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
August 15, 2010

Contents

1 Introduction 2

2 Factorial sets 4
  2.1 Recurrent sets ........................................ 4
  2.2 Recurrent words ...................................... 5
  2.3 Probability distributions ............................ 6

3 Prefix codes in factorial sets 10
  3.1 Prefix codes ........................................... 10
  3.2 Maximal prefix codes .................................. 11
  3.3 Average length ....................................... 13

4 Bifix codes in recurrent sets 14
  4.1 Indicator ............................................... 14
  4.2 Maximal bifix codes .................................. 16
  4.3 Degree ................................................. 21
  4.4 Kernel ................................................... 25
  4.5 Finite maximal bifix codes ......................... 27

5 Bifix codes in Sturmian sets 27
  5.1 Sturmian sets ......................................... 28
  5.2 Cardinality ............................................ 29
  5.3 Periodicity ............................................ 30

6 Basis of subgroups 32
  6.1 Sturmian basis ........................................ 32
  6.2 Syntactic groups ..................................... 38
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. 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 has $d+1$ elements. This generalizes the fact that a Sturmian set contains $d+1$ 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

The study of bifix codes goes back to founding papers by Schützenberger [21] and by Gilbert and Moore [9]. These papers already contain significant results. The first systematic study is in the papers of Schützenberger [22],[23]. 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 [6]. He also defined the notion of derived code which allows to build maximal bifix codes by increasing degrees [7]. This study lead to the Critical Factorization Theorem that we will meet again here.

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. We extend most properties of ordinary maximal bifix codes to bifix codes maximal in a recurrent set of words. We show in particular that the average length of a finite 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 4.3.1). 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 $n$. Our main result is that a maximal bifix code of degree $d$ in the set of factors of a Sturmian word is always a basis of a subgroup of index $d$ of the free group (Theorem 6.1.1). In particular, it has $d+1$ 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 6.2.1). This aspect was first considered by Schützenberger in [20]. 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.

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 [18] 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 [8] where the internal transformation is generalized. In [17], 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 [10]. A more general notion was considered in [1]. 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 [2]. Codes in sets of factors of Sturmian words have been studied in [5]. Finally, bifix codes $X$ contained in restricted sets of words are used in [15] 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 4.5.1). In the case of Sturmian sets, we prove our main results. First, a bifix code of degree $d$ maximal in a Sturmian set has $d + 1$ 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[12]).

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 an $F$-thin 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 6.2 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 [12] for example).

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

A set \( F \) is said to be 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 \( uv \in F \).

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 \( uaw \in F \).

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

Proposition 2.1.1 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 2.1.1 is not true as shown in the example below.

Example 2.1.3 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 \( x \in A^\infty \) 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 \( X \subset A^\ast \). A (one-sided) shift space is a set of infinite words of the form \( S_X \) for some \( X \subset A^\ast \).

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

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 \( uuv \in F \). Thus \( x \) is recurrent. Conversely, assume that \( x \) is recurrent. Let \( u, v \) be in \( F \). Then there is a factorization \( x = puv \) with \( p \in F \) and \( y \in A^\infty \). Since \( x \) is recurrent, the word \( u \) is a factor of \( y \). Set \( y = qez \) with \( q \in F \) and \( z \in A^\infty \). Then \( uqv \) is in \( F \). Thus \( F \) is recurrent.

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

**Example 2.2.1** Given a sequence \( (u_n)_{n \geq 0} \) of nonempty words, let \( (w_n)_{n \geq 0} \) be the sequence of words defined by \( u_0 = u_0 \) and \( w_n = w_{n-1}u_nw_{n-1} \). The infinite word \( x \) obtained as the limit of the sequence \( (w_n)_{n \geq 0} \) is always recurrent. Indeed, we know that for any \( n \geq 0 \), there is \( y \in A^\infty \) such that \( x = w_ny \). Then, for any \( u \in F(x) \), there exists \( n \in \mathbb{N} \) such that \( u \in F(w_n) \), that is \( w_n = uvu' \). Consequently, \( x = w_{n+1}y' = w_nu_{n+1}w_ny' = uvu'_{n+1}vuv'y' \) and \( uv'_{n+1}vuv'y' \in F(x) \). However, let \( x \) be obtained as above with \( u_n = ab^n \). Then \( x \) is not uniformly recurrent. Indeed, \( ab \in F(x) \) and if \( x \) were uniformly recurrent, there should be \( n \) such that \( ab \) is factor of every word in \( F(x) \cap A^n \). Since \( x = w_ny = w_{n-1}abnw_{n-1}y \), we have \( b^n \in F(x) \cap A^n \), so \( ab \in F(b^n) \), a contradiction.
We use indifferently the terms of *morphism* or *substitution* for a monoid morphism from $A^*$ into itself.

**Example 2.2.2** Set $A = \{a, b\}$. The *Thue-Morse morphism* is the substitution $f : A^* \to A^*$ defined by $f(a) = ab$ and $f(b) = ba$. The *Thue-Morse word* $x = abbabaab\cdot\cdot\cdot$ is the fixpoint $f^\omega(a)$ of $f$. It is uniformly recurrent (see [13] Example 1.5.10).

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

The following property is classical (see for example [13] Theorem 1.5.9).

**Proposition 2.2.3** An infinite word $x \in A^\mathbb{N}$ 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 $n \geq 1$.

**Example 2.2.3** 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* $x = abaababa\cdot\cdot\cdot$ is the fixpoint $f^\omega(a)$ of $f$. It is a Sturmian word (see [13] Example 2.1.1).

The following is Proposition 2.1.25 in [13].

**Proposition 2.2.4** If $x$ is Sturmian, then $S(x)$ is minimal and $x$ is uniformly recurrent.

The converse is false as shown by the following example.

**Example 2.2.4** The Thue-Morse word of Example 2.2.2 is not Sturmian. Indeed, it has four factors of length 2.

### 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 *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 probability distribution $\pi$ on $F$ and a set $X \subset F$, we denote $\pi(X) = \sum_{x \in X} \pi(x)$. See [3] for the elementary properties of probability distributions. Note in particular that for any $u \in F$ and $n \geq 0$, one has as a consequence of condition (ii)

$$\pi(ua^n \cap F) = \pi(u). \quad (2.1)$$

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

When $F$ is factorial, the distribution is said to be *invariant* if additionally
\( \sum_{a \in P(w)} \pi(aw) = \pi(w) \), for any \( w \in F \),

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

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

**Proposition 2.3.1** For any right essential set \( F \) of words, there exists a positive 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 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 \( \pi(aw) = 1/d_0d_1 \cdots d_n \). Since \( \text{Card}(S(w)) = d_n \), we obtain that \( \pi \) satisfies condition (ii) and thus it is a 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 \( A^N \) 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 [13]).

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

**Proof** It is clear that a shift space is both closed and invariant. Conversely, let \( S \subset A^N \) 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 \( F \) of Borel subsets of \( S \) is invariant if \( \mu(\sigma^{-1}B) = \mu(B) \) for any \( B \in F \).

The following result is from [16] (Krylov and Bogolioubov’s Theorem 4.2).
Theorem 2.3.1 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 2.2.3, the shift space \( S = S(x) \) is minimal.

By Theorem 2.3.1 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 \( U = \cup_{n \in \mathbb{Z}} \sigma^n(T) \) 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 \( wA^\infty \cap S \) is open, we have shown in particular that \( \mu(wA^\infty \cap S) > 0 \).

Let \( \pi \) be the map from \( F \) to [0, 1] defined by \( \pi(w) = \mu(wA^\infty \cap S) \). It is easy to verify that \( \pi \) is an invariant probability distribution which is positive. Indeed, one has \( \pi(1) = \mu(S) = 1 \). Next, for \( w \in F \)

\[
\sum_{a \in S(w)} \pi(aw) = \sum_{a \in S(w)} \mu(awA^\infty \cap S) = \mu(wA^\infty \cap S) = \pi(w).
\]

In the same way

\[
\sum_{a \in P(w)} \pi(aw) = \sum_{a \in P(w)} \mu(awA^\infty \cap S) = \mu(\sigma^{-1}(wA^\infty \cap S)) = \mu(wA^\infty \cap S) = \pi(w).
\]

Thus we have proved the following result.

Corollary 2.3.1 For any uniformly recurrent set \( F \subset A^* \), there exists positive invariant probability distributions on \( F \).

Corollary 2.3.1 is not true in general for a recurrent set (see [4]).

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 on the set \( F(x) \) ([16], Theorem 5.6). We illustrate this result by the following examples.

Example 2.3.1 Let \( x = abababaaba \cdots \) 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 [16]). 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

\[
M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}
\]
Figure 2.1: The invariant probability distribution on the factors of the Fibonacci word

This implies \( v = [\lambda \quad 1 - \lambda] \). The other values can be computed using conditions (ii) and (iii) of the definition of an invariant probability distribution.

**Example 2.3.2** 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.

Figure 2.2: The invariant probability distribution on the factors of the Thue-Morse word
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 3.2.4).

3.1 Prefix codes

Let $F$ be a prefix-closed set. We consider on $F$ the prefix order defined for $u, v \in F$ by $u \leq v$ if $u$ is a prefix of $v$. A set $X \subseteq F$ 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 on a suffix-closed set of words 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 $\underline{S}$ the characteristic $F$-series of $S$. By definition, for any $w \in F$,

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

Note that $S$ need not be included in $F$ but that $\underline{S} = \underline{S \cap F}$.

Note also that $\underline{A}$ is the inverse of $1 - \underline{A}$. More generally, for any prefix code, $\underline{X}^*$ is the inverse of $1 - \underline{X}$. The following is adapted from Proposition 3.1.6 in [3].

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

$$ F \subset X^*U. \quad (3.1) $$

If $X$ is a prefix code, then

$$ \underline{X} - 1 = U(A - 1) \quad \text{and} \quad \underline{F} = \underline{X}^*U \quad (3.2) $$
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, for any \( u \in U \) and \( a \in A \), one has either \( ua \in U \) or \( ua \in X \) but not both; Conversely, any nonempty element of \( U \cup X \) is of this form. Thus \( u(U) + u(X) = U + \frac{1}{A} + 1 \). This shows the first equality in (3.2). The second one is a consequence of the first one using \( u(X)^*(1 - u(X)) = 1 \) and \( u(F)(1 - u(A)) = 1 \).

### 3.2 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 3.2.1** Let \( F \) be a prefix closed set. 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 \( uv = xw \) 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 \( F \subset A^* \) be a prefix-closed set. 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 [3].

**Proposition 3.2.2** 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 [3].

**Proposition 3.2.3** 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 $XA^*$. Then $X \cup u$ is prefix, a contradiction.

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

**Example 3.2.1** The set $X = \{a, ba\}$ is a maximal prefix code in the set $F$ of factors of the Fibonacci word since $XA^*$ is right $F$-dense.

The following is a generalization of Propositions 3.7.1 and 3.7.2 in [3].

**Proposition 3.2.4** Let $F$ be a prefix-closed set. Let $\pi$ be a positive 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

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

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

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

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, provided the distribution is invariant.

**Example 3.2.2** Let $F$ be the set of factors of the Fibonacci word. The set $X = \{a, ba\}$ is a maximal prefix code (Example 3.2.1). One has $\pi_F(X) = 1$ where $\pi_F$ is defined in Example 2.3.1.
We will use the following result (see Theorem 4.2.3).

**Proposition 3.2.5** Let $F$ be a prefix-closed 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 $YA^*$ is right $F$-dense. This will imply that $Y$ is $F$-maximal by Proposition 3.2.3. Let $u \in F$. 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 3.2.5 is false for infinite prefix codes.

**Example 3.2.3** 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.

### 3.3 Average length

Let $F$ be a recurrent set and let $\pi$ be a probability distribution on $F$. The **average length** of a prefix code $X$ with respect to $\pi$ is the sum

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

**Proposition 3.3.1** Let $F$ be a prefix-closed right-essential set and let $\pi$ be a positive 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 3.2.4. Let us show that for any $p \in P$,

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

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 $pA^n \cap F = \cup_{x \in pA^n \cap X} xA^{n+|p|-|x|}$. Since $\pi(xA^{n-|x|} \cap F) = \pi(x)$, 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 3.3.1 holds for a suffix code and its set of proper suffixes, provided $\pi$ is invariant.
Example 3.3.1 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 $\pi_F(P) = 1 + (1 - \lambda) = 2 - \lambda$.

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 $u \in A^* \setminus XA^*$. 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 $(s, x, p)$ of $w$ such that $x = yz$ with $y, z \in X^*$, $sy = u$ and $v = zp$.

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 $u$ 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 4.1.1 Let $X = \emptyset$. Then $(L_X, w) = |w|$.

Note that

$$X \subset Y \Rightarrow L_Y \leq L_X.$$  \hfill (4.1)

The following is a reformulation of Proposition 6.1.6 in [3].
Proposition 4.1.2 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 $A^* \setminus X A^*$, 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 4.1.2 has a dual statement for suffix codes.

Proposition 4.1.3 Let $F$ be factorial set. Let $X \subset A^*$ be a prefix code and let $V = A^* \setminus A^* X$. Then

$$V = L(1 - A).$$  \quad (4.2)

If $X$ is bifix, one has

$$1 - X = (1 - A) L (1 - A)$$  \quad (4.3)

Proof Let $U = A^* \setminus X A^*$. By definition of the $F$-indicator, we have $L = V X^* U$. Since $X$ is prefix, we have by Proposition 3.1.1, the equality $F = X^* U$. Thus we obtain $L = V F$ (note that this is actually equivalent to Proposition 4.1.2). 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 $1 - X = (1 - A) V$, 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 [3]. The proof is quite similar.

Proposition 4.1.4 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 $aw \in F$

$$0 \leq (L, aw) - (L, w) \leq 1$$  \quad (4.4)

(ii) For any $w \in F$ and $a \in A$ such that $wa \in F$

$$0 \leq (L, wa) - (L, w) \leq 1$$  \quad (4.5)

(iii) For any $a, b \in A$ and $w \in F$ such that $awb \in F$

$$(L, aw) + (L, wb) \leq (L, w) + (L, awb)$$  \quad (4.6)

(iv) $(L, 1) = 1$

The following is a reformulation of Proposition 6.1.12 in [3].
Proposition 4.1.5 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

$$(L, ua) = \begin{cases} (L, u) & \text{if } ua \in A^+X \\ (L, u) + 1 & \text{otherwise} \end{cases}$$

(4.7)

Proof This follows directly from Proposition 4.1.3. \hfill \blacksquare

Proposition 4.1.5 has a dual for suffix codes expressing $(L, au)$ in terms of $(L, u)$.

Recall also that by Proposition 6.1.8 in [3], 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 2.2.2). Set $x_n = f^n(a)$ for $n \geq 1$. Note that $x_{n+1} = x_n x_n$ 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 $n \geq 2$, $x_{n+2} = x_{n+1}x_{n+1} = x_n x_n x_n x_n$ implies that $x_n x_n \in F$ and thus $x_n x_n \in F$. Next $X$ is a bifix code. Indeed, for $n < m$, $x_{2m}$ begins with $x_{2n} x_{2n}$ and thus cannot have $x_{2n}^2$ as a prefix. Similarly, since $x_{2m}$ ends with $x_{2n} x_{2n}$, 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 $x_n^2 \in X$. Otherwise, it is a factor of $x_{n+1}^2 = x_n x_n x_n x_n$.

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

$$H(X, F) = \{w \in F \mid A^+wA^+ \cap X \neq \emptyset\}$$

the set of internal factors of words in $X$. We denote $\bar{H}(X, F) = F \setminus H(X, F)$.

When $F$ is right essential and left essential, $X$ is $F$-thin if and only if $\bar{H}(X, F) \neq \emptyset$. Indeed, the condition is necessary. Conversely, if $w$ is in $\bar{H}(X, F)$, 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 [3].
Theorem 4.2.1 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 4.2.1, we introduce the following notation. Let $C(\mathcal{X}, F)$ be the set of pairs $(u, v)$ of words such that $u \in \bar{H}(\mathcal{X}, F)$, $v \in F$ and $uvu \in F$. We define for each pair $(u, v) \in C(\mathcal{X}, F)$ 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 $pvu \in X^*$ (see Exercise 6.2.1).

Figure 4.3: The relation $\varphi_{u,v}$

Lemma 4.2.1 The set $X$ is a prefix code if and only if, for all pairs $u, v \in C(\mathcal{X}, F)$, the relation $\varphi_{u,v}$ is a partial function from $\Pi(u)$ into itself.

Proof Assume first that $X$ is a prefix code. For $(u, v) \in C(\mathcal{X}, F)$, 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 $s' = s''w$. Since $p' \in X^*$, we have $w \in X^*$. Since $s' \notin A^*X$, this forces $s' = s''$. Furthermore, since $s' = s''$, the equality $s' = s''$ implies, by Equation (3.2), $x' = x''$, and $p' = p''$. Thus $\pi' = \pi''$.

Conversely, if $X$ is not a prefix code, let $x, y$ be distinct words in $X$ such that $x$ is a prefix of $y$. Since $X$ is $F$-thin, there is a word $w \in \bar{H}(\mathcal{X}, F)$. Since $F$ is recurrent, there is a word $s$ such that $ysw \in F$. Let $u = ysw$ and let $v$ be such that $uvu \in F$. Thus $(u, v) \in C(\mathcal{X}, F)$. By Equation (3.1) there exist $z', z'' \in X^*$ such that $u = xz'p' = yz''p''$. By the dual of Equation (3.1), there exist $s \in A^* \setminus A^*X$ and $z \in X^*$ such that $uv = sz$. Then $\pi = (s, z, 1)$, $\pi' = (1, xz', p')$ and $\pi'' = (1, yz'', p'')$ 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.

Lemma 4.2.1 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 $X^*$. 

17
Lemma 4.2.2 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 $XA$ 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 $swu \in F$. Since $X$ is right $F$-complete, there is word $w$ such that $swuw \in X^\ast$. Since $u \in \bar{H}(X, F)$, this implies that there is a parse $\pi' = (s', x', p')$ of $u$ such that $pvs', p'w \in X^\ast$. 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 $w \in \bar{H}(X, F)$ and let $v$ be such that $uvw \in F$. Set $r = uwv$. Let $t$ be such that $rtr \in F$. Then $(r, s) \in C(X, F)$. Let $\pi = (s, x, p)$ be a parse of $r$ such that $s = 1$. By the hypothesis, there is a parse $\pi' = (s', x', p')$ of $r$ such that $(\pi, \pi') \in \varphi_{r,t}$. Then $pts' \in X^\ast$. 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 $X^\ast$. Thus $X$ is right $F$-complete.

Lemma 4.2.2 has a dual formulation for left $F$-complete sets.

Proposition 4.2.1 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 4.2.1 and 4.2.2, 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 4.2.1, 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 a permutation of $\Pi(u)$. Thus it is also surjective for any pair $(u, v) \in C(X, F)$. This implies by the dual of Lemma 4.2.2 that $X$ is left $F$-complete and thus an $F$-maximal suffix code.

The converse implication is proved in an analogous way.

Proposition 4.2.1 has a dual formulation for an $F$-maximal suffix code.

Proof of Theorem 4.2.1. We first show that (i) implies (ii). If $X$ 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 4.2.1, it is left $F$-complete and thus (ii) holds. Finally assume that $X$ is neither an $F$-maximal prefix code nor an $F$-maximal suffix code. Let $y, z \in F$ be such that $X \cup y$ is prefix and $X \cup z$ is suffix. Since $F$ is uniformly recurrent, there is a word $u$ such that $yu \in F$. Then $X \cup yu$ is bifix and thus we reach a contradiction.

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

(ii) implies (iii). Consider the set $Y = X \setminus A^+ X$. It is a suffix code by definition. It is prefix since it is contained in $X$. It is left $F$ complete. Indeed, one has $A^+ X = A^+ Y$ and thus $A^+ Y$ is left $F$-dense by the dual of Proposition 3.2.3. Hence $Y$ is a maximal suffix code. By the dual of Proposition 4.2.1, 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). 

\textbf{Example 4.2.2} 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 4.2.1 is false if $F$ is not recurrent.

\textbf{Example 4.2.3} 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 [3].

\textbf{Theorem 4.2.2} Let $F$ be a recurrent set and let $X \subset F$ be a bifix code. Then $X$ is an $F$-thin and $F$-maximal bifix code if and only if the $F$-indicator $L = L_{X,F}$ of $X$ is bounded. In this case,

$$\bar{H}(X,F) = \{ w \in F \mid (L,w) = d_F \}$$

(4.9)

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

\textbf{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, $H(X,F)$ is not empty. Let $u \in H(X,F)$ and $w \in F$. Since $F$ is recurrent, there is a word $v \in F$ such that $uvw \in F$. Since $X$ is prefix, by Proposition 4.1.2, 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 $uvw$ which is in $S$ is a prefix of $u$. Thus $(L,uvw) = (\Sigma A^*,uvw) = (\Sigma A^*,u) = (L,u)$.

Since by Equation (4.8), $(L,w) \leq (L,uvw)$, we get $(L,w) \leq (L,u)$. This shows that $L$ is bounded and moreover that $\bar{H}(X,F)$ 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 $w' = pw$. Then $(L,w') > (L,w)$ 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, set $u = au'$ with $a \in A$ and $u' \in F$. Then $(L,au') \geq (L,u') \geq (L,w)$ by Equation (4.8). This implies $(L,au') = (L,u'w) = (L,w)$. By the dual of Equation (4.7) this implies that $uw \in XA^*$. This implies that $XA^*$ is right $F$-dense. Indeed, for any $v \in F$, since $F$ is recurrent, there is a word $u \in F$ such that $vwu \in F$. Then $vwu \in XA^*$ by the above argument. Thus $XA^*$ is right $F$-dense and $X$ is an $F$-maximal bifix code by Theorem 4.2.1.

Finally, $X$ is $F$-thin. Indeed, let $w \in F$ be such that $(L,w) \geq (L,w')$ for all $w' \in F$. Then, as we have seen above, for any nonempty word $u \in F$ such that $
\[ u \nu \in F, \text{ we have } u \nu \in XA^\ast. \] Suppose that \( w \) is an internal factor of a word in \( X \). Let \( p, s \in F \setminus \lbrace 1 \rbrace \) be such that \( pws \in X \). Then \( pw \in F \) implies \( pw \in XA^\ast \), a contradiction. Thus \( w \in \bar{H}(X, F) \).

The degree in \( F \), or \( F \)-degree, denoted \( d_F(X) \), of an \( F \)-thin and \( F \)-maximal bifix code \( X \subset F \) is the maximal number of parses of words of \( F \) with respect to \( X \). Thus an \( F \)-maximal bifix code with finite degree is the same as an \( F \)-thin and \( F \)-maximal bifix code.

**Example 4.2.4** Let \( F \) be the set of factors of the Fibonacci word. The set \( X = \lbrace a, bab, baab \rbrace \) 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 4.2.1, 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 4.2.5** Let \( F \) be the set of factors of the Fibonacci word. Then \( X = \lbrace aaba, ab, baa, baba \rbrace \) 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 \( \bar{H}(X, F) \).

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

**Theorem 4.2.3** Let \( F \) be a recurrent set. For any thin maximal bifix code \( X \subset A^+ \), the set \( Y = X \cap F \) is an \( F \)-thin and \( F \)-maximal bifix code. One has \( d_F(X \cap F) \leq d(X) \) with equality when \( X \) is finite.

The proof uses the following lemma, in which we denote \( L = L_{A^\ast} \).

**Lemma 4.2.3** Let \( X \subset A^+ \) be a thin maximal bifix code. For any words \( u \in A^+ \) and \( v \in A^* \) such that \((L, uvu) = (L, u)\), one has \( uvu \in XA^* \).

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

**Proof of Theorem 4.2.3.** Since \( X \) is thin, its indicator is bounded. Let \( w \in F \) be such that \((L_X, w)\) is maximal among the values of \( L_X \) on \( F \). Then \( w \) cannot be an internal factor of a word in \( Y \). Thus \( Y \) is \( F \)-thin.

We show that \( XA^\ast \) is right \( F \)-dense. Since \( XA^\ast \cap F = YA^\ast \cap F \), it will imply that \( YA^\ast \) is right \( F \)-dense and thus that \( Y \) is an \( F \)-maximal prefix code by Proposition 3.2.3.
Let \( u \in F \). We prove that there is a word \( v \in F \) such that \( uvu \in XA^* \). 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 4.2.3, 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 an parse of a word in \( F \) with respect to \( Y \) is also an 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 \((L_X,y)\) is maximal among all words \( y \) in \( Y \). If \((L_X,y) < d(X)\), then \( y \) is in \( H(X) \). Thus there exist \( p, s \in A^* \) such that \( pys \in X \). 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 3.2.5, 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 \( yv \in F \), it is left essential. By the dual of Proposition 3.2.5, 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 \((L_X,wyv) > (L_X,y)\). We conclude that \((L_X,y)\) is not maximal in \( Y \), a contradiction.

**Example 4.2.6** 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).

**Example 4.2.7** Let \( F \) be the set of factors of the Thue-Morse word. Consider again \( X = a \cup ba^*b \). Then \( Y = \{ a, baab, bab, bb \} \) is a finite \( F \)-maximal bifix code of \( F \)-degree 2 (see Figure 2.2).

We will see in Example 4.4.1 that a strict inequality can hold in Theorem 4.2.3.

### 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 [3].

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 \( Y = (HA \cap F) \setminus H \) and \( Z = (AH \cap F) \setminus H \). Then the set \( X' = K \cup (Y \cap Z) \) 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 4.3.1** One has $Y \subset S$ and $Z \subset P$.

*Proof* By Theorem 4.2.2, the $F$-indicator of $X$ is bounded and $\hat{H}(X, F) = F \setminus H$ is the set of words in $F$ with maximal value $d_F(X)$. Let $y = ha$ be in $Y$ with $h \in H$ and $a \in A$. Since $y \notin H$, we have $(L_X, ha) > (L_X, h)$. Thus, by Proposition 4.1.5, $y = ha$ 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 $Z \subset P$ is symmetrical.

**Lemma 4.3.2** 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 $x'$ 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 $x' \in Y = (HA \cap F) \setminus H$.

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 $x' \in Z$. Thus we conclude that $x' \in Y \cap Z \subset X'$.

There is a dual of Lemma 4.3.2 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 $Y \cap Z$. By Lemma 4.3.1, a word in $Z$ 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 $Y \cap Z$ 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 $y \in Y \cap Z$ can be a proper prefix of another word $y'$ in $Y \cap Z$, otherwise $y = y'z$, with $z \in A^+$. Therefore, since $Y \subset S$ by Lemma 4.3.1, there is $t \in A^+$ such that $ty = ty'z \in X$. Consequently, $y' \in Y \cap H$, 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 4.3.2, 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 \( Y \cap Z \neq \emptyset \). 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 \( X' \setminus K = Y \cap Z \neq \emptyset \). Consider \( x' \in Y \cap Z \). Since \( (Y \cap Z) \cap H(X) \) is empty, and since \( H(X') \subseteq H(X) \), \( x' \) 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 4.1.2.

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 \( x' \in Z \subset P \). 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 \( x'' \in Y \setminus Z \). 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 4.3.1** 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 4.2.4). We have \( K = \{a\} \), \( H = \{1, a, aa\} \), \( Y = \{b, ab, aab\} \) and \( Z = \{b, ba, baa\} \). Thus \( X' = \{a, b\} \).

The following is a generalization of Proposition 6.3.14 in [3].

**Proposition 4.3.1** 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 4.3.3** 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 \( st'a \notin S \), we have \( st'a \in A^*X \). Thus by Equation (4.7), this implies \( (L_X, st') = (L_X, st'a) \). 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 \( s \in F \setminus H \).
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 4.3.1.**

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 4.3.3, 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 \( x' \in Y \cap Z \). Since \( Y \subset S \) by Lemma 4.3.1, we have \( s \in S \). And since \( Z \subset AH \), 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 4.3.2, 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 4.3.1 concerning the set of proper prefixes of an \( F \)-thin and \( F \)-maximal bifix code \( X \subset F \).

The following property generalizes Theorem 6.3.15 in [3].

**Theorem 4.3.2** 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 4.3.1, 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 [3], 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 3.2.4). Next, it applies only for recurrent sets such that there exists a positive invariant probability distribution (in order to be able to use Proposition 3.3.1).

**Corollary 4.3.1** 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 3.3.1, one has \( \lambda(X) = \pi(S) \). In view of Theorem 4.3.2, 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 3.2.4, we have \( \pi(Y_i) = 1 \) for \( 1 \leq i \leq d - 1 \). Thus \( \lambda(X) = d \).

**Example 4.3.2** Let \( F \) be the set of factors of the Fibonacci word and let \( X = \{a, bab, baab\} \). The set \( X \) is an \( F \)-maximal bifix code of degree 2. With respect to the unique invariant probability distribution of \( F \) (Example 2.3.1), 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 [3].

**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 \( L_X = (F)((X) - 1)(F) \). 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 \( Y \). The words in \( F(w) \cap X \) other than \( x \) are all in \( K \). Thus we have again \( F(w) \cap X \subset F(w) \cap Y \).

To show Formula (4.11), assume first that \( w \in H(X,F) \). Then \((L_X, w) < d \) by Theorem 4.2.2. Moreover, \((L_X, w) = (L_Y, w)\) by Formula (4.10). Thus Equation (4.11) holds. Next, suppose that \( w \in H(Y,F) \). 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 [3].

**Theorem 4.4.1** 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 4.2.3.

**Example 4.4.1** 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 [3].

**Theorem 4.4.2** Let \( F \) be a factorial 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 4.1.4. 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 4.2.2. 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 the \( Y \) 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 \( K(X) = H(X, F) \cap X \subset Y \) and \( Y \) is the kernel of \( X \).

**Example 4.4.2** 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 4.4.
4.5 Finite maximal bifix codes

The following generalizes Theorem 6.5.2 of [3].

**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 $X'$. Since $X'$ is a finite $F$-maximal bifix code of degree $d$, there are only a finite number of kernels and we are done. □

**Example 4.5.1** 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 4.5.1** 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 $w \in \hat{H}(X,F)$. 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 $x = pqr$, with $q \in F_r \cap H(X,F)$, and $p$, $r$ 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. □

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 4.5.1). Next, in a
Sturmian set $F$, any $F$-thin and $F$-maximal bifix code of $F$-degree $d$ has $d + 1$
elements (Theorem 5.2.1). This generalizes the fact that $\text{Card}(F \cap A^n) = n + 1$
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

Set $A = \{a, b\}$. A set of words is called Sturmian if it is the set of factors of a
Sturmian word.

By Proposition 2.2.4 a Sturmian set if uniformly recurrent.

Let $F$ be a Sturmian set. A word $w$ is right-special if one has $wa, wb \in F$.

It is said to be left-special if $aw, bw \in F$.

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

As another consequence of the definitions, for every right-special word $w$, exactly one of $aw, bw$ 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 $wa, wb$ which is
left-special.

The following statement is a direct consequence of the definitions.

**Proposition 5.1.1** Let $F$ be a Sturmian set of words. One has $ab, ba \in F$ and $aa \in F$ if and only if $bb \notin F$. For each $n \geq 1$ there is exactly one word of $F$ of
length $n$ which is right-special. There is a unique left infinite word such that all
its suffixes are right-special.

A symmetric statement holds for left-special words.

**Proposition 5.1.2** Any word in a Sturmian set is a prefix of a right-special
word.

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

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

**Example 5.1.1** 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 5.1.2, the word $u$ is a prefix of a right special word $v$. Then we choose $x_{n+1} = va$. The set $X_{n+1} = X_n \cup x_{n+1}$ is a bifix code. It is not $F$-maximal since $vb$ is incomparable for the prefix order with the words of $X_{n+1}$. Moreover $(L_{X_{n+1}}, u) \geq n + 1$ 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 5.1.3** 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 5.1.3 holds for suffix codes and right-special words.

### 5.2 Cardinality

The following result shows that Proposition 4.5.1 can be made much more precise for Sturmian sets.

**Theorem 5.2.1** Let $F$ be a Sturmian set. For any $F$-thin and $F$-maximal bifix code $X \subset F$, one has $\text{Card}(X) = d_F(X) + 1$.

The proof uses three lemmas.

**Lemma 5.2.1** Let $F$ be a Sturmian set. Let $X \subset F$ be an $F$-thin and $F$-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 Proposition 5.1.1 there is a right-special word of length $n$. Then $u$ is not a factor of a word of $X$. By Theorem 4.2.2 it implies that $(L_X, u) = d_F(X)$.

The word $u$ has $d_F(X)$ suffixes which are in $P$. 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.

**Lemma 5.2.2** Let $X \subset \{a, b\}^*$ be a finite prefix code or the set $\{1\}$ and let $P$ be the set of proper prefixes of the words of $X$. Let $d = \text{Card}\{p \in P \mid pA \subset P \cup X\}$. Then, $\text{Card}(X) = d + 1$. 

29
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 then $X = \{1\}$ and $P = \emptyset$. Assume $n \geq 1$ and let $S_X = \{ p \in P \mid pA \subseteq P \cup X \}$. If $1 \notin S_X$, then all words of $X$ begin with the same letter. We may assume that $X = aY$. Then, $Y$ is a prefix code or the set $\{1\}$ and $\text{Card}(S_Y) = \text{Card}(S_X) = d$. Hence, by induction hypothesis $\text{Card}(X) = \text{Card}(Y) = d + 1$. Otherwise, $X = aY \cup bZ$. Set $t = \text{Card}(S_Y)$ and $q = \text{Card}(S_Z)$. We have $t + q = d - 1$. By induction hypothesis, $\text{Card}(Y) = t + 1$ and $\text{Card}(Z) = q + 1$. Therefore, $\text{Card}(X) = \text{Card}(Y) + \text{Card}(Z) = t + q + 2 = d + 1$.

Proof of Theorem 5.2.1 Let $P$ be the set of proper prefixes of the words of $X$. An element of $P$ satisfies $pA \subseteq P \cup X$ if and only it is right-special. Thus the proof results directly from Lemmas 5.2.1 and 5.2.2.

Example 5.2.1. Let $F$ be the set of factors of the Fibonacci word. We have seen in Example 4.4.2 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

\[
\begin{array}{|c|c|c|}
\hline
\text{code} & \text{kernel} & \text{derived code} \\
\hline
\text{aab, aba, baa, bab} & \emptyset & \text{aa, ab, ba} \\
\text{aa, aba, baab, bab} & \text{aa} & \\
\text{aaba, ab, baa, baba} & \text{ab} & \\
\text{aab, abaa, abab, ba} & \text{ba} & \\
\text{aa, ab, baaba, baba} & \text{aa, ab} & \\
\text{aa, abab, abab, ba} & \text{aa, ba} & \\
\text{aaba, aababa, ab, ba} & \text{ab, ba} & \\
\text{a, baabaab, baabab, babaaab} & \text{a} & \text{a, baab, bab} \\
\text{a, baab, bababaabab, bababab} & \text{a, baab} & \\
\text{a, baabaab, baababaab, bab} & \text{a, bab} & \\
\text{aaba, abaa, ababa, b} & \text{b} & \text{aa, aba, b} \\
\text{aaba, ababa, ababa, b} & \text{aa, b} & \\
\text{aaba, aababa, ab, ba} & \text{aba, b} & \\
\hline
\end{array}
\]

Figure 5.5: The 13 maximal bifix codes of degree 3 in the factors of the Fibonacci word

Theorem 5.2.1 on this example. Consider the code $X$ with kernel $K = \{a, baab\}$. There exactly three right-special words which are proper left factors of words of $X$, namely 1, $ba$ and $bababaaba$ (indicated in black on Figure 5.6).

5.3 Periodicity

The following result, due to Coven and Hedlund, is well-known (see [13], 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 \( d \).

The proof uses the critical factorization theorem (see [12]) that we recall below. For a pair of words \((u,v)\), the repetition \( \text{rep}(u,v) \) is the minimal length of a word \( w \) such that \( A^*u \cap A^*w \neq \emptyset \), \( vA^* \cap wA^* \neq \emptyset \).

Let \( w = a_1a_2\cdots a_n \) be a word with letters \( a_i \in A \). An integer \( p \geq 1 \) is a period of \( w \) if for \( 1 \leq i \leq j \leq n \), \( j - i = p \) implies \( a_i = a_j \). A factorization of a word \( w \in A^* \) is a pair \((u,v)\) of words such that \( w = uv \).

Theorem 5.3.3 For any word \( w \in A^+ \), the maximal value of \( \text{rep}(u,v) \) for all factorizations \((u,v)\) of \( w \) is the least period of \( w \).

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 \( v \) be a nonempty prefix of \( y \) and set \( y = vz \). Let \( w \) be a prefix of \( z \) of length larger than \( n \) and set \( z = wt \).

Let \((q,r)\) be a factorization of \( v \). We show that \( \text{rep}(q,r) \leq n \).

Since \( uq \) has \( d \) parses with respect to \( X \), there are \( d \) suffixes \( p_1, p_2, \ldots, p_d \) of \( uq \) which are in \( P \). We may assume that \( p_1 = 1 \). Similarly, there are \( d \) prefixes \( s_1, s_2, \ldots, s_d \) of \( rz \) which are in \( S \). We may assume that \( s_1 = 1 \). Since \( uqrw \) has \( d \) parses, for each \( p_i \) with \( 2 \leq i \leq d \) there is exactly one \( s_j \) with \( 2 \leq j \leq d \) such that \( p_is_j \in X \). We may renumber the \( s_i \) in such a way that \( p_is_i \in X \) for \( 2 \leq i \leq d \). Set \( x_i = p_is_i \). Since \( uq \notin S \), we have \( uq \in A^*X \). Let \( x_0 \) be the word of \( X \) which is a suffix of \( uq \). Similarly, let \( x_1 \) be the word of \( X \) which is a prefix of \( rw \) (see Figure 5.7).
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 $\text{rep}(q, r) \leq n$. If $x_0 = x_i$ for an index $i$ with $1 \leq i \leq d$, set $w = x_0 = x_i$ and $x_0 = sp_i$. We have $sw = ws_i$. Thus $|s_i| = |s|$ is a period of the word $sw$. Again $\text{rep}(q, r) \leq |s| \leq n$. 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 $1 \leq i, j \leq n$. Set $w = x_i = x_j$ and let $p_j = p_i t$, $s_i = s_j s$. Since $sw = wt$, $|s|$ is a period of the word $sw$. This implies that $\text{rep}(q, r) \leq |s| \leq n$.

By the critical factorization theorem, this implies that the least period of $v$ is at most equal to $n$. Since the least period of all prefixes of $y$ is at most $n$, an infinity of them have the same period. This implies that $y$ is periodic. Thus $x$ is ultimately periodic.

6 Basis of subgroups

The main result of this section is Theorem 6.1.1. It states that an $F$-thin and $F$-maximal bifix code $X \subset F$ of $F$-degree $d$ is a basis of a subgroup of index $d$ of the free group on $A$.

6.1 Sturmian basis

We denote by $A^\circ$ the free group generated by $A$.

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 an $F$-thin and $F$-maximal bifix code of $F$-degree $d$.

A basis of a subgroup of finite index of $A^\circ$ contained in a Sturmian set $F$ is called an $F$-basis or a Sturmian basis. As a consequence of Theorem 6.1.1 a subgroup of finite index admits for any Sturmian set $F$, an $F$-basis.

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 $r$, then

$$n - 1 = d(r - 1)$$

Let $X$ be an $F$-thin and $F$-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 $\text{Card}(X) = d + 1$ by Schreier’s formula (6.1).
Since $x_k$ we have $u_{x_k}$ any $\varepsilon$ one has $X$ generated by $H$. Then $H \cap F = X^* \cap F$.

The proof uses the following lemmas.

For a set $X \subset A^+$, we define

$$\text{Zigzag}(X) = \{ut \in A^* \mid uv, wv, wt \in X \text{ for some } u, v, w, t \in A^*\}.$$  

The definition of $\text{Zigzag}(X)$ is related to the conditions defining a Nielsen reduced set in a free group. Note that $X \subset \text{Zigzag}(X)$ since for $t = v = 1$ and $u = w \in X$, we have $u \in \text{Zigzag}(X)$. Note also that $\text{Zigzag}(X)$ is included in the subgroup generated by $X$. Indeed, $ut = (uv)(wv)^{-1}wt$. The following is part of Exercise 6.1.2 in [3].

**Lemma 6.1.1** Let $X \subset A^+$ be a bifix code and let $H$ be the subgroup of $A^*$ generated by $X$. Then $X^* = H \cap A^*$ if and only if $\text{Zigzag}(X) \subset X$.

**Proof** Suppose first that $X^* = H \cap A^*$. Let $u, v, w, t$ be such that $uv, wv, wt \in X$. Then $ut \in H \cap A^*$ and thus $ut \in X^*$. If $ut = xy$ with $x, y \in X^*$ and $x$ nonempty, then $x$ cannot be a prefix of $u$. Thus $t = t'y$ with $x = ut'$. Then $wt' = wv(ut)^{-1}ut'$ is in $H \cap A^*$. This implies $wt' \in X^*$. This forces $t' = t$ and $y = 1$. Thus $ut$ is in $X$ and this proves that $\text{Zigzag}(X) \subset X$.

Conversely, let $h \in H \cap F$. Set

$$h = x_0^\varepsilon_0 x_1^\varepsilon_1 \cdots x_n^\varepsilon_n$$

with $x_i \in X$, $\varepsilon_i = \pm 1$ and $x_i \neq x_{i+1}$ for $0 \leq i \leq n - 1$. Note that, since $X$ is bifix, no $x_j$ such that $\varepsilon_j = -1$ cancels completely with one of its neighbors. Assume that some indices $j$ are such that $\varepsilon_j = -1$ and that $n$ is minimal with this property.

For each $j$ with $0 \leq j \leq n$, we define

$$P(j) = \{(i, k) \mid 0 \leq i \leq j \leq k \leq n, x_i^\varepsilon_i \cdots x_k^\varepsilon_k \in A^*\}$$

The set is nonempty since $(0, n) \in P(j)$. Set

$$\text{Span}(j) = \min\{k - i \mid (i, k) \in P(j)\}.$$  

Let $j$ with $0 \leq j \leq n$ be such that $\varepsilon_j = -1$ with $\text{Span}(j)$ minimal. Let $(i, k) \in P(j)$ be such that $k - i = \text{Span}(j)$. Then, for any $\ell$ such that $i \leq \ell < j$, one has $\varepsilon_\ell = 1$. Indeed, since the cancellations in a free group are non-crossing, any $x_\ell$ such that $\varepsilon = -1$ has to cancel completely inside $x_i \cdots x_{j-1}$ and thus we have $u = x_i \cdots x_{j-1} \in A^*$. In the same way, we have $v = x_{j+1} \cdots x_k \in A^*$. Since $x_j$ cannot cancel completely with one of its neighbors, we have $i = j - 1$ and $k = j + 1$. 

33
This shows that \( \text{Span}(\ell) < \text{Span}(j) \) which is impossible. Similarly, for any \( m \) such that \( j < m \leq k \), one has \( \varepsilon_m = 1 \). Since \( x_j \) cancels completely with \( x_i \cdots x_{j-1} \) on the left and \( x_{j+1} \cdots x_k \) on the right, this forces \( i = j - 1 \) and \( k = i + 1 \).

Since \( x_j \) cancels with \( x_{j-1} \) and \( x_{j+1} \), we have \( x_{j-1} = uv, x_j = vw \) and \( x_{j+1} = wt \) and \( ut \in \text{Zigzag}(X) \). Thus \( ut \in X \). By the minimality of \( n \), we have \( x_0^* \cdots x_{j-2}^* ut x_{j+2}^* \cdots x_n^* \in X^* \). Thus \( h \) is in \( X^* \).

We denote by \( \text{Closure}(X) \) the bifix code generating the smallest biunitary submonoid containing \( \text{Zigzag}(X) \).

\( \text{Lemma 6.1.2} \) Let \( X \subset A^+ \) be a bifix code and let \( H \) be the subgroup of \( A^\circ \) generated by \( X \). Let \( (X_n)_{n \geq 0} \) be the sequence of bifix codes defined by \( X_0 = X \) and \( X_{i+1} = \text{Closure}(X_i) \) for \( i \geq 0 \). Then \( \cup_{n \geq 0} X_n^* = H \cap A^* \).

**Proof** Let us prove by induction on \( n \geq 0 \) that \( X_n^* \subset H \cap A^* \). It is true for \( n = 0 \). Assume that it holds for \( n \geq 0 \). For \( u, v, w, t \in A^* \) such that \( uv, wv, wt \in X_n \), we have \( ut = uv(wv)^{-1}wt \). Thus \( ut \) is in \( H \). This shows that \( \text{Zigzag}(X_n) \subset H \cap A^* \).

Let \( n \) be the subgroup generated by \( Y \). Since \( H \cap A^* \) is a biunitary submonoid, this implies that \( \text{Closure}(X_n)^* \subset H \cap A^* \). Thus \( X_{n+1}^* \subset H \cap A^* \).

Let us prove the converse inclusion. Since all submonoids \( X_n^* \) are biunitary, the set \( \cup_{n \geq 0} X_n^* \) is a biunitary submonoid. Let \( Y \) be the bifix code such that \( Y^* = \cup_{n \geq 0} X_n^* \) and let \( K \) be the subgroup generated by \( Y \). Since \( \text{Zigzag}(Y) \subset Y \), we have \( Y^* = K \cap A^* \) by Lemma 6.1.1. Since \( K \) is a subgroup containing \( X \), we have \( H \subset K \). Thus \( H \cap A^* \subset K \cap A^* = Y^* \).

\( \text{Lemma 6.1.3} \) Let \( F \) be a Sturmian set and let \( X \subset F \) be a bifix code. The set \( \text{Zigzag}(X) \) is a bifix code \( Y \) such that \( Y \cap F = X \).

**Proof** Let \( u, v, w, t \in F \) be such that \( uv, wv, wt \in X \). Let \( p \) be the longest common prefix of \( t, v \). Since \( y \neq z \), we have \( v \neq t \). We cannot have \( p = v \) or \( p = t \) since \( X \) is prefix. Thus \( t = pt' \) and \( v = pv' \) where \( t', v' \) are nonempty words which distinguish initial letters. Since \( wpv', wpt' \in F \), the word \( wp \) is right special. Similarly, since \( upt', upv' \in F \), the word \( up \) is right special. Thus \( u \) and \( w \) are comparable for the suffix order and so are \( x, y \) a contradiction.

**Proof of Proposition 6.1.1**. We have \( X^* \cap F \subset H \cap F \). Conversely,

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 \( x \), which are considered in \([24]\) and \([11]\), up to a left-right symmetry. Vuillon has shown 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. We modify slightly the proof of Lemma 2.2 in \([11]\) in order to prove the following sharpening of one direction in Vuillon’s theorem.

**Proposition 6.1.2** Let \( A = \{a, b\} \) and let \( F \) be a Sturmian set. For a word \( u \in F \), the set \( R_F(u) \) is a basis of the free group \( A^\circ \).

**Proof** Let \( x \) be a Sturmian word such that \( F = F(x) \). By Proposition 5.1.1, the set of factors of length 2 of \( x \) is either \( \{a^2, ab, ba\} \) or \( \{ab, ba, b^2\} \). Hence, either \( a^2 \in F(x) \) or \( b^2 \in F(x) \) but not both. For later use, we note that since \( x \) is recurrent, we may assume that \( x \) starts with \( a \).

We note also that, if \( a^2 \in F(x) \) (hence \( b^2 \not\in F(x) \)), and if \( x \) starts with \( a \), then \( x = f(y) \) where \( f \) is the morphism defined by \( f(a) = a \), \( f(b) = ab \) and where \( y \) is a Sturmian word by Proposition 2.3.2 in \([13]\).

We prove the proposition by induction on \( |u| \) starting with the words of length at most 1.

- If \( u = 1 \), then \( R_{F(u)}(u) = A \) which is a basis of \( A^\circ \).
- If \( u = a \), we have two cases. If \( a^2 \) is a factor of \( x \), then the first return words to \( a \) are clearly \( a \) and \( ba \), which form a basis of \( A^\circ \). If \( b^2 \) is a factor of \( x \), then it is known that between two consecutive occurrences of \( a \), there are \( k \) or \( k + 1 \) occurrences of \( b \) for a \( k \) depending only on \( x \). Hence the first returns to \( a \) are \( b^k a \) and \( b^{k+1} a \) which form a basis of \( A^\circ \).
- If \( u = b \), we proceed like in the preceding case.

Suppose now that we have proved the proposition for all words \( u \) of length at most \( n \) for \( n \geq 1 \), and all Sturmian sets \( F \). We take now a Sturmian set \( F \) and some word \( u \in F \) of length \( n + 1 \). Let \( x \) be a Sturmian word such that \( F = F(x) \). We have to consider the four following cases.

(i) \( u \in a A^* b \),
(ii) \( u \in a A^* a \),
(iii) \( u \in b A^* a \),
(iv) \( u \in b A^* b \).

By symmetry between \( a \) and \( b \), we may assume in the sequel that \( a^2 \in F(x) \) (hence \( b^2 \not\in F(x) \)) and moreover that \( x \) starts with \( a \), so that \( x = f(y) \) with \( y \) Sturmian.

(i) We have \( u = f(w) \) where \( |w| = |u|_a = |u| - |u|_b < |u| \). For later use, note that if \( |u|_b \geq 2 \), we have even \( |w| < |u| - 1 \). The occurrences of \( w \) in \( y \) are mapped bijectively by \( f \) onto the occurrences of \( u \) in \( x \). This is because the inverse mapping \( f^{-1} \) corresponds to the decoding of \( x \) by the code \( \{a, ab\} \),
and by hypothesis (i). We deduce that $R_{F(x)}(u) = f(R_{F(y)}(u))$, and therefore

$R_{F(x)}(u)$ is a basis of $A^\circ$, $f$ being an automorphism of $A^\circ$.

(ii) We may write $u = u'a$. Suppose first that $u'$ is not right-special. Then each occurrence of $u'$ in $x$ is followed by $a$. Hence, the first return words to $u'$ correspond bijectively to those of $u$ and more precisely

$$R_{F(x)}(u) = a^{-1}R_{F(x)}(u')a.$$  

Since $|u'| < |u|$, the induction hypothesis implies that $R_{F(x)}(u)$ is a basis of $A^\circ$.

Suppose now that $u'$ is right-special. Then $u'b \in F(x)$. Therefore $u' = u''a$ and $u = u''aa$. There exists $w \in F(y)$ such that $f(w) = u''a$, by considering the decoding of an occurrence of $u$ in $x$. Then $f$ induces a bijection from the set of occurrences of $w$ in $y$ onto the set of occurrences of $u$ in $x$, as follows from the alternative: if $w$ is followed by $a$, then $f(wa) = w''aa = u$; if $w$ is followed by $b$, then $f(wb) = w''aab = ub$. We deduce that $R_{F(x)}(u) = a^{-1}f(R_{F(y)}(u))a$, which proves that $R_{F(x)}(u)$ is a basis of $A^\circ$.

(iii) This case reduces to case (i), by the following remarks: the set $F(x)$ is closed under reversal; the bases of $A^\circ$ are closed under reversal (that is, the anti-automorphism of $A^\circ$ which fixes $a$ and $b$); the basis of $A^\circ$ are closed by conjugacy by an element $u$.

(iv) Since $b^2 \notin F(x)$, each occurrence of $u$ is preceded by an $a$. Then we have $R_{F(x)}(u) = R_{F(x)}(au)$. Since $au \in A^\circ b$, we reduced to case (i), since $|au|_b \geq 2$.

---

**Proof of Theorem 6.1.1.** Assume first that $X$ is an $F$-thin and $F$-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 $Hp = Hq$. We may assume that $p = uq$. Then $Huq = Hq$ implies $Hu = H$ and thus $u \in H$. By Proposition 6.1.1, 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 5.2.1 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 $Q$.

For any $w \in F$, we define a partial map $\varphi(w)$ from $P$ into itself as follows. Since $X$ is prefix, for $p \in P$, there is at most one $q \in P$ such that $pw \in X^*q$. Set $q = p\varphi(w)$.

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 $quy$ is in $F$. Thus there is a word $r \in P$ such that $quy \in X^*r$. Since $y \in R_{F}(u)$, the word $r$ is a suffix of $u$ and thus we have $r \in Q$ (see Figure 6.8). Each element of $Q$ is obtained exactly once in this way. This shows that the restriction of $\varphi(y)$ to $Q$ is a permutation.

Since, by Proposition 6.1.2, $R_{F}(u)$ is a basis of the free group $A^\circ$, the group generated by $R_{F}(u)$ is $A^\circ$. Let

$$U = \{u \in A^\circ \mid Qu \subset HQ\}$$

36
The set $U$ is a subgroup of $A^\circ$. Since $R_F(u) \subset U$, we have $U = A^\circ$. 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 4.2.3, the set $X = Y \cap F$ is an $F$-thin and $F$-maximal bifix code. By the preceding argument, the group generated by $X$ is of finite index equal to $d_F(X)$. Thus $d_F(X) = d$.

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
& u & a & b & aa & ab \\
\hline
R_F(u) & a, ba & ab, aab & baa, babaa & ab, aab \\
\hline
\end{array}
\]

**Example 6.1.1** 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 $F$.

**Example 6.1.2** Let $F$ be the set of factors of the Fibonacci word. Let $X$ be the bifix code shown on Figure 6.9. 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 6.10.

\[
\begin{array}{|c|c|c|c|c|}
\hline
& 1 & a & b & aba \\
\hline
& a & ba & b & ba \\
\hline
\end{array}
\]

Figure 6.9: An $F$-maximal bifix code of $F$-degree 4.
6.2 Syntactic groups

In this section, we describe results on syntactic groups obtained using Sturmian basis. We first recall the basic terminology on deterministic automata and transition monoids (see [3] for a more detailed exposition). We then prove our main result which states that for any transitive permutation group \( G \) of degree \( d \) and rank 2 is a syntactic group of a bifix code with \( d + 1 \) elements (Theorem 6.2.2).

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 \( p \cdot q = \emptyset \) otherwise.

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 \( u^{-1}X = \{v \in A^* \mid uv \in X\} \) for \( u \in A^* \). The initial state is the set \( X \) and the terminal states are the sets \( u^{-1}X \) for \( u \in X \).

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

Let \( X \subset A^* \) be a prefix code. Then there is a deterministic automaton \( A = (Q, 1, 1) \) such that \( X^* = \text{Stab}(1) \). In particular, the minimal automaton of \( X^* \) has a unique terminal state which coincides with the initial state.

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

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 6.2.1** 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 \( I \) and \( G \) is 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 \( I \). \( \blacksquare \)

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

Let \( X \) be a prefix code and let \( A = A(X^*) \). A syntactic group \( G \) of \( X \) is called special if \( \varphi_A^{-1}(G) \) is a cyclic submonoid. In particular a special syntactic group is cyclic.
Theorem 6.2.1 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 6.2.1 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 [15] for a bibliography on this problem).

Theorem 6.2.1 is clearly not true for special syntactic groups since the $\mathbb{Z}/n\mathbb{Z}$ is a syntactic group of $X = a^n$ for any $n \geq 1$.

We will use Theorem 6.1.1 to prove the following result. The rank of a group is the minimal cardinality of a set generating $G$.

Theorem 6.2.2 Any transitive permutation group of degree $d$ and rank 2 is a syntactic group of a bifix code with $d + 1$ elements.

Proof Let $G$ be a transitive group of degree $d$ on a set $R$ generated by a set $\{g, h\}$. Let $\psi$ be the morphism from $\{a, b\}^*$ onto $G$ defined by $\psi(a) = g$ and $\psi(b) = h$. Let $r$ be an element of $R$ and let $H$ be the subgroup fixing $r$. Let $Z$ be the bifix code such that $Z^* = \psi^{-1}(H)$. Let $F$ be a Sturmian set and let $X = Y \cap F$. By Theorem 6.1.1, the code $X$ is an $F$-thin and $F$ maximal bifix code of degree $d$. Let $A = (S, 1, 1)$ be the minimal automaton of $X^*$. Let $P$ be the set of proper prefixes of the words of $X$. Let $Q$ be the set of right-special elements of $P$. Set $I = \{1 \cdot q \mid q \in Q\}$. Let $u \in F$ be the longest element of $Q$.

Then $I$ is the image of $\varphi_{A}(u)$. Indeed, for any $p \in Q$ such that $p \cdot u \neq \emptyset$ there is a parse $(s, x, q)$ of $u$ such that $p \cdot s = 1$ and thus $p \cdot u = 1 \cdot q$.

By Proposition 6.1.2, the set $Y = R_{F}(u)$ is a basis of the free group $A^\mathbb{Z}$.

The set of elements of the monoid $M = \varphi_{A}(A^\mathbb{Z})$ with image $I$ contains the submonoid $\varphi_{A}(Y^*)$. Thus it contains an idempotent $e$. Since $\varphi^{-1}(e)$ contains elements of arbitrary length, the rank of $e$ cannot exceed $d$. Thus the image of $e$ is $I$.

The set $G' = e\varphi_{A}(Y^*)e$ is the maximal group contained in $M$ which contains $e$. Indeed, it is clearly a group of permutations on the common image of its elements. Moreover, if $I \cdot w = I$, then $uw \in A^\mathbb{Z}u \cap F$ and thus $w \in Y^*$.

Finally, $G$ and $G'$ are equivalent as permutation groups because the diagram below is commutative. In this diagram, the horizontal arrows correspond to the action of $G'$ and $G$ on $I$ and $R$ respectively. The vertical arrows correspond to the bijection from $I$ to $R$ which maps $1 \cdot q$ with $q \in Q$ to $r\psi(q)$.

Theorem 6.2.2 was known before only in particular cases. In [14] it is shown for the case of a group generated by a $d$-cycle and another permutation. In [19], it is proved for that for an Abelian group of rank $r$ and order $d$ there exists a

39
bifix code $X$ such that $\text{Card}(X) - 1 = (r - 1)d$. 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 6.1.2.

**Example 6.2.1** We consider again the code of Example 6.1.2. The minimal automaton of $X^*$ is represented on Figure 6.11. The action on the sets of states with four elements is shown on Figure 6.12. 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 $(14)(28)$.

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


