Aix-Marseille Université
École doctorale en Mathématiques et Informatique
Institut de Mathématiques de Marseille

Exceptional trajectories in the symbolic dynamics of multidimensional continued fraction algorithms

Trajectoires exceptionnelles dans la dynamique symbolique d'algorithmes de fractions continues multidimensionnelles.

Mélodie Andrieu

Thèse de doctorat en mathématiques

Composition du jury :

Pierre Arnoux	Aix-Marseille Université	codirecteur de thèse
Valérie Berthé	CNRS et Université de Paris	examinatrice
Julien Cassaigne	CNRS et Aix-Marseille Université	codirecteur de thèse
Karma Dajani	University of Utrecht	rapportrice
Fabien Durand	Université de Picardie Jules Verne	examinateur
Thomas Fernique	CNRS et Université Sorbonne Paris Nord	examinateur
Anna Frid	Aix-Marseille Université	examinatrice
Michel Rigo	Université de Liège	rapporteur.

Abstract

In this thesis, I develop a semi-algorithm consisting of an automaton, or rather, an ever-building family of automata, whose states contain all the information on the imbalances of the words belonging to a S-adic system. In particular, this semi-algorithm can be used as an exploration tool, giving strong intuitions for an upper bound for the imbalances of words in the system; or, when no such bound exists, helps in understanding where the imbalance grows. Thanks to this tool, I spotted families of C-adic words (the class of words associated with Cassaigne-Selmer multidimensional continued fraction algorithm) with arbitrary high imbalances; from which I constructed a C-adic word with infinite imbalance. Much stronger: I constructed Arnoux-Rauzy and C-adic words whose Rauzy fractals are unbounded in all directions of the plane. This result is conflicting with the intuition given by the Oseledets theorem on Lyapunov exponents.

On another hand, I introduce and discuss a topological definition of natural coding of a minimal rotation on the d-dimensional torus, inspired by the seminal works of Rauzy on the Tribonacci word. In particular, under the axiom of choice, it is possible to wisely complete the pseudo-fundamental domain of the torus into a fundamental domain, while preserving the property of piecewise translation and a weak form of sequential continuity. Under this good definition, being a natural coding of a minimal rotation of the d-torus passes through induction and its reverse process: exduction. As a consequence, (1) no uniformly recurrent tree word with infinite imbalance is a natural coding of a minimal rotation of the torus, with bounded fundamental domain; (2) being a natural coding of a minimal rotation of the 2-torus, both for Arnoux-Rauzy and C-adic words, is a property that only depends on the asymptotic behavior of the directive sequence.

At last, in collaboration with Anna Frid, we study the numerical morphic sequences associated with substitutive words.

Keywords: multidimensional continued fraction algorithm, symbolic dynamics, S-adic system, imbalance, natural coding of rotation, induction, discrepancy.

Résumé

Dans cette thèse, nous introduisons un semi-algorithme fondé sur l'exploration d'un automate, ou plutôt, d'une famille d'automates en construction, dont les états contiennent toute l'information sur le déséquilibre des mots d'un système S-adique. Ce semi-algorithme peut être détourné afin de conjecturer des majorants pour l'ensemble des déséquilibres des mots du système ou, lorsque ceux-ci ne sont pas bornés, pour comprendre d'où proviennent les grands déséquilibres. Nous utilisons cet outil pour identifier une famille de mots C-adiques (i.e. les mots associés à l'algorithme de fraction continue de Cassaigne-Selmer) dont le déséquilibre n'est pas borné, à partir de laquelle nous construisons un mot C-adique de déséquilibre infini. Plus remarquables encore, nous construisons des mots d'Arnoux-Rauzy et des mots C-adiques dont le fractal de Rauzy n'est borné dans aucune direction du plan : ce résultat contredit l'intuition donnée par le théorème d'Oseledets sur les exposants de Lyapounov.

D'autre part, nous proposons une définition topologique du codage naturel de rotation minimale du tore en dimension d, inspirée de l'article originel de Rauzy sur le mot de Tribonacci. Sous l'axiome du choix, nous complétons le pseudo-domaine fondamental, tout en préservant l'échange de morceaux et, sous une forme affaiblie, la continuité du codage. Grâce à cette définition, la propriété d'être codage naturel est préservée par induction et son processus réciproque : l'exduction. Nous montrons qu'alors, (1) aucun mot dendrique uniformément récurrent de déséquilibre infini n'est un codage naturel de rotation minimale du tore, pour un domaine fondamental borné ; (2) le fait d'être ou non un codage naturel, pour un mot d'Arnoux-Rauzy et un mot C-adique, dépend uniquement du comportement asymptotique de la suite directrice.

Enfin, en collaboration avec Anna Frid, nous étudions des suites numériques morphiques associées aux mots substitutifs.

Mots-clés : algorithme de fraction continue multidimensionnelle, dynamique symbolique, système S-adique, déséquilibre, codage naturel de rotation, induction, discrépance.

Table of contents

This dissertation is made up of an introduction and five chapters, which can be read independently. Each chapter corresponds to an existing or intended article. Chapters 1 to 4 focus on the symbolic dynamical systems generated by multidimensional continued fraction algorithms. Chapter 5, which results of a collaboration with Anna Frid, deals with morphic numerical sequences associated with substitutive words.

References

[1] Andrieu, M. Autour du déséquilibre des mots C-adiques (text in French). *Proceedings of Mons Days of Theoretical Computer Science* (2018), 51–54.

[2] Andrieu, M. Natural coding of minimal rotations of the torus, induction and exduction. *to be submitted* (2021).

[3] Andrieu, M. A Rauzy fractal unbounded in all directions of the plane. *to appear in Comptes Rendus de l'Académie des Sciences* (2021).

[4] Andrieu, M., and Frid, A. E. Morphic words and equidistributed sequences. *Theoretical Computer Science 804* (2020), 171 – 185. Online software: `https://www.i2m.univ-amu.fr/perso/anna.frid/MorphismsOnReals/mp.html`, accessed: 2021-01-01.

Introduction

Contents

1 General background

1.1 Motivations

To the continued fraction algorithm, which consists in the (infinite) iteration of the Farey map:

$$\begin{array}{rcll} (\mathbb{R}^+)^2 & \to & (\mathbb{R}^+)^2 & \\ (x,y) & \mapsto & (x-y,y) & \text{if } x \geq y, \\ & & (x,y-x) & \text{otherwise,} \end{array}$$

is associated a remarkable class of infinite binary words, called Sturmian words. A *word* is a finite or infinite sequence of elements (*letters*) picked in a finite set (*alphabet*). Sturmian words enjoy numerous characterizations in combinatorics, number theory and geometry (see [Lot97] for a general introduction). For instance, they are exactly the aperiodic words with minimal complexity (the *complexity* of a word w is the map which counts the number of different factors of length n appearing in w; a *factor* of w is a subword of w formed with consecutive letters). Equivalently, Sturmian words are the aperiodic binary words with imbalance equal to 1, i.e. those in which all factors of same length contain, up to one, the same number of 0s (and thereby, up to one again, the same number of 1s). For instance, a word starting with $w = 0010001001000100001001\dots$ could be Sturmian, whereas a word starting with $w = 011011100...$ is definitively not, since it contains both factors 11 and 00. This property implies in particular that the letters 0 and 1 are uniformly distributed with respect to a probability measure ν on $\{0,1\}$, and that the difference between the Birkhoff sum $1/N \sum_{n=0}^{N-1} \mathbb{1}_{\{0\}}(w[n])$, which quantifies the proportion of 0s observed among the N first letters of w, and its expected value $\nu(0)$ (called *frequency* of 0s) is bounded above by $1/N$. From a geometrical standpoint, this means that the points $P_N := \sum_{n=0}^{N} e_{w[n]}$, where (e_0, e_1) denotes the canonical basis of $\mathbb{R}^2$, remain at a bounded distance from the line carried by the frequency vector $(\nu(0), \nu(1))$. The sequence $(P_N)_{N\in\mathbb{N}}$ is called the *broken line* of w. In computer science, Sturmian broken lines are used to discretize straight lines with irrational slope.

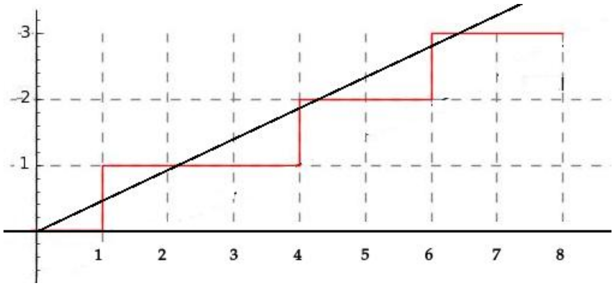


Figure 1: The broken line of 01000100100...

Since the work of Jacobi, several algorithms have been proposed to generalize continued fractions to triplets of positive numbers (see for instance [Sch00]). They are expected to yield good and simultaneous approximations of pairs of numbers.

A general question is: does the dynamics of these algorithms enjoy similar properties than the regular continued fractions? For instance, do the 3-d broken lines of the words they generate remain at bounded distance from their average directions? Is there a uniform bound for this distance? Numerically, such results would ensure that all approximations given by the algorithm converge (with possibly a minimal speed).

My work focuses on the Arnoux-Rauzy and Cassaigne-Selmer algorithms, respectively introduced in 1991 and 2017 in [AR91] and [CLL17], because they generate classes of words with complexity 2n+1, which is the lowest we can expect for a continued fraction algorithm running on triplets of numbers. The Cassaigne-Selmer algorithm has the additional advantage to be defined on all inputs, whereas the Arnoux-Rauzy algorithm is defined on almost none (its set of admissible inputs, called *Rauzy Gasket*, is studied in [AS13], [AHS16]).

Arnoux-Rauzy and Cassaigne-Selmer algorithms consist in the infinite iteration of the maps F_{AR} and F_C respectively:

$$\begin{array}{rcccl} F_{AR}: & (\mathbb{R}^+)^3 & \to & (\mathbb{R}^+)^3 & \\ & (x,y,z) & \mapsto & (x-y-z,y,z) & \text{if } x > y+z, \\ & & & (x,y-x-z,z) & \text{if } y > x+z, \\ & & & (x,y,z-x-y) & \text{if } z > x+y; \end{array}$$

and

$$\begin{array}{rcccl} F_C: & (\mathbb{R}^+)^3 & \to & (\mathbb{R}^+)^3 & \\ & (x,y,z) & \mapsto & (x-z,z,y) & \text{if } x \geq z, \\ & & & (y,x,z-x) & \text{otherwise.} \end{array}$$

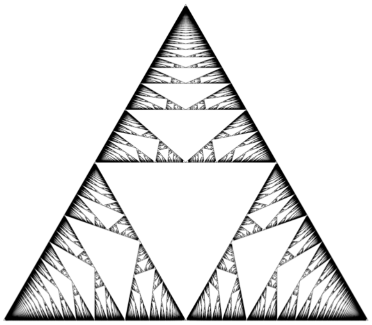


Figure 2: The Rauzy Gasket.

1.2 Some definitions

Let A be an alphabet. A *substitution* is an application mapping letters to finite words: $A \mapsto A^*$, that is extended into a morphism on A^* (the set of finite words) on one hand, and on $A^{\mathbb{N}}$ (the set of infinite words) on the other hand. For instance, the *Thue-Morse substitution* is defined over the alphabet $A = \{1,2\}$ by $\sigma_{TM}(1) = 12$ and $\sigma_{TM}(2) = 21$. Observe that $(\sigma_{TM})^2(1) = \sigma_{TM}(12) = 1221$, $(\sigma_{TM})^3(1) = 12212112$, ..., and that actually the sequence of the iterated images of 1 'converges' to an infinite word $w_{TM} = 1221211221121221211212211221...$ (called *Thue-Morse word*). Generally, the infinite words obtained as the limit of iterated images by a substitution are said *substitutive*. This construction is enlarged by allowing the substitution to differ at each step. Let S be a set of substitutions defined over a common alphabet A. An infinite word $w \in A^{\mathbb{N}}$ is *S-adic* if there exist a *directive sequence* $(s_n) \in S^{\mathbb{N}}$, together with a letter $a \in A$ such that the limit of finite words $(s_0 \circ ... \circ s_{n-1}(a))_{n\in\mathbb{N}}$ converges to w.

Example - definition 1. *Consider $S_{AR} := \{\sigma_1, \sigma_2, \sigma_3\}$, with:*

$$\begin{array}{lcclccclccclccl} \sigma_1: & 1 & \to & 1 & ; & \sigma_2: & 1 & \to & 21 & \textit{and} & \sigma_3: & 1 & \to & 31 \\ & 2 & \to & 12 & & & 2 & \to & 2 & & & 2 & \to & 32 \\ & 3 & \to & 13 & & & 3 & \to & 23 & & & 3 & \to & 3. \end{array}$$

An infinite word is an Arnoux-Rauzy word *if and only if it has the same set of factors than a S_{AR}-adic word whose directive sequence contains infinitely many occurrences of σ_1, σ_2 and σ_3.*

The set of substitutions S_{AR} has been constructed such that the symbolic trajectory of the vector of letter frequencies $(\nu(1), \nu(2), \nu(3))$ of an Arnoux-Rauzy word, with respect to the piecewise definition of F_{AR}, is exactly its directive sequence. *Episturmian words* are the natural generalization of Arnoux-Rauzy words to larger alphabets.

Example - definition 2. *We call* C-adic words *the S-adic words generated by the set of substitutions $C := \{c_1, c_2\}$ (associated with F_C):*

$$c_1 : \begin{array}{rcl} 1 & \mapsto & 1 \\ 2 & \mapsto & 13 \\ 3 & \mapsto & 2 \end{array} \qquad \textit{and} \qquad c_2 : \begin{array}{rcl} 1 & \mapsto & 2 \\ 2 & \mapsto & 13 \\ 3 & \mapsto & 3. \end{array}$$

The *imbalance* of an infinite word w is the quantity (possibly infinite):

$$\operatorname{imb}(w) = \sup_{n\in\mathbb{N}} \sup_{u,v\in\mathcal{F}_n(w)} \max_{a\in A} ||u|_a - |v|_a|,$$

where $\mathcal{F}_n(w)$ denotes the set of factors of length n of w, and $|u|_a$ the number of occurrences of the letter a in the (finite) word u. One sees that imbalance measures inequities in the distribution of letters in a given word. In fact, it is the combinatorial counterpart of the discrepancy function, which is the difference between the frequency of letters and their effective proportions in growing prefixes of w.

Proposition 3 ([Ada03])**.** *If w is an infinite word with vector of letter frequencies f_w, then:*

$$\frac{1}{4}\operatorname{imb}(w) \leq \sup_{n\in\mathbb{N}} ||\operatorname{ab}(p_n(w)) - nf_w||_\infty \leq \operatorname{imb}(w).$$

2 Contributions to the study of multidimensional continued fraction algorithms

2.1 A semi-algorithm to explore the set of imbalances in S-adic systems

As already mentioned, Sturmian words are exactly aperiodic binary words with imbalance equal to 1 [CH73]. Does such a property still hold for continued fractions algorithms in higher dimensions? Are there some algorithms for which the words produced have an imbalance bounded by 1? or by a greater constant? or for which the imbalance is always finite?

On one hand, words with imbalance equal to 1 have been fully described in [Hub00]. On the other hand, Cassaigne, Ferenczi and Zamboni for Arnoux-Rauzy algorithm - and later Delecroix, Hejda and Steiner for Brun algorithm - constructed, by *ad hoc* techniques, words with infinite imbalances ([CFZ00], [DHS13]).

Question 1 (status: solved) - Do C-adic words have finite imbalance? If it is the case, are these imbalances uniformly bounded?

In Chapter 1, I describe families of C-adic words with arbitrary high imbalances. Then, by a pumping process, I construct the directive sequence of a C-adic word with infinite imbalance.

Theorem A (Chapter 1)**.** *For all nonnegative integer n, there exists a finite sequence $\boldsymbol{s}$= $(s_i)_{0\leq i\leq k}$ of substitutions in C such that all C-adic words w whose directive sequence starts with $\boldsymbol{s}$ satisfy* $\operatorname{imb}(w) \geq n$*.*

Theorem B (Chapter 1)**.** *There exists a C-adic word whose imbalance is infinite.*

In Chapter 2, I formalize the method which led me to these results, in a much broader framework. I develop a semi-algorithm consisting of an automaton, or rather, an ever-building family of automata, whose states contain all the information on the imbalances of the words belonging to a S-adic system, thus proving that the property of having an imbalance bounded by D is semi-decidable. In particular, this semi-algorithm can be used as an exploration tool, giving strong intuitions for an upper bound for the imbalances of words in the system; or, when no such bound exists, helps in exhibiting families of words with growing imbalance.

Theorem C (Chapter 2)**.** *Let S denote a finite set of nonerasing substitutions over a common alphabet A, and assume that all letters in A appear in a word (not necessarily the same) in a S-adic word. If D_S denotes the quantity (possibly infinite):*

$$D_S = \sup_{w\ S\text{-adic}} \operatorname{imb}(w),$$

then a breadth first search in the automaton of imbalances, from its initial states, outputs, for any $d \leq D_S$, a S-adic word whose imbalance is greater than d.

In practice, this tool is difficult to use because of its exponential algorithmic complexity. Nonetheless, with some adjustments, it worked for the S-adic systems associated with Cassaigne-Selmer algorithm (Theorem A) and Arnoux-Rauzy algorithm (I rediscovered the families described in [CFZ00]).

2.2 Imbalance and natural coding of rotations

Sturmian words with slope α are the symbolic trajectories of the linear flow on the torus $\mathbb{R}/\mathbb{Z}$ given by the vector $(1, \alpha)$, with respect to a remarkable partition (Ω_1, Ω_2). This partition is such that each of the two pieces, once lifted to $[0, 1)$, is an interval; and the translation on the torus, once lifted, coincides with the exchange of these two intervals of $[0, 1)$. This property was first generalized by Rauzy in [Rau82] in the case of the Tribonacci word, and expected to remain true for all Arnoux-Rauzy words, although no general definition was proposed.

In [CFZ00], it is claimed that infinite imbalance is an obstruction to natural coding of rotation.

Claim 4. *[CFZ00] Let w be an Arnoux-Rauzy word with infinite imbalance. Then either w or one of its three derivated words is not a natural coding of a rotation of the torus.*

Unfortunately, the proof sketched relies on a mistake and the definition of natural coding proposed is inadequate (this is discussed in Chapter 3 with the agreement of two of the authors).

Question 2 (status: partially solved) - Can we fix that argument? Is infinite imbalance an obstruction to natural coding?

In Chapter 3, I introduce and discuss a topological definition of natural coding of a minimal rotation on the d-dimensional torus, inspired by the seminal work of Rauzy on the Tribonacci word. In particular, I show that under the axiom of choice, it is possible to wisely complete the pseudo-fundamental domain of the torus into a fundamental domain, while preserving the property of piecewise translation and a weak form of sequential continuity. An enjoyable consequence is that the natural coding of rotation is stable under the induction process.

Theorem D (Chapter 3)**.** *If w is a natural coding of a minimal rotation of the d-torus, and if w admits $d+1$ return words to a letter a, then its derivated word to the letter a is also a natural coding of a minimal rotation of the d-torus, that we can fully describe.*

In particular, Theorem D completes the argument of Cassaigne, Ferenczi and Zamboni: under this assumption, *and if furthermore the fundamental domain associated with the natural coding is bounded*, then the cylinder $[a]$ is a bounded remainder set for w (i.e. the empiric frequency with which the symbolic trajectory w visits the set $[a]$ tends to its expected value at speed at least $1/n$), which is equivalent to finite imbalance on the letter a. As a consequence:

Theorem E (Chapter 3)**.** *No Arnoux-Rauzy word with infinite imbalance is a natural coding of a minimal rotation of the* 2*-dimensional torus, with bounded fundamental domain.*

The same holds for primitive C-adic words and, more generally, for uniformly recurrent tree words.

Theorem E has been shown independently by Steiner in a late version of [Thu19], with a simpler proof, which relies exclusively on the boundedness of the fundamental domain.

A natural question is to determine whether Theorem E remains valid if we relax the hypothesis of boundedness of the fundamental domain. Indeed, the relevant fundamental domain for the Tribonacci word is given by its Rauzy fractal (i.e. the closure of the projection of its broken line); and the words with infinite imbalance are precisely those whose Rauzy fractal is unbounded.

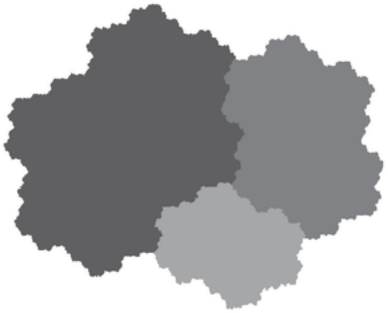


Figure 3: The Rauzy fractal of the Tribonacci word

By showing that the property of being a natural coding of a minimal rotation passes through *exduction* (a reciprocal operation to induction), and by studying the S-adic expression of return words, I obtained that being a natural coding of a minimal rotation of the 2-torus, both for Arnoux-Rauzy and C-adic words, is a property that only depends on the asymptotic behavior of the directive sequence.

Theorem F (Chapter 3)**.** *For Arnoux-Rauzy and primitive C-adic subshifts, the property of being a natural coding of a minimal rotation of the 2-torus does not depend on any prefix of the directive sequence.*

2.3 Construction of a Rauzy fractal unbounded in all directions of the plane.

The assumption of boundedness in Theorem E is inherited from a theorem of Rauzy on induction and bounded remainder sets [Rau84]. At this point, it is natural to look where the boundedness is used in Rauzy theorem proof, and examine if we can bypass it. In the proof of Rauzy, one needs a nonzero linear form to be bounded on (what will be later) the fundamental domain of the natural coding. This requires the fundamental domain to be *at least* trapped between two parallel hyperplanes.

Hence the question: what can be said about the geometric shape of Rauzy fractals of Arnoux-Rauzy words with infinite imbalance? As we said, they are not bounded; but are they spread over the whole plane, or trapped between two parallel lines? The Oseledets theorem suggests that the latter option is true. Indeed, if the Lyapunov exponents of the matrix product associated with a word exist, one of these exponents at least is nonpositive since their sum equals zero.

Question 3 (status: solved) Are all Rauzy fractals trapped between two parallel lines?

In Chapter 4, I prove that the intuition given by the Oseledets theorem is wrong.

Theorem G (Chapter 4)**.** *There exists an Arnoux-Rauzy word w_∞ whose Rauzy fractal is unbounded in all directions of the plane.*

I explicitly describe the directive sequence of such an Arnoux-Rauzy word. This construction relies on a thorough study of the automaton of imbalances for the set of substitutions S_{AR} associated with Arnoux-Rauzy words. Similar methods yield similar result for C-adic words on one hand, and for strict episturmian words on the other hand.

2.4 The Arnoux-Rauzy multidimensional continued fraction algorithm detects all kinds of rational dependencies.

Question 4 (status: solved) What can be said of the vector of letter frequencies of the remarkable word w_∞? For instance, are its entries rationally independent?

I actually obtained a much stronger result:

Theorem H (Chapter 4)**.** *The vector of letter frequencies of any Arnoux-Rauzy word has rationally independent entries.*

Theorem H was conjectured by Arnoux and Starosta in 2013 [AS13]. It has been independently proved by Dynnikov, Hubert and Skripchenko in [DHS]. My method extends to arbitrary dimension:

Theorem I (Chapter 4)**.** *Let $d \geq d' \geq 1$. Let w an episturmian word over A_d. Denote by $f = (f_1, ..., f_d)$ its vector of letter frequencies, and by $(s_n)_{n\in\mathbb{N}}$ one of its directive sequences. The following assertions are equivalent.*

1. *Exactly d' substitutions appear infinitely many times in $(s_n)_{n\in\mathbb{N}}$.*
2. *The dimension of the linear space $f_1\mathbb{Q} + ... + f_d\mathbb{Q}$ is d'.*

Theorem I says that the (generalized) Arnoux-Rauzy multidimensional continued fraction algorithm detects all kinds of rational dependencies. This is an enjoyable property for a continued fraction algorithm.

3 Contributions to the study of substitutive words

In collaboration with Anna Frid.

Problem: given an infinite word w on an ordered alphabet, construct the sequence $\nu_w = (\nu[n])_n$, equidistributed on $[0, 1]$, and such that $\nu[m] < \nu[n]$ if and only if $S^m(w) < S^n(w)$, where S is the *shift map*, which erases the first letter of an infinite word. The sequence ν_w exists and is unique for every word with well-defined positive uniform frequencies of every factor (equivalently: for every element of a uniquely ergodic subshift).

Example 5. *The equidistributed sequence associated with the Thue-Morse word* $w_{TM} = 122121122$ $1121221211212211221...$ *(see Subsection 1.2) is:*

$$\nu_{w_{TM}} = \frac{1}{2}, 1, \frac{3}{4}, \frac{1}{4}, \frac{5}{8}, \frac{1}{8}, \frac{3}{8}, \frac{7}{8}, ...$$

Observe that $\nu_{w_{TM}}$ *is the 'substitutive numerical sequence' arising from the* numerical substitution*:*

$$\begin{array}{cccll} f_{TM}: & [0,1] & \to & [0,1]^* & \\ & x & \mapsto & \begin{cases} \frac{x}{2}+\frac{1}{4}, \frac{x}{2}+\frac{3}{4} & \textit{if } 0 \leq x \leq 1/2, \\ \frac{x}{2}+\frac{1}{4}, \frac{x}{2}-\frac{1}{4} & \textit{otherwise.} \end{cases} \end{array}$$

In Chapter 5, we describe the construction of ν_w for a subclass of substitutive words, which contains all binary substitutive words. The numerical sequence ν_w is, in this case, constructed with a "numerical substitution". I wrote a software in Sage which, given a binary substitution σ, computes the equidistributed sequences ν_w associated with all substitutive words w generated by σ. This software can be tried online [AF20].

4 Future directions

Question 2 (is infinite imbalance an obstruction to natural coding?) is still open. To progress in this direction, it would be of interest to better understand the exceptional trajectories we spotted.

- Is the unbounded Rauzy fractal constructed in [CFZ00] (resp. Chapter 1) unbounded *in all directions* of the plane? Are all unbounded Arnoux-Rauzy (resp. C-adic) Rauzy fractals unbounded in all directions? (Probably not.)
- What can be said of the Lyapunov exponents of the remarkable words of [CFZ00], Chapter 1 and Chapter 4?
- What can be said about the topological properties of their (unbounded) Rauzy fractals? Do they have an inner point? Are they the closure of their interior?
- What can be said about the algebraic properties of the vectors which generate such fractals? Do these directions remain singular for other continued fraction algorithms?
- Can we find rules to pick up convergents that provide good approximations for the exceptional directions? If it is possible, what quality of approximation can we expect?
- Which dynamical systems do the words constructed in [CFZ00], Chapter 1 and Chapter 4 code?

References

[Ada03] Boris Adamczewski. Balances for fixed points of primitive substitutions. *Theoretical Computer Science*, 307:47–75, 2003.

[AF20] Mélodie Andrieu and Anna E. Frid. Morphic words and equidistributed sequences: a demonstration tool. *https://www.i2m.univ-amu.fr/perso/anna.frid/MorphismsOnReals/mp.html*, 2020. Accessed: 2020-01-01.

[AHS16] Artur Avila, Pascal Hubert, and Alexandra Skripchenko. On the Hausdorff dimension of the Rauzy gasket. *Bulletin de la Société Mathématique de France*, 144:539–568, 2016.

[AR91] Pierre Arnoux and Gérard Rauzy. Représentation géométrique de suites de complexité 2n+1. *Bulletin de la Société Mathématique de France*, 119:199–215, 1991.

[AS13] Pierre Arnoux and Štěpán Starosta. The Rauzy Gasket. In *Further Developments in Fractals and Related Fields*, pages 1–23. Springer, 2013.

[CFZ00] Julien Cassaigne, Sébastien Ferenczi, and Luca Q. Zamboni. Imbalances in Arnoux-Rauzy sequences. *Annales de l'Institut Fourier*, 50:1265–1276, 2000.

[CH73] Ethan M. Coven and G. A. Hedlund. Sequences with minimal block growth. *Mathematical systems theory*, 7:138–153, 1973.

[CLL17] Julien Cassaigne, Sébastien Labbé, and Julien Leroy. A set of sequences of complexity 2n+1. In *WORDS 2017 Proceedings*, pages 144–156. Springer, 2017.

[DHS] Ivan Dynnikov, Pascal Hubert, and Alexandra Skripchenko. Dynamical systems around the Rauzy gasket and their ergodic properties. *in preparation.*

[DHS13] Vincent Delecroix, Tomáš Hejda, and Wolfgang Steiner. Balancedness of Arnoux-Rauzy and Brun words. In Juhani Karhumäki, Arto Lepistö, and Luca Zamboni, editors, *Combinatorics on Words*, pages 119–131. Springer Berlin Heidelberg, 2013.

[Hub00] Pascal Hubert. Suites équilibrées. *Theoretical Computer Science*, 242:91–108, 07 2000.

[Lot97] Lothaire. *Combinatorics on Words.* Cambridge Mathematical Library. 2nd edition, 1997.

[Rau82] Gérard Rauzy. Nombres algébriques et substitutions. *Bulletin de la Société Mathématique de France*, 110:147–178, 1982.

[Rau84] Gérard Rauzy. Ensembles à restes bornés. *Séminaire de Théorie des Nombres de Bordeaux*, pages 1–12, 1984.

[Sch00] Fritz Schweiger. *Multidimensional Continued Fractions.* Oxford Science Publications. Oxford University Press, 2000.

[Thu19] Jörg M. Thuswaldner. S-adic sequences: a bridge between dynamics, arithmetic, and geometry. *arXiv:1908.05954*, 2019.

I. Construction of a C-adic word with infinite imbalance

The results of this chapter are published [in French] in the proceedings of the conference Mons Theoretical Computer Science Days 2018.

Table des matières

Autour du déséquilibre des mots C-adiques

Mélodie Andrieu

Résumé

Nous étudions une propriété combinatoire, le déséquilibre, d'une classe particulière de mots sur l'alphabet $\{a, b, c\}$: les mots C-adiques. En particulier, nous exhibons des familles de mots C-adiques de déséquilibres arbitrairement grands, et même des mots C-adiques de déséquilibre infini. Ces constructions ont été obtenues par l'exploration d'un automate et l'étude de ses chemins.

1 Motivations

À l'algorithme de fraction continue soustractif décrit par l'itération de l'application

$$\begin{array}{rcll} (\mathbb{R}^+)^2 & \to & (\mathbb{R}^+)^2 & \\ (x,y) & \mapsto & (x-y,y) & \text{si } x \geq y \\ & & (x,y-x) & \text{sinon} \end{array}$$

est associée une classe particulière de mots infinis binaires : les mots sturmiens. Ceux-ci jouissent de deux caractérisations combinatoires : d'une part, ce sont exactement les mots de complexité $n+1$, c'est-à-dire les mots qui admettent $n+1$ facteurs de longueur n pour tout entier n ; d'autre part, ce sont les mots apériodiques dont le déséquilibre vaut 1, c'est-à-dire les mots apériodiques dans lesquels chaque lettre apparaît, à une unité près, un même nombre de fois dans tous les facteurs d'une longueur donnée.

Plusieurs tentatives ont été faites pour généraliser les fractions continues à des triplets de réels positifs. Un tel algorithme pourrait permettre d'approcher simultanément deux réels par une suite de couples de nombres rationnels.

Dans ce document, nous nous interrogeons sur les mots C-adiques, qui sont les mots ternaires associés à l'algorithme [CLL17] :

$$\begin{array}{rcll} (\mathbb{R}^+)^3 & \to & (\mathbb{R}^+)^3 & \\ (x,y,z) & \mapsto & (x-z,z,y) & \text{si } x \geq z \\ & & (y,x,z-x) & \text{sinon.} \end{array} \tag{1}$$

Tout comme les mots d'Arnoux-Rauzy, ces mots sont de complexité $2n+1$. L'intérêt de cet algorithme est qu'il admet comme instance n'importe quel triplet de réels positifs, contrairement à l'algorithme d'Arnoux-Rauzy qui n'est défini que sur un ensemble de mesure de Lebesgue nulle.

Aussi, il est naturel de s'interroger sur l'existence d'une borne uniforme pour le déséquilibre. Hélas, comme pour les mots d'Arnoux-Rauzy [CFZ00], nous pouvons construire des familles de mots C-adiques de déséquilibre aussi grand que souhaité, et même, par un lemme de pompage, des mots de déséquilibre infini.

2 Déséquilibre, mots C-adiques

Soit u un mot fini sur l'alphabet ternaire $A = \{a, b, c\}$ et $\alpha \in A$ une lettre. On désigne par $|u|_\alpha$ le nombre d'occurrences de la lettre α dans le mot u. Le *vecteur de Parikh* de u est le vecteur $\chi(u) = (|u|_a, |u|_b, |u|_c)$, qui compte les multiplicités de chacune des lettres de l'alphabet. Remarquons que la somme des coordonnées de ce vecteur est égale à la longueur du mot u, que l'on note $|u|$. Étant donné deux mots de même longueur u et v, on appelle *vecteur de déséquilibre de u et v* la différence de leurs vecteurs de Parikh. La somme des coordonnées d'un tel vecteur est nulle. Le déséquilibre d'un mot infini w est la quantité (éventuellement infinie) :

$$d \;=\; \sup_{n\in\mathbb{N}} \; \sup_{u,v\in F_n(w)} \; ||\chi(u) - \chi(v)||_\infty,$$

qui s'écrit encore :

$$d \;=\; \sup_{n\in\mathbb{N}} \; \sup_{u,v\in F_n(w)} \; \max_{\alpha\in\{1,2,3\}} ||u|_\alpha - |v|_\alpha|.$$

Elle mesure les iniquités de répartition entre les lettres dans un mot donné.

Dans toute la suite, on s'intéressera aux substitutions $c_1 : 1 \mapsto 1, 2 \mapsto 13, 3 \mapsto 2$ et $c_2 : 1 \mapsto 2, 2 \mapsto 13, 3 \mapsto 3$, qui proviennent de l'algorithme de fraction continue 1 [CLL17]. On notera aussi $C = \{c_1, c_2\}$.

Un *mot C-adique* est un mot infini de la forme $w = \lim_{n\to\infty} s_1 \circ ... \circ s_n(w')$, où $(s_k)_{k\in\mathbb{N}} \in C^{\mathbb{N}}$ porte le nom de *suite directrice* de w, et où w' est un mot infini quelconque sur A ; avec la condition supplémentaire que chacune des substitutions c_1 et c_2 apparaisse une infinité de fois dans la suite directrice. Remarque : pour S un ensemble de substitutions, on peut étudier les mots S-adiques dans un cadre plus général [BST14].

Enfin, pour $s \in C$, nous introduisons les applications ${}^{-}s$ et s^{-} qui, à un mot fini non vide u, associent le mot $s(u)$ auquel on efface la première (resp. la dernière) lettre. Ces applications ne sont pas des morphismes.

3 Construction de mots C-adiques de déséquilibres arbitrairement grands

Pour tout entier n, nous souhaitons exhiber un mot C-adique de déséquilibre supérieur à n. Pour ce faire, nous allons construire par récurrence une suite $(w_n)_{n\in\mathbb{N}}$ de mots C-adiques, et donner sur chacun d'eux deux facteurs u_n et v_n de même longueur, dont le vecteur de déséquilibre est de norme n. Cela assure que w_n a pour déséquilibre au moins n. L'intuition de ces constructions provient de l'exploration d'un automate et de l'étude de ses chemins.

Construction de (w_n).

Soit w_0 n'importe quel mot C-adique, par exemple $c_1 \circ c_2 \circ c_1 \circ c_2 \circ ...(a)$. Posons $w_1 = c_2 \circ c_2 \circ c_2(w_0)$ et pour tout $n \geq 1$:

$$\begin{cases} w_{n+1} = c_1^{2n+2} \circ c_2(w_n) & \text{si } n \text{ est impair} \\ w_{n+1} = c_2^{2n+2} \circ c_1(w_n) & \text{sinon.} \end{cases}$$

Pour tout entier n, w_n est un mot C-adique.

Construction des suites de facteurs (u_n) et (v_n).

La lettre b apparaît dans w_0 (la complexité nous le garantit), donc $acc = c_2 \circ c_2 \circ c_2(b)$ apparaît dans w_1. Posons $u_1 = ac$, $v_1 = cc$ et, pour tout $n \geq 1$:

$$\begin{cases} u_{n+1} = c_1^- \circ (c_1 \circ c_1)^n \circ c_1 \circ c_2(u_n) \\ v_{n+1} = c_1 \circ (c_1 \circ^- c_1)^n \circ c_1 \circ c_2(v_n) \end{cases} \qquad \text{si } n \text{ est impair}$$

$$\begin{cases} u_{n+1} = c_2 \circ (c_2 \circ c_2^-)^n \circ c_2 \circ c_1(u_n) \\ v_{n+1} =^- c_2 \circ (c_2 \circ c_2)^n \circ c_2 \circ c_1(v_n) \end{cases} \qquad \text{sinon.}$$

Les mots u_n et v_n sont bien facteurs de w_n, pour chaque n.

Lemme 1. *Soit $n \geq 1$.*

- *Si n est impair, alors u_n commence par a et termine par c, v_n commence et termine par c, et $\chi(u_n) - \chi(v_n) = (n, 1-n, -1)$.*

- *Sinon, alors u_n commence et termine par a, v_n commence par a et termine par c, et $\chi(u_n) - \chi(v_n) = (1, n-1, -n)$.*

Theorème 1. *Il existe des mots C-adiques de déséquilibre arbitrairement grand.*

4 Construction de mots C-adiques de déséquilibre infini

Nous venons de construire une famille de mots C-adiques pour laquelle le déséquilibre n'est pas borné. Toutefois, le déséquilibre des mots pris individuellement peut l'être. Pour construire un mot dont le déséquilibre est infini, nous allons recourir à un lemme de pompage.

La première étape consiste à montrer que lorsque l'on compose un mot C-adique par l'une ou l'autre des substitutions c_1 et c_2, on ne rééquilibre pas le mot au-delà d'une certaine proportion.

Proposition 1. *Si w est un mot C-adique de déséquilibre $Des(w) \geq 3n$, alors $c_1(w)$ (resp. $c_2(w)$) est un mot C-adique de déséquilibre $Des(w) \geq n$.*

Pour tout entier naturel m, on note désormais $C^m := \{s_0 \circ ... \circ s_{m-1} \in \{c_1, c_2\}^m\}$, et $C^* := \cup_{m \in \mathbb{N}} C^m$.

Corollaire 1 (Lemme de pompage). $\forall s \in C^*, \forall n \in \mathbb{N}, \exists \sigma \in C^*$ *tel que pour tout mot C-adique w, $Des(s \circ \sigma(w)) \geq n$.*

Theorème 2. *Il existe un mot C-adique de déséquilibre infini.*

Démonstration. D'après le corollaire 1, je peux construire une suite $(\sigma_k)_{k \in \mathbb{N}} \in (C^*)^{\mathbb{N}}$ telle que pour tout entier naturel n et pour tout mot C-adique w, le déséquilibre du mot $\sigma_0 \circ ... \circ \sigma_n(w)$ vaut au moins n. Ainsi, le mot limite $w_\infty = \lim_{n \to \infty} \sigma_0 \circ ... \sigma_n(w_0)$ (s'il existe, sinon prendre une valeur d'adhérence de la suite), où w_0 est un mot C-adique quelconque, est lui même un mot C-adique (chaque σ_k fourni par le corollaire contient c_1 et c_2), et son déséquilibre vaut au moins n, pour tout n ; il est donc de déséquilibre infini. □

Références

La bibliographie de cet article est déplacée à la fin du chapitre, où elle est complétée par quelques références supplémentaires.

2 Détail des démonstrations

Nous conservons les notations et numérotations du résumé publié.

2.1 Preuve du Lemme 1

Rappel de l'énoncé. *Soit $n \geq 1$.*

— *Si n est impair, alors u_n commence par a et termine par c, v_n commence et termine par c, et $\chi(u_n) - \chi(v_n) = (n, 1-n, -1)$.*

— *Sinon, alors u_n commence et termine par a, v_n commence par a et termine par c, et $\chi(u_n) - \chi(v_n) = (1, n-1, -n)$.*

Remarque 1. *Si Sym désigne l'opération miroir suivie de l'échange des lettres a et c (par exemple $Sym(aabcb) = babcc$), alors le diagramme suivant est commutatif :*

$$\begin{tikzcd}
u \arrow[r, mapsto, "c_1"] \arrow[d, mapsto, "Sym"'] & c_1(u) \arrow[d, mapsto, "Sym"] \\
v \arrow[r, mapsto, "c_2"'] & c_2(v).
\end{tikzcd}$$

Autrement-dit, pour tout mot fini u : $c_2(Sym(u)) = Sym(c_1(u))$. De même, les applications c_1^- et ${}^-c_2$ d'une part, et ${}^-c_1$ et c_2^- d'autre part, sont conjuguées par Sym.

Démonstration. Nous démontrons le lemme par récurrence sur n.

Pour $n = 1$, $u_1 = ac$, $v_1 = cc$ et $\chi(u_1) - \chi(v_1) = (1, 0, -1)$ satisfont la propriété.

Soit $n \geq 1$; supposons la propriété vraie pour (u_n, v_n).

Si n est impair, alors $n+1$ est pair, $c_1 \circ c_2(u_n)$ commence par a et termine par b, $c_1 \circ c_2(v_n)$ commence et termine par b ; de plus, $\chi(c_1 \circ c_2(u_n)) - \chi(c_1 \circ c_2(v_n)) = (1, -n, n)$. Puis $(c_1 \circ c_1)^n[c_1 \circ c_2(u_n)]$ commence par a et termine par b, $(c_1 \circ^- c_1)^n[c_1 \circ c_2(v_n)]$ commence et termine par b, et $\chi((c_1 \circ c_1)^n[c_1 \circ c_2(u_n)]) - \chi((c_1 \circ^- c_1)^n[c_1 \circ c_2(v_n)]) = (1+n, -n, n)$. Enfin, $u_{n+1} = c_1^-[(c_1 \circ c_1)^n \circ c_1 \circ c_2(u_n)]$ commence et termine par a, $v_{n+1} = c_1[(c_1 \circ^- c_1)^n \circ c_1 \circ c_2(v_n)]$ commence par a et termine par c, et $\chi(u_{n+1}) - \chi(v_{n+1}) = (1, n, -n-1)$; il s'agit bien des résultats prédits par le lemme dans le cas pair.

Sinon, n est pair et $n+1$ impair, $Sym(v_n)$ commence par a et termine par c, $Sym(u_n)$ commence et termine par c et $\chi(Sym(v_n)) - \chi(Sym(u_n)) = (n, 1-n, -1)$. De plus, d'après la Remarque 1, $Sym(v_{n+1}) = c_1^-[(c_1 \circ c_1)^n \circ c_1 \circ c_2(Sym(v_n))]$ et $Sym(u_{n+1}) = c_1[(c_1 \circ^- c_1)^n \circ c_1 \circ c_2(Sym(u_n))]$. Le raisonnement du cas *impair* montre alors que $Sym(v_{n+1})$ commence et termine par a, $Sym(u_{n+1})$ commence par a et termine par c, et $\chi(Sym(v_{n+1})) - \chi(Sym(u_{n+1})) = (1, n, -n-1)$. D'où : u_{n+1} commence par a et termine par c, v_{n+1} commence et termine par c, puis $\chi(u_{n+1}) - \chi(v_{n+1}) = (n+1, -n, -1)$; ce qui achève la preuve du lemme. □

2.2 Preuve du Théorème 1

Rappel de l'énoncé. *Il existe des mots C-adiques de déséquilibre arbitrairement grand.*

Démonstration. Pour chaque entier $n \geq 1$, w_n est un mot C-adique qui admet pour facteur u_n et v_n. De plus, puisque la somme des coordonnées du vecteur $\chi(u_n) - \chi(v_n)$ est nulle, les mots u_n et v_n sont de même longueur, et $\chi(u_n) - \chi(v_n)$ est un vecteur de déséquilibre. Sa norme infinie étant égale à n, le déséquilibre de w_n vaut au moins n. □

Nous venons en fait de démontrer que, quel que soit le choix du mot C-adique w_0, le déséquilibre de w_n vaut au moins n. Autrement-dit, si on introduit la suite de substitutions $(\gamma_n)_{n\in\mathbb{N}^*}$ définie par $\gamma_1 = c_2 \circ c_2 \circ c_2$ puis

$$\begin{cases} \gamma_{n+1} = c_1^{2n+2} \circ c_2 \circ \gamma_n & \text{si } n \text{ est impair,} \\ \gamma_{n+1} = c_2^{2n+2} \circ c_1 \circ \gamma_n & \text{sinon.} \end{cases}$$

nous pouvons préciser l'énoncé du théorème :

Proposition 2 (précision du Théorème 1)**.** *Pour tout entier strictement positif n, pour tout mot C-adique w, le déséquilibre du mot $\gamma_n(w)$, qui est encore C-adique, vaut au moins n.*

2.3 Preuve de la Proposition 1

Rappel de l'énoncé. *Si w est un mot C-adique de déséquilibre $Des(w) \geq 3n$, alors $c_1(w)$ (resp. $c_2(w)$) est un mot C-adique de déséquilibre $Des(w) \geq n$.*

Démonstration. Soit w un mot C-adique de déséquilibre $Des(w) \geq 3n$. Il existe donc deux facteurs $u, v \in F(w)$ de même longueur, et une lettre $\alpha \in A$ tels que $|u|_\alpha - |v|_\alpha \geq 3n$. Sans perte de généralité, on peut supposer que $|u|_\beta - |v|_\beta \leq 0$ pour tout $\beta \in A\backslash\{\alpha\}$. L'un des quatre cas suivants est donc vérifié :

— $\alpha = c$

— $\alpha = b$

— $\alpha = a$ et $|u|_b - |v|_b \leq -2n$

— $\alpha = a$ et $-2n < |u|_b - |v|_b \leq 0$.

Nous allons respectivement construire $\tilde{u}$ et $\tilde{v} \in F(c_1(w))$ tels que $|\tilde{u}| = |\tilde{v}|$ et :

— $|\tilde{u}|_b - |\tilde{v}|_b \geq n$

— $|\tilde{u}|_c - |\tilde{v}|_c \geq n$

— $|\tilde{v}|_c - |\tilde{u}|_c \geq n$

— $|\tilde{u}|_a - |\tilde{v}|_a \geq n$.

Nous aurons ainsi montré que, dans tous les cas, le déséquilibre de $c_1(w)$ vaut au moins n.

Premier cas : $\alpha = c$. Nous avons donc

$$\begin{cases} |u|_a - |v|_a \leq 0 \\ |u|_b - |v|_b \leq 0 \\ |u|_c - |v|_c \geq 3n \end{cases} \qquad \text{et} \qquad |u| = |v|,$$

d'où :

$$\chi(c_1(u)) - \chi(c_1(v)) = \begin{pmatrix} |u|_a - |v|_a + |u|_b - |v|_b \\ |u|_c - |v|_c \\ |u|_b - |v|_b \end{pmatrix} = \begin{pmatrix} -(|u|_c - |v|_c) \\ |u|_c - |v|_c \\ |u|_b - |v|_b \end{pmatrix}$$

et $|c_1(u)| - |c_1(v)| = (1,1,1) \cdot (\chi(c_1(u) - \chi(c_1(v))) = |u|_b - |v|_b \leq 0$. En posant $\tilde{u} = c_1(u)$ et $\tilde{v}$ le *préfixe* de longueur $|\tilde{u}|$ de $c_1(v)$ (c'est-à-dire le mot formé par les $|\tilde{u}|$ premières lettres de $c_1(v)$), on obtient :

$$\begin{cases} |\tilde{u}| = |\tilde{v}| \\ |\tilde{u}|_b - |\tilde{v}|_b \geq |c_1(u)|_b - |c_1(v)|_b = |u|_c - |v|_c \geq 3n \geq n. \end{cases}$$

Deuxième cas : $\alpha = b$. Nous partons à présent de

$$\begin{cases} |u|_a - |v|_a \leq 0 \\ |u|_b - |v|_b \geq 3n \\ |u|_c - |v|_c \leq 0 \end{cases} \qquad \text{et} \qquad |u| = |v|,$$

d'où $|c_1(u)|_c - |c_1(v)|_c = |u|_b - |v|_b \geq 3n$ et $|c_1(u)| - |c_1(v)| = |u|_b - |v|_b \geq 0$.
Nous posons $\tilde{v} = c_1(v)$ et souhaitons "raccourcir" $c_1(u)$ tout en contrôlant le nombre d'occurrences de c que l'on perd. Mais en remarquant que dans tout suffixe de $c_1(u)$ (le *suffixe* de longueur l de $c_1(u)$ est le mot formé des l dernières lettres de $c_1(u)$), chaque occurrence de c est précédée par une occurrence de a, sauf éventuellement la première dans le cas où le suffixe commence par c, on déduit que parmi les $|u|_b - |v|_b$ dernière lettres de $c_1(u)$, il y a au plus $\lceil \frac{|u|_b - |v|_b}{2} \rceil$ occurrences de c. Ainsi, en notant $\tilde{u}$ le préfixe de $c_1(u)$ de longueur $|\tilde{v}| = |c_1(v)| = |c_1(u)| - (|u|_b - |v|_b)$, on obtient :
$\begin{cases} |\tilde{u}| = |\tilde{v}| \\ |\tilde{u}|_c - |\tilde{v}|_c \geq (|c_1(u)|_c - \lceil \frac{|u|_b - |v|_b}{2} \rceil) - |c_1(v)|_c \geq |u|_b - |v|_b - \lceil \frac{|u|_b - |v|_b}{2} \rceil = \lfloor \frac{|u|_b - |v|_b}{2} \rfloor \geq n. \end{cases}$

Troisième cas : $\alpha = a$ et $|u|_b - |v|_b \leq -2n$. Comme dans le cas précédent, nous allons nous intéresser au nombre d'occurrences de la lettre c. Nous partons de

$$\begin{cases} |u|_a - |v|_a \geq 3n \\ |u|_b - |v|_b \leq -2n \\ |u|_c - |v|_c \leq 0 \end{cases} \qquad \text{et} \qquad |u| = |v|,$$

d'où $|c_1(u)|_c - |c_1(v)|_c = |u|_b - |v|_b \leq 2n$ et $|c_1(u)| - |c_1(v)| = |u|_b - |v|_b \leq 0$.
C'est donc à présent $c_1(v)$ que nous souhaitons "raccourcir". En posant $\tilde{u} = c_1(u)$ et $\tilde{v}$ le préfixe de $c_1(v)$ de longueur $|\tilde{u}|$, et en remarquant, par l'argument précédent, que la lettre c apparaît au plus $\lceil \frac{|v|_b - |u|_b}{2} \rceil$ fois dans le suffixe raboté, on obtient :

$$\begin{cases} |\tilde{u}| = |\tilde{v}| \\ |\tilde{v}|_c - |\tilde{u}|_c \geq \lfloor \frac{|v|_b - |u|_b}{2} \rfloor \geq n. \end{cases}$$

Quatrième cas : $\alpha = a$ et $-2n < |u|_b - |v|_b \leq 0$.

Dans ce dernier cas, $|c_1(u)|_a - |c_1(v)|_a = |u|_a - |v|_a + |u|_b - |v|_b > n$ et $|c_1(u)| - |c_1(v)| = |u|_b - |v|_b \leq 0$. On pose donc $\tilde{u} = c_1(u)$ et $\tilde{v}$ le préfixe de $c_1(v)$ de longueur $|\tilde{u}|$. On obtient immédiatement :
$\begin{cases} |\tilde{u}| = |\tilde{v}| \\ |\tilde{u}|_a - |\tilde{v}|_a \geq |c_1(u)|_a - |c_1(v)|_a \geq n, \end{cases}$
ce qui achève la démonstration pour la substitution c_1. On obtient les même résultats pour c_2 à condition de permuter les lettres a et c et d'intervertir les rôles des préfixes et des suffixes dans la preuve. □

2.4 Preuve du Corollaire 1

Notons C^* l'ensemble des mots finis sur l'alphabet $C = \{c_1, c_2\}$ et $C^{*,\circ}$ l'ensemble des substitutions obtenues par un nombre fini de compositions de c_1 et c_2. Nous commençons par justifier que nous pouvons "confondre" C^* et $C^{*,\circ}$.

Lemme 2. *L'application $\psi : C^* \to C^{*,\circ}$ définie par $\psi(s_0 ... s_{n-1}) = s_0 \circ ... \circ s_{n-1}$ est bijective.*

Démonstration. Soit $\sigma = s_0 \circ ... \circ s_{m-1} \in C^{*,\circ}$. Nous allons montrer que les substitutions $s_{m-1}, ..., s_0 \in C$ sont uniquement déterminées. Observons que $ab = c_1 c_2(b) = c_1 c_1(b)$ mais que $ab \notin c_2(A^*)$; observons symétriquement que $bc = c_2 c_1(b) = c_2 c_2(b)$ mais que $bc \notin c_1(A^*)$. Aussi, et ces deux situations s'excluent mutuellement, si ab apparaît dans l'une au moins des images des lettres par la substitution σ, alors $s_{m-1} = c_1$, et si bc apparaît dans l'une au moins des images des lettres, alors $s_{m-1} = c_2$. Comme par ailleurs, toutes les lettres de l'alphabet apparaissent dans l'image d'au moins une lettre par c_1, et dans l'image d'au moins une lettre par c_2, toutes les lettres de l'alphabet (et en particulier la lettre b) apparaissent dans l'image d'au moins une lettre par n'importe quelle composition des substitutions c_1, c_2. Nous déduisons que si ni le facteur ab ni le facteur bc n'apparaît dans les images des lettres par σ, alors $m = 0$ ou 1 et σ est soit l'identité, soit c_1, soit c_2, que nous reconnaissons immédiatement.

Pour conclure la démonstration du lemme par récursion, nous justifions que si $v = c_1(u)$ (resp. $v = c_2(u)$), alors nous pouvons déduire u de la connaissance de v. Cela est possible car les mots $c_1(a), c_1(b)$ et $c_1(c)$ terminent par des lettres distinctes (resp. $c_2(a), c_2(b)$ et $c_2(c)$ commencent par des lettres distinctes.) Ci-dessous, nous précisons les algorithmes qui permettent d'obtenir u à partir de v.

Algorithme 1 : $v = c_1(u)$.

```
Entrée : v.
Initialisation : $u \leftarrow \epsilon$ (mot vide) ; $k \leftarrow |v| - 1$.
Tant que $k > -1$ :
    si $v[k] = a$, faire $u \leftarrow au$ (on concatène $a$ à la gauche de $u$) et $k \leftarrow k - 1$ ;
    sinon, si $v[k] = b$, faire $u \leftarrow cu$ et $k \leftarrow k - 1$ ;
    sinon (cas $v[k] = c$) $u \leftarrow bu$ et $k \leftarrow k - 2$.
Rendre $u$.
```

Algorithme 2 : $v = c_2(u)$.

```
Entrée : v.
Initialisation : $u \leftarrow \epsilon$ (mot vide) ; $k \leftarrow 0$.
Tant que $k < |v|$ :
    si $v[k] = a$, faire $u \leftarrow ub$ et $k \leftarrow k + 2$ ;
    sinon, si $v[k] = b$, faire $u \leftarrow ua$ et $k \leftarrow k + 1$ ;
    sinon (cas $v[k] = c$) $u \leftarrow uc$ et $k \leftarrow k + 1$.
Rendre $u$.
```

□

Rappel de l'énoncé (Corollaire 1)**.** $\forall s \in C^*, \forall n \in \mathbb{N}, \exists t \in C^*$ *tel que pour tout mot C-adique* w, $Des(s \circ t(w)) \geq n$.

Démonstration. Soit $s \in C^*$ et $n \in \mathbb{N}$. Notons m la longueur du mot s. D'après la Proposition 2, on peut construire une suite de substitutions $(\gamma_k)_{k \in \mathbb{N}}$ telle que pour tout mot C-adique w, le déséquilibre de $\gamma_k(w)$ vaut au moins k. Choisir $t := \gamma_{n3^m}$ convient. En effet, en appliquant m fois la Proposition 1, nous obtenons que pour tout mot C-adique w : $Des(s \circ t(w)) \geq 3^{-m} Des(t(w)) \geq n$. □

3 Commentaires

3.1 Sur la stratégie de construction d'un mot C-adique de déséquilibre infini à partir de mots C-adique de déséquilibres arbitrairement grands.

Nous avons exhibé une famille $\{w_n\}_{n \geq 1}$ de mots C-adiques dont le déséquilibre n'est pas uniformément borné. Toutefois, rien ne s'oppose à ce que le déséquilibre des mots pris individuellement le soit.

Pour "fabriquer" un mot C-adique de déséquilibre infini, on pourrait être tenté de passer à la limite ; mais la suite (w_n) que nous proposons, d'une part, n'a pas de limite (ceci indépendamment du choix de w_0) ; d'autre part, ses valeurs d'adhérence dans le compact $A^{\mathbb{N}}$ n'ont aucune raison d'être C-adiques ou déséquilibrées (en fait, ce sont des mots ultimement constants.)

Nous parvenons cependant à construire, en concaténant des préfixes bien choisis des suites directrices des w_n, une nouvelle suite directrice $(d_k)_{k \in \mathbb{N}}$ pour laquelle :

— la suite de mots infinis $(d_0 \circ ... \circ d_{k-1}(w_0))_k$, où w_0 est un mot C-adique quelconque (mais fixé), converge,

— c_1 et c_2 apparaissent une infinité de fois dans $(d_k)_{k\in\mathbb{N}}$,

— il existe une suite d'indices $(k_n)_{n\in\mathbb{N}}$ telle que pour tout mot C-adique w, le déséquilibre du mot $d_0 \circ ... \circ d_{k_n}(w)$ vaut au moins n.

Ainsi, le mot $w_\infty := \lim_{n\mapsto\infty} d_0 \circ ... \circ d_{k-1}(w_0)$ est encore un mot C-adique et son déséquilibre n'est pas borné.

Le basculement d'une "mauvaise" limite par compositions à gauche à une "bonne" limite par compositions à droite, qui est l'objet souhaitable lorsque l'on travaille avec un système S-adique, repose sur la Proposition 1 :

Si w est un mot C-adique de déséquilibre $Des(w) \geq 3n$, alors $c_1(w)$ (resp. $c_2(w)$) est un mot C-adique de déséquilibre $Des(w) \geq n$.

Autrement-dit, lorsque l'on compose un mot C-adique par l'une ou l'autre des substitutions c_1 et c_2, on ne rééquilibre pas le mot au-delà d'une certaine proportion. Ce résultat permet de construire, par un procédé de pompage (Corollaire 1 puis Théorème 2), des mots C-adiques de plus en plus déséquilibrés, et dont les suites directrices partagent un préfixe de plus en plus long. Le mot limite que l'on obtient alors (quitte à extraire une sous-suite) est C-adique ; son déséquilibre, ne pouvant être majoré, est infini.

3.2 Sur la croissance du déséquilibre

Nous pouvons reformuler la Proposition 2 par : *tout mot C-adique dont la suite directrice commence par le préfixe γ_n a pour déséquilibre au moins n.* Deux questions émergent naturellement :

Question I - 1. *Pour $n \in \mathbb{N}$, quel est le mot le plus court $\gamma_n^0 \in C^*$ tel que tout mot C-adique dont la suite directrice commence par le préfixe γ_n^0 ait un déséquilibre supérieur ou égal à n.*

Question I - 2. *A quelle vitesse croît $(|\gamma_n^0|)_{n\in\mathbb{N}}$, la suite des longueurs de ces préfixes ?*

Nous pensons que nous ne pouvons pas espérer mieux qu'une croissance quadratique. La suite γ_n que nous avons exhibée croît quadratiquement.

Lemme 3. *Pour tout entier $n \geq 1$, $|\gamma_n| = n^2 + 2n$.*

Démonstration. Par construction, la suite $(|\gamma_n|)_{n\in\mathbb{N}}$ a pour premier terme $|\gamma_1| = 3$ puis satisfait la relation de récurrence $|\gamma_{n+1}| = 2n + 3 + |\gamma_n|$. □

Cependant, la croissance des longueurs des préfixes $(\sigma_n)_{n\in\mathbb{N}} \in (C^*)^\infty$ qui garantissent un déséquilibre supérieur ou égal à n dans le mot w_∞ nous avons construit dans le Théorème 2 est beaucoup plus rapide. Pour a et n deux entiers naturel, on définit l'opération de *tétration*, notée $a \uparrow\uparrow n$, par :

$$\begin{cases} a \uparrow\uparrow 0 & = & 1 \\ a \uparrow\uparrow (n+1) & = & a^{a\uparrow\uparrow n}. \end{cases}$$

Par exemple :

$$\begin{array}{lcl} 3 \uparrow\uparrow 1 & = & 3 \\ 3 \uparrow\uparrow 2 & = & 3^3 \quad = \quad 27, \\ 3 \uparrow\uparrow 3 & = & 3^{3^3} \quad = \quad 3^{27} = 7625597484987, \\ 3 \uparrow\uparrow 4 & = & 3^{3^{3^3}} \quad = \quad 3^{7625597484987}. \end{array}$$

Lemme 4. *Pour tout entier naturel $n \geq 1$, $|\sigma_n| \geq 3 \uparrow\uparrow n$.*

Démonstration. Par construction, nous avons $\sigma_0 = id$ puis $\sigma_{n+1} = \sigma_n \cdot \gamma_{(n+1)3^{|\sigma_n|}}$. D'où $|\sigma_1| = |\gamma_1| = 3$ puis $|\sigma_{n+1}| = |\sigma_n| + |\gamma_{(n+1)3^{|\sigma_n|}}|$. Pour tout entier naturel n, nous minorons $|\gamma_n|$ par n : $|\sigma_{n+1}| \geq |\sigma_n| + |(n+1)3^{|\sigma_n|}| \geq 3^{|\sigma_n|}$. Nous avons donc, pour tout entier $n \geq 1$, $|\sigma_n| \geq 3 \uparrow\uparrow n$. □

On peut bien sûr accélérer la croissance du déséquilibre de w_∞ en initialisant σ_0 à γ_{10} ou γ_{1000}, ou bien en "pompant plus vite" : au lieu de construire à l'étape n, un préfixe σ_n qui garanti un déséquilibre au moins n, on construit un préfixe σ'_n qui garantit un déséquilibre n^2 (ou n'importe quelle fonction strictement croissante.) Mais ces accélérations ne permettent par de contrebalancer la croissance en tour d'exponentielles de la suite $(|\sigma_n|)_{n\in\mathbb{N}}$.

Références

[BST14] Valérie Berthé, Wolfgang Steiner, and Jörg Thuswaldner. Geometry, dynamics, and arithmetic of S-adic shifts. *preprint https ://arxiv.org/abs/1410.0331*, 2014.

[CFZ00] Julien Cassaigne, Sébastien Ferenczi, and Luca Q. Zamboni. Imbalances in Arnoux-Rauzy sequences. *Annales de l'Institut Fourier*, 50 :1265–1276, 2000.

[CLL17] Julien Cassaigne, Sébastien Labbé, and Julien Leroy. A set of sequences of complexity 2n+1. In *Combinatorics on Words - 11th International Conference, WORDS 2017, Montréal, QC, Canada, September 11-15, 2017, Proceedings*, pages 144–156, 2017.

II. A semi-algorithm to explore the set of imbalances in a S-adic system

The content of this chapter was presented during the workshop Dyadics 3 (Decidability and dynamical systems, Amiens 2019).

Contents

Abstract

In this chapter, we formalize the method which led to the construction of a word with infinite imbalance for the Cassaigne-Selmer multidimensional continued fraction algorithm. We introduce a general semi-algorithm consisting of an automaton, or rather, an ever-building family of automata, whose states contain all the possible imbalances of infinite words in a S-adic system, thus proving that the property of C-balancedness is semi-decidable.

1 Motivations

A S-adic word is an infinite word which can be written as the limit of the iterated images of a letter by an infinite composition of substitutions (not necessarily the same). S-adic systems are thus a large generalization of substitutive systems. They emerged from the study of the symbolic dynamics of continued fraction algorithms (see for instance [Fog02], [Ber11] or [BD14]).

The imbalance is a combinatorial quantity which measures inequities in the distribution of letters in an infinite word. This quantity is linked to the quality of the convergence of the frequency of letters in growing prefixes of the word, so, in the context of multidimensional continued fraction algorithms, to the quality of the convergence of the approximations proposed by the algorithm.

To our knowledge, two distinct behaviors have been observed.

On one hand, the regular continued fraction algorithm is associated with the S-adic system $X_{S_{st}}$ of Example 5. Every infinite word in X_{st} has its imbalance bounded above by 1.

On the other hand, the S-adic systems associated with Arnoux-Rauzy and Brun (2-dimensional) continued fraction algorithms contain words with infinite imbalance (see respectively [CFZ00] and [DHS13]). The existence of such words was unexpected; in fact, they have measure zero ([DHS13]). The construction of these exceptional words relied on thorough observations and an outstanding intuition.

The following questions have not been answered yet.

1. Does there exist a S-adic system that satisfies the 'intermediate situation', i.e. such that (i) all S-adic words have a finite imbalance; (ii) there is no upper bound for these imbalances.

2. Does there exist a multidimensional continued fraction algorithm which generate words with bounded, or at least finite, imbalances?

This document introduces a semi-algorithm that would help in developing intuitions for a wide range of S-adic systems. Roughly speaking, this semi-algorithm consists in tracking backwards, in an economical way, the common desubstitution history of pairs of factors. This tool is already behind the construction of a C-adic word with infinite imbalance (Chapter 1) and behind the construction of a Rauzy fractal unbounded in all directions (Chapter 4). Most probably, it will be useful to address other questions, such that Questions 1 and 2 above.

2 Preliminaries

2.1 General definitions

An *alphabet* A is a finite set, whose elements are called *letters.* We will work with $(A^* := \cup_{n\in\mathbb{N}} A^n, \cdot)$ the free monoid over A for the concatenation operation. Its elements are called *finite words*; its neutral element ϵ is *the empty word.* A *language* L is a subset of A^*. An *infinite word* over A is an element of $A^{\mathbb{N}}$. The *length* of a finite word u, denoted by $|u|$, is the total number of letters it is written with. A finite word $u = u[0]u[1]...u[n-1]$, where $u[k]$ denotes the $(k+1)$-th letter of u, is a *factor* of a (finite or infinite) word w if there exists a nonnegative integer i such that for all $k \in \{0, ..., n-1\}$, $w[i+k] = u[k]$. In the particular case $i = 0$, we say that u is the *prefix of length* n *of* w, and denote it by $u = p_n(w)$. Symmetrically, a (finite or infinite) word w' is a *suffix* of w if there exists $l \in \mathbb{N}$ such that $w = p_l(w) \cdot w'$. We immediately see that w' is finite if and only if w is finite; in that case, if n denotes the length of w', we say that w' is the suffix of length n of w, that we denote by $w' = s_n(w)$. At last, we denote by $\mathcal{F}_n(w)$ the set of factors of w of length n and by $\mathcal{F}(w)$ the *language of* w, which is the set of all its factors.

The *abelianized vector* (sometimes called *Parikh vector*) of a finite word $u \in A^*$ is the line vector $\text{ab}(u) := (|u|_a)_{a \in A}$ which counts the number of occurrences of each letter in u. We immediately see that the sum of its coordinates is equal to $|u|$, the length of the word u. If two words u and v have the same length, we call *imbalance vector of* u *and* v the difference of their abelianized vectors: $\text{imb}(u, v) = \text{ab}(u) - ab(v)$. The sum of its coordinates is of course equal to zero. The *imbalance* of a (finite or infinite) word w is the quantity, possibly infinite:

$$\begin{aligned} \text{imb}(w) &= \sup_{n \in \mathbb{N}} \sup_{u,v \in \mathcal{F}_n(w)} ||\text{imb}(u, v)||_\infty \\ &= \sup_{n \in \mathbb{N}} \sup_{u,v \in \mathcal{F}_n(w)} \max_{a \in A} ||u|_a - |v|_a|. \end{aligned}$$

The imbalance measures inequities in the distribution of letters in w.

A *substitution* is an application mapping letters to finite words: $A \mapsto A^*$, that we extend into a morphism on the free monoid A^* on one hand, and on the set of infinite words $A^{\mathbb{N}}$ on the other hand. For instance, consider the Thue-Morse substitution:

$$\begin{aligned} \sigma_{TM} :\ & \{a, b\} \to \{a, b\}^* \\ & a \mapsto ab \\ & b \mapsto ba. \end{aligned}$$

A substitution is *nonerasing* if no image of letter is the empty word. At last, the *incidence matrix* of a substitution σ defined over A is:

$$M_\sigma := (|\sigma(i)|_j)_{i,j \in A}.$$

Incidence matrices and abelianized vectors are made to satisfy $\text{ab}(\sigma(u)) = \text{ab}(u)M_\sigma$, for any finite word u.

In this document, we are interested in the imbalance of words associated with a set of substitutions.

2.2 S-adic systems

First of all, we endow the set of infinite words $A^{\mathbb{N}}$ with the distance δ, which makes it compact: for all $w, w' \in A^{\mathbb{N}}$, $\delta(w, w') = 2^{-n_0}$, where $n_0 = \min\{n \in \mathbb{N} | w[n] \neq w'[n]\}$ if $w \neq w'$, and $\delta(w, w') = 0$ otherwise. We say that a sequence of finite words $(u_n)_{n \in \mathbb{N}} \in (A^*)^{\mathbb{N}}$ *converges* to an infinite word $w \in A^{\mathbb{N}}$ if for any sequence of infinite words $(v_n)_{n \in \mathbb{N}} \in (A^{\mathbb{N}})^{\mathbb{N}}$, the sequence of infinite words $(u_n \cdot v_n)_{n \in \mathbb{N}} \in (A^{\mathbb{N}})^{\mathbb{N}}$ converges to w.

Let S be a finite set of substitutions, defined over a common alphabet A. An infinite word $w \in A^{\mathbb{N}}$ is *S-adic* if there exist a *directive sequence* $(d_n) \in S^{\mathbb{N}}$, together with a *seed* $a \in A$ such that the sequence of finite words $(d_0 \circ ... \circ d_{n-1}(a))_{n \in \mathbb{N}}$ converges to w.

Example 1. *The set* $S = \{\sigma\}$ *where* σ *is the substitution defined over* $\{1, 2\}$ *by* $\sigma(1) = 2$, $\sigma(2) = 1$, *generates no S-adic word.*

Example 2. *The set $S = \{\sigma_{TM}\}$, where σ_{TM} is the Thue-Morse substitution defined in Subsection 2.1, generates two S-adic words:*

$$w_a = \lim_n (\sigma_{TM})^n(a) = abbabaabbaababbabaababbaabbabaabbaababbaabbabaababbab...$$

$$w_b = \lim_n (\sigma_{TM})^n(b) = baababbaabbabaababbabaabbaababbaabbabaabbaababbabaaba...$$

Example 3. *The set $S = \{\sigma_{fib}\}$, where σ_{fib} is the substitution defined by $\sigma_{fib}(1) = 12$ and $\sigma_{fib}(2) = 1$ generates a unique S-adic word (called* Fibonacci word*):*

$$w_{fib} = \lim_n (\sigma_{fib})^n(1) = \lim_n (\sigma_{fib})^n(2) = 1211212112112121121211211212112112...$$

When the set S contains a unique substitution, S-adic words are said *substitutive*. Substitutive words have been intensively studied (see for instance [Fog02]).

Example - definition 4. *Let $S = C := \{c_1, c_2\}$ the set of Cassaigne-Selmer substitutions, defined over $\{1,2,3\}$ by:*

$$c_1 : \begin{array}{ll} 1 & \mapsto 1 \\ 2 & \mapsto 13 \\ 3 & \mapsto 2, \end{array} \qquad \textit{and} \qquad c_2 : \begin{array}{ll} 1 & \mapsto 2 \\ 2 & \mapsto 13 \\ 3 & \mapsto 3. \end{array}$$

As their name suggests, S-adic words for $S = C$ are exactly C-adic words, as they are defined in Chapter 1.

Example - definition 5. *Let $d \in \mathbb{N}$ such that $d \geq 2$. We consider the set of d substitutions $S_d = \{\sigma_{d,1}, ..., \sigma_{d,d}\}$ defined over the alphabet $A_d := \{1, ..., d\}$ by:*

$$\sigma_{d,i} : \begin{array}{lll} i & \mapsto & i \\ j & \mapsto & ij \quad \textit{if } j \neq i, \end{array}$$

for all $i \in \{1, ..., d\}$.

An infinite word over A_d is episturmian *if it has the same language than a S-adic word, for $S = S_d$. It is furthermore* strictly *episturmian if it has the same langage than a S-adic word whose directive sequence contains infinitely many occurrences of each substitutions in S_d.* Sturmian *words are exactly strict episturmian words for $d = 2$;* Arnoux-Rauzy *words are exactly strict episturmian words for $d = 3$. The Fibonacci word (see Example 3) is Sturmian: indeed, it is the limit of the sequence $((\sigma_{2,1} \circ \sigma_{2,2})^n(1))_{n\in\mathbb{N}}$.*

From these five examples, we deduce that:

1. Not all sets of substitutions generate S-adic words.
2. Given a directive sequence $(d_n)_{n\in\mathbb{N}} \in S^{\mathbb{N}}$ and a seed $a \in A$, the sequence of finite words $(d_0 \circ ... \circ d_{n-1}(a))_{n\in\mathbb{N}}$ may not even have an adherence value in $A^{\mathbb{N}}$.
3. In general, S-adic words are not substitutive. Indeed, there are only a countable number of substitutive words, whereas the set of Sturmian words, which is a subset of the class of all S-adic words, is already uncountable (the slope function maps them onto the irrational numbers).

4. Given a S-adic word, the directive sequence and the seed are not *a priori* unique. Nonetheless, for some sets of substitutions (e.g. standard episturmian words or C-adic words), it is possible to find sufficient conditions on $(d_n) \in S^{\mathbb{N}}$ and $a \in A$ to ensure both the convergence of the sequence $(d_0 \circ ... \circ d_{n-1}(a))_n$ to an infinite word w, and the uniqueness of its directive sequence (see for instance [AR91], [AS13] for Arnoux-words, [GLR09] for episturmian words in general, and [CLL17] for C-adic words).

Let T denote the *shift map*, which erases the first letter of infinite words: if $w = w[0]w[1]w[2]... \in A^{\mathbb{N}}$, then $T(w) = w[1]w[2]w[3]....$ The *S-adic system* associated with a set of substitutions S is the minimal close set containing all S-adic words and stable under the action of T:

$$X_S = \overline{\{T^n(w) \mid \text{ for } n \in \mathbb{N} \text{ and } w\, S\text{-adic word}\}}.$$

Lemma 6 (immediate)**.** *Let S be a set of substitutions. The language of X_S (i.e. the set of factors of words in X_S) is exactly the language of the set of all S-adic words.*

3 Main result

3.1 Teaser

Given a set S of substitutions, we want to answer the questions:

1. Are the imbalances of words in X_S bounded?

2. If they are, give an upper bound.

3. If they are not, for an arbitrary $d \in \mathbb{N}$, exhibit a word in X_S whose imbalance is greater than d.

Theorem A. *Let S denote a finite set of nonerasing substitutions over a common alphabet A, and assume that all letters in A belong to the language of X_S. If D_S denotes the quantity (possibly infinite):*

$$D_S = \sup_{w \in X_S} \operatorname{imb}(w),$$

then a breadth first search in the automaton of imbalances, from its initial states, outputs, for any $d \leq D_S$, a finite sequence of substitutions $(\sigma_i)_{i \in \{1,\dots,n\}}$ in S such that the imbalance of any word in X_S, whose directive sequence starts with $(\sigma_i)_{i \in \{1,\dots,n\}}$, is larger than d.

Remark 7. *We will see (Proposition 13) that the automaton of imbalances has:*

- *an infinite number of states;*

- *a bounded number of transitions from any state;*

- *a finite number of initial states, which only depends on the alphabet A;*

so its breadth first traversal is possible (but would, of course, never end).

Corollary 8 (immediate)**.** *Let $d \in \mathbb{N}$. The question "does there exists a word in X_S with imbalance greater than d?" is semi-decidable.*

3.2 Description of the automaton of imbalances

This subsection is devoted to the harsh (but complete) description of the automaton of imbalances. It is written for readers willing to program it. To understand its construction, it may be preferable to read Section 4 or to consider the Thue-Morse example, in Subsection 3.4.

Definition 9. *Let S be a finite set of nonerasing substitutions defined over a common alphabet A. Let $(l_1, l_2, l_3, l_4) \in (A \cup \{\epsilon\})^4$ and $(x_a)_{a\in A} \in \mathbb{Z}^A$; $(\delta_1, \delta_2, \delta_3, \delta_4) \in \mathbb{N}^4$ and $\sigma \in S$. We say that ${}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}$ is a* substitute and cut operation *allowed on $X := \left(\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \end{pmatrix}, (x_a)_{a\in A}\right)$ if it satisfies the two conditions:*

$$1.\ \begin{cases} (\delta_1, \delta_2) = (0, 0) & \text{if } (l_1, l_2) = (\epsilon, \epsilon), \\ \delta_1 < |\sigma(l_1)|,\ \delta_2 < |\sigma(l_1)| \text{ and } \delta_1 + \delta_2 < |\sigma(l_1)| & \text{if } (l_1, l_2) \in A \times \{\epsilon\}, \\ \delta_1 < |\sigma(l_1)| \text{ and } \delta_2 < |\sigma(l_2)| & \text{if } (l_1, l_2) \in A^2; \end{cases}$$

$$2.\ \begin{cases} (\delta_3, \delta_4) = (0, 0) & \text{if } (l_3, l_4) = (\epsilon, \epsilon), \\ \delta_3 < |\sigma(l_3)|,\ \delta_4 < |\sigma(l_3)| \text{ and } \delta_3 + \delta_4 < |\sigma(l_3)| & \text{if } (l_3, l_4) \in A \times \{\epsilon\}, \\ \delta_3 < |\sigma(l_3)| \text{ and } \delta_4 < |\sigma(l_4)| & \text{if } (l_3, l_4) \in A^2. \end{cases}$$

In this case, the image of X by the substitute and cut operation ${}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}$ is well defined and equals:

$$\left(\begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix}, (y_a)_{a\in A}\right),$$

where m_1, m_2, m_3, m_4 and $(y_a)_{a\in A}$ are given by:

- $m_1 = \sigma(l_1)[\delta_1]$;
- $\begin{cases} m_2 = \sigma(l_1)[-\delta_2 - 1] & \text{if } l_2 = \epsilon, \\ m_2 = \sigma(l_2)[-\delta_2 - 1] & \text{otherwise;} \end{cases}$
- $m_3 = \sigma(l_3)[\delta_3]$;
- $\begin{cases} m_4 = \sigma(l_3)[-\delta_4 - 1] & \text{if } l_4 = \epsilon, \\ m_4 = \sigma(l_4)[-\delta_4 - 1] & \text{otherwise;} \end{cases}$

where, following `Python`*, $u[k]$ denotes the $(k+1)$-th letter of u and $u[-k]$ its k-th letter, reading backwards;*

- $\begin{cases} (y_a) = (x_a)M_\sigma - \text{ab}(p_{\delta_1}(\sigma(l_1))) - \text{ab}(s_{\delta_2}(\sigma(l_1))) + \text{ab}(p_{\delta_3}(\sigma(l_3))) + \text{ab}(s_{\delta_4}(\sigma(l_3))) & \text{if } (l_2, l_4) = (\epsilon, \epsilon) \\ (y_a) = (x_a)M_\sigma - \text{ab}(p_{\delta_1}(\sigma(l_1))) - \text{ab}(s_{\delta_2}(\sigma(l_1))) + \text{ab}(p_{\delta_3}(\sigma(l_3))) + \text{ab}(s_{\delta_4}(\sigma(l_4))) & \text{if } (l_2, l_4) \in \{\epsilon\} \times A \\ (y_a) = (x_a)M_\sigma - \text{ab}(p_{\delta_1}(\sigma(l_1))) - \text{ab}(s_{\delta_2}(\sigma(l_2))) + \text{ab}(p_{\delta_3}(\sigma(l_3))) + \text{ab}(s_{\delta_4}(\sigma(l_3))) & \text{if } (l_2, l_4) \in A \times \{\epsilon\} \\ (y_a) = (x_a)M_\sigma - \text{ab}(p_{\delta_1}(\sigma(l_1))) - \text{ab}(s_{\delta_2}(\sigma(l_2))) + \text{ab}(p_{\delta_3}(\sigma(l_3))) + \text{ab}(s_{\delta_4}(\sigma(l_4))) & \text{otherwise.} \end{cases}$

In the last point, (x_a) and (y_a) are line vectors indexed by A, and M_σ denotes the incidence matrix of the substitution σ, indexed by $A \times A$.

Notation 10. *Let α and ω be the functions defined by:*

$$\begin{array}{rccc} \alpha : & A^* & \to & A \cup \{\epsilon\} \\ & u & \mapsto & \begin{cases} u[0] \text{ if } |u| \geq 1 \\ \epsilon \text{ otherwise;} \end{cases} \end{array} \qquad \begin{array}{rccc} \omega : & A^* & \to & A \cup \{\epsilon\} \\ & u & \mapsto & \begin{cases} u[-1] \text{ if } |u| \geq 2 \\ \epsilon \text{ otherwise.} \end{cases} \end{array}$$

They are constructed so as to output the extremal letters of a word (when this makes sense), but never point to the same letter. For instance, $\alpha(examples) = e$, $\omega(examples) = s$, $\alpha(a) = a$, $\omega(a) = \epsilon$, $\alpha(\epsilon) = \omega(\epsilon) = \epsilon$.

Definition 11. *Let S be a finite set of nonerasing substitutions defined over a common alphabet A. Assume that each letter in A appears in a S-adic word (not necessarily the same for all letters). The* automaton of imbalances *of the S-adic system X_S is the infinite oriented graph $\mathtt{G}$ whose vertices are:*

$$\mathtt{V} = \bigcup_{w \in X_S} \left\{ \left(\begin{pmatrix} \alpha(u) & \omega(u) \\ \alpha(v) & \omega(v) \end{pmatrix}, \mathrm{ab}(u) - \mathrm{ab}(v) \right) \mid u, v \in \mathcal{F}(w) \right\},$$

and whose edges map vertices to each of their images by all allowed substitute and cut operations ${}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}$, with $\sigma \in S$.

A vertex $X = \left(\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \end{pmatrix}, (x_a)_{a \in A} \right)$ is a final state *if $\sum_{a \in A} x_a = 0$. The set of final states of $\mathtt{G}$ is denoted by $\mathtt{F}$.*

The set of initial states *of $\mathtt{G}$ is:*

$$\mathtt{I} := \mathtt{V} \cap \left\{ \{((\begin{smallmatrix} \epsilon & \epsilon \\ \epsilon & \epsilon \end{smallmatrix}), (0)_{a \in A})\} \cup \bigcup_{a \in A} \{((\begin{smallmatrix} a & \epsilon \\ \epsilon & \epsilon \end{smallmatrix}), \mathrm{ab}(a)); ((\begin{smallmatrix} \epsilon & \epsilon \\ a & \epsilon \end{smallmatrix}), -\mathrm{ab}(a)); ((\begin{smallmatrix} a & \epsilon \\ a & \epsilon \end{smallmatrix}), (0)_{a \in A})\} \right\}.$$

Remark 12. *Contrary to what its name suggests, the automaton of imbalances is not an automaton. Indeed, it has a infinite number of vertices (see Proposition 13 hereafter). Nonetheless, since we are interested in the labeling of the paths between a particular finite class of vertices (the "initial states") and another subclass -possibly infinite- (the "final states"), we choose to keep the name "automaton".*

3.3 Properties and consequences

Proposition 13. *Let S be a finite set of nonerasing substitutions over a common alphabet A. Denote by $\mathtt{G}$ the automaton of imbalances of X_S. If X_S is empty, so is $\mathtt{G}$. Otherwise:*

1. *there are infinitely many vertices in $\mathtt{G}$;*
2. *there exists $N \in \mathbb{N}$ such that all vertices have less than N outgoing edges;*
3. *there are exactly $3\mathrm{card}(A) + 1$ initial states.*

Proof. (1) Assume that X_S is nonempty and pick $w \in X_S$. Then for all integers $n \geq 2$,

$$X_n := \left(\begin{pmatrix} w[0] & w[n-1] \\ \epsilon & \epsilon \end{pmatrix}, \mathrm{ab}(p_n(w)) \right),$$

is a vertex of $\mathtt{G}$, corresponding to the pair of factors of w: $(p_n(w), \epsilon)$. Furthermore, since the sum of the coordinates of the vector $\mathrm{ab}(p_n(w))$ is equal to n (the length of $p_n(w)$), the vertices X_n, for n in $\mathbb{N}$, are pairwise distinct. Therefore, $\mathtt{G}$ contains an infinite number of vertices.
(2) Let $l = \max\{|\sigma(a)| \text{ for } a \in A, \sigma \in S\}$. Then for all $i \in \{1, 2, 3, 4\}$, $\delta_i < l$; so the number of outgoing edges from any vertex is bounded above by $\mathrm{card}(S) \times l^4$.
(3) Immediate. □

Proposition 14. *Let* $(l_1, l_2, l_3, l_4) \in (A \cup \{\epsilon\})^4$ *and* $(x_a)_{a\in A} \in \mathbb{Z}^A$; $(\delta_1, \delta_2, \delta_3, \delta_4) \in \mathbb{N}^4$ *and* $\sigma \in S$ *such that the substitute and cut operation* ${}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}$ *is allowed on* $X = \left(\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \end{pmatrix}, (x_a)_{a\in A}\right)$ *and has image* $Y = \left(\begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix}, (y_a)_{a\in A}\right)$.

Then for all $\tilde{u}, \tilde{v} \in A^*$ *such that:*

1. $(\alpha(\tilde{u}), \omega(\tilde{u}), \alpha(\tilde{v}), \omega(\tilde{v})) = (l_1, l_2, l_3, l_4)$,

2. $\mathrm{ab}(\tilde{u}) - \mathrm{ab}(\tilde{v}) = (x_a)_{a\in A}$,

the truncations u *and* v *of* $\sigma(\tilde{u})$ *and* $\sigma(\tilde{v})$ *defined by:*

$$\begin{cases} p_{\delta_1}(\sigma(\tilde{u})) \cdot u \cdot s_{\delta_2}(\sigma(\tilde{u})) = \sigma(\tilde{u}), \\ p_{\delta_3}(\sigma(\tilde{v})) \cdot v \cdot s_{\delta_4}(\sigma(\tilde{v})) = \sigma(\tilde{v}); \end{cases}$$

satisfy:

1. $(\alpha(u), \omega(u), \alpha(v), \omega(v)) = (m_1, m_2, m_3, m_4)$,

2. $\mathrm{ab}(u) - \mathrm{ab}(v) = (y_a)_{a\in A}$.

Proposition 14 rules the construction of the automaton of imbalances, as detailed in Section 4.

3.4 Example: the imbalance automaton of the Thue-Morse subshift

Consider X_S the S-adic system for $S = \{\sigma\}$, where σ is the Thue-Morse substitution:

$$\begin{array}{cccc} \sigma: & a & \mapsto & ab \\ & b & \mapsto & ba. \end{array}$$

Observe that:
i) the substitution σ is nonerasing;
ii) the set S generates two S-adic words (described in Example 2);
iii) the letters a and b both appear in these words.

In what follows, we are going to breadth first traverse the automaton of imbalances of X_S, from its initial states. In order to reduce the growth of the tree (strictly, it is a forest with one main tree), we will:
– consider vertices up to some symmetries (see Lemma 16);
– trim the branches that will never lead to a final state (see Lemma 15).
We will obtain a finite graph (see Figure 1), on which we will read the maximal imbalance of words in X_w: 2, thus retrieving a well-known result.

Denote by $\mathtt{G} = (\mathtt{V}, \mathtt{I}, \mathtt{F})$ the automaton of imbalances of X_S. It has 7 initial states, namely:

$$\begin{pmatrix} \epsilon & \epsilon \\ \epsilon & \epsilon \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} a & \epsilon \\ \epsilon & \epsilon \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \epsilon & \epsilon \\ a & \epsilon \end{pmatrix}\begin{pmatrix} -1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} a & \epsilon \\ a & \epsilon \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} b & \epsilon \\ \epsilon & \epsilon \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} \epsilon & \epsilon \\ b & \epsilon \end{pmatrix}\begin{pmatrix} 0 \\ -1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} b & \epsilon \\ b & \epsilon \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

(In this subsection, it will be more convenient to represent the second vector as a column instead of a line.) Each vertex in `V` admits at most 16 outgoing edges, among:

$$\left\{ {}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}, \text{ for } (\delta_1,\delta_2,\delta_3,\delta_4) \in \{0,1\}^4 \right\}.$$

We recall that a vertex Y is *accessible* from a vertex X if there exists a sequence of edges in `G` which leads from X to Y - or, in other words, if Y is the image of X by a finite composition of substitute and cut operations.

Lemma 15. *No vertex in* `F` *is accessible from a vertex of the form*

$$\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

with $|x_a + x_b| \geq 2$.

Therefore, during the breadth first traversal of `G`, we ignore all vertices such that $|x_a + x_b| \geq 2$: Lemma 15 guarantees that we miss no finite state.

Proof of Lemma 15. Let $X = \left(\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \end{pmatrix}, \begin{pmatrix} x_a \\ x_b \end{pmatrix} \right)$ be a vertex in `V`, ${}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}$ a substitute and cut operation allowed from X, and denote by $Y = \left(\begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix}, \begin{pmatrix} y_a \\ y_b \end{pmatrix} \right)$ its image. We are going to prove that $|y_a + y_b| \geq 2$. This first implies that $Y \notin$ `F` (remind that a state is final if an only if it satisfies $y_a + y_b = 0$), and, iteratively, that no final state is the image of X by a finite composition of substitute and cut operations.

Let $\tilde{u}, \tilde{v} \in \{a,b\}^*$ such that $(\alpha(\tilde{u}), \omega(\tilde{u}), \alpha(\tilde{v}), \omega(\tilde{v})) = (l_1, l_2, l_3, l_4)$ and $\mathrm{ab}(\tilde{u}) - \mathrm{ab}(\tilde{v}) = (x_a, x_b)$. We have $||\tilde{u}| - |\tilde{v}|| = |x_a + x_b| \geq 2$; without loss of generality, assume that $|\tilde{u}| - |\tilde{v}| \geq 2$. Then we have $|\sigma(\tilde{u})| - |\sigma(\tilde{v})| \geq 4$. Denote by u and v the truncations of $\sigma(\tilde{u})$ and $\sigma(\tilde{v})$ given by:

$$\begin{cases} p_{\delta_1}(\sigma(\tilde{u})) \cdot u \cdot s_{\delta_2}(\sigma(\tilde{u})) = \sigma(\tilde{u}), \\ p_{\delta_3}(\sigma(\tilde{v})) \cdot v \cdot s_{\delta_4}(\sigma(\tilde{v})) = \sigma(\tilde{v}). \end{cases}$$

Then, by Proposition 14, we have $\mathrm{ab}(u) - \mathrm{ab}(v) = (y_a, y_b)$, from which we deduce that $|y_a + y_b| = |u| - |v|$. Finally, observing that $(\delta_1, \delta_2, \delta_3, \delta_4) \in \{0,1\}^4$ for the Thue-Morse substitution, we deduce that: $|u| - |v| \geq 2$; this concludes the proof. □

Lemma 16 (Symmetries)**.** *Let* $\left(\begin{pmatrix} l_1 & l_2 \\ l_3 & l_4 \end{pmatrix}, \begin{pmatrix} x_a \\ x_b \end{pmatrix} \right) \in$ `V`. *Then the vertices*

$$\begin{pmatrix} l_3 & l_4 \\ l_1 & l_2 \end{pmatrix} \begin{pmatrix} -x_a \\ -x_b \end{pmatrix}, \quad \begin{pmatrix} \neg l_1 & \neg l_2 \\ \neg l_3 & \neg l_4 \end{pmatrix} \begin{pmatrix} x_b \\ x_a \end{pmatrix} \quad \textit{and} \quad \begin{cases} \begin{pmatrix} l_2 & l_1 \\ l_4 & l_3 \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix} & \textit{if } (l_2, l_4) \neq (\epsilon, \epsilon), \\ \begin{pmatrix} l_2 & l_1 \\ l_3 & l_4 \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix} & \textit{if } l_2 \neq \epsilon \textit{ and } l_4 = \epsilon, \\ \begin{pmatrix} l_1 & l_2 \\ l_4 & l_3 \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix} & \textit{if } l_2 = \epsilon \textit{ and } l_4 \neq \epsilon. \end{cases}$$

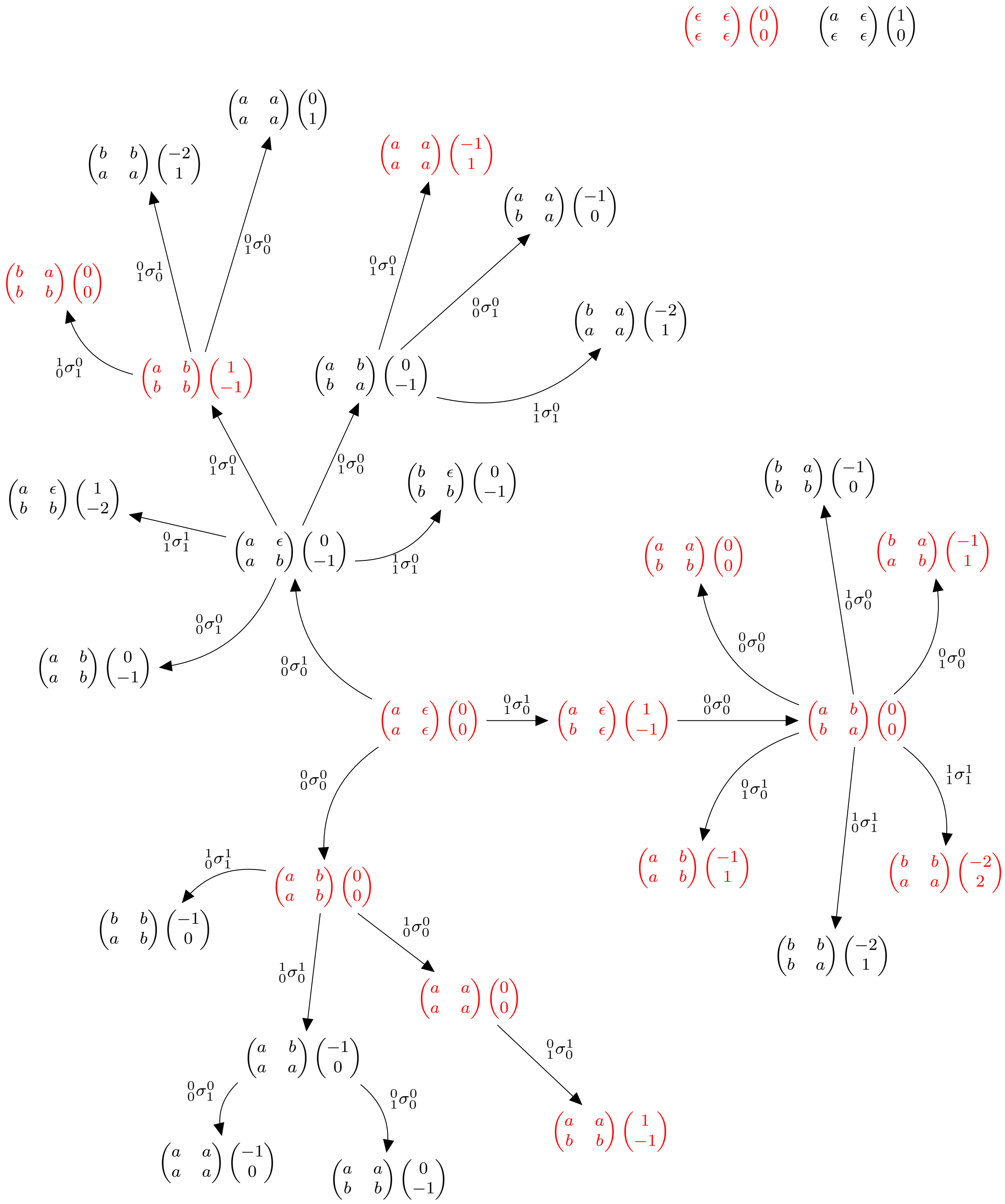
where $\neg a = b$ *and* $\neg b = a$, *also belong to* `V`.

Proof. Let $w \in X_S$ and $(u, v) \in \mathcal{F}(w)^2$ such that $(\alpha(u), \omega(u), \alpha(v), \omega(v)) = (l_1, l_2, l_3, l_4)$ and $\mathrm{ab}(u) - \mathrm{ab}(v) = (x_a, x_b)$. We immediately see that the pair (v, u) also belong to $\mathcal{F}(w)^2$, which proves that the first symmetric vertex is also in $\mathtt{V}$. One easily checks from Example 2 that if $w \in X_S$, then $\neg w$ also belongs to X_S. Since the pair $(\neg u, \neg v)$ belongs to $\mathcal{F}(\neg w)$, the second symmetric vertex is also in $\mathtt{V}$. At last, the Thue-Morse subshift is such that if $u \in \mathcal{F}(w)$, then $\mathrm{rev}(u)$, where $\mathrm{rev}(u)$ denotes the word u written backwards, is a factor of w too; we deduce that the third symmetric vertices belong to $\mathtt{V}$. □

Below, we draw the forest obtained after a breadth first traversal of $\mathtt{G}$, from its initial states. This forest is simplified by:

- considering the vertices up to the symmetries of Lemma 16 (as a consequence, there remain only three initial states);

- deleting all vertices such that $|x_a + x_b| \geq 2$ (see Lemma 15).


$\begin{pmatrix}\epsilon&\epsilon\\ \epsilon&\epsilon\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
$\begin{pmatrix}a&\epsilon\\ \epsilon&\epsilon\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}$
$\begin{pmatrix}a&a\\a&a\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}$
$\begin{pmatrix}b&b\\a&a\end{pmatrix}\begin{pmatrix}-2\\1\end{pmatrix}$
$\begin{pmatrix}a&a\\a&a\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}$
$\begin{pmatrix}a&a\\b&a\end{pmatrix}\begin{pmatrix}-1\\0\end{pmatrix}$
${}^0_1\sigma^0_0$
${}^0_1\sigma^1_0$
${}^0_1\sigma^0_1$
${}^0_0\sigma^0_1$
$\begin{pmatrix}b&a\\b&b\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
$\begin{pmatrix}b&a\\a&a\end{pmatrix}\begin{pmatrix}-2\\1\end{pmatrix}$
${}^1_0\sigma^0_1$
$\begin{pmatrix}a&b\\b&b\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}$
$\begin{pmatrix}a&b\\b&a\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}$
${}^1_1\sigma^0_1$
${}^0_1\sigma^0_1$
${}^0_1\sigma^0_0$
$\begin{pmatrix}b&\epsilon\\b&b\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}$
$\begin{pmatrix}b&a\\b&b\end{pmatrix}\begin{pmatrix}-1\\0\end{pmatrix}$
$\begin{pmatrix}a&\epsilon\\b&b\end{pmatrix}\begin{pmatrix}1\\-2\end{pmatrix}$
${}^0_1\sigma^1_1$
$\begin{pmatrix}a&\epsilon\\a&b\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}$
${}^1_1\sigma^0_1$
$\begin{pmatrix}a&a\\b&b\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
$\begin{pmatrix}b&a\\a&b\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}$
${}^1_0\sigma^0_0$
${}^0_0\sigma^0_1$
${}^0_0\sigma^0_0$
${}^0_1\sigma^0_0$
$\begin{pmatrix}a&b\\a&b\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}$
${}^0_0\sigma^1_0$
$\begin{pmatrix}a&\epsilon\\a&\epsilon\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
${}^0_1\sigma^1_0$
$\begin{pmatrix}a&\epsilon\\b&\epsilon\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}$
${}^0_0\sigma^0_0$
$\begin{pmatrix}a&b\\b&a\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
${}^0_0\sigma^0_0$
${}^1_1\sigma^1_1$
${}^0_1\sigma^1_0$
${}^1_0\sigma^1_1$
${}^1_0\sigma^1_1$
$\begin{pmatrix}a&b\\a&b\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}$
$\begin{pmatrix}b&b\\a&a\end{pmatrix}\begin{pmatrix}-2\\2\end{pmatrix}$
$\begin{pmatrix}a&b\\a&b\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
$\begin{pmatrix}b&b\\a&b\end{pmatrix}\begin{pmatrix}-1\\0\end{pmatrix}$
${}^1_0\sigma^0_0$
${}^1_0\sigma^1_0$
$\begin{pmatrix}a&a\\a&a\end{pmatrix}\begin{pmatrix}0\\0\end{pmatrix}$
$\begin{pmatrix}b&b\\b&a\end{pmatrix}\begin{pmatrix}-2\\1\end{pmatrix}$
$\begin{pmatrix}a&b\\a&a\end{pmatrix}\begin{pmatrix}-1\\0\end{pmatrix}$
${}^0_1\sigma^1_0$
${}^0_0\sigma^0_1$
${}^0_1\sigma^0_0$
$\begin{pmatrix}a&a\\b&b\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}$
$\begin{pmatrix}a&a\\a&a\end{pmatrix}\begin{pmatrix}-1\\0\end{pmatrix}$
$\begin{pmatrix}a&a\\b&b\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix}$

In red: final states.

Figure 1: The forest obtained after a breadth first traversal of G for the Thue-Morse subshift.

The maximal imbalance found is 2 and corresponds to the vertex $\left(\begin{smallmatrix} b & b \\ a & a \end{smallmatrix}\right)\left(\begin{smallmatrix} -2 \\ 2 \end{smallmatrix}\right)$.

4 Proof

In this Section, we describe step by step the construction of the automaton of imbalances.

4.1 General idea

Let S be a finite set of nonerasing substitutions over a common alphabet A. We consider the set of all imbalance vectors of words in X_S:

$$\mathtt{F}_3 := \bigcup_{w \in X_S} \{\mathrm{ab}(u) - \mathrm{ab}(v) \mid u, v \in \mathcal{F}(w) \text{ and } |u| = |v|\} \subset \mathbb{Z}^A.$$

Proposition 17 (immediate)**.** *The set* $\mathtt{F}_3$ *is finite if and only if there exists* $D \in \mathbb{N}$ *such that for all* $w \in X_S$*,* $\mathrm{imb}(w) \leq D$*.*

So we want to explore the set $\mathtt{F}_3$. To do so, we need to understand where do the imbalance vectors come from. A classic idea, in the substitutive framework, and we extend to S-adic systems, consists in tracking backwards the history of factors by successive desubstitutions. This is why we introduce the underlying set:

$$\mathtt{F}_2 := \bigcup_{w \in X_S} \{(u, v) \mid u, v \in \mathcal{F}(w) \text{ and } |u| = |v|\},$$

that we will explore through a (huge) graph $\mathtt{G}_2$, whose automaton of imbalances $\mathtt{G}$ is a quotient graph, obtained after partial abelianization.

4.2 Desubstitution

Let $u, \tilde{u} \in A^*$ and σ a substitution over A. We say that $\tilde{u}$ is a *predecessor* of u by σ if there exists a pair $p, s \in A^*$ such that $pus = \sigma(\tilde{u})$. We denote by $\mathcal{P}_\sigma(u)$ the set of all predecessors of u by σ, that we endow with the relation:

$$\tilde{u}_1 \leq \tilde{u}_2 \text{ if and only if } \tilde{u}_1 \in \mathcal{F}(\tilde{u}_2).$$

The set $\mathcal{P}_\sigma(u)$ is not totally ordered: for instance 1 and 22 are two incomparable predecessors of 2 for the Thue-Morse substitution σ_{TM} (defined in Example 2). However, the partial order is well-founded: all nonempty subset of $\mathcal{P}_\sigma(u)$ admit minimal elements. *Desubstituting* u by σ consists in returning one of its minimal predecessors.

Example 18. *The set of all predecessors of* 2 *by* σ_{TM} *is* $\{1, 2\}^* \backslash \{\epsilon\}$*; its minimal predecessors are* 1 *and* 2*. The word* 222 *has no predecessor by* σ_{TM}*, therefore, it can not be desubstituted.*

Observe that we always have $\mathcal{P}_\sigma(\epsilon) = A^*$, and that the only minimal predecessor of ϵ is itself.

Proposition 19. *Let* $u \in A^* \backslash \{\epsilon\}$ *and* $\tilde{u} \in \mathcal{P}_\sigma(u)$*. Then* $\tilde{u}$ *is nonempty. It is furthermore minimal if and only if all pairs* $(p, s) \in A^*$ *such that* $pus = \sigma(\tilde{u})$ *satisfy* $|p| < |\sigma(\tilde{u}[0])|$ *and* $|s| < |\sigma(\tilde{u}[-1])|$ $(\star)$*. Remainder:* $\tilde{u}[0]$ *and* $\tilde{u}[-1]$ *denote the first and last (possibly same) letter of* $\tilde{u}$*.*

Proof. We immediately see that if u is nonempty, then $\tilde{u}$ is also nonempty. We now prove both implications by contraposition.

Assume that $\tilde{u}$ is not minimal. There exists $\tilde{v} \in \mathcal{P}_\sigma(u)$ and $a, b \in A^*$ non both empty (say, without loss of generality, $a \neq \epsilon$), such that $\tilde{u} = a\tilde{v}b$. Then, by writing $\sigma(\tilde{v}) = pus$, we obtain: $\sigma(\tilde{u}) = \sigma(a)p \cdot u \cdot s\sigma(b)$, with $|\sigma(a)p| \geq |\sigma(a)| \geq |\sigma(\tilde{u}[0])|$, since $a \neq \epsilon$.

Conversely, assume that there exists a pair $p, s \in A^*$ which does not satisfy ($\star$) (say, without loss of generality, $|p| \geq |\sigma(\tilde{u}[0])|$). One easily checks that the suffix of length $|\tilde{u}| - 1$ of $\tilde{u}$ is still a predecessor of u, which implies that $\tilde{u}$ is not minimal. □

Counterintuitively, the existence of one pair satisfying ($\star$) is not sufficient to guarantee the minimality of a predecessor.

Example 20. *Consider the substitution σ_{cex} defined by $\sigma_{cex}(0) = 10$ and $\sigma_{cex}(1) = 1010$. Denote $\tilde{u} = 101$, $u = 0101$, $p = 101$ and $s = 010$. Observe that:*

$$\begin{cases} \sigma_{cex}(\tilde{u}) = 1010101010 = pus \\ |p| < \sigma(\tilde{u}[0]) \\ |s| < \sigma(\tilde{u}[-1]) \end{cases}$$

and, yet, $\tilde{u}$ is not minimal since its strict factor 10 *is also a predecessor of u:*

$$\sigma_{cex}(10) = 101010 = 1u0.$$

Corollary 21. *Let σ be a substitution and $u \in \mathcal{F}(\sigma(w)) \backslash \{\epsilon\}$, where w is a finite or infinite word. Then there exist $\tilde{u} \in \mathcal{F}(w) \backslash \{\epsilon\}$ together with a pair p, s of finite words satisfying ($\star$) such that $pus = \sigma(\tilde{u})$.*

Proof. The set $\mathcal{P}_\sigma(u)$ is nonempty (indeed, it contains w if w is finite, or one of its factors otherwise), so it admits minimal elements. We conclude with Proposition 19. □

Finally, if u and v are two factors of a S-adic word $w = \lim_{n\to\infty} d_0 \circ ... \circ d_{n-1}(a)$, we can simultaneously desubstitute u and v by d_0; unfortunately, there is no guarantee that, among the pairs of minimal predecessors $(\tilde{u}, \tilde{v})$, there is one at least which satisfies $\tilde{u} = \tilde{v}$, or equivalently, which still belongs to $\mathtt{F}_2$. So we introduce the larger set:

$$\mathtt{V}_2 = \bigcup_{w \in X_S} \{(u, v) \,|\, u, v \in \mathcal{F}(w)\} \subset A^* \times A^*,$$

that will be the vertices of the (huge) graph we construct.

4.3 An inverse to desubstitution: the *substitute and cut* operation

In order to explore the set of factors in a constructive way, we choose for transition not the desubstitution, but its "inverse", which consists in substituting and cropping the image according to some rules.

In the sequel, we refer to the functions α and ω defined in Notation 10.

Definition 22. *Let u and $\tilde{u}$ in A^*. A* substitute and cut operation *from $\tilde{u}$ to u is a triplet $(\sigma, \delta_1, \delta_2)$, where σ is a substitution and δ_1,δ_2 two nonnegative integers such that:*

1. $p_{\delta_1}(\sigma(\tilde{u})) \cdot u \cdot s_{\delta_2}(\sigma(\tilde{u})) = \sigma(\tilde{u})$.

2. $\begin{cases} \delta_1 = \delta_2 = 0 & \text{if } \tilde{u} = \epsilon, \\ \delta_1 < |\sigma(\tilde{u})|,\ \delta_2 < |\sigma(\tilde{u})| \text{ and } \delta_1 + \delta_2 < |\sigma(\tilde{u})| & \text{if } |\tilde{u}| = 1, \\ \delta_1 < |\sigma(\alpha(\tilde{u}))| \text{ and } \delta_2 < |\sigma(\omega(\tilde{u}))| & \text{otherwise.} \end{cases}$

We denote it by:

$$^{\delta_1}\sigma^{\delta_2}(\tilde{u}) = u.$$

Proposition - definition 23. *Let $\tilde{u} \in A^*$, σ a substitution defined over A, and δ_1, $\delta_2 \in \mathbb{N}$. The two assertions are equivalent:*

1. *$\tilde{u}$, σ, δ_1 and δ_2 satisfy Condition (2) of Definition 22;*

2. *there exists $u \in A^*$ such that $u = {}^{\delta_1}\sigma^{\delta_2}(\tilde{u})$.*

When these assertions are satisfied, we say that $(\sigma, \delta_1, \delta_2)$ is a substitute and cut operation allowed *from $\tilde{u}$.*

Proof. Since (2) trivially implies (1), we just need to check that (1)⇒(2). Let us assume that $\tilde{u}$, σ, δ_1 and δ_2 satisfy Condition (2) of Definition 22. If $\tilde{u} = \epsilon$, it suffices to take $u = \epsilon$. Otherwise, the bounds on δ_1 and δ_2 ensure that the equation: $p_{\delta_1}(\sigma(\tilde{u})) \cdot u \cdot s_{\delta_2}(\sigma(\tilde{u})) = \sigma(\tilde{u})$ defines a (unique and nonempty) word u. □

This definition is consistent with Definition 9, as we see later.

Example 24. *One easily checks that:*

1. *the triplet $(\sigma_{TM}, 1, 1)$ (with σ_{TM} the Thue-Morse substitution defined in Subsection 2.1) is a substitute and cut operation from* 21 *to* 11*;*

2. *the triplet $(\sigma_{cex}, 3, 3)$ (where σ_{cex} is the substitution defined in Example 20) is a substitute and cut operation from* 101 *to* 0101*.*

Example 24(2) shows that a finite word is not always a minimal predecessor of its substitute and cut image. Nonetheless:

Lemma 25. *Let $u \in A^*$ and σ a substitution over A. The set of all preimages of u by allowed substitute and cut operations built over σ is finite and contains all minimal predecessors of u by σ.*

Proof. Let $u \in A^*$, σ a substitution and (δ_1, δ_2) a pair of integer such that $(\sigma, \delta_1, \delta_2)$ is a substitute and cut operation allowed on $\tilde{u}$. Then we must have $\delta_1, \delta_2 \leq \max_{a \in A} |\sigma(a)|$; hence the finiteness of the set. Now, if $\tilde{u}$ is a minimal predecessor of u by σ, either $u = \tilde{u} = \epsilon$ and we have ${}^0\sigma^0(\epsilon) = \epsilon$, or $u \neq \epsilon$ and we know from Proposition 19 that all pairs of words p, s such that $\sigma(\tilde{u}) = pus$ will satisfy Condition (2), so we can write $u = {}^{|p|}\sigma^{|s|}(\tilde{u})$. □

4.4 Aside: the graph of all factors of a S-adic system

Let S be a set of substitutions defined over a common alphabet A. Denote by $\mathtt{V}_1$ the set of all factors of the S-adic system X_S:

$$\mathtt{V}_1 := \bigcup_{w \in X_S} \mathcal{F}(w) = \bigcup_{w \text{ S-adique}} \mathcal{F}(w) \qquad \text{(see Lemma 6).}$$

We consider the oriented graph $\mathtt{G}_1$ whose set of vertices is $\mathtt{V}_1$, and whose edges map vertices to their images by all allowed substitute and cut operations $(\sigma, \delta_1, \delta_2)$, with $\sigma \in S$.

Remark 26. *In the particular case $S = \{\sigma\}$, with σ a primitive substitution, the subgraph of $\mathtt{G}_1$ restricted to the vertices of length 1 (i.e. letters in A) is the automaton of prefixes-suffixes, with reversed edges, defined in [CS01].*

Lemma 27 (immediate). *(i) If it is not empty, the graph $\mathtt{G}_1$ has infinitely many vertices.*
(ii) The number of outgoing edges of any vertex is bounded above by $\mathrm{card}(S) \times l^2$*, where* $l = \max_{\sigma \in S, a \in A} |\sigma(a)|$.

We say that a factor $v \in \mathtt{V}_1$ is *accessible* from a factor $u \in \mathtt{V}$ if there exists a finite path in $\mathtt{G}$ which goes from u to v.

Proposition 28. *If $A \subset \mathtt{V}_1$, then any nonempty factor in $\mathtt{V}_1$ is accessible from a letter in A.*

Proof. Assume that $A \subset \mathtt{V}_1$ and pick a nonempty factor u in $\mathtt{V}_1$. By definition, there exists a letter $a \in A$ and a directive sequence $(d_n)_{n \in \mathbb{N}} \in S^{\mathbb{N}}$ such that $u \in \mathcal{F}(w)$ with $w = \lim_n d_0 \circ ... \circ d_{n-1}(a)$. Thus, there exists a [finite] prefix v of w such that $u \in \mathcal{F}(v)$; and, by definition of the convergence (see Subsection 2.1), there exists a rank $n_0 \in \mathbb{N}$ such that v is also a prefix of the finite words $d_0 \circ ... \circ d_{n-1}(a)$ for all n larger than n_0.

Now, we recursively construct three finite sequences: $(u_k)_{k \in \{0,...,n_0\}} \in \mathtt{V}^{n_0+1}$, $(\beta_k)_{k \in \{0,...,n_0-1\}}$ and $(\gamma_k)_{k \in \{0,...,n_0-1\}} \in \mathbb{N}^{n_0}$:

- $u_0 = u$,
- for $k \in \{0, ..., n_0 - 1\}$, we denote by $(p, \tilde{u}, s)$ the triplet of words given by the application of Corollary 21 to the substitution d_k and the word $u_k \in \mathcal{F}(d_k(w_{k+1}))$, with $w_{k+1} := d_{k+1} \circ ... \circ d_{n_0-1}(a)$; we then set $u_{k+1} = \tilde{u}$, $\beta_k = |p|$ and $\gamma_k = |s|$.

Thus, the sequences $(u_k)_{k \in \{0,...,n_0\}}$, $(\beta_k)_{k \in \{0,...,n_0-1\}}$ and $(\gamma_k)_{k \in \{0,...,n_0-1\}}$ are such that for all $k \in \{0, ..., n_0 - 1\}$, u_k is the image of u_{k+1} by the allowed substitute and cut operation (d_k, β_k, γ_k). Furthermore, Corollary 21 ensures that at each step of the recursion, $u_k \in \mathcal{F}(d_k \circ ... \circ d_{n_0-1}(a)) \backslash \{\epsilon\}$; so in particular $u_{n_0} \in \mathcal{F}(a) \backslash \{\epsilon\}$, which implies $u_{n_0} = a$.

Finally, we exhibited a finite path in $\mathtt{G}$, which goes from $a \in A \subset \mathtt{V}_1$ to u; this concludes the proof. □

This result suggests that we could obtain all factors in $\mathtt{V}_1$ by simple knowledge of the alphabet. Hence the question: are all words obtained after a finite composition of allowed substitute and cut operations still in $\mathtt{V}_1$?

Proposition 29. *If $A \subset \mathtt{V}_1$ and if u is the image of a letter by a finite composition of allowed substitute and cut operations, then $u \in \mathtt{V}_1$.*

Proof. Let $a \in A$ and $(\sigma_k, \beta_k, \gamma_k)_{k \in \{0,...,n-1\}} \in (S \times \mathbb{N}^2)^n$ such that for all $k \in \{0, ..., n_0 - 1\}$, the substitute and cut operation $(\sigma_k, \beta_k \gamma_k)$ is allowed on the word ${}^{\beta_{k-1}}\sigma_{k-1}{}^{\gamma_{k-1}} \circ ... \circ {}^{\beta_0}\sigma_0^{\gamma_0}(a)$. Denote $u := {}^{\beta_{n-1}}\sigma_{n-1}{}^{\gamma_{n-1}} \circ ... \circ {}^{\beta_0}\sigma_0^{\gamma_0}(a) \in A^*$; we want to prove that $u \in \mathtt{V}_1$. Since $a \in A \subset \mathtt{V}_1$, there exists w a S-adic word such that $a \in \mathcal{F}(w)$. But then, the infinite word $w' := \sigma_{n-1} \circ ... \circ \sigma_0(w)$ is also S-adic, and we immediately have $u \in \mathcal{F}(w')$, hence $u \in \mathtt{V}_1$. □

The assumption $A \subset \mathtt{V}_1$ is essential.

Example 30. *Consider $S = \{\sigma\}$, where the substitution σ is defined by $\sigma(a) = baa$ and $\sigma(b) = bb$. One easily checks that the S-adic system X_S contains a unique word, which is:*

$$w = bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb...$$

Therefore:

$$\mathtt{V}_1 = \{\epsilon, b, bb, bbb, bbbb, bbbbb, ...\}$$

and yet, the words a, baa, baaba, etc, can be obtained from letters by finite compositions of allowed substitute and cut operations.

4.5 The graph of all pairs of factors of S-adic words

4.5.1 Simultaneous substitute and cut operations

In order to study the imbalances of S-adic words, we do not want to reconstruct all pairs of factors in X_S, but all pairs of words (u, v) in which u and v are *simultaneous* factors of a S-adic word.

Definition 31. *Let $u, v, \tilde{u}$ and $\tilde{v}$ in A^*. A* [simultaneous] substitute and cut operation *from the pair $(\tilde{u}, \tilde{v})$ to the pair (u, v) is a quintuplet $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4)$ such that:*

- *$(\sigma, \delta_1, \delta_2)$ is a substitute and cut operation from $\tilde{u}$ to u,*
- *$(\sigma, \delta_3, \delta_4)$ is a substitute and cut operation from $\tilde{v}$ to v.*

We denote it by:

$$^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}(\tilde{u}, \tilde{v}) = (u, v).$$

A quintuplet $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4) \in S \times \mathbb{N}^4$ is an allowed substitute and cut operation from $(\tilde{u}, \tilde{v})$ if there exists a pair $(u, v) \in (A^*)^2$ such that $(u, v) = {}^{\delta_1}_{\delta_3}\sigma^{\delta_2}_{\delta_4}(\tilde{u}, \tilde{v})$.

The crucial point in Definition 31 is that we run the same substitution on both words, but we crop their images independently- these variations will generate imbalance.

Lemma 32. *Let $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4) \in S \times \mathbb{N}^4$ and $(\tilde{u}, \tilde{v}) \in (A^*)^2$. The following assertions are equivalent:*

1. *the quintuplet $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4)$ is an allowed substitute and cut operation from $(\tilde{u}, \tilde{v}) \in (A^*)^2$;*
2. *the triplets $(\sigma, \delta_1, \delta_2)$ and $(\sigma, \delta_3, \delta_4)$ are allowed on $\tilde{u}$ and $\tilde{v}$ respectively;*
3. *both triplets $(\sigma, \delta_1, \delta_2)$ and $(\sigma, \delta_3, \delta_4)$ satisfy Condition (2) in Definition 22.*

Proof. This comes immediately from Proposition-definition 23. □

Example 33. *We consider the set $C = \{c1, c2\}$ of Cassaigne-Selmer substitutions, defined in Example-Definition 4. Let $(\tilde{u}, \tilde{v}) = (23, 33)$.*
There are exactly four substitute and cut operations allowed from the pair $(\tilde{u}, \tilde{v})$:

$$^0_0c1^0_0 \quad , \quad ^1_0c1^0_0 \quad , \quad ^0_0c2^0_0 \quad \textit{and} \quad ^1_0c2^0_0.$$

Their respective images are: $(132, 22)$, $(32, 22)$,$(133, 33)$ *and* $(33, 33)$.

4.5.2 The graph of all pairs of factors of S-adic words

Let $\mathtt{G}_2$ be the oriented graph whose set of vertices is

$$\mathtt{V}_2 := \bigcup_{w \in X_S} \{(u,v)\,|u,v \in \mathcal{F}(w)\} = \bigcup_{w \text{ S-adique}} \{(u,v)\,|u,v \in \mathcal{F}(w)\}$$

and whose edges map vertices to their images by all allowed [simultaneous] substitute and cut operations.

We consider two remarkable subsets of $\mathtt{V}_2$:

$$\mathtt{I}_2 := \mathtt{V}_2 \cap \left(\{(\epsilon,\epsilon)\} \cup \cup_{a\in A}\{(a,\epsilon),(\epsilon,a),(a,a)\}\right), \qquad \text{called } \textit{initial states};$$

$$\mathtt{F}_2 := \bigcup_{w \text{ S-adique}} \{(u,v)\,|\,u,v \in \mathcal{F}(w) \text{ and } |u| = |v|\}, \qquad \text{called } \textit{finite states}.$$

Finite states corresponds to the pairs of factors we are interested in when studying the imbalances of infinite words in a S-adic system. To access them, Proposition 34 (below) suggests to traverse the graph $\mathtt{G}$ from its initial states.

Proposition 34. *1. If $\mathtt{G}_2$ is nonempty, then it has infinitely many vertices.*

2. The number of outgoing edges from any vertex is bounded above by $\mathrm{card}(S) \times l^4$*, where* $l = \max_{\sigma\in S, a\in A} |\sigma(a)|$.

3. If $A \subset \mathtt{V}_2$, then any pair of nonempty words in $\mathtt{V}_2$ is accessible from an initial state.

4. If $A \subset \mathtt{V}_2$ and if (u,v) is the image of an initial state by a finite composition of allowed [simultaneous] substitute and cut operations, then $(u,v) \in \mathtt{V}_2$.

Proof. Assertions (1) and (2) are immediate adaptations of Lemma 27.
Assertions (3) and (4) are easy adaptations of Propositions 28 and 29 respectively. The key point is to remember that $(u,v) \in \mathtt{V}_2$ means that u and v are factors of the same S-adic word w, so they can be simultaneously desubstituted by the first substitution of its directive sequence; conversely, the image of a pair in $\mathtt{V}_2$ by an allowed simultaneous substitute and cut operation is still made of factors of a same S-adic word. □

4.5.3 A semi-algorithm to detect imbalances in a S-adic system

Below, we explicitly describe a semi-algorithm to detect high imbalances in X_S.

Algorithm 35.
INPUTS:
- a finite set of substitutions S over a common alphabet A
- a nonnegative integer d

ALGORITHM

```
V ← $\mathtt{I}_2$ (initialization)
V' ← empty set
for $(\tilde{u},\tilde{v})$ in V:
    for $(\sigma,\delta_1,\delta_2,\delta_3,\delta_4)$ substitute and cut operation allowed on (u,v):
        (u,v) ← image of $(\tilde{u},\tilde{v})$ by the substitute and cut operation
        if (u,v) has not yet appeared in the processing:
```

```
        if $|u| = |v|$ and $\mathrm{imb}(u,v) \geq d$:
            return true
        otherwise:
            put (u,v) in V'
V ← V'
V' ← empty set
```

Corollary 36 (immediate consequence of Proposition 34)**.** *Let S be a finite set of nonerasing substitutions defined over a common alphabet A. Assume furthermore that $A \subset \mathtt{V}_2(S)$. Then for all $d \leq \sup_{w \in X_S} \mathrm{imb}(w)$, Algorithm 35, run on the entry (S, d), will finish.*

Remark 37. *If at each first visit of a pair (u, v), we record the preimage and the substitute and cut operation which gave (u, v), then we are able to retrace a shortest [finite] sequence of substitute and cut operations which lead from one of the initial states to the intended imbalance. Thus, if we denote by $(\sigma_0, ..., \sigma_{n-1})$ the substitutions appearing in the labels of this path, and by $\tau := \sigma_{n-1} \circ ... \circ \sigma_0$ the substitution obtained by backwards composition of them, then for all S-adic word w containing the letter a, where a is the letter appearing in the first [initial] state of the path, the imbalance of the S-adic word $\tau(w)$ is larger or equal than d.*

Remark 38. *By contrast, if $d > \sup_{w \in X_S} \mathrm{imb}(w)$, the algorithm will indefinitely keep running on the entry (S, d), hence the terminology "semi-algorithm".*

4.6 Construction of the automaton of imbalances

However, to study the imbalance of a S-adic word, it is not necessary to fully describe its factors: the simple knowledge of their abelianized vectors is sufficient. Put in another way, the graph $\mathtt{G}_2$ contains much more information than what is actually needed.

The aim of this subsection is to simplify it. This simplication (and some others...) will make its computation possible (see Section 5).

4.6.1 Obstructions to the abelianization of $\mathtt{G}_2$

We consider the binary relation $\mathcal{R}_{ab}$ defined over pairs of finite words:

$$(u, v)\, \mathcal{R}_{ab}\, (u', v') \text{ if and only if } \mathrm{ab}(u) - \mathrm{ab}(v) = \mathrm{ab}(u') - \mathrm{ab}(v'),$$

which partitions $\mathtt{V}_2$ into equivalence classes:

$$\mathtt{V}_3 := \mathtt{V}_2/\mathcal{R}_{ab} = \bigcup_{w \text{ S-adic}} \{\mathrm{ab}(u) - \mathrm{ab}(v) \mid u, v \in \mathcal{F}(w)\}.$$

Unfortunately, this partition is not compatible with the substitute and cut operation. Indeed, a same vector $x \in \mathtt{V}_3$ may represent two pairs of factors (u, v) and (u', v') for which the sets of allowed substitute and cut operations (and the results they give) are different (see Example 39 below).

Example 39. *We consider the set C of Cassaigne-Selmer substitutions, defined in Example-Definition 4. The two pairs $(u, v) = (132, 132)$ and $(u', v') = (\epsilon, \epsilon)$ in $\mathtt{V}_2$ both belong to the equivalence class $x = (0, 0, 0) \in \mathtt{V}_3$. However:*
- there are exactly eight substitute and cut operations allowed from the pair (u, v):

$$ {}^0_0c1^0_0 \quad , \quad {}^0_0c1^1_0 \quad , \quad {}^0_0c1^0_1 \quad , \quad {}^0_0c1^1_1$$
$$ {}^0_0c2^0_0 \quad , \quad {}^0_0c2^1_0 \quad , \quad {}^0_0c2^0_1 \quad \textit{and} \quad {}^0_0c2^1_1;$$

- only the first and the fifth ones are also allowed on (u', v').

Fundamentally, a vector $x \in \mathtt{V}_3$ contains no information on the initial and final letters of the pairs of factors it represents. This information is yet essential when applying Condition (2) of Definition 22.

4.6.2 Towards a partial abelianization of $\mathtt{G}_2$

This is why we consider a finer partition of $\mathtt{V}_2$. Let $\mathcal{R}_{pab}$ be the binary relation defined on pairs of finite words by:

$$(u,v)\,\mathcal{R}_{pab}\,(u',v') \text{ if and only if } \begin{cases} (\alpha(u),\omega(u),\alpha(v),\omega(v) = \alpha(u'),\omega(u'),\alpha(v'),\omega(v')) \\ \quad \text{and} \\ \mathrm{ab}(u)-\mathrm{ab}(v) = \mathrm{ab}(u')-\mathrm{ab}(v'). \end{cases}$$

We recall

Notation 10. *The functions α and ω are defined by:*

$$\begin{array}{rccl} \alpha: & A^* & \to & A\cup\{\epsilon\} \\ & u & \mapsto & \begin{cases} u[0] \ \textit{if}\ |u|\geq 1 \\ \epsilon \ \textit{otherwise;} \end{cases} \end{array} \qquad \begin{array}{rccl} \omega: & A^* & \to & A\cup\{\epsilon\} \\ & u & \mapsto & \begin{cases} u[-1] \ \textit{if}\ |u|\geq 2 \\ \epsilon \ \textit{otherwise;} \end{cases} \end{array}$$

where, following `Python`*, $u[0]$ and $u[-1]$ respectively denote the first and last letters of u. For instance, $\alpha(examples) = e$, $\omega(examples) = s$, $\alpha(a) = a$, $\omega(a) = \epsilon$, $\alpha(\epsilon) = \omega(\epsilon) = \epsilon$.*

The functions α and ω are constructed so as to save all the information necessary to apply substitute and cut operations.

Notation 40. *Let $(u,v) \in (A^*)^2$. The 2×2 table*

$$M_{u,v} := \begin{pmatrix} \alpha(u) & \omega(u) \\ \alpha(v) & \omega(v) \end{pmatrix}$$

is called matrix of extremities *of the pair (u,v).*

Example 41. *Let $A = \{a,b,c\}$.*
The pairs of words (acb, cc) and $(acab, cac)$ both belong to the equivalence class $\left(\begin{pmatrix} a & b \\ c & c \end{pmatrix}, (1,1,-1)\right)$.
The pair of words (a, cc) belongs to the class $\left(\begin{pmatrix} a & \epsilon \\ c & c \end{pmatrix}, (1,0,-2)\right)$.
The pair of words (acb, acb) belongs to the class $\left(\begin{pmatrix} a & b \\ a & b \end{pmatrix}, (0,0,0)\right)$, whereas the pair (ϵ,ϵ) belongs to the class $\left(\begin{pmatrix} \epsilon & \epsilon \\ \epsilon & \epsilon \end{pmatrix}, (0,0,0)\right)$. ‘

Let $\mathtt{V}$ denote the quotient set:

$$\mathtt{V} := \mathtt{V}_2/\mathcal{R}_{pab} = \bigcup_{w \text{ S-adic}} \{(M_{u,v}, \mathrm{ab}(u)-\mathrm{ab}(v)) \,|\, u,v \in \mathcal{F}(w)\}\,.$$

The [simultaneous] substitute and cut operation is compatible with the relation $\mathcal{R}_{pab}$.

Proposition 42. *Let S be a finite set of* ***nonerasing*** *substitutions defined over a common alphabet A. Let $X \in \mathtt{V}$. Let (u,v) and (u',v') be two pairs of words in the equivalence class X.*

i) A substitute and cut operation $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4) \in S \times \mathbb{N}^4$ is allowed from (u, v) if and only if it is allowed from (u', v'); in that case, we simply say that $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4)$ is allowed from X.

ii) Furthermore, the images of (u, v) and (u', v') by such a substitute and cut operation belong to a same equivalence class modulo $\mathcal{R}_{pab}$.

At last, the set of all allowed substitute and cut operations from X, as well as the equivalence classes of their images, can be computed from the simple knowledge of X.

It is essential for the substitutions to be non erasing.

Example 43. *Let σ be the [erasing] substitution:*

$$\begin{array}{cccc} \sigma & a & \mapsto & acb \\ & b & \mapsto & b \\ & c & \mapsto & \epsilon. \end{array}$$

We consider the two pairs of words $(\tilde{u}, \tilde{v}) = (cbab, \epsilon)$ and $(\tilde{u}', \tilde{v}') = (cabb, \epsilon)$, which belong to the same equivalence class modulo $\mathcal{R}_{pab}$, given by:

$$\left(\begin{pmatrix} c & b \\ \epsilon & \epsilon \end{pmatrix} (1, 2, 1) \right).$$

Their respective images by the allowed substitute and cut operation ${}^0_0\sigma^0_0$ are $(u, v) = (bacbb, \epsilon)$ and $(u', v') = (acbbb, \epsilon)$, which do not have the same matrix of extremities.

Proof of Proposition 42. i) Observe that the knowledge of $\alpha(u)$ and $\omega(u)$ in Condition (2) of Definition 22 is sufficient to determine the set of substitute and cut operations which are allowed from a word u. Therefore, the knowledge of $M_{u,v}$ is sufficient to determine the set of allowed simultaneous substitute and cut operations from a pair (u, v). Therefore, two pairs of words equal modulo $\mathcal{R}_{pab}$ must have the same set of allowed substitute and cut operations.

ii) Let $(\sigma, \delta_1, \delta_2)$ be an allowed substitute and cut operation from a word $\tilde{u}$. Denote $u = {}^{\delta_1}\sigma^{\delta_2}$. We first show that the knowledge of $\alpha(\tilde{u})$ and $\omega(\tilde{u})$ is sufficient to determine $\alpha(u)$ and $\omega(u)$.
- If $\omega(\tilde{u}) = \epsilon$, then $\tilde{u} \in A \cup \{\epsilon\}$, and the knowledge of $\alpha(\tilde{u}) = \tilde{u}$ is sufficient to determine the whole image $u = {}^{\delta_1}\sigma^{\delta_2}(\tilde{u})$, so in particular $\alpha(u)$ and $\omega(u)$.
- Otherwise, we know that $|\tilde{u}| \geq 2$. In this case, $l_1 = \alpha(\tilde{u})$ and $l_2 = \omega(\tilde{u})$ are respectively the first and last letter of $\tilde{u}$. Since σ is nonerasing, we have $|\sigma(l_1)| \geq 1$ and $|\sigma(l_2)| \geq 1$; since furthermore $\delta_1 < |\sigma(l_1)|$ and $\delta_2 < |\sigma(l_2)|$, the words ${}^{\delta_1}\sigma^0(l_1)$ and ${}^0\sigma^{\delta_2}(l_2)$ are nonempty. These two words are respectively prefix and suffix of u, with moreover $|u| \geq 2$, so we explicitly know the first and last letters of u. Finally, in this case too, the knowledge of $\alpha(\tilde{u})$ and $\omega(\tilde{u})$ is sufficient to determine $\alpha(u)$ and $\omega(u)$.
Now, we show that the knowledge of $\alpha(\tilde{u})$, $\omega(\tilde{u})$ and $\mathrm{ab}(\tilde{u})$ is sufficient to compute $\mathrm{ab}(u)$. Indeed, if M_σ denotes the incidence matrix of σ, we have:

$$\begin{array}{rcl} \mathrm{ab}(u) & = & \mathrm{ab}(\tilde{u})M_\sigma - \mathrm{ab}(p_{\delta_1}(\sigma(\tilde{u}))) - \mathrm{ab}(s_{\delta_2}(\sigma(\tilde{u}))) \\ & = & \mathrm{ab}(\tilde{u})M_\sigma - \mathrm{ab}(p_{\delta_1}(\sigma(\alpha(\tilde{u})))) - \mathrm{ab}(s_{\delta_2}(\sigma(\omega(\tilde{u})))). \end{array}$$

When applied to pairs of words, this says that the knowledge of $M_{\tilde{u},\tilde{v}}$ and $\mathrm{ab}(\tilde{u}) - \mathrm{ab}(\tilde{v})$ (we use here the linearity of the matrix product in the expression above) is sufficient to determine $M_{u,v}$ and $\mathrm{ab}(u) - \mathrm{ab}(v)$, where (u, v) is the image of $(\tilde{u}, \tilde{v})$ by a substitute and cut operation $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4)$. In particular, the images, by an allowed substitute and cut operation, of two pairs of words which equal modulo $\mathcal{R}_{pab}$ are still equal modulo $\mathcal{R}_{pab}$. □

Proposition 42 enable us to define the substitute and cut operations over the quotient set $\mathtt{V} = \mathtt{V}_2/\mathcal{R}_{pab}$. This is detailed in Definition 9.

Example 44. *We consider the set $C := \{c1, c2\}$ of Cassaigne-Selmer substitutions (which are nonerasing). We set $X = \big(\left(\begin{smallmatrix} 2 & 3 \\ 3 & 3 \end{smallmatrix}\right), (0,1,-1)\big)$. There are exactly four substitute and cut operations allowed from X:*

$$ {}^{0}_{0}c1^{0}_{0} \quad , \quad {}^{1}_{0}c1^{0}_{0} \quad , \quad {}^{0}_{0}c2^{0}_{0} \quad \textit{and} \quad {}^{1}_{0}c2^{0}_{0}\,; $$

which respectively give:

$$ \big(\left(\begin{smallmatrix} 1 & 2 \\ 2 & 2 \end{smallmatrix}\right), (1,-1,1)\big)\,,\ \big(\left(\begin{smallmatrix} 3 & 2 \\ 2 & 2 \end{smallmatrix}\right), (0,-1,1)\big)\,,\ \big(\left(\begin{smallmatrix} 1 & 3 \\ 3 & 3 \end{smallmatrix}\right), (1,0,0)\big)\ \textit{and}\ \big(\left(\begin{smallmatrix} 3 & 3 \\ 3 & 3 \end{smallmatrix}\right), (0,0,0)\big)\,. $$

One immediately checks that the pairs of images obtained in Example 33 belong to these equivalence classes.

4.6.3 The automaton of imbalances

As a consequence, we can factorize the graph $\mathtt{G}_2$ by $\mathcal{R}_{pab}$. The quotient graph we obtain is the graph $\mathtt{G}$ described in Section 3.2, that we call *automaton of imbalances*.

Its set of vertices is $\mathtt{V}$; its edges map vertices to their respective images by allowed substitute and cut operations. Among its vertices, we are interested in the subclasses:

- $\mathtt{I} := \mathtt{I}_2/\mathcal{R}_{pab} = \mathtt{V} \cap \big\{\{(\left(\begin{smallmatrix} \epsilon & \epsilon \\ \epsilon & \epsilon \end{smallmatrix}\right), (0)_{a\in A})\} \cup \bigcup_{a\in A} \{(\left(\begin{smallmatrix} a & \epsilon \\ \epsilon & \epsilon \end{smallmatrix}\right), \mathrm{ab}(a)); (\left(\begin{smallmatrix} \epsilon & \epsilon \\ a & \epsilon \end{smallmatrix}\right), -\mathrm{ab}(a)); (\left(\begin{smallmatrix} a & \epsilon \\ a & \epsilon \end{smallmatrix}\right), (0)_{a\in A})\}\big\}$,
- $\mathtt{F} := \mathtt{F}_2/\mathcal{R}_{pab} = \bigcup_{w \text{ S-adic}} \{(M_{u,v}, \mathrm{ab}(u) - \mathrm{ab}(v)) \mid u, v \in \mathcal{F}(w) \text{ and } |u| = |v|\}$;

respectively called *initial* and *final* states of $\mathtt{G}$.

The quotient graph $\mathtt{G}$ inherits from the properties of accessibility of the graph $\mathtt{G}_2$.

Proposition 45. *Let S be a finite set of nonerasing substitutions defined over a common alphabet A.*

1. *If $\mathtt{G}$ is nonempty, then it has infinitely many vertices.*
2. *The number of outgoing edges from any vertex is bounded above by $\mathrm{card}(S) \times l^4$, where $l = \max_{\sigma\in S, a\in A} |\sigma(a)|$.*
3. *If each letter in A appears in a S-adic word (not necessarily the same), then any vertex in $\mathtt{V}$ is accessible from a vertex in $\mathtt{I}$.*
4. *If each letter in A appears in a S-adic word and if X is the image of an element in $\mathtt{I}$ by a finite composition of allowed substitute and cut operations, then $X \in \mathtt{V}$.*

Proof. Assertions (1) and (2) are proved in Proposition 13. One easily checks that Assertions (3) and (4) are inherited from similar properties in the graph of pairs of factors $\mathtt{G}_2$, namely Assertions (3) and (4) in Proposition 34. □

Therefore, similarly to $\mathtt{G}_2$, the graph $\mathtt{G}$ can be breadth first traversed. Algorithm 46 (below) is an easy adaptation of Algorithm 35; it yet requires the substitutions in S to be nonerasing (remind of Example 43).

Algorithm 46.

INPUTS:

- *a finite set of nonerasing substitutions S over a common alphabet A*
- *a nonnegative integer d*

ALGORITHM

```
$V \leftarrow$ I (initialization)
$V' \leftarrow$ empty set
for $X$ in $V$:
    for $(\sigma, \delta_1, \delta_2, \delta_3, \delta_4)$ substitute and cut operation allowed on $X$:
        $Y \leftarrow$ image of $X$ by the substitute and cut operation
        if $Y$ has not yet appeared in the processing:
            $(M, (x_a)_{a \in A}) \leftarrow Y$
            if $\sum_{a \in A} x_a = 0$ and $\max_{a \in A} |x_a| \geq d$:
                return true
            otherwise:
                put $Y$ in $V'$
$V \leftarrow V'$
$V' \leftarrow$ empty set
```

Theorem A. *Let S denote a finite set of nonerasing substitutions over a common alphabet A, and assume that all letters in A appear in some S-adic word (not necessarily the same). If D_S denotes the quantity (possibly infinite):*

$$D_S = \sup_{w \in X_S} \operatorname{imb}(w),$$

then a breadth first search in the automaton of imbalances (viz. Algorithm 46), from its initial states, outputs, for any $d \leq D_S$, a finite sequence of substitutions $(\sigma_i)_{i \in \{1,\dots,n\}} \in S^*$ such that the imbalance of any word in X_S, whose directive sequence starts with $(\sigma_i)_{i \in \{1,\dots,n\}}$, is larger than d.

Proof. This comes from Corollary 36 and Remark 37 that hold for Algorithm 35. □

At last, observe that, like Algorithm 35, Algorithm 46 indefinitely keeps running for any instance (S, d) with $d > \sup_{w \in X_S} \operatorname{imb}(w)$.

5 Discussion on the implementation of the semi-algorithm

We saw that, as soon as it is nonempty, the automaton of imbalances has infinitely many vertices (Proposition 13). We implemented its breadth first traversal (Algorithm 46) for the S-adic systems associated with Arnoux-Rauzy (see Example-Definition 5) and Cassaigne-Selmer (see Example-Definition 4) substitutions. In both cases, the exponential growth of the spanning forest is an obstacle to the manifestation of non-trivial properties.

To slow it down, one needs to consider all the opportunities to spare calculations. In particular, we have to take advantage of all the symmetries enjoyed by the language of the system, and find tight sufficient conditions to kill as many unfruitful branches as possible in the trees. This approach is illustrated in Section 3.4 on the (easy) example of the Thue-Morse subshift. Despite such efforts (the symmetries and trim conditions are specific to the studied S-adic system), for interesting systems, the growth remains strong, as illustrated in Figures 2 and 3.

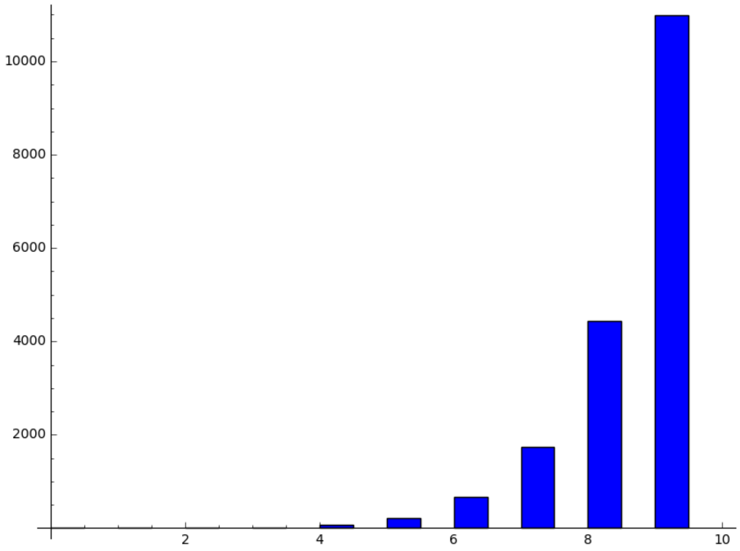


Figure 2: Number of vertices of the spanning forest of the Cassaigne-Selmer system, in function of exploration's depth. We already took advantage of symmetries.

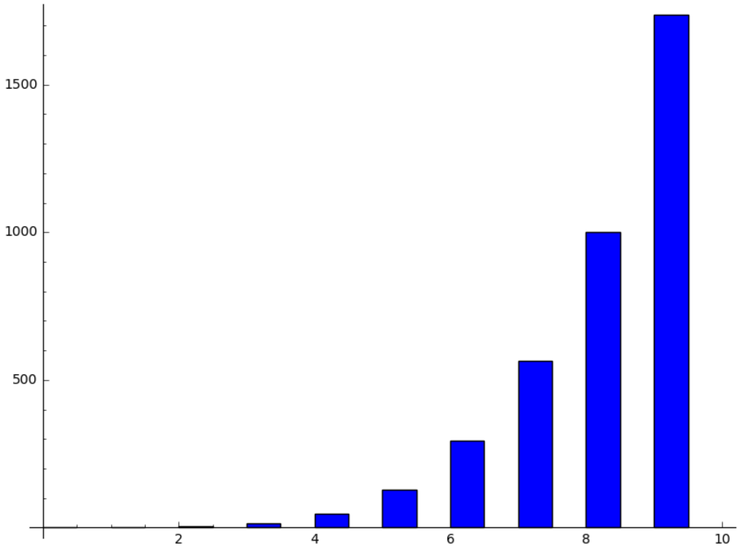
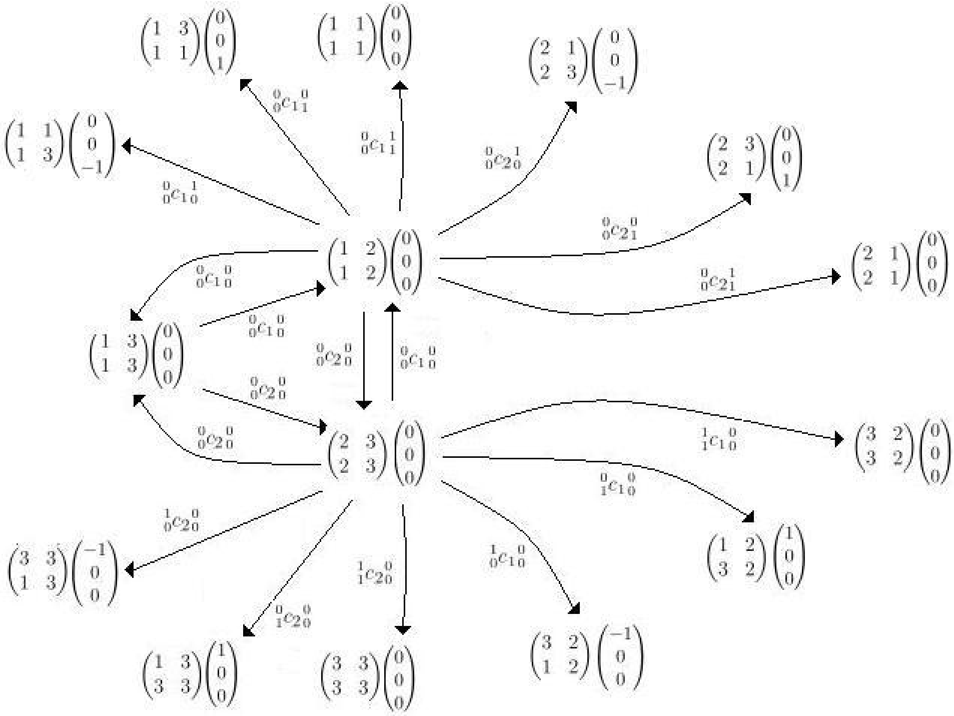


Figure 3: Number of vertices of the spanning forest of the Cassaigne-Selmer system, in function of exploration's depth. Here we take advantage of symmetries, plus an *ad hoc* trim condition.

Here are the results obtained with our final implementation of the automaton of imbalances for the Cassaigne-Selmer system:

- at depth 9, among 1 000 vertices, we found the first imbalance 3;
- at depth 16, among 80 000 vertices, we found the first imbalance 4; the computation (in `Python`) took 20 minutes.

Surprisingly, the study of the labels of the paths leading to these four first occurrences of imbalance turned to be sufficient to guess a general pattern. This pattern gave birth to the families of C-adic words with growing imbalances described in Chapter 1.



$\begin{pmatrix}1&3\\1&1\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}$
$\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
$\begin{pmatrix}2&1\\2&3\end{pmatrix}\begin{pmatrix}0\\0\\-1\end{pmatrix}$
${}^0_0c_1{}^0_1$
${}^0_0c_1{}^1_1$
${}^0_0c_2{}^1_0$
$\begin{pmatrix}1&1\\1&3\end{pmatrix}\begin{pmatrix}0\\0\\-1\end{pmatrix}$
$\begin{pmatrix}2&3\\2&1\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}$
${}^0_0c_1{}^1_0$
${}^0_0c_2{}^0_1$
$\begin{pmatrix}1&2\\1&2\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
$\begin{pmatrix}2&1\\2&1\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
${}^0_0c_1{}^0_0$
${}^0_0c_2{}^1_1$
${}^0_0c_1{}^0_0$
$\begin{pmatrix}1&3\\1&3\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
${}^0_0c_2{}^0_0$
${}^0_0c_1{}^0_0$
${}^0_0c_2{}^0_0$
${}^0_0c_2{}^0_0$
$\begin{pmatrix}2&3\\2&3\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
${}^1_1c_1{}^0_0$
$\begin{pmatrix}3&2\\3&2\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
${}^0_1c_1{}^0_0$
${}^1_0c_2{}^0_0$
$\begin{pmatrix}1&2\\3&2\end{pmatrix}\begin{pmatrix}1\\0\\0\end{pmatrix}$
$\begin{pmatrix}3&3\\1&3\end{pmatrix}\begin{pmatrix}-1\\0\\0\end{pmatrix}$
${}^1_1c_2{}^0_0$
${}^1_0c_1{}^0_0$
${}^0_1c_2{}^0_0$
$\begin{pmatrix}1&3\\3&3\end{pmatrix}\begin{pmatrix}1\\0\\0\end{pmatrix}$
$\begin{pmatrix}3&3\\3&3\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}$
$\begin{pmatrix}3&2\\1&2\end{pmatrix}\begin{pmatrix}-1\\0\\0\end{pmatrix}$

Figure 4: A portion of the automaton of imbalances for the Cassaigne-Selmer system.

References

[AR91] Pierre Arnoux and Gérard Rauzy. Représentation géométrique de suites de complexité 2n+1. *Bulletin de la Société Mathématique de France*, 119:199–215, 1991.

[AS13] Pierre Arnoux and Štěpán Starosta. The Rauzy Gasket. In *Further Developments in Fractals and Related Fields*, pages 1–23. Springer, 2013.

[BD14] Valérie Berthé and Vincent Delecroix. Beyond substitutive dynamical systems: S-adic expansions. In *Lecture note 'Kokyuroku Bessatu'*, pages 81–123, 2014.

[Ber11] Valérie Berthé. Multidimensional Euclidean algorithms, numeration and substitutions. *Integers [electronic only]*, 2011.

[CFZ00] Julien Cassaigne, Sébastien Ferenczi, and Luca Q. Zamboni. Imbalances in Arnoux-Rauzy sequences. *Annales de l'Institut Fourier*, 50:1265–1276, 2000.

[CLL17] Julien Cassaigne, Sébastien Labbé, and Julien Leroy. A set of sequences of complexity 2n+1. In *WORDS 2017 Proceedings*, pages 144–156. Springer, 2017.

[CS01] Vincent Canterini and Anne Siegel. Automate des préfixes-suffixes associé à une substitution primitive. *Journal de théorie des nombres de Bordeaux*, 13(2):353–369, 2001.

[DHS13] Vincent Delecroix, Tomáš Hejda, and Wolfgang Steiner. Balancedness of Arnoux-Rauzy and Brun words. In Juhani Karhumäki, Arto Lepistö, and Luca Zamboni, editors, *Combinatorics on Words*, pages 119–131. Springer Berlin Heidelberg, 2013.

[Fog02] N. Pytheas Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer-Verlag, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.

[GLR09] Amy Glen, Florence Levé, and Gwénaël Richomme. Directive words of episturmian words: equivalences and normalization. *RAIRO - Theoretical Informatics and Applications*, 43(2):299–319, 2009.

III. Natural coding of minimal rotations of the torus, induction and exduction

The content of this chapter will be submitted for publication.

Contents

Natural coding of minimal rotations of the torus, induction and exduction

Mélodie Andrieu

Abstract

We discuss a topological definition of natural coding of a minimal rotation on the d-dimensional torus, inspired by the seminal works of Rauzy on the Tribonacci word. In particular, we show that under the axiom of choice, it is possible to wisely complete the pseudo-fundamental domain of the torus into a fundamental domain, while preserving the property of piecewise translation and a weak form of sequential continuity. We prove then that if w is a natural coding of a minimal rotation of the d-torus, admitting furthermore $d+1$ return words to a letter a, then its derivated word to the letter a is still a natural coding of a minimal rotation of the d-torus, that we fully describe. In particular, this result completes an argument of Cassaigne, Ferenczi and Zamboni: under this assumption, and if furthermore the fundamental domain associated with the natural coding is bounded, then the cylinder $[a]$ is a bounded remainder set for w (i.e. the empiric frequency with which the symbolic trajectory w visits the set $[a]$ tends to its expected value at speed at least $1/n$), which is equivalent to finite imbalance on the letter a. As a consequence, no Arnoux-Rauzy word with infinite imbalance is a natural coding of a minimal rotation of the 2-dimensional torus, with bounded fundamental domain. The same holds for primitive C-adic words and, more generally, uniformly recurrent tree words.

Besides, we prove that to any natural coding of a minimal rotation of the d-torus we can associate other natural codings constructed by a reverse induction process, that we call *exduction*. We study the return words of Arnoux-Rauzy and primitive C-adic words within the S-adic framework and obtain that, for these two classes of words, being a natural coding of a minimal rotation of the 2-torus is a property that only depends on the asymptotic behavior of the directive sequence.

1 Introduction

Backgrounds

In [Rau82], Rauzy undertakes the study of symbolic systems associated with minimal rotations of the 2-dimensional torus by a remarkable example: the Tribonacci word. The "canonical association" he obtains, through a well-chosen partition, was refered to (though not written at that time) under the name of "natural coding". Later, when it appeared that the Tribonacci word was a remarkable element of a wide class of words generalizing Sturmian words on a 3-letter alphabet [AR91], it was believed that the canonical association property would extend to the whole class of words (now known as Arnoux-Rauzy words.)

In [CFZ00], Cassaigne, Ferenczi and Zamboni disproved this belief by exhibiting an Arnoux-Rauzy word satisfying a remarkable combinatorial property: infinite imbalance (see Definition 1.)

The first and main part of their paper is devoted to the construction of this unsuspected object; in the second part, relying on a theorem of Rauzy on bounded remainder sets (see [Rau84] or Theorem B below), they state that if w^* is an Arnoux-Rauzy word with infinite imbalance, then either w^* or one of its induced words (which are still Arnoux-Rauzy words with infinite imbalance) is not a natural coding of a rotation of the 2-torus (the definition of natural coding is discussed in Section 3.) Even if the proof is incorrect, their result is true, under the additional assumptions of boundedness of the fundamental domain and minimality of the rotation (it is a consequence of Corollary 43.) In Section 4 of the present document, we rectify and complete the proof that Cassaigne, Ferenczi and Zamboni sketched to achieve a more significant result: Theorem A. Besides, the existence of non-coding Arnoux-Rauzy words was established by other techniques in [CFM08].

Since then, substantial advances have been made in the counterpart. Under a measure theory definition, [AI01], [BJS12] and [BŠW13] show that purely substitutive Arnoux-Rauzy words are natural codings of rotation of the 2-torus; [BST19] extends this positive result in the generic S-adic case to a large subclass of Arnoux-Rauzy words.

Our work

First, we propose a topological definition of natural coding of a minimal rotation on the d-dimensional torus, inspired by the seminal works of Rauzy [Rau82] (Definition 2.) Under this framework, we cannot elude the question of borders.

We show that if w is a natural coding of a minimal rotation of the d-torus, then: 1) w is written with d+1 letters and is uniformly recurrent (see immediate Lemma 3); 2) under the axiom of choice, it is possible to wisely complete the pseudo-fundamental domain of the torus into a fundamental domain, while preserving the property of piecewise translation as well as a the continuity, in an of course weak sense, of the coding function (see Proposition 9 and Lemma 16.) If furthermore w admits $d+1$ return words to a letter a, then its derivated word to the letter a (see Definition 27) is still a natural coding of a minimal rotation of the d-torus, that we fully describe (see Theorem A.)

Fulfilling an argument of [CFZ00], we prove then that the cylinder $[a]$ is a bounded remainder set for w, which is equivalent to finite imbalance on the letter a (see Proposition 36.) As a consequence, no Arnoux-Rauzy word with infinite imbalance is a natural coding of a minimal rotation of the 2-torus. This consequence holds for primitive C-adic words as well, and more generally for tree words (see Corollary 43.)

On another hand, we show that the property of being a natural coding of a minimal rotation also passes through a reverse induction operation that we call *exduction* (Theorem G.)

In the case of Arnoux-Rauzy and primitive C-adic words, we link the induction and exduction processes to the action of a multidimensional continued fraction algorithm on the letter-frequencies vector of w, through the S-adic expression of return words (see Theorem D for Arnoux-Rauzy words, which is a restatement of [JV00], and Theorem E for primitive C-adic words.)

Finally, we show that for Arnoux-Rauzy and primitive C-adic subshifts, being a natural coding of a minimal rotation on the 2-torus only depends on the asymptotic behavior of the directive sequence (Theorem H.)

2 Preliminaries

Finite and infinite words

An *alphabet* $\mathcal{A}$ is a finite set; its elements are called *letters.* For instance, in what follows, we work with the $(d+1)$-letter alphabet $\mathcal{I} = \{1, ..., d+1\}$. A *finite word of length* n, where n is a nonnegative integer, is the concatenation of n letters: $u = a_0 \cdot a_1 \cdot ... \cdot a_{n-1} \in \mathcal{A}^n$. As soon as there is no ambiguity, the concatenation symbol $\cdot$ will be omitted. We denote by $\mathcal{A}^* = \cup_{n\in\mathbb{N}} \mathcal{A}^n$ the set of finite words; an infinite word is an element $w = a_0 a_1 ... \in \mathcal{A}^{\mathbb{N}}$. Following `Python`, we denote by $w[k]$, for $k \in \mathbb{N}$, the $(k+1)$-th letter of a nonempty (finite or infinite) word w.

A finite word u is a *factor of length* n of a (finite or infinite) word w if there exists a nonnegative integer i such that for all $k \in \{0, ..., n-1\}$, $w[i+k] = u[k]$; in the particular case $i = 0$, we say that u is the *prefix of length* n of w, and denote it by $u = \mathrm{pref}_n(w)$. We denote by $\mathcal{F}_n(w)$ the set of factors of w of length n and by $\mathcal{F}(w)$ the set of factors of all lengths. An infinite word is said *recurrent* if every factor occurs infinitely often.

We endow the set $\mathcal{A}^{\mathbb{N}}$ with the product topology, for which it is compact. Given a finite word $u \in \mathcal{A}^*$, we denote by $[u]$ the set of words in $\mathcal{A}^{\mathbb{N}}$ which admit u as prefix. The sets $[u]$ are called *cylinders*; they are clopen and form a neighborhood basis for the topology.

Symbolic dynamics

We denote by S the *shift map*, which acts on infinite words by 'erasing' the first letter: if w is an infinite word, for all $k \in \mathbb{N}, S(w)[k] = w[k+1]$. If w_0 is an infinite word, we call *subshift associated with* w_0, denoted by X_0, the closure set (for the product topology) of the trajectory of w_0 under the shift action: $X_0 = \overline{\{S^k(w_0) | k \in \mathbb{N}\}}$.

The shift map is at the core of symbolic dynamics. Given a dynamical system, one can choose to partition the space into a finite number of regions $A_1, ..., A_d$, and study the possible sequences of regions crossed over time (for a general introduction to symbolic dynamics, the reader should refer to [LM95].) The difficulty and the interest of the discrete dynamical system thus obtained highly depends on the choice made for the partition. In this paper, we study the behavior, under the induction and the exduction operations, of a family of remarkable partitions for the flow of a minimal rotation on a d-dimensional torus (called hereafter d-torus.)

Imbalance

Definition 1. *The* imbalance on the letter *a of an infinite word w is the quantity (possibly infinite):*

$$\mathrm{imb}_a(w) \ = \ \sup_{n\in\mathbb{N}} \quad \sup_{u,v\in\mathcal{F}_n(w)} \quad ||u|_a - |v|_a|,$$

where $|u|_a$ denotes the number of times the letter a appears in the word u.

The imbalance *of w is the quantity:*

$$\mathrm{imb}(w) \ = \ \sup_{a\in A} \mathrm{imb}_a(w).$$

The imbalance measures inequities in the distribution of letters in a given word. This notion appeared for the first time in the works of Morse and Hedlund ([MH38] and [MH40]); in [CH73] Coven and Hedlund showed that this quantity characterizes Sturmian words: a binary word is Sturmian if and only if it is aperiodic and its imbalance equals 1.

This quantity has been much studied since, through the notions of C-balancedness (a word is *C-balanced* if and only if its imbalance is lower than C), balancedness (originally, a word is *balanced* if and only if its imbalance is lower than 1; in recent papers, balanced words tend to denote words with finite imbalance.) In [Ada03], Adamczewski introduced the *balance function* of an infinite word w: $B_w(n) = \max\limits_{a\in\mathcal{A}} \max\limits_{u,v\in\mathcal{F}_n(w)} ||u|_a - |v|_a|$. The imbalance of w is the smallest upper bound for this balance function.

For instance, the imbalance of the Tribonacci word is 2 (see [RSZ09] for a proof, but this fact was mentioned before.) Because they were constructed as a generalization of Sturmian words from the combinatorial viewpoint, it was expected that Arnoux-Rauzy words would have bounded imbalance. This is not the case: Cassaigne, Ferenczi and Zamboni exhibited in [CFZ00] families of Arnoux-Rauzy words with arbitrary high imbalance, and even families of words with infinite imbalance (see also [And20] for an alternative construction.)

3 Natural coding of minimal rotations

3.1 A topological definition and its consequences

Let d a positive integer. Recall that L is a lattice of $\mathbb{R}^d$ if there exist $e_1, ..., e_d \in \mathbb{R}^d$, linearly independent, such that $L = \mathbb{Z}e_1 + ... + \mathbb{Z}e_d$. We denote by $\mathbb{T}_L := \mathbb{R}^d/L$ the d-torus associated with the lattice L, and by p_L the projection map on the torus. The torus is endowed with the quotient topology (consisting of all sets with an open preimage under p_L), which makes p_L open and continuous.

A set $\Omega \subset \mathbb{R}^d$ is *L-simple* if the map $p_L : \Omega \to \mathbb{T}_L$ is one-to-one; Ω is a *fundamental domain* of L if the map $p_L : \Omega \to \mathbb{T}_L$ is one-to-one and onto. As soon as Ω is L-simple, we introduce the cover map $r_{\Omega,L}$ which maps each point in $p_L(\Omega)$ to its unique preimage in Ω. If the set Ω is open, the cover map $r_{\Omega,L}$ is open and continuous for the topology on $\mathbb{R}^d$. Remark: in [Rau84], L-simple sets are furthermore assumed to be bounded - in our work we explicitly mention this assumption each time it is required.

Now, given $\alpha \in \mathbb{R}^d$, the rotation of the torus $\mathbb{T}_L$ through the angle α is the map $R_{\alpha,L} : \mathbb{T}_L \to \mathbb{T}_L,\ x \mapsto x + p_L(\alpha)$ (possibly denoted by R_α if there is no ambiguity on the lattice.) Following [Rau84], a pair (α, L) is said *minimal* if for all $\tilde{x} \in \mathbb{T}_L$, the sequence $(R_\alpha^n(\tilde{x}))_{n\in\mathbb{N}}$ is dense in $\mathbb{T}_L$ - or equivalently, if there exists one such $\tilde{x}$ in $\mathbb{T}_L$.

Definition 2. *A word $w_0 \in \mathcal{A}^{\mathbb{N}}$ is a* natural coding of a minimal rotation of the d-torus *if there exists a lattice $L \subset \mathbb{R}^d$ together with a vector $\alpha \in \mathbb{R}^d$ such that:*

- *(minimality) The pair (α, L) is minimal.*
- *(partition of a pseudo-fundamental domain) There exist Ω_1,...,Ω_{d+1} nonempty, open sets of $\mathbb{R}^d$ such that:*
  - *the sets Ω_1,...,Ω_{d+1} are pairwise disjoint;*
  - *the union set $\Omega = \cup_{i\in\{1,...,d+1\}}\Omega_i$ is L-simple;*
  - *the projection set $p_L(\Omega)$ is dense in the torus $\mathbb{T}_L$.*
- *(exchange of pieces) There exist $\alpha_1, ..., \alpha_{d+1} \in \mathbb{R}^d$ such that for all index $i \in \{1, ..., d+1\}$ and for all point $\tilde{x} \in p_L(\Omega_i) \cap R_\alpha^{-1}(p_L(\Omega))$, $r_{\Omega,L}(R_\alpha(\tilde{x})) = r_{\Omega,L}(\tilde{x}) + \alpha_i$.*
- *(a coding trajectory) There exists $\tilde{x}_0$ in $p_L(\Omega)$ such that, for all $n \in \mathbb{N}$, $R_\alpha^n(\tilde{x}_0) \in p_L(\Omega_{w_0[n]})$, where $w_0[n]$ denotes the (n+1)-th letter of w_0.*

  We set $x_0 = r_{\Omega,L}(\tilde{x}_0)$ and we call $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$ the elements *of the natural coding w.*

Lemma 3. *If w_0 is a natural coding of a minimal rotation of the d-torus, then w is written with exactly $d + 1$ letters and is uniformly recurrent.*

Proof. By minimality of the rotation, the trajectory of the point $\tilde{x}_0$ is dense in the torus and, thus, visits each open set $p_L(\Omega_i)$ - so w_0 contains each letter i in $\{1,...,d+1\}$.

Let u a factor of w_0. Then there exists a nonnegative integer n such that $S^n(w_0) \in [u]$. Observe that the set $\tilde{\Omega}_u := \cap_{l=0}^{|u|-1} R_\alpha^{-l}(p_L(\Omega_{u[l]}))$ is nonempty (it contains the point $R_\alpha^n(\tilde{x}_0)$) and open (the Ω_i are open, the projection p_L is open and the rotation is continuous.) Therefore, by minimality of the pair (α, L), we obtain a cover of the torus by a countable family of open sets: $\mathbb{T}_L = \cup_{n\in\mathbb{N}} R_\alpha^{-n}(\tilde{\Omega}_u)$, from which we extract, by compacity of $\mathbb{T}_L$, a finite cover. We conclude that there exists a nonnegative integer m such that $\mathbb{T}_L = \cup_{n=0}^{m} R_\alpha^{-n}(\tilde{\Omega}_u)$ and, finally, that w_0 is uniformly recurrent. □

At last, we say that a *subshift* is a *natural coding of a minimal rotation of the d-torus* if it is minimal and if one of its elements is a natural coding of a minimal rotation of the d-torus.

Hereafter, we denote by $\mathcal{I} = \{1,...,d+1\}$ the alphabet.

Notation 4. *In this document we shall work with a second lattice, called M. To avoid confusion, we will use the symbol $\tilde{.}$ (tilda) to refer to points or sets in the torus $\mathbb{T}_L$ whereas the symbol $\bar{.}$ (bar) will be devoted to elements in $\mathbb{T}_M$- the absence of symbol referring by default to the covering space $\mathbb{R}^d$. From now on, we denote $\tilde{\Omega} = p_L(\Omega)$, $\tilde{\Omega}_i = p_L(\Omega_i)$ and $x_0 = r_{\Omega,L}(\tilde{x}_0)$.*

Example 5. *A Sturmian subshift with slope α is a natural coding of the minimal rotation of the torus $\mathbb{R}/\mathbb{Z}$ through the angle α. One can take the pseudo-fundamental domain $\Omega =]0,1[$ together with the partition $\Omega_1 =]0, 1-\alpha[$, $\Omega_2 =]1-\alpha, 1[$.*

The Tribonacci subshift is a natural coding of the minimal rotation of the torus $\mathbb{R}^2/\mathbb{Z}^2$ through the angle (ζ, ζ^2), where ζ is is the unique real root of the polynomial $x^3 + x^2 + x - 1$. Furthermore, the Rauzy fractal gives a pseudo-fundamental domain for which the pieces of the partition are furthermore bounded and simply connected (see [Rau82].)

Given a natural coding w_0 with elements $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$, we introduce the numbering function $\nu : \tilde{\Omega} \to \mathcal{I}$, which maps all elements of $\tilde{\Omega}_i$ to the letter i; and we consider the coding function f given by $f(x) = (\nu(R_\alpha^n(p_L(x))))_{n\in\mathbb{N}}$ which makes sense each time the trajectory of $p_L(x)$ for the rotation action is included in $\tilde{\Omega}$. Let $\mathcal{D}$ denote the maximal subset of Ω on which the coding function f is defined. The covering rotation $T = r_{\Omega,L} \circ R_\alpha \circ p_L$ is well-defined on $\mathcal{D}$ and satisfies $T(\mathcal{D}) \subset \mathcal{D}$. Following Notation 4, we denote $\tilde{\mathcal{D}} = p_L(\mathcal{D})$.

Lemma 6. *The trajectory of x_0 under the action of T is included in $\mathcal{D}$ and dense in Ω.*

Proof. By definition of natural coding, the trajectory of $\tilde{x}_0$ under R_α is included in $\tilde{\mathcal{D}}$ (we even know that $f(x_0) = w_0$.) The rotation being minimal, the trajectory of $\tilde{x}_0$ is dense in the torus $\mathbb{T}_L$ and in particular in $\tilde{\Omega}$. By continuity of the cover map $r_{\Omega,L}$ (Ω is open), we conclude that this property is preserved in the cover space. □

Proposition 7. *The coding function f is continuous for the induced topology on $\mathcal{D}$. Furthermore, the diagram below is commutative:*

$$\begin{tikzcd}
\mathcal{D} \arrow[r, "T"] \arrow[d, "f"'] & \mathcal{D} \arrow[d, "f"] \\
\mathcal{I}^{\mathbb{N}} \arrow[r, "S"'] & \mathcal{I}^{\mathbb{N}},
\end{tikzcd}$$

and the image set $f(\mathcal{D})$ of f is included in X_0, the subshift generated by the word w_0.

Proof. First, observe that $f(x)$ belongs to the cylinder $[i_0...i_{n-1}]$ if and only if $x \in \mathcal{D}$ and for all k in $\{0, ..., n-1\}$, $R_\alpha^k(p_L(x)) \in \tilde{\Omega}_{i_k}$; if and only if $x \in \mathcal{D} \cap \cap_{k=0}^{n-1} p_L^{-1}(R_\alpha^{-k}(\tilde{\Omega}_{i_k}))$, which is open for the induced topology on $\mathcal{D}$ - hence the continuity of f.

Secondly, the diagram is commutative by definition of f.

Thirdly, let $w \in f(\mathcal{D})$ and $x \in \mathcal{D} \subset \Omega$ one of its preimages. By Lemma 6, there exists an extracted sequence $(T^{n_k}(x_0))_k$ in $\mathcal{D}^{\mathbb{N}}$ that tends to x. But f being continuous, the image sequence, which is $(S^{n_k}(w_0))_k$, tends to $w = f(x)$ - meaning that the word $w \in \mathcal{I}^{\mathbb{N}}$ actually belongs to X_0; we conclude that $f(\mathcal{D}) \subset X_0$. □

We will prove in Proposition 18 that the coding function f is one-to-one.

Lemma 8. *For all $x \in \mathcal{D}$, we have $T(x) = x + \alpha_i$, with $i = \nu(p_L(x))$.*

Proof. This is an immediate consequence of Item 2 (exchange of pieces) in Definition 2 (natural coding.) □

3.2 Borders assignment

In this subsection, we show that, under the axiom of choice, it is possible to wisely assign borders to the pieces Ω_i, in order to complete the L-simple set Ω into a fundamental domain Ω' and enlarge the remarkable property of exchange of pieces, while keeping (under a weak form) the continuity of the coding function f.

Proposition 9. *Let L a lattice of $\mathbb{R}^d$, and $\Omega_1, ..., \Omega_{d+1}$ nonempty, open, pairwise disjoint sets, such that moreover $\Omega = \cup_{i \in \mathcal{I}} \Omega_i$, where $\mathcal{I} = \{1, ..., d+1\}$, is L-simple (H1). Let $\alpha \in \mathbb{R}^d$ such that (α, L) is minimal (H2). Assume there exists $x_0 \in \Omega$ such that for all nonnegative integer n, $R_\alpha^n(p_L(x_0)) \in \Omega$(H3), and denote by $(i_n)_n \in \mathcal{I}^{\mathbb{N}}$ the unique sequence satisfying: for all $n \in \mathbb{N}$, $R_\alpha^n(p_L(x_0)) \in \Omega_{i_n}$. At last, assume that there exist $\alpha_1, ..., \alpha_{d+1} \in \mathbb{R}^d$ such that for all nonnegative integer n, $T^{n+1}(x_0) = T^n(x_0) + \alpha_{i_n}$, where $T = r_{\Omega,L} \circ R_\alpha \circ p_L$ (H4).*

Then, under the axiom of choice, there exist $\Omega'_1, ..., \Omega'_{d+1} \subset \mathbb{R}^{d+1}$ such that:

- *(C1) for all $i \in \mathcal{I}$, $\Omega_i \subset \Omega'_i$,*
- *(C2) the union set $\Omega' = \cup_{i \in \mathcal{I}} \Omega'_i$ is a fundamental domain of L,*
- *(C3) the sets Ω'_i are pairwise disjoint.*

Furthermore, if T' denotes the covered rotation $T' = r_{\Omega',L} \circ R_\alpha \circ p_L$, then:

- *(C4) for all $x \in \Omega'_i$, $T'(x) = x + \alpha_i$;*
- *(C5) for all $x \in \Omega'$, for all $q \in \mathbb{N}$, there exists τ an extraction (i.e. an increasing map from $\mathbb{N}$ to $\mathbb{N}$) such that: (i) the sequence $(T^{\tau(m)}(x_0))_{m \in \mathbb{N}}$ converges to x; (ii) for all $n \in \{0, ..., q\}$, for all nonnegative integer m, $T^{\tau(m)+n}(x_0) \in \Omega_{\iota_n}$, where ι_n is given by $T'^n(x) \in \Omega'_{\iota_n}$.*

Proof. **General idea.** The proof consists of a lifting process, based on the axiom of choice. Initially, the sets $\Omega'_1, ..., \Omega'_{d+1}$ are empty. We browse each orbit for the action of the rotation R_α to the future and back to the past, from a well-chosen point, in order to assign to each visited point of the torus $\mathbb{T}_L$ a covering point in $\mathbb{R}^d$, that we furthermore stow in one of the $d+1$ sets $\Omega'_1, ..., \Omega'_{d+1}$. The pair (α, L) being minimal (H2), each point of the torus is visited exactly once by this process, and the sets $\Omega'_1, ..., \Omega'_{d+1} \subset \mathbb{R}^d$ form a partition of a fundamental domain of the torus.

Following Notation 4, we denote $\tilde{x}_0 = p_L(x_0)$, $\tilde{\Omega}_i = p_L(\Omega_i)$ for all $i \in \mathcal{I}$, and $\tilde{\Omega} = p_L(\Omega)$.

Method to lift one orbit. Let $(\tilde{y}_n)_{n\in\mathbb{Z}}$ an orbit for the action of R_α on the torus. The pair (α, L) being minimal (H2), the pair $(-\alpha, L)$ is minimal as well, and there exists an increasing sequence of nonnegative indices $(n_k)_{k\in\mathbb{N}}$ such that for all $k\in\mathbb{N}$, $\tilde{y}_{-n_k}$ belongs to the nonempty open set $\tilde{\Omega}$ (H1). Without loss of generality, we assume $n_0 = 0$.

We now intend to construct, by a diagonal process, a lifted sequence $(y_n)_{n\in\mathbb{Z}}$ for the orbit $(\tilde{y}_n)_{n\in\mathbb{Z}}$, together with a numbering sequence $(j_n)_{n\in\mathbb{Z}}$, such that for all $n\in\mathbb{Z}$:

1. $p_L(y_n) = \tilde{y}_n$;

2. if moreover $\tilde{y}_n \in \tilde{\Omega}$, then $y_n = r_{\Omega,L}(\tilde{y}_n)$;

3. $y_{n+1} = y_n + \alpha_{j_n}$;

4. for all $q\in\mathbb{N}$, there exists an extraction τ such that $T^{\tau(m)}(x_0) \to_{m\to\infty} y_n$ and for all nonnegative integer m, $i_{\tau(m)}...i_{\tau(m)+q} = j_n...j_{n+q}$.

The lifted orbit $(y_n)_{n\in\mathbb{Z}}$ and its symbolic trajectory $(j_n)_{n\in\mathbb{Z}}$ will be obtained as the limit sequences, when k tends to infinity, of the finite sequences $(y_n)_{n\in\{-n_k,...,n_k\}}$ and $(j_n)_{n\in\{-n_k,...,n_k\}}$, as constructed below.

First, since $\tilde{y}_0 \in \tilde{\Omega}$, there exists a unique $i\in\mathcal{I}$ such that $\tilde{y}_0\in\tilde{\Omega}_i$. We set $y_0^0 = r_{\Omega,L}(\tilde{y}_0)$ and $j_0^0 = i$. By minimality of (α, L) (H2), and since $\Omega_{j_0^0}$ is open (H1), there exists an extraction σ_0 such that the sequence $(T^{\sigma_0(m)}(x_0))_{m\in\mathbb{N}}$ is included in $\Omega_{j_0^0}$ and tends to y_0^0.

Now, let $k\in\mathbb{N}$, and assume there exist an extraction σ_k together with a finite word $j^k_{-n_k}...j^k_{n_k} \in \mathcal{I}^{2n_k+1}$, such that $T^{\sigma_k(m)}(x_0) \to_{m\to\infty} r_{\Omega,L}(\tilde{y}_{-n_k})$ and for all nonnegative integer m, $i_{\sigma_k(m)}...i_{\sigma_k(m)+2n_k}$ $= j^k_{-n_k}...j^k_{n_k}$. We want to construct an extraction σ_{k+1} together with a finite word $j^{k+1}_{-n_{k+1}}...j^{k+1}_{n_{k+1}} \in \mathcal{I}^{2n_{k+1}+1}$, such that:

$$\begin{cases} T^{\sigma_{k+1}(m)}(x_0) \to_{m\to\infty} r_{\Omega,L}(\tilde{y}_{-n_{k+1}}), \\ \text{for all } m\in\mathbb{N}, \quad i_{\sigma_{k+1}(m)}...i_{\sigma_{k+1}(m)+2n_{k+1}} = j^{k+1}_{-n_{k+1}}...j^{k+1}_{n_{k+1}}. \end{cases}$$

Denote $l = n_{k+1} - n_k \in \mathbb{N}^*$, and let $m_0\in\mathbb{N}$ such that $\sigma_k(m_0) \geq l$. Denote $\gamma : m \mapsto m + m_0$. The sequence $(T^{\sigma_k\circ\gamma(m)-l}(x_0))_{m\in\mathbb{N}}$ is the image, by the continuous function $T^{-l} = r_{\Omega,L}\circ R_\alpha^{-l}\circ p_L$ (Ω is open by (H1)), of the convergent sequence $(T^{\sigma_k\circ\gamma(m)}(x_0))_{m\in\mathbb{N}}$; it thus admits a limit that we denote $y^{k+1}_{-n_{k+1}}$. Furthermore, the possible values of the sequence $(i_{\sigma_k\circ\gamma(m)-l}...i_{\sigma_k\circ\gamma(m)-l+2n_{k+1}})_{m\in\mathbb{N}}$ belong to the finite set $\mathcal{I}^{2n_{k+1}+1}$, so there exists $j^{k+1}_{-n_{k+1}}...j^{k+1}_{n_{k+1}} \in \mathcal{I}^{2n_{k+1}+1}$ together with an extraction υ such that for all nonnegative integer m, $i_{\sigma_{k+1}(m)}...i_{\sigma_{k+1}(m)+2n_{k+1}} = j^{k+1}_{-n_{k+1}}...j^{k+1}_{n_{k+1}}$, where $\sigma_{k+1}(m) = \sigma_k\circ\gamma\circ\upsilon(m) - l$, for all nonnegative integers m.

Observe that the sequence $(T^{\sigma_{k+1}(m)+l}(x_0))_m$ is a subsequence of $(T^{\sigma_k(m)}(x_0))_m$. Consequently, for all $n\in\{-n_k, ..., n_k\}$:

$$\begin{cases} y_n^{k+1} := \lim_m \; T^{\sigma_{k+1}(m)+l+n+n_k}(x_0) = \lim_m \; T^{\sigma_k(m)+n+n_k}(x_0) =: y_n^k; \\ j_n^{k+1} = j_n^k. \end{cases}$$

So we can construct $(y_n)_{n\in\mathbb{Z}}$ and $(j_n)_{n\in\mathbb{Z}}$ as the biinfinite limits of the finite words $(y_n^k)_{n\in\{-n_k,...,n_k\}}$ and $(j_n^k)_{n\in\{-n_k,...,n_k\}}$, for $k\in\mathbb{N}$.

Properties of the lifted orbit.

Now we show that the lifted orbit $(y_n)_{n\in\mathbb{Z}}$ satisfies the properties (1), (2), (3) and (4).

Lemma 10. *For all $n\in\mathbb{Z}$, $p_L(y_n) = \tilde{y}_n$; if moreover $\tilde{y}_n\in\tilde{\Omega}$, then $y_n = r_{\Omega,L}(\tilde{y}_n)$.*

Proof. Let $n \in \mathbb{Z}$ and $k \in \mathbb{N}$ such that $-n_k \leq n$. By continuity of p_L and R_α , we have:

$$\begin{array}{rcl} p_L(y_n) & = & \lim_m \, p_L \circ T^{\sigma_k(m)+n+n_k}(x_0) \\ & = & \lim_m \, R_\alpha^{n+n_k} \circ p_L \circ T^{\sigma_k(m)}(x_0) \\ & = & R_\alpha^{n+n_k}(\tilde{y}_{-n_k}) \\ & = & \tilde{y}_n. \end{array}$$

Furthermore, if $\tilde{y}_n \in \tilde{\Omega}$, then there exists $\lambda \in L$ such that $y_n = r_{\Omega,L}(\tilde{y}_n) + \lambda$. Since the sequence $(T^{\sigma_k(m)+n+n_k}(x_0))_{m\in\mathbb{N}}$ of elements in $\Omega^{\mathbb{N}}$ tends to y_n, which belongs to the open set $\Omega + \lambda$, we must have $\Omega \cap \Omega + \lambda \neq \emptyset$. By L-simplicity of Ω (H1), this implies $\lambda = 0$ and $y_n = r_{\Omega,L}(\tilde{y}_n)$. □

Lemma 11. *For all $n \in \mathbb{Z}$, $y_{n+1} = y_n + \alpha_{j_n}$.*

Proof. Let $n \in \mathbb{Z}$ and $k \in \mathbb{N}$ such that $-n_k \leq n < n_k$. For all nonnegative integer m, $i_{\sigma_k(m)+n+n_k} = j_n$, hence $T^{\sigma_k(m)+n+n_k+1}(x_0) = T^{\sigma_k(m)+n+n_k}(x_0) + \alpha_{j_n}$ (H4). Taking the limit when $m \to \infty$ on both sides gives $y_{n+1} = y_n + \alpha_{j_n}$. □

Lemma 12. *For all $n \in \mathbb{Z}$ and for all $q \in \mathbb{N}$, there exists an extraction τ such that $T^{\tau(m)}(x_0) \to_{m\to\infty} y_n$ and for all nonnegative integer m, $i_{\tau(m)}...i_{\tau(m)+q} = j_n...j_{n+q}$.*

Proof. Let $n \in \mathbb{Z}$ and $q \in \mathbb{N}$. We choose $k \in \mathbb{N}$ such that $-n_k \leq n$ and $q \leq n_k - n$. By construction of σ_k, the function $\tau : m \mapsto \sigma_k(m) + n + n_k$ suits. □

Construction of a fundamental domain with a "good" partition.

Thanks to the axiom of choice, we run this process for each orbit in the action of the rotation R_α on the torus $\mathbb{T}_L$. We denote $\Omega' \subset \mathbb{R}^d$ the set of all the lifted points we obtain. Since the orbits form a partition of the torus, and by minimality of (α, L) (H2), each point of the torus is lifted exactly once, meaning that Ω' is a fundamental domain of the torus (C2). Hereafter we denote by $r_{\Omega',L}(\tilde{y})$ the covering point of $\tilde{y}$ given by the process.

Again, because each point in the torus is visited exactly once, we can define the numbering map $\nu' : \mathbb{T}_L \to \mathcal{I}$, which maps the n-th point of a given orbit $(\tilde{y}_n)_{n\in\mathbb{Z}}$ (with respect to the indexation used for the lifting process) to the n-th term of the associated numbering sequence $(j_n)_{n\in\mathbb{Z}}$. For all $i \in \mathcal{I}$, we set $\Omega'_i = \{r_{\Omega',L}(\tilde{y}) \mid \tilde{y} \in \mathbb{T}_L \text{ s.t. } \nu'(\tilde{y}) = i\}$. The sets Ω'_1,...,Ω'_{d+1} form a partition of Ω' (C3).

At last, Lemma 10 implies that for all $\tilde{y} \in p_L(\Omega)$, $r_{\Omega',L}(\tilde{y}) = r_{\Omega,L}(\tilde{y})$, hence the inclusion $\Omega_i \subset \Omega'_i$ for all i in $\mathcal{I}$. Lemmas 11 and 12 respectively ensure (C4) and (C5).

□

Definition 13. *Let w_0 a natural coding of a minimal rotation of the d-torus with elements $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$. We say that $(\Omega' : \Omega'_1, ..., \Omega'_{d+1})$ is a* borders assignment *of w if:*

1. *for all i in $\mathcal{I} = \{1, ..., d + 1\}$, Ω_i is included in Ω'_i;*

2. *the sets $\Omega'_1, ..., \Omega'_{d+1}$ forms a partition of Ω';*

3. *the set Ω' is a fundamental domain of L;*

4. *for all i in $\mathcal{I}$, for all x in Ω'_i, $T'(x) = x+\alpha_i$, where T' denotes the covered map of the rotation to the fundamental domain Ω': $T' = r_{\Omega',L} \circ R_\alpha \circ p_L$.*

5. *for all $x \in \Omega'$, for all $q \in \mathbb{N}$, there exists an extraction τ such that: (i) $T^{\tau(m)}(x_0) \to_{m\to\infty} x$; (ii) for all $n \in \{0, ..., q\}$, for all nonnegative integer m, $T^{\tau(m)+n}(x_0) \in \Omega_{\iota_n}$, where ι_n is given by $T'^n(x) \in \Omega'_{\iota_n}$.*

Corollary 14. *If $(\Omega' : \Omega'_1, ..., \Omega'_{d+1})$ is a border assignment of a natural coding with elements $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$, then for all $i \in \mathcal{I} := \{1, ..., d+1\}$, the set Ω_i is dense in Ω'_i.*

Proof. Let $i \in \mathcal{I}$ and $x \in \Omega'_i$. Denote by τ the extraction given by Definition 13 for x and $q = 0$. Then the sequence $(T^{\tau(n)}(x_0))_{n \in \mathbb{N}}$ belongs to $\Omega_i^{\mathbb{N}}$ and converges to x - hence the density of Ω_i in Ω'_i. □

Corollary 15. *Under the axiom of choice, a natural coding of a minimal rotation of the d-torus admits borders assignments.*

Given a natural coding of a minimal rotation w_0 endowed with a borders assignment $(\Omega' : \Omega'_1, ..., \Omega'_{d+1})$, we extend the numbering and coding functions ν and f into $\nu' : \tilde{\Omega}' \mapsto \mathcal{I}$ and $f' : \Omega' \mapsto \mathcal{I}^{\mathbb{N}}$: for all $\tilde{x} \in \tilde{\Omega}'_i$, we set $\nu'(\tilde{x}) = i$; for all x in Ω', $f'(x) = (\nu'(R^n_\alpha(p_L(x))))_{n \in \mathbb{N}}$. The extended coding function is defined on the whole fundamental domain Ω' and coincides with f wherever f is defined, i.e. on the subset $\mathcal{D}$.

The following lemma, which is an immediate consequence of Definition 13, is the keystone of the paper.

Lemma 16 (Weak sequential continuity)**.** *For all $x \in \Omega'$, there exists a sequence $(y_n)_n \in \mathcal{D}^{\mathbb{N}}$ such that $y_n \to_{n \to \infty} x$ and $f'(y_n) = f(y_n) \to_{n \to \infty} f'(x)$.*

Proof. Let $x \in \Omega'$. For $n \in \mathbb{N}$, we set $y_n = T^{\tau_n(n)}(x_0)$, where τ_n is the extraction given for $q = n$ in Definition 13. The sequence $(y_n)_{n \in \mathbb{N}}$ belongs to $\mathcal{D}^{\mathbb{N}}$, tends to x and satisfies, for all nonnegative integer n, $f(y_n)[0...n] = f'(x)[0...n]$. □

This implies in particular that the image set of the extended coding function f' belongs to the subshift (which is a close set) generated by w_0: $f'(\Omega') \subset X_0$.

We finally show that the extended coding function f' is one-to-one. This results of the minimality of the covered dynamical system (Ω', T').

Lemma 17. *The nonnegative orbit of any x in Ω' under the action of the extended covered rotation T' is dense in Ω'.*

Proof. Let $x, z \in \Omega'$ and $\varepsilon > 0$. By density of Ω in Ω', one can pick y in Ω at distance less than $\varepsilon/2$ from z. Consider an open ball $\mathcal{B}$ with center y and diameter less than $\varepsilon/2$ included in the open set Ω. The projected set $p_L(\mathcal{B})$ is still a nonempty open set; by minimality of the pair (α, L), there exists $n \in \mathbb{N}$ such that $R^n_{\alpha,L}(p_L(x)) \in p_L(\mathcal{B})$. Back to the covering space, we have that $T'^n(x) \in \mathcal{B}$; the point $T'^n(x)$ is thus at distance less than $\varepsilon/2$ from y and less than ε from z. □

Proposition 18. *The extended coding function $f' : \Omega' \mapsto X_0$ is one-to-one.*

Proof. By contradiction, consider $x \neq y \in \Omega'$ such that $f'(x) = f'(y)$. An easy induction argument shows that $T'^n(y) = T'^n(x) + y - x$ for any nonnegative integer n. Taking the closure set of their nonnegative orbit, we obtain that $\overline{\{T'^n(y) | n \in \mathbb{N}\}} = \overline{\{T'^n(x) | n \in \mathbb{N}\}} + y - x$. Since $\overline{\{T'^n(z) | n \in \mathbb{N}\}} = \overline{\Omega}$ for any z in Ω' (immediate consequence of Lemma 17), this implies that the set $\overline{\Omega}$ is invariant under the translation by the vector $y - x$.

Now, consider $\mathcal{B}_0 \subset \Omega$ an open ball with diameter less than $y - x$. By compacity of $\mathbb{T}_L$, there exists a positive integer n such that the intersection $R^n_{y-x,L}(p_L(\mathcal{B}_0)) \cap p_L(\mathcal{B}_0)$ is nonempty. Denote $\mathcal{B}_1 = \mathcal{B}_0 + n(y - x)$. On one hand, due to their small diameter, the balls $\mathcal{B}_0$ and $\mathcal{B}_1$ are disjoint. On the other hand, the translated ball $\mathcal{B}_1$ is still included in $\overline{\Omega}$. Thus, the intersection $\mathcal{B}_1 \cap \Omega$ is dense in $\mathcal{B}_1$ and its projected set $p_L(\mathcal{B}_1 \cap \Omega)$ is dense in $p_L(\mathcal{B}_1)$. In particular, since $p_L(\mathcal{B}_1) \cap p_L(\mathcal{B}_0)$ is open and nonnempty (indeed, $p_L(\mathcal{B}_1) = R^n_{y-x,L}(p_L(\mathcal{B}_0))$ by definition of n) the intersection $p_L(\mathcal{B}_1 \cap \Omega) \cap p_L(\mathcal{B}_0)$ is also nonempty. Given that $\mathcal{B}_0 \subset \Omega$ and that the balls $\mathcal{B}_0$ and $\mathcal{B}_1$ are disjoint, this nonemptyness is conflicting with the L-simplicity of Ω. □

3.3 The underlying group of a natural coding

Let w_0 a natural coding of a minimal rotation of the d-torus with elements $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$, . We introduce the *underlying group* of the natural coding w_0:

$$G = \sum_{i \in \mathcal{I}} \alpha_i \mathbb{Z}.$$

We now state two general lemmas about the group G, that will be useful in Sections 4 and 6.

Lemma 19. *The group G is a free abelian group of rank $d+1$, which is dense in $\mathbb{R}^d$. The family $(\alpha_1, ..., \alpha_{d+1})$ forms a basis of G.*

Proof. We first show that the group G is dense in $\mathbb{R}^d$. By minimality of the pair (α, L), the orbit $(R^n_{\alpha,L}(p_L(x_0)))_{n\in\mathbb{N}}$, which is included in $p_L(\Omega)$, is dense in $\mathbb{T}_L$. The set Ω being open and L-simple, the covered map $r_{\Omega,L}$ is well-defined and continuous, and thus, the covered orbit $(r_{\Omega,L} \circ R^n_{\alpha,L} \circ p_L(x_0))_{n\in\mathbb{N}} \subset G + x_0$ is dense in Ω. Finally, the group G is dense in the nonempty open set $\Omega - x_0 \subset \mathbb{R}^d$, and thus, in the whole space $\mathbb{R}^d$.

Now, we show that the family $(\alpha_1, ..., \alpha_{d+1})$ is free over $\mathbb{Z}$, which proves that G is a free abelian group of rank $d+1$, with basis $(\alpha_1, ..., \alpha_{d+1})$. By contradiction, assume that there exist $n_1, ..., n_{d+1} \in \mathbb{Z}$, non simultaneously equal to zero, and such that $\sum_{i\in\mathcal{I}} n_i \alpha_i = 0$. Without loss of generality, assume that $n_{d+1} \in \mathbb{N}^*$; so we have $n_{d+1}\alpha_{d+1} = -\sum_{i=1}^d n_i \alpha_i$. Now, denote $\mathcal{V}$ the vectorial space over $\mathbb{R}$ generated by the vectors $\alpha_1, ..., \alpha_d$. If $\mathcal{V}$ is a strict subspace of $\mathbb{R}^d$, then we can find a vector $e \in \mathbb{R}^d$ such that the distance between e and the subspace $\mathcal{V}$ is greater than 1 - which is impossible since G is included in $\mathcal{V}$ and dense in $\mathbb{R}^d$. Therefore, $\alpha_1, ..., \alpha_d$ form a basis of $\mathbb{R}^d$ and $N = \sum_{i=1}^d \alpha_i \mathbb{Z}$ is a lattice of $\mathbb{R}^d$. We now show that

$$G = \bigcup_{r \in \{0,...,n_{d+1}-1\}} N + r\alpha_{d+1}.$$

Let $g = \sum_{i\in\mathcal{I}} m_i \alpha_i$ an element of G. Denote respectively by q and r the quotient and the rest in the euclidean division of m_{d+1} by n_{d+1}. Then we have $g = \sum_{i=1}^d (m_i - qn_i)\alpha_i + r\alpha_{d+1}$. We conclude that G is included in the union set $\cup_{r\in\{0,...,n_{d+1}-1\}} N + r\alpha_{d+1}$; since the converse inclusion is trivially true, we have the equality. Then, as the finite union of discrete sets, G is discrete - a contradiction. Finally, the vectors $\alpha_1, ..., \alpha_{d+1}$ form a basis of the group G.

□

Lemma 20. *We have $G = L + \alpha\mathbb{Z}$.*

Proof. Since $\alpha_i = \alpha \mod L$ for all $i \in \mathcal{I}$, we immediately have $G \subset L + \alpha\mathbb{Z}$. Conversely, let $x \in L + \alpha\mathbb{Z} + x_0$. We are going to show that $x \in G + x_0$ - which will conclude the proof. By density of G and openness of Ω, there exists $g \in G$ such that $x - g \in \Omega$. Since $G \subset L + \alpha\mathbb{Z}$, we deduce that $p_L(x-g) = R^l_{\alpha,L}(\tilde{x}_0)$ for a certain $l \in \mathbb{Z}$. But then, there exists $l_1, ..., l_{d+1} \in \mathbb{Z}$ such that $x - g = x_0 + \sum l_i \alpha_i$; hence $x \in x_0 + G$. □

3.4 Discussion on the definition

The notion of natural coding of rotation, sometimes better called *natural coding of translation of the torus*, goes back to the works of Morse and Hedlund [MH40], and to the seed paper of Rauzy [Rau82] (study of the Tribonacci word) in dimension 2. Nonetheless, the terminology appears later (for instance in [CFZ00] and [Fog02].) As far as we know, the terminology was introduced, through not written, by Rauzy [Arn20].

Roughly speaking, a "natural" coding of rotation denotes a word obtained as the coding trajectory of a point of the torus, under the action of a rotation, with respect to a remarkable partition that can be covered such that the induced rotation on the associated fundamental domain coincides, on each covered piece, with a translation. Of course, we are interested in partitions with as few pieces as possible; moreover we would like coding words with minimal complexity (the complexity of an infinite word w is the function which maps each nonnegative integer n to the number of factors of length n in w.) The study of dimension 1, and the results obtained in dimension 2 (for instance [Rau82], [AI01], [BB12]) lead us to hope for a generic coding strategy with classes of words of complexity $dn + 1$ (see also [BST20] and [Fog20].)

We start by discussing the definition proposed in the article [CFZ00] (which inspired the present work), where no assumption is made on the topological nature of the partition. This definition is also used, under a weaker form (pieces are assumed disjoint up to measure 0) in [BST19].

Definition 21. *([CFZ00] No topological assumption) Let L a lattice of $\mathbb{R}^d$. A word $w \in \{1, ..., d+1\}^{\mathbb{N}}$ is a* natural coding "with no topological assumption" *of the rotation $R_{\alpha,L}$ if there exist a fundamental domain Ω of $\mathbb{T}_L$, together with a partition $\Omega = \Omega_1 \cup ... \cup \Omega_{d+1}$, such that on each Ω_i the induced rotation is a translation by a vector α_i; and the sequence w is the symbolic coding of the orbit of a point $x \in \Omega$ with respect to the Ω_i.*

This definition is not restrictive enough, as illustrated by the following proposition.

Proposition 22. *Under the axiom of choice, for any lattice $L \subset \mathbb{R}^d$ and any $\alpha \in \mathbb{R}^d$ such that (α, L) is minimal, any word in $\{1, ..., d+1\}^{\mathbb{N}}$ is a natural coding "with no topological assumption" of the rotation $R_{\alpha,L}$.*

Proof: a stupid Cantor example. Let $L \subset \mathbb{R}^d$ a lattice, and $\alpha \in \mathbb{R}^d$ such that (α, L) is minimal. Let w a word in $\{1, ..., d+1\}^{\mathbb{N}}$. Thanks to the axiom of choice, we cover each orbit $(\tilde{y}_n)_{n\in\mathbb{Z}}$ for the action of R_α on $\mathbb{T}_L$ as follows. We choose y_0 in $p_L^{-1}(\tilde{y}_0)$ and set, for any integer n, $y_n = y_0 + n\alpha$. We thus have $p_L(y_n) = p_L(y_0) + np_L(\alpha) = R_\alpha^n(\tilde{y}_0) = \tilde{y}_n$. Furthermore, we put the point y_n into the set $\Omega_{w[n]}$ if $n \geq 0$, or Ω_1 otherwise. By minimality of (α, L), each point of the torus is visited exactly once by this process, and the sets $\Omega_1, ..., \Omega_{d+1}$ form a partition of a fundamental domain of L. At last, by construction, the induced rotation coincides with the translation by the vector α on each set Ω_i, and the point indexed by 0 on each orbit admits w as symbolic coding. □

We thus have to restrict what we accept for the partition. As evidenced by Proposition 32 further, natural codings are made to preserve rotations while inducing on pieces, so the first property should ensure that these inductions are well-defined: nonempty interior for a topological study. This is what Berthé, Steiner and Thuswaldner require in [BST20]. We state here their definition with our notations.

Definition 23. *([BST20] Topological and metric assumptions, eluding borders.) Let $\alpha \in \mathbb{R}^d$ such that $(\alpha, \mathbb{Z}^d)$ is minimal. A* measurable fundamental domain *of $\mathbb{R}^d/\mathbb{Z}^d$ is a set $\Omega \subset \mathbb{R}^d$ with Lebesgue measure 1 that satisfies $\Omega + \mathbb{Z}^d = \mathbb{R}^d$. A collection $\{\Omega_1, ..., \Omega_h\}$ is said to be a* natural measurable partition *of Ω with respect to $R_{\alpha,\mathbb{Z}^d}$ if the sets Ω_i are measurable, they are the closure of their* *interior and zero measure boundaries, $\cup_{i=1}^{h}\Omega_i = \Omega$, the (Lebesgue) measure of $\Omega_i \cap \Omega_j$ is 0 for all $i \neq j$, and moreover there exist vectors $\alpha_1, ..., \alpha_h$ in $\mathbb{R}^d$ such that $\alpha_i + \Omega_i \subset \Omega$ with $\alpha_i \equiv \alpha \mod \mathbb{Z}^d$, $1 \leq i \leq h$. This allows to define a map T (which depends on the partition) as an exchange of domains defined a.e. on Ω as $T(x) = x + \alpha_i$ whenever $x \in \mathring{\Omega}_i$.*

A sequence $(i_n)_{n\in\mathbb{N}} \in \{1, ..., h\}^{\mathbb{N}}$ is said to be a natural coding *of $(\mathbb{R}^d/\mathbb{Z}^d, R_\alpha)$ w.r.t. the natural measurable partition $\Omega_1, ..., \Omega_h$ if there exists $x \in \Omega$ such that $(i_n)_{n\in\mathbb{N}}$ codes the orbit of x under the action of T , i.e. $T^n(x) \in \Omega_{i_n}$ for all $n \in \mathbb{N}$.*

By introducing objects up to measure 0, they manage to elude the question of borders, for which several arbitrary choices are admissible. However, this definition leads to induce the covered rotation on the interior of the pieces (where it is defined) instead of on the whole pieces. In our mind, this induction does not behave as well as it could be, as evidenced in Example 24.

Example 24. *For d=1, consider $L = \mathbb{Z}$ and α an irrational number in $[0, 1/2]$. We introduce $\Omega = \Omega_0 \cup \Omega_1$, where $\Omega_0 = [0, 1-\alpha]$ and $\Omega_1 = [1-\alpha, 1]$, which turns to be a natural partition according to Definition 23. Let $T_{0,ind}$ denotes the first return map to $\mathring{\Omega}_0$ of the covered rotation T. We have: $T_{0,ind}(x) = x+\alpha$ if $x \in A_0 =]0, 1-2\alpha[$ and $T_{0,ind}(x) = x+2\alpha-1$ if $x \in A_1 =]1-2\alpha, 1-\alpha[$. Hence, for all $x \in A_0 \cup A_1$, $T_{0,ind}(x)$ coincides with the rotation through the angle α modulo $(1-\alpha)$. But looking at the remaining point $x = 1-2\alpha$, we have $T_{0,ind}(x) = x + 3\alpha - 1 \not\equiv \alpha \mod (1-\alpha)$. In other words, the induced map of T on $\mathring{\Omega}_0$, which is defined everywhere, is a rotation almost everywhere, but* not *everywhere.*

This is why, following [Rau82] and [Rau84], we chose to work with an exclusively topological background. This requires to carefully examine what happen with the borders. By doing so, on one hand, we guarantee that the assignment we choose for the borders enjoys the weak continuity property (Lemma 16) which turns to be central; on the other hand, we open the definition to sets whose borders have positive Lebesgue measure. As far as we know, it is an open question to determinate if the (S-adic) Rauzy fractal of all Arnoux-Rauzy words (which are good candidates for coding of rotation of the 2-torus) have borders of measure zero.

Finally, the conditions we ask for the pieces Ω_i (to be open, and they union set to be dense) are inherited from [Rau82] (see the definition of 'morcellement'.) They are all needed in our work. In particular, we did not assume the pieces to be bounded. Indeed, this assumption is required only when resorting to Rauzy's theorem on bounded remainder sets (Theorem B further.) It would be of high interest to know (1) what remains of Rauzy's theorem if we remove this assumption; (2) if a word with infinite imbalance could be a natural coding of a minimal rotation with an unbounded pseudo-fundamental domain.

4 Stability under induction

4.1 Main result for induction

Definition 25. *[Dur98] A finite word u is a* return word *to the letter a in the recurrent word w if u starts with the letter a, contains no other occurrence of a and the word ua is a factor of w.*

Lemma 26. *[Dur98] Let w a uniformly recurrent word and $\mathcal{U}$ the set of return words to the letter a in w. Then $\mathcal{U}$ is finite. Furthermore, if w' is an element of the subshift generated by w, then the set of return words to the letter a in the word w' is again $\mathcal{U}$. If furthermore w' starts with the letter a, then it can be written in a unique way as a concatenation of elements in $\mathcal{U}$.*

Definition 27. *[Dur98] Let w a uniformly recurrent word and a a letter. Denote by $\mathcal{U}$ the set of return words of w to the letter a, that we enumerate: $u_1, ..., u_n$. Let l the index of the first occurrence of a in w. The* derivated word of w to a, *with respect to the chosen numeration, is the unique word $D_a(w)$ in $\{1, ..., n\}^{\mathbb{N}}$ satisfying: $\sigma(D_a(w)) = S^l(w)$, where σ is the substitution that maps k to the word u_k, for all $k \in \{1, ..., n\}$.*

Remark 28. *The derivated word of w to a is unique up to the choice made for the numeration. Whenever this choice has no significance, we will talk about* the *derivated word $D_a(w)$. However, it can happen that the choice of the numeration is of interest: see for instance the case of primitive C-adic word in [Section 5, Example 50.]*

Theorem A. *Let w_0 a natural coding of a minimal rotation of a d-dimensional torus, and denote by $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$ its elements, and by $(\Omega' : \Omega'_1, ..., \Omega'_{d+1})$ a borders assignment. Assume that w admits $d+1$ return words $u_1, ..., u_{d+1}$ to a letter a.*

Then there exist a second lattice M together with an angle $\beta \in \mathbb{R}^d$ such that:

1. *(β, M) is minimal;*
2. *the set Ω'_a is a fundamental domain of M;*
3. *for all x in Ω'_a, $T_{ind,a}(x) = x + \beta \mod M$, where $T_{ind,a}$ denotes the first return map of the covered rotation $T' = r_{\Omega',L} \circ R_\alpha \circ p_L$ to the set Ω'_a;*
4. *the derivated word to the letter a, $D_a(w_0)$, is a natural coding of the rotation of $\mathbb{T}_M$ through the angle β, whose elements and borders assignment are explicit (they are given in Proposition 34.)*

Furthermore, the return words $u_1, ..., u_{d+1}$ to the letter a form a basis of the free group generated by $\mathcal{I} = \{1, ..., d+1\}$.

Remark 29. *The induced natural coding $D_a(w_0)$ inherits from the choice made for the borders assignment of the original natural coding w_0. There is no need to resort to the axiom of choice a second time.*

4.2 Proof of Theorem A

Let $a \in \mathcal{I} = \{1, ..., d+1\}$ and $w_0 \in \mathcal{I}^{\mathbb{N}}$ a word admitting $d+1$ return words $u_1, ..., u_{d+1}$ to the letter a. We assume that w_0 is a natural coding of a minimal rotation of the d-torus with elements $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$ and borders assignment $(\Omega' : \Omega'_1, ..., \Omega'_{d+1})$. We keep the notations and tools developed in Section 3. In particular, we denote by f the coding function with respect to the partition $(\Omega_1, ..., \Omega_{d+1})$ of the L-simple set Ω, $\mathcal{D}$ the maximal set on which f is defined, and f' the extended coding function on Ω'. We still denote $T = r_{\Omega,L} \circ R_\alpha \circ p_L$ the covered rotation on the L-simple set Ω, which is defined on $r_{\Omega,L}(\tilde{\Omega} \cap R_\alpha^{-1}(\tilde{\Omega}))$ and its extension $T' = r_{\Omega',L} \circ R_\alpha \circ p_L$ defined on the whole fundamental domain Ω'.

The dynamical system (Ω', T') inherits from the minimality of the dynamical system $(\mathbb{T}_L, R_\alpha)$. Since the set Ω'_a has nonempty interior (it contains the open set Ω_a), all trajectories end up with passing through it. We can define the first return map to Ω'_a:

$$\begin{array}{rccl} T_{ind,a}: & \Omega'_a & \to & \Omega'_a \\ & x & \mapsto & T'^{n_0}(x), \qquad\qquad \text{where} \quad n_0 = \inf\{n \in \mathbb{N} | T'^n(x) \in \Omega'_a\} < \infty. \end{array}$$

For $i \in \mathcal{I}$, we introduce the sets:

$$A_i = \bigcap_{k=0}^{|u_i|} T'^{-k}(\Omega_{u_i a[k]}) \quad \subset \Omega_a \qquad \text{and} \qquad A'_i = \bigcap_{k=0}^{|u_i|} T'^{-k}(\Omega'_{u_i a[k]}) \quad \subset \Omega'_a.$$

Lemma 30. *[A subpartition]*

- *The sets A'_i, for i in $\mathcal{I}$, form a partition of Ω'_a.*
- *For all i in $\mathcal{I}$, A_i is a nonempty open set included and dense in A'_i.*

Proof. By definition of f', we have $A'_i = f'^{-1}([u_i a])$ for all $i \in \mathcal{I}$. Since the words u_i contain a unique occurrence of the letter a, at the first position, the cylinders $[u_i a]$ are pairwise disjoint, and so are the preimage sets A'_i. Moreover, for all $y \in \Omega'_a$, $f'(y) \in X_0 \cap [a] \subset \cup_{i \in \mathcal{I}}[u_i a]$ by Lemmas 16 and 26; consequently $y \in \cup_{i\in\mathcal{I}} A'_i$.

The set A_i is the preimage, by the continuous map $r_{\Omega,L}$, of the open set $\cap_{k=0}^{|u_i|} R_\alpha^{-k}(\Omega_{u_i a[k]}) \subset \Omega$ - it is thus open. Since the word $u_i a$ is a factor of w_0, there exists a nonnegative integer n such that $S^n(w_0) \in [u_i a]$. The set A_i is nonempty since it contains the set $f^{-1}([u_i a])$, which contains itself the point $T^n(x_0)$. The inclusion $A_i \subset A'_i$ is inherited from the inclusions $\Omega_j \subset \Omega'_j$ for j in $\mathcal{I}$. At last, for the density, consider $y \in A'_i$. Applying Lemma 16, we can find a sequence $(x_n)_n \in \mathcal{D}^{\mathbb{N}}$ such that $x_n \to y$ and $\mathrm{pref}_{|u_i a|}(f'(x_n)) = \mathrm{pref}_{|u_i a|}(f'(y)) = u_i a$.

□

The following lemma states that the induced map $T_{ind,a}$ acts on the sets $A'_1, ..., A'_{d+1}$ as an exchange of pieces.

Lemma 31 (Exchange of pieces)**.** *Let i in $\mathcal{I}$. For all $x \in A'_i$, $T_{ind,a}(x) = x + \beta_i$, where $\beta_i = \sum_{j\in\mathcal{I}} |u_i|_j \, \alpha_j$.*

Proof. If x belongs to A'_i, then its coding word $f'(x)$ belongs to $[u_i a]$, meaning that the $|u_i|$ first steps of the trajectory of x are fully known. More precisely, starting from $\Omega'_{u_i[0]} = \Omega'_a$, x is translated by the vector α_a and falls into $\Omega'_{u_i[1]}$; then it is translated by $\alpha_{u_i[1]}$ and falls into $\Omega'_{u_i[2]}$, and so on; until arriving into $\Omega'_{u_i[|u_i|-1]}$ from where it is translated by $\alpha_{u_i[|u_i|-1]}$ and falls at last - and for the first time - into Ω'_a. All in all, from $A'_i \subset \Omega'_a$ to its first return into Ω'_a, the point x was translated by the vector $\beta_i = \sum_{j\in\mathcal{I}} |u_i|_j \, \alpha_j$. □

We introduce the vectors of $\mathbb{R}^d$:

$$v_k = \beta_k - \beta_{d+1} \qquad \text{for } k \in \{1, ..., d\}$$

and the subgroup $M = \sum_{k=1}^d \mathbb{Z} v_k$.

Proposition 32 (A rotation on a new torus)**.** *(i) The subgroup M is a lattice of $\mathbb{R}^d$.*

(ii) The vectors β_j, $j \in \mathcal{I}$, are equals modulo M.
From now on, we denote $\beta = \beta_{d+1}$.

(iii) The pair (β, M) is minimal.

(iv) For all x in Ω'_a, $T_{ind,a}(x) = x + \beta \mod M$.

Proof. The assertions (ii) and (iv) stem from the definition of M and the Lemma 31. We now propose to show that the subgroup $\sum_{i\in\mathcal{I}}\beta_i\mathbb{Z} = \beta\mathbb{Z} + \sum_{k=0}^{d} v_k\mathbb{Z}$ is dense in $\mathbb{R}^d$. This fact implies that:

- the vectors v_k, $k \in \{1, ...d\}$, are linearly independent - thus proving (i);
- the trajectory of $p_M(0)$ is dense in the torus $\mathbb{T}_M$ for the action of the rotation $R_{\beta,M}$ - thus proving (iii).

Let l such that $x := T^l(x_0) \in \Omega'_a$. The trajectory of x under the map T being dense in Ω', the sequence $(T^n_{ind,a}(x))_{n\in\mathbb{N}}$ - consisting of all the points falling into Ω'_a - is dense in the open subset Ω_a and is, by Lemma 31, included in $x + \sum_{i\in\mathcal{I}}\beta_i\mathbb{Z}$. We conclude that the subgroup $\sum_{i\in\mathcal{I}}\beta_i\mathbb{Z}$ is dense in $\Omega_a - x$, which is a nonempty open set of $\mathbb{R}^d$, and thus, is actually dense in $\mathbb{R}^d$ itself.

□

Proposition 33. *The set Ω'_a is a fundamental domain of M.*

Proof. We prove that the projection map $p_M : \Omega'_a \to \mathbb{T}_M$ is one-to-one and onto.

Be $x, y \in \Omega'_a$ such that $p_M(x) = p_M(y)$, i.e. $y = x + \sum_{j=1}^{d} b_j v_j$ for some $b_j \in \mathbb{Z}$. Since each β_i is a linear combination of $\alpha_1, ..., \alpha_{d+1}$ (Lemma 31), which are all congruent to α modulo L, it comes that $\beta_i = k_i\alpha \mod L$, where k_i is the length of the associated return word u_i. The previous equality can then be rewritten $y = x + \sum_{j=1}^{d} b_j(k_j - k_{d+1})\alpha + l$, for some $l \in L$; hence $p_L(y) = R^n_{\alpha,L}(p_L(x))$, with $n = \sum_{k=1}^{d} b_j(k_j - k_{d+1}) \in \mathbb{N}$ (if needed, we swap x and y), and thus $y = T^n(x)$. But, given that both x and y belong to Ω'_a, y is not only on the trajectory of x for the action of T, but also for the action of the first return map $T_{ind,a}$: there exists $m \in \mathbb{N}$ s.t. $y = T'^m_{ind,a}(x)$. Finally, we had $y = x \mod M$ and now, we have $y = x + m\beta \mod M$, meaning that either $m = 0$, or the trajectory of $p_M(x)$ under $R_{\beta,M}$ is periodic, which is forbidden by minimality of (β, M). It eventually comes that $m = 0$, and $x = y$ - hence the injectivity.

Now, let $\overline{y} \in \mathbb{T}_M$. By minimality of $(-\beta, M)$, and because $p_M(\Omega'_a)$ has nonempty interior, there exist an element $\overline{x} \in p_M(\Omega'_a)$ and a nonnegative integer n such that $R_{\beta,M}(\overline{x}) = \overline{y}$. Denote $x = r_{\Omega'_a,M}(\overline{x}) \in \Omega'_a$ (the covering map $r_{\Omega'_a,M}$ is well-defined by the previous paragraph.) The trajectory of x under the map $T_{ind,a}$ remains in Ω'_a; in particular, $y := T'^n_{ind,a}(x)$ belongs to Ω'_a. Then, $p_M(y) = p_M(T'^n_{ind,a}(x)) = R^n_\beta(p_M(x)) = R^n_\beta(\overline{x}) = \overline{y}$. We conclude that $\overline{y}$ admits a preimage by p_M in Ω'_a - hence the surjectivity. □

Proposition 34 (An induced natural coding.)**.** *The derivated word of w_0 to the letter a, $D_a(w_0)$, is a natural coding of the minimal rotation of $\mathbb{T}_M$ through the angle β, with elements $(\beta, M); (A : A_1, ..., A_{d+1});\ T^l(x_0); (\beta_1, ..., \beta_{d+1}))$, where l denotes the minimal nonnegative integer such that $S^l(w_0)$ starts with the letter a. Furthermore, $(\Omega'_a : A'_1, ..., A'_{d+1})$ is a borders assignment of this natural coding.*

Proof. Partition of a pseudo-fundamental domain.

By Lemma 30, the sets $A_1, ..., A_{d+1}$ are nonempty, open and pairwise disjoint. Furthermore, their union set $A = \cup_{i\in\mathcal{I}} A_i$ inherits from the M-simplicity of the set Ω'_a it is included in (Proposition 33.) We now show that the projection set $p_M(A)$ is dense in the torus $\mathbb{T}_M$. Denote $y_0 = T^l(x_0) \in \Omega_a$. The trajectory of y_0 under the action of T is included in $\mathcal{D}$; so its trajectory under the action of the induced map $T_{ind,a}$ is included in $\Omega_a \cap \mathcal{D} \subset A$. From Proposition 32, we deduce that the trajectory of $\overline{y}_0 := p_M(y_0)$ under the rotation $R_{\beta,M}$, which is dense in $\mathbb{T}_M$ by minimality of (β, M), is included in $p_M(A)$; this implies that the set $p_M(A)$ is dense in $\mathbb{T}_M$.

Exchange of pieces. By Lemma 31, for all $i \in \mathcal{I}$ and for all $\overline{x} \in p_M(A_i) \cap R_{\beta,M}^{-1}(p_M(A))$, we have $r_{A,M} \circ R_{\beta,M}(\overline{x}) = T_{ind,a}(r_{A,M}(\overline{x})) = r_{A,M}(\overline{x}) + \beta_i$.

A coding trajectory. By construction of the sets $A_1, ..., A_{d+1}$, for all nonnegative integer n, we have $R_{\beta,M}^n(\overline{y}_0) \in p_M(A_i)$ if and only if $T_{ind,a}'^n(y_0) \in A_i$ if and only if $f(T_{ind,a}'^n(y_0))$ starts with the word u_i if and only if the n-th letter of the derived word $D_a(w_0)$ is i.

A borders assignment. (1) & (2) By Lemma 30, for all $i \in \mathcal{I}$, A_i is included in A_i' and the sets A_i' form a partition of Ω_a'. (3) The set Ω_a' is a fundamental domain of M (Proposition 33.) (4) For all $i \in \mathcal{I}$ and for all $x \in A_i'$, $r_{\Omega_a',M} \circ R_{\beta,M} \circ p_M(x) = T_{ind,a}(x) = x + \beta_i$ (Proposition 32 and Lemma 31.) (5) Let $x \in \Omega_a'$ and $q \in \mathbb{N}$. Denote $l = \max_{i \in \mathcal{I}} |u_i|$ and σ the extraction given by the definition of borders assignment associated with the natural coding w_0 for the point x (seen as an element of Ω') and the integer $q(l+1)$. For any nonnegative integer m, the prefix of length $q(l+1)+1$ of the words $f(T^{\sigma(m)}(x_0))$ and $f'(x)$ coincide. In particular, the sequence $(T^{\sigma(m)}(x_0))_{m \in \mathbb{N}}$ is included in Ω_a, so it is a subsequence of the trajectory of x_0 under the action of the first return map to Ω_a'. Denote by τ the extraction such that for all nonnegative integer m, $T_{ind,a}^{\tau(m)}(x_0) = T^{\sigma(m)}(x_0)$. We immediately have that $T_{ind,a}^{\tau(m)}(x_0) = T^{\sigma(m)}(x_0) \to_{m \to \infty} x$. Furthermore, by definition of l, the prefix of length $q(l+1)+1$ of $f'(x)$ contains at least $q+2$ occurrences of the letter a; we deduce that, for any $m \in \mathbb{N}$, the first $q+1$ return words to a of the symbolic trajectory of x and $T_{ind,a}^{\tau(m)}(x_0)$ for the action of $T_{ind,a}$ coincide, i.e.: for all $n \in \{0, ..., q\}$, for all nonnegative integer m, $T_{ind,a}^{\tau(m)+n}(x_0) \in A_{\iota_n}$, where the index ι_n is defined by $T_{ind,a}^n(x) \in A_{\iota_n}'$. □

Proposition 35. *The return words $u_1, ..., u_{d+1}$ to the letter a form a basis of the free group generated by $\mathcal{I}$.*

Proof. We are going to show that $\mathcal{M} = (|u_j|_i)_{i,j} \in GL(\mathbb{Z})$. Since $(\beta_1, ..., \beta_{d+1}) = (\alpha_1, ..., \alpha_{d+1})\mathcal{M}$, and since the vectors $\alpha_1, ..., \alpha_{d+1}$ form a basis of the $\mathbb{Z}$-module G (Lemma 19), it is sufficient to show that the $d+1$ vectors $(\beta_1, ..., \beta_{d+1})$ are free over $\mathbb{Z}$. This is the case by Lemma 19 again, given that the word $D_a(w_0)$ is a natural coding of the minimal rotation with elements $((\beta, M); (A : A_1, ..., A_{d+1}); T^l(x_0); (\beta_1, ..., \beta_{d+1}))$ (Proposition 34.) □

Propositions 32, 33, 34 and 35 prove Theorem A.

4.3 Correction of the proof of [CFZ00]

We now complete the idea of [CFZ00] to prove, resorting to Rauzy's theorem on remainder sets (Theorem B below), that being a natural coding of a minimal rotation, with a bounded fundamental domain, implies finite imbalance.

A direct and general proof of this fact can be found in the latest version of [Thu19].

Proposition 36. *Let w_0 a natural coding of a minimal rotation of a d-dimensional torus, and denote by $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$ its elements, and by $(\Omega' : \Omega_1', ..., \Omega_{d+1}')$ a borders assignment. Assume that w_0 admits $d+1$ return words to a letter a. Assume furthermore that the pseudo-fundamental domain Ω is bounded. Then the set Ω_a is a bounded remainder set for for any trajectory $(R_\alpha(\tilde{x}))_{n \in \mathbb{N}}$, and the imbalance of w_0 on the letter a is finite.*

Definition 37 (Following [Rau84])**.** *A set A is a* bounded remainder set for a sequence $(u_n)_{n \in \mathbb{N}}$ *if there exist two real numbers (ν, C) such that, for all positive integer N:*

$$|\sum_{n=0}^{N-1} 1_A(u_n) - N\nu| < C.$$

The numbers ν and C can be understood as a frequency and a tolerance margin for the event 'falling into A'. So, A is a bounded remainder set for the sequence (u_n) means that (u_n) is well-distributed relatively to A: the observed frequency of visits to A converges to it expected value at speed $1/n$.

Theorem B. *[Rau84] Be d a positive integer, L a lattice of $\mathbb{R}^d$, α an element of $\mathbb{R}^d$ such that (α, L) is minimal. Be $A \subset \mathbb{R}^d$, L-simple, bounded, with nonempty interior. Let T denotes the transformation on A induced by the rotation R_α.*

If there exist a lattice M of $\mathbb{R}^d$, together with an element $\beta \in \mathbb{R}^d$, such that:

(i) A is M-simple

then $p_L(A)$ is a bounded remainder set for all sequence $(R_\alpha^n(\tilde{x}))_n$, with $\tilde{x} \in \mathbb{T}_L$.

Remark 38. *In [Rau84], Rauzy includes the assumption of boundedness of A in the definition of* L-simplicity*. This assumption is crucial at two stages in the proof of his theorem.*

Remark 39. *This theorem gives a sufficient condition for a set to be a bounded remainder set. A necessary and sufficient condition generalizing this criterion is given in [Fer92] under the framework of measurable dynamical systems. Though not mentioned, the assumption of boundedness is still needed.*

Proof of Proposition 36. The pseudo-fundamental domain Ω being bounded, so are the fundamental domain Ω' and its subset Ω'_a. Therefore, by Theorem A and B, for all $\tilde{x} \in \mathbb{T}_L$, $p_L(\Omega'_a)$ is a bounded remainder set for the sequence $(R_\alpha(\tilde{x}))_{n\in\mathbb{N}}$.

On another hand, by Definition 2 (natural coding), for all nonnegative integer n, we have $R_\alpha(\tilde{x}_0) \in \Omega'_a$ if and only if $w_0[n] = a$ if and only if $S^n(w_0) \in [a]$. We deduce from this equivalence that the cylinder $[a]$ is also a bounded remainder set for the sequence $(S^n(w_0))_{n\in\mathbb{N}}$: there exist two real numbers ν and C such that for all positive integer N:

$$|\sum_{n=0}^{N-1} 1_{[a]}(S^n(w_0)) - N\nu| < C.$$

In other words, for all positive integer N, $|\text{pref}_N(w_0)|_a \in]\nu N - C; \nu N + C[$, from which we deduce that for all factor $u \in \mathcal{F}(w_0)$, $|u|_a \in]\nu|u| - 2C; \nu|u| + 2C[$. This implies that the imbalance of w_0 on the letter a is lower than the constant $4C$. □

Remark 40. *In fact, finite imbalance is equivalent to the cylinder $[a]$ being a bounded remainder set for the sequence $(S^n(w_0))_n$ (see [Ada03].)*

Remark 41. *The main mistake in the original proof of [CFZ00] is that no information on the second lattice M is given, and thereby, one cannot guarantee that the set A is M-simple. This confusion is still present in the first versions of the lecture notes [Thu19].*

5 Applications

5.1 Tree words

Theorem A claims that being a natural coding of a minimal rotation of the d-torus is a property preserved by the derivation operation. This is why good candidates should be families of words stable under this operation. This is the case for the class of infinite words admitting d return words to any factor [BPS08]; this is also the case of its remarkable subclass comprised of tree words.

A finite word u is a *return word to the factor* v in the recurrent word w if $u = w[i]...w[j-1]$, where i and j are two consecutive occurrences of v; or equivalently, if $uv \in \mathcal{F}(w)$, v is a prefix of uv and if there are exactly two occurrences of v in uv [Dur98]. This definition is of course consistent with Definition 25 (return word to a letter.)

Let w an infinite word over an alphabet A, and u one of its factor. Following [BFD$^+$15a], we denote $L(u)$ (resp. $R(u)$) the set of letters a in A such that au (resp. ua) is still a factor of w. The *extension graph* of u is the undirected graph whose vertices are the disjoint union of $L(u)$ and $R(u)$, and whose edges are the pairs $(a, b) \in L(u) \times R(u)$ such that aub is a factor of w. An infinite word w is a *tree word* (or a *dendric word*, in recent texts) if the extension graph of each of its factors is acyclic and connected (viz. a tree.)

On the two-letter alphabet, the set of infinite words admitting two return words to any factor, the set of uniformly recurrent tree words, the set of Sturmian words and the set of words whose subshift is a natural coding of a minimal rotation of the circle coincide ([Vui01], [JV00].)

More generally:

- Uniformly recurrent tree words on the alphabet $\mathcal{I} = \{1, ..., d+1\}$ admit $d+1$ return words to any factor (so in particular to each letter), that moreover form a basis of the free group over $\mathcal{I}$ [BFD$^+$15a] ; but we also have examples of infinite words admitting $d+1$ return words to any factors that are not tree words.

- Strict episturmian words are uniformly recurrent tree words, but there exists other families of words, such as primitive C-adic words (see Definition 46 below), that belong to this class too.

The following proposition and corollary are immediate applications of Theorem A.

Proposition 42. *If a uniformly recurrent tree word w_0 on the alphabet $\mathcal{I} = \{1, ...d+1\}$ is a natural coding of a minimal rotation of the d-torus, then all its derivated sequences to a letter are also natural codings of a minimal rotation of the d-torus.*

Corollary 43. *No uniformly recurrent tree word with infinite imbalance is a natural coding of a minimal rotation of the* 2*-torus with a bounded pseudo-fundamental domain.*

In particular, no Arnoux-Rauzy word with infinite imbalance is a natural coding of a minimal rotation of the 2-torus. This result strengthens the one stated in [CFZ00]. An other construction of an Arnoux-Rauzy word with infinite imbalance is given in [And20]. Likewise, no primitive C-adic word with infinite imbalance is a natural coding of a minimal rotation of the 2-torus. Primitive C-adic words with infinite imbalance have been constructed in [And18].

On the counterpart, remember that a lot of Arnoux-Rauzy words and C-adic words are natural codings of rotation [BST20].

Once a uniformly recurrent tree word is a natural coding of a minimal rotation of the d-torus, then its derivated words to the d letters of the alphabet are again tree words (see [BFD$^+$15b]) and natural codings of minimal rotations (Theorem A) - in particular, they are again uniformly recurrent by Lemma 3. We can thus iterate the derivation, and study the trajectory of words under this operation. In the remarkable cases of Arnoux-Rauzy and primitive C-adic words, these trajectories are driven by generalized euclidean maps (often referred to as *multidimensional continued fraction algorithms*), as evidenced through the S-adic frameworks (see the book [Sch00] for a general introduction to multidimensional continued fractions, and for instance the surveys [Ber11] or [BD14] for their study from the symbolic dynamical viewpoint.)

5.2 Return words for Arnoux-Rauzy words (under the S-adic framework)

We recall that *Arnoux-Rauzy words* are infinite words on the alphabet $\mathcal{I} = \{1, 2, 3\}$ with complexity $p(n) = 2n + 1$, such that for each n there is exactly one right and one left special factor of length n [AR91]. By a result of Boshernitzan [Bos84], Arnoux-Rauzy words are uniquely ergodic; hence the existence of frequencies, which are positive, for each factor. We introduce the set $AR = \{\sigma_i | i \in \mathcal{I}\}$ of Arnoux-Rauzy substitutions:

$$\begin{aligned} \sigma_i : \ & \mathcal{I} \to \mathcal{I}^* \\ & i \mapsto i \\ & j \mapsto ij \text{ for } j \in \mathcal{I} \backslash \{i\}. \end{aligned}$$

The following theorem evidences the link between Arnoux-Rauzy words and a generalized Euclid map.

Theorem C ([AS13])**.** *Let w an Arnoux-Rauzy word. Then:*

1. *there exists a unique sequence of substitutions (called* directive sequence*) $d = (\sigma_{i_n})_n$ in $AR^{\mathbb{N}}$, and a unique Arnoux-Rauzy word w' such that: 1. each prefix of w' is a left-special factor 2. the sets of factors of w and w' are equal; 3. $w' = lim_{n \to \infty} \sigma_{i_0} \circ ... \circ \sigma_{i_{n-1}}(1)$.*

2. *we also have $w' = lim_{n \to \infty} \sigma_{i_0} \circ ... \circ \sigma_{i_{n-1}}(2)$ and $w' = lim_{n \to \infty} \sigma_{i_0} \circ ... \circ \sigma_{i_{n-1}}(3)$.*

3. *each Arnoux-Rauzy substitution appears infinitely many times in d;*

4. *d is uniquely defined by the frequencies of letters in w: the sequence $(i_n)_{n \in \mathbb{N}}$ is the symbolic trajectory of the letters frequency vector under the action of the generalized Euclid map:*

$$F_{AR} : \quad (x, y, z) \quad \mapsto \quad \begin{cases} (x - y - z, y, z) & \text{if } x \geq y + z, \\ (x, y - x - z, z), & \text{if } y \geq x + z, \\ (x, y, z - x - y), & \text{if } z \geq x + y, \end{cases}$$

with respect to the partition given by its piecewise definition.

As evidenced by Lemma 31 and Proposition 32, the action of the induction/derivation operation of a natural coding on the lattice and the angle of the rotation is driven by the abelianized vectors of the return words to a letter.

We now describe how to get the return words to a letter. This result comes from [JV00]; we just state it under the S-adic formalism, i.e. as a function of the sequence of substitutions given by Theorem C.

Notation 44. *We denote by s the* circular shift on (nonempty) finite words*: $s(u) = a_2 ... a_n a_1$, where $a_1, ..., a_n$ are letters and $u = a_1 ... a_n$. The map s is bijective.*

Theorem D ([JV00], under a slighly different formalism.)**.** *If w is an Arnoux-Rauzy word with directive sequence $d = (\sigma_{i_n})_n$, and $a \in \{1, 2, 3\}$ is a letter, then w admits three return words to a, namely: $s^{-1} \circ d_0 \circ ... \circ d_{n_0 - 1} \circ s \circ d_{n_0}(b)$, for $b \in \{1, 2, 3\}$, where s is the circular shift on finite words and $n_0 = min\{n \in \mathbb{N} | \, i_n = a\}$. Furthermore, the derivated word of w to a is an Arnoux-Rauzy word with directive sequence $d' = (\sigma_{i_n})_{n > n_0}$.*

Return words to any factor are described in [JV00].

At last, we denote by $M_\sigma = (|\sigma(j)|_i)_{(i,j) \in \mathcal{I}^2}$ the *incidence matrix* of a substitution σ.

Corollary 45. *Let w an Arnoux-Rauzy word with directive sequence $d = (d_n)_n$, and $a \in \{1, 2, 3\}$ a letter. If w is a natural coding of a minimal rotation of the 2-torus, with elements $((\alpha, L); (\Omega : \Omega_1, \Omega_2, \Omega_3); x_0; (\alpha_1, \alpha_2, \alpha_3))$, then the vectors β_1, β_2 and β_3 describing the induced rotation on Ω_a are given by:*

$$(\beta_1, \beta_2, \beta_3) = (\alpha_1, \alpha_2, \alpha_3)M_{d_0}...M_{d_{n_0}},$$

where $n_0 = min\{n \in \mathbb{N}|\ i_n = a\}$.

Proof. We have $(\beta_1, \beta_2, \beta_3) = (\alpha_1, \alpha_2, \alpha_3)\mathcal{M}$, with $\mathcal{M} = (|u_j|_i)_{i,j}$. By theorem D, for all $i, j \in \{1, 2, 3\}$, $|u_j|_i = |d_0 \circ ... \circ d_{n_0}(j)|_i$; therefore $\mathcal{M}$ is the incidence matrix of the substitution $d_0 \circ ... \circ d_{n_0}$, and $(\beta_1, \beta_2, \beta_3) = (\alpha_1, \alpha_2, \alpha_3)M_{d_0}...M_{d_{n_0}}$. □

5.3 Return words for primitive C-adic words

We now deal with primitive C-adic words. This class of words was introduced in [CLL17], as resulting from the research of a generalized Euclid map defined on $(\mathbb{R}^+)^3$ defined for any projective direction -contrary to F_{AR} which is defined for almost none (see [AR91], [AS13] and [AHS13])- and producing words with the lowest complexity possible: $p(n) = 2n + 1$. This led to the map:

$$F_C : \quad (x, y, z) \quad \mapsto \quad \begin{cases} (x - z, z, y) & \text{if } x \geq z \\ (y, x, z - x) & \text{otherwise} \end{cases},$$

and to the associated substitutions $C = \{c_1, c_2\}$ given by:

$$\begin{array}{rrcl} c_1 : & 1 & \mapsto & 1 \\ & 2 & \mapsto & 13 \\ & 3 & \mapsto & 2 \end{array} \qquad \begin{array}{rrcl} c_2 : & 1 & \mapsto & 2 \\ & 2 & \mapsto & 13 \\ & 3 & \mapsto & 3. \end{array}$$

Definition 46 ([CLL17])**.** *An infinite word w is* C-adic *if there exist a directive sequence $d = (d_n) \in C^{\mathbb{N}}$, together with a letter $a \in \{1, 2, 3\}$, such that w can be written $w = lim_{n\to\infty} d_0 \circ ... \circ d_{n-1}(a)$.*

As long as d contains infinitely many occurrences of c_1 and c_2, the sequence of finite words $(d_0 \circ ... \circ d_{n-1}(a))_n$ converges to an infinite words w that, furthermore, does not depend on the letter a [CLL17].

Proposition 47 ([CLL17])**.** *Let w a C-adic word with directive sequence d. If $d \notin C^*\{c_1^2, c_2^2\}^\omega$, then w is a uniformly recurrent tree word.*

We will call *primitive* a C-adic word whose directive sequence does not belong to $C^*\{c_1^2, c_2^2\}^\omega$. In fact, this condition is equivalent to primitivity in the S-adic sense (see [CLL17].) The class of primitive C-adic words will naturally emerge while studying return words.

Lemma 48. *A primitive C-adic word w admits a unique directive sequence, that can be deduced from the knowledge of its set of factors $\mathcal{F}(w)$.*

Proof. Let $d = (d_n)_{n\geq 0} \in \{c_1, c_2\}^{\mathbb{N}}$ such that $d \notin C^*\{c_1^2, c_2^2\}^\omega$, and w its (unique) associated C-adic word, which is primitive. Denote by w' and w'' the C-adic words obtained with the directive sequences $(d_n)_{n\geq 1}$ and $(d_n)_{n\geq 2}$ respectively, which are also primitive. We then have that 2 is factor of w'', which implies that 13 is factor of w', which implies in turn that 12 or 23 (exclusively) is a factor of w. If $12 \in \mathcal{F}(w)$, then $d_0 = c_1$; if $23 \in \mathcal{F}(w)$, then $d_0 = c_2$. Furthermore, the word w' can be deduced from the knowledge of w and d_0, so that in the end, the entire sequence d is determined by iterating the process. □

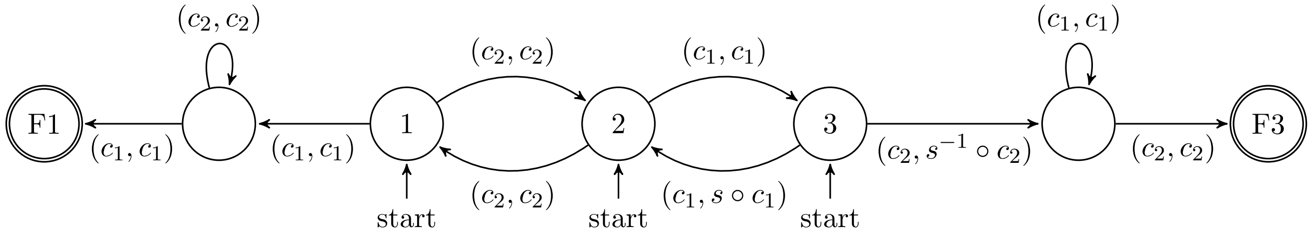
Notation 49. *We denote by l (resp. r) the map which extracts the first (resp. the second) component x (resp. y) of a pair (x, y).*

Theorem E. *Let w a primitive C-adic word with directive sequence d, and $a \in \{1, 2, 3\}$ a letter. The following assertions are true.*

1. *There exists in the automaton of partial quotients of C-adic words (Figure 5.3) a unique accepted path $e = (e_0, ..., e_{n_1})$ starting from the initial state a and such that the finite sequence $l_e = l(e_0)...l(e_{n_1}) \in \{c_1, c_2\}^*$ is a prefix of d.*
   We denote by n_2 the length of this prefix, and by w' the primitive C-adic word with directive sequence $(d_n)_{n \geq n_2}$.

2. *The set of return words to the letter a of w, denoted $\mathcal{U}$, is the image set of the alphabet $\{1, 2, 3\}$ by the application $r_e = r(e_0) \circ ... \circ r(e_{n_1})$.*

3. *The set $\mathcal{U}$ contains 3 elements, and if we denote them by $u_i = r_e(i)$ for i in $\{1, 2, 3\}$, then the derivated word of w to a, with respect to this numeration, is the word w' if the final state of e is F_1, and $S(w')$ (where S denotes the shift map) if the final state of e is F_3.*



with s the circular shift on finite words: $s(a_1...a_n) = a_2...a_n a_1$.

Figure 1: Automaton of partial quotients for C-adic words.

Example 50. *We consider the primitive C-adic word w with directive sequence $d = c_1c_1c_2c_2(c_1c_2)^\omega$, and the letter $a = 3$.*

$$w = 113131213113121311313113121311313121311313113121311313121311...$$

Applying Theorem E, we obtain $l(e) = c_1c_1c_2c_2$ and $r_e = s \circ c_1 \circ c_1 \circ s^{-1} \circ c_2 \circ c_2$; hence $u_1 = 311$, $u_2 = 3121$, $u_3 = 31$ and w' is the (primitive) C-adic word with directive sequence $(c_1c_2)^\omega$. Since the path e leads to the final state F_3, the derivated word (with respect to the chosen numeration) is:

$$D_3(w) = S(w') = 32121312132131213212132131213212131213212132131213212131213213121321...$$

A final remark: if we had chosen another numeration, say $\tilde{u}_1 = 3121$, $\tilde{u}_2 = 311$ and $\tilde{u}_3 = 31$, we would have obtained:

$$\tilde{D}_3(w) = 312123212312321231212312321231212321231212312321231212321231232123121...$$

which is not is the subshift of a primitive C-adic word, since it does not contain the factor $13 = c_1(2) = c_2(2)$.

Proof of Theorem E. Let w a primitive C-adic word with directive sequence d. Since d contains in particular infinitely many occurrences of c_1 and c_2, so does any sequence of the form $(d_n)_{n\geq n_0}$ for $n_0 \in \mathbb{N}$, so that the sequence of finite words $(d_{n_0} \circ ... \circ d_{n-1}(b))_{n\geq n_0}$ converges to an infinite word w', which does not depend on the letter b, and is again a primitive C-adic word.

1. If d starts with c_1, then the set of return words to 2 is the image set by c_1 of the return words to 3 of the word with directive sequence $(d_n)_{n\geq 1}$. Symmetrically, if d starts with c_2, then the set of return words to 2 is the image set by c_2 of the return words to 1 of the word with directive sequence $(d_n)_{n\geq 1}$.

2. If d starts with c_1, then there exists n_0 such that d starts with $c_1 \circ c_2^{n_0} \circ c_1$. If $n_0 = 2k+1$, the images of the letters 1, 2 and 3 by $c_1 \circ c_2^{n_0} \circ c_1$ are respectively $132^k, 132^{k+1}$ and 12^{k+1}; if $n_0 = 2k$, they are respectively $12^k, 12^{k+1}$ and 132^k. Since furthermore the word w' with directive sequence $(d_n)_{n\geq n_0+2}$ contains the three letters 1, 2 and 3, w contains three return words to 1, which are the images of the letters by the substitution $c_1 \circ c_2^{n_0} \circ c_1$. Otherwise, if d starts with c_2, the set of return words to 1 is the image set by c_2 of return words to 2 of the word with directive sequence $(d_n)_{n\geq 1}$.

3. Symmetrically, if d starts with c_2, then there exists n_0 such that d starts with $c_2 \circ c_1^{n_0} \circ c_2$. If $n_0 = 2k+1$, the images of the letters 1, 2 and 3 by $c_2 \circ c_1^{n_0} \circ c_2$ are respectively $2^{k+1}3, 2^{k+1}13$ and 2^k13; if $n_0 = 2k$, they are respectively $2^k13, 2^{k+1}3$ and 2^k3. Since furthermore the word w' with directive sequence $(d_n)_{n\geq n_0+2}$ contains the three letters 1, 2 and 3, w contains three return words to 3, which are the images of the letters by the application $s^{-1} \circ c_2 \circ c_1^{n_0} \circ c_2$. Otherwise, if d starts with c_1, the return words to 3 are the images by the map $s \circ c_1$ of the return words to 2 of the word with directive sequence $(d_n)_{n\geq 1}$.

Since $d \notin C^*\{c_1^2, c_2^2\}^\omega$, this recursive process ends. We conclude by observing that the images by c_1 (resp. c_2) of two distinct finite words are again distinct. Indeed, two distinct words u and v can always be written $u = u's$ and $v = v's$ (resp. $u = pu'$ and $v = pv'$), where u' and v' end (resp. start) with distinct letters (one of them at most could possibly be empty); then $c_1(u')$ and $c_1(v')$ also end (resp. start) with distinct letters, implying $c_1(u) \neq c_1(v)$ (resp. $c_2(u) \neq c_2(v)$.) □

Corollary 51. *Let w a primitive C-adic word with directive sequence $d = (d_n)_n$, and $a \in \{1, 2, 3\}$ a letter. If w is a natural coding of a minimal rotation of the 2-torus, with elements $((\alpha, L); (\Omega : \Omega_1, \Omega_2, \Omega_3); x_0; (\alpha_1, \alpha_2, \alpha_3))$, then the vectors β_1, β_2 and β_3 describing the induced rotation on Ω_a are given by:*

$$(\beta_1, \beta_2, \beta_3) = (\alpha_1, \alpha_2, \alpha_3)M_{d_0}...M_{d_{n_2-1}},$$

where n_2 is the length of the unique prefix of d accepted by the partial quotients automaton for C-adic words from the initial state a.

Proof. The proof is identical to the proof of Corollary 45 for Arnoux-Rauzy words. □

At last, we deduce from Theorem E an algorithm which, given a primitive C-adic words w and v one of its factor, outputs the three return words to v of w.

Theorem F. *Let w a primitive C-adic word with directive sequence d, and $v \in \mathcal{F}(w)\backslash\{1, 2, 3\}$ one of its factors of length at least 2. Let $n_0 = \min\{n \in \mathbb{N} | u \in \mathcal{F}(d_0 \circ ... \circ d_{n_0-1}(2))\}$. Let p and s such that $(d_0 \circ ... \circ d_{n_0-1}(2)) = pvs$. At last, let $\mathcal{P} = \{p^{-1}d_0 \circ ... \circ d_{n_0-1}(u)p\ |u\in \mathcal{U}\}$, where $\mathcal{U}$ is the set of return words to the letter 2 of the C-adic word w' with directive sequence $(d_n)_{n\geq n_0}$. Then $\mathcal{P}$ contains three words, which start with v and pave a suffix of w.*

Proof. Since $v \in \mathcal{F}(w)$ and since the sequence $(d_0 \circ ... \circ d_{n-1}(1))_n$ shares a growing common prefix with w, we can define the nonnegative integer:

$$n_0 = \min\{n \in \mathbb{N} | u \in \mathcal{F}(d_0 \circ ... \circ d_{n-1}(1)) \cup \mathcal{F}(d_0 \circ ... \circ d_{n-1}(2)) \cup \mathcal{F}(d_0 \circ ... \circ d_{n-1}(3))\}.$$

Since v contains at least two letters, n_0 is actually positive. Observe that if $v \in \mathcal{F}(d_0 \circ ... \circ d_{n-1}(1))$ for $n \geq 1$, then $v \in \mathcal{F}(d_0 \circ ... \circ d_{n-2}(1))$ if $d_{n-1} = c_1$ and $v \in \mathcal{F}(d_0 \circ ... \circ d_{n-2}(2))$ otherwise. Symmetrically, if $v \in \mathcal{F}(d_0 \circ ... \circ d_{n-1}(3))$ for $n \geq 1$, then $v \in \mathcal{F}(d_0 \circ ... \circ d_{n-2}(2))$ if $d_{n-1} = c_1$ and $v \in \mathcal{F}(d_0 \circ ... \circ d_{n-2}(3))$ otherwise. We deduce from the minimality of n_0 that $v \in \mathcal{F}(d_0 \circ ... \circ d_{n_0-1}(2))$ and $v \notin \mathcal{F}(d_0 \circ ... \circ d_{n_0-1}(a))$ for $a \in \{1, 3\}$.

The C-adic word w with directive sequence $(d_n)_{n\in\mathbb{N}}$ being primitive, so is the C-adic word w' with directive sequence $(d_n)_{n\geq n_0}$; the word w' thus admits three return words to the letter 2, whose set is denoted by $\mathcal{U}$. Let $k_0 = \min\{k \in \mathbb{N} | S^k(w') \in [2]\}$, where S denotes the shift map. Then the words in $\mathcal{U}$ pave the infinite word $S^{k_0}(w')$. Denote by k_1 the length of the image by the substitution $d_0 \circ ... \circ d_{n_0-1}$ of the prefix of length k_0 of w', and $k_2 = k_1 + |p|$, where p is such that $d_0 \circ ... \circ d_{n-1}(2) = pvs$. Then the set $\mathcal{P} = \{p^{-1} d_0 \circ ... \circ d_{n_0-1}(u)p \ |u \in \mathcal{U}\}$, which contains three elements (the images of distinct words by c_1 or c_2 remaining distinct - see the end of the proof of Theorem E) that start with v, pave the infite word $S^{k_2}(w)$. □

The set $\mathcal{P}$ is not always the set of return words to the factor v of w, as illustrated by Example 53. Nonetheless, the set of return words to the factor v of w is easily deduced from $\mathcal{P}$.

Corollary 52. *If we denote $\mathcal{P} = \{p_1, p_2, p_2\}$, the set of return words to the factor v of w is exactly the set of return words to v of the finite word $p_1p_2p_3p_1$.*

Example 53. *We consider the primitive C-adic word w with directive sequence $d = c_2c_2c_1c_2c_1c_1(c_1c_2)^\omega$, and the factor $v = 31 \in \mathcal{F}(w)$.*

$$w = 13313323133132313313313231331332313313313231331332313313231331332313 31...$$

Applying Theorem F, we obtain $n_0 = 6$, and $\sigma = d_0 \circ ... \circ d_5 = c_2 \circ c_2 \circ c_1 \circ c_2 \circ c_1 \circ c_1$ is given by:

$$\begin{array}{rrcl} \sigma : & 1 & \mapsto & 133 \\ & 2 & \mapsto & 13\underline{31}323 \\ & 3 & \mapsto & 13323, \end{array}$$

hence $p = 13$ and $s = 323$. By Theorem E, the three return words to the letter 2 of the primitive C-adic word w' with directive sequence $(d_n)_{n\geq n_0} = (c_1c_2)^\omega$ are 21, 213 and 2131. We finally obtain the paving set: $\mathcal{P} = \{313231331332313313,\ 3132313313,\ 313231331332313\}$, from which we deduce, following Corollary 52, the three return words to 31 of w: 313, 3132 and 31332.

In this example, no element of $\mathcal{P}$ is a return word to the factor v. This is a consequence of v appearing in $w = \sigma(w')$ not only as factor of $\sigma(2)$, but also at each junction of images of letters by σ: indeed, here, all images by the substitution σ starts with 1 and ends with 3.

6 Stability under exduction (reverse induction)

We now prove that being a natural coding of a minimal rotation is a property which passes through the reverse operation of induction, that we call, following Rauzy (see [AI01]), *exduction.*

We start by an example in dimension 1 (Sturmian case.)

Example 54. *For d=1, consider $M = \mathbb{Z}$ and β an irrational number. We introduce $A_1 =]0, 1-\alpha[$, $A_2 =]1-\alpha, 1[$, $A'_1 = [0, 1-\alpha[$ and $A'_2 = [1-\alpha, 1[$. Then the standard Sturmian word with slope β, that we denote w_{st}, is a natural coding with elements $((\beta, M), A : (A_1, A_2), \beta, (\beta_1, \beta_2))$, where $\beta_1 = \beta$ and $\beta_2 = \beta - 1$ and borders assignment $(A' : (A'_1, A'_2))$.*

Now, we consider the substitution σ given by $\sigma(1) = 1$ and $\sigma(2) = 12$, and its incidence matrix M_σ:

$$M_\sigma = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

We set $(\alpha_1, \alpha_2) = (\beta_1, \beta_2)M_\sigma^{-1}$, which gives: $\begin{cases} \alpha_1 = \beta \\ \alpha_2 = -1. \end{cases}$

Denote $\alpha = \alpha_1 = \beta$ and $L = (\alpha_2 - \alpha_1)\mathbb{Z} = (\beta + 1)\mathbb{Z}$. At last, we introduce $\Omega_1 = A =]0, 1[$, $\Omega'_1 = A' = [0, 1[$ $\Omega_2 =]1, 1 + \beta[$ and $\Omega'_2 = [1, 1 + \beta[$ (see Figure 2.)

Observe that the pair (α, L) is minimal, that the sets $\{\Omega'_1, \Omega'_2\}$ form a partition of a fundamental domain for the lattice L, and that the rotation $R_{\alpha,L}$ acts on the piece Ω'_1 (resp. Ω'_2) of the fundamental domain as a translation by the vector α_1 (resp. α_2.) In fact, the word $\sigma(w_{st})$ is a natural coding of a minimal rotation of the circle with elements $((\alpha, L), \Omega : (\Omega_1, \Omega_2), \beta, (\alpha_1, \alpha_2))$ and borders assignment $(\Omega' : (\Omega'_1, \Omega'_2))$.

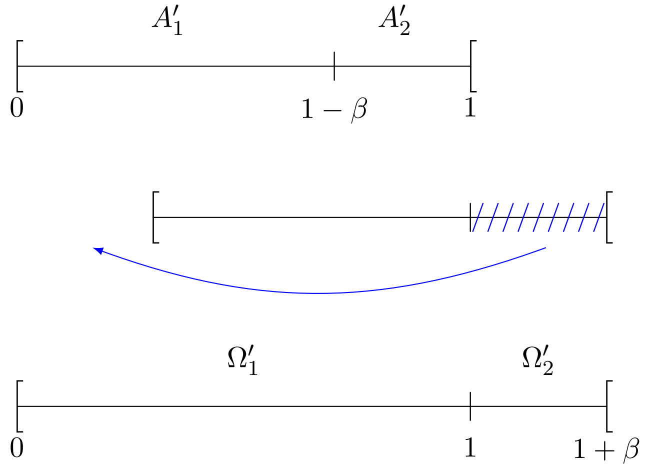


Figure 2: Example of exduction in dimension 1.

Our aim is to show that this construction is valid in a more general context.

6.1 Main result for exduction

Assumptions. Let w_0 a natural coding of a minimal rotation of the d-torus, with elements $((\beta, M); (A : A_1, ..., A_{d+1}); x_0; (\beta_1, ..., \beta_{d+1}))$ and borders assignment $(A' : A'_1, ..., A'_{d+1})$. Let $a \in \mathcal{I} = \{1, ..., d + 1\}$ a letter, and u_1,...,$u_{d+1} \in \mathcal{I}^*$ such that:

1. For all i in $\mathcal{I}$, the word u_i starts with the letter a and admits no other occurrence of a;
2. the matrix $\mathcal{M} = (|u_j|_i)_{i,j} \in GL_{d+1}(\mathbb{Z})$.

Construction. Let $(\alpha_1, ..., \alpha_{d+1}) = (\beta_1, ..., \beta_{d+1})\mathcal{M}^{-1}$. We denote $\alpha = \alpha_a$ and L the additive subgroup of $\mathbb{R}^d$ given by: $L = \sum_{i=1, i\neq a}^{d+1}(\alpha_i - \alpha_a)\mathbb{Z}$. Furthermore, for $i \in \mathcal{I}$ and $k \in \{0, ..., |u_i|\}$, we denote:

$$v_{i,k} = \sum_{l=0}^{k-1} \alpha_{u_i[l]} \in \mathbb{R}^d.$$

For $k \in K_i := \{0, ..., |u_i| - 1\}$, we set $A_{i,k} = A_i + v_{i,k}$ and $A'_{i,k} = A'_i + v_{i,k}$. Observe that for $k = |u_i|$ we have $v_{i,k} = \beta_i$, and that $A'_i + \beta_i \subset A'$. Now, let:

$$\Omega_j = \bigcup_{\substack{(i,k)\in\mathcal{I}\times K_i \\ u_i[k]=j}} A_{i,k} \qquad \text{and} \qquad \Omega'_j = \bigcup_{\substack{(i,k)\in\mathcal{I}\times K_i \\ u_i[k]=j}} A'_{i,k},$$

and set $\Omega := \cup_{j\in\mathcal{I}}\Omega_j$ and $\Omega' := \cup_{j\in\mathcal{I}}\Omega'_j$. At last, denote by σ the substitution given by $\sigma(i) = u_i$.

Theorem G. *The word $\sigma(w_0)$ is a natural coding of a minimal rotation of the d-torus, with elements $((\alpha, L); (\Omega : \Omega_1, ..., \Omega_{d+1}); x_0; (\alpha_1, ..., \alpha_{d+1}))$ and borders assignment $(\Omega' : \Omega'_1, ..., \Omega'_{d+1})$. Furthermore, we have $\Omega_a = A$, $\Omega'_a = A'$, and the induced map of $T_{sus} := r_{\Omega',L} \circ R_{\alpha,L} \circ p_L$ on the set with nonempty interior Ω_a is the map $T = r_{A',M} \circ R_{\beta,M} \circ p_M$.*

6.2 Proof of Theorem G

Let $G = \sum_{i\in\mathcal{I}} \beta_i\mathbb{Z}$ the underlying group of the natural coding w_0.

Lemma 55. *The vectors $\alpha_1, ..., \alpha_{d+1}$ form a basis of G.*

Proof. By Lemma 19, the vectors $\beta_1, ..., \beta_{d+1}$ form a basis of the $\mathbb{Z}$-module G and, since $\mathcal{M} \in GL(\mathbb{Z})$, so do the vectors $\alpha_1, ..., \alpha_{d+1}$. □

Corollary 56. *The subgroup L is a lattice of $\mathbb{R}^d$ and the pair (α, L) is minimal.*

Proof. We write $G = \sum_{i\in\mathcal{I}} \alpha_i\mathbb{Z} = \sum_{i\in\mathcal{I}\setminus\{a\}}(\alpha_i - \alpha)\mathbb{Z} + \alpha\mathbb{Z}$. The group G being dense, with a similar argument than in Lemma 19, we show that the vectors $\alpha_i - \alpha$, for $i \in \mathcal{I}\setminus\{a\}$, form a basis of $\mathbb{R}^d$, and thus, that the group $L = \sum_{i\in\mathcal{I}\setminus\{a\}}(\alpha_i - \alpha)\mathbb{Z}$ is a lattice of $\mathbb{R}^d$. By density of G again, we obtain that the pair (α, L) is minimal. □

Lemma 57. *Let $x, y \in \mathbb{R}^d$. The three following assertions are equivalent:*

(i) the points x and y are equal modulo G;

(ii) there exists a unique $n \in \mathbb{Z}$ such that $p_L(y) = R^n_{\alpha,L}(p_L(x))$.

(iii) There exists a unique $m \in \mathbb{Z}$ such that $p_M(y) = R^m_{\beta,M}(p_M(x))$.

Proof. $(i \Rightarrow ii)$ Let $x, y \in \mathbb{R}^d$ such that $y - x \in G$. Then, $\alpha_1, ...\alpha_{d+1}$ being a basis of G (Lemma 55), there exist $n_1, ..., n_{d+1} \in \mathbb{Z}$ such that $y - x = \sum_{i\in\mathcal{I}} n_i\alpha_i$. Therefore, $p_L(y) = R^n_{\alpha,L}(p_L(x))$, with $n = \sum_{i\in\mathcal{I}} n_i$. The pair (α, L) being minimal (56), the integer n is unique. $(ii \Rightarrow i)$ Let $x, y \in \mathbb{R}^d$ and $n \in \mathbb{Z}$ such that $p_L(y) = R^n_{\alpha,L}(p_L(x))$. Then there exists $l \in L \subset G$ such that $y = x + n\alpha_a + l$, hence $y = x + g$, with $g = n\alpha_a + l \in G$. The equivalence between (i) and (iii) is given by Lemma 20 and by the minimality of (β, M).

□

Lemma 58. *Let $x, y \in A'$ that are equal modulo G. Denote by $(l_1, ..., l_{d+1})$ and $(m_1, ..., m_{d+1})$ the coordinates of the element $y - x$ with respect to the bases $(\alpha_1, .., .\alpha_{d+1})$ and $(\beta_1, ...\beta_{d+1})$ respectively. Then the integers $l_1, ..., l_{d+1}, m_1, ..., m_{d+1}$ are simultaneously nonnegative or nonpositive. Furthermore, if we set $l = \sum_{i\in\mathcal{I}} l_i$ and $m = \sum_{i\in\mathcal{I}} m_i$, we have $|l| \geq |m|$ and $l_a = m$.*

Corollary 59 (immediate)**.** *If $x, y \in A'$ are equal modulo G, then there exists two unique integer, l and m, such that $p_L(y) = R^l_{\alpha,L}(p_L(x))$ and $p_M(y) = R^m_{\beta,M}(p_M(x))$. Furthermore, l and m are simultaneously positive, negative or equal to zero.*

Proof of Lemma 58. Let $x, y \in A'$ that are equal modulo G. By Lemma 57, the exists $m \in \mathbb{Z}$ such that $p_M(y) = R^m_{\beta,M}(p_M(x))$. First, assume that $m \geq 0$. Since $x, y \in A'$, by definition of natural coding, there exist $m_1, ..., m_{d+1} \in \mathbb{N}$ such that:

$$\begin{aligned} y &= x + (\beta_1, ..., \beta_{d+1})(m_1, ..., m_{d+1})^t \\ &= x + (\alpha_1, ..., \alpha_{d+1})\mathcal{M}(m_1, ..., m_{d+1})^t \\ &= x + (\alpha_1, ..., \alpha_{d+1})(l_1, ..., l_{d+1})^t. \end{aligned}$$

Since $\mathcal{M} \in GL(\mathbb{Z})$ and has nonnegative entries, we have $l_i \geq 0$ for all $i \in \mathcal{I}$, and $l_1, ...l_{d+1}$ are simultaneously equal to zero if and only if $m_1, ..., m_{d+1}$ are simultaneously equal to zero if and only if $m = 0$; we also obtain $l \geq m$. We lead a symmetric argument for $m \leq 0$ and conclude that in both cases, $p_L(y) = R^l_{\alpha,L}(p_L(x))$, with $l_1, ..., l_{d+1}, m_1, ..., m_{d+1}$ simultaneously nonnegative or nonpositive, and $|l| \geq |m|$. At last, since the words u_i, for $i \in \mathcal{I}$, contain a unique occurrence of the letter a, the $a - th$ line of the matrix $\mathcal{M}$ only contains the entry 1, and $m = \sum_{i \in \mathcal{I}} m_i = l_a$. □

Proposition 60. *The following assertions are true.*

1. *The sets $\Omega_1, ..., \Omega_{d+1}$ are nonempty and open.*
2. *For all $i \in \mathcal{I}$, the set Ω_i is included and dense in Ω'_i.*
3. *The sets $\Omega'_1, ..., \Omega'_{d+1}$ are pairwise disjoint.*

Proof. Let $i \in \mathcal{I}$, and $k \in K_i$. The sets $A_{i,k}$ and $A'_{i,k}$ are the translated sets, by the vector $v_{i,k}$, of A_i and A'_i respectively, from which they inherit of the following properties: $A_{i,k}$ is nonempty, open, included and dense in $A'_{i,k}$. Now, let $j \in \mathcal{I}$. As the finite and nonempty union, for some $i \in \mathcal{I}$ and some $k \in K_i$, of the sets $A_{i,k}$, the set Ω_j is nonempty, open and furthermore included and dense in and Ω'_j, which is the union, for the same indices, of the sets $A'_{i,k}$.

We now prove that the sets $A'_{i,k}$ are pairwise disjoint. Let $y \in A'_{i_1,k_1} \cap A'_{i_2,k_2}$, with $i_1, i_2 \in \mathcal{I}$, $0 \leq k_1 < |u_{i_1}|$ and $0 \leq k_2 < |u_{i_2}|$. Denote $x_j = y - v_{i_j,k_j}$, which belongs to A'_{i_j}, for $j = 1, 2$. Since $x_1, x_2 \in A'$, and since $x_2 - x_1$ is an integer linear combination of $\alpha_1, ..., \alpha_{d+1}$, it comes that $x_2 - x_1 \in G$ and by Lemma 58 and Corollary 59, that the coordinates $(l_1, ..., l_{d+1})$ of $x_2 - x_1$ with respect to the basis $(\alpha_1, ..., \alpha_{d+1})$ are simultaneously nonnegative or nonpositive (w.l.o.g. say nonnegative.) Therefore, if $k_1 = 0$, we have successively $k_2 = 0$, $x_1 = x_2$ and, since the sets $A'_1, ..., A'_{d+1}$ form a partition of A', $i_1 = i_2$. Now, assume that k_1 is positive. If $k_2 = 0$, then $y = x_2$, $l_a = 1$ and thereby $p_M(y) = R_{\beta,M}(p_M(x_1))$. Since $x_1 \in A'_{i_1}$, by definition of natural coding we have $y = x_1 + \beta_{i_1}$, which is conflicting with the hypothesis $k_1 < |u_1|$. So, if k_1 is positive, then k_2 is positive too. In this case, we have $l_a = |\text{pref}_{k_1}(u_{i_1})|_a - |\text{pref}_{k_2}(u_{i_2})|_a = 0$ since the word u_j, for $j \in \{1, 2\}$, contains exactly one occurrence of the letter a, at the first position. By Lemma 58 and Corollary 59 again, we have $p_M(x_2) = p_M(x_1)$; by M-simplicity of A', we obtain that $x_1 = x_2$ and $i_1 = i_2$. Then we have $0 = \sum_{l=k_1}^{k_2-1} \alpha_{u_{i_1}[l]}$, which implies $k_1 = k_2$. So the sets $A'_{i,k}$ are pairwise disjoint. We conclude, by observing that each $A'_{i,k}$, for $i \in \mathcal{I}$ and $k \in K_i$, belongs to exactly one set Ω'_j, that the sets $\Omega'_1, ..., \Omega'_{d+1}$ are pairwise disjoint. □

Proposition 61. *The set Ω' is a fundamental domain of the torus $\mathbb{T}_L$.*

Proof. We first show that the set A' is L-simple, from which we deduce that Ω' is L-simple; we conclude by proving that the projection map $p_L : \Omega' \mapsto \mathbb{T}_L$ is onto.

Let $x, y \in A'$ such that $p_L(x) = p_L(y)$. Since $L \subset G$, the points x and y are equal modulo G and, by Corollary 59, there exist two integers l and m, that are unique, such that $p_L(y) = R^l_{\alpha,L}(p_L(x))$

and $p_M(y) = R^m_{\beta,M}(p_M(x))$. Here, we already have $l = 0$, and since $|m| \leq |l|$, we obtain $m = 0$; hence $p_M(y) = p_M(x)$ and by M-simplicity of A', $x = y$. This proves the L-simplicity of A'.

Now, let z_1 and $z_2 \in \Omega'$ such that $p_L(z_1) = p_L(z_2)$. By construction, there exists two 3-tuples (x_1, i_1, k_1) and (x_2, i_2, k_2) with $i_1, i_2 \in \mathcal{I}$, such that for $j = 1, 2$, $x_j \in A'_{i_j} \subset A'$, $k_j \in K_{i_j}$ and $y_j = x_j + v_{i_j,k_j}$. We are going to show that $x_1 = x_2$. On one hand, the points x_1 and x_2 are in the same equivalent class modulo G (which is the class of z_1 and z_2), hence:

$$\left\{\begin{array}{lll} & p_M(x_2) = R^{l+1}_{\beta,M}(p_M(x_1)) & \text{with } l \geq 0, \\ \text{or} & p_M(x_2) = R^{l}_{\beta,M}(p_M(x_1)) & \text{with } l \leq 0. \end{array}\right.$$

This implies, since x_1 and x_2 belongs to A', and $x_1 \in A'_{i_1}$:

$$\left\{\begin{array}{lll} & p_L(x_2) = R^{m+|u_{i_1}|}_{\alpha,L}(p_L(x_1)) & \text{with } m \geq 0, \\ \text{or} & p_L(x_2) = R^{m}_{\alpha,L}(p_L(x_1)) & \text{with } m \leq 0. \end{array}\right.$$

On the other hand, without loss of generality, assume that $k_2 \leq k_1$. Then, if we set $y_1 = x_1 + v_{i_1,k_1-k_2}$, we have $p_L(x_2) = p_L(y_1) = R^{k_1-k_2}_{\alpha,L}(p_L(x_1))$ with $0 \leq k_1 - k_2 \leq |u_{i_1}| - 1$. The only possibility is thus $k_1 = k_2$, from which we deduce successively $p_L(x_1) = p_L(x_2)$, the equality $x_1 = x_2$ by L-simplicity of A', the equality $i_1 = i_2$ by pairwise disjointedness of the sets A'_j for $j \in \mathcal{I}$, and in the end, $z_1 = z_2$. Therefore, the set Ω' is L-simple.

At last, we show that $p_L : \Omega' \mapsto \mathbb{T}_L$ is onto. Let $\tilde{y} \in \mathbb{T}_L$ and denote $k = min\{n \in \mathbb{N} | R^{-n}_{\alpha,L}(\tilde{y}) \in p_L(A')\}$, which is finite since (α, L) is minimal (Corollary 56) and $p_L(A')$ has nonempty interior. Denote also $\tilde{x} = R^{-k}_{\alpha,L}(\tilde{y})$ and $x = r_{A',L}(\tilde{x})$ its covering into A' (which is L-simple by the first part of the proof), and $i \in \mathcal{I}$ such that $x \in A'_i$. Then, since $R^{-k+|u_i|}_{\alpha,L}(\tilde{y}) \in p_L(A')$, by minimality of k we must have $k < |u_i|$. So if we set $y = x + v_{i,k}$, which belongs to Ω' by construction, it comes that $p_L(y) = p_L(x) + k\alpha = R^k_{\alpha,L}(\tilde{x}) = \tilde{y}$. We conclude that Ω' is a fundamental domain of the torus $\mathbb{T}_L$. □

Hereafter, we denote by $T_{sus} = r_{\Omega',L} \circ R_{\alpha,L} \circ p_L$ the covered rotation in the fundamental domain Ω'.

Proposition 62. *The following assertions are true.*

1. *For all $j \in \mathcal{I}$, for all $y \in \Omega'_j$, we have $T_{sus}(y) = y + \alpha_j$.*
2. *We have $\Omega_a = A$ and $\Omega'_a = A'$. Furthermore, the induced map of T_{sus} on the set with nonempty interior Ω'_a is the map T.*
3. *For all nonnegative integer n, $T^n_{sus}(x_0) \in \Omega_{\sigma(w)[n]}$.*
4. *For all $y \in \Omega'$ and for all $q \in \mathbb{N}$, there exists an extraction τ such that: (i) $T^{\tau(m)}_{sus}(x_0) \to_{m\to\infty} y$; (ii) for all $n \in \{0, ..., q\}$ and for all nonnegative integer m, $T^{\tau(m)+n}_{sus}(x_0) \in \Omega_{\iota_n}$, where ι_n is defined by $T^m_{sus}(y) \in \Omega'_{\iota_n}$.*

Proof. (1) Let $j \in \mathcal{I}$ and $y \in \Omega'_j$. Since $R_{\alpha,L}(p_L(y)) = p_L(y + \alpha_j)$, to prove the assertion, we need to show that $y + \alpha_j \in \Omega'$. Let (i, k) the unique pair, with $i \in \mathcal{I}$ and $k \in K_i$, such that $y \in A'_{i,k}$. By definition of Ω'_j, the indices i and k satisfy $u_i[k] = j$. Thus, if $k < |u_i| - 1$, then $y + \alpha_j \in A'_{i,k+1} \subset \Omega'$. Otherwise, let $x = y - v_{i,k}$. Then we have $y + \alpha_j = x + v_{i,|u_i|} = x + \beta_i \in A' \subset \Omega'$, which ends the proof.

(2) Since each u_i, for $i \in \mathcal{I}$, admits exactly one occurrence of the letter a, at the first position, we have by construction $\Omega_a = A$ and $\Omega'_a = A'$. Moreover, for all $i \in \mathcal{I}$ and for all $x \in A'_i$, we have $min\{n \in \mathbb{N}^* | T^n_{sus}(x) \in \Omega'_a\} = |u_i|$, and by (1), $T^{|u_i|}_{sus}(x) = x + v_{i,|u_i|} = x + \beta_i = T(x)$.

(3) Let $(n_k)_{k\in\mathbb{N}}$ the sequence of indices such that for all $k \in \mathbb{N}$, $T^{n_k}_{sus}(x_0) = T^k(x_0) \in A'$ (it actually belongs to A), which is well-defined by (2). Let $n \in \mathbb{N}$. Denote by k the unique nonnegative integer such that $n_k \leq n < n_{k+1}$ and by i the unique index in $\mathcal{I}$ such that $T^k(x_0) \in A_i$ (consequence of Proposition 60.) Then, on one hand we have $\sigma(w)[n] = u_i[n - n_k]$, and on the other hand, $T^n_{sus}(x_0) = T^{n_k}_{sus}(x_0) + v_{i,n-n_k} \in A_{i,n-n_k} \subset \Omega_{u_i[n-n_k]}$. Finally, we have $T^n_{sus}(x_0) \in \Omega_{\sigma(w)[n]}$.

(4) Let $y \in \Omega'$ and $q \in \mathbb{N}$. By construction, there exist $i \in \mathcal{I}$ and $k \in K_i$ such that $x := y - v_{i,k} \in A'$. Denote by φ the extraction given by the definition of border assignments associated with the natural coding w_0 for the point $x \in A'$ and the integer $k+q$. Then, we have that $T^{\varphi(m)}(x_0) \to_{m\to\infty} x$ and for any nonnegative integer m, the first $k+q+1$ letters of $f'(x)$ and $f(T^{\varphi(m)}(x_0))$ coincide, which immediately implies, since no image of letters by the substitution σ is the empty word, that for any m, the $k+q+1$ first letters of the words $\sigma(f'(x))$ and $\sigma(f(T^{\varphi(m)}(x_0))$ coincide as well. Since the sequence $(T^m(x_0))_{m\in\mathbb{N}}$ is a subsequence of $(T^m_{sus}(x_0))_{m\in\mathbb{N}}$, we can define an extraction ψ such that for all m, $T^{\psi(m)}_{sus}(x_0) = T^{\varphi(m)}(x_0)$. We finally set $\tau(m) = \psi(m) + k$. Thus, on one hand we have that $T^{\tau(m)}_{sus}(x_0) \to_{m\to\infty} y$; on the other hand, for all $n \in \{0, ..., q\}$ and for all nonnegative integer m, $T^{\tau(m)+n}_{sus}(x_0) \in \Omega_{\iota_n}$, where ι_n is defined by $T^n_{sus}(y) \in \Omega'_{\iota_n}$.

□

Proof of Theorem G. Proof of Theorem G results of Corollary 56 and Propositions 60, 61 and 62. □

6.3 Consequences for Arnoux-Rauzy and primitive C-adic words

Theorem G applies in particular to Arnoux-Rauzy and primitive C-words.

Proposition 63. *Let w an Arnoux-Rauzy word and $\sigma \in AR^*$. Assume that w is a natural coding of a minimal rotation of the 2-torus. Then $\sigma(w)$ is also a natural coding of a minimal rotation of the 2-torus, whose elements and borders assignment can be explicitly described from the elements and the choice made for the borders assignment of the natural coding w. In particular, the piecewise translation vectors $\alpha_1, \alpha_2, \alpha_3$ of $\sigma(w)$ satisfy:*

$$(\alpha_1, \alpha_2, \alpha_3) = (\beta_1, \beta_2, \beta_3)M_\sigma^{-1},$$

where $M_\sigma = (|\sigma(j)|_i)_{i,j}$ is the incidence matrix of the substitution σ, and β_1, β_2 and β_3 are the piecewise translation vectors of w.

Proof. It is sufficient to prove the proposition for $\sigma \in AR$. For $\sigma \in AR$, we immediately have that the words $u_i = \sigma(i)$, for $i \in \{1, 2, 3\}$ satisfy the two assumptions of Theorem G. □

Proposition 64. *Let w a primitive C-adic word and $\sigma \in C^*$. Assume that w is a natural coding of a minimal rotation of the 2-torus. Then, there exists $k \in \mathbb{N}$ such that $S^k(\sigma(w))$ is also a natural coding of a minimal rotation of the 2-torus, whose elements and borders assignment can be explicitly described from the elements and the choice made for the borders assignment of the natural coding w. In particular, the piecewise translation vectors $\alpha_1, \alpha_2, \alpha_3$ of $S^k(\sigma(w))$ satisfy:*

$$(\alpha_1, \alpha_2, \alpha_3) = (\beta_1, \beta_2, \beta_3)M_\sigma^{-1},$$

where $M_\sigma = (|\sigma(j)|_i)_{i,j}$ is the incidence matrix of the substitution σ, and β_1, β_2 and β_3 are the piecewise translation vectors of w.

Proof. Again, it is sufficient to prove the proposition for the substitutions c_1 and c_2. Let w a primitive C-adic word, and assume that w is a natural coding of a minimal rotation of the 2-torus. Denote by $(d_n)_{n\in\mathbb{N}}$ its directive sequence. By Theorem E, there exists a unique accepted path $e = (l_e, r_e)$ in the automaton of partial quotients for C-adic words (see Figure 5.3) that starts from the initial state 2 and such that l_e is a prefix of $(d_n)_{n\in\mathbb{N}}$. Denote by $w' = D_2(w)$ the derivated word of w relatively to the letter 2. By Theorem A, the word w' is a natural coding of a minimal rotation of the 2-torus. Besides, the path $e_1 = (c_1, s \circ c_1) \cdot e$ (where the symbol $\cdot$ denotes the concatenation operation), starting from the initial state 3, is accepted by the automaton. Therefore, the words $u_i = r_{e_1}(i)$, for $i \in \{1,2,3\}$, are the three return words to the letter 3 in the primitive C-adic word $c_1(w)$ and, thereby, satisfy the assumptions of Theorem G. Thus, the word $\sigma(w)$, where the substitution σ is given by $\sigma(i) = u_i$ for $i \in \{1,2,3\}$, is equal to the word $S^k(c_1(w))$ for a certain $k \in \mathbb{N}$, and is a natural coding of a minimal rotation of the 2-torus, whose elements and borders assignment are explicitly given by those of w', which themselves are explicitly given by the elements and the choice made for borders assignment of the natural coding w. In particular, if $(\alpha_1, \alpha_2, \alpha_3)$, $(\beta_1, \beta_2, \beta_3)$ and $(\gamma_1, \gamma_2, \gamma_3)$ respectively denote the piecewise translation vectors of the natural codings $S^k(c_1(w))$, w and w', then we have, on one hand $(\gamma_1, \gamma_2, \gamma_2) = (\beta_1, \beta_2, \beta_3) M_{d_0} ... M_{d_{n_0-1}}$, where n_0 is the length of the path e, and on the other hand $(\gamma_1, \gamma_2, \gamma_2) = (\alpha_1, \alpha_2, \alpha_3) M_{c_1} M_{d_0} ... M_{d_{n_0-1}}$. Each matrix being invertible, we conclude that $(\alpha_1, \alpha_2, \alpha_3) = (\beta_1, \beta_2, \beta_3) M_{c_1}^{-1}$. A symmetric argument applies to $c_2(w)$. □

Theorem H. *For Arnoux-Rauzy and primitive C-adic subshifts, the property of being a natural coding of a minimal rotation of the* 2*-torus does not depend on any prefix of the directive sequence* $(d_n)_{n\in\mathbb{N}}$.

Proof. Let w an Arnoux-Rauzy (resp. primitive C-adic) word, and denote by $(d_n)_{n\in\mathbb{N}}$ its directive sequence. Assume that w is a natural coding of a minimal rotation of the 2-torus. On the first hand, we showed in Proposition 63 (resp. Proposition 64) that for all $\sigma \in AR^*$ (resp. $\sigma \in C^*$), the subshift generated by $\sigma(w)$ is a natural coding of a minimal rotation of the 2-torus. On the other hand, we claim that for all $n_0 \in \mathbb{N}$, the Arnoux-Rauzy (resp. primitive C-adic) subshift with directive sequence $(d_n)_{n\geq n_0}$ is again a natural coding of a minimal rotation of the 2-torus. Indeed, by inducing on letters as many times as needed, we can find $n_1 \geq n_0$ such that the word with directive sequence $(d_n)_{n\geq n_1}$ is a natural coding of a minimal rotation of the 2-torus (Proposition 42, Theorems D and E.) But then, by applying Proposition 63 (resp. 64) with $\sigma = d_{n_0} \circ ... \circ d_{n_1-1}$, we obtain that the Arnoux-Rauzy (resp. primitive C-adic) subshift with directive sequence $(d_n)_{n\geq n_0}$ is also a natural coding of a minimal rotation of the 2-torus. □

References

[Ada03] Boris Adamczewski. Balances for fixed points of primitive substitutions. *Theoretical Computer Science*, 307(1):47–75, 2003.

[AHS13] Artur Avila, Pascal Hubert, and Alexandra Skripchenko. On the Hausdorff dimension of the Rauzy gasket. *Bulletin de la Société mathématique de France*, 144:539–568, 2013.

[AI01] Pierre Arnoux and Shunji Ito. Pisot substitutions and Rauzy fractals. *Bulletin of the Belgian Mathematical Society - Simon Stevin*, 8(2):181–207, 2001.

[And18] Mélodie Andrieu. Autour du déséquilibre des mots C-adiques. *Journées Montoises d'informatique théorique*, 2018.

[And20] Mélodie Andrieu. Thesis (in preparation). 2020.

[AR91] Pierre Arnoux and Gérard Rauzy. Représentation géométrique de suites de complexité 2n+1. *Bulletin de la Société Mathématique de France*, 119:199–215, 1991.

[Arn20] Pierre Arnoux. Private communication. 2020.

[AS13] Pierre Arnoux and Štěpán Starosta. The Rauzy Gasket. In *Further Developments in Fractals and Related Fields*, pages 1–23. Springer, 2013.

[BB12] Nicolas Bédaride and Jean-François Bertazzon. Minoration of the complexity function associated to a translation on the torus. *Monatshefte für Mathematik*, 171, 2012.

[BD14] Valérie Berthé and Vincent Delecroix. Beyond substitutive dynamical systems: S-adic expansions. In *Lecture note 'Kokyuroku Bessatu'*, pages 81–123, 2014.

[Ber11] Valérie Berthé. Multidimensional Euclidean algorithms, numeration and substitutions. *Integers [electronic only]*, 2011.

[BFD$^+$15a] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Acyclic, connected and tree sets. *Monatshefte für Mathematik*, 176:521–550, 2015.

[BFD$^+$15b] Valérie Berthé, Clelia De Felice, Francesco Dolce, Julien Leroy, Dominique Perrin, Christophe Reutenauer, and Giuseppina Rindone. Maximal bifix decoding. *Discrete Mathematics*, 338:725–742, 2015.

[BJS12] Valérie Berthé, Timo Jolivet, and Anne Siegel. Substitutive Arnoux-Rauzy sequences have pure discrete spectrum. *Uniform Distribution Theory*, 7(1):173–197, 2012.

[Bos84] Michael Boshernitzan. A unique ergodicity of minimal symbolic flows with linear block growth. *J. Analyse Math.*, 44:77–96, 1984.

[BPS08] L'Ubomíra Balková, Edita Pelantová, and Wolfgang Steiner. Sequences with constant number of return words. *Monatshefte für Mathematik*, 155, 2008.

[BST19] Valérie Berthé, Wolfgang Steiner, and Jörg M. Thuswaldner. Geometry, dynamics, and arithmetic of S-adic shifts. *Annales de l'Institut Fourier*, 69:1347–1409, 2019.

[BST20] Valérie Berthé, Wolfgang Steiner, and Jörg M. Thuswaldner. Multidimensional continued fractions and symbolic codings of toral translations. *preprint arXiv:2005.13038v1*, 2020.

[BŠW13] Marcy Barge, Sonja Štimac, and R. F. Williams. Pure discrete spectrum in substitution tiling spaces. *Discrete and Continuous Dynamical Systems*, 33(2):579–597, 2013.

[CFM08] Julien Cassaigne, Sébastien Ferenczi, and Ali Messaoudi. Weak mixing and eigenvalues for Arnoux-Rauzy sequences. *Annales de l'Institut Fourier*, 58(6), 2008.

[CFZ00] Julien Cassaigne, Sébastien Ferenczi, and Luca Q. Zamboni. Imbalances in Arnoux-Rauzy sequences. *Annales de l'Institut Fourier*, 50:1265–1276, 2000.

[CH73] Ethan M. Coven and Gustav A. Hedlund. Sequences with minimal block growth. *Mathematical systems theory*, 7:138–153, 1973.

[CLL17] Julien Cassaigne, Sébastien Labbé, and Julien Leroy. A set of sequences of complexity 2n+1. In *WORDS 2017 Proceedings*, pages 144–156. Springer, 2017.

[Dur98] Fabien Durand. A characterization of substitutive sequences using return words. *Discrete Mathematics*, 179(1):89 – 101, 1998.

[Fer92] Sébastien Ferenczi. Bounded remainder sets. *Acta Arithmetica*, 61, 01 1992.

[Fog02] N. Pytheas Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer-Verlag, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.

[Fog20] N. Pytheas Fogg. Symbolic coding of linear complexity for generic translations of the torus, using continued fractions. *preprint arXiv:2005.12229v1*, 2020.

[JV00] Jacques Justin and Laurent Vuillon. Return words in Sturmian and episturmian words. *Theoretical Informatics and Applications*, 34:343–356, 2000.

[LM95] Douglas Lind and Brian Marcus. *An introduction to symbolic dynamics and coding.* Cambridge University Press, 1995.

[MH38] Marston Morse and Gustav A. Hedlund. Symbolic dynamics. *American Journal of Mathematics*, 60(4):815–866, 1938.

[MH40] Marston Morse and Gustav A. Hedlund. Symbolic dynamics ii. Sturmian trajectories. *American Journal of Mathematics*, 62(1):1–42, 1940.

[Rau82] Gérard Rauzy. Nombres algébriques et substitutions. *Bulletin de la Société Mathématique de France*, 110:147–178, 1982.

[Rau84] Gérard Rauzy. Ensembles à restes bornés. *Séminaire de Théorie des Nombres de Bordeaux*, pages 1–12, 1984.

[RSZ09] Gwénaël Richomme, Kalle Saari, and Luca Zamboni. Balance and abelian complexity of the Tribonacci word. *Advances in Applied Mathematics*, 45(2):212–231, 2009.

[Sch00] Fritz Schweiger. *Multidimensional Continued Fractions.* Oxford Science Publications. Oxford University Press, 2000.

[Thu19] Jörg M. Thuswaldner. S-adic sequences: a bridge between dynamics, arithmetic, and geometry. *arXiv:1908.05954*, 2019.

[Vui01] Laurent Vuillon. A characterization of Sturmian words by return words. *European Journal of Combinatorics*, 22:363–375, 2001.

IV. A Rauzy fractal unbounded in all directions of the plane

The first section of this chapter is published in Comptes Rendus de l'Académie des Sciences.

Contents

A Rauzy fractal unbounded in all directions of the plane

Mélodie Andrieu

Abstract

We construct an Arnoux-Rauzy word for which the set of all differences of two abelianized factors is equal to $\mathbb{Z}^3$. In particular, the imbalance of this word is infinite - and its Rauzy fractal is unbounded in all directions of the plane.

Résumé

Nous construisons explicitement un mot d'Arnoux-Rauzy pour lequel l'ensemble des différences possibles des facteurs abélianisés est égal à $\mathbb{Z}^3$. En particulier, le déséquilibre de ce mot est infini, et son fractal de Rauzy n'est borné dans aucune direction du plan.

1 Introduction

À l'algorithme de fraction continue soustractif décrit par l'itération de l'application (dite de Farey)

$$\begin{array}{rcll} (\mathbb{R}^+)^2 & \to & (\mathbb{R}^+)^2 & \\ (x,y) & \mapsto & (x-y,y) & \text{si } x \geq y, \\ & & (x,y-x) & \text{sinon,} \end{array}$$

est associée une classe particulière de mots infinis binaires appelés mots sturmiens. Rappelons qu'un *mot* est une suite finie ou infinie d'éléments (*lettres*) pris dans un ensemble fini (*alphabet*). Les mots sturmiens jouissent de nombreuses caractérisations combinatoires, arithmétiques et géométriques (consulter [Lot97] pour une introduction générale). En particulier, ce sont exactement les mots apériodiques binaires dont le déséquilibre vaut 1, c'est-à-dire dans lesquels tous les facteurs de même longueur (un *facteur* de longueur n est un sous-mot constitué de n lettres consécutives) contiennent, à 1 près, le même nombre de 0 (et donc, à 1 près également, le même nombre de 1). Par exemple, un mot commençant par $w = 0010001001000100010001001\dots$ pourrait être sturmien, tandis qu'un mot commençant par $w = 011011100...$ ne l'est pas, car il contient les facteurs 11 et 00. Cette propriété garantit en particulier que les lettres 0 et 1 sont uniformément distribuées par rapport à une mesure de probabilité ν sur $\{0,1\}$, et que l'écart entre la somme de Birkhoff $1/N\sum_{n=0}^{N-1}\mathbb{1}_{\{0\}}(w[n])$, qui mesure la fréquence de 0 observée parmi les N premières lettres du mot w, et sa valeur attendue $\nu(0)$ (appelée fréquence de 0) est majoré par $1/N$. D'un point de vue géométrique, cela signifie que les points $P_N := \sum_{n=0}^{N} e_{w[n]}$, où (e_0, e_1) désigne la base canonique de $\mathbb{R}^2$, restent à une distance bornée de la droite portée par le vecteur fréquence $(\nu(0), \nu(1))$. On appelle *ligne brisée* associée à w la suite $(P_N)_{N\in\mathbb{N}}$. En informatique, les lignes brisées associées aux mots sturmiens sont utilisées pour discrétiser les droites de pente irrationnelles.

Depuis Jacobi, plusieurs algorithmes ont été proposés pour généraliser les fractions continues à des triplets de réels positifs (on peut consulter à ce sujet le livre [Sch00]). De tels algorithmes devraient permettre d'approcher simultanément et efficacement deux réels par une suite de couples de nombres rationnels.

Dans ce document, nous nous intéressons aux mots d'Arnoux-Rauzy, introduits par Arnoux et Rauzy dans [AR91], qui sont les mots ternaires associés à l'algorithme (défini sur un ensemble de mesure nulle) :

$$\begin{array}{llcll} F_{AR}: & (\mathbb{R}^+)^3 & \to & (\mathbb{R}^+)^3 & \\ & (x,y,z) & \mapsto & (x-y-z,y,z) & \text{si } x \geq y+z, \\ & & & (x,y-x-z,z) & \text{si } y \geq x+z, \\ & & & (x,y,z-x-y) & \text{si } z \geq x+y. \end{array}$$

Parce qu'ils conservent de nombreuses propriétés combinatoires des mots sturmiens, les mots d'Arnoux-Rauzy sont souvent présentés comme leur généralisation. En particulier, on peut montrer qu'ils admettent un vecteur fréquence des lettres. Aussi, une façon d'étudier la ligne brisée (tridimensionnelle) associée à un mot d'Arnoux-Rauzy consiste à la projeter, parallèlement au vecteur fréquence, sur le plan diagonal $\Delta_0 : x+y+z=0$. On appelle *fractal de Rauzy de w* l'adhérence de cet ensemble de points.

Jusqu'en 2000, on a pensé que, comme pour les mots sturmiens, le déséquilibre des mots d'Arnoux-Rauzy était borné, ou au moins fini. Cassaigne, Ferenczi et Zamboni [CFZ00] ont contredit cette conjecture en construisant un mot d'Arnoux-Rauzy de déséquilibre infini - un mot donc, dont la ligne brisée s'écarte régulièrement et de plus en plus loin de sa direction moyenne, ou, dit encore autrement, un mot dont le fractal de Rauzy n'est pas borné.

Aujourd'hui, on ne sait presque rien sur les propriétés géométriques et topologiques de ces fractals de Rauzy déséquilibrés. Le théorème d'Oseledets [Ose68] suggère toutefois que ces fractals sont contenus dans une bande du plan ; en effet, si les exposants de Lyapounov associés au produit de matrices donné par l'algorithme existent, l'un de ces exposants au moins doit être négatif puisque leur somme est nulle.

Dans cette note, nous prouvons que cette intuition est fausse.

Théorème 1. *Il existe un mot d'Arnoux-Rauzy dont le fractal de Rauzy n'est borné dans aucune direction du plan.*

La construction que nous présentons s'adapte sans effort à la classe des mots associés à l'algorithme de fraction continue multidimensionnelle de Cassaigne-Selmer, introduite dans [CLL17].

Par ailleurs, nous proposons une preuve élémentaire du :

Théorème 2. *Le vecteur fréquence des lettres d'un mot d'Arnoux-Rauzy a des coordonnées rationnellement indépendantes.*

Ce résultat, conjecturé par Arnoux et Starosta en 2013 [AS13], a été démontré très récemment par des moyens plus sophistiqués par Dynnikov, Hubert et Skripchenko [DHS].

Enfin, les résultats que nous présentons s'étendent aux dimensions supérieures : (1) pour tout entier $d \geq 3$, il existe un mot *épisturmien strict* sur l'alphabet $\{1, ..., d\}$ (que l'on appelle aussi mot d'Arnoux-Rauzy généralisé) dont la ligne brisée s'éloigne arbitrairement loin de tout hyperplan de $\mathbb{R}^d$; (2) pour tout entier $d \geq 1$, le vecteur fréquence d'un mot épisturmien strict sur l'alphabet $\{1, ..., d\}$ a des coordonnées rationnellement indépendantes.

2 Introduction (short English version)

Until 2000, it was believed that, as for Sturmian words, the imbalance of Arnoux-Rauzy words was bounded - or at least finite. Cassaigne, Ferenczi and Zamboni disproved this conjecture by constructing an Arnoux-Rauzy word with infinite imbalance, i.e. a word whose broken line deviates regularly and further and further from its average direction [CFZ00]. Today, we know virtually nothing about the geometrical and topological properties of these unbalanced Rauzy fractals. The Oseledets theorem suggests that these fractals are contained in a strip of the plane: indeed, if the Lyapunov exponents of the matrix product associated with the word exist, one of these exponents at least is nonpositive since their sum equals zero. This article aims at disproving this belief.

Theorem 1. *There exists an Arnoux-Rauzy word whose Rauzy fractal is unbounded in all directions of the plane.*

Theorem 1 holds also for C-adic words, which are the infinite words associated with the Cassaigne-Selmer multidimensional continued fraction algorithm introduced in [CLL17]. It can also be adapted to generalized Arnoux-Rauzy words: for $d \geq 3$, we can construct a strict episturmian word over a d-letter alphabet whose broken line deviates regularly and further and further from any hyperplane of $\mathbb{R}^d$.

Besides, we propose an elementary proof of:

Theorem 2. *The vector of letter frequencies of an Arnoux-Rauzy word has rationally independent entries.*

This theorem completes the works of Arnoux and Starosta, who conjectured it in 2013, to prove that the Arnoux-Rauzy continued fraction algorithm detects all kind of rational dependencies [AS13]. Note that it has been recently proved by Dynnikov, Hubert and Skripchenko using quadratic forms [DHS].

3 Preliminaries

We denote by $\mathfrak{A}^*$ the set of all finite words over an alphabet $\mathfrak{A}$. A finite word $u = u[0]u[1]...u[n-1]$, where $u[k]$ denotes the $(k+1)$-th letter of u, is a *factor of length* n of a (finite or infinite) word w if there exists a nonnegative integer i such that for all $k \in \{0, ..., n-1\}$, $w[i+k] = u[k]$; in the particular case $i = 0$, we say that u is the *prefix of length* n of w, and denote it by $u = p_n(w)$. We denote by $\mathcal{F}_n(w)$ the set of factors of w of length n and by $\mathcal{F}(w)$ its set of factors of all lengths.

A *substitution* is an application mapping letters to finite words: $\mathfrak{A} \mapsto \mathfrak{A}^*$, that we extend into a morphism on the free monoid for the concatenation operation $\mathfrak{A}^*$ on one hand, and on the set of infinite words $\mathfrak{A}^{\mathbb{N}}$ on the other hand. Three substitutions will be of high interest in this paper: σ_1, σ_2 and σ_3 defined over $A = \{1, 2, 3\}$ by:

$$\begin{aligned} \sigma_i :\ & A \to A^* \\ & i \mapsto i \\ & j \mapsto ij \text{ for } j \in A \backslash \{i\}. \end{aligned}$$

They are called *Arnoux-Rauzy substitutions*; we denote $AR = \{\sigma_1, \sigma_2, \sigma_3\}$. The set AR can be seen as a three letter alphabet -it should not be confused with $A = \{1, 2, 3\}$ over which the substitutions are defined. As much as we can, we refer to the elements of AR^* or $AR^{\mathbb{N}}$ as "sequences" instead of "words"; nonetheless, some tools like the notions of factor and prefix will turn out to be useful for this second alphabet as well, especially in Section 4.

The set $\mathfrak{A}^{\mathbb{N}}$ of infinite words over $\mathfrak{A}$ is endowed with the distance δ: for all $w, w' \in \mathfrak{A}^{\mathbb{N}}$, $\delta(w,w') = 2^{-n_0}$, where $n_0 = \min\{n \in \mathbb{N} | w[n] \neq w'[n]\}$ if $w \neq w'$, and $\delta(w,w') = 0$ otherwise. We say that a sequence of finite words $(u_n)_{n\in\mathbb{N}} \in (\mathfrak{A}^*)^{\mathbb{N}}$ *converges* to an infinite word $w \in \mathfrak{A}^{\mathbb{N}}$ if for any sequence of infinite words $(v_n)_{n\in\mathbb{N}} \in (\mathfrak{A}^{\mathbb{N}})^{\mathbb{N}}$, the sequence of infinite words $(u_n \cdot v_n)_{n\in\mathbb{N}} \in (\mathfrak{A}^{\mathbb{N}})^{\mathbb{N}}$ converges to w.

If $(s_n)_{n\in\mathbb{N}} \in AR^{\mathbb{N}}$ is a sequence containing infinitely many occurrences of each Arnoux-Rauzy substitution σ_1, σ_2 and σ_3, then the sequence of finite words $(s_0 \circ ... \circ s_{n-1}(\alpha))$, with $\alpha \in A$, converges to an infinite word w_0 which does not depend on α. The infinite words w_0 obtained this way are called *standard Arnoux-Rauzy words.* An infinite word w is an *Arnoux-Rauzy word* if it has the same set of factors than a standard Arnoux-Rauzy word w_0. One can show that the standard Arnoux-Rauzy word w_0 and the *directive sequence* $(s_n)_{n\in\mathbb{N}}$ associated with w are unique. This definition of Arnoux-Rauzy words is equivalent to the more usual one: an infinite word is an *Arnoux-Rauzy word* if it has complexity $2n+1$ and admits exactly one right and one left special factor of each length.

Given a finite word $u \in \mathfrak{A}^*$ and a letter $\alpha \in \mathfrak{A}$, we denote by $|u|_\alpha$ the number of occurrences of α in u. The *abelianized vector* of u, sometimes called *Parikh vector* of u, is the vector $\mathrm{ab}(u) = (|u|_\alpha)_{\alpha\in\mathfrak{A}}$, which counts the number of times that each letter occurs in the finite word u. At this point, it is useful to order the alphabet. For the convenience of typing, we choose to represent abelianized words as line vectors. Observe that the sum of the entries of $\mathrm{ab}(u)$ is equal to the *length* of the word u, that we denote by $|u|$. Now, given a substitution $s : \mathfrak{A} \to \mathfrak{A}^*$, the *incidence matrix* of s is the matrix M_s whose the $i-th$ line is the abelianized of the image by s of the $i-th$ letter in the alphabet. For instance, the incidence matrices of the Arnoux-Rauzy substitutions are:

$$M_{\sigma_1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad M_{\sigma_2} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad M_{\sigma_3} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \in \mathrm{GL}_3(\mathbb{Z}).$$

Abelianized words and incidence matrices are made to satisfy: $\mathrm{ab}(s(u)) = \mathrm{ab}(u)M_s$ for any substitution $s : \mathfrak{A} \to \mathfrak{A}^*$ and any finite word $u \in \mathfrak{A}^*$.

If $w \in \mathfrak{A}^{\mathbb{N}}$ is an infinite word and $\alpha \in \mathfrak{A}$ is a letter, the frequency of α in w is the limit, if it exists, of the proportion of α in the sequence of growing prefixes of w: $f_w(\alpha) = \lim_{n\to\infty} \frac{|p_n(w)|_\alpha}{n}$. We denote by $f_w = (f_w(\alpha))_{\alpha\in\mathfrak{A}}$ the *vector of letter frequencies* of w, if it exists. When the vector of letter frequencies exists, as it is the case for any Arnoux-Rauzy word, it is natural to study the difference between the predicted frequencies of letters and their observed occurrences. Given an infinite word $w \in \mathfrak{A}^{\mathbb{N}}$ for which the vector of letter frequencies is defined, we consider the *discrepancy function*:

$$\begin{aligned} &\mathbb{N} \to \mathbb{R} \\ &n \mapsto \max_{\alpha\in\mathfrak{A}} \big| \, |p_n(w)|_\alpha - n f_w(\alpha) \big|. \end{aligned}$$

The discrepancy is linked to a combinatorial property: the imbalance. The imbalance of an infinite word w is the quantity (possibly infinite) :

$$\mathrm{imb}(w) \; = \; \sup_{n\in\mathbb{N}} \; \sup_{u,v\in\mathcal{F}_n(w)} \; ||\mathrm{ab}(u) - \mathrm{ab}(v)||_\infty.$$

The imbalance of an infinite word w is finite if and only if its discrepancy function is bounded. Geometrically, the discrepancy is linked to the diameter of the Rauzy fractal.

Let Δ_0 denotes the plane of $\mathbb{R}^3$ with equation $x+y+z=0$. For w an Arnoux-Rauzy word, denote by f_w its letter frequencies vector and by π_w the (oblique) projection onto Δ_0 associated with the direct sum: $\mathbb{R}f_w \oplus \Delta_0 = \mathbb{R}^3$. The *Rauzy fractal* of w, denoted by $\mathcal{R}_w$, is the closure of the image of the set of abelianized prefixes of w (the *broken line of* w) by the projection π_w: $\mathcal{R}_w = \overline{\cup_{k\in\mathbb{N}}\{\pi_w(\mathrm{ab}(p_k(w)))\}} \subset \Delta_0$. Note that the statement of our main result (Theorem 1) does not depend on the choice of the plane we project onto.

4 Results

Lemma 1. *For any $(a,b,c) \in \mathbb{Z}^3$, there exists $s \in AR^*$ and there exist $u, v \in \mathcal{F}(s(1))$ that satisfy* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a,b,c)$.

Remark 1 (Abuse of notation)**.** *If $s = s_0 \cdot ... \cdot s_{n-1} \in AR^*$, and if $w \in A^* \cup A^{\mathbb{N}}$, then $s(w)$ denotes the image of the word w by the substitution $s_0 \circ ... \circ s_{n-1}$.*

Proof. Section 5 is devoted to the proof of Lemma 1. □

Therefore, all standard Arnoux-Rauzy words -and thereby all Arnoux-Rauzy words- whose directive sequence starts with the prefix s will admit (a,b,c) as difference of abelianized factors.

Lemma 2. *For any $p \in AR^*$ and any $(a,b,c) \in \mathbb{Z}^3$, there exists $s \in AR^*$ and there exist $u, v \in \mathcal{F}(p \cdot s(1))$ that satisfy* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a,b,c)$.

Proof. Let $p \in AR^*$ and $(a,b,c) \in \mathbb{Z}^3$. Denote by M_p the incidence matrix of the substitution associated with p (following Remark 1), which is a product of the Arnoux-Rauzy matrices $M_{\sigma_1}, M_{\sigma_2}$ and M_{σ_3}, and thus belongs to $\mathrm{GL}_3(\mathbb{Z})$. By Lemma 1, there exists $s \in AR^*$ and there exist u and $v \in \mathcal{F}(s(1))$ such that $\mathrm{ab}(u) - \mathrm{ab}(v) = (a,b,c)\, M_p^{-1}$. But then, $p(u)$ and $p(v)$ are factors of $\mathcal{F}(p \cdot s(1))$ and satisfy $\mathrm{ab}(p(u)) - \mathrm{ab}(p(v)) = (\mathrm{ab}(u) - \mathrm{ab}(v))M_p = (a,b,c)$. □

We now construct a standard Arnoux-Rauzy word for which all triplets of integers can be obtained as a difference of two of its abelianized factors.

Proposition 1. *There exists an Arnoux-Rauzy word w_∞ such that for all $(a,b,c) \in \mathbb{Z}^3$, there exist u and $v \in \mathcal{F}(w_\infty)$ satisfying* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a,b,c)$.

Proof. Let $\varphi : \mathbb{N} \to \mathbb{Z}^3$ a bijection (that can be chosen explicitly). We construct an infinite word $d \in AR^{\mathbb{N}}$ as the limit of the sequence of finite words $(p_k)_{k\in\mathbb{N}} \in (AR^*)^{\mathbb{N}}$ that we define by recurrence as follows. We first set p_0 as the prefix given by Lemma 1 for $(a,b,c) = \varphi(0)$. Now, for $k \in \mathbb{N}$, we set $p_{k+1} = p_k.\sigma_1.\sigma_2.\sigma_3.s$, where $s \in AR^*$ is given by applying Lemma 2 to the word $p_k.\sigma_1.\sigma_2.\sigma_3 \in AR^*$ and the vector $\varphi(k+1) \in \mathbb{Z}^3$. By construction, the sequence of finite words $(p_k)_{k\in\mathbb{N}}$ converges to an infinite sequence d which contains infinitely many occurrences of σ_1, σ_2 and σ_3. This guarantees that the sequence of finite words $(d_0 \circ ... \circ d_{n-1}(1))_{n\in\mathbb{N}}$ converges to an Arnoux-Rauzy word, that we denote by w_∞. Finally, for any $k \in \mathbb{N}$, since the directive sequence of w_∞ starts with the prefix p_k, there exist $u_k, v_k \in \mathcal{F}(w_\infty)$ such that $\mathrm{ab}(u_k) - \mathrm{ab}(v_k) = \varphi(k)$. □

Corollary 1. *The imbalance of the word w_∞ is infinite.*

Proof. For any $n \in \mathbb{N}$, there exist u_n and $v_n \in \mathcal{F}(w_\infty)$ such that $\mathrm{ab}(u_n) - \mathrm{ab}(v_n) = (n, 0, -n)$; this implies both $|u_n| = |v_n|$ and $|u_n|_1 - |v_n|_1 = n$. The imbalance of w_∞ is thus infinite. □

The imbalance of a word, which is a combinatorial quantity, is linked to the geometrical shape of its associated broken line. More precisely: a word w admitting frequencies has an infinite imbalance if and only if its Rauzy fractal is unbounded. We now propose to show that the word w_∞ actually satisfies a stronger property: its Rauzy fractal is unbounded in *all directions* of the plane. This relies on the following proposition.

Proposition 2. *Let $w \in A^{\mathbb{N}}$. If for all $\boldsymbol{d} \in \mathbb{Z}^3 \cap \Delta_0$, where Δ_0 denotes the plane of $\mathbb{R}^3$ with equation $x+y+z=0$, there exist u and $v \in \mathcal{F}(w)$ such that $\mathrm{ab}(u) - \mathrm{ab}(v) = \boldsymbol{d}$, then, for any plane Π and for any $D \in \mathbb{R}^+$, there exists $k \in \mathbb{N}$ such that the euclidean distance between the point $\mathrm{ab}(p_k(w))$ and the plane Π is larger than D.*

Proof. Without loss of generality, we can assume that Π contains $(0,0,0)$.
If $\Pi = \Delta_0$, then for any $D \in \mathbb{R}^+$, $\mathrm{dist}(\mathrm{ab}(p_k(w)), \Pi) \geq D$, with $k = \lceil D\sqrt{3}/3 \rceil$.
Let $\Pi \neq \Delta_0$. By contradiction, assume that there exists $D \in \mathbb{R}^+$ such that for all nonnegative integer k, $\mathrm{dist}(\mathrm{ab}(p_k(w)), \Pi) < D$. Let $\mathbf{d} \in \mathbb{Z}^3 \cap \Delta_0$ with $\mathrm{dist}(\mathbf{d}, \Pi) \geq 4D$, and factors $u, v \in \mathcal{F}(w)$ such that $\mathrm{ab}(u) - \mathrm{ab}(v) = \mathbf{d}$. Then, without loss of generality, we have $\mathrm{dist}(ab(u), \Pi) \geq 2D$. Let $t \in A^*$ be such that tu is a prefix of w. Then we have $\mathrm{dist}(\mathrm{ab}(t), \Pi) \geq D$ or $\mathrm{dist}(\mathrm{ab}(tu), \Pi) \geq D$, a contradiction. □

Remark 2. *Proposition 2 and its proof remain valid by replacing Δ_0 by any other plane whose intersection with $\mathbb{Z}^3$ is not trapped between two parallel lines.*

Theorem 1. *There exists an Arnoux-Rauzy word whose Rauzy fractal is unbounded in all directions of the plane.*

Proof. We obtain, by applying Proposition 2 to the word w_∞ described in Proposition 1 and to planes spanned by f_w and a vector of Δ_0, that the Rauzy fractal associated with w_∞ cannot be trapped between two parallels lines. □

5 Proof of Lemma 1

We consider the infinite oriented graph whose vertices are the elements of $\mathbb{Z}^3$ and whose edges map triplets to their images by one the 15 following applications. For $\delta \in \{-2,-1,0,1,2\}$ and $i \in \{1,2,3\}$, consider:

$$\begin{array}{llll} \tau_{i,\delta}: & \mathbb{Z}^3 & \to & \mathbb{Z}^3 \\ & (x_j)_{j\in\{1,2,3\}} & \mapsto & (y_j)_{j\in\{1,2,3\}} \quad \text{where } y_i = x_1+x_2+x_3+\delta \text{ and } y_j = x_j \text{ for } j \neq i. \end{array}$$

Our aim is to show that all vertices can be reached from the triplet $O = (0,0,0) \in \mathbb{Z}^3$, moving through a finite number of edges (see Definition 1 and Proposition 3 below). The motivation lies in the following lemma.

Lemma 3. *Let $d \in \mathbb{Z}^3$. If there exist $n \in \mathbb{N}$ and a finite sequence $(i_l, \delta_l)_{0 \leq l \leq n-1} \in (\{1,2,3\} \times \{-2,-1,0,1,2\})^n$ such that $d = \tau_{i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{i_0,\delta_0}(O)$, then there exist $s_1 \in AR^*$ and $u, v \in \mathcal{F}(\sigma_{i_{n-1}} \circ ... \circ \sigma_{i_0}(s_1(1)))$ satisfying $\mathrm{ab}(u) - \mathrm{ab}(v) = d$.*

Proof. Let $d \in \mathbb{Z}^3$. Assume that there exist $n \in \mathbb{N}$ and $(i_l, \delta_l)_{0 \leq l \leq n-1} \in (\{1,2,3\} \times \{-2,-1,0,1,2\})^n$ such that $d = \tau_{i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{i_0,\delta_0}(O)$. We are going to build iteratively two finite sequences of finite words (u_l) and (v_l), where $l \in \{0, ..., n\}$, and $s_1 \in AR^*$, such that for all l, the words u_l and v_l are factors of $\sigma_{i_{l-1}} \circ ... \circ \sigma_{i_0}(s_1(1))$, and such that $\mathrm{ab}(u_n) - \mathrm{ab}(v_n) = d$.

First, we choose $s_1 \in AR^*$ that satisfies $|s_1(1)| \geq 2n$, and we set $u_0 = v_0 = p_n(s_1(1))$ (prefix of length n of $s_1(1)$). Then, assuming that u_l and $v_l \in \mathcal{F}(\sigma_{i_{l-1}} \circ ... \circ \sigma_{i_0}(s_1(1)))$ are built, we set $\tilde{u}_{l+1} = \sigma_{i_l}(u_l)$ and $\tilde{v}_{l+1} = \sigma_{i_l}(v_l)$. From $\tilde{u}_{l+1}$ and $\tilde{v}_{l+1}$, we define u_{l+1} and v_{l+1} according to the following table.

δ	choice for u_{l+1}	choice for v_{l+1}
0	$\tilde{u}_{l+1}$	$\tilde{v}_{l+1}$
1	$\tilde{u}_{l+1}.i_l$	$\tilde{v}_{l+1}$
2	$\tilde{u}_{l+1}.i_l$	v_{l+1} such that $i_l.v_{l+1} = \tilde{v}_{l+1}(*)$
-1	$\tilde{u}_{l+1}$	$\tilde{v}_{l+1}.i_l$
-2	u_{l+1} such that $i_l.u_{l+1} = \tilde{u}_{l+1}(*)$	$\tilde{v}_{l+1}.i_l$

We now justify that the steps marked with $(*)$ (removal of the initial i_l) are well-defined, and that u_{l+1} and v_{l+1} are, in all cases, factors of $\sigma_{i_l} \circ ... \circ \sigma_{i_0}(s_1(1))$.

Observe that for any step $l \in \{0, ..., n-1\}$, we remove at most one letter from the left and add at most one letter to the right of $\tilde{u}_{l+1}$ (resp. $\tilde{v}_{l+1}$). The Arnoux-Rauzy substitutions being nonerasing, we recursively check that (these properties hold symmetrically for v_l) :

- the length of u_l and its image $\tilde{u}_{l+1}$ is at least $n - l$; so we can always perform step $(*)$;

- there is an occurrence of u_l which is followed by at least $n - l$ letters in $\sigma_{i_{l-1}} \circ ... \circ \sigma_{i_0}(s_1(1))$; so its image $\tilde{u}_{l+1}$ has also an occurrence in $\sigma_{i_l} \circ ... \circ \sigma_{i_0}(s_1(1))$ which is followed by at least $n - l$ letters, and whose first following letter is i_l.

Finally, in all cases, the words u_{l+1} and v_{l+1} are factors of $\sigma_{i_l} \circ ... \circ \sigma_{i_0}(s_1(1))$ and satisfy $\text{ab}(u_{l+1}) - \text{ab}(v_{l+1}) = \tau_{i_l,\delta_l}(\text{ab}(u_l) - \text{ab}(v_l))$. In particular, at step $l = n-1$, the finite words u_n and v_n are factors of $\sigma_{i_{n-1}} \circ ... \circ \sigma_{i_0}(s_1(1))$ and satisfy $\text{ab}(u_n) - \text{ab}(v_n) = \tau_{i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{i_0,\delta_0}(O) = d$. □

In the sequel, it is convenient to introduce some vocabulary from graph theory.

Definition 1. *A triplet $(a, b, c) \in \mathbb{Z}^3$ is* accessible *from a triplet (d, e, f) if there exist a nonnegative integer n and a finite sequence $(i_l, \delta_l)_{0 \leq l \leq n-1} \in (\{1, 2, 3\} \times \{-2, -1, 0, 1, 2\})^n$ such that $(a, b, c) = \tau_{i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{i_0,\delta_0}((d, e, f))$.*

Proposition 3. *All triplets in $\mathbb{Z}^3$ are accessible from O.*

The proof of Proposition 3 lies on the three following lemmas.

Lemma 4. *The triplet $(a, b, c) \in \mathbb{Z}^3$ is accessible from O if and only if $(-a, -b, -c)$ is also accessible from O. Similarly, $(x_j)_{j\in\{1,2,3\}}$ is accessible from O if and only if for all $s \in \mathfrak{S}_3$, where $\mathfrak{S}_3$ denotes the symmetric group acting on three elements, the triplet $(x_{s(j)})_{j\in\{1,2,3\}}$ is accessible from O.*

Proof. For the first assertion, change δ_l into $-\delta_l$ in the finite sequence of edges going from O to (a, b, c). For the second assertion, change i_l into $s(i_l)$ in the finite sequence of edges going from O to $(x_j)_{j\in\{1,2,3\}}$. □

Lemma 5. *Let $a \in \mathbb{N}$. The triplet $(a, -a, -a) \in \mathbb{Z}^3$ is accessible from O.*

Proof. The lemma is trivially true for $a = 0$. By recurrence, consider an arbitrary nonnegative integer a such that the triplet $(a, -a, -a)$ is accessible from O. One can check that $(a + 1, -a - 1, a + 1) = \tau_{1,1} \circ (\tau_{3,2})^{2a+1} \circ \tau_{2,-1}((a, -a, -a))$. So the triplet $(a + 1, -a - 1, a + 1)$ is accessible from O. But then, Lemma 4 indicates that $(a + 1, -a - 1, -a - 1)$ is accessible from O. □

Proof of Proposition 3. Let $(a, b, c) \in \mathbb{Z}^3$. Without loss of generality (Lemma 4), we assume that $a \geq b \geq c \geq -a$. One easily check that $(a, b, -a) = (\tau_{2,1})^{b+a}((a, -a, -a))$; we deduce from Lemma 5 that the vertex $(a, b, -a)$ is accessible from O. We now prove that (a, b, c) is also accessible from O if $c > -a$.

Let E denote the subtractive Euclid map:

$$\begin{array}{llll} E: & \mathbb{N}^2 & \longrightarrow & \mathbb{N}^2 & \\ & (x,y) & \longmapsto & (x-y,y) & \text{if } x \geq y \\ & & & (x,y-x) & \text{otherwise.} \end{array}$$

Let $(x_0, y_0) = (a+b, a+c) \in (\mathbb{N}^*)^2$ and $(x_k, y_k) = E^k(x_0, y_0)$ for any nonnegative integer k. The sequence $(x_k, y_k)_{k\in\mathbb{N}}$ is ultimately constant; denote by $n = \min\{k \in \mathbb{N} | (x_k, y_k) = (0, \gcd(x_0, y_0)\}$ the number of iterations required by the Euclid algorithm on the input (x_0, y_0). For all k in $\{0, ..., n\}$, let $(b_k, c_k) = (x_k - a, y_k - a)$. Observe that:

(i) $(b_0, c_0) = (b, c)$,
(ii) $(a, b_n, c_n) = (a, -a, -a + \gcd(a+b, a+c))$,
(iii) the triplet (a, b_n, c_n) is accessible from O,
(iv) for all $k \in \{0, .., n-1\}$, the triplet (a, b_k, c_k) is accessible from (a, b_{k+1}, c_{k+1}).

The assertions (i) and (ii) are immediate; (iii) comes from the facts that $(a, -a, -a + \gcd(a + b, a + c) = (\tau_{3,1})^{\gcd(a+b,a+c)}((a, -a, -a))$ and that $(a, -a, -a)$ is accessible from O (Lemma 5.) We now prove (iv). Let k in $\{0, ..., n-1\}$. Since $(x_{k+1}, y_{k+1}) = E(x_k, y_k)$, two cases may happen: $(x_{k+1}, y_{k+1}) = (x_k - y_k, y_k)$ or $(x_{k+1}, y_{k+1}) = (x_k, y_k - x_k)$. In the first case, we have $y_k = y_{k+1}$ and $x_k = x_{k+1} + y_k = x_{k+1} + y_{k+1}$, and then, $(a, b_k, c_k) = (a, x_{k+1} + y_{k+1} - a, y_{k+1} - a) = \tau_{2,0}((a, x_{k+1} - a, y_{k+1} - a)) = \tau_{2,0}((a, b_{k+1}, c_{k+1}))$. In the second case, we have $x_k = x_{k+1}$ and $y_k = x_k + y_{k+1} = x_{k+1} + y_{k+1}$, and then, $(a, b_k, c_k) = (a, x_{k+1} - a, y_{k+1} + x_{k+1} - a) = \tau_{3,0}((a, x_{k+1} - a, y_{k+1} - a)) = \tau_{3,0}((a, b_{k+1}, c_{k+1}))$. In both cases, the triplet (a, b_k, c_k) is accessible from (a, b_{k+1}, c_{k+1}) in one step. Finally, the triplet $(a, b, c) = (a, b_0, c_0)$ is accessible from (a, b_n, c_n), and then, from O □

Proof of Lemma 1. Lemma 1 follows from Proposition 3, Definition 1 and Lemma 3. □

Remark 3. *The graph $\mathcal{G}$ is a simplification, exploiting the remarkable properties of the substitutions σ_1, σ_2 and σ_3, of the imbalance automaton, introduced in [And20] (chapter 2) for a much wider range of S-adic systems (ie class of words obtained from a set of substitutions through* directive sequences*).*

6 The vector of letter frequencies of w_∞ has rationally independent entries

We sketch an elementary proof of the much wider result:

Theorem 2. *The vector of letter frequencies of an Arnoux-Rauzy word has rationally independent entries.*

The proof is inspired from a similar result that holds for C-adic words [CLL17].

Proof. Let w an Arnoux-Rauzy word; denote by $(s_n)_{n\in\mathbb{N}}$ its directive sequence and by f its letter frequencies vector. We recall that for all nonnegative integer n, $s_n = \sigma_i$ if and only if the $i-th$ entry of $F_{AR}^n(f)$ (F_{AR} is defined in Section 1) is greater than the sum of the two others. By contradiction, assume that the entries of f are not rationally independent.

First, observe that if for some $r \in \mathbb{N}$, the $i-th$ entry of $F_{AR}^r(f)$ is zero, then it will remain zero; and from this point on the directive sequence will not contain the substitution σ_i, which is conflicting with the definition of Arnoux-Rauzy words and the uniqueness of the directive sequence. Thus, for all $n \in \mathbb{N}$, all entries of $F_{AR}^n(f)$ are positive. Let l_0 a nonzero integer column vector such that $fl_0 = 0$

(recall that f is a line vector). Let $l_m = M_{s_{m-1}}...M_{s_0}l_0$. The Arnoux-Rauzy matrices being invertible, l_m is also a nonzero integer column vector; it satisfies $F_{AR}^m(f)l_m = f.M_{s_0}^{-1}...M_{s_{m-1}}^{-1}l_m = fl_0 = 0$. Denote $l_m = (a, b, c)^t$ and consider $D_m = \max(|b-a|, |c-b|, |c-a|) \in \mathbb{N}$ the difference between the maximum and the minimum entry of l_m, that we call *spread* of l_m. We claim that the sequence of nonnegative integers $(D_m)_{m\in\mathbb{N}}$ is non-increasing and that it furthermore decreases infinitely often - and here will be the contradiction.

Indeed, the vector l_{m+1} is of the form Ml_m, where M is one the the three Arnoux-Rauzy matrices M_{σ_1}, M_{σ_2} or M_{σ_3}, which give respectively: $l_{m+1} = (a, a+b, a+c)^t$, $l_{m+1} = (a+b, b, c+b)^t$ and $l_{m+1} = (a+c, b+c, c)^t$. One can easily show, observing that the extreme entries of l_m have opposite signs, that in all cases $D_m \geq D_{m+1}$. Similarly, we write $l_{m+2} = M_{s_{m+1}}M_{s_m}l_m$. A quick argument show that as soon as $s_{m+1} \neq s_m$, which happens infinitely many times by definition of Arnoux-Rauzy words, we have $D_m > D_{m+2}$. □

Remark 4. *Theorem 2 actually holds in arbitrary dimension: with a similar proof, one obtain that the vector of letter frequencies of a strict episturmian word has rationally independent entries.*

References

The bibliography of the article is moved to the end of the chapter, and completed with some additional references.

2 Extension to arbitrary dimension (episturmian words)

In this section, we generalize our results to arbitrary dimension d.

2.1 Results

Lemma 6. *Let $d \geq 3$, and $A_d = \{1, ..., d\}$. For any $(a_1, ..., a_d) \in \mathbb{Z}^d$, there exists $s \in (AR_d)^*$ and $u, v \in (A_d)^*$ satisfying* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a_1, ..., a_d)$*, such that for all episturmian words w over A_d, the finite words u and v are factors of $s(w)$.*

Theorem 3. *Let $d \geq 3$, and $A_d = \{1, ..., d\}$. There exists a strict episturmian word over A_d whose broken line is not trapped between two parallel hyperplanes of $\mathbb{R}^d$.*

Theorem 4. *Let $d \geq 2$. The vector of letter frequencies of a d-letter strict episturmian word has rationally independent entries.*

The proofs of Theorems 3 and 4 are simple adaptations of the demonstrations of Theorems 1 and 2 that we wrote in dimension 3. Lemma 6 requires some additional work.

Theorem 4 spearheads the classification of episturmian words.

Corollary 2. *Let $d \geq d' \geq 1$. Let w an episturmian word over A_d. Denote by $f = (f_1, ..., f_d)$ its vector of letter frequencies, and by $(s_n)_{n \in \mathbb{N}}$ one of its directive sequences. The following assertions are equivalent.*

1. *Exactly d' substitutions appear infinitely many times in $(s_n)_{n \in \mathbb{N}}$.*
2. *The dimension of the vectorial space $f_1\mathbb{Q} + ... + f_d\mathbb{Q}$ is d'.*

Furthermore, in the case $d' \geq 2$, the word w is nonperiodic and the assertions (1) *and* (2) *are equivalent to:*

3. *There exists τ a finite composition of substitutions in AR_d, a subset $A' \subset A_d$ containing d' elements and w'_0 a standard strict episturmian word over the alphabet A' such that $w_0 = \tau(w'_0)$, where w_0 denotes the standard word associated with w.*

In other words, the Arnoux-Rauzy d-dimensional continued fraction algorithm detects all kinds of rational dependencies.

2.2 Definitions

For a complete introduction to episturmian words, the reader should refer to [DJP01] and [JP02], or to the survey [GJ09].

Let w a (finite or infinite) word. A factor $u \in \mathcal{F}(w)$ is *left special* (resp *right special*) if there exist two letters $a \neq b$ such that $au, bu \in \mathcal{F}(w)$ (resp. $ua, ub \in \mathcal{F}(w)$). Let $d \in \mathbb{N}^*$. An infinite word over $A_d := \{1, ..., d\}$ is *episturmian* if it admits at most one left special factor of each length, and if for any factor $u = u[0]...u[|u| - 1] \in \mathcal{F}(w)$, the reversed word $u[|u| - 1]...u[0]$ is also a factor of w. An immediate consequence is that an episturmian word admits at most one right special factor fo each length. Sturmian and Arnoux-Rauzy words are episturmian words. An episturmian word is *standard* if all its prefixes are left special.

Proposition 4 (immediate consequence of the Morse-Hedlund theorem (see for instance [Lot97]))**.** *An infinite word over A_d is not ultimately periodic if and only if it admits right special factors of all lengths.*

In particular, a standard episturmian word is not ultimately periodic. Episturmian words are uniformly recurrent. Therefore, an episturmian word is ultimately periodic if and only if it is periodic; when it is not the case, we say that the word is nonperiodic. For each nonperiodic episturmian word w, there exists a standard episturmian word w_0 such that $\mathcal{F}(w_0) = \mathcal{F}(w)$, which is furthermore unique.

We now define directive sequences for episturmian words. Let $d \in \mathbb{N}^*$. We consider, over the alphabet $A_d = \{1, ..., d\}$, the d substitutions:

$$\begin{aligned} \sigma_{d,i} : \quad & A_d \to A_d^* \\ & i \mapsto i \\ & j \mapsto ij \text{ for } j \in A_d \backslash \{i\}. \end{aligned}$$

In the sequel, we (still) call them *Arnoux-Rauzy substitutions*; we denote $AR_d = \{\sigma_{d,1}, ..., \sigma_{d,d}\}$. The set AR_d will be seen as a second d-letter alphabet.

Lemma 7 (Summary of [AR91], [JP02] and [AS13])**.** *Let $d \geq 1$. For any episturmian word w over A_d, there exists a letter $i \in A_d$ and an episturmian word w' such that $w = \sigma_{d,i}(w')$ or $iw = \sigma_{d,i}(w')$.*

Furthermore, if w is not periodic with shorter period of length 2, the pair (i, w') is unique; otherwise, if $w = (ij)^\omega$, with $i \neq j \in A_d$, there are exactly two admissible pairs, which are (i, j^ω) and (j, i^ω).

By iterating this lemma, we construct an infinite word $(i_n)_{n\in\mathbb{N}}$ and a sequence of episturmian words $(w^{(n)})_{n\in\mathbb{N}}$, with also $w^{(-1)} = w$, such that for any $n \in \mathbb{N}$, we have $w^{(n-1)} = \sigma_{d,i_n}(w^{(n)})$. The sequences of substitutions $(\sigma_{d,i_n})_{n\in\mathbb{N}}$ thus obtained are called *directive sequences* of w.

Proposition 5 (Summary of [AR91], [JP02] and [AS13])**.**

- *If w is nonperiodic, the directive sequence $(s_n)_{n\in\mathbb{N}}$ is unique, not ultimately periodic, and for all $\alpha \in A_d$, the sequence of finite words $s_0 \circ ... \circ s_{n-1}(\alpha)$ converges to w_0, the standard word associated with w.*

- *If $w = i^\omega$, then the directive sequence is unique: it is the constant sequence $(\sigma_{d,i})_{n\in\mathbb{N}}$. For all $\alpha \in A_d \backslash \{i\}$, the sequence of finite words $s_0 \circ ... \circ s_{n-1}(\alpha)$ converges to w.*

- *If w is periodic with shorter period of length at least 2, then w admits exactly two directive sequences, $(s_n)_{n\in\mathbb{N}}$ and $(t_n)_{n\in\mathbb{N}}$, which are both ultimately constant. More precisely, there exists $i \neq j \in A_d$ and $n_0 \in \mathbb{N}$ such that for all $n < n_0$, $s_n = t_n$, and then, $s_{n_0} = \sigma_{d,i}$ and $t_{n_0} = \sigma_{d,j}$, and at last, for all $n > n_0$, $s_n = \sigma_{d,j}$ and $t_n = \sigma_{d,i}$. Furthermore, there exist two episturmian words w_i and w_j, which satisfy $\mathcal{F}(w_i) = \mathcal{F}(w_j) = \mathcal{F}(w)$, and such that for all $\alpha \in A_d \backslash \{j\}$ and for all $\beta \in A_d \backslash \{i\}$, the sequence of finite words $(s_0 \circ ... \circ s_{n-1}(\alpha))_{n\in\mathbb{N}}$ and $(t_0 \circ ... \circ t_{n-1}(\beta))_{n\in\mathbb{N}}$ converges to w_i and w_j respectively.*

In particular, the first point guarantees that this general definition of directive sequence is equivalent to the one we wrote, in Section 1, in the restrictive case of Arnoux-Rauzy words. Generally, if $d \geq 2$ and if $(s_n)_{n\in\mathbb{N}} \in (AR_d)^{\mathbb{N}}$ is a sequence containing infinitely many occurrences of each substitution $\sigma_{d,1}, ..., \sigma_{d,d}$, then the sequence of finite words $(s_0 \circ ... \circ s_{n-1}(\alpha))$, with $\alpha \in A_d$, converges to a standard episturmian word w_0 which does not depend on α. The standard episturmian words w_0 obtained this way are called *standard strict episturmian words* over A_d. An infinite word w is a *strict episturmian word* over A_d if it has the same set of factors than a standard strict episturmian word w_0 over A_d [JP02]. Strict episturmian words over A_2 are exactly Sturmian words. Strict episturmian words over A_3 are exactly Arnoux-Rauzy words.

2.3 Proof of Lemma 6

We recall:

Lemma 6. *Let $d \geq 3$, and $A_d = \{1, ..., d\}$. For any $(a_1, ..., a_d) \in \mathbb{Z}^d$, there exists $s \in (AR_d)^*$ and $u, v \in (A_d)^*$ satisfying* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a_1, ..., a_d)$*, such that for all episturmian words w over A_d, the finite words u and v are factors of $s(w)$.*

Let $d \geq 3$. We consider the infinite oriented graph $\mathcal{G}_d$ whose vertices are the elements of $\mathbb{Z}^d$ and whose edges maps triplets to their images by one the $5 \times d$ following applications. For $\delta \in \{-2, -1, 0, 1, 2\}$ and $i \in A_d$, consider:

$$\begin{array}{llll} \tau_{d,i,\delta}: & \mathbb{Z}^d & \to & \mathbb{Z}^d \\ & (x_j)_{j \in A_d} & \mapsto & (y_j)_{j \in A_d} \quad \text{where } y_i = \sum_{k \in A_d} x_k + \delta \text{ and } y_j = x_j \text{ for } j \neq i. \end{array}$$

Our aim is to show that all vertices are accessible from the vertex $O_d := (0, .., 0) \in \mathbb{Z}^d$.

Lemma 8. *Let $d \geq 3$. Let $(a_1, ..., a_d) \in \mathbb{Z}^d$. If there exist $n \in \mathbb{N}$ and a finite sequence $(i_l, \delta_l)_{0 \leq l \leq n-1} \in (A_d \times \{-2, -1, 0, 1, 2\})^n$ such that $(a_1, ..., a_d) = \tau_{d,i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{d,i_0,\delta_0}(O_d)$, then there exist $u, v \in (A_d)^*$ satisfying* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a_1, ..., a_d)$*, such that for all episturmian words w over A_d, the finite words u and v are factors of $s(w)$, where $s = \sigma_{d,i_{n-1}} \circ ... \circ \sigma_{d,i_0}$.*

Proof. The proof is similar to the proof of Lemma 3. □

Lemma 9. *The vertex $(a_1, ..., a_d) \in \mathbb{Z}^d$ is accessible from O_d if and only if $(-a_1, ..., -a_d)$ is also accessible from O_d. Similarly, $(a_j)_{j \in \{1,...,d\}}$ is accessible from O_d if and only if for all $s \in \mathfrak{S}_d$, where $\mathfrak{S}_d$ denotes the symmetric group acting on d elements, the vertex $(a_{s(j)})_{j \in \{1,...,d\}}$ is accessible from O_d.*

Proof. For the first assertion, change δ_l into $-\delta_l$ in the finite sequence of edges going from O_d to $(a_1, ..., a_d)$. For the second assertion, change i_l into $s(i_l)$ in the finite sequence of edges going from O_d to $(a_j)_{j \in \{1,...,d\}}$. □

Proposition 6. *Let $d \geq 3$. All d-tuples in $\mathbb{Z}^d$ are accessible from O_d.*

Proof. By recurrence on $n \in \{3, ..., d\}$, we show that all vertices in $\mathbb{Z}^n \times \{0\}^{d-n}$ are accessible from O_d.

The case $n = 3$ is an immediate consequence of Proposition 3: indeed, if $(a_1, a_2, a_3) = \tau_{i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{i_0,\delta_0}(O)$ (with notations of Section 1), then $(a_1, a_2, a_3, 0, ..., 0) = \tau_{d,i_{n-1},\delta_{n-1}} \circ ... \circ \tau_{d,i_0,\delta_0}(O_d)$.

Let $n \in \{3, ..., d-1\}$ and assume that all vertices in $\mathbb{Z}^n \times \{0\}^{d-n}$ are accessible from O_d. Let $(a_1, ..., a_{n-1}, b, c, 0, ..., 0) \in Z^{n+1} \times \{0\}^{d-n-1}$. Denote by s the sum $s = \sum_{i=1}^{n-1} a_i$ and by $\mathcal{V}_0$ the vertex $(a_1, ..., a_{n-1}, -s, 0, ..., 0) \in \mathbb{Z}^n \times \{0\}^{d-n}$. By recurrence, $\mathcal{V}_0$ is accessible from O_d. Therefore, the vertex $\mathcal{V}_1 := (a_1, ..., a_{n-1}, -s, -s, 0, ..., 0) = \tau^s_{d,n+1,-1}(\mathcal{V}_0)$ is also accessible from O_d.

In a first time, we assume that $b \geq -s$ and $c \geq -s$. Denote by E the subtractive Euclid map:

$$\begin{array}{llll} E: & \mathbb{N}^2 & \longrightarrow & \mathbb{N}^2 \\ & (x, y) & \longmapsto & (x - y, y) \qquad \text{if } x \geq y \\ & & & (x, y - x) \qquad \text{otherwise.} \end{array}$$

Let $(x_0, y_0) = (s + b, s + b) \in \mathbb{N}^2$ and $(x_k, y_k) = E^k(x_0, y_0)$ for any nonnegative integer k. The sequence $(x_k, y_k)_{k \in \mathbb{N}}$ is ultimately constant; denote by $n = \min\{k \in \mathbb{N} | (x_k, y_k) = (0, \gcd(x_0, y_0)\}$ the number of iterations required by the Euclid algorithm on the input (x_0, y_0). For all k in $\{0, ..., n\}$, let $(b_k, c_k) = (x_k - s, y_k - s)$. Observe that:

(i) $(b_0, c_0) = (b, c)$,

(ii) $(b_n, c_n) = (-s, -s + \gcd(s+b, s+c))$,

(iii) the vertex $\mathcal{V}_2 := (a_1, ..., a_{n-1}, b_n, c_n, 0, ..., 0)$ is accessible from O_d,

(iv) for all $k \in \{0, .., n-1\}$, the vertex $(a_1, ..., a_{n-1}, b_k, c_k)$ is accessible from $(a_1, ..., a_{n-1}, b_{k+1}, c_{k+1})$.

The assertions (i) and (ii) are immediate; (iii) comes from the facts that $\mathcal{V}_2 = (a_1, ..., a_{n-1}, -s, -s + \gcd(s+b, s+c), 0, ..., 0) = (\tau_{d,n+1,1})^{\gcd(s+b,s+c)}(\mathcal{V}_1)$ and that $\mathcal{V}_1$ is accessible from O_d. We now prove (iv). Let k in $\{0, ..., n-1\}$. Since $(x_{k+1}, y_{k+1}) = E(x_k, y_k)$, two cases may happen: $(x_{k+1}, y_{k+1}) = (x_k - y_k, y_k)$ or $(x_{k+1}, y_{k+1}) = (x_k, y_k - x_k)$. In the first case, we have $y_k = y_{k+1}$ and $x_k = x_{k+1} + y_k = x_{k+1} + y_{k+1}$, and then, $(a_1, ..., a_{n-1}, b_k, c_k, 0, ..., 0) = (a_1, ..., a_{n-1}, x_{k+1} + y_{k+1} - s, y_{k+1} - s, 0, ..., 0) = \tau_{d,n,0}((a_1, ..., a_{n-1}, x_{k+1} - s, y_{k+1} - s, 0, ..., 0)) = \tau_{d,n,0}((a_1, ..., a_{n-1}, b_{k+1}, c_{k+1}, 0, ..., 0))$. In the second case, we have $x_k = x_{k+1}$ and $y_k = x_k + y_{k+1} = x_{k+1} + y_{k+1}$, and then, $(a_1, ..., a_{n-1}, b_k, c_k, 0, ..., 0) = (a_1, ..., a_{n-1}, x_{k+1} - s, y_{k+1} + x_{k+1} - s, 0, ..., 0) = \tau_{d,n+1,0}((a_1, ..., a_{n-1}, x_{k+1} - s, y_{k+1} - s, 0, ..., 0)) = \tau_{d,n+1,0}((a_1, ..., a_{n-1}, b_{k+1}, c_{k+1}, 0, ..., 0))$. In both cases, the vertex $(a_1, ..., a_{n-1}, b_k, c_k, 0, ..., 0)$ is accessible from $(a_1, ..., a_{n-1}, b_{k+1}, c_{k+1}, 0, ..., 0)$ in one step. Finally, the vertex $(a_1, ..., a_{n-1}, b, c, 0, ..., 0) = (a_1, ..., a_{n-1}, b_0, c_0, 0, ..., 0)$ is accessible from $(a_1, ..., a_{n-1}, b_n, c_n, 0, ..., 0) = \mathcal{V}_2$, and then, from O_d.

Now, we show that the result remains true if $b < -s$ or $c < -s$. If $b < -s$ and $c < -s$, the previous paragraph indicates that the vertex $(-a_1, ..., -a_{n-1}, -b, -c, 0, ..., 0)$ is accessible from O_d; thus the symmetric vertex $(a_1, ..., a_{n-1}, b, c, 0, ..., 0)$ is also accessible from O_d (Lemma 9). Now, assume that $b \geq -s$ and $c < -s$ (the case $b < -s$ and $c \geq -s$ being treated the same way). In this case, the sequence $(b - k(s+c))_{k \in \mathbb{N}}$ is increasing; therefore, there exists $k_0 \in \mathbb{N}$ such that $\tilde{b} := b - k_0(s+c) \geq -\sum_{i=0}^{n-2} a_i - c$. This implies that the vertex $\mathcal{V}_3 := (a_1, ..., a_{n-2}, \tilde{b}, a_{n-1}, c, 0, ..., 0) \in \mathbb{Z}^{n+1} \times \{0\}^{d-n-1}$ satisfies the conditions of the previous paragraph, which are: $a_{n-1} \geq -\sum_{i=1}^{n-2} a_i - \tilde{b}$ (indeed, $\tilde{b} \geq b$) and $c \geq -\sum_{i=1}^{n-2} a_i - \tilde{b}$. Therefore, the vertex $\mathcal{V}_3$ is accessible from O_d, and by Lemma 9, so is the vertex $\mathcal{V}_4 := (a_1, ..., a_{n-1}, \tilde{b}, c, 0, ..., 0)$. Since furthermore $(a_1, ..., a_{n-1}, b, c, 0, ..., 0) = (\tau_{d,n,0})^{k_0}(\mathcal{V}_4)$, we conclude that the vertex $(a_1, ..., a_{n-1}, b, c, 0, ..., 0)$ is accessible from O_d. Finally, all vertices in $\mathbb{Z}^{n+1} \times \{0\}^{d-n-1}$ are accessible from O_d. This concludes the proof by recurrence. □

Proof of Lemma 6. Lemma 6 follows from Proposition 6 and Lemma 8. □

2.4 Proof of Theorem 3

Theorem 3. *Let $d \geq 3$, and $A_d = \{1, ..., d\}$. There exists a strict episturmian word over A_d whose broken line is not trapped between two parallel hyperplanes of $\mathbb{R}^d$.*

Proof. An immediate adaptation of Lemma 2 and Proposition 1 leads to the following results.

1. For any $p \in (AR_d)^*$ and any $(a_1, ..., a_d) \in \mathbb{Z}^d$, there exist $s \in (AR_d)^*$ and $u, v \in (A_d)^*$ satisfying $\text{ab}(u) - \text{ab}(v) = (a_1, ..., a_n)$, such that for all episturmian words w, the words u and v are factors of $p \cdot s(w)$.

2. There exists a strict episturmian word w_∞ over A_d such that for all $(a_1, ..., a_d) \in \mathbb{Z}^d$, there exist u and $v \in \mathcal{F}(w_\infty)$ satisfying $\text{ab}(u) - \text{ab}(v) = (a_1, ..., a_d)$.

We now show that for any hyperplane Π of $\mathbb{R}^d$, the broken line $(P_n)_{n \in \mathbb{N}}$, where $P_n := \text{ab}(p_n(w_\infty)) \in \mathbb{N}^d$, does not remain at bounded distance from Π. By contradiction, assume that there exists φ a nonzero linear form on $\mathbb{R}^d$, and $C \in \mathbb{R}^+$, such that for all nonnegative integer n, $|\varphi(P_n)| \leq C$. Then, for all $n, m, p, q \in \mathbb{N}$, we have on one hand $|\varphi(P_n) - \varphi(P_m) + \varphi(P_p) - \varphi(P_q)| \leq 4C$; hence by linearity $|\varphi((P_n - P_m) - (P_q - P_p))| \leq 4C$. On the other hand, since the linear form φ is not everywhere equal to zero, it is unbounded on $\mathbb{R}^d$, so we can find $(a_1, ..., a_d) \in \mathbb{Z}^d$ such that $\varphi((a_1, ..., a_d)) > 4C$. By (2), there exists u and $v \in \mathcal{F}(w_\infty)$ that satisfy $\text{ab}(u) - \text{ab}(v) = (a_1, ..., a_d)$. If we denote by n, m, p, q the nonnegative integers such that $u = w[m+1]...w[n]$ and $v = w[p+1]...w[q]$, we have $P_n - P_m = \mathrm{ab}(u)$ and $P_q - P_p = \mathrm{ab}(v)$ in so that, finally, $|\varphi((P_n - P_m) - (P_q - P_p))| > 4C$, a contradiction. So, for any nonzero linear form on $\mathbb{R}^d$, the sequence $|\varphi(P_n)|$ is unbounded, which means that for any hyperplane Π of $\mathbb{R}^d$, the broken line $(P_n)_{n\in\mathbb{N}}$ does not remain at bounded distance from Π. Finally, the broken line $(P_n)_{n\in\mathbb{N}}$ of w_∞ is not trapped between two parallel hyperplanes of $\mathbb{R}^d$. □

2.5 Proofs of Theorem 4 and Corollary 2

Theorem 4. *Let $d \geq 2$. The vector of letter frequencies of a d-letter strict episturmian word has rationally independent entries.*

Proof. Let w a strict episturmian word; denote by $(s_n)_{n\in\mathbb{N}}$ its directive sequence (which is unique) and by f its letter frequencies vector. We recall that for all nonnegative integer n, $s_n = \sigma_{d,i}$ if and only if the $i-th$ entry of $F^n_{d,AR}(f)$ is greater than the sum of $d-1$ others, where:

$$\begin{array}{llll} F_{d,AR}: & (\mathbb{R}^+)^d & \to & (\mathbb{R}^+)^d \\ & (x_1, ..., x_d) & \mapsto & (y_{i,1}, ..., y_{i,d}) \qquad \text{if} \quad x_i \geq \sum_{j\neq i} x_j, \end{array}$$

with $y_{i,i} = x_i - \sum_{j\neq i} x_j$, and $y_{i,j} = x_j$ for all $j \neq i$.

By contradiction, assume that the entries of f are not rationally independent.

First, observe that if for some $r \in \mathbb{N}$, the $i-th$ entry of $F^r_{d,AR}(f)$ is zero, then it will remain zero; and from this point on the directive sequence will not contain the substitution $\sigma_{d,i}$, which is conflicting with the definition of strict episturmian words and the uniqueness of the directive sequence. Thus, for all $n \in \mathbb{N}$, all entries of $F^n_{d,AR}(f)$ are positive. Let l_0 a nonzero integer column vector such that $fl_0 = 0$ (recall that f is a line vector). Let $l_m = M_{s_{m-1}}...M_{s_0}l_0$. The generalized Arnoux-Rauzy matrices being invertible, l_m is also a nonzero integer column vector; it satisfies $F^m_{d,AR}(f)l_m = f.M^{-1}_{s_0}...M^{-1}_{s_{m-1}}l_m = fl_0 = 0$.

Denote $l_m = (a_{m,1}, ..., a_{m,d})^t$ and consider $D_m = \max_{i,j\in\{1,...d\}^2}(|a_{m,i} - a_{m,j}|) \in \mathbb{N}$ the difference between the maximum and the minimum entry of l_m, that we call (again) *spread* of l_m. We claim that the sequence of nonnegative integers $(D_m)_{m\in\mathbb{N}}$ is non-increasing and that it furthermore decreases infinitely often - here will be the contradiction.

For any pair of indices $(i, j) \in \{1, ..., d\}^2$, we have:

$$a_{m+1,i} - a_{m+1,j} = \begin{cases} a_{m,i} - a_{m,j} & \text{if } s_m \notin \{\sigma_{d,i}, \sigma_{d,j}\}, \\ -a_{m,j} & \text{if } s_m = \sigma_{d,i}, \\ a_{m,i} & \text{if } s_m = \sigma_{d,j}. \end{cases}$$

Observing that the extreme entries of l_m have opposite signs, we deduce that $|a_{m,i}|, |a_{m,j}| < D_m$; hence $D_{m+1} \leq D_m$. Furthermore, if $n \geq m$ is such that $\{s_m, ..., s_{n-1}\} = AR_d$, then for any pair of indices $(i, j) \in \{1, .., d\}^2$, we have:

$$|a_{n,i} - a_{n,j}| = |a_{k_0+1,i} - a_{k_0+1,j}| < D_{k_0} \leq D_m$$

where $k_0 = \max\{k \in \{m, .., n-1\} \,|\, s_k \in \{\sigma_{d,i}, \sigma_{d,j}\}\}$. Since these inequalities hold for any pair of indices $(i, j) \in \{1, ..., d\}^2$, we conclude that $D_n < D_m$. Therefore, $(D_m)_{m\in\mathbb{N}}$ is a non-increasing sequence of nonnegative integers that decreases infinitely often - which is impossible. □

We now classify episturmian words according to the dimension of the linear space generated by their vector of letter frequencies.

Corollary 2. *Let $d \geq d' \geq 1$. Let w an episturmian word over A_d. Denote by $f = (f_1, ..., f_d)$ its vector of letter frequencies, and by $(s_n)_{n\in\mathbb{N}}$ one of its directive sequences. The following assertions are equivalent.*

1. *Exactly d' substitutions appear infinitely many times in $(s_n)_{n\in\mathbb{N}}$.*

2. *The dimension of the vectorial space $f_1\mathbb{Q} + ... + f_d\mathbb{Q}$ is d'.*

Furthermore, in the case $d' \geq 2$, the word w is nonperiodic and the assertions (1) *and* (2) *are equivalent to:*

3. *There exists τ a finite composition of substitutions in AR_d, a subset $A' \subset A_d$ containing d' elements and w'_0 a standard strict episturmian word over the alphabet A' such that $w_0 = \tau(w'_0)$, where w_0 denotes the standard word associated with w.*

Proof. We first prove that (1) $\iff$ (2) in the case $d' = 1$.

If the directive sequence $(s_n)_{n\in\mathbb{N}}$ is ultimately constant, then w is periodic, and the frequency of any letter is rational. Conversely, assume that $\dim(f_1\mathbb{Q} + ... + f_d\mathbb{Q}) = 1$. Since $f_1 + ... + f_d = 1$, this implies that $f \in \mathbb{Q}^d$. Let $q \in \mathbb{N}^*$ such that $qf \in \mathbb{N}^d$. For all nonnegative integer n, we have that $F^n_{d,AR}(qf) = qF^n_{d,AR}(f)$. While at least 2 entries of $F^n_{d,AR}(qf)$ are nonzero, the sum of the entries decreases at each iteration; since this sum is a nonnegative integer, we conclude that there exists a rank r, together with a letter $i \in A_d$, such that the vector $F^r_{d,AR}(f)$ has all its entries, excepted the $i - th$, equal to zero. Then, for all $n \geq r$, all the entries of $F^r_{d,AR}(f)$, excepted the $i - th$, are to equal zero, and $s_n = \sigma_{d,i}$. Thus, the substitution $\sigma_{d,i}$ is the only to appear infinitely many times in $(s_n)_{n\in\mathbb{N}}$.

From now on and until the end of the proof, we assume that $d' \geq 2$. We are going to prove the equivalence between the three assertions by showing that (1) $\Rightarrow$ (3) $\Rightarrow$ (2) $\Rightarrow$ (1).

(1) $\Rightarrow$ (3) Without loss of generality, assume that these d' substitutions are $\sigma_{d,1}, ..., \sigma_{d,d'}$. Let $n_0 \in \mathbb{N}$ such that $(s_n)_{n\geq n_0} \in \{\sigma_{d,1}, ..., \sigma_{d,d'}\}^{\mathbb{N}}$ and contains infinitely many occurrences of each of these substitutions. Denote $w'_0 = \lim_{n\to\infty} s_{n_0} \circ ... \circ s_{n-1}(1)$. This word can also be written $w'_0 = \lim_{n\to\infty} s'_0 \circ ... \circ s'_{n-1}(1)$, with $s'_k = \sigma_{d',j_k} \in AR_{d'}$ where $j_k \in A_{d'}$ is determined by $s_{n_0+k} = \sigma_{d,j_k}$. Since each substitution in $AR_{d'}$ appears infinitely many times in $(s'_n)_{n\in\mathbb{N}}$, the word w'_0 is a standard strict episturmian word over $A_{d'}$. We conclude by observing that the standard word w_0 associated with w satisfies $w_0 = \tau(w'_0)$, with $\tau = s_0 \circ ... \circ s_{n_0-1}$.

(3) $\Rightarrow$ (2) Without loss of generality, assume that $A' = A_{d'}$. Since w'_0 is a strict episturmian word over $A_{d'}$, its letter frequencies vector $(f'_1, ..., f'_{d'})$ has rationally independent entries (Theorem 4), i.e. $\dim(f'_1\mathbb{Q} + ... + f'_{d'}\mathbb{Q}) = d'$. The vector of letter frequencies of w'_0, seen as a word over the larger alphabet A_d is $f' := (f'_1, ..., f'_{d'}, 0, ..., 0) \in \mathbb{R}^d$. Since the incidence matrix M_τ of the substitution τ belongs to $\mathrm{GL}_d(\mathbb{Z})$ (as the product of elements in $\mathrm{GL}_d(\mathbb{Z})$), and since $f = f'M_\tau$ (the vectors of letter frequencies of an episturmian word and its standard word are equal), we conclude that $\dim(f_1\mathbb{Q} + ... + f_d\mathbb{Q}) = \dim(f'_1\mathbb{Q} + ... + f'_{d'}\mathbb{Q}) = d'$.

(2) $\Rightarrow$ (1) By contraposition. Assume that $d'' \neq d'$ substitutions appear infinitely many times in $(s_n)_{n\in\mathbb{N}}$. The (already proven) implication (1) $\Rightarrow$ (2) indicates that $\dim(f_1\mathbb{Q} + ... + f_d\mathbb{Q}) = d'' \neq d'$. □

3 Extension to Cassaigne-Selmer continued fraction algorithm (C-adic words)

In this section, we show that Theorem 1 holds for C-adic words.

3.1 Results

Theorem 5. *There exists a C-adic word whose Rauzy fractal is unbounded in all directions of the plane.*

The construction is, at each step, similar to the one we made for Arnoux-Rauzy words. It relies on the

Lemma 10. *For any $(a, b, c) \in \mathbb{Z}^3$, there exist $s \in C^*$ and $u, v \in A^*$ such that* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a, b, c)$ *and for all C-adic words w which contain the factor* 13*, the finite words u and v are factors of $s(w)$.*

and on the invertibility of the incidence matrices of the C-adic substitutions c_1 and c_2. Lemma 10 is obtained by the study of accessibility in an adapted graph.

3.2 Definitions

The class of C-adic words is introduced in [CLL17], as resulting from the research of a generalized Euclid map defined for any projective direction in $(\mathbb{R}^+)^3$ -contrary to F_{AR} which is defined for almost none (see [AR91], [AS13] and [AHS16])- and producing words with the lowest complexity possible: $p(n) = 2n + 1$. This leads to the map:

$$F_C : \quad (x, y, z) \quad \mapsto \quad \begin{cases} (x - z, z, y) & \text{if } x \geq z \\ (y, x, z - x) & \text{otherwise} \end{cases},$$

and to the associated substitutions $C = \{c_1, c_2\}$ given by:

$$\begin{array}{rrcl} c_1 : & 1 & \mapsto & 1 \\ & 2 & \mapsto & 13 \\ & 3 & \mapsto & 2 \end{array} \qquad\qquad \begin{array}{rrcl} c_2 : & 1 & \mapsto & 2 \\ & 2 & \mapsto & 13 \\ & 3 & \mapsto & 3. \end{array}$$

The incidence matrices of c_1 and c_2 belong to $\mathrm{GL}_3(\mathbb{Z})$.

An infinite word w is *C-adic* if there exist a directive sequence $(s_n)_{n\in\mathbb{N}} \in C^{\mathbb{N}}$, together with a letter $\alpha \in A := \{1, 2, 3\}$, such that w can be written $w = lim_{n\to\infty} s_0 \circ ... \circ s_{n-1}(\alpha)$. One can show that, as long as $(s_n)_{n\in\mathbb{N}}$ contains infinitely many occurrences of c_1 and c_2, the sequence of finite words $(s_0 \circ ... \circ s_{n-1}(\alpha))_n$ converges to an infinite word w that, furthermore, does not depend on the letter $\alpha \in A$. Moreover, if the sequence $(s_n)_{n\in\mathbb{N}}$ does not belong to the set $C^*.\{c_1^2, c_2^2\}^{\mathbb{N}}$, then the associated C-adic word w admits $2n + 1$ factors of length n, for each integer n; in particular, w contains all letters in $A = \{1, 2, 3\}$ [CLL17].

3.3 Proof of Lemma 10

We recall:

Lemma 10. *For any $(a, b, c) \in \mathbb{Z}^3$, there exist $s \in C^*$ and $u, v \in A^*$ such that* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a, b, c)$ *and for all C-adic words w which contain the factor* 13*, the finite words u and v are factors of $s(w)$.*

This subsection is dedicated to the proof of Lemma 10.

On the set $A^4 \times \mathbb{Z}^3$, we consider the 32 involutory functions defined as follows. For $(\delta_1, \delta_2, \delta_3, \delta_4) \in \{0,1\}^4$:

$$\begin{array}{lllll} \tau_{c1,\delta_1,\delta_2,\delta_3,\delta_4}: & A^4 \times \mathbb{Z}^3 & \to & A^4 \times \mathbb{Z}^3 \\ & (l_1, l_2, l_3, l_4, a, b, c) & \mapsto & (l'_1, l'_2, l'_3, l'_4, a + b - \delta_1 + \delta_3, c, b - \delta_2 + \delta_4) \end{array}$$

$$\text{where } \left\{ \begin{array}{lllll} l'_1 = 1 & \text{if} & \delta_1 = 0 & \text{and } l'_1 = 3 & \text{otherwise,} \\ l'_2 = 3 & \text{if} & \delta_2 = 0 & \text{and } l'_2 = 1 & \text{otherwise,} \\ l'_3 = 1 & \text{if} & \delta_3 = 0 & \text{and } l'_3 = 3 & \text{otherwise,} \\ l'_4 = 3 & \text{if} & \delta_4 = 0 & \text{and } l'_4 = 1 & \text{otherwise,} \end{array} \right.$$

and

$$\begin{array}{lllll} \tau_{c2,\delta_1,\delta_2,\delta_3,\delta_4}: & A^4 \times \mathbb{Z}^3 & \to & A^4 \times \mathbb{Z}^3 \\ & (l_1, l_2, l_3, l_4, a, b, c) & \mapsto & (l'_1, l'_2, l'_3, l'_4, b - \delta_1 + \delta_3, a, b + c - \delta_2 + \delta_4) \end{array}$$

$$\text{where } \left\{ \begin{array}{lllll} l'_1 = 1 & \text{if} & \delta_1 = 0 & \text{and } l'_1 = 3 & \text{otherwise,} \\ l'_2 = 3 & \text{if} & \delta_2 = 0 & \text{and } l'_2 = 1 & \text{otherwise,} \\ l'_3 = 1 & \text{if} & \delta_3 = 0 & \text{and } l'_3 = 3 & \text{otherwise,} \\ l'_4 = 3 & \text{if} & \delta_4 = 0 & \text{and } l'_4 = 1 & \text{otherwise.} \end{array} \right.$$

Let $\mathcal{G}_C$ the infinite oriented graph whose set of vertices is $A^4 \times \mathbb{Z}^3$, and such that the outgoing edges from a vertex $V = (l_1, l_2, l_3, l_4, a, b, c) \in A^4 \times \mathbb{Z}^3$ are the maps $\tau_{\sigma,\delta_1,\delta_2,\delta_3,\delta_4}$, with $\sigma \in C$ and $\delta_i \in \{0,1\}$ if $l_i = 2$ and $\delta_i = 0$ otherwise, which respectively lead to the vertices $\tau_{\sigma,\delta_1,\delta_2,\delta_3,\delta_4}(V)$.

Remark 5. *There are always at least two outgoing edges from a vertex $V = (l_1, l_2, l_3, l_4, a, b, c)$, which are $\tau_{c1,0,0,0,0}$ and $\tau_{c2,0,0,0,0}$. There are at most 32 outgoing edges; this case happens if and only if $l_1 = l_2 = l_3 = l_4 = 2$.*

Exemple 1. *The outgoing edges from the vertex $V = (2,3,3,3,0,1,-1)$ are $\tau_{c1,0,0,0,0}$, $\tau_{c1,1,0,0,0}$, $\tau_{c2,0,0,0,0}$ and $\tau_{c2,1,0,0,0}$. They respectively lead to the vertices $(1,2,2,2,1,-1,1)$, $(3,2,2,2,0,-1,1)$, $(1,3,3,3,1,0,0)$ and $(3,3,3,3,0,0,0)$.*

Our aim is to show that for all $(a,b,c) \in \mathbb{Z}^3$, there exists $(l_1,l_2,l_3,l_4) \in A^4$ such that the vertex $(l_1,l_2,l_3,l_4,a,b,c) \in A^4 \times \mathbb{Z}^3$ is accessible from the vertex $O_C = (1,3,1,3,0,0,0)$. The motivation lies in Lemma 11.

Lemma 11. *Let $(l_1,l_2,l_3,l_4,a,b,c) \in A^4 \times \mathbb{Z}^3$. Assume that there exist $n \in \mathbb{N}$ and a finite sequence $(\tau_i)_{i\in\{0,\ldots,n-1\}}$, where τ_i denotes $\tau_{\sigma_i,\delta_{1i},\delta_{2i},\delta_{3i},\delta_{4i}}$ with $(\sigma_i,\delta_{1i},\delta_{2i},\delta_{3i},\delta_{4i}) \in C \times \{0,1\}^4$, such that $(l_1,l_2,l_3,l_4,a,b,c) = \tau_{n-1} \circ \ldots \circ \tau_0(O_C)$. Then, there exist $u, v \in A^*$ that satisfy $\mathrm{ab}(u) - \mathrm{ab}(v) = (a,b,c)$, $u[0] = l_1$, $u[-1] = l_2$, $v[0] = l_3$ and $v[-1] = l_4$ - where $u[-1]$ and $v[-1]$ denote the last letter of u and v respectively - and such that for all C-adic words w which contain the factor 13, u and v are factors of $\sigma_{n-1} \circ \ldots \circ \sigma_0(w)$.*

Proof. We are going to build iteratively two finite sequences of finite words (u_i) and (v_i), where $i \in \{0,\ldots,n\}$, so that for all i, the words u_i and v_i are factors of $\sigma_{i-1} \circ \ldots \circ \sigma_0(13)$, and such that the ultimate elements u_n and v_n satisfy $u_n[0] = l_1$, $u_n[-1] = l_2$, $v_n[0] = l_3$, $v_n[-1] = l_4$ and $\mathrm{ab}(u_n) - \mathrm{ab}(v_n) = (a,b,c)$.

First, we choose $u_0 = v_0 = 13$. We immediately have $u_0[0] = 1$, $u_0[-1] = 3$, $v_0[0] = 0$, $v_0[-1] = 3$ and $\mathrm{ab}(u_0) - \mathrm{ab}(v_0) = (0,0,0)$. Let $i \in \{0,\ldots,n-1\}$, and denote $(l_{1i}, l_{2i}, l_{3i}, l_{4i}, a_i, b_i, c_i) = \tau_{i-1} \circ \ldots \circ \tau_0(O_C)$. Assume that u_i and v_i are factors of $\sigma_{i-1} \circ \ldots \circ \sigma_0(13)$ and satisfy $u_i[0] = l_{1i}, u_i[-1] = l_{2i}, v_i[0] = l_{3i}, v_i[-1] = l_{4i}$ and $\mathrm{ab}(u_i) - \mathrm{ab}(v_i) = (a_i, b_i, c_i)$. From $\sigma_i(u_i)$ and $\sigma_i(v_i)$, we define u_{i+1} and v_{i+1} according to the following tables.

δ_1	δ_2	choice for u_{i+1}
0	0	$\sigma_i(u_i)$
0	1	u_{i+1} such that $u_{i+1}.3 = \sigma_i(u_i)$
1	0	u_{i+1} such that $1.u_{i+1} = \sigma_i(u_i)$
1	1	u_{i+1} such that $1.u_{i+1}.3 = \sigma_i(u_i)$

δ_3	δ_4	choice for v_{i+1}
0	0	$\sigma_i(v_i)$
0	1	v_{i+1} such that $v_{i+1}.3 = \sigma_i(v_i)$
1	0	v_{i+1} such that $1.v_{i+1} = \sigma_i(v_i)$
1	1	v_{i+1} such that $1.v_{i+1}.3 = \sigma_i(v_i)$

In all cases, the words u_{i+1} and v_{i+1} are factors of $\sigma_i \circ ... \circ \sigma_0(13)$ and satisfy, by definition of the maps $\tau_{\sigma,\delta_1,\delta_3,\delta_3,\delta_4}$, the five equalities: $u_{i+1}[0] = l_{1i+1}$,$u_{i+1}[-1] = l_{2i+1}$,$v_{i+1}[0] = l_{3i+1}$,$v_{i+1}[-1] = l_{4i+1}$ and $\mathrm{ab}(u_{i+1}) - \mathrm{ab}(v_{i+1}) = (a_{i+1}, b_{i+1}, c_{i+1})$, where $(l_{1i+1}, l_{2i+1},\ l_{3i+1}, l_{4i+1}, a_{i+1}, b_{i+1}, c_{i+1}) = \tau_i \circ ... \circ \tau_0(O_C)$. Finally, at step $l = n$, we obtain $u_n, v_n \in \mathcal{F}(\sigma_{n-1} \circ ... \circ \sigma_0(13))$ that satisfy $u_n[0] = l_1$, $u_n[-1] = l_2$, $v_n[0] = l_3$, $v_n[-1] = l_4$ and $\mathrm{ab}(u_n) - \mathrm{ab}(v_n) = (a, b, c)$. In particular, if w denotes a C-adic word which contains the factor 13, we have u_n, $v_n \in \mathcal{F}(\sigma_{n-1} \circ ... \sigma_0(w))$. □

Proposition 7. *For all $(a, b, c) \in \mathbb{Z}^3$, there exists $(l_1, l_2, l_3, l_4) \in A^4$ such that the vertex $(l_1, l_2, l_3, l_4, a, b, c)$ is accessible from O_C.*

The proof of Proposition 7 lies on the 3 following lemmas.

Lemma 12. *For all $a \in \mathbb{N}$, the vertices $V_a = (1, 2, 1, 1, -a, a, -a+1)$ and $V'_a = (3, 3, 2, 3, a-1, -a, a)$ are accessible from O_C.*

Proof. Let $a \in \mathbb{N}$. Observe that the vertex V_a is accessible from O_C if and only of the vertex V'_a is accessible as well. Indeed, let U and $\tilde{U}$ such that $\{U, \tilde{U}\} = \{V_a, V'_a\}$. If there exist $n \in \mathbb{N}$ and a finite sequence $(\tau_i)_{i \in \{0,...,n-1\}}$, where τ_i denotes $\tau_{\sigma_i,\delta_{1i},\delta_{2i},\delta_{3i},\delta_{4i}}$ with $(\sigma_i, \delta_{1i}, \delta_{2i}, \delta_{3i}, \delta_{4i}) \in C \times \{0, 1\}^4$, such that $U = \tau_{n-1} \circ ... \circ \tau_0(O_C)$, then we have $\tilde{U} = \tilde{\tau}_{n-1} \circ ... \circ \tilde{\tau}_0(O_C)$, where $\tilde{\tau}_i$ denotes $\tau_{\tilde{\sigma}_i,\delta_{4i},\delta_{3i},\delta_{2i},\delta_{1i}}$ with $\{\sigma_i, \tilde{\sigma}_i\} = \{c_1, c_2\}$.

We prove by induction on $a \in \mathbb{N}$ that the vertex V_a is accessible from O_C. For $a = 0$, we have $V_0 = \tau_{c_1,0,0,0,1} \circ \tau_{c_1,0,0,0,0} \circ \tau_{c_2,1,0,0,0} \circ \tau_{c_2,0,1,0,0} \circ \tau_{c_1,0,0,1,0} \circ \tau_{c_2,0,0,0,0}(O_C)$.

Let $a \in \mathbb{N}$. Assume that the vertex V_a is accessible from O_C. We want to prove that the vertex V_{a+1} is also accessible from O_C. One can check that $V'_{a+1} = \tau_{c_2,0,0,0,0} \circ \tau_{c_2,1,0,0,0} \circ \tau_{c_2,0,0,0,0} \circ (\tau_{c_2,0,0,0,1} \circ \tau_{c_2,0,0,0,0})^a(V_a)$, meaning that the vertex V'_{a+1} is accessible from V_a, and thereby, from O_C. We conclude, resorting to the first paragraph, that V_{a+1} is accessible from O_C as well. Finally, for any $a \in \mathbb{N}$, the vertices V_a and V'_a are accessible from O_C. □

Lemma 13. *For all $a \in \mathbb{N}$ and for all $k \in \mathbb{N}^*$, the vertices $V_{a,k} = (1, 3, 1, 1, -a, a, -a + 1 + k)$ and $V'_{a,k} = (2, 3, 2, 2, a, -a, k)$ are accessible from O_C.*

Proof. Let $a \in \mathbb{N}$ and $k \in \mathbb{N}^*$. Following Lemma 12, denote by V_a the vertex $(1, 2, 1, 1, -a, a, -a+1)$, which is accessible from O_C. One can check that $V_{a,k} = (\tau_{c_2,0,0,0,1} \circ \tau_{c_2,0,0,0,0})^k(V_a)$. Similarly, $V'_{a,k} = (\tau_{c_2,0,0,0,0} \circ \tau_{c_2,0,0,0,1})^{k-1} \circ \tau_{c_2,0,0,0,0}(V_a)$. Finally, the vertices $V_{a,k}$ and $V'_{a,k}$ are accessible from O_C. □

Corollary 3. *For all $a, b \in \mathbb{Z}$, there exists $(l_1, l_2, l_3, l_4) \in A^4$ such that the vertex $(l_1, l_2, l_3, l_4, -a, a, b)$ is accessible from O_C.*

Proof. It is sufficient to prove the corollary for $a \in \mathbb{N}$ and $b \in \mathbb{Z}$. Indeed, one check that if there exist $n \in \mathbb{N}$ and a finite sequence $(\tau_i)_{i \in \{0,...,n-1\}}$, where τ_i denotes $\tau_{\sigma_i,\delta_{1i},\delta_{2i},\delta_{3i},\delta_{4i}}$ with $(\sigma_i, \delta_{1i}, \delta_{2i}, \delta_{3i}, \delta_{4i}) \in C \times \{0, 1\}^4$, such that$(l_1, l_2, l_3, l_4, -a, a, b) = \tau_{n-1} \circ ... \circ \tau_0(O_C)$, then we have that $(l_3, l_4, l_1, l_2, a, -a, -b) = \tilde{\tau}_{n-1} \circ ... \circ \tilde{\tau}_0(O_C)$, where $\tilde{\tau}_i$ denotes $\tau_{\sigma_i,\delta_{3i},\delta_{4i},\delta_{1i},\delta_{2i}}$.

Let $a \in \mathbb{N}$. One one hand, if $b \in -\mathbb{N}^*$, we know by Lemma 13 that the vertex $V'_{a,-b} = (2,3,2,2,a,-a,-b)$ is accessible from O_C. Then, due to the symmetry described in the first paragraph of the proof, we obtain that the vertex $(2,2,2,3,-a,a,b)$ is accessible from O_C. On the other hand, if $b \geq 2-a$, then the vertex $V_{a,b+a-1} = (1,3,1,1,-a,a,b)$ is accessible from O_C by Lemma 13. To conclude the proof of the corollary, we need to check the three remaining cases: $(a,b) = (0,0), (0,1)$ and $(1,0)$. This is easy: $O_C = (1,3,1,3,0,0,0)$ is trivially accessible from itself, and we know from Lemma 12 that the vertices $V_0 = (1,2,1,1,0,0,1)$ and $V_1 = (1,2,1,1,-1,1,0)$ are accessible from O_C. □

Lemma 14. *For all $a,b,c \in \mathbb{Z}$ satisfying $a,c \geq -b$, there exists $(l_1,l_2,l_3,l_4) \in A^4$ such that the vertex (l_1,l_2,l_3,l_4,a,b,c) is accessible from O_C.*

Proof. Let $a,b,c \in \mathbb{Z}$ such that $a,c \geq -b$. Denote by E the subtractive Euclid map:

$$\begin{array}{rlcll} E: & \mathbb{N}^2 & \longrightarrow & \mathbb{N}^2 & \\ & (x,y) & \longmapsto & (x-y,y) & \text{if } x \geq y \\ & & & (x,y-x) & \text{otherwise.} \end{array}$$

Let $(x_0,y_0) = (b+a,b+c) \in \mathbb{N}^2$ and $(x_k,y_k) = E^k(x_0,y_0)$ for any nonnegative integer k. The sequence $(x_k,y_k)_{k\in\mathbb{N}}$ is ultimately constant; denote by $n = \min\{k \in \mathbb{N} | (x_k,y_k) = (0, \gcd(x_0,y_0)\}$ the number of iterations required by the Euclid algorithm on the input (x_0,y_0). For all k in $\{0,...,n\}$, let $(a_k,c_k) = (x_k-b, y_k-b)$. Observe that:

(i) $(a_0,c_0) = (a,c)$,
(ii) $(a_n,b,c_n) = (-b,b,-b+\gcd(b+a,b+c))$,
(iii) there exists $(l_{1n},l_{2n},l_{3n},l_{4n}) \in A^4$ such that the vertex $(l_{1n},l_{2n},l_{3n},l_{4n},a_n,b,c_n)$ is accessible from O_C,
(iv) for all $k \in \{0,..,n-1\}$, for all $(l_{1k+1},l_{2k+1},l_{3k+1},l_{4k+1}) \in A^4$, there exists $(l_{1k},l_{2k},l_{3k},l_{4k}) \in A^4$ such that the vertex $(l_{1k},l_{2k},l_{3k},l_{4k},a_k,b,c_k)$ is accessible from $(l_{1k+1},l_{2k+1},l_{3k+1},l_{4k+1}, a_{k+1},b,c_{k+1})$.

The assertions (i) and (ii) are immediate; (iii) comes from Corollary 3. We now prove (iv). Let k in $\{0,...,n-1\}$ and $(l_{1k+1},l_{2k+1},l_{3k+1},l_{4k+1}) \in A^4$. Since $(x_{k+1},y_{k+1}) = E(x_k,y_k)$, two cases may happen: $(x_{k+1},y_{k+1}) = (x_k-y_k,y_k)$ or $(x_{k+1},y_{k+1}) = (x_k,y_k-x_k)$. In the first case, we have $y_k = y_{k+1}$ and $x_k = x_{k+1}+y_k = x_{k+1}+y_{k+1}$; hence $(a_k,b,c_k) = (x_{k+1}+y_{k+1}-b,b,y_{k+1}-b) = (a_{k+1}+c_{k+1}+b,b,c_{k+1}) = (a_{k+1},b,c_{k+1})(M_{c_1})^2$. So there exists $(l_{1k},l_{2k},l_{3k},l_{4k}) \in A^4$ such that $(l_{1k},l_{2k},l_{3k},l_{4k},a_k,b,c_k) = (\tau_{c_1,0,0,0,0})^2((l_{1k+1},\ l_{2k+1},l_{3k+1},l_{4k+1},a_{k+1},b,c_{k+1}))$. In the second case, we have $x_k = x_{k+1}$ and $y_k = x_k+y_{k+1} = x_{k+1}+y_{k+1}$; hence $(a_k,b,c_k) = (x_{k+1}-b,b,y_{k+1}+x_{k+1}-b) = (a_{k+1},b,c_{k+1}+a_{k+1}+b) = (M_{c_2})^2$. So there exists $(l_{1k},l_{2k},l_{3k},l_{4k}) \in A^4$ such that $(l_{1k},l_{2k},l_{3k},l_{4k},a_k,b,c_k) = (\tau_{c_2,0,0,0,0})^2((l_{1k+1},l_{2k+1},l_{3k+1},l_{4k+1},a_{k+1},b,c_{k+1}))$. In both cases, the triplet $(l_{1k},l_{2k},l_{3k},l_{4k},a_k,b,c_k)$ is accessible from $(l_{1k+1},l_{2k+1},l_{3k+1},l_{4k+1},a_{k+1},b,c_{k+1})$ in one step.

Finally, the vertex $(l_{10},l_{20},l_{30},l_{40},a,b,c) = (l_{10},l_{20},l_{30},l_{40},a_0,b,c_0)$ is accessible from $(l_{1n},l_{2n}, l_{3n},l_{4n},a_n,b,c_n)$, and therefore, from O_C. □

Proof of Proposition 7. Let $(a,b,c) \in A^4$. Without lost of generality, we assume that two entries at least are nonnegative. Indeed, if there exist $n \in \mathbb{N}$ and a finite sequence $(\tau_i)_{i\in\{0,...,n-1\}}$, where τ_i denotes $\tau_{\sigma_i,\delta_{1i},\delta_{2i},\delta_{3i},\delta_{4i}}$ with $(\sigma_i,\delta_{1i},\delta_{2i},\delta_{3i},\delta_{4i}) \in C \times \{0,1\}^4$, such that $(l_1,l_2,l_3,l_4,a,b,c) = \tau_{n-1} \circ ... \circ \tau_0(O_C)$, then we have that $(l_3,l_4,l_1,l_2,-a,-b,-c) = \tilde{\tau}_{n-1} \circ ... \circ \tilde{\tau}_0(O_C)$, where $\tilde{\tau}_i$ denotes $\tau_{\sigma_i,\delta_{3i},\delta_{4i},\delta_{1i},\delta_{2i}}$.

Case 1. If $b,c \geq 0$, we write $(a,b,c) = (b-c+a,c-a,a)M_{c_1}M_{c_2}$. Since $a \geq a-c$ and $b-c+a = a-c+b \geq a-c$, we deduce from Lemma 14 that there exists $(l'_1,l'_2,l'_3,l'_4) \in A^4$ such that $(l'_1, l'_2, l'_3, l'_4, b-c+a, c-a, a)$ is accessible from O_C. Therefore, the vertex $(l_1, l_2, l_3, l_4, a, b, c) = \tau_{c_2,0,0,0,0} \circ \tau_{c_1,0,0,0,0}((l'_1, l'_2, l'_3, l'_4, b-c+a, c-a, a))$ is also accessible from O_C.

Case 2. If $a, b \geq 0$, we write $(a, b, c) = (c, a-c, b+c-a)M_{c_2}M_{c_1}$. Since $c \geq c-a$ and $b+c-a \geq c-a$, we deduce from Lemma 14 that there exists $(l'_1, l'_2, l'_3, l'_4) \in A^4$ such that $(l'_1, l'_2, l'_3, l'_4, c, a-c, b+c-a)$ is accessible from O_C. Therefore, the vertex $(l_1, l_2, l_3, l_4, a, b, c) = \tau_{c_1,0,0,0,0} \circ \tau_{c_2,0,0,0,0}((l'_1, l'_2, l'_3, l'_4, c, a-c, b+c-a))$ is also accessible from O_C.

Case 3. At last, assume that $a, c \geq 0$. If $a \geq c$, then we write $(a, b, c) = (a-c, c, b)M_{c_1}$. Since $(a-c, c, b) \in \mathbb{N} \times \mathbb{N} \times \mathbb{Z}$, we know from *case 2* that there exists $(l'_1, l'_2, l'_3, l'_4) \in A^4$ such that $(l'_1, l'_2, l'_3, l'_4, a-c, c, b)$ is accessible from O_C. We deduce that the vertex $(l_1, l_2, l_3, l_4, a, b, c) = \tau_{c_1,0,0,0,0}((l'_1, l'_2, l'_3, l'_4, a-c, c, b))$ is also accessible from O_C. Similarly, if $c \geq a$, then we write $(a, b, c) = (b, a, c-a)M_{c_2}$. Since $(b, a, c-a) \in \mathbb{Z} \times \mathbb{N} \times \mathbb{N}$, we know from *case 1* that there exists $(l'_1, l'_2, l'_3, l'_4) \in A^4$ such that $(l'_1, l'_2, l'_3, l'_4, b, a, c-a)$ is accessible from O_C. We deduce that the vertex $(l_1, l_2, l_3, l_4, a, b, c) = \tau_{c_2,0,0,0,0}((l'_1, l'_2, l'_3, l'_4, b, a, c-a))$ is also accessible from O_C. □

Proof of Lemma 10. Let $(a, b, c) \in \mathbb{Z}^3$. By Proposition 7, there exists $(l_1, l_2, l_3, l_4) \in A^4$ such that the vertex $(l_1, l_2, l_3, l_4, a, b, c)$ is accessible from O_C. Then, by Lemma 11, there exist $u, v \in A^*$ which satisfy $\mathrm{ab}(u) - \mathrm{ab}(v) = (a, b, c)$ and such that for any C-adic word w which contains the factor 13, the words u and v are factors of $\sigma_{n-1} \circ ... \circ \sigma_0(w)$, where the substitutions σ_i are given by the path that goes from O_C to $(l_1, l_2, l_3, l_4, a, b, c)$. □

3.4 Proof of Theorem 5

We recall:

Theorem 5. *There exists a C-adic word whose Rauzy fractal is unbounded in all directions of the plane.*

Lemma 15. *For any $p \in C^*$ and any $(a, b, c) \in \mathbb{Z}^3$, there exist $s \in C^*$ and $u, v \in A^*$ satisfying* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a, b, c)$, *such that for all C-adic words w which contain the factor* 13, *the words u and v are factors of $p \cdot s(w)$.*

Proof. Let $p \in C^*$ and $(a, b, c) \in \mathbb{Z}^3$. Denote by M_p the incidence matrix of p, which is a finite product of the incidence matrices M_{c_1} and M_{c_2}, and thus belongs to $\mathrm{GL}_3(\mathbb{Z})$. By Lemma 10, there exist $s \in C^*$ and $u, v \in A^*$ satisfying $\mathrm{ab}(u) - \mathrm{ab}(v) = (a, b, c)(M_p)^{-1}$, such that for all C-adic words which contain 13, the words u and v are factors of $s(w)$. But then, $p(u)$ and $p(v)$ are factors of $p \cdot s(w)$ and satisfy $\mathrm{ab}(p(u)) - \mathrm{ab}(p(v)) = (\mathrm{ab}(u) - \mathrm{ab}(v))M_p = (a, b, c)$. □

We now construct a C-adic word for which all triplets of integers can be obtained as a difference of two of its abelianized factors.

Proposition 8. *There exists a C-adic word w_∞ such that for all $(a, b, c) \in \mathbb{Z}^3$, there exist u and $v \in \mathcal{F}(w_\infty)$ satisfying* $\mathrm{ab}(u) - \mathrm{ab}(v) = (a, b, c)$.

Proof. Let $\varphi : \mathbb{N} \to \mathbb{Z}^3$ a bijection (that can be chosen explicitly). We construct an infinite word $d \in C^{\mathbb{N}}$ as the limit of the sequence of finite words $(p_k)_{k\in\mathbb{N}} \in (C^*)^{\mathbb{N}}$ that we define by recurrence as follows. First, we set p_0 the prefix given by Lemma 10 for $(a, b, c) = \varphi(0)$. Now, for $k \in \mathbb{N}$, we set $p_{k+1} = p_k.c_1.c_2.c_1.s$, where $s \in C^*$ is given by applying Lemma 15 to the word $p_k.c_1.c_2.c_1 \in C^*$ and the vector $\varphi(k+1) \in \mathbb{Z}^3$. By construction, the sequence of finite words $(p_k)_{k\in\mathbb{N}} \in (C^*)^{\mathbb{N}}$ converges to an infinite sequence $(d_n)_{n\in\mathbb{N}} \in C^{\mathbb{N}}$ which contains infinitely many occurrences of c_1 and c_2. This guarantees that the sequence of finite words $(d_0 \circ ... \circ d_{n-1}(1))_{n\in\mathbb{N}}$ converges to a C-adic word, that we denote by w_∞. Now, observe that for any $k \in \mathbb{N}$, the directive sequence of

w_∞ starts with the prefix p_k. Let $n_k \in \mathbb{N}$ such that $p_k = d_0...d_{n_k-1}$, and denote by w_0 and w_1 the C-adic words with directive sequences, respectively, $(d_n)_{n \geq n_k+1}$ and $(d_n)_{n \geq n_k}$. By construction, the sequence $(d_n)_{n \geq n_k+1}$ contains infinitely many occurrences of the factor $c_1.c_2.c_1$, and thus, does not belong to the set $C^*.\{c_1^2, c_2^2\}^{\mathbb{N}}$. This implies successively that the word w_0 contains the letter 2 and that the word $w_1 = d_{n_k}(w_0)$ contains the factor 13. Finally, by Lemma 10, there exist $u_k, v_k \in \mathcal{F}(w_\infty) = \mathcal{F}(p_k(w_1))$ such that $\mathrm{ab}(u_k) - \mathrm{ab}(v_k) = \varphi(k)$. □

Corollary 4. *The imbalance of the word w_∞ is infinite.*

Proof. For any $n \in \mathbb{N}$, there exist u_n and $v_n \in \mathcal{F}(w_\infty)$ such that $\mathrm{ab}(u_n) - \mathrm{ab}(v_n) = (n, 0, -n)$; this implies both $|u_n| = |v_n|$ and $|u_n|_1 - |v_n|_1 = n$. The imbalance of w_∞ is thus infinite. □

Proof of Theorem 5. We obtain, by applying Proposition 2 to the word w_∞ described in Proposition 8, that the Rauzy fractal associated with w_∞ cannot be trapped between two parallels lines. □

References

[AHS16] Artur Avila, Pascal Hubert, and Alexandra Skripchenko. On the Hausdorff dimension of the Rauzy gasket. *Bulletin de la Société mathématique de France*, 144:539–568, 2016.

[And20] Mélodie Andrieu. Thesis. *in preparation*, 2020.

[AR91] Pierre Arnoux and Gérard Rauzy. Représentation géométrique de suites de complexité 2n+1. *Bulletin de la Société Mathématique de France*, 119:199–215, 1991.

[AS13] Pierre Arnoux and Štěpán Starosta. The Rauzy Gasket. In *Further Developments in Fractals and Related Fields*, pages 1–23. Springer, 2013.

[CFZ00] Julien Cassaigne, Sébastien Ferenczi, and Luca Q. Zamboni. Imbalances in Arnoux-Rauzy sequences. *Annales de l'Institut Fourier*, 50:1265–1276, 2000.

[CLL17] Julien Cassaigne, Sébastien Labbé, and Julien Leroy. A set of sequences of complexity 2n+1. In *WORDS 2017 Proceedings*, pages 144–156. Springer, 2017.

[DHS] Ivan Dynnikov, Pascal Hubert, and Alexandra Skripchenko. Dynamical systems around the Rauzy gasket and their ergodic properties. *in preparation.*

[DJP01] Xavier Droubay, Jacques Justin, and Giuseppe Pirillo. Episturmian words and some constructions of de Luca and Rauzy. *Theoretical Computer Science*, 255(1):539–553, 2001.

[GJ09] Amy Glen and Jacques Justin. Episturmian words: a survey. *RAIRO - Theoretical Informatics and Applications*, 43(3):403–442, 2009.

[JP02] Jacques Justin and Giuseppe Pirillo. Episturmian words and episturmian morphisms. *Theoretical Computer Science*, 276(1):281–313, 2002.

[Lot97] M. Lothaire. *Combinatorics on Words.* Cambridge Mathematical Library. Cambridge University Press, 1997.

[Ose68] Valery Iustinovich Oseledets. A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:197–231, 1968.

[Sch00] Fritz Schweiger. *Multidimensional Continued Fractions.* Oxford Science Publications. Oxford University Press, 2000.

V. Morphic words and equidistributed sequences

This chapter is made of the eponymous article, written with Anna Frid and published in Theoretical Computer Science*.*

Contents

Morphic words and equidistributed sequences

Mélodie Andrieu, Anna E. Frid

Abstract

The problem we consider is the following: Given an infinite word w on an ordered alphabet, construct the sequence $\nu_w = (\nu[n])_n$, equidistributed on $[0, 1]$ and such that $\nu[m] < \nu[n]$ if and only if $\sigma^m(w) < \sigma^n(w)$, where σ is the shift operation, erasing the first symbol of w. The sequence ν_w exists and is unique for every word with well-defined positive uniform frequencies of every factor, or, in dynamical terms, for every element of a uniquely ergodic subshift. In this paper we describe the construction of ν_w for the case when the subshift of w is generated by a morphism of a special kind; then we overcome some technical difficulties to extend the result to all binary morphisms. The sequence ν_w in this case is also constructed with a morphism.

At last, we introduce a software tool which, given a binary morphism φ, computes the morphism on extended intervals and first elements of the equidistributed sequences associated with fixed points of φ.

1 Introduction

Consider an infinite word $w = w[0]w[1] \cdots w[n] \cdots$ on an ordered alphabet Σ; here $w[n] \in \Sigma$. Suppose that the uniform frequency $\mu(u)$ of every factor u of w exists and is strictly positive, that is, that the dynamical system (subshift) associated with w is uniquely ergodic, and μ is the unique ergodic measure on it (see [12] for a discussion of this notion). Now for a factor u of w, define $\nu(u)$ as the sum of measures μ of all words of the same length as u which are lexicographically less than u, and $\nu(w)$ as the limit $\nu(w) = \lim_{n\to\infty} \nu(w[0] \cdots w[n])$.

The function ν on infinite words has been considered by Lopez and Narbel [19] from the dynamical point of view. On the other hand, as it was proven in [5], for every appropriate word w, the sequence $(\nu(\sigma^n(w)))_{n=0}^{\infty}$, where σ is the shift operation, is uniformly distributed on $[0, 1]$, and moreover, for $n \neq m$, we have $\nu(\sigma^n(w)) \neq \nu(\sigma^m(w))$. This makes it possible to call the sequence $(\nu(\sigma^n(w)))_{n=0}^{\infty}$ the canonical representative of the *infinite permutation* defined by the shifts of w. Infinite permutations in this sense were introduced in [13]; as for permutations defined by words, their study was initiated independently by Makarov [21, 22, 23] and by Bandt, Keller and Pompe [7]; see also [11] and the monograph [4] summarizing that approach. The fact that the sequence $\{\nu(\sigma^n(w))\}_{n=0}^{\infty}$ is uniformly distributed means in particular that the respective permutation is also equidistributed (see [6] for the definition and discussion of an equidistributed permutation).

In this paper, given a morphism φ with several nice properties, we describe how to find $\nu(w)$ for any infinite word w from the respective subshift L_φ, and in particular for a fixed point w_φ of φ. This result generalizes the Makarov's construction for the Thue-Morse word [22]. A previous result in this direction, stated in combinatorial terms and considering not the whole subshift but just a fixed point of the morphism, can be found in [5].

The next result of the paper concerns the binary case: if the morphism is binary, even if it does not belong to the "nice" class, our technique can be adapted to it. We also support the binary case by a software tool.

After introducing usual definitions in Section 2 and the object of our study in Section 3, we have to devote Section 4 to a discussion of properties of morphic subshifts. Section 5 starts with a correct extension of the interval $[0, 1]$ to a wider set, which is needed to distinguish images of consecutive elements of the subshift. It also contains the first of main results of the paper, Theorem A, giving a way to construct a morphism on numbers corresponding to the initial morphism φ, and a sequence ν_w for any element $w \in L_\varphi$. The construction is supported by examples.

Section 6 contains a discussion of the case when the morphism φ is k-uniform and thus its fixed point w is k-automatic [1]. It is proved that in this case, the sequence ν_w is k-regular (see [2] for the respective definitions).

The construction from Theorem A works only for a restricted class of morphisms. However, in Section 7 we use some additional machinery to extend this result to any binary morphism. So, given a binary morphism, we know how to construct an equidistributed sequence corresponding to its fixed point(s) and to any element of the respective subshift.

At last, in Section 8, we discuss and refer to a software tool developed to compute the morphisms on numbers and sequences described in the paper.

2 Definitions and notation

We consider infinite words $w = w[0]w[1] \cdots w[n] \cdots$, where $w[i] \in \Sigma$, on an ordered alphabet Σ. In this paper, we usually take $\Sigma = \{a, b, c, \ldots\}$, under the convention that $a < b < c < \cdots$. The order of symbols of Σ naturally extends to the lexicographic order on finite and infinite words.

The factor $w[i] \cdots w[j]$ of a finite or infinite word w, where $j \geq i$, is denoted also by $w[i..j]$. The set of all factors of w of length n is denoted by $\mathrm{Fac}_n(w)$.

The set of infinite words over Σ is denoted by Σ^ω. As usual, the shift operation σ corresponds to erasing the first symbol: $\sigma(w[0]w[1] \cdots w[n] \cdots) = w[1]w[2] \cdots w[n + 1] \cdots$. Given an infinite word $w \in \Sigma^\omega$, we denote by L_w the closure of the orbit of w under σ. The dynamical system (L_w, σ) is called a *subshift* generated by w.

An infinite word w and its subshift L_w are called *ultimately periodic* if $w = uvvvv \cdots$ for some finite words u and v. If a word (or subshift) is not ultimately periodic, it is called *aperiodic*.

Given a set S, we denote by S^* the set of finite concatenations of elements of S. In particular, if S is an alphabet, S^* is the set of finite words on S; but if S is an interval of reals, S^* is the set of finite sequences of numbers from I. A *morphism* $f : S^* \to S^*$ is a mapping which preserves concatenation, so that $f(xy) = f(x)f(y)$ for all $x, y \in S$. Clearly, a morphism is defined by its values on all elements on S.

Consider a morphism $\varphi : \Sigma^* \mapsto \Sigma^*$, where Σ is an alphabet. If the image of a symbol of $x \in \Sigma$ starts with x, the morphism φ admits a finite or right infinite fixed point w_x starting with x and defined as the limit $\lim_{n\to\infty} \varphi^n(x)$. If in addition $\varphi(x) \neq x$, φ has no other fixed points starting with x.

If the fixed point $w_x = \varphi(w_x) = \lim_{n\to\infty} \varphi^n(x)$ is infinite, it is called also a *pure morphic* infinite word, and the associated subshift (L_{w_x}, σ) is called a *pure morphic* subshift. In most cases considered in this paper (in particular, when the morphism is *primitive*, see the definition in Section 4), the subshift does not depend on the letter x and the fixed point w_x but is uniquely defined by φ. In this case, it is denoted by (L_φ, σ), and its set of factors of length n is denoted by $\mathrm{Fac}_n(L_\varphi)$.

Note that every element u of a pure morphic subshift (L_φ, σ) can be obtained from the φ-image of another element $v = v[0]v[1]\cdots \in L_\varphi$ by the shift operation applied p times, where $p \geq 0$. Moreover, we can choose p to be less than the length of $\varphi(v[0])$: here of course we suppose that $\varphi(v[0])$ is not empty. So, $u = \sigma^p(\varphi(v))$, where $0 \leq p < |\varphi(v[0])|$. Note that in the general case, v and p, as well as $v[0]$ for a given p, are not unique.

A word w is called *recurrent* if each of its factors $s = w[i..j]$ appears in it an infinite number of times. If in addition the distances between successive occurrences of s are bounded, the word is called *uniformly recurrent.* As it is well-known, an infinite word w is uniformly recurrent if and only if the associated subshift (L_w, σ) is *minimal*, meaning that L_w does not contain any proper subset which would be closed under σ. An even stronger condition is the existence of the unique σ-invariant probability measure μ on L_w, which is equivalent to the existence of uniform positive frequencies of all factors. In this case, the word w and the dynamical system (L_w, σ) generated by it are called *uniquely ergodic.*

3 Equidistributed sequences arising from words

Note that an infinite word w is ultimately periodic if and only if $\sigma^m(w) = \sigma^n(w)$ for some $m \neq n$; so, if w is aperiodic, for each $m \neq n$ we have either $\sigma^m(w) > \sigma^n(w)$ or $\sigma^m(w) < \sigma^n(w)$.

Consider an aperiodic word w with well-defined non-zero uniform frequency $\mu(u)$ of every factor u. The subshift L_w is uniquely ergodic, and its unique ergodic measure μ is completely determined by the frequencies $\mu(u)$ which can be interpreted as the values of the measure on cylinders: here a cylinder $[u]$ is the set of elements of L_w starting with a word u.

Now let us associate with an infinite word $v \in L_w$, that is, with an infinite word with the same set of factors that w, the measure

$$\nu(v) = \mu([w_{min}, v]),$$

of the interval $[w_{min}, v]$: here w_{min} is the lexicographically minimal element of the subshift L_w, existing since the set L_w is closed, and the interval $[w_{min}, v]$ is defined as the set of all infinite words from L_w which are greater than or equal to w_{min} and less than or equal to v.

The mapping $\nu : L_w \mapsto [0, 1]$ is well-defined, and moreover, since among the shift images of w the frequency of words from the interval $[w_{min}, v]$ is the same as in the whole subshift,

$$\nu(v) = \lim_{n\to\infty} \frac{\#\{k|\sigma^k(w) < v, k \leq n\}}{n}.$$

Comparing it to the definition of a uniformly distributed (or, which is the same, equidistributed) sequence $(x[n])$ on an interval $[a, b]$, meaning that for every interval $[c, d] \subseteq [a, b]$, we have

$$\lim_{n\to\infty} \frac{\#(\{x[0], \cdots, x[n]\} \cap [c, d])}{n} = \frac{d-c}{b-a},$$

we see that for all $v \in L_w$, the sequence $\nu_v = (\nu(\sigma^n(v)))_{n=0}^{+\infty}$ is equidistributed on $[0, 1]$. Indeed, since our subshift is uniquely ergodic, the proportion of words which are less than or equal to $\sigma^n(v)$ is the same in v, w and the shift L_w in total; see also a discussion in [5].

We will denote the real number $\nu(\sigma^n(w))$ by $\nu[n]$. By the construction, the sequence $\nu_w = (\nu[0], \nu[1], \ldots)$ is unique for every uniquely ergodic infinite word w.

The mapping ν has been considered by Lopez and Narbel in [19]. On the other hand, the equidistributed sequence ν_w for a sequence w, under another notation, was considered in [6] because of its relation to the *infinite permutation* generated by w; see [21, 23, 11] for a discussion of infinite permutations defined by words.

Example 1. The famous Thue-Morse word $w_{tm} = abbabaabbaababba\cdots$ [3] is defined as the fixed point starting with a of the morphism $\varphi_{tm} : a \to ab, b \to ba$. The sequence $\nu_{tm} = \nu_{w_{tm}}$ is equal to the fixed point

$$\frac{1}{2}, 1, \frac{3}{4}, \frac{1}{4}, \frac{5}{8}, \frac{1}{8}, \frac{3}{8}, \frac{7}{8}, \cdots,$$

of the morphism $f_{tm} : [0,1]^* \mapsto [0,1]^*$:

$$f_{tm}(x) = \begin{cases} \frac{x}{2} + \frac{1}{4}, \frac{x}{2} + \frac{3}{4}, \text{ if } 0 \leq x \leq \frac{1}{2}, \\ \frac{x}{2} + \frac{1}{4}, \frac{x}{2} - \frac{1}{4}, \text{ if } \frac{1}{2} < x \leq 1. \end{cases} \tag{1}$$

Note that morphisms on intervals have been discussed in Section 2 and, as any other morphisms considered in this paper, they transform a concatenation into a concatenation.

This construction (or, more precisely, a similar construction on the interval $[-1,1]$) was found by Makarov in 2009 [22]; below in Section 5 we shall prove its correctness as a corollary of a more general statement, Theorem A.

In particular, we see from that construction that $\nu(0) = 1/2$ and $\nu(1) = 1$, which means that the Thue-Morse word is the maximal element of its subshift starting with a, and if we erase the first symbol from it, the result is the lexicographically maximal element of the subshift L_{tm}. These are known results (see, e. g., [8, 15]).

Note also that due to the symmetry between a in b, the value of $\nu(w'_{tm})$ of the other fixed point $w'_{tm} = baababbaabba\cdots$ of the same morphism φ_{tm} is also 1/2. So, the mapping ν is not injective on the respective subshift L_{tm}. As it was discussed in [19], this is a typical situation and it can be resolved by extending $[0,1]$ to a new wider domain described later in Section 5. However, if we consider just a recurrent infinite word w and its orbit, that is, the set of its shifts, and not the whole dynamical system to which it belongs, it is not necessary. Indeed, two infinite words with the same value of ν can never appear in the same orbit due to the following statement.

Proposition 2. [5] *Let w be a recurrent aperiodic word and u and v be two of its factors. Then the orbit of w cannot contain at the same time the lexicographically maximal word from L_w starting with u and the lexicographically minimal word from L_w starting with v.*

In this paper, we consider only uniformly recurrent and, moreover, uniquely ergodic words, so, Proposition 2 can always be used.

4 Properties of morphic symbolic subshifts

In this section, we define the class of morphisms such that we can directly generalize the Thue-Morse construction above to their subshifts: namely, these are *primitive order-preserving* morphisms with *separable* subshifts. For such a morphism, we construct an interval morphism similar to the Thue-Morse construction from Example 1, and prove its correctness. The considered family of morphisms includes in particular all morphisms considered by Valyuzhenich [29], and much more. Note that similar definitions have been introduced in [5], but here they are updated to better fit our wider goals.

Consider an alphabet $\Sigma = \{a_1, \ldots, a_q\}$ and let $\varphi : \Sigma^* \mapsto \Sigma^*$ be a morphism with an aperiodic fixed point $u = \varphi(u)$ starting with a letter a.

The matrix M of a morphism φ on a q-letter alphabet is a $q \times q$-matrix whose element m_{ij} is equal to the number of occurrences of a_i in $\varphi(a_j)$. The matrix M and the morphism φ are called *primitive* if in some power M^n of M all the entries are positive, i.e., for every $b \in \Sigma$ all the symbols of Σ appear in $\varphi^n(b)$ for some n. The classical Perron-Frobenius theorem says that every primitive matrix has a dominant positive *Perron-Frobenius eigenvalue* θ such that $\theta > |\lambda|$ for any other eigenvalue λ of M. It is also well-known [25] that a fixed point of a primitive morphism is uniquely ergodic; moreover, for every sequence of factors (v_n) of L_φ of increasing length, the limit

$$\lim_{n\to\infty} \frac{|\varphi(v_n)|}{|v_n|}$$

exists and is equal to θ.

Note in particular that every primitive morphism is non-erasing, which means that the images of all symbols are non-empty.

Example 3. The Thue-Morse morphism is primitive with the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. The Fibonacci morphism $\varphi_f : a \to ab, b \to a$ is primitive with the matrix $M = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$: M is not positive, but $M^2 = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ is. The Sierpinski morphism $a \to aba, b \to bbb$ is not primitive since in all the powers of its matrix $\left(\begin{smallmatrix} 2 & 1 \\ 0 & 3 \end{smallmatrix}\right)$, the left lower element is 0, and indeed, a never appears in images of b.

We say that a morphism φ is *order-preserving on an infinite word* u if for any $n, m > 0$ we have $\sigma^n(u) < \sigma^m(u)$ if and only if $\varphi(\sigma^n(u)) < \varphi(\sigma^m(u))$; here $<$ denotes the lexicographic order. A morphism is called *order-preserving* if it is order-preserving on all infinite words, or, equivalently, if for any infinite words u and v we have $u < v$ if and only if $\varphi(u) < \varphi(v)$. If this property holds only for $u, v \in L_\varphi$, we say that φ is *order-preserving on its subshift*. Note that in [5], order-preserving morphisms were called *monotone*.

Example 4. The Thue-Morse morphism φ_{tm} is order-preserving since $ab = \varphi_{tm}(a) < \varphi_{tm}(b) = ba$. The Fibonacci morphism from Example 3 is not order-preserving since $ab = \varphi_f(a) > \varphi_f(ba) = aab$, whereas $a < ba$. At the same time, $\varphi_f^2 : a \to aba, b \to ab$ is order-preserving since for all $x, y \in \{a, b\}$ we have $\varphi_f^2(ax) = abaax' < ababy' = \varphi_f(by)$, where $x', y' \in \{a, b\}^*$. So, to use our construction on the Fibonacci word $u_f = \varphi_f(u_f) = abaab\cdots$ we should consider u_f as the fixed point of φ_f^2 which is order-preserving.

The last condition on the morphism φ, or, more precisely, on the subshift (L_φ, σ), is to be *separable*. To define this property, consider an element $u \in L_\varphi$ and all the ways to represent it as $\sigma^p(\varphi(u'))$ with $u' \in L_\varphi$ and $0 \le p < |\varphi(a)|$, where $a = u'[0]$ is the first symbol of u'. At least one such pair (u', p) exists by the definition of L_φ. If this pair is unique, we call the pair (a, p) the *type* $\tau(u)$ of u. A subshift is *typable* if for all elements $u \in L_\varphi$, the type of u exists. If in addition the $\sum_{a\in\Sigma} |\varphi(a)|$ possible types can be ordered so that for all $u, v \in L_\varphi$ with $\tau(u) < \tau(v)$, we always have $u < v$, we say that the subshift L_φ is *separable*.

Example 5. The Thue-Morse subshift L_{tm} is separable. Indeed, first, any two consecutive as (or bs) in its element determine a boundary between images of letters and thus all such boundaries. Also, the last symbols of $\varphi_{tm}(a)$ and $\varphi_{tm}(b)$ are different, the incomplete image of a symbol in the beginning can also be uniquely reconstructed, so, the morphism is typable. Moreover, if $\tau(u) = (a, 0)$ and $\tau(v) = (b, 1)$, we always have $u > v$, i.e., all as which are first symbols of $\varphi_{tm}(a) = ab$ give greater words than as which are second symbols of $\varphi_{tm}(b) = ba$. The situation with bs is symmetric, so, we can order the types as $(b, 1) < (a, 0) < (b, 0) < (a, 1)$ to have $u < v$ whenever $\tau(u) < \tau(v)$ for $u, v \in L_{tm}$.

Example 6. The subshift L generated by the morphism $\varphi : a \to aab, b \to abb$ is not typable because of the common suffix b of images of letters. Indeed, consider a *special* infinite word u such that $au, bu \in L$: such a word exists since the subshift is not periodic. Then the word $b\varphi(u)$ belongs both to $\sigma^2(\varphi(au))$ and to $\sigma^2(\varphi(bu))$, so that its type is not well-defined.

Example 7. The subshift generated by the morphism $\varphi : a \to aabab, b \to bba$ is typable but not separable. Indeed, consider $u_1 = abaa \cdots = \sigma^3(\varphi(aa \cdots))$, $u_2 = ababaa \cdots = \sigma(\varphi(aa \cdots))$, $u_3 = abbb \cdots = \sigma^3(\varphi(ab \cdots))$. Then $u_1 < u_2 < u_3$ whereas $\tau(u_1) = \tau(u_3) = (a, 3)$ and $\tau(u_2) = (a, 1)$.

For recurrence and, in some cases, precise formulas for the frequencies of factors in fixed points of morphisms, see [25, 14].

In what follows, given a primitive order-preserving separable morphism φ and the respective minimal subshift L_φ, we define a mapping which allows to build the sequence ν_w, and in particular its first value $\nu(w)$, for any infinite word $w \in L_\varphi$. However, to do it, we first have to consider the extended domain to make the mapping ν injective.

5 Extended intervals and morphisms

As Lopez and Narbel showed in [19], and as we discussed above just after Example 1, the mapping $\nu : L \mapsto [0, 1]$ defined in Section 3 for any minimal subshift L is surjective but not injective. In the Thue-Morse example, the image of the greatest word starting from a, which is w_{tm} itself, is $1/2$, as well as the image of the smallest word starting from b. As it was proved in [19], this happens exactly with *consecutive* words, or, equivalently, for consecutive cylinders. Recall that for a finite word u, a cylinder $[u]$ here is the set of all infinite words from L starting from u. Finite words u_1, u_2 (and their cylinders $[u_1]$, $[u_2]$) are called *consecutive* if $u_1 < u_2$ and there is no word $w \in L$ such that $w_1 < w < w_2$, where w_1 is the greatest element of L starting with u_1 and w_2 is the smallest element of L starting with u_2. The infinite words w_1 and w_2 are also called consecutive. As it was proved in [19], consecutive infinite words are exactly words $w_1 \neq w_2$ for which $\nu(w_1) = \nu(w_2)$. Every pair of infinite consecutive words corresponds to a pair (or, more precisely, a countable number of pairs) of consecutive cylinders. For example, in the Thue-Morse subshift, the words w_{tm} and w'_{tm} are consecutive, as well as the respective cylinders $[a]$ and $[b]$, or $[abb]$ and $[baab]$, or any other pair of cylinders corresponding to prefixes of respectively w_{tm} and w'_{tm}.

Let Z be the set of ν-images of elements of consecutive pairs of words: it is a countable set since a consecutive pair can be defined by two finite (consecutive) words. To make the mapping ν injective and following [19], we replace Z in $[0, 1]$ by two copies Z^- and Z^+, and thus consider ν as a mapping from L to the associated extended interval $X = X_L = ([0, 1]\backslash Z) \cup Z^- \cup Z^+$. Here for each pair $w_1 < w_2$ of consecutive words with $\nu(w_1) = \nu(w_2) = x$ we denote $\nu(w_1) = a^- \in Z^-$ and $\nu(w_2) = a^+ \in Z^+$. It is natural to set $a^- < a^+$ and to make them both inherit from $[0, 1]$ the relation with other elements of X. To unify the notation, we may also say for a number $a \in [0, 1]\backslash Z$ that $a^- = a^+ = a$.

Let φ be a primitive order-preserving morphism on an ordered alphabet $\Sigma = \{a_1, \dots, a_q\}$, $a_1 < \cdots < a_q$, with a separable subshift (L, σ), and X_L be the associated extended interval defined above. We will define a morphism on X_L corresponding to φ, thus extending to X_L a construction from [5].

Denote by $\mu = (\mu_1, \dots, \mu_q)$ the vector of measures of cylinders $[a_i]$ in L, or, which is the same, of frequencies of symbols in any element of L. Since the morphism is primitive, these measures exist and are not equal to 0. Denote the intervals $I_{a_1} = [0, \mu_1^-]$, $I_{a_2} = [\mu_1^+, (\mu_1 + \mu_2)^-]$, …, $I_{a_q} = [(1 - \mu_q)^+, 1]$, $I_a \subset X_L$.

Now let us take all the $k = \sum_{i=1}^q |\varphi(a_i)|$ types of elements of L and denote them according to their order:

$$\tau_1 < \tau_2 < \cdots < \tau_k,$$

with $\tau_i = (b_i, p_i)$. Types and their order exist since the subshift is separable.

For each $\tau_i = (b_i, p_i)$, the frequency of factors of type τ_i in the subshift is equal to $l_i = \mu_{b_i}/\theta$, where θ is the Perron-Frobenius eigenvalue of φ. Indeed, given a word u from the subshift L_φ, consider its φ-image v interpreted as a word on the alphabet $\varphi(\Sigma)$. By the construction, occurrences of $\varphi(b_i)$ to v correspond to occurrences of b_i to u. But if now we interpret v as a word on Σ, $v \in L_\varphi$, we see that its prefix corresponding to the prefix of u of length n has a length which grows as θn with $n \to \infty$. So, factors of type τ_i occur in v exactly θ times rarer than b_i in u. Since the frequencies do not depend on the choice of an element of L_φ, we get the formula $l_i = \mu_{b_i}/\theta$.

Denote

$$J_1 = [0, l_1^-], J_2 = [l_1^+, (l_1 + l_2)^-], \ldots, J_k = [(1 - l_k)^+, 1];$$

so that in general, $J_i = [(\sum_{m=1}^{i-1} l_m)^+, (\sum_{m=1}^{i} l_m)^-]$. We will also denote $J_i = J_{b_i,p_i}$.

The interval J_i is the range of $\nu(u)$ corresponding to elements of $u \in L$ of type τ_i. Note that the first symbol of such a word u is always the symbol number $p_i + 1$ of $\varphi(b_i)$ (the range of p_i for a given b_i is from 0 to $|\varphi(b_i)| - 1$). So, the union of elements J_i corresponding to this element equal to a_m is exactly I_m for every m. In particular, each J_i is a subinterval of some I_m. By the construction, all the ends of these intervals are in Z^- or Z^+, and thus the intervals J_i do not intersect: the ends a^- and a^+ of consecutive intervals correspond to consecutive words from L.

Now we define the morphism $f : X_L^* \mapsto X_L^*$ as follows: For $x \in I_a$ we have

$$f(x) = f_{a,0}(x), \ldots f_{a,|\varphi(a)|-1}(x).$$

Here $f_{a,p}$ is the increasing affine bijection $f_{a,p} : I_a \mapsto J_{a,p}$: If $I_a = [x_1^+, x_2^-]$ and $J_{a,p} = [y_1^+, y_2^-]$, then

$$f_{a,p}(x) = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1. \tag{2}$$

Here, by the convention, the image of any $x \in Z^-$ ($x \in Z^+$) is $(f_{a,p}(x))^-$ (respectively, $(f_{a,p}(x))^+$). Note that the slope $\frac{y_2 - y_1}{x_2 - x_1}$ of the affine mapping $f_{a,p}(x)$ is equal to $1/\theta$ since the interval $J_{a,p}$ is θ times shorter than I_a.

The meaning of intervals I_a and of the morphism f is explained in the following proposition following directly from the construction.

Proposition 8. *Let $d : X_L^* \mapsto \Sigma^*$ be the morphism defined by $d(x) = a$ whenever $x \in I_a$. Then for all $x \in X_L$ we have $d(f(x)) = \varphi(d(x))$.*

This proposition means in particular that the lengths of φ-images of letters and of f-images of reals from the respective intervals are synchronized. So, the following statement holds.

Proposition 9. *Given a word $w \in L_\varphi$, where φ is primitive and order-preserving and L_φ is separable, the following statements are equivalent:*

- *The letter $\varphi(w)[n]$ is the letter indexed p of the φ-image of the letter indexed n' of w, and*
- *The number $f(\nu_w)[n]$ is the number indexed p of the f-image of the number indexed n' of ν_w.*

Example 10. The Thue-Morse morphism on $[0, 1]$ from Example 1 can now be more correctly redefined on the respective set X_{tm}. Here $1/2$, which is the frequency of a, is one of the numbers which is doubled, as well as all binary rationals from $(0, 1)$. So, we have $I_a = [0, 1/2^-]$, $I_b = [1/2^+, 1]$, $J_{a,0} = [1/4^+, 1/2^-]$, $J_{a,1} = [3/4^+, 1]$, $J_{b,0} = [1/2^+, 3/4^-]$, $J_{b,1} = [0, 1/4^-]$, and (1) can now be rewritten as

$$f_{tm}(x) = \begin{cases} f_{a,0}(x), f_{a,1}(x) \text{ for } x \in I_a, \\ f_{b,0}(x), f_{b,1}(x) \text{ for } x \in I_b, \end{cases} \tag{3}$$

where the used linear mappings on X_{tm} are defined by (2) and, of course, coincide on $[0,1]$ with those from (1).

For an example concerning the Fibonacci word as the fixed point of the square morphism φ_f^2 from Example 4, see [5]; the only difference in the presentation should be extended intervals.

The following statements is one of the main results of the paper.

Theorem A. *Let φ be a primitive morphism defining a separable subshift (L,σ) and order-preserving on it, f be the morphism on extended intervals associated with φ (and L), and $\nu_w \in X_L^\omega$ be the equidistributed sequence corresponding to a sequence $w \in L$. Then $f(\nu_w) = \nu_{\varphi(w)}$ (see the commutative diagram below).*

$$\begin{tikzcd}
w \arrow[r, "\varphi"] \arrow[d, "\nu"] & \varphi(w) \arrow[d, "\nu"] \\
\nu_w \arrow[r, "f"] & \nu_{\varphi(w)}
\end{tikzcd}$$

Proof. We shall prove first, that the two sequences have the same order among elements, and second, that $f(\nu_w)$ is equidistributed on $[0,1]$. Since the equidistributed sequence on $[0,1]$ corresponding to any ordering of elements is at most unique (if it exists, each of its element is uniquely defined as the fraction of elements in the ordering smaller than it, see also a discussion after Definition 2.3 in [6]), this is sufficient.

Suppose that $f(\nu_w)[n] < f(\nu_w)[m]$ for some $n, m \in \mathbb{N}$; our goal is to prove that $\nu_{\varphi(w)}[n] < \nu_{\varphi(w)}[m]$.

Suppose first that $f(\nu_w)[n]$ and $f(\nu_w)[m]$ are situated in the same interval $J_{c,p}$. Since all such intervals are disjoint, this means that $f(\nu_w)[n]$ is obtained as $f_{c,p}(\nu_w[n'])$ and $f(\nu_w)[m]$ is obtained as $f_{c,p}(\nu_w[m'])$ for some $n', m' \in \mathbb{N}$, where $\nu_w[n']$, $\nu_w[m'] \in I_c$. So, by the definition of I_c, we have $w[n'] = w[m'] = c$. Moreover, since f and φ are synchronized as described in Proposition 9, the symbols of φ-images of $w[n']$ and $w[m']$ numbered p are $\varphi(w)[n]$ and $\varphi(w)[m]$. So, the inequality $f(\nu_w)[n] < f(\nu_w)[m]$ implies that $\nu_w[n'] < \nu_w[m']$ since $f_{c,p}$ is an affine mapping with a positive slope, then that $\sigma^{n'}(w) < \sigma^{m'}(w)$ by the definition of ν_w, then that $\varphi(\sigma^{n'}(w)) < \varphi(\sigma^{m'}(w))$ since φ is order-preserving on L, and $\sigma^p(\varphi(\sigma^{n'}(w))) < \sigma^p(\varphi(\sigma^{m'}(w)))$ since first p symbols of $\varphi(\sigma^{n'}(w))$ and $\varphi(\sigma^{m'}(w))$ are equal (to the first p symbols of $\varphi(c)$). But since φ and f are synchronized by Proposition 9, $\sigma^p(\varphi(\sigma^{n'}(w))) = \sigma^n(\varphi(w))$ and $\sigma^p(\varphi(\sigma^{m'}(w))) = \sigma^m(\varphi(w))$. So, $\sigma^n(\varphi(w)) < \sigma^m(\varphi(w))$, and, by the definition of $\nu_{\varphi(w)}$, $\nu_{\varphi(w)}[n] < \nu_{\varphi(w)}[m]$.

Now suppose that $f(\nu_w)[n] \in J_{c_n,p_n}$ and $f(\nu_w)[m] \in J_{c_m,p_m}$, where $J_{c_n,p_n} \neq J_{c_m,p_m}$. These intervals are disjoint and correspond to types $(c_n,p_n) < (c_m,p_m)$. So, $f(\nu_w)[n]$ is the number indexed p_n in the f-image of some $\nu_w[n'] \in I_{c_n}$, and $f(\nu_w)[m]$ is the number indexed p_m in the f-image of some $\nu_w[m'] \in I_{c_m}$. By the construction of intervals I_c, this means that $w[n'] = c_n$ and $w[m'] = c_m$, and moreover, due to Proposition 9, $\varphi(w)[n]$ is the symbol indexed p_n of the φ-image of $w[n']$ and $\varphi(w)[m]$ is the symbol indexed p_m of the φ-image of $w[m']$. Since the morphism φ is separable, the type of $\sigma^n(\varphi(w))$ is thus equal to (c_n,p_n), and it is less than the type of $\sigma^m(\varphi(w))$ equal to (c_m,p_m). So, $\sigma^n(\varphi(w)) < \sigma^m(\varphi(w))$ and thus $\nu_{\varphi(w)}[n] < \nu_{\varphi(w)}[m]$, which was to be proved.

We have proved that the sequences $f(\nu_w)$ and $\nu_{\varphi(w)}$ have the same order among elements; it remains to prove that $f(\nu_w)$ is equidistributed on X_L. Indeed, let us consider any subinterval I of length l of some interval I_c. The frequency of elements of ν_w which are in I is l since ν_w is equidistributed. Due to the definition of f, the images of the interval I are $f_{c,1}(I), \ldots, f_{c,|\varphi(c)|}(I)$. These are intervals from X_L of length l/θ each, where θ is the Perron-Frobenius eigenvalue of φ. The frequency of elements of $f(\nu_w)$ from each of these intervals is also l/θ, since f and φ are synchronized in length and since θ is the limit of the ratio $|\varphi(w[0]\cdots w[n])|/n$ with $n \to \infty$; we use also the fact that the intervals $J_{x,p}$ form a disjoint partition of X_L. On the other hand, every subinterval J of some $J_{c,p}$, where $p = 1, \ldots, |\varphi(c)|$ is the $f_{c,p}$-image of a respective subinterval I of I_c which is θ times longer than it. So, the frequency of elements from J in $f(\nu_w)$ is equal to the length of J. This is true for all subintervals of $J_{c,p}$ and thus by union for all subintervals of X_L, meaning exactly that the sequence $f(\nu_w)$ is equidistributed on X_L. □

Corollary 11. *Let $w = w_a \in L_\varphi$ be the fixed point of φ starting with a. Then ν_w is the unique fixed point of f starting from a number from I_a, which is the fixed point of $f_{a,0}$.*

Proof. First of all, the fixed point w_a of φ starting with a is unique since the morphism is primitive and thus $\varphi(a) \neq a$. We know from Theorem A that $\nu_{\varphi(w_a)} = f(\nu_{w_a})$, but since $w_a = \varphi(w_a)$, here it means that $\nu_{w_a} = f(\nu_{w_a})$, and so w_a is a fixed point of f starting with a number from I_a which is a fixed point of $f_{a,0}$, the first applied interval morphism. Since $f_{a,0}$ is an affine function with the slope $1/\theta < 1$, this fixed point is unique. □

Example 12. The sequence ν_{tm}, $\nu_{tm}[k] \in X_{tm}$, corresponding to the Thue-Morse word w_{tm} starting with a, is the fixed point starting with $1/2^-$ of the morphism (3):

$$1/2^-, 1, 3/4^-, 1/4^-, 5/8^-, 1/8^-, 3/8^-, 7/8^-, \cdots.$$

The other fixed point w'_{tm} of φ_{tm}, starting with b, corresponds to the fixed point of the same morphism (3) starting with $1/2^+$:

$$1/2^+, 0, 1/4^+, 3/4^+, 3/8^+, 7/8^+, 5/8^+, 1/8^+, \cdots.$$

Compared to Example 1, we see that the fixed points here differ from the original definitions of ν_{tm} and $\nu_{tm'}$ by signs $-$ or $+$ added to numbers.

Remark 13. The set Z existing in X_{tm} in two copies, Z^+ and Z^-, contains not only binary rationals from the fixed points above. As another example, consider the number $1/6$. We have $1/6 = f_{b,1}(f_{a,1}(1/6))$, and thus $1/6$ corresponds to words au starting from a and satisfying $au = \sigma(\varphi_{tm}(\sigma(\varphi_{tm}(au)))) = \sigma^3(\varphi^2_{tm}(au)) = a\varphi^2_{tm}(u)$. This equation has two solutions $aw_{tm} < aw'_{tm}$. So, we have $\nu(aw_{tm}) = 1/6^-$ and $\nu(aw'_{tm}) = 1/6^+$. Since $1/6$ is the frequency of aa in the Thue-Morse word, we see that aw_{tm} is the maximal element of L_{tm} starting from aa and aw'_{tm} is the minimal element of L_{tm} starting from ab.

Other examples, starting with the Fibonacci morphism, can be treated with the software tool described below in Section 8 and available online.

6 Morphism f and k-regular sequences

Let φ be a primitive order-preserving morphism, w be some of its fixed points, and L be the separable subshift generated by w. Since the morphism is primitive, the subshift L does not depend on the choice of the fixed point of φ and is minimal.

Consider the morphism $f : X_L^* \mapsto X_L^*$ described above and the sequence ν_w which is its fixed point corresponding to w. As we have discussed above, each mapping $f_{a,p}(x)$ from the definition of f is an affine mapping sending the interval I_a of length μ_a to the interval $J_{a,p}$ of length μ_a/θ, where θ is the Perron-Frobenius eigenvalue of φ. So, in the definition (2) of $f_{a,p}$, we have $(y_2 - y_1)/(x_2 - x_1) = 1/\theta$. We get the following

Corollary 14. *Under the conditions of Theorem A, the mappings $f_{a,p}$ from the definition of the morphism f are of the form*

$$f_{a,p}(x) = x/\theta + C_{a,p}, \tag{4}$$

where θ is the Perron-Frobenius eigenvalue of φ and $C_{a,p}$ is a constant defined, in the notation of (2), by

$$C_{a,p} = y_1 - x_1/\theta.$$

This statement is particularly interesting when the morphism φ is k-uniform, that is, the length of all φ-images of letters is the same and equal to $k \geq 2$. Since θ is the limit of the ratio $|\varphi(w[0..n])|/n$, here we have $\theta = k$. The word w is k-automatic (for the definition and discussion on k-automatic words and k-regular sequences, the reader is referred to [1, 2]).

Lemma 15. *Under the conditions of Theorem A, if the morphism φ is k-uniform, then the sequence ν_w is k-regular.*

Proof. First of all, the morphism f is also k-uniform. With the definition (4) of each mapping $f_{a,p}$, we can write the following expression for an element $\nu_w[kn+p]$ of the sequence $(\nu_w[n])_{n=0}^{\infty}$, where $n \geq 0$ and $p \in \{0, \dots, k-1\}$:

$$\begin{aligned}
\nu_w[kn+p] &= \frac{1}{k}\nu_w[n] + C_{w[n],p} \\
&= \frac{1}{k}\nu_w[n] + \sum_{q=0}^{k-1}\sum_{a\in\Sigma} C_{a,q}X(w[n]=a)X(q=p).
\end{aligned}$$

Here $X(P)$ is the characteristic sequence of a property P, equal to 1 if the property holds and to 0 otherwise.

The sequence w is k-automatic and thus k-regular, as well as the sequences $X(w[n]=a)$ and $X(w[n]=a)X(q=p)$. So, the sequence ν_w is also k-regular by Theorem 16.1.3 from [2]. □

7 The binary case

In this section, we adapt Theorem A to all binary morphisms with aperiodic uniformly recurrent fixed points. To do it, we first discuss what may happen. In this section, we consider the alphabet $\Sigma_2 = \{a, b\}$ with $a < b$ and a morphism $\varphi : \Sigma_2^* \mapsto \Sigma_2^*$.

7.1 Non-primitive case

First of all, let us discuss the condition of the morphism φ to be primitive. In fact, everywhere in the proof we used not the primitivity itself but the facts that the subshift L_φ is uniquely ergodic, and for any finite factor u of L_φ, the relation $|\varphi(u)|/|u|$ has a limit, denoted θ with $|u| \to \infty$. We also need the subshift to be aperiodic.

If φ is primitive, all the conditions except for perhaps aperiodicity hold. Consider the case of non-primitive φ. It may have several fixed points with different orbit closures, but without loss of generality, suppose that φ has an infinite fixed point w starting with $\varphi(a)$ in which both letters appear. The condition that φ is not primitive means then that $\varphi(b) \in b^*$.

Proposition 16. *The fixed point w and its orbit closure can be minimal and aperiodic only if $\varphi(b) = b$ and $\varphi(a) = axa$, where x is a finite word containing b. In this case, w is also uniquely ergodic, and there exists a limit*

$$\lim_{u\in\mathrm{Fac(w)}, |u|\to\infty} \frac{|\varphi(u)|}{|u|} = \theta. \tag{5}$$

Proof. If $\varphi(b)$ is the empty word, then $w = \varphi(a)^\omega$ is periodic. If $\varphi(b) = b^k$ with $k \geq 2$, then L_w is not minimal since contains b^ω. The same is true if $\varphi(b) = b$ and $\varphi(a)$ ends with b. So, the only case when L_w is minimal and can be aperiodic is $\varphi(b) = b$ and $\varphi(a) = axa$, where x is a word containing b. In this case, w is uniformly recurrent: indeed, the distance between two consecutive occurences of a is bounded by $|x| + 1$ and thus the distance between two consecutive occurrences of any factor of $\varphi^n(a)$ is bounded by $|\varphi^n(ax)|$. Moreover, due to Theorem 3 of [10], w is a primitive morphic sequence. In particular, it is uniquely ergodic, and the limit (5) exists. □

Note that in this case, the word w and its subshift can be periodic if $\varphi(a) = (ab)^k a$ for some k. In other cases, we can work with φ, w and L_w exactly as if φ were primitive.

Example 17. If $\varphi(a) = aaba, \varphi(b) = b$, the minimal subshift L_w generated by w is the orbit closure of the fixed point w of φ starting with a:

$$w = aabaaababaabaaabaaababaaba \cdots .$$

We can work with L_w exactly as if φ were primitive.

7.2 Preserving order

How can we work with a morphism which is not order-preserving? The following proposition shows that in the binary case, we can just replace it by its square, as we did in Example 4 for the Fibonacci morphism.

Proposition 18. *For every binary morphism φ with an aperiodic subshift, if φ is not order-preserving, then φ^2 is.*

Proof. Borchert and Rampersad proved (see Theorem 15 in [9]) that every aperiodic binary morphism is either order-preserving or order-reversing, the latter property meaning that $\varphi(u) > \varphi(v)$ whenever $u < v$ for infinite words u, v. Clearly, the square of an order-reversing morphism is order-preserving. □

So, if by chance a binary morphism is not order-preserving, its square is, and we can consider the subshift as generated by the square morphism. Now let us consider two different types of inseparability.

7.3 Common suffixes

Here we consider the situation when the morphism φ is not typable because of a non-empty common suffix of images of letters, like in Example 6. Since we consider only the minimal (or, which is the same, uniformly recurrent) aperiodic case, we can use Proposition 16 and see that if φ is not primitive, the common suffix is empty. So, in our case, the morphism φ is primitive.

The following classical statement will be useful. We give its proof for the sake of completeness.

Proposition 19. *Suppose that $\varphi : a \to p_a s, b \to p_b s$, where s is any common suffix of $\varphi(a)$ and $\varphi(b)$. Then $L_\varphi = L_{\varphi'}$, where $\varphi' : a \to s p_a, b \to s p_b$.*

Proof. Clearly, both subshifts generated by φ and φ' contain the word s and are closed under the operation sending a word $u = u_1 u_2 \cdots u_n$ to the word $p_{u_1} s p_{u_2} s \cdots s p_{u_n}$. So, the intersection of the two sets of factors is infinite, and since both subshifts are minimal, they are equal. □

Example 20. For the morphism $\varphi : a \to aab, b \to abb$ from Example 6, we have $\varphi' : a \to baa, b \to bab$.

Clearly, if one of the images of symbols is not a suffix of the other, it is sufficient to apply Proposition 19 once to get a new morphism, which is also primitive and with images of letters ending with different symbols, like in the previous example. If it is not the case, however, it can be necessary to apply the operation from Proposition 19 several times:

Example 21. If $\varphi : a \to ab, b \to babab$, then, to get a morphism with images of letters ending by different symbols, we follow three steps:

$$\varphi : \begin{cases} a \to ab \\ b \to babab \end{cases} \quad \to \quad \varphi' : \begin{cases} a \to ab \\ b \to abbab \end{cases}$$

$$\to \quad \varphi'' : \begin{cases} a \to ab \\ b \to ababb \end{cases} \quad \to \varphi''' : \begin{cases} a \to ba \\ b \to babab. \end{cases}$$

Note that in terms used in [26], φ' is a *left conjugate* of φ; for conjugacy of morphisms see also [28]. The next proposition follows from results of [26], but it does not take more space to give a new proof than to explain the relationship between used terminology. The meaning of the proposition is that the number of necessary steps is always finite.

Proposition 22. *Let φ be a primitive binary morphism with aperiodic subshift and, without loss of generality, $|\varphi(a)| \geq |\varphi(b)|$. Then after applying the operation from Proposition 19 with the maximal common suffix at most k times, where $k \leq \lfloor |\varphi(a)|/|\varphi(b)| \rfloor + 1$, we get a morphism whose images of letters end by different symbols.*

Proof. If one image of letter is not a suffix of the other, the statement is obvious and $k = 1$ is enough. Suppose now the opposite: let $\varphi(a) = ps, \varphi(b) = s$. Suppose that we can continue to exchange the prefix and the suffix s of the image of a at least $(|\varphi(a)| + |\varphi(b)|)/|\varphi(b)|$ times. It would mean that the word $sps = \varphi(ba)$ is periodic with the period $|s|$. But it is also periodic with the period $|ps|$, so, due to the Fine and Wilf theorem, it is periodic with the period $\gcd(|s|, |ps|)$. In particular, it means that $\varphi(a) = ps$ and $\varphi(b) = s$ are powers of the same word, and thus the subshift generated by φ is periodic, a contradiction. So, if k is the maximal number of replaced suffixes, we have $k < (|\varphi(a)| + |\varphi(b)|)/|\varphi(b)|$, which means $k \leq \lfloor |\varphi(a)|/|\varphi(b)| \rfloor + 1$. □

Example 23. Note that nevertheless, successive transfers of the longest common suffixes to the left can touch more symbols than there are in the longer image of a letter. For example, consider $\varphi : a \to abbab, b \to bab$; then

$$\varphi : \begin{cases} a \to abbab \\ b \to bab \end{cases} \quad \to \varphi' : \begin{cases} a \to babab \\ b \to bab \end{cases} \quad \to \varphi'' : \begin{cases} a \to babba \\ b \to bab. \end{cases}$$

Here the longer image of a letter is of length 5, and there are 3+3=6 letters replaced.

So, we may assume that given a primitive binary morphism φ, we can always transfer common suffixes of images of letters to the left until we get a morphism $\psi = \varphi^{(k)}$ with the same subshift and with images of letters ending by different symbols. To justify this passage completely, we should also describe how we can apply Theorem A to φ if we know how to do it for ψ. The following proposition gives a recipe for that.

Proposition 24. *Suppose that a binary morphism φ is transformed to another morphism ψ by a series of transfers of common suffixes to the left: $\varphi \to \varphi' \to \cdots \to \varphi^{(k)} = \psi$, where the suffix transferred at the passage from $\varphi^{(i)}$ to $\varphi^{(i+1)}$ is of length p_{i+1}. Then for every infinite word w, we have $\psi(w) = \pi\varphi(w)$, where the word π of length $p = p_1 + \cdots + p_k$ is the concatenation of all replaced common suffixes in order from right to left.*

Proof. For each step i, it is not difficult to see that $\varphi^{(i+1)}(w) = \pi_{i+1}\varphi^{(i)}(w)$, where π_{i+1} is the replaced common suffix of length p_{i+1}. It remains to combine these arguments for all i and to set $\pi = \pi_k \cdots \pi_2\pi_1$. □

Corollary 25. *The morphism φ is order-preserving if and only if ψ is order-preserving.*

This corollary means just that we can successively take a square of our morphism if it is necessary to make it order-preserving, and then transfer common suffixes as we need.

Example 26. Consider the fixed point $w = aabaababb\cdots$ of the morphism $\varphi : a \to aab, b \to abb$ from Examples 6 and 20. To find the value $\nu(w)$ and all the sequence ν_w, we pass to the morphism $\psi = \varphi' : a \to baa, b \to bab$. The morphism ψ falls into conditions of Theorem A, and gives rise to the the following morphism on intervals:

$$\begin{aligned} f(x) &= \begin{cases} f_{a,0}(x), f_{a,1}(x), f_{a,2}(x) \text{ for } x \in [0, 1/2^-] \\ f_{b,0}(x), f_{b,1}(x), f_{b,2}(x) \text{ for } x \in [1/2^+, 1] \end{cases} \\ &= \begin{cases} x/3 + 1/2, x/3, x/3 + 1/6 \text{ for } x \in [0, 1/2^-] \\ x/3 + 1/2, x/3 + 1/6, x/3 + 2/3 \text{ for } x \in [1/2^+, 1]. \end{cases} \end{aligned}$$

The only fixed point $v = babbaabab\cdots$ of ψ corresponds to the only fixed point ν_v of f starting with the fixed point $\nu(v) = 3/4$ of $f_{b,0}(x) = x/3 + 1/2$: $\nu_v = 3/4, 5/12, 11/12, 23/36, \ldots$. At the same time, the fixed point w of φ due to the previous proposition satisfies $w = \varphi(w) = \sigma(\psi(w))$, and due to Proposition 9 and Theorem A, corresponds to the fixed point ν_w of $\sigma(f)$: $\nu_w = \sigma(f(\nu_w))$. In particular, it starts with the fixed point $\nu(w) = 0$ of $f_{a,1}(x) = x/3$: $\nu_w = 0, 1/6^+, 10/18^+, 1/18^+, 2/9^+, \cdots$. Here we have to add pluses to values since these are lower ends of intervals.

Example 27. Consider the morphism $\varphi : a \to ab, b \to babab$ from Example 21. After moving $p = 2 + 2 + 1 = 5$ symbols from right to left, we get an order-preserving separable morphism $\psi = \varphi''' : a \to ba, b \to babab$ inducing the same subshift L.

Note that by the construction, for every $u \in L$, we have $\varphi(u) = \sigma^5(\psi(u))$. In particular, it is true for both fixed points w_a and w_b of φ. Let us start with the fixed point $w = w_a$ starting with a:

$$w = ab.babab.babab.ab.babab.ab.babab.babab.\ldots = \varphi(w) = \sigma^5(\psi(w))$$

(dots are put between φ-images of symbols for readability). Since w starts with a and $\psi(a)$ is of length 2, we have

$$w = \sigma^5(\psi(w)) = \sigma^3(\psi(\sigma(w))).$$

Passing to the sequences ν, we see that

$$\nu_w = \sigma^3(f(\sigma(\nu_w))),$$

where f is the morphism on extended intervals corresponding to ψ. Denote $\nu_w = \nu[1]\nu[2]\nu[3]\cdots$. Here $\nu[1] \in I_a$, $\nu[2], \nu[3] \in I_b$ and so on along the word w. Then $\sigma(\nu_w)) = \nu[2]\nu[3]\nu[4]\cdots$, and

$$f(\sigma(\nu_w)) = f_{b,0}(\nu[2])f_{b,1}(\nu[2])\cdots f_{b,4}(\nu[2])f_{b,0}(\nu[3])f_{b,1}(\nu[3])\cdots$$

At last, applying σ^3, we see that

$$\nu_w = \nu[1]\nu[2]\nu[3]\nu[4]\cdots = f_{b,3}(\nu[2])f_{b,4}(\nu[2])f_{b,0}(\nu[3])f_{b,1}(\nu[3])\cdots$$

So, the number $\nu[2]$ can be reconstructed as the fixed point of the mapping $f_{b,4}$, and $\nu[3]$ as the fixed point of the mapping $f_{b,0}$. All the other elements of ν_w, including $\nu[1]$, can be computed one-by-one as functions of previously known values. In particular, $\nu[1] = f_{b,3}(\nu[2])$.

For the other fixed point

$$w = w_b = babab.ab.babab.ab.babab.ab.babab.babab.\dots = \varphi(w) = \sigma^5(\psi(w)),$$

we also have $w = \sigma^5(\psi(w))$, but since the length of the image of the first symbol b is 5, it means just that

$$w = \psi(\sigma(w)).$$

For the sequence ν, it means that

$$\nu_w = f(\sigma(\nu_w)).$$

So, $\nu[2]$ is the fixed point of $f_{a,1}$, $\nu[3]$ is the fixed point of $f_{b,0}$, and all the other numbers of the sequence ν can be found starting from them.

The same idea can be used for every morphism obtained from a "good" one by transferring common prefixes of images of symbols to the right: the needed values can be reconstructed from one or several fixed points of mappings $f_{x,i}$.

7.4 Inseparable types

In this subsection, we propose a method to avoid the situation described in Example 7, when the types of elements of the orbit of w are well-defined but cannot be ordered since the relations between words of two given types can be different. We shall show that it happens because of prefixes of these words of bounded length, which can be classified and considered as symbols of a new larger alphabet. The sequence on this new alphabet will inherit all good properties of w and will be separable.

As we have seen above, we may restrict ourselves to a binary order-preserving morphism φ such that the last symbols of $\varphi(a)$ and $\varphi(b)$ are different, and the subshift of φ is minimal with aperiodic fixed points. Due to Proposition 16, the morphism φ is either primitive or of the form $\varphi(a) = axa$, $\varphi(b) = b$ for a finite word x containing b.

To discuss the subject, we have to introduce yet another property of morphic subshifts called *circularity*. There exist several very close definitions of this property discussed in particular in [18]; we shall use the following one. The fixed point w of a morphism φ (and the whole subshift L_φ) are called *circular* if there exists a positive constant D called a *synchronization delay* such that in any factor u of w (L_φ) of length at least D, there exists a *synchronization point*. Here a synchronization point is a place in u where in any occurrence of u to w (L_φ) there is a boundary between images of two symbols: $u = ps$, where p is a suffix of $\varphi(p')$, s is a prefix of $\varphi(s')$, $p's'$ is a factor of w (L_φ).

The smallest value of a synchronization delay can be called *the* synchronization delay.

Example 28. The Thue-Morse word w_{tm} is circular with $D \leq 5$. Indeed, each factor of w_{tm} of length 5 contains one of factors aa or bb, and thus a synchronization point between two letters a or two letters b. For example, $aabba$ contains two synchronization points, after the first and the third symbol, and appears in w_{tm} only as a suffix of $\varphi_{tm}(bab)$.

The Sierpinski morphism $a \to aba, b \to bbb$ is not circular since for all n, the word b^n appears in it and has no synchronization points: the boundaries between images of b in it can pass anywhere.

Note that in our case, when the last letters of images of symbols are all different, a synchronization point $u = ps$ determines a unique decomposition to images of symbols of the whole preceeding prefix p of u: we reconstruct it from right to left taking each time the image of symbol ending by the given last letter. At the same time, the suffix s, if it is short, may leave some ambiguity if s is the prefix of both images of letters or, in the case of $\varphi(a)$ prefix of $\varphi(b)$ (or vice versa), s is a prefix of $\varphi(ab)$ and of $\varphi(ba)$.

It is well-known that a fixed point of a primitive morphism is circular [24, 18]. It is also not difficult to extract from the main result of [17] and Theorem 12 from [18] that the non-primitive fixed points from Proposition 16 are also circular. So, all morphisms we consider are circular. To be accurate, we redefine the (smallest) synchronization delay D so that, in addition to the main property, each word of length D or more contains both letters: since w is uniformly recurrent, there is no problems with that. We need it to have $|\varphi(u)| \geq D + m - 1$ for all u with $|u| \geq D$, where m is the maximum of $|\varphi(a)|, |\varphi(b)|$.

Now let us define another morphism over a greater alphabet preserving all good properties of φ and with separable types. To do it, we consider the alphabet A_D of all factors of w of length D and define the trivial isomorphism π :$\mathrm{Fac}_D(w) \mapsto A_D$ which can be naturally extended to π :$\mathrm{Fac}_{D+n}(w) \mapsto A_D^{n+1}$ for all n and to $\pi : L_w \mapsto A_D^{\omega}$, by $\pi(x_1 \cdots x_D y) = \pi(x_1 \cdots x_D)\pi(x_2 \cdots x_D y)$ for all letters x_i and all words y. Clearly, π commutes with the shift σ and thus we can consider the subshift $(\pi(L_w), \sigma)$. Moreover, in addition to $\pi^{-1} : A_D \mapsto \mathrm{Fac}_D(w)$, it is reasonable to consider a simpler mapping $\rho : A_D \mapsto \Sigma_2$, where for each $a \in A_D$, the symbol $\rho(a)$ is its first symbol. The alphabet A_D and words over it inherit the lexicographic order on Σ_2.

Now, given a morphism $\varphi : \Sigma_2^* \mapsto \Sigma_2^*$, let us define the morphism $\chi : A_D^* \mapsto A_D^*$ as follows: for all $a \in A_D$ such that $a = \pi(u)$ and $\rho(a) = x$ (so that x is the first symbol of u), the image $\chi(a)$ is defined as the first $|\varphi(x)|$ symbols of $\pi(\varphi(u)) = \pi(\varphi(\pi^{-1}(a)))$. This mapping is well-defined since by the definition of D, we have $|\varphi(u)| \geq D + |\varphi(x)| - 1$.

Example 29. Let us continue Example 7 (the subshift L_φ defined by the morphism $\varphi : a \to aabab$, $b \to bba$) and define the respective morphism χ. To do it, we first observe that the synchronization delay of L_φ is 5: indeed, the longest word without the synchronization point is $babb$ which can be decomposed both as the factor of $\varphi(ab)$ without the prefix and the suffix of length 2 each, and the factor of $\varphi(bb)$ without the prefix and the suffix of length 1 each. Its continuations $babba$ and $babbb$ disambiguate the situation.

The set $\mathrm{Fac}_5(w)$ is of cardinality 17: in the lexicographic order, $\mathrm{Fac}_5(L_\varphi) = \{aaaba, aabab, abaab, ababa, ababb, ab$
$baaab, baaba, babaa,$
$babba, babbb, bbaaa, bbabb, bbbaa, bbbab\}$. We denote the elements of the alphabet A_5, in the same order, as $A_5 = \{a_1, \ldots, a_8, b_1, \ldots, b_9\}$: so, $\pi(aaaba) = a_1$ and so on till $\pi(bbbab) = b_9$. Note that by the notation, $\rho(a_i) = a$ and $\rho(b_j) = b$ for all well-defined i and j.

To define χ, we take the φ-images of words from $\mathrm{Fac}_5(L_\varphi)$ and then their prefixes of length 9 for words starting from a and of length 7 for words starting from b. Then we take π-images of these words. For example, to find $\chi(a_1)$, we take $\varphi(aaaba) = aababaababaababbbaaabab$, then its prefix of length 9 which is $aababaaba$, then its π-image $a_2a_4b_3a_3b_2$. So, $\chi(a_1) = a_2a_4b_3a_3b_2$, and continuing the same method, we get

$$\chi : \begin{cases} a_1, a_2 \to a_2a_4b_3a_3b_2, \\ a_3, a_4, a_5 \to a_2a_5b_5a_8b_8, \\ a_6, a_7, a_8 \to a_2a_5b_5a_8b_9, \\ b_1, b_2, b_3, b_4, b_5 \to b_6b_1a_1, \\ b_6, b_7 \to b_7b_4a_6, \\ b_8, b_9 \to b_7b_4a_7. \end{cases}$$

Note that for all $x \in A_5$, $\varphi(\rho(x)) = \rho(\chi(x))$, so, $\varphi \circ \rho = \rho \circ \chi$. In particular, the two fixed points of

φ, starting from a and from b, are ρ-images of the two fixed points of χ, starting from a_2 and b_7.

The next proposition, following directly from the construction, claims that this is a general situation.

Proposition 30. *1. If the morphism φ is order-preserving on its subshift, then so is χ.*

2. *If φ-images of letters end by different symbols, then different χ-images of symbols of A_D end by different letters of A_D. Moreover, these last letters do not occur anywhere else in χ-images of symbols.*

3. *There are as many fixed points of χ as of φ, and any fixed point of φ can be obtained as the ρ-image of a fixed point of χ.*

4. *If φ-images of letters end by different symbols, then ρ is an isomorphism between (L_χ, σ) and (L_φ, σ), and moreover, it preserves the lexicographic order of words on the respective subshifts.*

5. *If φ-images of letters end by different symbols, then the subshift (L_χ, σ) is separable.*

Proof. The first four properties follow directly from the construction. It remains to show the separability of (L_χ, σ). First let us prove that this subshift is typable. Indeed, for any representation $u = \sigma^p(\chi(u'))$, where $u' \in L_\chi$ and $0 \leq p < |\chi(u'[0])|$, consider the ρ-images $v = \rho(u) \in L_\varphi$ and $v' = \rho(u') \in L_\varphi$. Clearly, $v = \sigma^p\varphi(v')$. The morphism φ is circular, the last symbols of two images of letters are different, which means that any synchronization point determines all preceeding synchronization points, and so the type of v is well-defined: it is $(v'[0], p)$. But $u = \pi(v)$ and $u' = \pi(v')$ since v and v' belong to L_φ and by the definition of π, so, since v' and p are unique, so is u' (and the same p). The type of u is thus well-defined as $(u'[0], p)$. Note that we uniquely reconstruct $u'[0]$ even if there are several symbols in A_D with the same χ-image, like in the example above.

It remains to prove that the types in L_χ are comparable. Consider two elements $u_1, u_2 \in L_\chi$ of different types (x_1, p_1) and (x_2, p_2). If their first symbols (uniquely defined by types) are different, the order is determined by them. If by contrary the first symbols are the same, let us compare the suffixes of $\chi(x_1)$ and $\chi(x_2)$ which are prefixes of u_1 and u_2. If $\chi(x_1) \neq \chi(x_2)$, then, by a previous property, the last symbols of these two images are different and do not appear anywhere else in χ-images of letters. So, the order between u_1 and u_2 is again uniquely determined by the prefixes of u_1 and u_2 which are suffixes of $\chi(x_1)$ and $\chi(x_2)$. If $\chi(x_1) = \chi(x_2)$ but $p_1 \neq p_2$, the same argument holds. At last, if $\chi(x_1) = \chi(x_2)$, $p_1 = p_2$, but $x_1 \neq x_2$, we have $u_1 < u_2$ if and only if $x_1 < x_2$ since the morphism χ is order-preserving on L_χ. This completes the proof of separability of the subshift (L_χ, σ). □

This construction completes the algorithm allowing to treat every binary pure morphic uniquely ergodic subshift (L_φ, σ). First, if the morphism φ is not order-preserving, we pass to its square according to Proposition 18. Then, if the resulting morphism is not typable because of common suffixes of images of letters, we transfer these common suffixes to to the left as many times as needed due to Proposition 22. At last, we find a synchronization delay and construct the morphism χ on the extended alphabet which, due to Proposition 30, has all the desired properties. So, the morphism f constructed as described in the beginning of Section 5 due to Theorem A gives a equidistributed morphic subshift on the interval $[0, 1]$. Again due to propositions from this section, this is the same subshift as it would be for the initial morphism φ instead of χ. The value and the infinite sequence corresponding to a fixed point of the initial morphism, we can use the method described in Examples 26 and 27.

Remark 31. Note that the restriction to the binary alphabet is crucial. On the three-letter alphabet, it is easy to construct a morphism which does not become order-preserving even when we consider its powers:

$$g : a \to ac, b \to ab, c \to cb.$$

It can be easily seen that $g^n(b) < g^n(a)$ for all $n \geq 1$. Moreover, it is not possible to transfer the common suffix b of $g(b)$ and $g(c)$ to the left since $g(a)$ does not end with b. So, we see two technical problems in one example. The extension of the result to general morphisms on larger alphabets is thus an open problem.

8 Computational tool

We conclude the paper by a presentation of the software which, given a binary morphism φ with an aperiodic uniformly recurrent fixed point, computes the respective morphism f on numbers and first l elements of the sequence corresponding to each of the fixed points of φ. The code is available at
`https://www.i2m.univ-amu.fr/perso/anna.frid/MorphismsOnReals/mp.py`.
A web page where the computation can be done online for a relatively small input (images of letters not longer than about 5 letters, if the morphism is not order-preserving, or about 25 letters, if it is) is
`https://www.i2m.univ-amu.fr/perso/anna.frid/MorphismsOnReals/mp.html`.

1) Given a binary morphism φ, we first check if it is primitive: in the binary case, it is clearly sufficient to check if the square of its matrix is positive. If the morphism is not primitive, we continue to consider it if it falls into the case of Proposition 16. Then we check if φ admits a fixed point starting with each letter, that is, if $\varphi(x)$ starts with x for some letter x. To check fixed points for aperiodicity, we use the result of [27] and eliminate periodic fixed points corresponding $\varphi(a) = a(ba)^m$ and $\varphi(b) = b(ab)^n$ for some $m, n \geq 0$ such that $m + n \geq 1$, and those whose images are powers of a common word of length at least 2. Among the non-primitive morphisms from Proposition 16, we eliminate those of the form $\varphi(a) = a(b^m a)^n$ and $\varphi(b) = b$ (and of course the symmetric case).

2) In the primitive case, we check if the morphism is order-preserving. If one image of a letter is not a prefix of the other one, the check is straighforward; if it is, we can transfer the common prefix to the end of both images until they can be directly compared. Corollary 25 assures that the property of being order-preserving is stable under this operation. If the morphism is not order-preserving, we pass to its square due to Proposition 18, even though it considerably increases the complexity of the computation. Then we check if the two images of letters have a common suffix, at if it is the case, we transfer it to the beginning of the images as it was described in Proposition 19. Due to further results of Subsection 7.3, it is sufficient to repeat this procedure a finite number of times, and the resulting morphism ψ remains order-preserving.

3) The morphism μ considered at this stage (here μ can be the initial φ, or its square, or the ψ obtained from φ or φ^2 by transferring common suffixes to the left) is circular. Now we have to compute a synchronization delay D of μ to check if μ is separable; if it is not, we will have to use yet another morphism on a greater alphabet. This is a slow part of the computation, especially as in the general case, there is no known upper bound for D. We only know from a recent paper by Klouda and Medková [16] that for a uniform binary morphism, D is bounded by m^3, where m is the morphism length. So, in the general case, we unfortunately do not have an upper bound for the complexity of the following procedure.

First, we find all factors of length 2 of the subshift as follows: starting from the set A of all factors of length 2 of images of letters, we expand A while there are new words of length 2 situated at the boundary between images of letters in words $\mu(a)$, $a \in A$. Clearly, as soon as there are no new words of length 2 obtained like that, the set of factors of length 2 of the subshift L is complete.

Now, starting with the set of factors of length 2, we use the following fact: If m is the shorter length of an image of a letter, then every factor of L of length $ml + 1$ is contained in a μ-image of a factor of L of length $l + 1$. If $m \geq 2$, this fact is sufficient to find all factors of L together with their types for any given length greater than 2. If $m = 1$ and μ is primitive, we can pass to its square to get $m > 1$. At last, for the non-primitive case of $\varphi(a) = axa$, $\varphi(b) = b$, we need the following two lemmas which allow to initialize and continue the process of finding all factors of any given length.

Lemma 32. *Let φ be a binary morphism such that $\varphi(a) = axa$ and $\varphi(b) = b$, where x is a finite word containing b. If we denote by h_b the highest power of b appearing in x, then the factors of length $h_b + 2$ of L_φ are exactly those of $\varphi^2(a)$.*

Proof. Any factor of $\varphi^2(a)$ is also a factor of L_φ. Conversely, let u be a factor of length $h_b + 2$ of the subshift. Then u contains an occurrence of a and thus any its given occurrence has common letters with an occurrence of $\varphi(a)$. If it is contained in $\varphi(a)$, the statement is proved; if not, since $|\varphi(a)| \geq h_b + 2$, the word u overruns $\varphi(a)$ from at most one side, so, it is a factor of $\varphi(a)b^k\varphi(a)$ for some $k \leq h_p$. In particular, $\varphi(a)b^k\varphi(a)$ is a factor of the subshift, and for that, its preimage ab^ka had to appear, at some point, in $\varphi(a)$. So, $\varphi(a)b^k\varphi(a)$ and its factor u are factors of $\varphi(\varphi(a))$. □

Lemma 33. *Let φ be a binary morphism such that $\varphi(a) = axa$ and $\varphi(b) = b$, where x is a finite word containing b. If we denote by h_b the highest power of b appearing in x, then for every $l \geq h_b+2$, every factor of L_φ of length less than $p(l)$ is a factor of some $\varphi(y)$, where y is a factor of L_φ of length at most l. Furthermore, we have $p(l) > l$. Here $p(l)$ is defined by*

$$p(l) = 1 + q(h_b + |\varphi(a)|) + r,$$

where q and r stand for the quotient and the rest in the euclidean division of $l - 1$ by $h_b + 1$.

Proof. The shortest word that can be written by concatenating $l - 1$ φ-images of letters and which does not contain the word b^{h_b+1} is $(\varphi(b)^{h_b}\varphi(a))^q\varphi(b)^r$ of length $q(h_b + |\varphi(a)|) + r$. As a consequence, the length of the shortest factors of L_φ that lie on $l + 1$ (and not less) φ-images of letters is at least $q(h_b + |\varphi(a)|) + r + 2 = p(l) + 1$, meaning that all factors of length $p(l)$ or less lie on at most l φ-images of letters. It is easy to check that for $l \geq h_b + 2$ we have $p(l) > l$. □

So, in each situation, we can find all factors of any given length together with their types. A synchronization delay is reached as soon as every word appears in the list with only one type. It is reasonable to check all the shorter lengths and find *the* smallest synchronization delay D not an upper bound for it, to consider a smaller alphabet and to have a nicer output (and probably to gain in computation time).

4) We have obtained the set of factors of length D, and each of them corresponds to a type. However, the same type may correspond to several words, and if they are not lexicographically consecutive, the morphism is not separable. If it is the case, we pass to a morphism χ on a larger alphabet as described in Subsection 7.4.

5) The morphism considered at this stage satisfies the conditions of Theorem A and so we construct the morphism on intervals as described in Section 5.

6) To find numeric sequences corresponding to fixed points of the initial morphism φ, we start with finding fixed points of respective mappings $f_{c,p}$ as described in Examples 1, 26, 27.

At last, note that the Perron-Frobenius eigenvalue is an algebraic number, making it possible for a mathematical software to do exact computations at each stage, and to print the outputs with arbitrary precision.

As we have discussed, the algorithm we use is not very fast. First, there is no general upper bound for the synchronization delay D, and at the same time, computing D is the slowest part of the process. To avoid it, it would be nice to invent a faster way to check the subshift for separability. It would be also helpful to learn how to deal directly with order-reversing morphisms, since taking the morphism square before looking for separability slows down this slowest part of computation. We leave these questions to further research.

References

[1] J.-P. Allouche, J. Shallit, Automatic sequences — theory, applications, generalizations. Cambridge University Press, 2003.

[2] J.-P. Allouche, J. Shallit, The ring of k-regular sequences, Theoret. Comput. Sci. 98 (1992), 163–197.

[3] J.-P. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, Sequences and their Applications, Discrete Mathematics and Theoretical Computer Science, Springer, London, 1999. P. 1–16.

[4] J. Amigó, Permutation Complexity in Dynamical Systems - Ordinal Patterns, Permutation Entropy and All That. Springer Series in Synergetics, 2010.

[5] S. V. Avgustinovich, A. Frid, S. Puzynina, Canonical representatives of morphic permutations, Proc. WORDS 2015. LNCS V. 9304, 2015, 59–72.

[6] S. V. Avgustinovich, A. Frid, S. Puzynina, Minimal complexity of equidistributed infinite permutations, European J. Combin. 65 (2017) 24–36.

[7] C. Bandt, G. Keller and B. Pompe, Entropy of interval maps via permutations, Nonlinearity 15 (2002), 1595–1602.

[8] J. Berstel, Axel Thue's Papers on Repetitions in Words: a Translation, Publications du Laboratoire de Combinatoire et d'Informatique Mathématique 20, Université du Québec à Montréal, 1995.

[9] A. Borchert, N. Rampersad, Permutation complexity of images of Sturmian words by marked morphisms, Discrete Mathematics & Theoretical Computer Science 20:1 (2018), 9pp.

[10] F. Durand, Decidability of uniform recurrence of morphic sequences, Int. J. Found. Comput. Sci. 24 (2013) 123–146.

[11] S. Elizalde, The number of permutations realized by a shift, SIAM J. Discrete Math. 23 (2009), 765–786.

[12] S. Ferenczi, T. Monteil. Infinite words with uniform frequencies, and invariant measures. Combinatorics, automata and number theory, 373–409, Encyclopedia Math. Appl., 135, Cambridge Univ. Press, Cambridge, 2010.

[13] D. G. Fon-Der-Flaass, A. E. Frid, On periodicity and low complexity of infinite permutations, European J. Combin. 28 (2007), 2106–2114.

[14] A. Frid. On the frequency of factors in a D0L word. J. Autom. Lang. Comb. 3 (1998), 29–41.

[15] K. Johnson, Beta-shift dynamical systems and their associated languages, Ph.D. Thesis, University of North Carolina at Chapel Hill, 1999.

[16] K. Klouda, K. Medková, Synchronizing delay for binary uniform morphisms. Theoret. Comput. Sci. 615 (2016) 12–22.

[17] K. Klouda, Š. Starosta, An algorithm for enumerating all infinite repetitions in a D0L-system. J. Discrete Algorithms 33 (2015), 130–138.

[18] K. Klouda, Š. Starosta, Characterization of circular D0L systems. Theoret. Comput. Sci. 790 (2019), 131–137.

[19] L.-M. Lopez, Ph. Narbel, Infinite interval exchange transformations from shifts, Ergod. Th. & Dynam. Sys. 37 (2017), 1935–1965.

[20] Lothaire, M.: Algebraic combinatorics on words. Cambridge University Press, 2002.

[21] M. Makarov, On permutations generated by infinite binary words, Sib. Elektron. Mat. Izv. 3 (2006), 304–311.

[22] M. Makarov, On an infinite permutation similar to the Thue–Morse word, Discrete Math. 309 (2009), 6641–6643.

[23] M. Makarov, On the permutations generated by Sturmian words. Sib. Math. J. 50 (2009), 674–680.

[24] B. Mossé, Puissance de mots et reconnaissabilité des points fixes d'une substitution, Theoret. Comput. Sci. 99 (1992), 327–334.

[25] M. Queffélec, Substitution Dynamical Systems — Spectral Analysis, Lecture Notes in Mathematics 1294, Springer, 2010 (2nd edition).

[26] G. Richomme, Conjugacy and episturmian morphisms, Theoret. Comput. Sci. 302 (2003), 1–34.

[27] P. Séébold, An effective solution to the D0L periodicity problem in the binary case, EATCS Bull. 36 (1988), 137–151.

[28] P. Séébold, On the conjugation of standard morphisms, Theoret. Comput. Sci. 195 (1998), 91–109.

[29] A. Valyuzhenich, On permutation complexity of fixed points of uniform binary morphisms, Discr. Math. Theoret. Comput. Sci. 16 (2014), 95–128.