^*u$. Thus $y$ is a permutation on $I$ which belongs to $G_F(X)$.

Let $w \in A^*$ be such that $w$ is a permutation $g$ on $I$ with respect to the action in the automaton $B$. Let $y_1, \ldots, y_n \in Y$ be such that $w = y_1^{g_1} \cdots y_n^{g_n}$ with $\varepsilon_1 = \pm 1$. Then $g$ is the product of the permutations $g_i^\varepsilon$, where $g_i$ is the restriction of $\varphi_A(y)$ to $I$. Since each $g_i$ is in $G_F(X)$, this proves that $G(Z) \subset G_F(X)$. ■
The rank of a group is the minimal cardinality of a set generating $G$. We use Theorem 7.3.1 to prove the following result.

**Theorem 7.3.2** Any transitive permutation group of degree $d$ and rank $k$ is a syntactic group of a bifix code with $(k - 1)d + 1$ elements.

**Proof.** Let $Z$ be a group code on an alphabet 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 7.3.1, $G_F(X) = G$. □

Theorem 7.3.2 was known before only in particular cases. In [24] it is shown for the case of a group generated by a $d$-cycle and another permutation. In [31], it is proved that for an Abelian group of rank 2 and order $d$ there exists a bifix code $X$ such that $\text{Card}(X) - 1 = d$. The proof is based on the fact that the Cayley graph of an Abelian group contains a Hamiltonian cycle. Curiously, in the case of the group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, the result is the Sturmian basis of Example 6.3.2.

**Example 7.3.3** We consider again the code of Example 6.3.2. The minimal automaton of $X^*$ is represented on Figure 7.1. The action on the sets of states with four elements is shown on Figure 7.2. The set $\{1, 2, 4, 8\}$ corresponds to the states reached by proper prefixes which are right-special. The set of first returns to this set of states is $\{ba, aba\}$ which is just $R_F(aba)$ in agreement with the fact that $aba$ is the longest proper prefix which is right special. The word $ba$ defines the permutation $(18)(24)$ and the word $aba$ the permutation $(14)(28)$. 49
Acknowledgments  We wish to thank Mike Boyle, Aldo De Luca, Thierry Monteil, Patrice Séébold, Martine Queffélec and Gwenael Richomme for their help in the preparation of this manuscript.

References


