Acyclic, connected and tree sets

Valérie Berthé[1], Clelia De Felice[2], Francesco Dolce[3], Julien Leroy[4], Dominique Perrin[3], Christophe Reutenauer[5], Giuseppina Rindone[3]

[1]CNRS, Université Paris 7, [2]Università degli Studi di Salerno, [3]Université Paris Est, LIGM, [4]Université du Luxembourg, [5]Université du Québec à Montréal

December 30, 2013 1 h 36

Abstract

Given a set F of words, one associates to each word w in F an undirected graph, called its extension graph, and which describes the possible extensions of w in F on the left and on the right. We investigate the family of sets of words defined by the property of the extension graph of each word in the set to be acyclic or connected or a tree. We prove that in a uniformly recurrent tree set, the sets of first return words are bases of the free group on the alphabet. Concerning acyclic sets, we prove as a main result that a set F is acyclic if and only if any bifix code included in F is a basis of the subgroup that it generates.

Contents

1 Introduction

This paper studies properties of classes of sets which occur as the set of factors of infinite words of linear factor complexity. It is part of a series of papers devoted to this subject initiated in [3]. These classes of sets, called acyclic, connected or tree sets, are defined by a limitation to the possible two-sided extensions of a word of the set. We will see that Sturmian sets are tree sets (by Sturmian we mean the sets of factors of strict episturmian words, also called Arnoux-Rauzy words). Moreover, the sets obtained by coding a regular interval exchange set are also tree sets (see [5]). Any word w in a tree set is neutral in the sense that the number of pairs (a, b) of letters such that $awb \in F$ is equal to the number of letters a such that $aw \in F$ plus the number of letters b such that $wb \in F$ minus 1. We express this property saying that it is a neutral set.

We study sets of first return words in a tree set F. For this, we use Rauzy graphs, which are restrictions of a de Bruijn graph to the set of vertices formed by the words of given length in a set F. We first show that if F is a recurrent connected set, the group described by any Rauzy graph of F containing the alphabet A, with respect to some vertex is the free group on A (Theorem 5.2). Next, we prove that in a uniformly recurrent connected set containing A, the set of first return words to any word in F generates the free group on A (Theorem 5.6). Next, we prove that if F is a uniformly recurrent tree set containing A, the set of first return words to any word of F is a basis of the free group on A (Corollary 5.8). The proof uses the fact that in a uniformly recurrent neutral set F containing the alphabet A, the number of first return words to any word of F is equal to $\mathrm{Card}(A)$, a result obtained in [1].

Our main result is that a set F is acyclic if and only if any bifix code contained in F is a basis of the subgroup that it generates (Theorem 6.1 referred to as the Freeness Theorem). This is related to the main result of [3], referred to as the Finite Index Basis Theorem, proving that, in a Sturmian set F, a finite bifix code is F-maximal of F-degree d if and only if it is a basis of a subgroup of index d. This result is generalized in [5] to uniformly recurrent tree sets. The proof uses the results of this paper and, in particular Corollary 5.8. In the case of an acyclic set, the subgroup generated by a bifix code need not be of finite index, even if the bifix code is F-maximal (and even if the set F is uniformly recurrent, see Example 6.4).

We also prove a more technical result. We say that a submonoid M of the free monoid is saturated in a set F if the subgroup H of the free group generated by M satisfies $M \cap F = H \cap F$. We prove that if F is acyclic, the submonoid generated by a bifix code contained in F is saturated in F (Theorem 6.2 referred to as the Saturation Theorem). This property plays an important role in the proof of the Finite Index Basis Theorem.

Our paper is organized as follows.

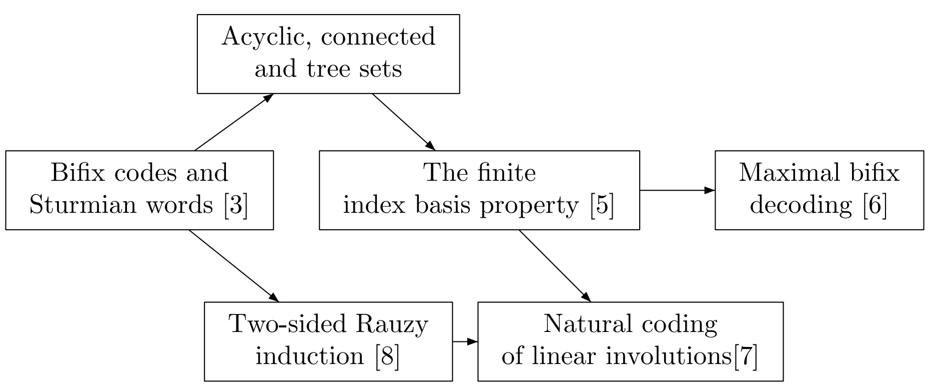
In Section 2 we present the definitions and basic properties used in the paper.

In Section 3, we introduce strong, weak and neutral sets. We prove a result on the cardinality of sets of first return words (Theorem 3.6) which is a generalization of a result from [1].

In Section 4, we define the extension graph of a word with respect to a set F. We define acyclic, connected and tree sets by the corresponding property of the extension graph of each word in the set to be acyclic, connected or a tree. We also introduce more general extension graphs where left (resp. right) extensions are relative to a finite suffix (resp. prefix) code. We prove that in acyclic sets, these more general extension graphs are also acyclic (Proposition 4.7).

In Section 5, we study sets of first return words in tree sets. We first show that if F is a recurrent connected set, the group described by any Rauzy graph of F containing the alphabet A, with respect to some vertex is the free group on A (Theorem 5.2). Next, we prove that in a uniformly recurrent connected set F containing A, the set of first return words to any word of F generates the free group on A (Theorem 5.6). We use Theorem 3.6 to prove that if F is additionally acyclic, then every set of first return words is a basis of the free group on A (Corollary 5.8).

In Section 6 we state and prove our main results (Theorem 6.1 and Theorem 6.2). The proof uses the notion of incidence graph of a bifix code (already introduced in [3]).



Some results used in this paper are proved in our first paper [3]. In turn, the results of this paper are used in other papers in preparation on similar objects. We include for clarity the logical dependency between these papers.

Ackowledgement This work was supported by grants from Region Ile-de-France and ANR project Equinocs.

2 Preliminaries

sectionPreliminaries

In this section, we first recall some definitions concerning words, codes and automata (see [4] for a more complete presentation). We give the definition of recurrent and uniformly recurrent sets of words. We also give the definitions and basic properties of bifix codes (see [3] for a more detailed presentation). We define basic notions concerning automata. We present the class of reversible automata and its connection with the Stallings automaton of a subgroup of a free group.

2.1 Recurrent sets

Let A be a finite nonempty alphabet. All words considered below, unless stated explicitly, are supposed to be on the alphabet A. We denote by A^* the set of all words on A. We denote by 1 or by ε the empty word. We denote by $|x|$ the length of a word x. A set of words is said to be *factorial* if it contains the factors of its elements.

For a set X of words and a word u, we denote

$$u^{-1}X = \{v \in A^* \mid uv \in X\}.$$

the right *residual* of X with respect to u.

Let F be a set of words on the alphabet A. For $w \in F$, we denote

$$\begin{aligned} L(w) &= \{a \in A \mid aw \in F\} \\ R(w) &= \{a \in A \mid wa \in F\} \\ E(w) &= \{(a,b) \in A \times A \mid awb \in F\} \end{aligned}$$

and further

$$\ell(w) = \mathrm{Card}(L(w)), \quad r(w) = \mathrm{Card}(R(w)), \quad e(w) = \mathrm{Card}(E(w)).$$

A word w is *right-extendable* if $r(w) > 0$, *left-extendable* if $\ell(w) > 0$ and *biextendable* if $e(w) > 0$. A factorial set F is called *right-essential* (resp. *left-essential*, resp. *biessential*) if every word in F is right-extendable (resp. left-extendable, resp. biextendable).

A word w is called *right-special* if $r(w) \geq 2$. It is called *left-special* if $\ell(w) \geq 2$. It is called *bispecial* if it is both right and left-special.

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

A set of words F is said to be *uniformly recurrent* if it is 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 of F of length n. A uniformly recurrent set is recurrent.

A *morphism* $f : A^* \to B^*$ is a monoid morphism from A^* into B^*. If $a \in A$ is such that the word $f(a)$ begins with a and if $|f^n(a)|$ tends to infinity with n, there is a unique infinite word denoted $f^\omega(a)$ which has all words $f^n(a)$ as prefixes. It is called a *fixpoint* of the morphism f.

A morphism $f : A^* \to A^*$ is called *primitive* if there is an integer k such that for all $a, b \in A$, the letter b appears in $f^k(a)$. If f is a primitive morphism, the set of factors of any fixpoint of f is uniformly recurrent (see [13] Proposition 1.2.3 for example).

An infinite word is *episturmian* if the set of its factors is closed under reversal and contains for each n at most one word of length n which is right-special (see [3] for more references). It is a *strict episturmian* word if it has exactly one right-special word of each length and moreover each right-special factor u is such that $r(u) = \mathrm{Card}(A)$.

A *Sturmian set* is a set of words which is the set of factors of a strict episturmian word. Any Sturmian set is uniformly recurrent (see [3]).

exampleFibonacci

Example 2.1 Let $A = \{a, b\}$. The *Fibonacci morphism* is the morphism $f : A^* \to A^*$ defined by $f(a) = ab$ and $f(b) = a$. The *Fibonacci word*

$$x = abaababaabaababaababa\ldots$$

is the fixpoint $x = f^\omega(a)$ of the Fibonacci morphism. It is a Sturmian word (see [17]). The set $F(x)$ of factors of x is the *Fibonacci set*.

exampleTribonacci

Example 2.2 Let $A = \{a, b, c\}$. The Tribonacci word

$$x = abacabaabacababacabaabacaba\cdots$$

is the fixpoint $x = f^\omega(a)$ of the morphism $f : A^* \to A^*$ defined by $f(a) = ab$, $f(b) = ac$, $f(c) = a$. It is a strict episturmian word (see [14]). The set $F(x)$ of factors of x is the *Tribonacci set*.

2.2 Free groups

In this section, we fix our notation concerning free groups (see [18] for example).

We denote by A° the free group on the alphabet A. It is the set of all words on the alphabet $A \cup A^{-1}$ which are *reduced*, in the sense that they do not have any factor aa^{-1} or $a^{-1}a$ for $a \in A$. Note that the exponent -1 used here should not be confused with the one used to define the residual of a set of words. We extend the bijection $a \mapsto a^{-1}$ to an involution on $A \cup A^{-1}$ by defining $(a^{-1})^{-1} = a$.

For any word w on $A \cup A^{-1}$ there is a unique reduced word equivalent to w modulo the relations $aa^{-1} \equiv a^{-1}a \equiv 1$ for $a \in A$. If u is the reduced word equivalent to w, we say that w *reduces* to u and we denote $w \equiv u$. We also denote $u = \rho(w)$. The product of two elements $u, v \in A^\circ$ is the reduced word w equivalent to uv, namely $\rho(uv)$. If $w = a_1 \cdots a_n$ with $a_i \in A \cup A^{-1}$ is a reduced word, its inverse is the reduced word denoted w^{-1} and defined by $w^{-1} = a_n^{-1} \cdots a_1^{-1}$. It is easy to verify that indeed $ww^{-1} \equiv w^{-1}w \equiv 1$.

For a set X of reduced words, we denote $X^{-1} = \{x^{-1} \mid x \in X\}$.

2.3 Bifix codes

A *prefix code* is a set of nonempty words which does not contain any proper prefix of its elements. A suffix code is defined symmetrically. A *bifix code* is a set which is both a prefix code and a suffix code.

We denote by X^* the submonoid generated by a set X of words. The submonoid M generated by a prefix code satisfies the following property: if $u, uv \in M$, then $v \in M$. Such a submonoid is said to be *right unitary*. The definition of a left unitary submonoid is symmetric and the submonoid generated by a suffix code is left unitary. Conversely, any right unitary (resp. left unitary) submonoid of A^* is generated by a prefix code (resp. a suffix code) (see [4]).

A *coding morphism* for a prefix code $X \subset A^+$ is a morphism $f : B^* \to A^*$ which maps bijectively B onto X (Note that in this paper we use $\subset$ to denote the inclusion allowing equality).

Let F be a set of words. A prefix code $X \subset F$ is F-maximal if it is not properly contained in any prefix code $Y \subset F$.

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

For a factorial set F, a prefix code is F-maximal if and only if it is right F-complete (Proposition 3.3.2 in [3]).

Similarly a bifix code $X \subset F$ is F-maximal if it is not properly contained in a bifix code $Y \subset F$. For a recurrent set F, a finite bifix code is F-maximal as a bifix code if and only if it is an F-maximal prefix code (see [3], Theorem 4.2.2). For a uniformly recurrent set F, any finite bifix code $X \subset F$ is contained in a finite F-maximal bifix code (Theorem 4.4.3 in [3]).

A *parse* of a word w with respect to a bifix code X is a triple (v, x, u) such that $w = vxu$ where v has no suffix in X, u has no prefix in X and $x \in X^*$. We denote by $\delta_X(w)$ the number of parses of w. By definition, the F-degree of X, denoted $d_F(X)$, is the maximal number of parses of a word in F. It can be finite or infinite.

Let X be a bifix code. The number of parses of a word w is also equal to the number of suffixes of w which have no prefix in X and to the number of prefixes of w which have no suffix in X (see Proposition 6.1.6 in [4]).

The set of *internal factors* of a set of words X, denoted $I(X)$ is the set of words w such that there exist nonempty words u, v with $uwv \in X$.

Let F be a recurrent set and let X be a finite bifix code. By Theorem 4.2.8 in [3], X is F-maximal if and only if its F-degree d is finite. Moreover, in this case, a word $w \in F$ is such that $\delta_X(w) < d$ if and only if it is an internal factor of X, that is

$$I(X) = \{w \in F \mid \delta_X(w) < d\}.$$

In particular, any word of X of maximal length has d parses.

exampleUniform

Example 2.3 Let F be a recurrent set. For any integer $n \geq 1$, the set $F \cap A^n$ is an F-maximal bifix code of F-degree n.

2.4 Automata and groups

sectionAutomata

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

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

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

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

An automaton is called *simple* if it is trim and if it has a unique terminal state which coincides with the initial state. The set recognized by a simple automaton is a right unitary submonoid. Thus it is generated by a prefix code.

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

For a nonempty set $L \subset A^*$, we denote by $\mathcal{A}(L)$ the *minimal automaton* of L. The states of $\mathcal{A}(L)$ are the nonempty residuals $u^{-1}L$ for $u \in A^*$. For $u \in A^*$ and $a \in A$, one defines $(u^{-1}L) \cdot a = (ua)^{-1}L$. The initial state is the set L itself and the terminal states are the sets $u^{-1}L$ for $u \in L$.

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

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

One verifies that this automaton recognizes X^*. Thus for any prefix code $X \subset A^*$, there is a simple automaton $\mathcal{A} = (Q, 1, 1)$ which recognizes X^*. Moreover, the minimal automaton of X^* is simple. Note that the literal automaton is not minimal in general (see Example 2.4).

exampleLiteral

Example 2.4 Let $X = \{aa, ab, bba, bbb\}$. The literal and the minimal automata of X^* are represented in Figure 2.1 (the initial state is indicated by an incoming arrow and the terminal states by an outgoing one).

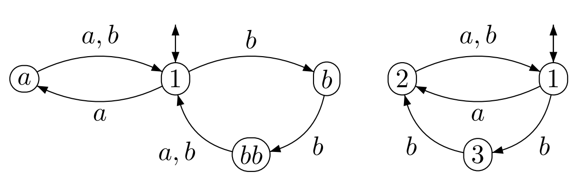


Figure 2.1: The literal and the minimal automata of X^*.

figLiteralMinimal

A simple automaton $\mathcal{A} = (Q, 1, 1)$ is said to be *reversible* if for any $a \in A$, the partial map $\varphi_{\mathcal{A}}(a) : p \mapsto p \cdot a$ is injective. This condition allows to construct the *reversal* of the automaton as follows: whenever $q \cdot a = p$ in $\mathcal{A}$, then $p \cdot a = q$ in the reversal automaton. The state 1 is the initial and the unique terminal state of this automaton. Any reversible automaton is minimal [20]. The set recognized by a reversible automaton is a submonoid generated by a bifix code.

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

The following result is from [20] (see also Exercise 6.1.2 in [4]). We denote by $\langle X \rangle$ the subgroup of the free group A° generated by X.

lemmaExercise612

Proposition 2.5 *Let $X \subset A^+$ be a bifix code. The following conditions are equivalent.*

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

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

The following example shows that for a bifix code X, the minimal automaton of X^* is not reversible in general.

Example 2.6 Let $X = \{aa, ab, ba, bbb\}$. Then X is a bifix code. The minimal automaton of X^* is represented in Figure 2.2. It is not reversible since $2 \cdot a =$

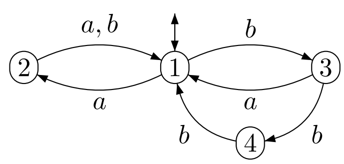


Figure 2.2: The minimal automaton of X^* figNonReversible

$3 \cdot a = 1$. Condition (i) of Proposition 2.5 is not either true since $bb = ba(aa)^{-1}ab$ is in $\langle X \rangle \cap A^*$ but not in X^*.

Let $\mathcal{A} = (Q, i, T)$ be a deterministic automaton. A *generalized path* is a sequence $(p_0, a_1, p_1, a_2, \ldots, p_{n-1}, a_n, p_n)$ with $a_i \in A \cup A^{-1}$ and $p_i \in Q$, such that for $1 \le i \le n$, one has $p_{i-1} \cdot a_i = p_i$ if $a_i \in A$ and $p_i \cdot a_i^{-1} = p_{i-1}$ if $a_i \in A^{-1}$. The *label* of the generalized path is the reduced word equivalent to $a_1 a_2 \cdots a_n$. It is an element of the free group A°. The set *described* by the automaton is the set of labels of generalized paths from i to a state in T. Since a path is a particular case of a generalized path, the set recognized by an automaton $\mathcal{A}$ is a subset of the set described by $\mathcal{A}$.

The set described by a simple automaton is a subgroup of A°. It is called the *subgroup described* by $\mathcal{A}$.

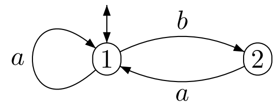


Figure 2.3: A simple automaton describing the free group on $\{a, b\}$. figDescribed

exGroupRecognized **Example 2.7** Let $\mathcal{A} = (Q, 1, 1)$ be the automaton represented in Figure 2.3 (figDescribed). The submonoid recognized by $\mathcal{A}$ is $\{a, ba\}^*$. Since $\{a, ba\}$ is a basis of the free group on A, the subgroup described by $\mathcal{A}$ is the free group on A.

The following result is Proposition 6.1.3 in [3].

propGeneratedGroup **Proposition 2.8** *Let $\mathcal{A}$ be a simple automaton and let X be the prefix code generating the submonoid recognized by $\mathcal{A}$. The subgroup described by $\mathcal{A}$ is generated by X. If moreover $\mathcal{A}$ is reversible, then $X^* = \langle X \rangle \cap A^*$.*

For any subgroup H of A°, the submonoid $H \cap A^*$ is right and left unitary and thus it is generated by a bifix code (see [4], Example 2.2.6). A subgroup H of the free group on A is *positively generated* if there is a subset of A^* which generates H. In this case, the set $H \cap A^*$ generates the subgroup H. Let X be the bifix code which generates the submonoid $H \cap A^*$. Then X generates the subgroup H. This shows that, for a positively generated subgroup H, there is a bifix code which generates H.

A subgroup of finite index of the free group is positively generated. This is well-known (see e.g. Proposition 6.1.6 in [3]) but it can be verified directly as follows.

Indeed let H be a subgroup of finite index of the free group. Let ψ be the morphism from A° onto the finite group G which is the representation of A° on the cosets of H. Let φ be the restriction of ψ to A^*. Since G is finite, and since any submonoid of a finite group is a subgroup, φ is surjective. Let us show that H is generated by the set $X = H \cap A^*$. Consider a reduced word $w \in H$. If w contains no occurrence of a letter in A^{-1}, then w is in X. Otherwise, set $w = ua^{-1}v$ for $a \in A$ and u, v reduced words. Since φ is surjective, there exist words $r, s \in A^*$ such that $\varphi(r) = \psi(u)^{-1}$ and $\varphi(s) = \psi(v)^{-1}$. Arguing by induction on the number of occurrences of letters in A^{-1}, we may assume that $ur, sv \in \langle X \rangle$. But $sar = svw^{-1}ur$ and $w = ur(sar)^{-1}sv$. The first equality shows that $sar \in H$ and consequently $sar \in X$. The second one thus implies $w \in \langle X \rangle$.

The following result is contained in Proposition 6.1.4 and 6.1.5 in [3].

propStallings **Proposition 2.9** *For any positively generated subgroup H of the free group on A, there is a unique reversible automaton $\mathcal{A}$ such that H is the subgroup described by $\mathcal{A}$. The subgroup is of finite index if and only if this automaton is a finite group automaton.*

The reversible automaton $\mathcal{A}$ such that H is the subgroup described by $\mathcal{A}$ is called the *Stallings automaton* of the subgroup H. It can also be defined for a subgroup which is not positively generated (see [2] or [15]).

The Stallings automaton of the subgroup H generated by a bifix code $X \subset A^*$ can be obtained as follows. Start with the minimal automaton $\mathcal{A} = (Q, 1, 1)$ of X^*. Then, if there are distinct states $p, q \in Q$ and $a \in A$ such that $p \cdot a = q \cdot a$, merge p, q (such a merge is called a *Stallings folding*). Iterating this operation leads to a reversible automaton which is the Stallings automaton of H (see [15]).

A subgroup H of the free group has finite index if and only if its Stallings automaton is a finite group automaton (see Proposition 2.9). In this case, the index of H is the number of states of the Stallings automaton.

Example 2.10 Let $X = \{aa, ab, ba\}$. The minimal automaton of X^* is represented in Figure 2.4 on the left. It is not reversible because $2 \cdot a = 3 \cdot a$. Merging the states 2 and 3, we obtain the reversible automaton of Figure 2.4 on the right. It is actually a group automaton, which is the Stallings automaton of the subgroup $H = \langle X \rangle$. Since the automaton describes the group $\mathbb{Z}/2\mathbb{Z}$, we

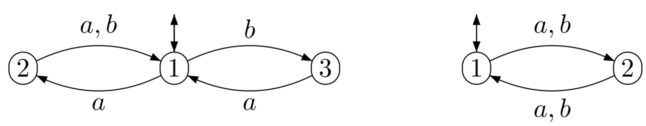


Figure 2.4: A Stallings folding.

figReversible

conclude that the subgroup generated by X is of index 2 in the free group on A.

3 Strong, weak and neutral sets

sectionNeutrality

In this section, we introduce strong, weak and neutral sets. We first prove some results concerning the factor complexity of acyclic, connected and tree sets. We prove a result on the cardinality of sets of first return words (Theorem 3.6) which is a generalization of a result from [1].

3.1 Strong, weak and neutral words

Let F be a factorial set. For a word $w \in F$, let

$$m(w) = e(w) - \ell(w) - r(w) + 1.$$

We say that, with respect to F, w is *strong* if $m(w) > 0$, *weak* if $m(w) < 0$ and *neutral* if $m(w) = 0$.

A biextendable word w is called *ordinary* if $E(w) \subset a \times A \cup A \times b$ for some $(a, b) \in E(w)$ (see [9], Chapter 4). If F is biessential any ordinary word is neutral. Indeed, one has $E(w) = (a \times (R(w) \setminus b)) \cup ((L(w) \setminus a) \times b) \cup (a, b)$ and thus $e(w) = \ell(w) + r(w) - 1$.

Example 3.1 In a Sturmian set, any word is ordinary. Indeed, for any bispecial word w, there is a unique letter a such that aw is right-special and a unique letter b such that wb is left-special. Then $awb \in F$ and $E(w) = a \times A \cup A \times b$.

We say that a set F is *strong* (resp. *weak*, resp. *neutral*) if it is factorial and every word $w \in F$ is strong or neutral (resp. weak or neutral, resp. neutral).

The sequence $(p_n)_{n\geq 0}$ with $p_n = Card(F \cap A^n)$ is called the *factor complexity* (or complexity) of F. Set $k = \mathrm{Card}(F \cap A) - 1$.

Proposition 3.2 *The factor complexity of a strong (resp. weak, resp. neutral) set F is at least (resp. at most, resp. exactly) equal to $kn + 1$.*

Given a factorial set F with complexity p_n, we denote $s_n = p_{n+1} - p_n$ the first difference of the sequence p_n and $b_n = s_{n+1} - s_n$ its second difference. The following is from [11] (it is also part of Theorem 4.5.4 in [9, Chapter 4]).

Lemma 3.3 *We have*

$$b_n = \sum_{w \in A^n \cap F} m(w) \quad \text{and} \quad s_n = \sum_{w \in A^n \cap F} (r(w) - 1)$$

for all $n \geq 0$.

Proof. Since F is factorial, we have for all n

$$\sum_{w \in A^n \cap F} e(w) = p_{n+2}, \qquad \sum_{w \in A^n \cap F} \ell(w) = \sum_{w \in A^n \cap F} r(w) = p_{n+1}.$$

Thus

$$\begin{aligned} \sum_{w \in A^n \cap F} m(w) &= \sum_{w \in A^n \cap F} (e(w) - \ell(w) - r(w) + 1) \\ &= p_{n+2} - p_{n+1} - p_{n+1} + p_n = s_{n+1} - s_n = b_n, \end{aligned}$$

giving the first formula. Next

$$\sum_{w \in A^n \cap F} (r(w) - 1) = \sum_{w \in A^n \cap F} (\mathrm{Card}(wA \cap F) - 1) = \mathrm{Card}(F \cap A^{n+1}) - \mathrm{Card}(F \cap A^n)$$

giving the second formula. ∎

Proposition 3.2 follows easily from the following lemma.

Lemma 3.4 *If F is strong (resp. weak, resp. neutral), then $s_n \geq k$ (resp. $s_n \leq k$, resp. $s_n = k$) for all $n \geq 0$.*

Proof. Assume that F is strong. Then $m(w) \geq 0$ for all $w \in F$ and thus, by Lemma 3.3, the sequence s_n is nondecreasing. Since $s_0 = k$, this implies $s_n \geq k$ for all n. The proof of the other cases is similar. ∎

We now give an example of a set of complexity $2n + 1$ on an alphabet with three letters which is not neutral.

exampleChacon **Example 3.5** Let $A = \{a, b, c\}$. The *Chacon word* on three letters is the fixpoint $x = f^\omega(a)$ of the morphism f from A^* into itself defined by $f(a) = aabc$, $f(b) = bc$ and $f(c) = abc$. Thus $x = aabcaabcbcabc \cdots$. The *Chacon set* is the set F of factors of x. It is of complexity $2n + 1$ (see [13] Section 5.5.2).

It contains strong, neutral and weak words. Indeed, $F \cap A^2 = \{aa, ab, bc, ca, cb\}$ and thus $m(\varepsilon) = 0$ showing that the empty word is neutral. Next $E(abc) = \{(a, a), (c, a), (a, b), (c, b)\}$ shows that $m(abc) = 1$ and thus abc is strong. Finally, $E(bca) = \{(a, a), (c, b)\}$ and thus $m(bca) = -1$ showing that bca is weak.

3.2 Return words

Let F be a set of words. For $w \in F$, let

$$\Gamma_F(w) = \{x \in F \mid wx \in F \cap A^+w\} \quad \text{and} \quad \Gamma'_F(w) = \{x \in F \mid xw \in F \cap wA^+\}$$

be respectively the set of *right return words* and of *left return words* to w. If F is recurrent, the sets $\Gamma_F(w)$ and $\Gamma'_F(w)$ are nonempty. Let

$$\mathcal{R}_F(w) = \Gamma_F(w) \setminus \Gamma_F(w)A^+ \quad \text{and} \quad \mathcal{R}'_F(w) = \Gamma'_F(w) \setminus A^+\Gamma'_F(w)$$

be respectively the set of *first right return words* and the set of *first left return words* to w. Note that $w\mathcal{R}_F(w) = \mathcal{R}'_F(w)w$.

Note that a recurrent set F is uniformly recurrent if and only if the set $\mathcal{R}_F(w)$ is finite for any $w \in F$. Indeed, if N is the maximal length of the words in $\mathcal{R}_F(w)$ for a word w of length n, then two successive occurrences of w in a word of F are separated by a word of length at most $N - n$. Thus any word in F of length $N + n$ contains an occurrence of w. The converse is obvious.

The following result has been proved for neutral sets in [1].

theoremCardReturn **Theorem 3.6** *Let F be a uniformly recurrent set containing the alphabet A. If F is strong (resp. weak, resp. neutral), then for every $w \in F$, the set $\mathcal{R}_F(w)$ has at least (resp. at most, resp. exactly)* $\mathrm{Card}(A)$ *elements.*

We will consider rooted trees with the usual notions of root, node, child and parent. The following lemma is well-known as a lemma on trees relating the number of its leaves to the sum of the degrees of its internal nodes.

lemmaArity **Lemma 3.7** *Let F be a prefix-closed set. Let X be a finite F-maximal prefix code and let P be the set of its proper prefixes. Then* $\mathrm{Card}(X) = 1 + \sum_{p \in P}(r(p) - 1)$.

The following lemma is also well known.

lemmaCombinat **Lemma 3.8** *Let T be a finite tree with root r and a set P of leaves, let π be a function assigning to each node an integer such that for each internal node n, $\pi(n) \le \sum \pi(m)$ where the sum runs over the children of n. Then $\sum_{n \in P} \pi(n) \ge \pi(r)$.*

A symmetric statement holds if π is such that $\pi(n) \geq \sum \pi(m)$ for each internal node n with the conclusion that $\sum_{n \in P} \pi(n) \leq \pi(r)$.

Proof of Theorem 3.6. For a word x, we denote $\pi(x) = r(x) - 1$ and for a set X of words, $\pi(X) = \sum_{x \in X} \pi(x)$.

Assume first that F is strong. Let $w \in F$ and let $n = |w|$. Set $S = F \cap A^n$. By Lemmas 3.3 and 3.4, and since F contains A, we have $\pi(S) \geq \mathrm{Card}(A) - 1$.

For $s \in S$, let P_s be the set of proper prefixes of $w\mathcal{R}_F(w)$ ending with s.

For each $s \in S$, the set P_s is a suffix code. Indeed, since a word of P_s is a proper prefix of $w\mathcal{R}_F(w)$ of length at least equal to the length of w, the word w occurs in a word of P_s exactly once and as a prefix. Let $p, q \in P_s$ with p suffix of q, we have $q = tp$. Then $p = wv$ and thus $q = twv$. Since the only occurrence of w in q is as a prefix, we have $t = 1$. Thus P_s is a suffix code.

Since F is uniformly recurrent, the set P_s is finite. We apply Lemma 3.8 to the tree T_s formed of the suffixes of P_s ending with s, considering each word $z \in T_s$ as the father of az for $a \in A$. The root of the tree is s. Since each $t \in T_s$ is strong or neutral, we have

$$\sum_{a \in L(t)} \pi(at) = \sum_{a \in L(t)} (r(at) - 1) = e(t) - \ell(t) \geq \pi(t).$$

Thus we have $\pi(P_s) \geq \pi(s)$ by Lemma 3.8.

Let $P = \cup_{s \in S} P_s$. Since the sets P_s are pairwise disjoint, we have $\pi(P) = \sum_{s \in S} \pi(P_s)$. Thus $\pi(P) \geq \pi(S)$.

Let Q be the set of proper prefixes of $\mathcal{R}_F(w)$ and set $G = w^{-1}F$. Since F is recurrent, the set $\mathcal{R}_F(w)$ is a G-maximal prefix code. Thus we may apply Lemma 3.7 to the prefix-closed set G and the G-maximal prefix code $\mathcal{R}_F(w)$. Since for any letter a, $xa \in G$ if and only if $wxa \in F$, we obtain $\mathrm{Card}(\mathcal{R}_F(w)) = 1 + \pi(wQ)$.

Next, $P = wQ$. Indeed, if $q \in Q$ then $wq \in P$, hence $wQ \subset P$. Conversely, each word in P has the form wq with $q \in Q$, so $P \subset wQ$.

We conclude that

$$\mathrm{Card}(\mathcal{R}_F(w)) - 1 = \pi(P) \geq \pi(S) \geq \mathrm{Card}(A) - 1.$$

If F is weak, then by Lemma 3.4, $\pi(S) \leq \mathrm{Card}(A) - 1$. The dual of Lemma 3.8 gives $\pi(P_s) \leq \pi(s)$ and thus $\pi(P) \leq \pi(S)$. Thus

$$\mathrm{Card}(\mathcal{R}_F(w)) - 1 = \pi(P) \leq \pi(S) \leq \mathrm{Card}(A) - 1.$$

■

The following example shows that in a set of complexity $kn + 1$ the number of first right return words need not be equal to $k + 1$.

Example 3.9 Let F be the Chacon set (see Example 3.5). We have $\mathcal{R}_F(a) = \{a, bca, bcbca\}$ but $\mathcal{R}_F(ab) = \{caab, cbcab\}$.

4 Acyclic, connected and tree sets

We introduce in this section the notion of extension graph of a word. We define acyclic (resp. connected, resp. tree) sets by the fact that all the extension graphs of its elements are acyclic (resp. connected, resp. trees). We give examples showing that a uniformly recurrent acyclic set may not be a tree set (Example 4.4) and that a uniformly recurrent neutral set may not be acyclic (Example 4.5). We introduce a generalization of the extension graphs called generalized extension graphs. We give conditions under which generalized extension graphs are acyclic (Proposition 4.7).

4.1 Extension graphs

Let F be a set of words. For a word $w \in F$, we consider an undirected graph $G(w)$ called its *extension graph* in F and defined as follows. The set of vertices is the disjoint union of $L(w)$ and $R(w)$ and its edges are the pairs $(a, b) \in E(w)$.

Example 4.1 Let F be the Tribonacci set (see Example 2.2). The graphs $G(\varepsilon)$ and $G(ab)$ are represented in Figure 4.1.

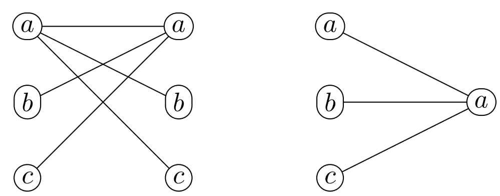


Figure 4.1: The extension graphs $G(\varepsilon)$ and $G(ab)$ in the Tribonacci set.

We say that F is an *acyclic* (resp. a connected, resp. a tree) set if it is biessential and if for every word $w \in F$, the graph $G(w)$ is acyclic (resp. connected, resp. a tree). Obviously, a tree set is acyclic and connected.

Note that a biessential set F is acyclic (resp. connected) if and only if the graph $G(w)$ is acyclic (resp. connected) for every bispecial word w. Indeed, if w is not bispecial, then $G(w) \subset a \times A$ or $G(w) \subset A \times a$, thus it is always acyclic and connected.

If the extension graph $G(w)$ of w is acyclic, then $m(w) \leq 0$. Thus w is weak or neutral. More precisely, one has in this case, $m(w) = -c + 1$ where c is the number of connected components of the graph $G(w)$.

Similarly, if $G(w)$ is connected, then w is strong or neutral. Thus, if F is an acyclic (resp. a connected, resp. a tree) set, then F is a weak (resp. strong, resp. neutral) set.

Example 4.2 A Sturmian set F is a tree set. Indeed, any word $w \in F$ is ordinary (Example 3.1), which implies that $G(w)$ is a tree.

Since a tree set is neutral, we deduce from Proposition 3.2 the following statement, where $k = \mathrm{Card}(F \cap A) - 1$.

Proposition 4.3 *The factor complexity of a tree set is $kn + 1$.*

One may wonder whether the notion of a tree set is of a topological or of a measure-theoretic nature for the associated symbolic dynamical system. In particular, one may wonder if uniformly recurrent tree sets have the property of unique ergodicity, which means that they have a unique invariant probability measure (see [3] or [9] for the definition of these notions). An element of answer is provided by interval exchange sets. Regular interval exchange sets form a special case of uniformly recurrent tree sets (see [5]).

It is well-known since [16] that there exist regular interval exchange sets that are not uniquely ergodic. This shows that the tree property does not imply unique ergodicity. However having complexity $p_n = kn + 1$, which is a priori of a topological nature, implies information on invariant measures. Indeed, according to [10], a minimal symbolic dynamical system for which $\liminf p_n/n \leq k$ is such that there exist at most k ergodic invariant measures. The bound can even be refined to $k-2$ [19] by a careful inspection of the evolution of the Rauzy graphs. For $k \leq 2$, that is for an alphabet of size at most 3 in our case, one gets the following [10]: a minimal symbolic system such that $\limsup p_n/n < 3$ is uniquely ergodic. We thus conclude that any uniformly recurrent word whose set of factors is a tree set on an alphabet of size at most 3 is uniquely ergodic.

4.2 Two examples

sectionExamples

We present two examples, due to Julien Cassaigne [12]. The first one is a uniformly recurrent acyclic set which is not a tree set.

exampleJulienAcyclic

Example 4.4 Let $A = \{a, b, c, d\}$ and let σ be the morphism from A^* into itself defined by

$$\sigma(a) = ab,\ \sigma(b) = cda,\ \sigma(c) = cd,\ \sigma(d) = abc.$$

Let F be the set of factors of the infinite word $\sigma^\omega(a)$ (see Figure 4.3 on the left). Since σ is primitive, F is uniformly recurrent. The graph $G(\varepsilon)$ is represented in Figure 4.2. It is acyclic with two connected components (and thus $m(\varepsilon) = -1$).

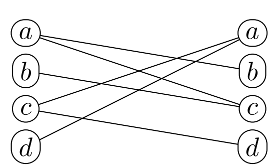


Figure 4.2: The graph $G(\varepsilon)$.

figureGepsilonJulien

We will show that for any nonempty word $w \in F$, the graph $G(w)$ is a tree. This will prove that F is acyclic. We will use some properties of the set $X = \sigma(A)$. Observe first that X is a suffix code. It has even the stronger property that distinct words of X end with distinct letters. The set X is not a prefix code but satisfies the following weaker property. If $x, x', y \in X$ and $y' \in X^*$ are such that xy is a prefix of $x'y'$, then $x = x'$ (the set X said to be *weakly prefix*).

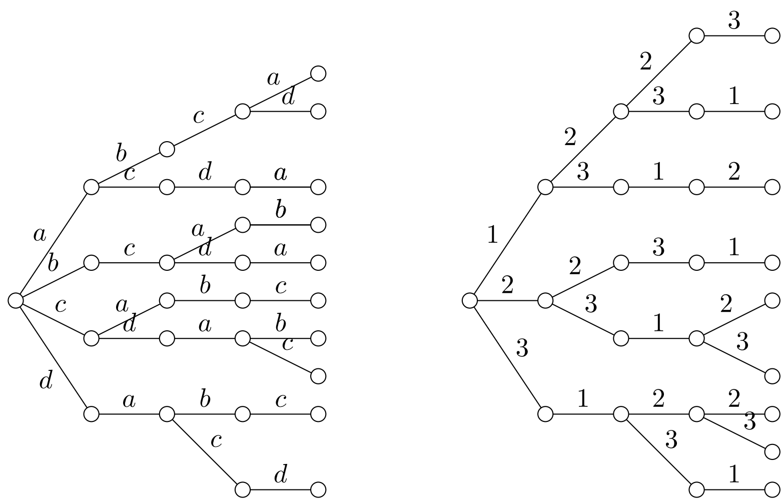


Figure 4.3: The words of length at most 4 of the sets F and G. figureCassaigne

As a third property, the set X has *synchronizing pairs*. A pair u, v of words is synchronizing if for all words p, q, if $puvq \in X^*$, then $pu, vq \in X^*$. For example (c, a) is a synchronizing pair.

Note that if (r, s) and (u, v) are synchronizing pairs, then $qrstuvw \in X^*$ implies $stu \in X^*$.

We first show the following properties.

1. If a left-special word of length at least 5 begins with a (resp. c), it begins with $abcda$ (resp. $cdabc$).
2. If a right-special word of length at least 5 ends with a (resp. c), it ends with $abcda$ (resp. $cdabc$).

Indeed, the left-special words of length at most 5 beginning with a are the prefixes of $abcda$. This implies that any left-special word of length at least 5 beginning with a begins with $abcda$.

The three other assertions can be proved in an analogous way.

Let us now show that for any nonempty bispecial word $w \in F$ the graph $G(w)$ is a tree. We use an induction on the length of the word to prove that the graph of a nonempty bispecial word is, according to its first and last letter, equal to one of the eight graphs of Figure 4.4. The assertion is true for words of length at most 4 since a, c, abc and cda are the bispecial words of length at most 4.

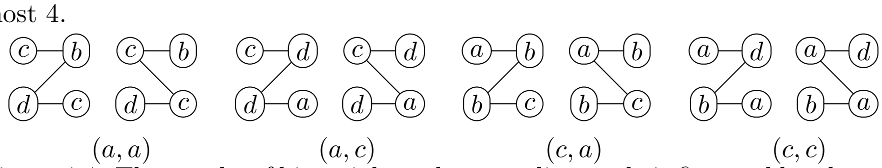


Figure 4.4: The graphs of bispecial words, according to their first and last letter. figureGraphs

Assume that v is a bispecial word of length at least 5. Assume first that v begins and ends with a. As seen previously, v begins and ends with $abcda$.

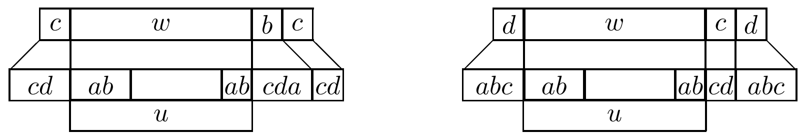


Figure 4.5: A bispecial word beginning and ending with a. figureBispecial

Set $v = ucda$. Since (d, ab) and (b, cd) are synchronizing and since $cducda \in F$, we have $u \in X^*$. Since X is a suffix code, there is a unique $w \in F$ such that $u = \sigma(w)$ and moreover $cw \in F$ (see Figure 4.5 on the left). Since $cv \in F$, and since (c, a) is synchronizing, we have also $abcv \in F$. Thus $dw \in F$ (see Figure 4.5 on the right).

Next, we have $vb \in F$. Since (d, a) is synchronizing, we have $wcd \in F$ and $vbc \in F$ (see Figure 4.5 on the right). Similarly, since $vc \in F$, we have $wbc \in F$ and $vcd \in F$ (see Figure 4.5 on the left). Thus w is a bispecial word shorter than v which begins and ends with a. By induction hypothesis, the graph of w is equal to one of the two first graphs of Figure 4.4. In both cases, we have $cwb, dwc \in F$ and thus $dvc, cvb \in F$. Next $dwb \in F$ if and only if $cvc \in F$. Thus $G(w)$ is one of the graphs if and only if $G(v)$ is the other one. This proves the property in this case.

The other cases are treated similarly.

We have thus shown that all extension graphs in F are acyclic and more precisely that $G(\varepsilon)$ is a union of two trees and all other graphs are trees. This shows, in view of Lemma 3.3 that $b_0 = -1$ and $b_n = 0$ for all $n \geq 1$. Accordingly, the complexity p_n of F is given by $p_0 = 1$ and $p_n = 2n + 2$ for $n \geq 1$.

The second example is a uniformly recurrent set which is neutral but is not a tree set (it is actually not even acyclic).

exampleJulien

Example 4.5 Let $B = \{1, 2, 3\}$ and let $\tau : A^* \to B^*$ be defined by

$$\tau(a) = 12, \quad \tau(b) = 2, \quad \tau(c) = 3, \quad \tau(d) = 13.$$

Let $G = \tau(F)$ where F is the set of Example 4.4 (see Figure 4.3 on the right). Thus G is also the set of factors of the infinite word $\tau(\sigma^\omega(a))$.

The set $Y = \tau(A)$ is a prefix code. It is not a suffix code but it is *weakly suffix* in the sense that if $x, y, y' \in X$ and $x' \in X^*$ are such that xy is a suffix of $x'y'$, then $y = y'$.

Let $g : \{a, c\}A^* \cap A^*\{a, c\} \to B^*$ be the map defined by

$$g(w) = \begin{cases} 3\tau(w) & \text{if } w \text{ begins and ends with } a \\ 3\tau(w)1 & \text{if } w \text{ begins with } a \text{ and ends with } c \\ 2\tau(w) & \text{if } w \text{ begins with } c \text{ and ends with } a \\ 2\tau(w)1 & \text{if } w \text{ begins with } c \text{ and ends with } c \end{cases}$$

It can be verified, using the fact that Y is a prefix and weakly suffix code, that the set of nonempty bispecial words of G is the union of 2, 31 and of the set

$g(S)$ where S is the set of nonempty bispecial words of F. One may verify that the words of $g(S)$ are neutral. Since the words 2, 31 are also neutral, the set G is neutral.

It is uniformly recurrent since F is uniformly recurrent and τ is a nontrivial morphism. The set G is not a tree set since the graph $G(\varepsilon)$ is neither acyclic nor connected (see Figure GepsilonJC 4.6).

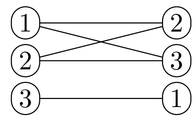


Figure 4.6: The graph $G(\varepsilon)$ for the set G. GepsilonJC

4.3 Generalized extension graphs

Let F be a set. For $w \in F$, and $U, V \subset F$, let $U(w) = \{\ell \in U \mid \ell w \in F\}$ and let $V(w) = \{r \in V \mid wr \in F\}$. The *generalized extension graph* of w relative to U, V is the following undirected graph $G_{U,V}(w)$. The set of vertices is made of two disjoint copies of $U(w)$ and $V(w)$. The edges are the pairs (ℓ, r) for $\ell \in U(w)$ and $r \in V(w)$ such that $\ell wr \in F$. The extension graph $G(w)$ defined previously corresponds to the case where $U, V = A$.

Example 4.6 Let F be the Fibonacci set. Let $w = a$, $U = \{aa, ba, b\}$ and let $V = \{aa, ab, b\}$. The graph $G_{U,V}(w)$ is represented in Figure figureStrongTree 4.7.

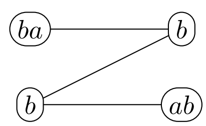


Figure 4.7: The graph $G_{U,V}(w)$. figureStrongTree

The following property shows that in an acyclic set, not only the extension graphs but, under appropriate hypotheses, all generalized extension graphs are acyclic.

PropStrongTreeCondition **Proposition 4.7** *Let F be an acyclic set. For any $w \in F$, any finite suffix code U and any finite prefix code V, the generalized extension graph $G_{U,V}(w)$ is acyclic.*

The proof uses the following lemma.

lemmaTree **Lemma 4.8** *Let F be a biessential set. Let $w \in F$ and let $U, V, T \subset F$. Let $\ell \in F \setminus U$ be such that $\ell w \in F$. Set $U' = (U \setminus T\ell) \cup \ell$. If the graphs $G_{U',V}(w)$ and $G_{T,V}(\ell w)$ are acyclic then $G_{U,V}(w)$ is acyclic.*

Proof. Assume that $G_{U,V}(w)$ contains a cycle C. If the cycle does not use a vertex in U', it defines a cycle in the graph $G_{T,V}(\ell w)$ obtained by replacing each vertex $t\ell$ for $t \in T$ by a vertex t. Since $G_{T,V}(\ell w)$ is acyclic, this is impossible. If it uses a vertex of U' it defines a cycle of the graph $G_{U',V}(w)$ obtained by replacing each possible vertex $t\ell$ by ℓ (and suppressing the possible identical successive edges created by the identification). This is impossible since $G_{U',V}(w)$ is acyclic. Thus $G_{U,V}(w)$ is acyclic. ∎

Proof of Proposition 4.7. We show by induction on the sum of the lengths of the words in U, V that for any $w \in F$, the graph $G_{U,V}(w)$ is acyclic.

Let $w \in F$. We may assume that $U = U(w)$ and $V = V(w)$ and also that $U, V \neq \emptyset$. If $U, V \subset A$, the property is true since F is acyclic.

Otherwise, assume for example that U contains words of length at least 2. Let $u \in U$ be of maximal length. Set $u = a\ell$ with $a \in A$. Let $T = \{b \in A \mid b\ell \in U\}$. Then $U' = (U \setminus T\ell) \cup \ell$ is a suffix code and $\ell w \in F$ since $U = U(w)$.

By induction hypothesis, the graphs $G_{U',V}(w)$ and $G_{T,V}(\ell w)$ are acyclic. By lemma 4.8, the graph $G_{U,V}(w)$ is acyclic. ∎

We prove now a similar statement concerning tree sets.

Proposition 4.9 *Let F be a tree set. For any $w \in F$, any finite F-maximal suffix code $U \subset F$ and any finite F-maximal prefix code $V \subset F$, the generalized extension graph $G_{U,V}(w)$ is a tree.*

The proof uses the following lemma, analogous to Lemma 4.8.

Lemma 4.10 *Let F be a biessential set. Let $w \in F$ and let $U, V \subset F$. Let $\ell \in F \setminus U$ be such that $\ell w \in F$ and $A\ell \cap F \subset U$. Set $U' = (U \setminus A\ell) \cup \ell$. If the graphs $G_{U',V}(w)$ and $G_{A,V}(\ell w)$ are connected then $G_{U,V}(w)$ is connected.*

Proof. Since F is left essential, there is a letter a such that $a\ell w \in F$ and thus $a\ell \in U(w)$. We proceed by steps.

Step 1. As a preliminary step, let us show that for each $b \in A$ such that $b\ell w \in F$, and each $v \in V(\ell w)$, there is a path from $b\ell$ to v in $G_{U,V}(w)$. Indeed, since the graph $G_{A,V}(\ell w)$ is connected there is a path from b to v in this graph. Thus, since $b\ell \in U(w)$, there is a path from $b\ell$ to v in $G_{U,V}(w)$.

Step 2. As a second step, let us show that for any $m \in U'(w) \setminus \ell$ and $v \in V(w)$, there is a path from m to v in $G_{U,V}(w)$. Indeed there is a path from m to v in $G_{U',V}(w)$. For each edge of this path of the form (ℓ, s), s is also in $V(\ell w)$ and thus, by Step 1, there is a path from $a\ell$ to s in the graph $G_{U,V}(w)$. Thus there is a path from m to v in $G_{U,V}(w)$.

Step 3. For each $b \in A$ such that $b\ell \in U(w)$, for each $v \in V(w)$, there is a path from $b\ell$ to v in $G_{U,V}(w)$. Indeed, since $G_{A,V}(\ell w)$ is connected, there is a path from b to a in $G_{A,V}(\ell w)$, thus a path from $b\ell$ to $a\ell$ in $G_{U,V}(w)$. Then there is a path from ℓ to v in $G_{U',V}(w)$ and, in the same way as in Step 2, there is a path from $a\ell$ to v in $G_{U,V}(w)$.

Consider now $m \in U(w)$ and $v \in V(w)$. If $m \notin A\ell$, then $m \in U'(w) \setminus \ell$ and thus, by Step 2, there is a path from m to v in $G_{U,V}(w)$. Next, assume that $m = b\ell$ with $b \in A$. By Step 3, there is a path from m to v in $G_{U,V}(w)$. This shows that the graph $G_{U,V}(w)$ is connected. ■

Proof of Proposition 4.9. The fact that $G_{U,V}(w)$ is acyclic follows from Proposition 4.7.

We show by induction on the sum of the lengths of the words in U, V that for any $w \in F$, the graph $G_{U,V}(w)$ is connected.

Assume first that $U(w), V(w) \subset A$. Since U is an F-maximal suffix code, we have $U(w) = L(w)$. Similarly, $V(w) = R(w)$. Thus the property is true since F is a tree set.

Otherwise, assume for example that $U(w)$ contains words of length at least 2. Let $u \in U(w)$ be of maximal length. Set $u = a\ell$ with $a \in A$. Then $U' = (U \setminus A\ell) \cup \ell$ is an F-maximal suffix code and $\ell w \in F$ since $a\ell \in U(w)$. Moreover, we have $A\ell \cap F \subset U$ since U is an F-maximal suffix code. Thus ℓ satisfies the hypotheses of Lemma 4.10.

By induction hypothesis, the graphs $G_{U',V}(w)$ and $G_{A,V}(\ell w)$ are connected. By Lemma 4.10, the graph $G_{U,V}(w)$ is connected. ■

Let F be a factorial set and let f be a coding morphism for a finite bifix code $X \subset F$. The set $f^{-1}(F)$ is called a *bifix decoding* of F. When X is an F-maximal bifix code, it is called a *maximal bifix decoding* of F.

decodingAcyclic

Theorem 4.11 *Any biessential set which is the bifix decoding of an acyclic set is acyclic.*

Proof. Let F be an acyclic set and let $f : B^* \to A^*$ be a coding morphism for a finite bifix code $X \subset F$ such that $f^{-1}(F)$ is biessential. Let $u \in f^{-1}(F)$ and let $v = f(u)$. Since X is a finite bifix code, it is both a suffix code and a prefix code. Thus the generalized extension graph $G_{X,X}(v)$ is acyclic by Proposition 4.7. Since $G(u)$ is isomorphic with $G_{X,X}(v)$, it is also acyclic. Thus $f^{-1}(F)$ is acyclic. ■

The previous statement is not satisfactory because of the assumption that $f^{-1}(F)$ is biessential which is added to obtain the conclusion. The following example shows that the condition is necessary.

exampleNotEssential

Example 4.12 Let F be the Fibonacci set and let f be the coding morphism for $X = \{aa, ab\}$ defined by $f(u) = aa$, $f(v) = ab$. Then $f^{-1}(F)$ is the finite set $\{u, v, vu, vv, vvu\}$ and thus not biessential. Note however that for any biextendable $w \in f^{-1}(F)$, the graph $G(w)$ is acyclic.

One may verify that a sufficient condition for $f^{-1}(F)$ to be biessential is that X is an F-maximal prefix code and an F-maximal suffix code.

The following result is a consequence of Proposition 4.9.

InverseImageTree

Theorem 4.13 *Any maximal bifix decoding of a recurrent tree set is a tree set.*

Proof. Let $f : B \to X$ be a coding morphism for a finite F-maximal bifix code X. Since F is recurrent, it is biessential. It implies that $f^{-1}(F)$ is also biessential. Indeed, let $u \in f^{-1}(F)$ and let $v = f(u)$. Let r, s be words of F longer than all words of X such that $rvs \in F$. Let r' (resp. s') be the suffix of r (resp. the prefix of s) which is in X. Then $f^{-1}(r')uf^{-1}(s')$ is in $f^{-1}(F)$. This shows that $f^{-1}(F)$ is biessential.

Let $u \in f^{-1}(F)$ and let $v = f(u)$. Since F is a tree set, it satisfies Proposition 4.9. Since F is recurrent and X is a finite F-maximal bifix code, X is both an F-maximal suffix code and an F-maximal prefix code. Thus the graph $G_{X,X}(v)$ is a tree. Since $G(u)$ is isomorphic with $G_{X,X}(v)$, it is also a tree. Thus $f^{-1}(F)$ is a tree set. ■

We have no example of a maximal bifix decoding of a recurrent tree set which is not recurrent.

exampleF2

Example 4.14 Let F be the Fibonacci set and let $X = A^2 \cap F = \{aa, ab, ba\}$. Let $B = \{u, v, w\}$ and let f be the coding morphism for X defined by $f(u) = aa$, $f(v) = ab$ and $f(w) = ba$. Then the set $f^{-1}(F)$ is a recurrent tree set which is actually a regular interval exchange set (see [5]). Part of the set $f^{-1}(F)$ is represented in Figure 4.8.

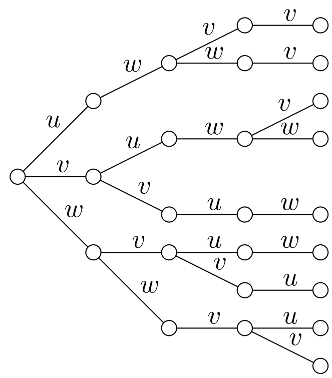


Figure 4.8: The set of words of $f^{-1}(F)$ of length at most 4.

figureSetF

5 Return words in tree sets

sectionReturnTreeSets

We study sets of first return words in tree sets. We first show that if F is a recurrent connected set, the group described by any Rauzy graph of F containing the alphabet A, with respect to some vertex is the free group on A (Theorem 5.2). Next, we prove that in a uniformly recurrent tree set containing A, the set of first return words to any word of F is a basis of the free group on A (Theorem 5.6).

5.1 Stallings foldings of Rauzy graphs

sectionRauzyGraphs

We first introduce the notion of a Rauzy graph (for a more detailed exposition, see [9]). Let F be a factorial set. The *Rauzy graph* of F of order $n \geq 0$ is the following labeled graph $G_n(F)$. Its vertices are the words in the set $F \cap A^n$. Its edges are the triples (x, a, y) for all $x, y \in F \cap A^n$ and $a \in A$ such that $xa \in F \cap Ay$.

propositionRauzy

Proposition 5.1 *Let $u \in F \cap A^n$. For any word w such that $uw \in F$, there is a path labeled w in $G_n(F)$ from u to the suffix of length n of uw.*

Conversely, the label of any path of length at most $n+1$ in $G_n(F)$ is in F.

Proof. We prove the first assertion by induction on the length of w. It is true if w is empty. Next, set $w = w'a$ with $a \in A$ and let v' be the suffix of length n of uw'. By induction hypothesis, there is a path labeled w' in $G_n(F)$ from u to the suffix v'. By definition, there is an edge from v' to the suffix of length n of $v'a$, whence the conclusion.

Next, let w be the label of a path of length $n+1$ from x to y in $G_n(F)$. Set $w = ua$ with $a \in A$. Then we have a path from x to u labeled u and an edge from u to y labeled a. Thus $ua \in F$ by definition of $G_n(F)$. ∎

When F is recurrent, all Rauzy graph $G_n(F)$ are strongly connected. Indeed, let $u, w \in F \cap A^n$. Since F is recurrent, there is a $v \in F$ such that $uvw \in F$. Then there is a path in $G_n(F)$ from u to w labeled vw by Proposition 5.1.

The Rauzy graph $G_n(F)$ of a recurrent set F with a distinguished vertex v can be considered as a simple automaton $\mathcal{A} = (Q, v, v)$ with set of states $Q = F \cap A^n$ (see Section 2.4).

Let G be a labeled graph on a set Q of vertices. The group described by G with respect to a vertex v is the subgroup described by the simple automaton (Q, v, v). We will prove the following statement.

proposition3

Theorem 5.2 *Let F be a recurrent connected set containing the alphabet A. The group described by a Rauzy graph of F with respect to any vertex is the free group on A.*

A *morphism* φ from a labeled graph G onto a labeled graph H is a map from the set of vertices of G onto the set of vertices of H such that (u, a, v) is an edge of H if and only if there is an edge (p, a, q) of G such that $\varphi(p) = u$ and $\varphi(q) = v$. An *isomorphism* of labeled graphs is a bijective morphism.

The *quotient* of a labeled graph G by an equivalence θ, denoted G/θ, is the graph with vertices the set of equivalence classes of θ and an edge from the class of u to the class of v labeled a if there is an edge labeled a from a vertex u' equivalent to u to a vertex v' equivalent to v. The map from a vertex of G to its equivalence class is a morphism from G onto G/θ.

We consider on a Rauzy graph $G_n(F)$ the equivalence θ_n formed by the pairs (u, v) with $u = ax$, $v = bx$, $a, b \in L(x)$ such that there is a path from a to b in the extension graph $G(x)$ (and more precisely from the vertex corresponding to a to the vertex corresponding to b in the copy corresponding to $L(x)$ in the bipartite graph $G(x)$).

Proposition 5.3 *If F is connected, for each $n \geq 1$, the quotient of $G_n(F)$ by the equivalence θ_n is isomorphic to $G_{n-1}(F)$.*

Proof. The map $\varphi : F \cap A^n \to F \cap A^{n-1}$ mapping a word of F of length n to its suffix of length $n-1$ is clearly a morphism from $G_n(F)$ onto $G_{n-1}(F)$. If $u, v \in F \cap A^n$ are equivalent modulo θ_n, then $\varphi(u) = \varphi(v)$. Thus there is a morphism ψ from $G_n(F)/\theta_n$ onto $G_{n-1}(F)$. It is defined for any word $u \in F \cap A^n$ by $\psi(\bar{u}) = \varphi(u)$ where $\bar{u}$ denotes the class of u modulo θ_n. But since F is connected, the class modulo θ_n of a word ax of length n has $\ell(x)$ elements, which is the same as the number of elements of $\varphi^{-1}(x)$. This shows that ψ is a surjective map from a finite set onto a set of the same cardinality and thus that it is one-to-one. Thus ψ is an isomorphism. ■

Let G be a strongly connected labeled graph. Recall from Section 2.4 that a Stallings folding at vertex v relative to letter a of G consists in identifying the edges coming into v labeled a and identifying their origins. A Stallings folding does not modify the group described by the graph with respect to some vertex. Indeed, if $p \xrightarrow{a} v$, $p \xrightarrow{b} r$ and $q \xrightarrow{a} v$ are three edges of G, then adding the edge $q \xrightarrow{b} r$ does not change the group described since the path $q \xrightarrow{a} v \xrightarrow{a^{-1}} p \xrightarrow{b} r$ has the same label. Thus merging p and q does not add new labels of generalized paths.

Proof of Theorem 5.2. The quotient $G_n(F)/\theta_n$ can be obtained by a sequence of Stallings foldings from the graph $G_n(F)$. Indeed, a Stallings folding at vertex v identifies vertices which are equivalent modulo θ_n. Conversely, consider $u = ax$ and $v = bx$, with $u, v \in F \cap A^n$ and $a, b \in A$ such that a and b (considered as elements of $L(x)$), are connected by a path in $G(x)$. Let $a_0, \ldots a_k$ and $b_1, \cdots b_k$ with $a = a_0$ and $b = a_k$ be such that (a_i, b_{i+1}) for $0 \leq i \leq k-1$ and (a_i, b_i) for $1 \leq i \leq k$ are in $E(x)$. The successive Stallings foldings at $xb_1, xb_2, \ldots, xb_k$ identify the vertices $u = a_0x, a_1x, \ldots, a_kx = v$. Indeed, since $a_ixb_{i+1}, a_{i+1}xb_{i+1} \in F$, there are two edges labeled b_{i+1} going out of a_ix and $a_{i+1}x$ which end at xb_{i+1}. The Stallings folding identifies a_ix and $a_{i+1}x$. The conclusion follows by induction.

Since the Stallings foldings do not modify the group described, we deduce from Proposition 5.3 that the group described by the Rauzy graph $G_n(F)$ is the same as the group described by the Rauzy graph $G_0(F)$. Since $G_0(F)$ is the graph with one vertex and with loops labeled by each of the letters, it describes the free group on A. ■

Example 5.4 Let F be the tree set obtained by decoding the Fibonacci set into blocks of length 2 (see Example 4.14). Set $u = aa$, $v = ab$, $w = ba$. The graph $G_2(F)$ is represented on the left of Figure 5.1. The classes of θ_2 are $\{wv, vv\}$ $\{vu\}$ and $\{ww, uw\}$. The graph $G_1(F)$ is represented on the right.

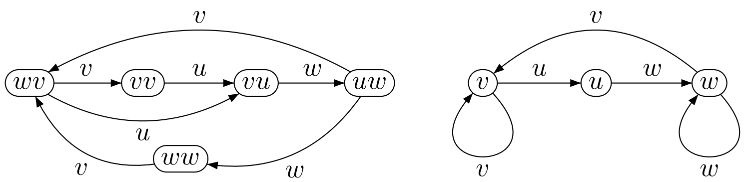


Figure 5.1: The Rauzy graphs $G_2(F)$ and $G_1(F)$ for the decoding of the Fibonacci set into blocks of length 2.

The following example shows that Proposition 5.3 is false for sets which are not connected.

Example 5.5 Consider again the Chacon set (see Example 3.5).

The Rauzy graph $G_1(F)$ corresponding to the Chacon set is represented in Figure 5.2 on the left. The graph $G_1(F)/\theta_1$ is represented on the right. It is not isomorphic to $G_0(F)$ since it has two vertices instead of one.

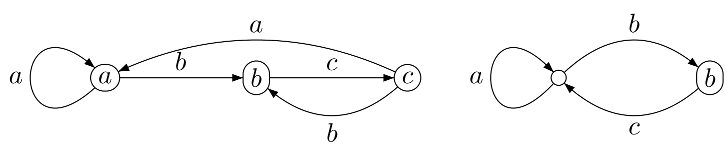


Figure 5.2: The graphs $G_1(F)$ and $G_1(F)/\theta_1$.

5.2 Return words and bases of free groups

We will prove the following result.

Theorem 5.6 *Let F be a uniformly recurrent connected set containing the alphabet A. For any $w \in F$, the set $\mathcal{R}_F(w)$ generates the free group on A.*

Proof. Since F is uniformly recurrent, the set $\mathcal{R}_F(w)$ is finite. Let n be the maximal length of the words in $w\mathcal{R}_F(w)$. In this way, any word in $F \cap A^n$ beginning with w has a prefix in $wR_F(w)$. Moreover, recall from Proposition 5.1 that the label of any path of length $n+1$ in the Rauzy graph $G_n(F)$ is in F.

Let $x \in F$ be a word of length n ending with w. Let $\mathcal{A}$ be the simple automaton defined by $G_n(F)$ with initial and terminal state x. Let X be the prefix code generating the submonoid recognized by $\mathcal{A}$. Since the automaton $\mathcal{A}$ is simple, by Proposition 2.8, the set X generates the group described by $\mathcal{A}$.

We show that $X \subset \mathcal{R}_F(w)^*$. Indeed, let $y \in X$. Since y is the label of a path starting at x and ending in x, the word xy ends with x and thus the word wy ends with w. Let $\Gamma = \{z \in A^+ \mid wz \in A^*w\}$ and let $R = \Gamma \setminus \Gamma A^+$. Then R is a prefix code and $\Gamma \cup 1 = R^*$, as one may verify easily. Since $y \in \Gamma$, we can write $y = u_1u_2 \cdots u_m$ where each word u_i is in R. Since F is recurrent and since

$x \in F$, there is $v \in F \cap A^n$ such that $vx \in F$ and thus there is a path labeled x ending at the vertex x by Proposition 5.1. Thus there is a path labeled xy in $G_n(F)$. This implies that for $1 \le i \le m$, there is a path in $G_n(F)$ labeled wu_i.

Assume that some u_i is such that $|wu_i| > n$. Then the prefix p of length n of wu_i is the label of a path in $G_n(F)$. This implies, by Proposition 5.1, that p is in F and thus that p has a prefix in $wR_F(w)$. But then wu_i has a proper prefix in $wR_F(w)$, a contradiction. Thus we have $|wu_i| \le n$ for all $i = 1, 2, \ldots, m$. But then the wu_i are in F by Proposition 5.1 and thus the u_i are in $\mathcal{R}_F(w)$. This shows that $y \in \mathcal{R}_F(w)^*$.

Thus the group generated by $\mathcal{R}_F(w)$ contains the group generated by X. But, by Theorem 5.2, the group described by $\mathcal{A}$ is the free group on A. Thus $\mathcal{R}_F(w)$ generates the free group on A. ∎

We illustrate the proof in the following example.

Example 5.7 Let F be the Fibonacci set. We have $\mathcal{R}_F(aa) = \{baa, babaa\}$. The Rauzy graph $G_7(F)$ is represented in Figure 5.3. The set recognized by the automaton obtained using $x = aababaa$ as initial and terminal state is X^* with $X = \{babaa, baababaa\}$. In agreement with the proof of Theorem 5.6, we have $X \subset \mathcal{R}_F(aa)^*$.

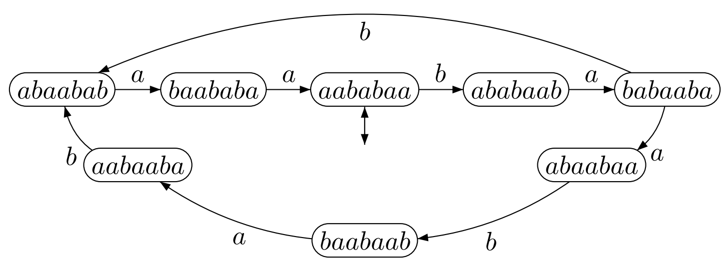


Figure 5.3: The Rauzy graph $G_7(F)$

Note that Theorem 5.6 implies that $\mathrm{Card}(\mathcal{R}_F(w)) \ge \mathrm{Card}(A)$. This is also a consequence of Theorem 3.6. When F is a tree set, Theorem 3.6 implies that $\mathrm{Card}(\mathcal{R}_F(w)) = \mathrm{Card}(A)$. Thus we have the following corollary.

Corollary 5.8 *Let F be a uniformly recurrent tree set containing the alphabet A. Then for any $w \in F$, the set $\mathcal{R}_F(w)$ is a basis of the free group on A.*

We show an example of a neutral set which is not a tree set and for which Corollary 5.8 does not hold.

Example 5.9 Consider the set F of Example 4.5. Then $\mathcal{R}_F(1) = \{2231, 31, 231\}$. This set has 3 elements, in agreement with Theorem 3.6 but it is not a basis of the free group on $\{1, 2, 3\}$ since it generates the same group as $\{2, 31\}$.

6 Bifix codes in acyclic sets

We prove in this section our main results. Bifix codes in acyclic sets are bases of the subgroup that they generate (Theorem 6.1, referred to as the Freeness Theorem). Moreover, the submonoid generated by a finite bifix code X included in an acyclic set F is such that $X^* \cap F = \langle X \rangle \cap F$ (Theorem 6.2, referred to as the Saturation Theorem). As a preliminary to the proof, we first define the incidence graph of a finite bifix code (already used in [3]). We prove a result concerning this graph, implying in particular that it is acyclic (Proposition 6.6). We then define the coset automaton whose states are connected components of the incidence graph. We prove that this automaton is the Stallings automaton of the subgroup $\langle X \rangle$ (Proposition 6.10). Finally, we prove the Freeness and the Saturation Theorems.

6.1 Freeness and Saturation Theorems

Let X be a subset of the free group. We say that X is *free* if it is a basis of the subgroup $\langle X \rangle$ generated by X. This means that if $x_1, x_2, \ldots, x_n \in X \cup X^{-1}$ are such that $x_1 x_2 \cdots x_n$ is equivalent to 1, then $x_i x_{i+1}$ is equivalent to 1 for some $1 \leq i < n$.

We will prove the following result (Freeness Theorem).

Theorem 6.1 *A set F is acyclic if and only if any bifix code $X \subset F$ is a free subset of the free group A°.*

Let M be a submonoid of A^* and let H be the subgroup of A° generated by M. Given a set of words F, the submonoid M is said to be *saturated* in F if $M \cap F = H \cap F$. If M is generated by X, then M is saturated in F if and only if $X^* \cap F = \langle X \rangle \cap F$.

Thus, for example, the submonoid recognized by a reversible automaton is saturated in A^* (Proposition 2.8).

We will prove the following result (Saturation Theorem).

Theorem 6.2 *Let F be an acyclic set. The submonoid generated by a bifix code included in F is saturated in F.*

We note the following corollary, which shows that bifix codes in acyclic sets satisfy a property which is stronger than being bifix (or more precisely that the submonoid X^* satisfies a property stronger than being right and left unitary).

Corollary 6.3 *Let F be an acyclic set, let $X \subset F$ be a bifix code and let $H = \langle X \rangle$. For any $u, v \in F$,*

(i) *if $u, uv \in H \cap F$, then $v \in X^*$.*
(ii) *if $v, uv \in H \cap F$, then $u \in X^*$.*

Proof. Assume that $u, uv \in H \cap F$. Since $v \equiv u^{-1}(uv)$, we have $v \in H$. But $v \in H \cap F$ implies $v \in X^*$ by Theorem 6.2. This proves (i). The proof of (ii) is symmetric. ■

We can express Corollary 6.3 in a different way. Let F be an acyclic set and let $X \subset F$ be a bifix code. Then no nonempty word of $\langle X \rangle$ can be a proper prefix (or suffix) of a word of X. Indeed, assume that $u \in \langle X \rangle$ is a prefix of a word of X. Then u is in $\langle X \rangle \cap F$ and thus in X^* since X^* is saturated in F. This implies $u = 1$ or $u \in X$.

We illustrate Theorem 6.1 in the following example.

Example 6.4 Let F be as in Example 4.4 and let $X = F \cap A^2$. We have

$$X = \{ab, ac, bc, ca, cd, da\}$$

The set X is an F-maximal bifix code. It is a basis of a subgroup of infinite index. Indeed, the minimal automaton of X^* is represented in Figure 6.1 on the left. The Stallings automaton of the subgroup H generated by X is obtained by merging 3 with 4 and 2 with 5. It is represented in Figure 6.1 on the right. Since it is not a group automaton, the subgroup has infinite index (see Proposition 2.9). The set X is a basis of H by Theorem 6.1. This can

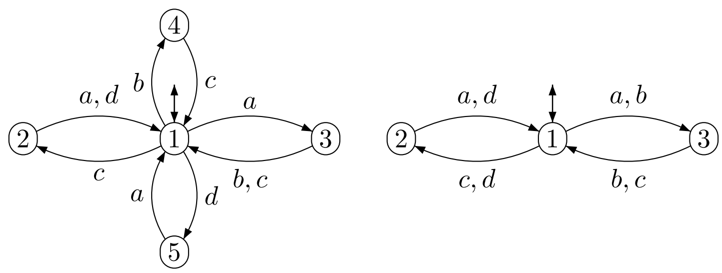


Figure 6.1: The minimal automaton of X^* and the Stallings automaton of $\langle X \rangle$.

also be seen by performing Nielsen transformations on the set X (see [18] for example). Indeed, replacing bc and da by $bc(ac)^{-1}$ and $da(ca)^{-1}$, we obtain $X' = \{ab, ac, ba^{-1}, ca, cd, dc^{-1}\}$ which is Nielsen reduced. Thus X' is a basis of H and thus also X.

Note that, in agreement with Theorem 6.2, the two words of length 2 which are in H but not in X^*, namely bb and dd, are not in F.

Theorem 6.1 is false if X is prefix but not bifix, as shown in the following example.

Example 6.5 Let F be the Fibonacci set and let $X \subset F$ be the prefix code $X = \{aa, ab, b\}$. Then $a = (ab)b^{-1}$ is in $\langle X \rangle$ and thus X generates the free group on A. Thus X is not a basis and $X^* \cap F$ is strictly included in $\langle X \rangle \cap F$ (for example $a \notin X^*$).

6.2 Incidence graph

Let X be a set, let P be the set of its proper prefixes and S be the set of its proper suffixes. Set $P' = P \setminus \{1\}$ and $S' = S \setminus \{1\}$. Recall from [3] that the incidence graph of X is the undirected graph G defined as follows. The set of vertices is the *disjoint union* of P' and S'. The edges of G are the pairs (p, s) for $p \in P'$ and $s \in S'$ such that $ps \in X$. As in any undirected graph, a connected component of G is a maximal set of vertices connected by paths.

The following result is proved in [3] in the case of a Sturmian set (Lemma 6.3.3). We give here a proof in the more general case of an acyclic set. We call a path reduced if it does not use equal consecutive edges.

newLemma633

Proposition 6.6 *Let F be an acyclic set, let $X \subset F$ be a bifix code and let G be the incidence graph of X. Then the following assertions hold.*

(i) *The graph G is acyclic.*

(ii) *The intersection of P' (resp. S') with each connected component of G is a suffix (resp. prefix) code.*

(iii) *For every reduced path $(v_1, u_1, \ldots, u_n, v_{n+1})$ in G with $u_1, \ldots, u_n \in P'$ and $v_1, \ldots, v_{n+1}$ in S', the longest common prefix of v_1, v_{n+1} is a proper prefix of all $v_1, \ldots, v_n, v_{n+1}$.*

(iv) *Symmetrically, for every reduced path $(u_1, v_1, \ldots, v_n, u_{n+1})$ in G with $u_1, \ldots, u_{n+1} \in P'$ and $v_1, \ldots, v_n \in S'$, the longest common suffix of u_1, u_{n+1} is a proper suffix of $u_1, u_2, \ldots, u_{n+1}$.*

Proof. Assertions (iii) and (iv) implies assertions (i) and (ii). Indeed, assume that (iii) holds. Consider a reduced path $(v_1, u_1, \ldots, u_n, v_{n+1})$ in G with $u_1, \ldots, u_n \in P'$ and $v_1, \ldots, v_{n+1}$ in S'. If $v_1 = v_{n+1}$, then the longest common prefix of v_1, v_{n+1} is not a proper prefix of them. Thus G is acyclic and (i) holds. Next, if v_1, v_{n+1} are comparable for the prefix order, their longest common prefix is one of them, a contradiction with (iii) again. The assertion on P' is proved in an analogous way using assertion (iv).

We prove (iii) and (iv) by induction on $n \geq 1$.

The assertions holds for $n = 1$. Indeed, if $u_1 v_1, u_1 v_2 \in X$ and if $v_1 \in S'$ is a prefix of $v_2 \in S'$, then $u_1 v_1$ is a prefix of $u_1 v_2$, a contradiction with the hypothesis that X is a prefix code. The same holds symmetrically for $u_1 v_1, u_2 v_1 \in X$ since X is a suffix code.

Let $n \geq 2$ and assume that the assertions hold for any path of length at most $2n-2$. We treat the case of a path $(v_1, u_1, \ldots, u_n, v_{n+1})$ in G with $u_1, \ldots, u_n \in P'$ and $v_1, \ldots, v_{n+1}$ in S'. The other case is symmetric.

Let p be the longest common prefix of v_1 and v_{n+1}. We may assume that p is nonempty since otherwise the statement is obviously true. Any two elements of the set $U = \{u_1, \ldots, u_n\}$ are connected by a path of length at most $2n-2$ (using elements of $\{v_2, \ldots v_n\}$). Thus, by induction hypothesis, U is a suffix code. Similarly, any two elements of the set $V = \{v_1, \ldots, v_n\}$ are connected by a path of length at most $2n-2$ (using elements of $\{u_1, \ldots u_{n-1}\}$). Thus V is a prefix code. We cannot have $v_1 = p$ since otherwise, using the fact that $u_n p$ is a prefix of u_nv_{n+1} and thus in F, the generalized extension graph $G_{U,V}(\varepsilon)$ would have the cycle $(p, u_1, v_2, \ldots, u_n, p)$, a contradiction since $G_{U,V}(\varepsilon)$ is acyclic by proposition 4.7. Similarly, we cannot have $v_{n+1} = p$.

Set $W = p^{-1}V$ and $V' = (V \setminus pW) \cup p$. Since V is a prefix code and since p is a proper prefix of V, the set V' is a prefix code. Suppose that p is not a proper prefix of all $v_2, \ldots, v_n$. Then there exist i, j with $1 \le i < j \le n+1$ such that p is a proper prefix of v_i, v_j but not of any $v_{i+1}, \ldots, v_{j-1}$. Then $v_{i+1}, \ldots, v_{j-1} \in V'$ and there is the cycle $(p, u_i, v_{i+1}, u_{i+1}, \ldots, v_{j-1}, u_{j-1}, p)$ in the graph $G_{U,V'}(\varepsilon)$. This is in contradiction with Proposition 4.7 because, V' being a prefix code, $G_{U,V'}(\varepsilon)$ is acyclic. Thus p is a proper prefix of all $v_2, \ldots, v_n$. ■

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

The following property relates the equivalence θ_X with the right cosets of $H = \langle X \rangle$. It is Proposition 6.3.5 in [3].

propTheta

Proposition 6.7 *Let X be a bifix code, let P be the set of proper prefixes of X and let H be the subgroup generated by X. For any $p, q \in P$, $p \equiv q \bmod \theta_X$ implies $Hp = Hq$.*

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

The following is proved in [3] (Lemma 6.3.6 and Lemma 6.4.2) in the case of a Sturmian set F.

lemmaCompatible

Proposition 6.8 *Let F be an acyclic set. Let $X \subset F$ be a bifix code and let P be the set of proper prefixes of X. Let $p, q \in P$ and $a \in A$ be such that $pa, qa \in P \cup X$. Then in the literal automaton of X^*, one has $p \equiv q \bmod \theta_X$ if and only if $p \cdot a \equiv q \cdot a \bmod \theta_X$.*

Proof.

Assume first that $p \equiv q \bmod \theta_X$. We may assume that p, q are nonempty. Let $(u_0, v_1, u_1, \ldots, v_n, u_n)$ be a reduced path in the incidence graph G of X with $p = u_0$, $u_n = q$. The corresponding words in X are $u_0v_1, u_1v_1, u_1v_2, \ldots, u_nv_n$. We may assume that the words u_i are pairwise distinct, and that the v_i are pairwise distinct. Moreover, since $pa, qa \in P \cup X$ there exist words v, w such that $pav, qaw \in X$. Set $v_0 = av$ and $v_{n+1} = aw$.

By Proposition 6.6, a is a proper prefix of $v_0, v_1, \ldots, v_{n+1}$. Set $v_i = av'_i$ for $0 \le i \le n+1$.

If $pa, qa \in P$, then $(u_0a, v'_1, u_1a, \ldots, v'_n, u_na)$ is a path from pa to qa in G. This shows that $pa \equiv qa \bmod \theta_X$.

Next, suppose that $pa \in X$ and thus that $v_0 = a$. By Proposition 6.6, we have $w = \varepsilon$ since otherwise $v_0 = a$ is a proper prefix of v_{n+1}. Thus $qa \in X$ and $p \cdot a = q \cdot a$.

Conversely, if $p \cdot a \equiv q \cdot a \bmod \theta_X$, assume first that $pa, qa \in P$. Then $pa \equiv qa \bmod \theta_X$ and thus there is a reduced path $(u_0, v_1, \ldots, v_n, u_n)$ in G with $u_0 = pa$ and $u_n = qa$. By Proposition 6.6, a is a proper suffix of $u_1, \ldots, u_n$. Set $u_i = u_i'a$. Thus $(p, av_1, u_1', \ldots, q)$ is a path in G, showing that $p \equiv q \bmod \theta_X$.

Finally, if $pa, qa \in X$, then (p, a, q) is a path in G and thus $p \equiv q \bmod \theta_X$.

■

6.3 Coset automaton

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

Observe first that the definition is consistent since, by Proposition 6.8, if $p \cdot a$ and $p' \cdot a$ are nonempty and p, p' are in the same class r, then $p \cdot a$ and $p' \cdot a$ are in the same class.

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

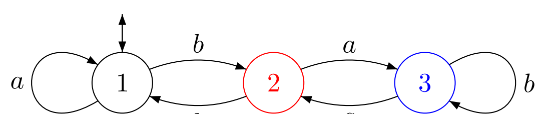


Figure 6.2: The automaton $\mathcal{B}_X$.

Example 6.9 Let F be the Fibonacci set and let

$$X = \{a, baab, babaabab, babaabaabab\}.$$

The set X is an F-maximal bifix code of F-degree 3 (see [3], Example 6.3.1). The automaton $\mathcal{B}_X$ has three states. It is a group automaton. State 2 is the class containing b, and state 3 is the class containing ba. The bifix code generating the submonoid recognized by this automaton is $Z = a \cup b(ab^*a)^*b$.

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

Proposition 6.10 *Let F be an acyclic set, and let $X \subset F$ be a bifix code. The coset automaton $\mathcal{B}_X$ is reversible and describes the subgroup generated by X.*

Moreover $X \subset Z$, where Z is the bifix code generating the submonoid recognized by $\mathcal{B}_X$.

Proof. Let $\mathcal{A} = (P, 1, 1)$ be the literal automaton of X^* and set $\mathcal{B}_X = (R, 1, 1)$. By Proposition 6.8, the automaton $\mathcal{B}_X$ is reversible.

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

Let us finally show that the coset automaton describes the group $H = \langle X \rangle$. By Proposition 2.8, the subgroup described by $\mathcal{B}_X$ is equal to $\langle Z \rangle$. Set $K = \langle Z \rangle$. Since $X \subset Z$, we have $H \subset K$. To show the converse inclusion, let us show by induction on the length of $w \in A^*$ that if, for $p, q \in P$, there is a path from the class of p to the class of q in $\mathcal{B}_X$ with label w then $Hpw = Hq$. By Proposition 6.7, this holds for $w = 1$. Next, assume that it is true for w and consider wa with $a \in A$. Assume that there are states $p, q, r \in P$ such that there is a path from the class of p to the class of q in $\mathcal{B}_X$ with label w, and an edge from the class of q to the class of r in $\mathcal{B}_X$ with the label a. By induction hypothesis, we have $Hpw = Hq$. Next, by definition of $\mathcal{B}_X$, there is an $s \equiv q \bmod \theta_X$ such that $s \cdot a \equiv r \bmod \theta_X$. If $sa \in P$, then $s \cdot a = sa$, and by Proposition 6.7, we have $Hs = Hq$ and $Hsa = Hr$. Otherwise, $sa \in X \subset H$ and $s \cdot a = r = 1$ because the class of 1 is a singleton and thus $Hqa = Hsa = H = Hr$. In both cases, $Hpwa = Hqa = Hsa = Hr$. This property shows that if $z \in Z$, then $Hz = H$, that is $z \in H$. Thus $Z \subset H$ and finally $H = K$. ∎

6.4 Proof of the main results

We can now prove Theorem 6.1. The proof uses Proposition 6.6. We will also use the elementary fact that if X is a bifix code, and $x, y \in X$ with $x \neq y$, then x cannot cancel completely with y^{-1}, which means that $\rho(xy^{-1})$ cannot be a prefix of x or a suffix of y^{-1}. Indeed, if xy^{-1} is equivalent to a prefix of x, then y is a suffix of x and if xy^{-1} is equivalent to a suffix of y^{-1} then x is a suffix of y. A symmetric argument holds for x^{-1} and y.

Proof of Theorem 6.1. To prove the necessity of the condition, assume that for some $w \in F$ the graph $G(w)$ contains a cycle $(a_1, b_1, \ldots, a_p, b_p, a_1)$ with $p \geq 2$, $a_i \in L(w)$ and $b_i \in R(w)$ for $1 \leq i \leq p$. Consider the bifix code $X = AwA \cap F$. Then $a_1wb_1, a_2wb_1, \ldots, a_pwb_p, a_1wb_p \in X$. But

$$a_1wb_1(a_2wb_1)^{-1}a_2wb_2 \cdots a_pwb_p(a_1wb_p)^{-1} \equiv 1,$$

contradicting the fact that X is free.

Let us now show the converse. Assume that F is acyclic and let $X \subset F$ be a bifix code. Set $Y = X \cup X^{-1}$. Let $y_1, \ldots, y_n \in Y$. We intend to show that provided $y_iy_{i+1} \not\equiv 1$ for $1 \leq i < n$, we have $y_1 \cdots y_n \not\equiv 1$. We may assume $n \geq 3$.

We say that a sequence $(u_i, v_i, w_i)_{1\le i\le n}$ of elements of the free group on A is *admissible* with respect to $y_1, \ldots, y_n$ if the following conditions are satisfied (see Figure figurey_i 6.3).

(i) $y_i = u_i v_i w_i$ for $1 \le i \le n$.

(ii) $u_1 = w_n = 1$ and $v_1, v_n \ne 1$.

(iii) $w_i u_{i+1} \equiv 1$ for $1 \le i \le n-1$.

(iv) For $1 \le i < j \le n$, if $v_i, v_j \ne 1$ and $v_k = 1$ for $i+1 \le k \le j-1$, then $v_i v_j$ is reduced.

Note that if $(u_i, v_i, w_i)_{1\le i\le n}$ is an admissible sequence with respect to $y_1, \ldots, y_n$, then $y_1 \cdots y_n$ is equivalent to the word $v_1 \cdots v_n$ which is a reduced nonempty word. Thus, in particular $y_1 \cdots y_n \not\equiv 1$.

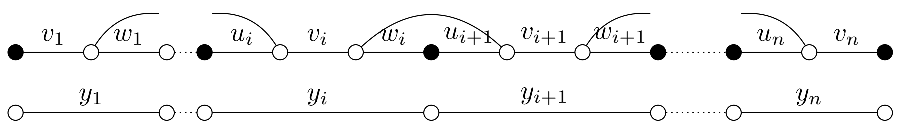


Figure 6.3: The word $y_1 \cdots y_n$. figurey_i

Let us show by induction on n that for any $y_1, \ldots, y_n$ such that $y_i y_{i+1} \not\equiv 1$ for $1 \le i \le n-1$, there exists an admissible sequence with respect to $y_1 \ldots, y_n$.

The property is true for $n = 1$. Indeed, we take $u_1 = w_1 = 1$.

Assume that the property is true for n. Among the possible admissible sequences with respect to the $y_1, \ldots, y_n$, we choose one such that $|v_n|$ is maximal.

Set $v_n = v'_n w'_n$ and $y_{n+1} = u_{n+1} v_{n+1}$ with $|w'_n| = |u_{n+1}|$ maximal such that $w'_n u_{n+1} \equiv 1$. Note that $v_{n+1} \ne 1$ since otherwise y_{n+1} would cancel completely with y_n.

If $v'_n \ne 1$, the sequence

$$(1, v_1, w_1), \ldots, (u_{n-1}, v_{n-1}, w_{n-1}), (u_n, v'_n, w'_n), (u_{n+1}, v_{n+1}, 1)$$

is admissible with respect to $y_1, \ldots, y_{n+1}$.

Otherwise, let i with $1 \le i < n$ be the largest integer such that $v_i \ne 1$. Observe that $w_i, w_{i+1}, \ldots, w_{n-1}, w'_n$ are nonempty. Indeed, if $w_j = 1$ with $i \le j \le n-1$, then $u_{j+1} = 1$ and thus y_{j+1} cancels completely with y_{j+2}. Next, if $v_n = w'_n = 1$, then y_n cancels completely with y_{n-1}.

Assume that $y_i \in X$ (the other case is symmetric).

If $y_{n+1} \in X$ (and thus $n - i$ is odd), then $v_i v_{n+1}$ is reduced because they are both in A^* and $v_{n+1} \ne 1$ as we have already seen. Thus the sequence

$$(1, v_1, w_1), \ldots, (u_{n-1}, v_{n-1}, w_{n-1}), (u_n, 1, w'_n), (u_{n+1}, v_{n+1}, 1)$$

is admissible with respect to $y_1, \ldots, y_{n+1}$.

Otherwise, let s be the longest common suffix of $u_i v_i$ and v_{n+1}^{-1}.

There is a path in the incidence graph $G(X)$ from $u_i v_i$ to v_{n+1}^{-1} (see Figure figureGX 6.4). By Proposition newLemma633 6.6, s is a proper suffix of $u_i v_i, w_{i+1}^{-1}, \ldots, w_{n-1}^{-1}, v_{n+1}^{-1}$. This implies that s^{-1} is a proper prefix of $w_{i+1}, \ldots, w_{n-1}, v_{n+1}$.

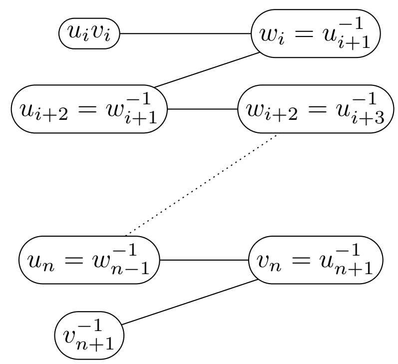


Figure 6.4: The graph $G(X)$. figureGX

It is not possible that v_i is a suffix of s. Indeed, this would imply that v_i^{-1} is a proper prefix of $w_{i+1}, \ldots, w_{n-1}, v_{n+1}$. But then we could change the $n-i+1$ last terms of the sequence $(u_j, v_j, w_j)_{1 \le j \le n}$ into $(u_i, 1, v_iw_i)$, $(u_{i+1}v_i^{-1}, 1, \rho(v_iw_{i+1})), \ldots,$ $(\rho(u_nv_i^{-1}), v_iv_n, 1)$ resulting in an admissible sequence with a longer v_n.

Thus s is a proper suffix of v_i. Since s is a proper suffix of v_i and v_{n+1}^{-1}, there are nonempty words $p, q \in A^*$ such that $v_i = ps$ and $v_{n+1}^{-1} = qs$. Moreover, the word pq^{-1} is reduced since s is the longest common suffix of v_i and v_{n+1}^{-1}. Thus we can change the last $n-i+2$ terms of the sequence formed by $(u_j, v_j, w_j)_{1 \le j \le n-1}$ followed by $(u_n, 1, v_n), (u_{n+1}, v_{n+1}, 1)$ into

$$(u_i, p, sw_i), (u_{i+1}s^{-1}, 1, \rho(sw_{i+1})), \ldots, (\rho(u_ns^{-1}), 1, sv_n), (u_{n+1}s^{-1}, q^{-1}, 1)$$

(see Figure 6.5). Since the word pq^{-1} is reduced, the new sequence is admissible.

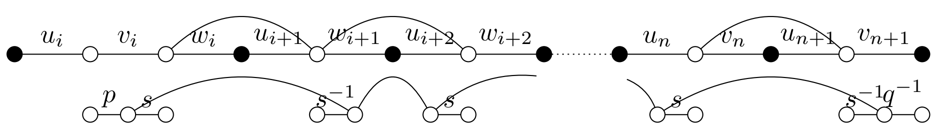


Figure 6.5: The word $y_i \cdots y_{n+1}$. figurey_is

This shows that $y_1 \cdots y_n \not\equiv 1$ for any sequence $y_1, \ldots, y_n \in X \cup X^{-1}$ such that $y_iy_{i+1} \not\equiv 1$ for $1 \le i < n$. Thus X is free. ■

We now give a proof of Theorem 6.2. It uses Proposition 6.10.

Proof of Theorem 6.2. Let F be an acyclic set and let $X \subset F$ be a bifix code. We have to prove that $X^* \cap F = \langle X \rangle \cap F$. Since $X^* \cap F \subset \langle X \rangle \cap F$, we only need to prove the reverse inclusion.

Consider the bifix code Z generating the submonoid recognized by the coset automaton $\mathcal{B}_X$ associated to X. Set $Y = Z \cap F$. By Theorem 6.1, Y is a basis of $\langle Y \rangle$.

By Proposition 6.10, we have $X \subset Z$ and thus $X \subset Y$.

Since any reversible automaton is minimal and since the automaton $\mathcal{B}_X$ is reversible by Proposition 6.10, it is equal to the minimal automaton of Z^*. Let K be the subgroup generated by Z. By Proposition 2.5, we have $K \cap A^* = Z^*$.

This shows that

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

The first inclusion holds because $X \subset Z$ implies $\langle X \rangle \subset K$. The last equality follows from the fact that if $z_1 \cdots z_n \in F$ with $z_1, \ldots, z_n \in Z$, then each z_i is in F (because F is factorial) and hence in $Z \cap F = Y$. Thus $\langle X \rangle \cap F \subset Y^*$. Consider $x \in \langle X \rangle \cap F$. Then $x \equiv x_1 \cdots x_n$ with $x_i \in X \cup X^{-1}$. But since $\langle X \rangle \cap F \subset Y^*$, we have also $x = y_1 \cdots y_m$ with $y_i \in Y$. Since $X \subset Y$ and since Y is free, this forces $n = m$ and $x_i = y_i$. Thus all x_i are in X and x is in X^*. This shows that $\langle X \rangle \cap F \subset X^*$ which was to be proved. ■

The proof of Theorem 6.1 proves not only that bifix codes in acyclic sets are free, but also that, in a sense made more precise below, the associated reductions are of low complexity.

We first define the *heigth* of w on $A \cup A^{-1}$ equivalent to 1 as the least integer h such that w is a concatenation of words of the form $w = uvu^{-1}$ where u is a word on $A \cup A^{-1}$ and v is a word of heigth $h-1$ equivalent to 1. The empty word is the only word equivalent to 1 of heigth 0.

We then define the height of an arbitrary word w on $A \cup A^{-1}$ as the least integer h such that $w = z_0 v_1 z_1 \cdots v_n z_n$ with $z_0, \ldots, z_n$ equivalent to 1 of height at most h and $v_1 \cdots v_n$ reduced.

In this way, any word on $A \cup A^{-1}$ has finite height. For example, the word $aa^{-1}cbb^{-1}$ has heigth 1 and $aaa^{-1}bb^{-1}a^{-1}$ has height 2. The words of height 0 are the reduced words.

Proposition 6.11 *Let F be an acyclic set and let $X \subset F$ be a bifix code. Any word $y = y_1 \cdots y_n$ with $y_i \in X \cup X^{-1}$ for $1 \le i \le n$ such that $y_i y_{i+1} \not\equiv 1$ for $1 \le i \le n-1$ has height at most* 1.

Proof. The proof of Theorem 6.1 shows that $y = z_0 v_1 z_1 \cdots z_{n-1} v_n z_n$ where

(i) $z_0, \ldots, z_n$ have height at most 1,

(ii) $v_1 \cdots v_n$ is reduced.

Thus y has height at most 1. ■

Example 6.12 Let X be as in Example 6.4. The word $bc(ac)^{-1}ab$, which reduces to bb, has height 1.

References

[1] L'ubomíra Balková, Edita Pelantová, and Wolfgang Steiner. Sequences with constant number of return words. *Monatsh. Math.*, 155(3-4):251–263, 2008. 2, 3, 10, 12

[2] Laurent Bartholdi and Pedro Silva. Rational subsets of groups. In *Handbook of Automata*. European science Foundation, 2011. 10

[3] Jean Berstel, Clelia De Felice, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Bifix codes and Sturmian words. *J. Algebra*, 369:146–202, 2012. 2, 3, 4, 5, 6, 9, 15, 26, 28, 29, 30

[4] Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Cambridge University Press, 2009. 4, 6, 8, 9

[5] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. The finite index basis property. 2013. 2, 3, 15, 21

[6] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Maximal bifix decoding. 2013. 3

[7] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Natural coding of linear involutions. 2013. 3

[8] Valérie Berthé, Clelia De Felice, Francesco Dolce, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Two-sided rauzy induction. 2013. `http://arxiv.org/abs/1305.0120`. 3

[9] Valérie Berthé and Michel Rigo, editors. *Combinatorics, automata and number theory*, volume 135 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010. 10, 11, 15, 22

[10] Michael Boshernitzan. A unique ergodicity of minimal symbolic flows with linear block growth. *J. Analyse Math.*, 44:77–96, 1984/85. 15

[11] Julien Cassaigne. Complexité et facteurs spéciaux. *Bull. Belg. Math. Soc. Simon Stevin*, 4(1):67–88, 1997. Journées Montoises (Mons, 1994). 11

[12] Julien Cassaigne. 2013. Personal communication. 15

[13] N. Pytheas Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel. 5, 12

[14] Jacques Justin and Laurent Vuillon. Return words in Sturmian and episturmian words. *Theor. Inform. Appl.*, 34(5):343–356, 2000. 5

[15] Ilya Kapovich and Alexei Myasnikov. Stallings foldings and subgroups of free groups. *J. Algebra*, 248(2):608–668, 2002. 10

[16] Michael Keane. Non-ergodic interval exchange transformations. *Israel J. Math.*, 26(2):188–196, 1977. 15

[17] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002. 5

[18] Roger C. Lyndon and Paul E. Schupp. *Combinatorial Group Theory*. Classics in Mathematics. Springer-Verlag, 2001. Reprint of the 1977 edition. 5, 27

[19] T. Monteil. An upper bound for the number of ergodic invariant measures of a minimal subshift with linear complexity. 2013. In preparation. 15

[20] Christophe Reutenauer. Une topologie du monoïde libre. *Semigroup Forum*, 18(1):33–49, 1979. 8