Embedding a θ-invariant code into a complete one

Jean Néraud, Carla Selmi

Embedding a θ-invariant code into a complete one

Jean Néraud*, Carla Selmi*

Laboratoire d'Informatique, de Traitemement de l'Information et des Systèmes (LITIS), Université de Rouen Normandie, UFR Sciences et Techniques, Avenue de l'université, 76830 Saint Etienne du Rouvray, France

Abstract

Let A be an arbitrary alphabet and let θ be an (anti-)automorphism of A^* (by definition, such a correspondence is determinated by a permutation of the alphabet). This paper deals with sets which are invariant under θ (θ-invariant for short) that is, languages L satisfying $\theta(L) \subseteq L$. We establish an extension of the famous defect theorem. With regards to the so-called notion of completeness, we provide a series of examples of finite complete θ-invariant codes. Moreover, we establish a formula which allows to embed any non-complete θ-invariant code into a complete one. As a consequence, in the family of the so-called thin θ-invariant codes, maximality and completeness are two equivalent notions.

Keywords: antimorphism, anti-automorphism, automorphism, (anti-)automorphism, Bernoulli distribution, bifix, code, complete, context-free, defect, equation, finite, invariant, involutive, label, maximal, morphism, order, overlap, overlapping-free, prefix, regular, suffix, thin, tree, θ-invariant, θ-code, uniform, variable-length code, word

1. Introduction

In the free monoid theory, during the last decade, research involving one-to-one *morphic* or *antimorphic* substitutions has played a particularly important part: this is due to the powerful applications of these objects, in particular in the framework of DNA-computing. In the case of automorphisms or anti-automorphisms -for short we write *(anti-)automorphisms*- given an arbitrary *alphabet*, say A, any such mapping is completely determined by extending a unique permutation of A to A^*, the *free monoid* that is generated by A.

In the special case of *involutive* (anti-)automorphisms, lots of successful investigations have been done for extending most of the now classical combinatorial properties on words. The topics of the so-called *pseudo-palindromes* [9], that of *θ-episturmian* words [3], and the one of *pseudo-repetitions* [7, 14] have been particularly involved. The framework of some peculiar families of *variable-length codes* [15] and that of *equations in words* [5, 8, 16, 21] have been concerned. Generalizations of the famous theorem of Fine and Wilf ([13],[18, Proposition 1.3.5]) were also established [6, 20].

Equations in words are also the starting point of the study in the present paper, which consists in some full version of [22]. Let A be an arbitrary alphabet and let θ be an (anti-)automorphism of A^*; we adopt the point of view from [18, Ch. 9], by considering a finite

*Corresponding author: neraud.jean@gmail.com

Email addresses: jean.neraud@univ-rouen.fr, neraud.jean@gmail.com (Jean Néraud), carla.selmi@univ-rouen.fr (Carla Selmi)

collection of unknown words, say Z. We assume that a (minimum) positive integer k (i.e. the so-called *order* of θ) exists such that $\theta^k = id_{A^*}$. This condition is particularly satisfied by every (anti-)automorphism whenever A is finite. In view of making the present foreword more easily readable, in the first instance let us take θ as an involutive (anti-)automorphism (that is, $\theta^2 = id_{A^*}$). We assign that the words in Z and their images by θ to satisfy a given equation, and we ask for the computation of a finite set of words, say Y, such that all the words of Z can be expressed as a concatenation of words in Y. Actually, such a question appears more complex than in the classical configuration, where θ does not interfer: in this classical case, according to the famous defect theorem [18, Theorem 1.2.5], it is well known that at most $|Z| - 1$ words allow to compute the words in Z. At the contrary, due to the interference of (anti-)automorphisms, in [16], examples where $|Y| = |Z|$ are provided by the authors.

Along the way, for solving our problem, applying the defect theorem to the set $X = Z \cup \theta(Z)$ might appear natural. Such a methodology garantees the existence of a set Y, with $|Y| \leq |X| - 1$ and whose elements allow by concatenation to rebuilt all the words in X. It is also well known that Y can be chosen in such a way that only trivial equations may hold among its elements: with the terminology of [1, 18, 19], Y is a *code*, or equivalently Y^*, the submonoid that it generates, is *free*. Unfortunately, since both the words in Z and $\theta(Z)$ are expressed as concatenations of words in Y, among the words of $Y \cup \theta(Y)$ non-trivial equations can still hold. In other words, by applying that methodology, the initial problem would be transferred among the words in $Y \cup \theta(Y)$.

An alternative methodology will consist in asking for codes Y which are invariant under θ (θ-*invariant* for short) that is, satisfying $\theta(Y) = Y$. Returning to the general case, where θ is an arbitrary (anti-)automorphism, this is equivalent to say that the union of the sets $\theta^i(Y)$, for all $i \in \mathbb{Z}$, is θ-invariant. By the way, it is straightforward to show that the intersection of an arbitrary family of free θ-invariant submonoids is itself a free θ-invariant submonoid. In the present paper we prove the following result:

Theorem 1. *Let θ be an (anti-)automorphism of A^* and let X be a finite θ-invariant set. If X it is not a code, then the smallest θ-invariant free submonoid of A^* containing X is generated by a θ-invariant code Y, which furthermore satisfies $|Y| \leq |X| - 1$.*

For illustrating this result in terms of equation, we refer to [5, 21], where the authors considered generalizations of the famous three unkown variables equation of Lyndon-Shützenberger [18, § 9.2]. They proved that, an involutive (anti-)automorphism θ being fixed, given such an equation with sufficiently long members, a word t exists such that any 3-uple of "solutions" can be expressed as a concatenation of words in $\{t\} \cup \{\theta(t)\}$. With the notation of Theorem 1, the elements of the θ-invariant set X are $x, y, z, \theta(x), \theta(y), \theta(z)$ and those of Y are t and $\theta(t)$: we verify that, in every case Y is a θ-invariant code, furthermore we have $|Y| \leq |X| - 1$.

With regards to the theory of codes, completeness is one of the most challenging notions: a subset X of the free monoid A^* is *complete* if any word is a factor of some word in X^*. Maximality is another important notion: a code is *maximal* if it cannot be strictly included in some other code of A^*. Actually, according to Zorn's Lemma, any code is included in a maximal one moreover, a famous result due to Schützenberger states that, for the family of the so-called *thin* codes (which contains the regular codes), maximality and completeness are two equivalent notions [1, Theorem 2.5.16]. From this point of view, in the second part of our study we are interested in *complete* θ-invariant codes. It is natural to prealably examine the case of finite codes. Clearly, the well-known complete *uniform codes* that is, the codes A^n (with $n \geq 1$), are invariant under every (anti-)automorphism. Beside that, non-trivial finite complete θ-invariant codes exist: for instance, take for A the binary alphabet $\{a, b\}$,

choose for θ the anti-automorphism that swaps the letters a and b, and consider the complete code which was introduced in [4]:

$$X = \{a^3, ab, a^2ba, a^2b^2, ba^2, baba, bab^2, b^2a, b^3\}.$$

It is staightforward to verify that X is θ-invariant. In our paper, we provide some other examples: each of the classes of bifix codes, prefix codes, and non-prefix non-suffix codes is concerned.

Despite that, the question of describing a general structure for finite complete θ-invariant codes remains largely open: this is not surprizing since, with the exception of certain special families (eg. [11, 12, 24]), no general structure that could embrace finite complete codes is described in the literature.

Another issue could consist in developing methods for embedding a code into a complete one. However, in [23], the author presents a class of codes that cannot be embedded into any finite complete one. With regards to θ-invariance, as far as we know, the question of embedding finite codes into complete ones remains open.

Actually, in [23], the question whether any finite code can be embedded into a regular one was implicitely asked: a positive answer was brought in [10], where the authors provided a now classical formula for embedding any regular code into a complete one. In the present paper, we put a corresponding problem in the framework of θ-invariant codes. Actually, by establishing the following result, we bring a positive answer:

Theorem 2. *Any non-complete θ-invariant code $X \subseteq A^*$, can be embedded into a complete one. Moreover, if A is finite and X regular, then X can be embedded into a regular complete θ-invariant code.*

As a consequence, we obtain the following result: it states that, in the framework of θ-invariant codes, a property similar to a famous one due to Schützenberger [1, Theorem 2.5.16] holds:

Theorem 3. *Given a thin θ-invariant code $X \subseteq A^*$, the five following conditions are equivalent:*
(i) *X is complete.*
(ii) *X is a maximal code.*
(iii) *X is maximal in the family of the θ-invariant codes.*
(iv) *A positive Bernoulli distribution π exists such that $\pi(X) = 1$.*
(v) *For any positive Bernoulli distribution π, we have $\pi(X) = 1$.*

We now describe the contents of our paper. Section 2 contains the preliminaries: the terminology of the free monoid is settled, and we recall some classical notions and results concerning the codes. The preceding Theorem 1 is established in Section 3, where an original example of equation is studied. In Section 4, we present several examples of finite complete θ-invariant codes. The problem of embedding a finite θ-invariant code into a complete one is also discussed: this ensures a transition to the question of embedding a regular θ-invariant code into a complete one. This last question is studied in Section 5, where the preceding Theorem 2 and Theorem 3 are established.

2. Preliminaries

2.1. Words and free monoid

We adopt the notation of the free monoid theory. In the whole paper, we consider an alphabet A, and we denote by A^* the free monoid that it generates. Given a word $w \in A^*$,

we denote by $|w|$ its length, the empty word, which we denote by ε, being the word with length 0. Given a subset X of A^*, we denote by X^* the submonoid of A^* that is generated by X, moreover we set $X^+ = X^* \setminus \{\varepsilon\}$.

Let $x \in A^*$ and $w \in A^+$. We say that x is a *prefix* (*suffix*) of w if a word u exists such that $w = xu$ ($w = ux$). Similarly, x is a *factor* of w if two words u, v exist such that $w = uxv$. Given a non-empty set $X \subseteq A^*$, we denote by $P(X)$ ($S(X)$, $F(X)$) the set of the words that are prefix (suffix, factor) of some word in X. Clearly, we have $X \subseteq P(X) \subseteq F(X)$ ($X \subseteq S(X) \subseteq F(X)$). Given a pair of non-empty words w, w', we say that it *overlaps* if words u, v exist such that $uw' = wv$ or $w'u = vw$, with $1 \leq |u| \leq |w| - 1$ and $1 \leq |v| \leq |w'| - 1$; otherwise, the pair is *overlapping-free* (in such a case, if $w = w'$, we simply say that w is overlapping-free).

2.2. Variable length codes

It is assumed that the reader has a fundamental understanding with the main concepts of the theory of variable-length codes: we only recall some of the main definitions and we suggest, if necessary, that he (she) report to [1]. A subset X of A^* is a *variable-length code* (a *code* for short) if any equation among the words of X is trivial that is, for any pair of sequences of words in X, say $(x_i)_{1 \leq i \leq n}$, $(y_j)_{1 \leq j \leq p}$, the equation $x_1 \cdots x_n = y_1 \cdots y_p$ implies $n = p$ and $x_i = y_i$, for each integer $i \in [1, n]$. By definition X^* is a *free* submonoid of A^*.

In the present paper the so-called *prefix*, *suffix* and *bifix* codes play an noticeable part: a code $X \subseteq A^*$ is prefix (suffix) if $X \cap XA^+ = \emptyset$ ($X \cap A^+X = \emptyset$). A code is bifix if it is both prefix and suffix.

A code $X \subseteq A^*$ is *maximal* if it is not strictly included in some other code of A^*. Given a set $X \subseteq A^*$, it is *complete* if $A^* = F(X^*)$; X is *thin* if $A^* \neq F(X)$. Regular codes are well known examples of thin codes [1, Proposition 2.5.20].

A *positive Bernoulli distribution* is a morphism π from the free monoid A^* onto the multiplicative monoid $[0, 1]$, such that we have $\pi(a) > 0$ for every $a \in A$, and such that $\sum_{a \in A} \pi(a) = 1$. The *uniform* distribution corresponds to $\pi(a) = 1/|A|$, for every letter a. For any subset X of A^*, we set $\pi(X) = \sum_{x \in X} \pi(x)$. Clearly, the last sum may be finite or not, however if X is a thin subset we have $\pi(X) < \infty$ [1, Proposition 2.5.12]; moreover for every code $X \subseteq A^*$, we have $\pi(X) \leq 1$. From this point of view, the following result was established by Shützenberger (eg. [1, Theorem 2.5.16]):

Theorem 2.1. *Given a thin code $X \subseteq A^*$, the four following conditions are equivalent:*
(i) *X is complete.*
(ii) *X is a maximal code.*
(iii) *A positive Bernoulli distribution π exists such that $\pi(X) = 1$.*
(iv) *For any positive Bernoulli distribution π, we have $\pi(X) = 1$.*

2.3. (Anti-)automorphisms

In the whole paper, we fix an alphabet A and a mapping θ onto A^* which is either an *automorphism* or an *anti-automorphism*: it is an anti-automophism if it is one-to-one, with $\theta(\varepsilon) = \varepsilon$ and $\theta(xy) = \theta(y)\theta(x)$, for any pair of words x, y. For short in any case we write that θ is an *(anti-)automorphism*.

We say that the (anti-)automorphism θ is of *finite order* if some positive integer k exist such that $\theta^k = id_{A^*}$, the smallest one being the so-called *order* of θ (trivially id_{A^*} is of order 1). It is well known that such a condition is satisfied whenever A is a finite set; in particular, over a two letter alphabet, any non-trivial (anti-)automorphism is of order 2 that is, it is *involutive*.

In the whole paper, we are interested in the family of sets $X \subseteq A^*$ that are invariant under θ (θ-*invariant* for short) that is, which satisfy $\theta(X) \subseteq X$; the mapping θ being one-to-one, this is equivalent to $\theta(X) = X$.

Example 1. Let $A = \{a, b, c, d\}$. Consider the (unique) anti-automorphism θ that is defined by $\theta(a) = a, \theta(b) = b, \theta(c) = d, \theta(d) = c$. It is straightforward to verify that the mapping θ is involutive, moreover the sets $\{cd\}$ and $\{abcd, cdba\}$ are θ-invariant.

Remark 1. In the spirit of the families of codes that were introduced in [15], given an (anti-)automorphism θ, define a θ-*code* as a set X such that $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$ is a code. Clearly, with this definition any θ-code is a code; the converse is false, as attested below by Example 2.

Actually, any θ-code that is a maximal code, is necessarily θ-invariant. Indeed, assuming X not θ-invariant, we have $X \subsetneq X \cup \theta(X)$, thus X is strictly included in the code $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$.

A similar argument proves that if X is maximal as a θ-code, then it is θ-invariant (indeed, $\bigcup_{i\in\mathbb{Z}} \theta^i(X)$ itself is a θ-code).

Taking account of the fundamental importance of the concept of maximality in the theory of codes, such properties reinforces the relevance of the notion of θ-invariant code.

Example 2. Let $A = \{a, b\}$ and θ be the so-called *mirror antimorphism*: $\theta(a) = a$, $\theta(b) = b$. Take for X the finite (prefix) code $\{a, ba\}$. We have $X \cup \theta(X) = \{a, ab, ba\}$, which is not a code ($ab \cdot a = a \cdot ba$).

3. A defect effect for invariant sets

We start with some considerations about θ-invariant submonoids of A^*. Clearly the intersection of a non-empty family of θ-invariant free submonoids of A^* is itself a θ-invariant free submonoid. Given a submonoid M of A^*, recall that its *minimal generating set* is $(M \setminus \{\varepsilon\}) \setminus (M \setminus \{\varepsilon\})^2$. The following property holds:

Proposition 3.1. *Given an alphabet A and given an (anti-)automorphism θ of A^*, let M be a submonoid of A^* and let $S \subseteq A^*$ such that $M = S^*$. Then the two following properties hold:*

(i) *If S is θ-invariant then the same holds for M.*

(ii) *If S is the minimal generating set of M and if M is θ-invariant then S is θ-invariant.*

Proof. (i) Assume that the set S is θ-invariant, and let $w \in M$. Since $M = S^*$, a finite sequence of words in S, namely $(s_i)_{1\leq i\leq n}$, exists such that $w = s_1 \cdots s_n$. Since θ is an (anti-)automorphism, in every case $\theta(w)$ is some concatenation of the words $\theta(s_i)$ $(1 \leq i \leq n)$, therefore we have $\theta(w) \in S^* = M$. Consequently M is θ-invariant.

(ii) Assume that M is θ-invariant and let $s \in S$. It follows from $S \subseteq M$ that we have $\theta(s) \in \theta(M) = M$ therefore, a sequence of words in S, namely $(s_i)_{1\leq i\leq n}$, exists such that $\theta(s) = s_1 \cdots s_n$. Since θ is an (anti-)automorphism, s is in fact some concatenation of the words $\theta^{-1}(s_1), \cdots, \theta^{-1}(s_n) \in M$. Moreover, for each integer $i \in [1, n]$, we have $\theta^{-1}(s_i) = s_i^1 \cdots s_i^{n_i}$, with $s_i^j \in S$ $(1 \leq j \leq n_i)$. It follows from the definition of S that we have $n = 1$ and $s = s_1^1 = \theta^{-1}(s_1)$, thus $\theta(s) = s_1 \in S$. As a consequence, S itself is θ-invariant. □

Informally, the famous defect theorem says that if some words in a set X satisfy a non-trivial equation, then these words can be written upon an alphabet of smaller size. In this section, we will examine whether a corresponding result may be stated in the framework of θ-invariant sets.

Theorem 3.2. *Given an alphabet A and given an (anti-)automorphism θ of A^*, let $X \subseteq A^*$ be a θ-invariant set. Let Y be the minimal generating set of the smallest θ-invariant free submonoid of A^* that contains X. If X is not a code, then we have $|Y| \leq |X| - 1$.*

With the notation of Theorem 3.2, since Y is a code, each word $x \in X$ has a unique factorization upon the words of Y, namely $x = y_1 \cdots y_n$, with $y_i \in Y$ $(1 \leq i \leq n)$. In a classical way, we say that y_1 (y_n) is the *initial* (*terminal*) factor of x (with respect to such a factorization). From this point of view, before to prove Theorem 3.2, we need to establish the following statement:

Lemma 3.3. *With the preceding notation, each word in Y is the initial (terminal) factor of some word in X.*

Proof. By contradiction, assume that a word $y \in Y$ that is never initial of any word in X exists. Set $Z_0 = (Y \setminus \{y\})\{y\}^*$ and $Z_i = \theta^i(Z_0)$, for each integer $i \in \mathbb{Z}$. In a classical way (see eg. [18, p. 7]), since Y is a code, Z_0 itself is a code.
For each integer $i \in \mathbb{Z}$, since θ^i is itself an (anti-)automorphism, Z_i is a code that is, Z_i^* is a free submonoid of A^*. Consequently, the intersection, say M, of the family $(Z_i^*)_{i\in\mathbb{Z}}$ is itself a free submonoid of A^*.
Let $w \in M$. For each integer $i \in \mathbb{Z}$, we have $w \in Z_i^*$, thus $\theta(w) \in \theta(Z_i^*) \subseteq (\theta(Z_i))^* = (\theta^{i+1}(Z_0))^* = Z_{i+1}^*$. Consequently we have $\theta(w) \in \bigcap_{i\in\mathbb{Z}} Z_{i+1}^* = \bigcap_{i\in\mathbb{Z}} Z_i^* = M$, whence we have $\theta(M) \subseteq M$ therefore, since θ is onto, we obtain $\theta(M) = M$.
Let x be an arbitrary word in X. Since $X \subseteq Y^*$, and according to the definition of y, we have $x = (z_1 y^{k_1})(z_2 y^{k_2}) \cdots (z_n y^{k_n})$, with $n \geq 1$, $z_1, \cdots z_n \in Y \setminus \{y\}$ and $k_1, \cdots k_n \geq 0$. Consequently x belongs to Z_0^*, therefore we have $X \subseteq Z_0^*$. Since X is θ-invariant, this implies $X = \theta^i(X) \subseteq \theta^i(Z_0^*) \subseteq Z_i^*$, for each $i \in \mathbb{Z}$, thus $X \subseteq M$.
But the word y belongs to Y^* and does not belong to Z_0^* thus, it doesn't belong to M. This implies $X \subseteq M \subsetneq Y^*$: a contradiction with the minimality of Y^*. Clearly, similar arguments may be applied to words $y \in Y$ that are never terminal of any word in X: this completes the proof. □

Proof of Theorem 3.2. Let α be the mapping from X onto Y which, with every word $x \in X$, associates the initial factor of x in its (unique) factorization over Y^*. According to Lemma 3.3, α is onto. We will prove that it is not one-to-one. Classically, since X is not a code, a non-trivial equation may be written among its words, say: $x_1 \cdots x_n = x'_1 \cdots x'_p$, with $x_i, x'_j \in X$ $x_1 \neq x'_1$ $(1 \leq i \leq n, 1 \leq j \leq p)$. Since Y is a code, a unique sequence of words in Y, namely $y_1, \cdots, y_m$ $(m \geq 1)$ exists such that: $x_1 \cdots x_n = x'_1 \cdots x'_p = y_1 \cdots y_m$. This implies $y_1 = \alpha(x_1) = \alpha(x'_1)$ and completes the proof. □

In what follows we discuss some interpretation of Theorem 3.2 with regards to equations in words. For this purpose, we assume that A is finite, θ being of order k, and we consider a finite set of words, say Z. Let X be the union of the sets $\theta^i(Z)$, for $i \in [0, k-1]$, and assume that a non-trivial equation holds among the words of X, namely $x_1 \cdots x_m = y_1 \cdots y_p$. By construction X is θ-invariant therefore, according to Theorem 3.2, a θ-invariant code Y exists such that $X \subseteq Y^*$, with $|Y| \leq |X| - 1$. This means that each of the words in X can be expressed by making use of at most $|X| - 1$ words of type $\theta^i(u)$, with $u \in Y$ and $0 \leq i \leq k - 1$. It will be easily verified that the examples from [5, 16, 21] corroborate this fact; moreover, below we mention an original one:

Example 3. Let θ be an anti-automorphism of order 3. Consider two different words x, y, with $|x| > |y| > 0$, satisfying the equation: $x\theta(y) = \theta^2(y)\theta(x)$. With this condition, a pair of words u, v exists such that $x = uv$, $\theta(x) = v\theta(y)$, $\theta^2(y) = u$, thus $y = \theta(u)$. It follows

from $x = uv$ that $v\theta(y) = \theta(x) = \theta(v)\theta(u)$, thus $v = \theta(v)$ and $\theta(y) = \theta(u)$. This implies $y = u = \theta(u) = \theta^2(u)$ and $v = \theta(v) = \theta^2(v)$. Moreover, we have $\theta(x) = vu$, $\theta^2(x) = uv$: we obtain $x = \theta^2(x)$ thus, $x = \theta(x) = \theta^2(x)$; hence we have $uv = vu$. Consequently, a non-empty word t and integers i, j exist such that $u = t^i$, $v = t^j$. With the preceding notation, we have $Z = \{x, y\}$, $X = Z \cup \theta(Z) \cup \theta^2(Z) = \{x, y\}$, $Y = \{t\}$. We verify that $|Y| \leq |X| - 1$.

4. Finite complete θ-invariant codes

In this section we are interested in finite complete θ-invariant codes over an alphabet A. Given an arbitrary letter $a \in A$, since for every non-negative integer n, we have $a^n \in F(X^*)$, necessarily a (unique) positive integer p exists such that $a^p \in X$; therefore, A is necessarily finite. Several examples of finite complete θ-invariant codes will be presented. We start with prefix codes, which certainly constitute the best-known class of them.

4.1. Finite complete prefix θ-invariant codes

Actually finite complete prefix codes play a peculiar part in the framework of codes. A famous result due to Schützenberger [25] (cf. also [2]) states that any finite complete code with a *finite deciphering delay* (eg. [1, Ch. 5]) is necessarily prefix. In particular, over A^* only one finite complete *circular* code (or, equivalently, finite complete *uniformly synchronized code*) can exist, namely the alphabet A itself (cf. [1, Ch. 7, Ch. 10], [17]).

It is well-known that each prefix set, say X, can be represented by a tree, say $\mathcal{T}(X)$, of arity $|A|$: in this representation, each node (i.e. vertice) is a prefixes of some word in X (i.e. the elements of $P(X)$), the root being ε, the empty word. Moreover, given two nodes u, v and a letter $a \in A$, an edge with label a exists from u to v in $\mathcal{T}(X)$ if, and only if, we have $v = ua$: we denote such a labelled edge by (u, a, v) and we say that v is a sucessor of u. In that representation, complete prefix codes correspond to *complete trees*, in the sense where each interior node has exactly $|A|$ successors.

We start with the case where θ is an automorphism of A^*. Given a prefix set $X \subseteq A^*$, we say that the corresponding tree $\mathcal{T}(X)$ is *invariant* under θ whenever (u, a, v) is an edge of $\mathcal{T}(X)$ if, and only if, $(\theta(u), \theta(a), \theta(v))$ is an edge of $\mathcal{T}(X)$. With this notion, a characterization of θ-invariant prefix codes may be stated:

Claim 1. *Let A be a finite alphabet, let θ be an automorphism of A^* and let X be a prefix code. Then X is θ-invariant if, and only if, the tree $\mathcal{T}(X)$ itself is invariant under θ.*

Proof. Assume that X is θ-invariant, and let (u, a, ua) an arbitrary edge in $\mathcal{T}(X)$. By construction a word $s \in S(X)$ exists such that $uas \in X$. Since X is a θ-invariant set, this implies $\theta(u)\theta(as) = \theta(ua)\theta(s) \in X$, thus $\theta(u)$ and $\theta(ua) \in P(X)$. Consequently, $(\theta(u), \theta(a), \theta(u)\theta(a))$ is an edge of $\mathcal{T}(X)$, therefore $\mathcal{T}(X)$ is invariant under θ.

Conversely, assume that $\mathcal{T}(X)$ is invariant under θ. Let $w = w_1 \cdots w_n \in X$, with $w_i \in A$ $(1 \leq i \leq n)$. By construction, the following sequence of edges exists in $\mathcal{T}(X)$ (for $i = 0$, we set $w_1 \cdots w_i = \varepsilon$):

$$(w_1 \cdots w_i, w_{i+1}, w_1 \cdots w_{i+1}) \quad (0 \leq i \leq n-1),$$

moreover the node $w = w_1 \cdots w_n$ has no sucessor. Since $\mathcal{T}(X)$ is invariant under θ, a corresponding sequence of edges exists in $\mathcal{T}(X)$, namely:

$$(\theta(w_1) \cdots \theta(w_i), \theta(w_{i+1}), \theta(w_1) \cdots \theta(w_{i+1})) \quad (0 \leq i \leq n-1).$$

Since the node $w_1 \cdots w_n$ has no sucessor, the same holds for the corresponding node $\theta(w) = \theta(w_1 \cdots w_n)$: this implies $\theta(w) \in X$. □

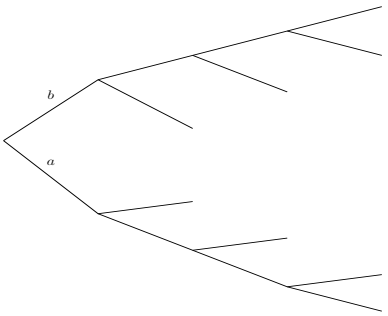

Figure 1: Example 4 with $n = 4$. In the tree $\mathcal{T}(X)$, each bottom-up (top-down) branch represents an edge with label b (a).

Example 4. Let $A = \{a, b\}$, and θ be the automorphism defined by $\theta(a) = b$, $\theta(b) = a$. Given an arbitrary integer $n \geq 3$, consider the following set:

$$X = \bigcup_{1 \leq i \leq n-1} \{a^i b, b^i a\} \cup \{a^n, b^n\}.$$

By construction, X is a prefix code. Moreover, X is complete: this can be directly verified by examining $\mathcal{T}(X)$ (an alternative method consists in applying Theorem 2.1 (iii), with π the uniform Bernoulli distribution). It is also straightforward to verify that X is θ-invariant.

Note that X is not bifix: indeed, for each integer $i \in [2, n-1]$, the word $ab \in X$ is a suffix of $a^i b \in X$. Figure 1 illustrates the corresponding tree $\mathcal{T}(X)$ for $n = 4$.

In the case where θ is an anti-automorphism, the following property is noticeable:

Claim 2. *Let θ be an anti-automorphism onto A^* and let $X \subseteq A^*$ be a finite θ-invariant code. If X is prefix, then it is necessarily bifix.*

Proof. By contradiction, assume X not bifix, thus not suffix: words $p \in A^*$, $s \in A^+$ exist such that $s, ps \in X$. Since X is θ-invariant, we have $\theta(s), \theta(ps) \in X$, thus $\theta(s), \theta(s)\theta(p) \in X$: this contradicts the fact that X is a prefix code. □

The result of Claim 2 directly leads to examine the behavior of finite complete bifix codes with regards to (anti-)automorphisms.

4.2. Finite complete bifix θ-invariant codes

At first, it is worth mentioning a well-known class of finite bifix codes:

Example 5. A set X is *uniform* if a positive integer n exists such that $X \subseteq A^n$. Trivially, such a set is a bifix code moreover, it is complete if, and only if, we have $X = A^n$. It is straightforward to verify that X is invariant under every (anti-)automorphism of A^*: indeed, the restriction of such a mapping on words of length n induces a permutation of A^n.

It is a natural question to ask whether non-uniform finite complete bifix θ-invariant codes exist. By exhibiting infinite classes of convenient codes, the three following examples allow to bring a positive answer. Actually, the two first families of codes have been constructed by applying a famous internal transformation to some uniform code [4] (cf. also [1, § 6.2])

Example 6. Let $A = \{a, b\}$ and θ be the anti-automorphism of A^* that is defined by $\theta(a) = b$, $\theta(b) = a$.
Let $n = 2k + 1$, with $k \geq 1$. Consider the following set:

$$X = (A^n \setminus (Aa^kb^k \cup a^kb^kA)) \cup \{a^kb^k\} \cup Aa^kb^kA.$$

The set Aa^kb^kA is a (uniform) bifix code. Since the condition $a^kb^k \in P(X)$ ($a^kb^k \in S(X)$) necessarily implies $a^kb^k \in P(a^kb^kA)$ ($a^kb^k \in S(Aa^kb^k)$), X is a finite (non-uniform) bifix code. The code X is complete: indeed, we have $Aa^kb^k \cap a^kb^kA = \emptyset$ therefore, given an arbitrary positive Bernoulli distribution π over A^*, we have:

$$\pi(X) = \pi(A^n \setminus (Aa^kb^k \cup a^kb^kA)) + \pi(a^kb^k) + \pi(Aa^kb^kA) = 1 - 2\pi(a^kb^k) + 2\pi(a^kb^k) = 1.$$

Furthermore, since we have $\theta(A) = A$ and $\theta(a^kb^k) = a^kb^k$, X is θ-invariant.
For $n = 3$, the preceding construction leads to the following finite complete bifix θ-invariant code [4, (1)]:

$$X = \{a^3, ba^2, b^2a, b^3, ab, a^2ba, a^2b^2, baba, bab^2\}.$$

Example 7. Let $A = \{a, b\}$ and θ be the so-called mirror-image, which is in fact the anti-automorphism defined by $\theta(a) = a$, $\theta(b) = b$.
Take $n = 3k + 1$, with $k \geq 1$. We have $a^kb^ka^k \notin P(Aa^kb^ka^k) \cup S(a^kb^ka^kA)$; therefore an examination similar to the one we applied at Example 6 leads to verify that the following set is a finite complete bifix θ-invariant code:

$$X = (A^n \setminus (Aa^kb^ka^k \cup a^kb^ka^kA)) \cup \{a^kb^ka^k\} \cup Aa^kb^ka^kA.$$

For $n = 4$ (i.e. $k = 1$) the corresponding binary tree $\mathcal{T}(X)$ is represented in Figure 2.

We observe that, in view of constructing arbitrarily large non-uniform finite bifix θ-invariant codes over arbitrarily large finite alphabets, the two last constructions can be generalized, as illustrated by the following example:

Example 8. 1) Let $A = \{a, b, c\}$ and θ be the anti-automorphism defined by $\theta(a) = b$, $\theta(b) = c$, $\theta(c) = a$.

Take $n = 2k + 1$, with $k \geq 1$, and

$$W = \bigcup_{i \in \mathbb{Z}} \{\theta^i(a^kb^k)\} = \{a^kb^k, c^kb^k, c^ka^k, b^ka^k, b^kc^k, a^kc^k\}.$$

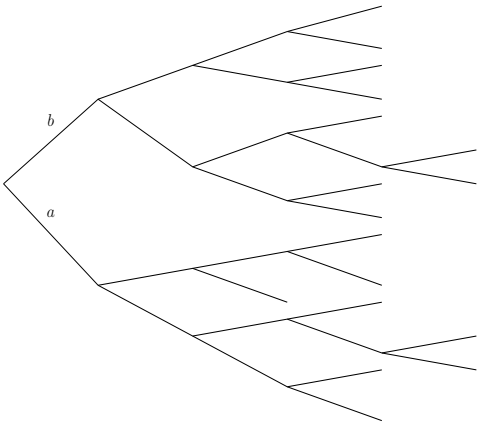

Figure 2: Example 7: the case where $n = 4$, thus $k = 1$.

By construction we have $W \cap (P(AW) \cup S(WA)) = \emptyset$ therefore, the following set is a θ-invariant finite bifix code:

$$X = (A^n \setminus (AW \cup WA)) \cup W \cup AWA. \tag{1}$$

Moreover, X is complete: indeed, by construction we have $AW \cap WA = \emptyset$ therefore, for any positive Bernoulli distribution over A^*, we have $\pi(X) = 1 - 2\pi(W) + 2\pi(W) = 1$.

2) Similarly, over $A = \{a, b, c\}$ take for θ the anti-automorphism onto A^* defined by $\theta(a) = b$, $\theta(b) = c$, $\theta(c) = a$. Set $n = 3k + 1$ and

$$W = \bigcup_{i \in \mathbb{Z}} \{\theta^i(a^k b^k a^k)\} = \{a^k b^k a^k, b^k c^k b^k, c^k a^k c^k\}.$$

Applying the construction from (1) also leads to obtain a finite complete bifix θ-invariant code.

4.3. Non-prefix non-suffix finite complete θ-invariant codes

In the most general case, given an (anti-)automorphism θ of A^*, we are looking to finite codes $X \subseteq A^*$ which are both complete and θ-invariant. With regards to the last condition, the following statement brings some characterization:

Claim 3. *Let θ be an (anti-)automorphism onto A^* and let X be a finite subset of A^*. Then X is θ-invariant if, and only if, it is the disjoint union of a finite family of uniform θ-invariant codes.*

Proof. Let $\ell_1 < \cdots < \ell_n$ be the unique increasing finite sequence of the lengths of the words in X. For each $i \in [1, n]$, set $X_i = X \cap A^{\ell_i}$. By construction, each set X_i is a uniform code, moreover we have:

$$X = \bigcup_{1 \leq i \leq n} X_i.$$

Clearly, the set X is θ-invariant if, and only if, for each integer $i \in [1, n]$, θ induces a permutation of X_i itself. □

When X is a required to be a code, Claim 3 only leads to some necessary condition. For instance, the set $\{a, ab, b\} = \{a, b\} \cup \{ab\}$, which satisfies the condition of the claim, is θ-invariant, but clearly it is not a code. Actually, despite that in any case θ-invariance is preserved with respect to the union of sets, the main obstacle is that, given two (disjoint) codes, there is no characterization that can guarantee that their union remains a code.

Of course, one can wonder about the impact of θ-invariance itself on the structure of a finite complete code. Indeed, in view of the above, such an influence is very strong with regards to two special families of codes: the uniform ones and, with respect to automorphisms, the family of prefix non-suffix codes. However, the part of θ-invariance appeared in fact of lesser importance in the construction of our families of bifix codes, where it essentially involved the structure of a few convenient words (eg. the elements of W).

Things become even more complex when attempting to construct finite complete θ-invariant codes that are neither prefix nor suffix. Indeed, with regards to finite complete codes, although that some famous families have been exhibited (eg. [11, 12]), no general structure is known. However, finite complete θ-invariant codes that are neither prefix nor suffix exist as attested by the following example:

Example 9. With the anti-automorphism θ that was introduced in Example 6 (which swaps the letters a and b), consider the classical finite complete code $X = \{a^2, ab, a^2b, ab^2, b^2\}$ [24, Example 2], which is neither prefix, nor suffix. It is straightforward to verify that it is θ-invariant (we have $\theta(ab) = ab$).

4.4. Toward the construction of regular complete θ-invariant codes

In [23], by making use of factorizations of the so-called cyclotomic polynomials, the author provided a family of non-finitely completable codes. It is therefore a natural question to ask whether corresponding objects exist in the framework of θ-invariant codes.

Let A be a finite alphabet, and let θ be an (anti)-automorphism of A^*. Given a finite code X, if X is embeddable into a complete θ-invariant code, say Y, then, with the terminology of Remark 1, it has to be a θ-code. Indeed the set $\bigcup_{i \in \mathbb{Z}} \theta^i(X)$ is necessarily a θ-invariant code that is included in Y. Therefore, our problem comes down to wonder whether a given finite θ-invariant code can be embedded into a complete one.

We begin by strictly restraining the problem to the framework of prefix codes. Given a (non-trivial) automorphism θ, according to the preceding Claim 1 any θ-invariant prefix

code can be embedded into a θ-invariant complete one. Informally, if suffices to complete the corresponding tree with convenient ones of arity $|A|$ that are invariant under θ.

In the case of anti-automorphisms, according to Claim 2, for being embeddable into a complete one, a finite prefix θ-invariant code has to be bifix. However, the converse is false; indeed there are finite bifix θ-invariant codes that cannot be included into any complete one, as attested by the following example:

Example 10. 1) Let $A = \{a, b\}$, and be θ be the mirror anti-automorphism of Example 7. At first, we observe that the finite θ-invariant bifix code $X = \{aa, b\}$ cannot embedded into any finite complete bifix (not necessarily θ-invariant) code. Indeed, assume that such a complete code, say Z, exists: necessarily Z is prefix and complete, hence for any positive integer p, we have $ab^p \in P(Z^*)$. Therefore a positive integer n exists such that ab^n belongs to Z; since b belongs to Z this contradicts the fact that Z is bifix.
As a consequence X cannot be included in any finite complete prefix θ-invariant code. Indeed, according to Claim 2, such a code should be bifix.

2) Note that the infinite (regular) set $Z = \{b\} \cup \{ab^n a : n \in \mathbb{N}\}$ is a θ-invariant bifix code which contains X. Moreover, taking for π the uniform Bernoulli distribution, it is straighforward to verify that we have $\pi(Z) = 1/2+1/4\sum_{n\in\mathbb{N}}(1/2^n) = 1$, thus Z is complete.

We do not know whether there are finite θ-invariant complete codes that contain the code X of Example 10. Actually, as far as we know, the question of embedding a finite θ-invariant code into a complete one remains open.

From another angle, the study in [23] led its author to conclude that the study of all finite codes requires also investigations on the infinite ones. From that, the question of embedding a finite code into a regular one was open. A positive answer was given in [10], where a now famous method for embedding a regular code into a complete one was published.

From this last point of view, in the next section, we will interest in the problem of embedding a regular θ-invariant code into a regular complete one.

5. Embedding a regular θ-invariant code into a complete one

5.1. Some notation

In this section we consider an (anti-)automorphism θ of A^*, and a non-complete θ-invariant code $X \subseteq A^*$. We ask for a complete regular θ-invariant code Y such that $X \subseteq Y$. We will bring a positive answer: let's begin by describing our construction.

Let X be a non-complete θ-invariant code, and let $y \notin F(X^*)$. Necessarily, we have $|A| \geq 2$ (otherwise, X should be complete). Without loss of generality, we may assume that the initial and the terminal letters of y are different (otherwise, substitute to y the word $ay\overline{a}$, with $a, \overline{a} \in A$ and $a \neq \overline{a}$): in particular, we have $|y| \geq 2$. Set:

$$z = \overline{a}^{|y|} y a^{|y|} \quad (\text{with} \quad y \in aA^*\overline{a}). \tag{2}$$

Since θ is an (anti-)automorphism, for each integer $i \in \mathbb{Z}$, two different letters $b, \overline{b}$ exist such that the following property holds:

$$\theta^i(z) = \overline{b}^{|y|}\theta^i(y)b^{|y|} \quad (\text{with} \quad \theta^i(y) \in bA^*\overline{b}). \tag{3}$$

Finally, we introduce the three following sets:

$$Z = \bigcup_{i\in\mathbb{Z}} \{\theta^i(z)\}, \tag{4}$$

$$W = ZA^* \cap A^*Z, \tag{5}$$

$$T = W \setminus (W \cup X)(W \cup X)^+. \tag{6}$$

By construction, the following inclusion holds:

$$W \subseteq (X \cup T)^+. \tag{7}$$

5.2. Basic properties of Z

By construction, each element of the preceding set Z has length $3|y|$. Given two (not necessarily different) integers $i, j \in \mathbb{Z}$, we will accurately study how the two words $\theta^i(z), \theta^j(z)$ may overlap.

Lemma 5.1. *With the notation in (3), let $u, v \in A^+$, $i, j \in \mathbb{Z}$ such that $|u| \leq |z| - 1$ and $\theta^i(z)v = u\theta^j(z)$. Then we have $|u| = |v| \geq 2|y|$, moreover a letter b and a unique positive integer k (depending of $|u|$) exist such that we have $\theta^i(z) = ub^k$, $\theta^j(z) = b^k v$, with $k \leq |y|$.*

Proof. According to (3), we set $\theta^i(z) = \overline{b}^{|y|}bx'\overline{b}b^{|y|}$ and $\theta^j(z) = \overline{c}^{|y|}cx''\overline{c}c^{|y|}$, with $b, \overline{b}, c, \overline{c} \in A$, $b \neq \overline{b}, c \neq \overline{c}$ and $|x'| = |x"| = |y| - 2$. Since θ is an (anti-)automorphism, we have $|\theta^i(z)| = |\theta^j(z)|$, thus $|u| = |v|$; since we have $1 \leq |u| \leq 3|y| - 1$, exactly one of the following cases occurs:
Case 1: $1 \leq |u| \leq |y| - 1$. With this condition, we have $(\theta^i(z))_{|u|+1} = \overline{b} = \overline{c} = (u\theta^j(z))_{|u|+1}$ and $(\theta^i(z))_{|y|+1} = b = \overline{c} = (u\theta^j(z))_{|y|+1}$, which contradicts $b \neq \overline{b}$.
Case 2: $|u| = |y|$. This condition implies $(\theta^i(z))_{|u|+1} = b = \overline{c} = (u\theta^j(z))_{|u|+1}$ and $(\theta^i(z))_{2|y|} = \overline{b} = \overline{c} = (u\theta^j(z))_{2|y|}$, which contradicts $b \neq \overline{b}$.
Case 3: $|y| + 1 \leq |u| \leq 2|y| - 1$. We obtain $(\theta^i(z))_{2|y|} = \overline{b} = \overline{c} = (u\theta^j(z))_{2|y|}$ and $(\theta^i(z))_{2|y|+1} = b = \overline{c} = (u\theta^j(z))_{2|y|+1}$ which contradicts $b \neq \overline{b}$.
Case 4: $2|y| \leq |u| \leq |z| - 1 = 3|y| - 1$. With this condition, necessarily we have $(\theta^i(z))_{|u|+1} = b = \overline{c} = (u\theta^j(z))_{|u|+1}$, therefore an integer $k \in [1, |y|]$ exists such that $\theta^i(z) = ub^k$ and $\theta^j(z) = b^k v$. □

Lemma 5.2. *With the preceding notation, we have $A^+ZA^+ \cap ZX^*Z = \emptyset$.*

Proof. By contradiction, assume that $z_1, z_2, z_3 \in Z$, $x \in X^*$ and $u, v \in A^+$ exist such that $uz_1v = z_2xz_3$. By comparing the lengths of u, v with $|z|$, exactly one of the three following cases occurs:
Case 1: $|z| \leq |u|$ and $|z| \leq |v|$. With this condition, we have $z_2 \in P(u)$ and $z_3 \in S(v)$, therefore the word z_1 is a factor of x: this contradicts $Z \cap F(X^*) = \emptyset$.
Case 2: $|u| < |z| \leq |v|$. We have in fact $u \in P(z_2)$ and $z_3 \in S(v)$. We are in the condition of Lemma 5.1: the words z_2, z_1 overlap. Consequently, $u, z_1' \in A^+$ and $b \in A$ exist such that $z_2 = ub^k$ and $z_1 = b^k z_1'$, with $1 \leq k \leq |y|$ and $|z_1'| = |u|$. But, by construction, we have $|uz_1| = |z_2xz_3| - |v|$. Since we assume $|v| \geq |z|$, this implies $|uz_1| \leq |z_2xz_3| - |z| = |z_2x|$, hence we obtain $uz_1 = ub^k z_1' \in P(z_2x)$. It follows from $z_2 = ub^k$ that $z_1' \in P(x)$. Since we have $z_1 \in Z$ and according to (3), $i \in \mathbb{Z}$ and $\overline{b} \in A$ exist such that we have $z_1 = b^k z_1' = b^{|y|}\theta^i(y)\overline{b}^{|y|}$. Since by Lemma 5.1 we have $|z_1'| = |u| \geq 2|y|$, we obtain $\theta^i(y) \in F(z_1')$, thus $\theta^i(y) \in F(x)$, which contradicts $y \notin F(X^*)$.
Case 3: $|v| < |z| \leq |u|$. Same arguments on the reversed words lead to a conclusion similar to that of Case 2.
Case 4: $|z| > |u|$ and $|z| > |v|$. With this condition, both the pairs of words z_2, z_1 and z_1, z_3

overlap. Once more we are in the condition of Lemma 5.1: letters c, d, words u, v, s, t, and integers h, k exist such that the two following properties hold:

$$z_2 = uc^h, \quad z_1 = c^h s, \quad |u| = |s| \geq 2|y|, \quad h \leq |y|, \tag{8}$$

$$z_1 = td^k, \quad z_3 = d^k v, \quad |v| = |t| \geq 2|y|, \quad k \leq |y|. \tag{9}$$

It follows from $uz_1v = z_2xz_3$ that $uz_1v = (uc^h)x(d^kv)$, thus $z_1 = c^hxd^k$. But according to (3), $i \in \mathbb{Z}$ and $\overline{c} \in A$ exist such that we have $z_1 = c^{|y|}\theta^i(y)\overline{c}^{|y|}$. Since we have $h, k \leq |y|$, this implies $d = \overline{c}$ moreover $\theta^i(y)$ is a factor of x. Once more, this contradicts $y \notin F(X^*)$. □

As a direct consequence of Lemma 5.2, we obtain the following result:

Corollary 5.3. *With the preceding notation, X^*Z is a prefix code.*

Proof. Let $z_1, z_2 \in Z$, $x_1, x_2 \in X^*, u \in A^+$, such that $x_1z_1u = x_2z_2$. For any word $z_3 \in Z$, we have $(z_3x_1)z_1(u) = z_3x_2z_1$, a contradiction with Lemma 5.2. □

5.3. The consequences for the set $X \cup T$

Lemma 5.4. *The set $X \cup T$ is a θ-invariant code.*

Proof. The fact that $X \cup T$ is θ-invariant comes from its construction. For proving that it is a code, we consider an arbitrary equation among the words in $X \cup T$. Since X is a code, and since $z \notin F(X^*)$, we may assume that at least one occurrence of a word in T appears in each side of the equation, therefore this equation takes the following form:

$$x_0t_0x_1t_1 \cdots t_{n-1}x_n = x'_0t'_0 \cdots t'_{p-1}x'_p, \tag{10}$$

with $x_i, x'_j \in X^*$ $(0 \leq i \leq n,\ 0 \leq j \leq p)$ and $t_i, t'_j \in T$ $(0 \leq i \leq n-1,\ 0 \leq j \leq p-1)$. Since by construction we have $T \subseteq W \subseteq ZA^*$, each side of the equation has a prefix in X^*Z. According to Corollary 5.3 and since all the words in Z have a common length, this implies $x_0 = x'_0$, therefore our equation is equivalent to:

$$t_0x_1t_1 \cdots t_{n-1}x_n = t'_0x'_1 \cdots t'_{p-1}x'_p. \tag{11}$$

Without loss of generality, we assume that $|t_0| \leq |t'_0|$; let k be the greatest non-negative integer such that a word s exists with $t_0x_1t_1 \cdots x_kt_ks = t'_0$, with $s \in A^*$. By contradiction, we assume $s \neq \varepsilon$. Let $z_0 \in Z$ $(z_1 \in Z)$ be the unique word such that $t'_0 \in A^*z_0$ $(t_k \in A^*z_1)$. According to the preceding property (3), an integer $i \in \mathbb{Z}$ and two letters $b, \overline{b}$ exist such that $z_0 = \overline{b}^{|y|}\theta^i(y)b^{|y|}$. Moreover, since we have $y \notin F(X^*)$, and since X is θ-invariant, we have $\theta^i(y) \notin F(X^*)$. By construction, the set with elements $t_0x_1 \cdots t_kx_{k+1}$ and t'_0 is not prefix; more precisely, exactly one of the two following main conditions holds:

1. At first, we assume that $t'_0 \in P(t_0x_1 \cdots t_kx_{k+1})$ that is, $s \in P(x_{k+1})$ (cf. Figure 5). By construction, at least one of the two words s, z_0 is a suffix of the other one. Actually, since we have $z_0 \notin F(X^*)$, necessarily s is a proper suffix of z_0, therefore $(z_1,\ z_0)$ is an overlapping pair of words. According to Lemma 5.1, necessarily we have $|s| \geq 2|y|$, which implies $\theta^i(y) \in F(s)$: a contradiction with $\theta^i(y) \notin F(X^*)$.

2. Now, we assume that $t_0x_1 \cdots t_kx_{k+1}$ is a proper prefix of t'_0, thus we have $s = x_{k+1}s_1$, with $s_1 \neq \varepsilon$. Let $z_2 \in Z$ be the unique word such that $t_{k+1} \in z_2A^*$. By construction the set with elements $t_0x_1 \cdots t_kx_{k+1}z_2$ and t'_0 is not prefix. More precisely, exactly one of the two following cases occurs:

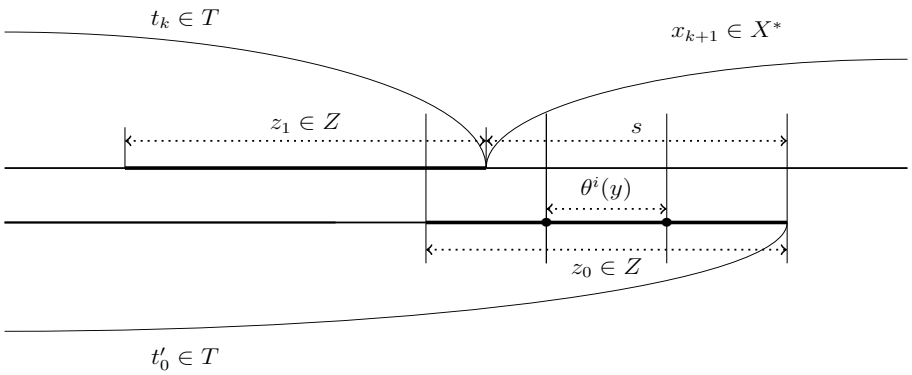

Figure 3: Proof of Lemma 5.4 - the case where $s \in P(x_{k+1})$.

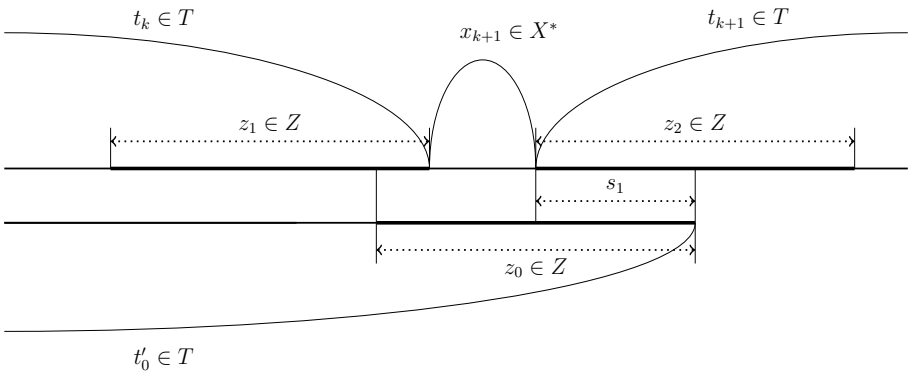

Figure 4: Proof of Lemma 5.4 -the case where $x_{k+1} \in P(s)$ and $s_1 \in P(z_2) \setminus \{z_2\}$.

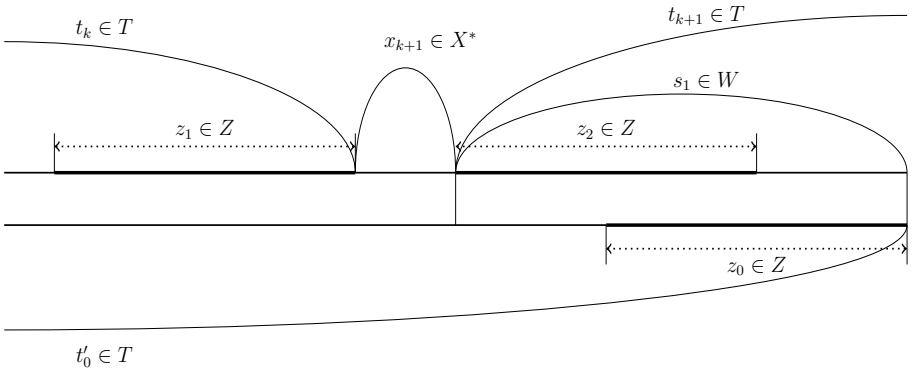

Figure 5: The case where $x_{k+1} \in P(S)$, with $z_2 \in P(s_1)$ in the proof of Lemma 5.4.

2.1 The first case corresponds to t'_0 being a proper prefix of $t_0x_1 \cdots t_kx_{k+1}z_2$, that is s_1 being a proper prefix of z_2 (cf. Figure 6). With this condition, the word z_0 is necessarily a factor of $z_1x_{k+1}z_2$. According to Lemma 5.2, since we have $s_1 \neq \varepsilon$, this implies $z_1 = z_0$, which contradicts $s \neq \varepsilon$.

2.2 It remains to consider the case where z_2 is a prefix of s_1 (cf. Figure 7). Actually, since $t_0x_1 \cdots t_kx_{k+1}z_2$ is a prefix of $t_0x_1 \cdots t_kx_{k+1}s_1 = t'_0$, with $|z_2| = |z_0|$, necessarily z_0 is a suffix of s_1, hence we have $s_1 \in z_2A^* \cap A^*z_0$, thus $s_1 \in W$ according to (5). We obtain $t'_0 = t_0x_1 \cdots t_kx_ks_1 \in (TX^*)^+W$, thus $t'_0 \in (T \cup X)^+W$: this contradicts (6).

In each case we obtain a contradiction: as a consequence we have $s = \varepsilon$, thus $t'_0 = t_0x_1 \cdots x_kt_k$. Once more according to (6), it follows from $t'_0 \in T$ that we have $k = 0$, thus $t'_0 = t_0$. As a consequence, Equation (11) is equivalent to the following one:

$$x_1t_1 \cdots t_{n-1}x_n = x'_1 \cdots t'_{p-1}x_p. \tag{12}$$

By iterating these arguments, we shall obtain: $n = p$ and $x_i = x'_i$, $t'_j = t_j$ $(0 \leq i \leq n$, $0 \leq j \leq n-1)$, therefore $X \cup T$ is a code: this completes the proof of Lemma 5.4. □

Lemma 5.5. *The code $X \cup T$ is complete.*

Proof. Let $w \in A^*$. According to the construction of W we have $ZwZ \subseteq W$. According to (7) this implies $ZwZ \subseteq (X \cup T)^*$, therefore we have $w \in F((X \cup T)^*)$. □

In the case where A is a finite alphabet, the (anti-)automorphism θ is of finite order. If X is a regular code, in starting with $y \notin F(X^*)$, the construction in (4) leads to a finite set Z: this guarantees the regularity of the sets W and T. As a direct consequence, we obtain the following result:

Theorem 5.6. *Given a non-complete θ-invariant code $X \subseteq A^*$, the two following properties hold:*

(i) In any case, X can be embedded into a complete θ-invariant code in A^*.

(ii) If A is finite and X regular, then X can be embedded into a regular complete θ-invariant code in A^*.

Example 11. Let $A = \{a, b\}$, and θ be the anti-automorphism such that $\theta(a) = b$, $\theta(b) = a$, and let $X = \{a^4, a^2b^2, a^2b^4, a^4b^2, ba, ba^4, b^4a, b^4\}$
Trivially, X is θ-invariant. By applying Sardinas-Patterson algorithm [1, § 2.3], one can easily verify that X is a (non-prefix) code. It is non complete: by making use of the uniform Bernoulli distribution π, we obtain $\pi(X) < 1$.
Let $y = ba^3ba$. Firstly, we note that we have $y \notin F(X)$, hence $y \in F(X^*)$ implies $y = sp$, with $s \in S(X)$ and $p \in P(X^*)$. Secondly, we have $S(X) \cap P(y) = \{b, ba\}$, but since $\{a^3ba, a^2ba\} \cap P(X^*) = \emptyset$, necessarily we have $y \notin F(X^*)$. Thirdly, in view of obtaining an overlapping-free word, we substitute $by = b^2a^3ba$ to y (we have $by \notin F(X^*)$.
With the notation (2, 4), we have $z = a^7b^2a^3bab^7$, thus:
$Z = \bigcup_{i\in\mathbb{Z}}\{\theta^i(z)\} = \{a^7b^2a^3bab^7, a^7bab^3a^2b^7\}$. Moreover, the sets W and T shall be constructed according to (5,6).

Example 11 provides a (non-finite) regular complete θ-invariant code; in the sequel we give an example of a non-regular one:

Example 12. Let $A = \{a, b\}$, and θ be an arbitrary (anti-)automorphism of A^*. Consider the famous Dyck language $D_1^* = \{w \in A^* : |w|_a = |w|_b\}$. Each of its elements is classically represented by a so-called Dyck path in the grid $\mathbb{N} \times \mathbb{Z}$. To be more precise, with each word $w = w_1 \cdots w_n$ (with $w_i \in A$, for $1 \leq i \leq n$), a unique path is associated, namely $(i, y_i)_{0\leq i\leq n}$, with $y_0 = y_n = 0$ and such that, for each $i \in [1, n]$:

$$w_i = a \Longrightarrow y_i = y_{i-1} + 1 \quad \text{and} \quad w_i = b \Longrightarrow y_i = y_{i-1} - 1.$$

By construction, D_1^* is a free submonoid of A^*. Its minimal generating set is the so-called Dyck code D_1, whose elements are represented by those of the preceding non-empty paths which satisfy the following condition:

$$n \geq 2 \quad \text{and} \quad (\forall i \in [1, n-1]) \;\; y_i \neq 0.$$

The code D_1 is well known for being a (non-thin) complete context-free language (eg. [1, Example 2.5.3]). Moreover, according to Proposition 3.1, since D_1^* is θ-invariant, the same holds to the Dyck code.

As a consequence of Theorem 5.6, we obtain the following result, which states a property similar to [1, Theorem 2.5.16] in the framework of θ-invariant code:

Theorem 5.7. *Given a thin θ-invariant code $X \subseteq A^*$, the following conditions are equivalent:*
(i) *X is complete.*
(ii) *X is a maximal code.*
(iii) *X is maximal in the family θ-invariant codes.*
(iv) *A positive Bernoulli distribution π exists such that $\pi(X) = 1$.*
(v) *For any positive Bernoulli distribution π, we have $\pi(X) = 1$.*

Proof. According to [1, Theorem 2.5.16], the conditions (i), (ii) , (iv), (v) are equivalent. Trivially, Condition (ii) implies Condition (iii). By contradiction, we prove that Condition (iii) implies Condition (i). Starting with a non-complete θ-invariant code X, according to Theorem 5.6 the existence of a complete θ-invariant code that strictly contains X is guaranteed, thus X is not maximal in the family of θ-invariant codes: this completes the proof. □

Aknowledgement

We would like to thank the anonymous reviewers for their fruitful suggestions and comments.

[1] J. Berstel, D. Perrin, and C. Reutenauer. *Codes and Automata*. Cambridge University Press, 2010.

[2] V. Bruyère. Automata and codes with a bounded deciphering delay. In I. Simon, editor, *LATIN'92*, volume 583, pages 99–107. Lect. Notes in Comp. Sci., 1992.

[3] M. Bucci, A. de Luca, A. De Luca, and L. Q. Zamboni. On θ-episturmian words. *Eur. J. of Comb.*, 30:473–479, 2009.

[4] V. Césari. Sur un algorithme donnant les codes bipréfixes finis. *Theory of Computing Systems*, 26:221–225, 1972.

[5] E. Czeizler, E. Czeizler, L. Kari, and S. Seki. An extension of the Lyndon-Schützenberger result to pseudoperiodic words. *Inf. Comput.*, 209:717–730, 2011.

[6] E. Czeizler, L. Kari, and S. Seki. On a special class of primitive words. *Theoret. Comp. Sci.*, 411:617–630, 2010.

[7] C. Annal Deva Priya Darshini, V. Rajkumar Dare, I. Venkat., and K.G. Subramanian. Factors of words under an involution. *J. of Math. and Inf.*, 1:52–59, 2013–2014.

[8] J.D. Day, P. Fleishmann, F. Manea F., and D. Nowotka. Equations enforcing repetitions under permutations. In S. Brlek, F. Dolce, C. Reutenauer, and E. Vandomme, editors, *Combinatorics on Words*, volume 10432, pages 72–84. Lect. Notes in Comp. Sci., 2017.

[9] A. de Luca and A. De Luca. Pseudopalindrome closure operators in free monoids. *Theoret. Comp. Sci.*, 362:282–300, 2006.

[10] A. Ehrenfeucht and S. Rozenberg. Each regular code is included in a regular maximal one. *RAIRO - Theor. Inform. Appl.*, 20:89–96, 1986.

[11] C. De Felice. Construction of a family of finite maximal codes. *Theoret. Comput. Sci.*, 63:157–184, 1989.

[12] C. De Felice and A. Restivo. Some results on finite maximal codes. *RAIRO - Theoret. Informatics and Appl*, 9:383–403, 1985.

[13] N.J. Fine and H. S. Wilf. Uniqueness theorem for periodic functions. *Proc. Am. Math. Soc.*, 16:109–114, 1965.

[14] P. Gawrychowski, F. Manea, R. Mercas, D. Nowotka, and C. Tiseanu. Finding pseudo-repetitions. In N. Portier and T. Wilke, editors, *30th International Symposium on Theoretical Aspects of Computer Science (STACS 2013)*, volume 20 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 257–267, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[15] L. Kari and K. Mahalingam. Dna codes and their properties. In C. Mao and T. Yokomori, editors, *12th International Meeting on DNA Computing (DNA12)*, volume 4287, pages 127–142. Lect. Notes in Comp. Sci., june 2006.

[16] L. Kari and K. Mahalingam. Watson-Crick conjugate and commutative words. In M.H. Garzon and H. Yan, editors, *DNA Computing. DNA 2007*, volume 4848, pages 273–283. Lect. Notes in Comp. Sci., 2008.

[17] J.-L Lassez. Circular codes and synchronization. *Internat. J. Computer Syst. Sci.*, 5:201–208, 1976.

[18] M. Lothaire. *Combinatorics on Words*. Addison-Wesley Publishing Company (2nd edition Cambridge University Press 1997), 1983.

[19] M. Lothaire. *Algeraic Combinatorics on Words*. Cambridge University Press, 2002.

[20] F. Manea, R. Mercas, and D. Nowotka. Fine and Wilf's theorem and pseudo-repetitions. In B. Rovan, V. Sassone, and P. Widmayer, editors, *Acts of Mathematical Foundations of Computer Science 2012 (MFCS 2012)*, volume 7464, pages 668–680. Lect. Notes in Comp. Sci., 2012.

[21] F. Manea, M. Müller, D. Nowotka, and S. Seki. Generalised Lyndon-Schützenberger equations. In E. Csuhaj-Varjú, M. Dietzfelbinger, and Z. Ésik, editors, *Mathematical Foundations of Computer Science 2014 (MFCS 2014)*, volume 8634, pages 402–413. Lect. Notes in Comp. Sci., 2014.

[22] J. Néraud and C. Selmi. Invariance: a theoretical approach for coding sets of words modulo literal (anti)morphisms. In S. Brlek, F. Dolce, C. Reutenauer, and E. Vandomme, editors, *Combinatorics on Words*, volume 10432, pages 214–227. Lect. Notes in Comp. Sci., 2017.

[23] A. Restivo. On codes having no finite completion. *Discr. Math.*, 17:309–316, 1977.

[24] A. Restivo, S. Salemi, and T. Sportelli. Completing codes. *RAIRO - Theoret. Informatics and Appl.*, 23:135–147, 1989.

[25] M.-P. Schützenberger. On a question concerning certain free submonoids. *J. Combin. Theory*, 1:437–442, 1966.