SYMBOLIC DYNAMICS
M2 MPRI 2021

V. BERTHÉ
BERTHE@IRIF.FR

Contents

1. Introduction 1
2. General notions 2
2.1. Discrete dynamical systems 2
2.2. Word combinatorics and factor complexity 2
2.3. Frequencies and balance 5
2.4. Symbolic dynamical systems 7
2.5. More on measure-theoretic dynamical systems 10
3. Substitutions 11
3.1. First properties 11
3.2. More on Pisot substitutions 17
4. Graphs of Words 18
4.1. First definitions 18
4.2. More on graphs of words and frequencies 20
5. Sturmian words 21
5.1. Factors and intervals 22
5.2. Frequencies of factors 23
5.3. More on Sturmian words and substitutions 24
6. Hints and corrections for exercices 24
References 31

1. Introduction

The object of study of symbolic dynamics are discrete dynamical systems made of infinite sequences of symbols with values in a finite alphabet, with the shift map T acting on them: the shift T maps an infinite word $(u_n)_{n\geq 0}$ onto this infinite word from which the first letter has been taken away, that is, $T((u_n)_{n\in\mathbb{N}}) = (u_{n+1})_{n\in\mathbb{N}}$. Symbolic dynamical systems come in a natural way as codings of trajectories of points in a dynamical system according to a finite partition. They are used as discretization tools for analyzing such trajectories, but they also occur in a natural way in arithmetics for instance for the representation of numbers (real, complex), vectors, or else polynomials or Laurent formal power series with coefficients in a finite field.

Date: February 25, 2025.

Symbolic dynamics originates in the work of Jacques Hadamard [23], in 1898, through the study of geodesics on surfaces of negative curvature (see also [15]). It was then also applied by Marston Morse in 1921 in [31] to the construction of a nonperiodic recurrent geodesic and for the symbolic representations of geodesics. The study of combinatorics on words originates at the same time in papers of Axel Thue from 1906 and 1912 (see [36, pp. 139-158 and 413-477]), in particular with the study of the Thue-Morse word. Symbolic dynamics and Sturmian words then were developed by Morse and Hedlund in 1938 in [32, 33]. Substitutions are central objects of symbolic dynamics. They play a prominent role in the study of aperiodic tilings and in aperiodic order for the mathematical formalization of quasicrystals.

Acknowledgement. I would like to thank warmly Léonard Brice, Antonin Callard, Clément Ducros, Thiago Felicissimo Cesar, Simon Jeanteur, Jana Lepsova, Rémi Morvan and Delphine Salvy for their careful reading and for their constructive additions, and for various proofs for the exercices in Section 6.

2. General notions

2.1. **Discrete dynamical systems.** By *discrete dynamical system* we mean here a map $T\colon X \to X$ that acts on a space X that will be usually assumed to be compact. The map T is also usually assumed to be continuous or piecewise continuous.

The (one-sided) *orbit* of $x \in X$ under the action of T is defined as $\{T^n x \mid n \in \mathbb{N}\}$. If T is assumed to be invertible (e.g., if T is a homeomorphism), then the two-sided orbit of $x \in X$ under the action of T is defined as $\{T^n x \mid n \in \mathbb{Z}\}$. Orbits are also called *trajectories*.

The terminology discrete refers here to the time: we consider trajectories of points of X under the discrete-time deterministic action of the mapping T. Discrete dynamical systems can be of a geometric nature (e.g., $X = [0,1]$), or of a symbolic nature (e.g., $X = \{0,1\}^{\mathbb{N}}$), such as described below. More precisely, as examples of dynamical systems, let us mention

- symbolic dynamical systems: these are dynamical systems defined on sets of symbols and words; we consider them in Section 2.4;
- the translation R_α by α on the one-dimensional torus, that is, $R_\alpha\colon x \mapsto x + \alpha \bmod 1$ (see Section 5).

The notion of dynamical system can be considered in a topological context (this is what we have considered so far), we get *topological dynamics*, but this notion can be extended to measurable spaces: we thus get *measure-theoretic dynamical systems*, that is, dynamical systems endowed with a probabilistic structure (an invariant measure). We will consider them in Section 2.5.

2.2. **Word combinatorics and factor complexity.** An *alphabet* is a finite set of *symbols* (or *letters*). Let $\mathcal{A}$ be an alphabet. A *finite* word over $\mathcal{A}$ is a finite sequence of letters in $\mathcal{A}$ (that is, a word of length n is a map u from $\{1, \cdots, n\}$ to $\mathcal{A}$). We write it as $u = u_1 \cdots u_n$ to express u as the concatenation of the letters u_i.

Let $u = u_1 \cdots u_m$ and $v = v_1 \cdots v_n$ be two words over $\mathcal{A}$. The *concatenation* of u and v is the word $w = w_1 \cdots w_{m+n}$ defined by $w_i = u_i$ if $1 \le i \le m$, and $w_i = v_{i-m}$ otherwise. We write $u \cdot v$ or simply uv to express the concatenation of u and v. The set of all (finite) words over $\mathcal{A}$ is denoted by $\mathcal{A}^*$. Endowed with the concatenation of words as product operation, $\mathcal{A}^*$ is a monoid with ε as identity element. It is the *free* monoid generated by $\mathcal{A}$. We thus have endowed the set of finite words with an algebraic structure.

We also consider infinite words, that is, elements of $\mathcal{A}^{\mathbb{N}}$, as well as bi-infinite (also called two-sided) words in $\mathcal{A}^{\mathbb{Z}}$. In all that follows we restrict ourselves to infinite words over a finite alphabet indexed by the set $\mathbb{N}$ of non-negative integers. All the notions defined below extend to two-sided words in $\mathcal{A}^{\mathbb{Z}}$.

A *factor* of the infinite word $u = (u_n)_n$ is a finite block w of consecutive letters of u, say $w = u_{n+1} \cdots u_{n+l}$ for some index n (called an index of occurrence of w in u); l is called the *length* of w, denoted by $|w|$.

Definition 1 (Factors and language)**.** A word $w_1 \cdots w_\ell$ is a *factor* of the word u (finite, infinite or bi-infinite) if there exists k such that $u_k \cdots u_{k+\ell-1} = w_1 \cdots w_\ell$. The set of factors $\mathcal{L}_u$ of an infinite word u is called its *language*. The set of factors of length n is denoted as $\mathcal{L}_u(n)$.

Definition 2 (Recurrence)**.** An infinite word is said to be *recurrent* if every factor appears infinitely often.

Note that every factor occurs infinitely often is equivalent to every factor occurs at least twice.

Definition 3 (Uniform recurrence)**.** An infinite word is said to be *uniformly recurrent* if every factor appears infinitely often and with bounded gaps (or, equivalently, if for every integer n, there exists an integer m such that every factor of u of length m contains every factor of length n). This gives

$$\forall n \in \mathbb{N}, \exists m \in \mathbb{N}, \forall w \in \mathcal{L}_u \cap \mathcal{A}^m, \mathcal{L}_u \cap \mathcal{A}^n \subseteq \mathcal{L}_w.$$

Exercise 4. Take $\mathcal{A} = \{a, b\}$ and consider the concatenation of words of $\mathcal{A}^*$ ordered by length and, when two words are of equal length, ordered by the lexicographic order, i.e.,

$$\varepsilon \cdot a \cdot b \cdot aa \cdot ab \cdot ba \cdot bb \cdot aaa \cdot aab \cdots \in \mathcal{A}^{\mathbb{N}}$$

and show that it is recurrent but not uniformly recurrent.

Definition 5 (Linear recurrence)**.** An infinite word u is said to be *linearly recurrent* if there exists C such that every factor of u of length Cn contains every factor of u of length n.

Observe that linear recurrence implies uniform recurrence which also implies recurrence.

Exercise 6. Construct an example of a word that is uniformly recurrent but not linearly recurrent.

Let us introduce a combinatorial measure of disorder for infinite words over a finite alphabet: this notion is called *factor complexity*.

Definition 7 (Factor complexity)**.** The (*factor*) *complexity* of an infinite word u counts the number of distinct factors of a given length: there are exactly $p_u(n)$ factors of length n in u.

For more on this function, see for instance [4]. The factor complexity is obviously non-decreasing and for any integer n, one has $1 \leq p_u(n) \leq d^n$, where d denotes the cardinality of the alphabet.

This function can be considered to measure the predictability of an infinite word. Indeed, the first difference of the factor complexity counts the number of possible extensions in the infinite word u of factors of a given length.

Definition 8 (Right and left extensions)**.** We call *right extension* (respectively *left extension*) of a factor w of the infinite word u a letter x such that wx (respectively xw) is a factor of the infinite word u.

Let u be an infinite word. Let w^+ (respectively w^-) denote the number of right (respectively left) extensions of w in u. (One may have $w^- = 0$ but always $w^+ \geq 1$.) We have

$$p_u(n+1) = \sum_{w \in \mathcal{L}_u(n)} w^+ = \sum_{w \in \mathcal{L}_u(n)} w^-,$$

and thus

$$p_u(n+1) - p_u(n) = \sum_{w \in \mathcal{L}_u(n)} (w^+ - 1) = \sum_{w \in \mathcal{L}_u(n)} (w^- - 1).$$

Indeed,

$$\sum_{w \in \mathcal{L}_u(n)} (w^+ - 1) = \sum_{w \in \mathcal{L}_u(n)} w^+ - \sum_{w \in \mathcal{L}_u(n)} 1 = p_u(n+1) - \#\mathcal{L}_u(n) = p_u(n+1) - p_u(n).$$

Definition 9 (Periodicity)**.** An infinite word u is *periodic* if there exists a positive integer T such that $\forall n,\ u_n = u_{n+T}$. It is *ultimately periodic* if there exist a positive integer T and $n_0 \in \mathbb{N}$ such that $\forall n \geq n_0,\ u_n = u_{n+T}$.

Proposition 10. *Let $(u_n)_{n\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}}$. The following assertions are equivalent:*

(1) $(u_n)_{n\in\mathbb{N}}$ is ultimately periodic;
(2) $\exists n, p_u(n) \leq n$;
(3) $\exists C, \forall n, p_u(n) \leq C$.

Proof. If $p_u(1) = 1$, then u is constant. Otherwise, we assume $p_u(1) \geq 2$. Then, $p_u(n) \leq n$ implies that $p_u(k+1) = p_u(k)$ for some k. For each word w of length k occurring in u, there exists at least one word of the form wa occurring in u, for some letter $a \in \mathcal{A}$. As $p_u(k+1) = p_u(k)$, there can be only one such word. Hence, if $u_i...u_{i+k-1} = u_j...u_{j+k-1}$, then $u_{i+k} = u_{j+k}$. As the set $\mathcal{L}_u(k)$ is finite, there exist $j > i$ such that $u_i...u_{i+k-1} = u_j...u_{j+k-1}$, and hence $u_{i+p} = u_{j+p}$ for every $p \geq 0$, one period being $j - i$. □

Exercise 11. What happens in the case of a binfinite word defined over $\mathbb{Z}$?

The complexity function is a measure of disorder connected to the topological entropy: the *topological entropy* is defined as the exponential growth rate of the complexity as the length increases

$$H_{top}(u) = \lim_{n \to +\infty} \frac{\log_d(p_u(n))}{n}.$$

Remember that d stands for the cardinality of the alphabet. It is easy to check that this limit exists because of the subadditivity of the function $n \mapsto \log(p_u(n))$. Note that the word *entropy* is used here as a measure of randomness or disorder.

Indeed, $p_u(n+m) \leq p_u(n)p_u(m)$ because a factor of length $n+m$ can be seen as the concatenation of a factor length n with a factor of length m: consider the map from $\mathcal{L}_u(m+n)$ to $\mathcal{L}_u(m) \times \mathcal{L}_u(n)$ which maps w to $(u, v) \in \mathcal{L}_u(m) \times \mathcal{L}_u(n)$, where $w = u \cdot v$; this map is clearly injective, which implies that $p_u(m+n) \leq p_u(m)p_u(n)$. Therefore, $\log p_u(n+m) \leq \log p_u(n) + \log p_u(m)$. Now, we prove the subadditive lemma. Let $(u_n)_{n\in\mathbb{N}}$ be a subadditive

sequence: we prove that (u_n/n) admits a limit in $\mathbb{R} \cup \{-\infty\}$. To do so, let n and q be two nonnegative integers. Let $n = pq + r$ be the Euclidean division of n by q. Then we have:

$$\frac{u_n}{n} = \frac{u_{pq+r}}{n} \leq \frac{(p-1)u_q + u_{q+r}}{n} \leq \frac{n-q-r}{n}\frac{u_q}{q} + \frac{\max_{0\leq i<q} u_{q+i}}{n}.$$

By going to lim sup on n on both sides, we obtain that $\limsup u_n/n \leq u_q/q$. And then going to lim inf on q, we obtain that $\limsup u_n/n \leq \liminf u_q/q$, which means that $\limsup u_n/n = \liminf u_n/n$, which concludes the proof. Because $p_u(n) \geq 1$ for all n, this limit is non-negative in the case of $\log p_u(n)$.

The study of the complexity is mainly concerned with the following three questions.

- How to compute the complexity of an infinite word?
- Which functions can be obtained as the complexity function of some infinite word?
- Can one deduce from the complexity a geometrical representation of infinite words?

We will see how to answer the first question by introducing special factors, in some particular cases of substitutive words. Although the complexity function is in general not sufficient to describe an infinite word , we will see that much can be said on the geometrical properties in the case of lowest complexity, i.e., in the case of *Sturmian words*: these words are defined to have exactly $n+1$ factors of length n, for any integer n. By Proposition 10 an infinite word u with complexity satisfying $p_u(n) \leq n$ for some n is ultimately periodic. Sturmian words have thus the minimal complexity among all infinite words that are not ultimately periodic.

Exercise 12. Deduce from Proposition 10 that every factor of a Sturmian word appears at least two times in the infinite word. Deduce that the factors of every Sturmian word appear infinitely often (we recall that such an infinite word is called *recurrent*).

Correction. We prove the first point by contradiction: assume that w is a prefix of a Sturmian word u that appears only once in u. Decompose u as $u = a \cdot u'$, where $a \in \mathcal{A}$. Obviously, any factor of u' is a factor of u, so $p_{u'}(n) \leq p_u(n)$ for all n. Consider now $n = |w|$: because w only appears once in u as a prefix, it does not occur in u', which means that $p_{u'}(n) \leq p_u(n) - 1 = n$. This implies that u' is ultimately periodic by Proposition 1, and so is u.

Suppose now that a factor w appears only once in the Sturmian word u. Then, there exists N such that $w \notin \mathcal{L}_{T^N u}$. Then $T^n u$ has at most $|w|$ factors of size $|w|$ (u had $|w|+1$ factors, w included), so $T^n u$ is ultimately periodic, so u is too, which is a contradiction to u being aperiodic.

2.3. **Frequencies and balance.** Let $w \in \mathcal{A}^*, u \in \mathcal{A}^{\mathbb{N}}$. Let $|u_k \dots u_{k+n}|_w$ denote the number of occurrences of w in $u_k \dots u_{k+n}$ (with possibly overlaps).

Example 13. One has $|abaaabaaaa|_{aa} = 5$.

Definition 14 (Frequencies)**.** Let u be a word in $\mathcal{A}^{\mathbb{N}}$. The *frequency* f_i of a letter $i \in \mathcal{A}$ in u is defined as the limit when n tends towards infinity, if it exists, of the number of occurrences of i in $u_0 u_1 \cdots u_{n-1}$ divided by n, i.e.,

$$f_i = \lim_{n\to\infty} \frac{|u_0 \dots u_{n-1}|}{n}$$

exists. In the case that the limit exists, we say that u *admits a frequency for the letter* i.

The word u admits *uniform letter frequencies* if, for every letter i of u, the number of occurrences of i in $u_k \cdots u_{k+n-1}$ divided by n has a limit when n tends to infinity, uniformly in k, i.e., for all letter i, there exists f_i such that

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N},\ \forall n \geq n_0,\ \forall k,\ \left| \frac{1}{n} |u_k \dots u_{k+n}|_i - f_i \right| \leq \varepsilon.$$

Similarly, we can define the frequency f_w and the uniform frequency of a factor w, and we say that u has *uniform frequencies* if all its factors have uniform frequency, i.e., for all factor w, there exists f_w such that

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N},\ \forall n \geq n_0,\ \forall k,\ \left| \frac{1}{n} |u_k \dots u_{k+n}|_w - f_w \right| \leq \varepsilon.$$

Definition 15 (Unique ergodicity)**.** An infinite word is said to be *uniquely ergodic* if it admits uniform frequencies for all it factors. It implies that it admits a frequency for every factor.

We will revisit this definition in Section 2.5.

Exercise 16. Construct an infinite word for which the frequencies of letters do not exist.

Correction. Consider the infinite word

$$u = 010^2 1^2 0^4 1^4 0^8 \cdots$$

The infinite word u admits no frequencies for the letter 0. Indeed, if u is cut after a 1, $\frac{|u_0 \cdots u_n|_0}{n} = \frac{1}{2}$, but if it is cut just after a sequence of zeroes, the value changes. For instance, if we cut after a sequence of $N+2$ zeroes, we have $n = 2\sum_{k=0}^{N} 2^k + 2^{N+1}$ and

$$\frac{|u_0 \cdots u_{n-1}|_0}{n} = \frac{\sum_{k=0}^{N} 2^k + 2^{N+1}}{2\sum_{k=0}^{N} 2^k + 2^{N+1}} = \frac{2^{N+1} - 1 + 2^{N+1}}{2^{N+2} + 2^{N+1} - 1} = \frac{1 - 2^{-(N+2)}}{1 + \frac{1}{2} - 2^{-(N+1)}} \to \frac{2}{3}.$$

More generally, on the alphabet $\mathcal{A} = \{0, 1\}$, define the function f from $\mathbb{N}$ to $\mathbb{N}$ as $f : n \mapsto 2^{2^n}$ and:

$$u = 0^{f(1)} \cdot 1^{f(2)} \cdot 0^{f(3)} \cdot 1^{f(4)} \ \dots\ 0^{f(2k+1)} \cdot 1^{f(2k+2)} \ \dots$$

For any $n \in \mathbb{N}$, one has:

$$\sum_{k=1}^{n} f(k) \leq n 2^{2^n} \leq 2^{2^n + \log n}$$

Hence

$$\frac{f(n+1)}{\sum_{k=1}^{n} f(k) + f(n+1)} \to_{n \to +\infty} 1$$

In particular, this implies that the frequencies of both letters 0 and 1 oscillate between 0 and 1 without ever stabilizing (to be more precise, $\limsup_n \frac{|u_0 \cdots u_n]|_a}{n} = 1$ and $\liminf_n \frac{|u_0 \cdots u_n|_a}{n} = 0$ for any $a \in \{0, 1\}$). So u is a word such that the frequencies of letters do not exist.

Definition 17 (Discrepancy)**.** Let $u \in \mathcal{A}^{\mathbb{N}}$ be an infinite word and assume that u admits the frequency f_i for each letter i. The *discrepancy* of u is

$$\Delta(u) = \limsup_{i \in \mathcal{A},\ n \in \mathbb{N}} ||u_0 u_1 \dots u_{n-1}|_i - n f_i|.$$

The quantity $\Delta(u)$ is considered e.g. in [1, 2]. We also consider

$$\Delta_n(u) = \sup_{i \in \mathcal{A}} ||u_0 u_1 \dots u_{n-1}|_i - n f_i|.$$

Definition 18 (Balancedness)**.** An infinite word $u \in \mathcal{A}^{\mathbb{N}}$ is said to be *C-balanced* if for any pair v, w of factors of the same length of u, and for any letter $i \in \mathcal{A}$, one has $||v|_i - |w|_i| \leq C$.

It is said *balanced* if there exists $C > 0$ such that it is C-balanced.

The vector f whose components are given by the frequencies of the letters is called the *letter frequency vector*.

Proposition 19. *An infinite word $u \in \mathcal{A}^{\mathbb{N}}$ is balanced if and only if it has uniform letter frequencies and there exists a constant B such that for any factor w of u, we have $||w|_i - f_i|w|| \leq B$ for all letter i in $\mathcal{A}$, where f_i is the frequency of i. Moreover, if u has letter frequencies, then u is balanced if and only if its discrepancy $\Delta(u)$ is finite.*

Proof. Let u be an infinite word with letter frequency vector f and such that $||w|_i - f_i|w|| \leq B$ for every factor w and every letter i in $\mathcal{A}$. Then, for every pair of factors w_1 and w_2 with the same length n, we have by triangular inequality

$$||w_1|_i - |w_2|_i| \leq ||w_1|_i - nf_i| + ||w_2|_i - nf_i| \leq 2B.$$

Hence L is $2B$-balanced (see also [1, Proposition 7]).

Conversely, assume that u is C-balanced for some $C > 0$. We fix a letter $i \in \mathcal{A}$. For every non-negative integer p, let N_p be defined as an integer N such that for every word of length p of u, the number of occurrences of the letter i belongs to the set $\{N, N+1, \cdots, N+C\}$.

We first observe that the sequence $(N_p/p)_{p\in\mathbb{N}}$ is a Cauchy sequence. Indeed consider a factor w of length pq, where $p, q \in \mathbb{N}$. The number $|w|_i$ of occurrences of i in w satisfies

$$pN_q \leq |w|_i \leq pN_q + pC, \qquad qN_p \leq |w|_i \leq qN_p + qC.$$

We deduce that $-qC \leq qN_p - pN_q \leq pC$ and thus $-C \leq N_p - pN_q/q \leq pC/q$, so $-C/p \leq N_p/p - N_q/q \leq C/q$.

Let f_i stand for $\lim_q N_q/q$. By letting q tend to infinity, one then deduces that $-C \leq N_p - pf_i \leq 0$. Thus, for any factor w of u we have

$$\left|\frac{|w|_i}{|w|} - f_i\right| \leq \frac{C}{|w|},$$

which was to be proved.

If u has letter frequencies, the equivalence between balancedness and finite discrepancy comes from triangular inequality. □

Sturmian words (see Section 5) are known to be 1-balanced [29]; they even are exactly the 1-balanced infinite words that are not eventually periodic.

2.4. **Symbolic dynamical systems.** For detailed introductions to symbolic dynamics and word combinatorics, see [4, 6, 10, 11, 25, 27, 28, 29, 22] and the references therein.

We first endow the set of infinite words $\mathcal{A}^{\mathbb{N}}$ with a topology. This topology is given by the usual metric on infinite words: two infinite words are close if they coincide on their first terms. More precisely, the set $\mathcal{A}^{\mathbb{N}}$ shall be equipped with the product topology of the discrete topology on each copy of $\mathcal{A}$. Thus, by Tychonov theorem, this set is a compact space. This topology is also the topology defined by the following distance:

$$\text{for } u \neq v \in \mathcal{A}^{\mathbb{N}},\ d(u,v) = 2^{-\min\{n\in\mathbb{N};\ u_n \neq v_n\}}.$$

Then, the sequence of infinite words $(u^{(n)})_n$ converges to $u \in \mathcal{A}^{\mathbb{N}}$ if $\forall N \in \mathbb{N}, \exists n_0 \in \mathbb{N} / \forall n \geq n_0$, $u^{(n)}$ and u share the same prefix of length N.

Note that the space $\mathcal{A}^{\mathbb{N}}$ is complete as a compact metric space. Furthermore, it is a *Cantor set*, that is, a totally disconnected compact set without isolated points.

Note that the topology extends in a natural way to $\mathcal{A}^{\mathbb{N}} \cup \mathcal{A}^*$. Indeed, let $\mathcal{B}$ be a new alphabet obtained by adding a further letter to the alphabet $\mathcal{A}$; finite words in $\mathcal{A}^*$ can be considered as infinite words in $\mathcal{B}^{\mathbb{N}}$, by extending them by the new letter in $\mathcal{B}$. The set $\mathcal{A}^{\mathbb{N}} \cup \mathcal{A}^*$ is thus metric and compact, as a closed subset of $\mathcal{B}^{\mathbb{N}}$.

The mapping T acting on sets of infinite words is the (one-sided) *shift map* acting on $\mathcal{A}^{\mathbb{N}}$, given by $T((u_n)_{n\in\mathbb{N}}) = (u_{n+1})_{n\in\mathbb{N}}$. It is continuous.

Definition 20 (Subshift)**.** A *subshift (also called shift)* is a closed shift invariant set included in some $\mathcal{A}^{\mathbb{N}}$.

If Y is a subshift, there exists a set $\mathcal{F} \subset \mathcal{A}^*$ of finite words such that an infinite word u belongs to Y if, and only if, none of its factors belongs to $\mathcal{F}$. This is by definition of the product topology: if Y is closed, its complement is open, and as such can be written as a union of cylinders: the words of these cylinders form a familly $\mathcal{F}$ of forbidden patterns. Reciprocally, any family $\mathcal{F}$ of words defines a subshift, possibly empty. A subshift X is called a *subshift of finite type* if one can choose the set $\mathcal{F}$ to be finite. A subshift is said to be *sofic* if the set $\mathcal{F}$ is a regular language. A subshift X is called a *subshift of finite type* if one can choose the set $\mathcal{F}$ to be finite. A subshift is said to be *sofic* if the set $\mathcal{F}$ is a regular language.

Definition 21 (Cylinder)**.** For a word $w = w_0 \cdots w_r$, the *cylinder set* $[w]$ is the set $\{v \in X_u \mid v_0 = w_0, \cdots, v_r = w_r\}$.

The cylinder sets are *clopen* (open and closed) sets in $\mathcal{A}^{\mathbb{N}}$ and form a basis of open sets for the topology of X_u. Indeed, if the cylinder $[w]$ is nonempty and v is a point in it, $[w]$ is identified with both the open ball $\{v' \mid d(v, v') < 2^{-r}\}$ and the closed ball $\{v' \mid d(v, v') \leq 2^{-r-1}\}$.

Exercise 22. As an exercise, prove that a clopen set is a finite union of cylinders.

As an example of a shift, take the closure in $\mathcal{A}^{\mathbb{N}}$ of the positive orbit of $u = (u_n)_{n\geq 0}$ under the action of the shift T, with u being some infinite word in $\mathcal{A}^{\mathbb{N}}$. This gives $X_u := \overline{\mathcal{O}(u)}$. One checks that

$$\overline{\mathcal{O}(u)} = \{v \in \mathcal{A}^{\mathbb{N}},\ \mathcal{L}_v \subset \mathcal{L}_u\},$$

where $\mathcal{L}_v$ is recalled to be the set of factors of the word v. Indeed, let $v \in \overline{\mathcal{O}(u)}$, and let w be a factor of v. Without any loss of generality, we can assume that w is a prefix of v (if it is not the case, we can shift v until it is). Then $[w]$ is a cylinder which contains v, so $[w]$ is a neighborhood of v. Because $v \in \overline{\mathcal{O}(u)}$, one has $[w] \cap \mathcal{O}(u) \neq \emptyset$. So there exists some n such that w is a prefix of $T^n(u)$. In other words, w is a factor of u.

Reciprocally, let v be an infinite word such that $\mathcal{L}_v \subseteq \mathcal{L}_u$, and consider an open set containing u. As cylinders form a basis of $\mathcal{A}^{\mathbb{N}}$, one can assume that such an open set is of the form $[w]$ for some $w \in \mathcal{A}^*$. Then w is a prefix of v, so $w \in \mathcal{L}_v \subseteq \mathcal{L}_u$. This implies that w is a factor of u, so there exists some n such that w is a prefix of $T^n(u)$. This implies that $[w] \cap \mathcal{O}(u) \neq \emptyset$. With this, we conclude that $v \in \overline{\mathcal{O}(u)}$.

We can generalize this fact for any shift Y with a language $\mathcal{L}_Y = \{w \in \mathcal{A}^* : \exists v \in Y, w \text{ is a factor of v}\}$, i.e.

$$Y = \{v \in \mathcal{A}^{\mathbb{N}},\ \mathcal{L}_v \subset \mathcal{L}_Y\},$$

This can be summarized by a claim that *every shift is described by its language.*

Let us come back to the general case of a discrete dynamical system $T\colon X \to X$. In order to understand the behavior of trajectories, it is natural to partition the set X into a finite number (say d) of subsets $(X_i)_{1\leq i\leq d}$: $X = \cup_{i=1}^{d} X_i$. We then *code* the trajectory of a point $x \in X$ with respect to the finite partition $(X_i)_{1\leq i\leq d}$. One thus associates with each point $x \in X$ an infinite word with values in the finite alphabet $\{1, \dots, d\}$ defined as follows:

$$\forall n \in \mathbb{N},\ u_n = i \text{ if and only if } T^n(x) \in X_i.$$

Coding trajectories allows one to go from dynamical systems (X, T) defined on 'geometric' spaces X to symbolic dynamical systems and backwards, provided the coding has been chosen in an efficient way. Section 5 devoted to Sturmian words, provides an example of such a fruitful coding since Sturmian words code discrete lines. If the partition is well-chosen, these symbolic codings allow the statistical analysis (via ergodic theory) on the underlying dynamical systems. This is the object of next section.

Consider the (one-sided shift) T defined on $\mathcal{A}^{\mathbb{N}}$ as $T((u_n)_n) = (u_{n+1})_n$. The map T is uniformly continuous, onto but not necessarily one-to-one on $\mathcal{A}^{\mathbb{N}}$.

Exercise 23. We recall that an infinite word is *recurrent* if every factor every factor appears an infinite number of times in this infinite word.

Prove that an infinite word u is recurrent if and only if there exists a strictly increasing sequence $(n_k)_k$ such that

$$u = \lim_{k\to+\infty} T^{n_k} u.$$

Definition 24 (Minimality)**.** Let (X, T) be a subshift. It is said minimal if the only subsets of X that are closed and stable by the shift are the empty set and X.

Exercise 25. Prove that (X, T) is minimal if and only if $X = \overline{\mathcal{O}(u)}$, for every element u of X.

We recall that an infinite word is said to be *uniformly recurrent* if every factor appears infinitely often and with bounded gaps (or, equivalently, if for every integer n, there exists an integer m such that every factor of u of length m contains every factor of length n).

Proposition 26. *An infinite word u is uniformly recurrent if and only if $(\overline{\mathcal{O}(u)}, T)$ is minimal*

Proof. The idea is that if w is a factor of u, write

$$\overline{\mathcal{O}(u)} = \bigcup_{n\in\mathbb{N}} T^{-n}[w],$$

and conclude by a compactness argument.

- Assume that $(\overline{\mathcal{O}(u)}, T)$ is minimal. Let $w \in \mathcal{L}_u$. Let us consider $T^{-n}[w] = \{v \in \overline{\mathcal{O}(u)} \mid w \text{ is a prefix of } T^n v\}$.

  Let $v \in \overline{\mathcal{O}(u)}$. Since $\overline{\mathcal{O}(v)} = \overline{\mathcal{O}(u)}$ by minimality, one has $\mathcal{L}_u = \mathcal{L}_v$. This implies that for any $w \in \mathcal{L}_u$, if $v \in \overline{\mathcal{O}(u)}$, then w must appear somewhere in v. In other words, one has

  $$\overline{\mathcal{O}(u)} \subseteq \bigcup_{n\in\mathbb{N}} T^{-n}[w].$$

  We recall that T is continuous and that w is a clopen set, so $T^{-n}[w]$ is a clopen set as well. The set $\bigcup_{n\in\mathbb{N}} T^{-n}[w]$ is closed. It is also non-empty ($u \in T^{-n_0}[w]$ for some

n_0), shift-invariant, and a subset of $\overline{\mathcal{O}(u)}$. So, by minimality,

$$\overline{\mathcal{O}(u)} = \bigcup_{n\in\mathbb{N}} T^{-n}[w].$$

As $\overline{\mathcal{O}(u)}$ is closed in a compact space, it is compact. By compactness, if there is a cover by open sets, a finite cover can be extracted. Hence there exists some $m \in \mathbb{N}$ such that

$$\overline{\mathcal{O}(u)} \subseteq \bigcup_{n=0}^{m} T^{-n}[w].$$

Since, for every $k \in \mathbb{N}$, $T^k(u) \in \overline{\mathcal{O}(u)}$, one has

$$\forall k \in \mathbb{N}, T^k(u) \in \bigcup_{n=0}^{m} T^{-n}[w].$$

This means that w appears infinitely often in u, and that the gaps between two of its occurrences are bounded by m, i.e., u is uniformly recurrent.

- Reciprocally, assume that u is uniformly recurrent. Let $v \in \overline{\mathcal{O}(u)}$. Let us show that for all $w \in \mathcal{L}_u$, one has $w \in \mathcal{L}_v$. Then, we would have $\overline{\mathcal{O}(u)} \subset \overline{\mathcal{O}(v)}$ and since $\overline{\mathcal{O}(v)} \subset \overline{\mathcal{O}(u)}$, we would conclude that $\overline{\mathcal{O}(u)} = \overline{\mathcal{O}(v)}$, which proves the minimality.

  Let $w \in \mathcal{L}_u$. Since u is uniformly recurrent, there exists N such that every factor of length N of u contains w. We know that $v \in \overline{\mathcal{O}(u)} = \{T^n u | n \in \mathbb{N}\} \cup \{\lim_{k\to\infty} T^{n_k} u | (n_k)_k \text{ an increasing sequence}\}$.
  - Suppose that there exists n such that $v = T^n u$. Then $w \in \mathcal{L}_v$.
  - Suppose that there exists (n_k) such that $v = \lim_{k\to\infty} T^{n_k} u$. We know $\exists k_0 / \forall k \geq k_0, T^{n_k} u$ starts with the same prefix of length N as v. This prefix must contain w, so $w \in \mathcal{L}_v$.

□

2.5. **More on measure-theoretic dynamical systems.** General references on the subject are [12, 18, 24, 34, 35, 37]. See [19] for connections with number theory and Diophantine approximation.

A *measure-theoretic dynamical system* is defined as a system $(X, T, \mu, \mathcal{B})$, where μ is a probability measure defined on the σ-algebra $\mathcal{B}$ of subsets of X, and $T : X \to X$ is a measurable map which preserves the measure μ, that is, $\mu(T^{-1}(B)) = \mu(B)$ for all $B \in \mathcal{B}$. The measure μ is said to be *T-invariant*.

An invariant probability measure on X is said *ergodic* if for every set $B \in \mathcal{B}$ such that $T^{-1}(B) = B$, B has either zero or full measure. The system $(X, T, \mu, \mathcal{B})$ is then said to be *ergodic*. This implies that almost all orbits are dense in X (almost all means that the set of elements $x \in X$ whose orbit is not dense is contained in a set of zero measure). More generally a property is said to hold *almost everywhere* (abbreviated as a.e.) if the set of elements for which the property does not hold has zero measure; this property is said to be *generic* (the points that satisfy this property are then also said to be generic). This helps us to give a meaning to the notion of typical behavior for a dynamical system.

Ergodicity yields furthermore the following striking convergence result. Indeed, measure-theoretic ergodic dynamical system satisfy the *Birkhoff ergodic theorem*, also called *individual ergodic theorem*, which relates spatial means to temporal means.

Theorem 27 (Birkhoff Ergodic Theorem)**.** *Let $(X, T, \mu, \mathcal{B})$ be an ergodic measure-theoretic dynamical system. Let $f \in L^1(X, \mathbb{R})$. Then*

$$\frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k \xrightarrow[n\to\infty]{\mu-a.e.} \int_X f \, d\mu \ .$$

Points for which this convergence property holds for a given f are generic.

In the case of a symbolic dynamical system $(\overline{\mathcal{O}(u)}, T)$ generated by an infinite word u, the following special case of the Daniell-Kolmogorov consistency theorem (see for instance [37]) provides probability measures on $(\overline{\mathcal{O}(u)}, T)$.

Theorem 28. *Let $\mathcal{A} = \{1, \dots, d\}$ and $u \in \mathcal{A}^{\mathbb{N}}$. Consider a family of maps $(p_n)_{n\geq 1}$, where p_n is a map from $\mathcal{A}^n$ to $\mathbb{R}$, such that for any word w in $\mathcal{A}^n$, $p_n(w) \geq 0$, $p_n(w) = \sum_{i=1}^{d} p_{n+1}(w_1 \dots w_n i)$, and $\sum_{i=1}^{d} p_1(i) = 1$. Then there exists a unique probability measure μ on $\mathcal{A}^{\mathbb{N}}$ defined on the cylinders by $\mu([w_1 \dots w_n]) = p_n(w_1 \dots w_n)$.*

Furthermore, if for any n and for any word $w = w_1 \dots w_n$ in $\mathcal{A}^n$, $p_n(w) = \sum_{i=1}^{d} p_{n+1}(i w_1 \dots w_n)$, then this measure is T-invariant (shift-invariant).

If the frequencies of all factors exist for a given $u \in \mathcal{A}^{\mathbb{N}}$, then, according to Theorem 28, there exists a unique T-invariant probability measure μ which assigns to each cylinder $[w]$ the frequency $f(w)$ of the corresponding factor $[w]$, by setting $\mu[w] := f(w)$. Thus a precise knowledge of the frequencies allows a complete description of the measure μ. One can similarly define a shift-invariant measure for a subshift $X \subset \mathcal{A}^{\mathbb{N}}$ provided that any factor w in the langage of X (i.e., the set of factors of its elements) has the same frequency in all the infinite words of X. We have seen the notion of unique ergodicity in Definition 15. In fact, unique ergodicity is equivalent to the fact that there is a unique invariant measure. Unique ergodicity corresponds in the case of continuous functions to uniform convergence for *all* points (and not only for a.e. point) in ergodic sums, in Birkhoff's ergodic theorem. For more details on invariant measures and ergodicity, we refer to [35] and [10, Chap. 7].

Natural questions that can be addressed now concerning discrete dynamical systems are the following. What is a good coding? How to describe the invariant measures? Can one find geometric representations of a given symbolic dynamical system? How to measure the disorder of a dynamical system? The next sections provides some elements of answer.

3. Substitutions

3.1. **First properties.** We consider a finite set of *letters* $\mathcal{A}$, called *alphabet*. A (finite) *word* is an element of the free monoid $\mathcal{A}^*$ generated by $\mathcal{A}$. A *substitution* σ over the alphabet $\mathcal{A}$ is a non-erasing endomorphism of the free monoid $\mathcal{A}^*$ (non-erasing means that the image of any letter is not equal to the empty word but contains at least one letter).

Let $\mathcal{A} = [\![1, d]\!]$. For $i \in \mathcal{A}$ and for $w \in \mathcal{A}^*$, let $|w|_i$ stand for the number of occurrences of the letter i in the word w. Let σ be a substitution. Its *incidence matrix* $M_\sigma = (m_{i,j})_{1\leq i,j\leq d}$ is defined as the square matrix with entries $m_{i,j} = |\sigma(j)|_i$ for all i, j.

Note that if σ and τ are substitutions, then $\sigma \circ \tau$ is still a substitution and $M_{\sigma\circ\tau} = M_\sigma \cdot M_\tau$. (Sending a substitution to its incidence matrix defines a monoid homomorphism from the

monoid of substitutions to the monoid of square matrices equiped with multiplication.) In particular, $M_{\sigma^n} = (M_\sigma)^n$ for all $n \in \mathbb{N}$.

Definition 29 (Primitive substitution)**.** A substitution is said *primitive* if there exists a power of its incidence matrix whose entries are all positive.

A *fixed point* of a substitution σ is an infinite word $u = (u_n)_n$ with $\sigma(u) = u$.

Substitutions are very efficient tools for producing infinite words. Let σ be a substitution over the alphabet $\mathcal{A}$, and a be a letter such that $\sigma(a)$ begins with a and $|\sigma(a)| \geq 2$. Then there exists a unique fixed point u of σ beginning with a. This infinite word is obtained as the limit in $\mathcal{A}^* \cup \mathcal{A}^{\mathbb{N}}$ (when n tends toward infinity) of the sequence of words $(\sigma^n(a))_n$, which is easily seen to converge (we recall that the topology on $\mathcal{A}^{\mathbb{N}}$ is extended to $\mathcal{A}^* \cup \mathcal{A}^{\mathbb{N}}$ by adding an extra symbol to the alphabet $\mathcal{A}$).

Example 30 (Fibonacci substitution)**.** We consider the substitution σ on $\mathcal{A} = \{a, b\}$ defined by $\sigma(a) = ab$ and $\sigma(b) = a$. Its incidence matrix is

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then, the sequence of finite words $(\sigma^n(a))_n$ starts with

$$\sigma^0(a) = a,\ \sigma^1(a) = ab,\ \sigma^2(a) = aba,\ \sigma^3(a) = abaababa,\ \ldots$$

Each $\sigma^n(a)$ is a prefix of $\sigma^{n+1}(a)$, and the limit word in $\mathcal{A}^{\mathbb{N}}$ is

$$abaababaabaababaababaabaababaabaababaababaabaababaababaabaab\ldots$$

The above limit word is called the *Fibonacci word* (for more on the Fibonacci word, see e.g. [29, 22]).

Definition 31. A substitution is right prolongable if there exists a letter a such that $\sigma(a) = av$ with v non-empty word.

Proposition 32. *Any primitive substitution admits a power σ^k that is right prolongable, that is, there exists a letter a such that $\sigma^k(a) = av$ with v non-empty word. One then has $\lim_n \sigma^{kn}(a) = +\infty$. It thus generates a fixed point.*

Proof. Let σ be a substitution. Consider the oriented graph having as vertices the letters of the alphabet and an arrow between a and b if b if the first letter of $\sigma(a)$. Since the graph is finite, there exists a letter a and a non-negative integer n such that the first letter of $\sigma^n(a)$ is a. We conclude by using the fact that σ is primitive.

A redaction in more details. If $\#\mathcal{A} = 1$, i.e., $\mathcal{A} = \{a\}$, then for any substitution σ the configuration a^ω is a fixed point. In what follows, we assume that $\#\mathcal{A} \geq 2$.

- First, we prove that there exists some $n \in \mathbb{N}$ and a letter $b \in \mathcal{A}$ such that $\sigma^n(b)$ starts with b and is of length at least 2. As σ is a primitive substitution, there exists some $k \in \mathbb{N}$ such that for every $a, b \in \mathcal{A}$, a appears in $\sigma^k(b)$. In particular, for every letter $b \in \mathcal{A}$, $|\sigma^k(b)| \geq 2$. Now, let $a \in \mathcal{A}$ and consider the sequence $(\sigma^{kp}(a))_{p\geq 0}$. It is a sequence of non-empty words, and because $\mathcal{A}$ is a finite set, there exists some $p < p'$ such that $\sigma^{kp}(a)$ starts with the same letter b as $\sigma^{kp'}(a)$ (by the pigeon-hole principle). Then we prove that $\sigma^{k(p'-p)}(b)$ is a finite word which starts with the letter b. Indeed, $\sigma^{k(p'-p)}(\sigma^{kp}(a)) = \sigma^{kp'}(a)$ starts with the letter b, and $\sigma^{kp}(a)$ also starts with the letter b. Define $n = k(p' - p)$. As σ^n is a power of σ^k, we have $|\sigma^n(b)| \geq 2$ and $\sigma^n(b)$ starts with the letter b.

- The sequence $(\sigma^{np}(b))_{p\geq 0}$ is a converging sequence. Indeed, the sequence of lengths $(|\sigma^{np}(b)|)$ tends towards $+\infty$, and for every $p \geq 0$ the word $\sigma^{np}(b)$ is a prefix of $\sigma^{n(p+1)}(b)$.

□

We recall that an infinite word $u = (u_n)_n$ is *uniformly recurrent* if every word occurring in u occurs in an infinite number of positions with bounded gaps, that is, if for every factor w, there exists s such that for every n, w is a factor of $u_n \dots u_{n+s-1}$.

Proposition 33. *If σ is primitive, then any infinite word u with $\sigma^k(u) = u$ for some $k > 0$ is uniformly recurrent.*

Proof. Let p be such an integer that M_σ^p has only positive entries and let k be such an integer that $\sigma^k(a)$ starts with $a \in \mathcal{A}$ (as σ is primitive, we know there exist such k and a and there exists a fixed point $u = \sigma^k(u)$, $u_0 = a$). As the matrix M_σ has non-negative entries, it follows that M_σ^{p+n} has strictly positive entries for every integer $n \geq 1$. Also, it holds that $(\sigma^k)^n(u) = u$ for any $n \geq 1$. Let us take n such that $nk \geq p$ and let $\ell := nk$. Then, $\sigma^\ell(u) = \sigma^\ell(u_0)\sigma^\ell(u_1)...$ and every $\sigma^\ell(u_i)$ contains all the letters from the alphabet. Therefore the letters appear in u with bounded gaps, where the size of the gap is bounded by the largest of the numbers $\sigma^\ell(u_i)$. Let us consider a factor $w \in \mathcal{L}_u$. Then, w is a factor of $\sigma^m(u_0)$ for some $m \in \mathbb{N}$. As we can write $u = (\sigma^\ell)^m(u) = (\sigma^\ell)^m(u_0)(\sigma^\ell)^m(u_1)...$, every $\sigma^\ell(u_i)$ contains u_0 and therefore every $(\sigma^\ell)^m(u_i)$ contains w, w occurs in u with bounded gaps. There are finitely many of factors of each length and therefore we can take for N the maximum of the gaps for factors with the same length for proving uniform recurrence. □

According to Perron–Frobenius' theorem, if a substitution is primitive, then its incidence matrix admits a dominant eigenvalue λ (it dominates strictly in modulus the other eigenvalues) that is (strictly) positive. It is called its *Perron–Frobenius eigenvalue*, or else its *expansion* factor. Then, for all i, j, there exists $c_{i,j}$ such that $M_{i,j}^n/\lambda^n \to c_{i,j}$.

Proposition 34. *Let σ be a primitive substitution over the finite alphabet $\mathcal{A}$. Let λ stand for its Perron–Frobenius eigenvalue. Then, there exist $C, C' > 0$ such that for all letters in $\mathcal{A}$*

$$C'\lambda^n \leq |\sigma^n(a)| \leq C\lambda^n.$$

Proof. We use the fact that $M_{\sigma^n} = (M_\sigma)^n$, and that $|\sigma^n(a)| = \sum_{b\in\mathcal{A}} |\sigma^n(a)|_b$. In more details.

(1) Let M be the matrix associated to σ, a primitive substitution. Then if $a \in \mathcal{A}$ is the i^{th} letter, then $|\sigma^n(a)| = \sum_{j=1}^d (M^n)_{i,j}$.
(2) Let λ be the Perron-Frobenius eigenvalue, and v be an associated Perron eigenvector ($\forall 1 \leq j \leq d, v_j > 0$). Then:

$$\forall n, \sum_{j=1}^{d} (M^n)_{i,j} v_j = \lambda^n v_i$$

So if $M = \max v_j$ and $m = \min v_j$, one has:

$$\frac{m}{M}\lambda^n \leq \frac{v_i}{M}\lambda^n = \sum_{j=1}^{d} (M^n)_{i,j}\frac{v_j}{M} \leq \sum_{j=1}^{d} (M^n)_{i,j} = |\sigma^n(a)|$$

and

$$\frac{M}{m}\lambda^n \geq \frac{v_i}{m}\lambda^n = \sum_{j=1}^{d}(M^n)_{i,j}\frac{v_j}{m} \geq \sum_{j=1}^{d}(M^n)_{i,j} = |\sigma^n(a)|$$

So $C' = \frac{m}{M}$ and $C = \frac{M}{m}$ verify the aforementioned inequality.

□

Theorem 35. *(1) The factor complexity of a fixed point u of a primitive substitution or of a fixed point u of a substitution of constant length satisfies*

$$\exists C,\ \forall n,\ p_u(n) \leq Cn.$$

(2) The factor complexity of a fixed point u of a substitution satisfies

$$\exists C,\ \forall n,\ p_u(n) \leq Cn^2.$$

Proof. Let us prove (1). We show the result for σ primitive. Let $n \in \mathbb{N}$ and u be a fixed point of σ. Take $k \in \mathbb{N}$ satisfying

$$\min_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq n < \min_{a\in\mathcal{A}} |\sigma^k(a)|.$$

This index exists by Proposition 34 since the sequence $(\min_{a\in\mathcal{A}} |\sigma^j(a)|)_{j\geq 0}$ is non-decreasing and tends towards $+\infty$.

Let w be a factor of length n of u. As $\sigma(u) = u$, we write u as

$$u = \sigma^k(u) = \sigma^k(u_0)\sigma^k(u_1)\sigma^k(u_2)\cdots$$

Now note that a factor w of length n occurs either is some $\sigma^k(u_i)$ for some i, or in some $\sigma^k(u_i)\sigma^k(u_{i+1})$, since $n < \min_{a\in\mathcal{A}} |\sigma^k(a)|$. In both cases we conclude that such a factor occurs entirely in some $\sigma^k(u_iu_{i+1})$ for some i. Any factor of size n is thus uniquely determined by two letters u_i, u_{i+1} and an offset l with $0 \leq l < \max_{a\in\mathcal{A}} |\sigma^k(a)|$, i.e., the number of factors of length n is bounded by the number of possibilities for the choice of two letters and a starting position. We therefore have the inequality

$$p_n(u) \leq |\mathcal{A}|^2 \max_{a\in\mathcal{A}} |\sigma^k(a)|\,.$$

We now would like to bound $\max_{a\in\mathcal{A}} |\sigma^k(a)|$ by a linear factor. First note that $\max_{a\in\mathcal{A}} |\sigma^k(a)| \leq \max_{a\in\mathcal{A}} |\sigma(a)| \cdot \max_{a\in\mathcal{A}} |\sigma^{k-1}(a)|$ and thus

$$p_n(u) \leq |A|^2 \max_{a\in\mathcal{A}} |\sigma^k(a)| \leq |\mathcal{A}|^2 \max_{a\in\mathcal{A}} |\sigma(a)| \max_{a\in\mathcal{A}} |\sigma^{k-1}(a)|\,.$$

By Proposition 34, there are C, C' such that for all $a \in \mathcal{A}$,

$$C'\lambda^{k-1} \leq |\sigma^{k-1}(a)| \leq C\lambda^{k-1}\,,$$

where λ is the Perron Frobenius eigenvalue. This inequality implies

$$C'\lambda^{k-1} \leq \min_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq \max_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq C\lambda^{k-1}$$

and an easy manipulation gives

$$\max_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq C\lambda^{k-1} = \frac{C}{C'}C'\lambda^{k-1} \leq \frac{C}{C'} \min_{a\in\mathcal{A}} |\sigma^{k-1}(a)|\,.$$

Finally, remember that we have taken k satisfying $\min_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq n$, and therefore we also have

$$\max_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq \frac{C}{C'} \min_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq \frac{C}{C'} n\,.$$

By putting all of this together,

$$p_n(u) \leq |\mathcal{A}|^2 \max_{a\in\mathcal{A}} |\sigma(a)| \max_{a\in\mathcal{A}} |\sigma^{k-1}(a)| \leq |\mathcal{A}|^2 \max_{a\in\mathcal{A}} |\sigma(a)| \frac{C}{C'} n$$

and as $|\mathcal{A}|^2 \max_{a\in\mathcal{A}} |\sigma(a)| \frac{C}{C'}$ does not depend on n, we have shown the result.

□

We associate a symbolic dynamical system (X_σ, T) with the primitive substitution σ over $\mathcal{A}$. Let $u \in \mathcal{A}^{\mathbb{N}}$ be such that $\sigma^k(u) = u$ for some $k \geq 1$. We recall that such an infinite word exists by primitivity of σ. Indeed, there exist a letter a and a positive integer k such that $\sigma^k(a)$ begins with a and $|\sigma^k(a)| \geq 2$; consider as first letter of u this letter a; take $u = \lim_{n\to\infty} \sigma^{kn}(a)$. Such an infinite word exists by primitivity of σ. Let again $\overline{\mathcal{O}(u)}$ be the positive orbit closure of the infinite word u under the action of the shift T, i.e., the closure of the set $\mathcal{O}(u) = \{T^n(u) \mid n \geq 0\}$. The substitutive symbolic dynamical system (X_σ, T) generated by σ is defined as $X_\sigma := \overline{\mathcal{O}(u)}$.

One checks by primitivity that (X_σ, T) does not depend on the choice of the infinite word u fixed by some power of σ. For more details, see e.g. [35]. Indeed, first, one should note that because σ is primitive, if u is fixed by some power k of σ, then every letter of $\mathcal{A}$ appears in u. Indeed, let N be a non-negative integer such that M_σ^N is positive; then every letter must appear in $\sigma^{kN}(u) = u$. Now, assume that u and v are fixed by some power k of σ (we can assume that they are fixed by the same power; otherwise, take their product). We have that $\mathcal{L}_u \subseteq \mathcal{L}_v$: indeed, if w is a factor of u, we can assume that it is a prefix of u (otherwise, shift u enough); then if a denotes the first letter of u, there exists some $n \in \mathbb{N}$ such that $|\sigma^n(a)| \geq |w|$, and then $\sigma^{nk}(u) = u$ and $|\sigma^{nk}(a)| \geq w$. This means that w is a factor of $\sigma^{nk}(a)$. Because by our first consideration, a must appear in v, and that $\sigma^{nk}(v) = v$, we obtain that w is a factor of v. Finally, by symmetry, $\mathcal{L}_u = \mathcal{L}_v$, which leads to $\overline{\mathcal{O}(u)} = \overline{\mathcal{O}(v)}$.

The dynamical system (X_σ, T) associated with a primitive substitution σ can be endowed with a Borel probability measure μ *invariant* under the action of the shift T, that is, $\mu(T^{-1}B) = \mu(B)$, for every Borel set B. Indeed, this measure is uniquely defined by its values on the cylinders. For a given (finite) word w of the language of X_σ, the cylinder $[w]$ is the set of infinite words in X_σ that have w as a prefix. We can define a measure by defining the measure of the cylinder $[w]$ as the frequency of the finite word w in any element of X_σ, which does exist (by primitivity of σ).

Before we proceed to the next statement about primitive substitutions, let us summarize what we know about them so far. If a substitution σ is primitive, there exists $k \in \mathbb{N}$ such that the k-th power of σ admits a fixed point u. The fixed point u is uniformly recurrent and this is equivalent to the fact that the dynamical system $(\overline{\mathcal{O}(u)}, T)$ is minimal. We define a dynamical system X_σ as $X_\sigma = \overline{\mathcal{O}(u)}$ and this term is well-defined because $\overline{\mathcal{O}(u)} = \overline{\mathcal{O}(v)}$ for any u, v fixed points of any of the powers of σ. We also know that the incidence matrix M_σ has a Perron-Frobenius eigenvalue λ which dominates all the other eigenvalues ($|\lambda'| < \lambda$ for all other eigenvalues λ') and there exist constants $C, C' > 0$ such that for every letter $a \in \mathcal{A}$ we have $C'\lambda^n \leq |\sigma^n(a)| \leq C\lambda^n$. Finally, we have that there exists a constant $C > 0$ such that $p_u(n) \leq Cn$ for all integers n.

Theorem 36. *Let σ be a primitive substitution. Then for any infinite word u which is a fixed point of some σ^k, for some $k \geq 1$, the frequencies of all the factors exist in u. Moreover, (X_σ, T) is minimal and uniquely ergodic. Any of its elements has at most linear factor complexity and it is linearly recurrent.*

Proof. Let us prove the linear recurrence. We assume w.l.o.g. that $\sigma(u) = u$. Let n, k as in the proof of Theorem 35. A factor of u of length n is contained in some $\sigma^k(a)$ or some $\sigma^k(ab)$ since $n < \min_{a\in\mathcal{A}} |\sigma^k(a)|$. Let R be an upper bound on gaps between successive occurences for any letter and any factor of size 2 (it exists by uniforme recurrence). Then, using the same method as in proof of Theorem 35

$$R_u(n) = \min\{N| \text{ every factor of length } N \text{ contains every factor of length } n\} \leq R\frac{C}{C'}\lambda n.$$

Therefore, every factor of length $R\frac{C}{C'}\lambda n$ contains all the factors of length n.

In more details.

- First, we prove that if u is fixed by σ^k, then u is linearly recurrent. Let w be a factor of length n in u, and let p be an integer such that:
$$\min_{a\in\mathcal{A}} |\sigma^{k(p-1)}(a)| \leq n \leq \min_{a\in\mathcal{A}} |\sigma^{kp}(a)|.$$
Then similarly to the proof of Theorem 35, there exist two letters $a, b \in \mathcal{A}$ such that w is a factor of $\sigma^{kp}(ab)$. Additionally, since u is uniformly recurrent, there exists some $m \geq 0$ such that every factor of length 2 belongs in every factor of u of length $\geq m$. Consider now the length between two occurrences of the word w. As the length between two occurrences of ab is bounded by m, we conclude that the length of between two occurrences of w is bounded by $m \cdot \sup_{a\in\mathcal{A}} |\sigma^{kp}(a)|$.
There exist $C, C', \lambda > 0$ such that:
$$\sup_{a\in\mathcal{A}} |\sigma^{kp}(a)| \leq C\lambda^{kp} = \lambda^k \frac{C}{C'}\left(C'\lambda^{k(p-1)}\right) \leq \lambda^k \frac{C}{C'} \inf_{a\in\mathcal{A}} |\sigma^{k(p-1)}(a)| \leq \left(\lambda^k \frac{C}{C'}\right) n.$$
We conclude that if w' is a factor of u of length $\geq \left(m\lambda^k \frac{C}{C'}\right) n$, then every factor of u of length n appears in w'. This means that u is linearly recurrent.
- We conclude by minimality. Let $v \in X_\sigma$. There exists some fixed point u of σ^k such that $X_\sigma = \overline{\mathcal{O}(u)}$. By minimality, $\overline{\mathcal{O}(v)} = \overline{\mathcal{O}(u)}$, which is equivalent to $\mathcal{L}_v = \mathcal{L}_u$.
Consider w' a factor of v of length $\geq \left(m\lambda^k \frac{C}{C'}\right) n$. Then w' is also a factor of u: by the first point, it contains every factor of u of length n, which are exactly the factors of v of length n.

□

Exercise 37. Let $\mathcal{A} = \{a, b\}$ and let $\sigma\colon \mathcal{A} \to \mathcal{A}^*$ be the Fibonacci substitution defined by $\sigma(a) = ab$ and $\sigma(b) = a$.

Prove the following decomposition property: every factor w of u can be written in a unique way as
$$w = r_1\sigma(v)r_2$$
where v is a factor of u (possibly empty), $r_1 \in \{\varepsilon, b\}$, and $r_2 = a$ if the last letter of w is a, and $r_2 = \varepsilon$, otherwise.

Prove that if w is a non-empty left special factor of u, then there exists a unique non-empty left special factor v of u such that $w = \sigma(v)r_2$, where $r_2 = a$ if the last letter of w is a, and

$r_2 = \varepsilon$, otherwise. Give a description of left special factors. Deduce that this infinite word is Sturmian.

Exercise 38. Let u be the Thue-Morse word defined as the fixed point beginning by 0 of the following substitution: $\sigma(0) = 01$ and $\sigma(1) = 10$.

Prove that every factor w can be written as follows: $w = r_1\sigma(x)r_2$, where x is a factor of u (possibly empty) and $r_i \in \{\varepsilon, 0, 1\}$. If $|w| \geq 5$, then this decomposition is unique.

Prove that $p(2n) = p(n) + p(n+1)$ and that $p(2n+1) = 2p(n+1)$, for $n \geq 2$. Give an expression for the complexity function.

3.2. **More on Pisot substitutions.** An algebraic integer λ is a root of a polynomial whose leading coefficient is 1 (it is said monic) with integer coefficients. Its algebraic conjugates are the roots of the unique monic polynomial with integer coefficients with lowest degree having λ as a root (it is called the minimal polynomial of λ). An algebraic integer $\lambda > 1$ is a *Pisot-Vijayaraghavan number* or a *Pisot number* if all its algebraic conjugates λ' other than λ itself satisfy $|\lambda'| < 1$. This class of numbers has been intensively studied and has some special Diophantine properties.

Example 39. The largest roots of X^2-X-1, X^3-X^2-X-1, X^3-X-1 or else X^3-X^2-1 are Pisot numbers.

A primitive substitution is said to be *Pisot* if its expansion number (i.e., its Perron–Frobenius eingenvalue) is a Pisot number.

Example 40. The Fibonacci substitution is a Pisot irreducible substitution.

Theorem 41. *Primitive Pisot substitutions are balanced, and have finite discrepancy.*

Proof. The proof follows the proof of [2, Proposition 11] and uses the Dumont-Thomas prefix-suffix numeration [20].

Let σ be a primitive Pisot substitution over the alphabet $\mathcal{A}$. Let us prove that σ has finite discrepancy. Let $(f_i)_i$ stand for its letter frequency vector. One has in particular $\sum_i f_i = 1$. We consider the abelianization map[1] l defined as the map

$$l\colon \mathcal{A}^* \to \mathbb{N}^d,\ w \mapsto (|w|_1, |w|_2, \cdots, |w|_d).$$

Note that

$$l(\sigma(w)) = M_\sigma l(w),$$

for any word w.

We first consider a fixed word w of the form $w = \sigma^n(j)$, for j letter in $\mathcal{A}$. If i is a fixed letter in $\mathcal{A}$, the sequence $(|\sigma^n(j)|_i)_n$ satisfies a linear recurrence whose coefficients are provided by the minimal polynomial of M_σ. Hence, there exists $C_{i,j}$ such that

$$|\sigma^n(j)|_i = C_{i,j} f_i \lambda^n + O(n^{\alpha_2}|\lambda_2|^n).$$

By applying the Perron–Frobenius Theorem, one checks that there exists C_j such that $C_{i,j} = C_j f_i$ for all i, hence

$$|\sigma^n(j)|_i = C_j f_i \lambda^n + O(n^{\alpha_2}|\lambda_2|^n).$$

We then deduce from $\sum_i f_i = 1$ that

$$|\sigma^n(j)|_i - f_i|\sigma^n(j)| = O(n^{\alpha_2}|\lambda_2|^n).$$

[1]Abelianization comes from the fact that $\mathbb{N}^d$ is the greatest abelian monoid contained in $\mathcal{A}^*$ in the sense that if M is an abelian monoid, every morphism $f : \mathcal{A}^* \to M$ can be uniquely factored as $\bar{f} \circ l$ where $\bar{f} : \mathbb{N}^d \to M$.

The conclusion follows from $|\lambda_2| < 1$ since σ is a Pisot substitution.

It remains to check that this result also holds for all the prefixes of the fixpoint w, and not only for prefixes of the form $\sigma^n(a)$. Indeed, it is easy to prove that any prefix w of u can be expanded as:

$$w = \sigma^k(w_k)\sigma^{k-1}(w_{k-1})\dots w_0,$$

where the w_i belong to a finite set of words. (This corresponds to a "numeration system" on words; there are some admissibility conditions on the possible sequences (w_i), which can be worked out explicitly: they are given by a finite automaton.) This numeration is called Dumont-Thomas numeration. □

In fact, more can be said concerning balance properties of primitive substitutions. Let σ be a primitive substitution and λ be its Perron–Frobenius eigenvalue. Consider the set of eigenvalues of M_σ whose modulus is strictly smaller than λ. Let λ_2 be one of those eigenvalues with maximal multiplicity $\alpha_2 + 1$ in the minimal polynomial of M_σ. Note that several eigenvalues might satisfy this condition.

Theorem 42 ([1, 2])**.** *Let σ be a primitive substitution. Let u be a fixed point of σ.*

- *If $|\lambda_2| < 1$, then the discrepancy $\Delta(u)$ is finite.*
- *If $|\lambda_2| > 1$, then $\Delta_n(u) = (O \cap \Omega)((\log n)^{\alpha_2} n^{(\log_\lambda |\lambda_2|)})$.*
- *If $|\lambda_2| = 1$, and λ_2 is not a root of unity[2] , then*

$$\Delta_n(u) = (O \cap \Omega)((\log n)^{\alpha_2+1}).$$

If λ_2 is a root of unity, then either

$$\Delta_n(u) = (O \cap \Omega)((\log n)^{\alpha_2+1}), \text{ or } \Delta_n(u) = (O \cap \Omega)((\log n)^{\alpha_2}).$$

In particular there exist balanced fixed points of substitutions for which $|\theta_2| = 1$. All eigenvalues of modulus one of the incidence matrix have to be roots of unity.

Observe that the Thue-Morse word is 2-balanced, but if one considers generalized balances with respect to factors of length 2 instead of letters, then it is not balanced anymore.

4. Graphs of Words

4.1. First definitions.

Definition 43 (Graphs of words)**.** Let u be an infinite word over the finite alphabet $\mathcal{A}$ (of cardinality d). The Rauzy graph Γ_n of words of length n of the infinite word u is an oriented graph which is a subgraph of the de Bruijn graph of words[3]. Its vertices are the factors of length n of the infinite word u and the edges are defined as follows: there is an edge from U to V if V follows U in u, i.e., if there exists a word W and two letters x and y such that $U = xW$, $V = Wy$ and xWy is a factor of the infinite word. Hence, it is the graph $\Gamma_n = (V, E)$ such that:

$$\begin{cases} V = \mathcal{L}_u \cap \mathcal{A}^n \\ E = \{(U, V),\ \exists W,\ \exists x, y \in \mathcal{A}, U = xW, V = Wy \text{ and } xWy \in \mathcal{L}_u\}. \end{cases}$$

There are $p_u(n+1)$ edges and $p_u(n)$ vertices, where $p_u(n)$ denotes the factor complexity.

[2]A root of unity α is such that there exists n such that $\lambda^n = 1$.

[3]The de Bruijn graph of words corresponds to the graph of words of an infinite word of maximal complexity ($\forall n$, $p(n) = d^n$) and was introduced by de Bruijn in order to construct circular finite words of length d^n with values in $\{0, 1, \dots, d-1\}$ such that every factor of length n appears once and only once: such a word corresponds to a Hamiltonian closed path in de Bruijn graph.

Exercise 44. Prove that the graphs of words of an infinite word are always connected. Prove the following equivalence:

- the infinite word u is recurrent,
- every factor of u appears at least twice,
- the graphs of words are strongly connected.

Let u be an infinite word over the finite alphabet $\mathcal{A}$ (of cardinality d). Let U be a vertex of the graph Γ_n, for some n. Denote by U^+ the number of edges of Γ_n with origin U and by U^- the number of edges of Γ_n with end vertex U. In other words, U^+ (respectively U^-) counts the number of right (respectively left) extensions of U. Recall that

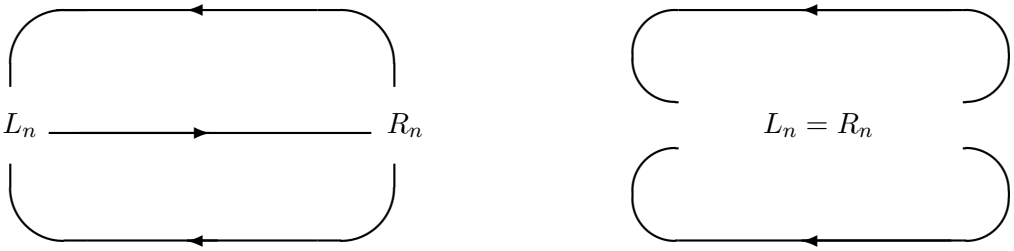
$$p_u(n+1) = \sum_{|U|=n} U^+ = \sum_{|U|=n} U^-,$$

and thus

$$p_u(n+1) - p_u(n) = \sum_{|U|=n} (U^+ - 1) = \sum_{|U|=n} (U^- - 1).$$

Exercise 45. Recall that a Sturmian word u is defined as an infinite word of factor complexity $p_u(n) = n + 1$, for every positive integer n, and that it is recurrent (Exercise 12).

- For any positive integer n, prove that there exists a unique factor of length n having two right (respectively left) extensions: such a factor is called a *right (*respectively *left) special factor* and is denoted from now on by R_n (respectively L_n).
- Prove that the graph of words Γ_n of a Sturmian word has the two following possible forms.



- Deduce from the morphology of the graph of words Γ_n that every Sturmian word is uniformly recurrent. One can first prove that every factor of a Sturmian word is a subfactor of a factor of the form R_n and then deduce from the morphology of the graph Γ_n that R_n appears with bounded gaps.

Exercise 46.

- Prove that if the infinite word u is uniformly recurrent and non-constant, then the graph Γ_n has no edge of the form $U \to U$, for n large enough.

- Suppose that the infinite word u is uniformly recurrent. Prove that if the graph of words Γ_{n+1} is Hamiltonian (i.e., there exists a closed oriented path passing exactly once through every vertex), then the graph Γ_n is Eulerian (there exists a closed path passing exactly once through every edge) and that $U^+ = U^-$, for every vertex of Γ_n. Is the converse true?
- Give the graphs of words of order $1, 2, 3, 4$ for the Thue-Morse word.

4.2. **More on graphs of words and frequencies.** Let us see how to deduce from the morphology of the graphs of words results concerning the frequencies of factors.

In this section we restrict ourselves to infinite words for which the frequencies exist. Observe that the function which associates to an edge labelled by xWy the frequency of the factor xWy is a *flow*. Indeed, it satisfies Kirchhoff's current law: the total current flowing into each vertex is equal to the total current leaving the vertex. This common value is equal to the frequency of the word corresponding to this vertex.

Lemma 47. *Let u be an infinite word which admits all frequencies of words, i.e., all the frequencies of words exist. Let U and V be two vertices linked by an edge such that $U^+ = 1$ and $V^- = 1$. Then the two factors U and V have the same frequency.*

Proof. Write $U = xW$ and $V = Wy$, where x and y are letters. As $U^+ = 1$, U has a unique right extension y. Similarly, V has a unique left extension x. Thus $f_U = f_{Uy} = f_{xWy} = f_{xV} = f_V$, where f denotes the frequency. □

A *branch* of the graph Γ_n is a sequence of maximal length $(U_1, \ldots, U_m)$ of connected edges of Γ_n, possibly empty, satisfying

$$U_i^+ = 1, \quad \text{for} \quad i < m, \quad U_i^- = 1, \quad \text{for} \quad i > 1.$$

Therefore, the edges of a branch have the same frequency and the number of frequencies of factors of given length is bounded by the number of branches of the corresponding graph, as expressed below (see [13]).

Theorem 48. *For a recurrent word of factor complexity $p_u(n)$, the frequencies of factors of given length, say n, take at most $3(p_u(n+1) - p_u(n))$ values.*

Proof. Let V_1 denote the set of factors of length n having more than one extension. In other words V_1 is the subset of vertices of the graph Γ_n defined as follows: $U \in V_1$ if and only if $U^+ \geq 2$. The cardinality of V_1 satisfies

$$\mathrm{card}(V_1) = \sum_{|U|=n,\ U^+\geq 2} 1 \leq \sum_{|U|=n} (U^+ - 1) = p_u(n+1) - p_u(n).$$

Let V_2 denote the subset of vertices of the graph Γ_n defined as follows: $U \in V_2$ if and only if $U^+ = 1$ and if V denotes the unique vertex such that there is an edge from U to V in Γ_n, then $V^- \geq 2$. In other words, U belongs to V_2 if and only if $U = xW$, where x is a letter and where the factor W of the infinite word u has a unique right extension but at least two left extensions. The cardinality of V_2 satisfies:

$$\mathrm{card}(V_2) \leq \sum_{V^-\geq 2} V^- = \sum_{V^-\geq 2} (V^- - 1) + \sum_{V^-\geq 2} 1 \leq 2(p_u(n+1) - p_u(n)).$$

Thus there are at most $3(p_u(n+1) - p_u(n))$ factors in $V_1 \cup V_2$.

Let U be a factor of length n belonging neither to V_1 nor to V_2: $U^+ = 1$ and the unique word V such that there is an edge from U to V in Γ_n satisfies $V^- = 1$. The two factors U

and V thus have the same frequency. Now consider the path of the graph beginning at U and consisting of vertices which do not belong to V_1 nor to V_2. The last vertex of this path belongs to either V_1 or to V_2, and has the same frequency as U. □

Remark 49. In fact we have proved that the frequencies of factors of length n take at most $p_u(n+1) - p_u(n) + r_n + l_n$ values, where r_n (respectively l_n) denotes the number of factors having more than one right (respectively left) extension.

We deduce from this result that if $p_u(n+1) - p_u(n)$ is uniformly bounded with n, the frequencies of factors of given length take a finite number of values. Indeed, using a theorem of Cassaigne quoted below (see [16]), we can easily state the following corollary.

Theorem 50. *If the complexity $p_u(n)$ of an infinite word u on a finite alphabet is sub-affine, i.e.,*

$$\exists (a,b),\ \forall n,\ p_u(n) \leq an + b,$$

then $p_u(n+1) - p_u(n)$ is bounded.

Corollary 51. *If an infinite word has a sub-affine complexity then the frequencies of its factors of given length take a uniform (with respect to the length) finite number of values.*

Note that this does not hold anymore for the second-difference of the complexity $p_u(n+2) + p_u(n) - 2p_u(n+1)$ in the case of a sub-quadratic complexity (see the counterexample in [21]).

5. Sturmian words

Sturmian words provide symbolic codings of translations R_α of the unit circle (that is, the one-dimensional torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$) with

$$R_\alpha \colon \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z},\ x \mapsto x + \alpha \bmod 1.$$

They have been introduced in [33] and widely studied. For more on Sturmian words, see the corresponding chapters in [29, 22] and the references therein.

Definition 52. A word $u \in \{0,1\}^{\mathbb{N}}$ is *Sturmian* if and only if it has exactly $n+1$ factors of length n.

Theorem 53. *The infinite word $u = (u_n)_{n\in\mathbb{N}} \in \{0,1\}^{\mathbb{N}}$ is a Sturmian word if there exist $\alpha \in (0,1)$, $\alpha \notin \mathbb{Q}$, $x \in \mathbb{R}$ such that*

$$\forall n \in \mathbb{N},\ u_n = i \iff R_\alpha^n(x) = n\alpha + x \in I_i \ (mod\ 1),$$

with either $I_0 = [0, 1-\alpha[,\ I_1 = [1-\alpha, 1[$, or $I_0 =]0, 1-\alpha],\ I_1 =]1-\alpha, 1]$.

A Sturmian word is thus a coding of the dynamical system $(\mathbb{T}, R_\alpha)$ with respect either to the two-interval partition $\{I_0 = [0, 1-\alpha[,\ I_1 = [1-\alpha, 1[\}$ or to $\{I_0 =]0, 1-\alpha],\ I_1 =]1-\alpha, 1]\}$.

Sturmian sequences are also characterized by the following properties.

- Sturmian sequences are exactly the non-ultimately periodic balanced sequences over a two-letter alphabet.
- Sturmian sequences are codings of trajectories of irrational initial slope in a square billiard obtained by coding horizontal sides by the letter 0 and vertical sides by the letter 1.
- One can also consider Sturmian sequences as approximations of a line of irrational slope in the upper half-plane.

5.1. **Factors and intervals.** The following lemma is crucial for the study of Sturmian words.

Lemma 54 (Factors lemma)**.** *The word $w = w_1 \cdots w_n$ over the alphabet $\{0,1\}$ is a factor of the Sturmian word u if and only if $I_w := I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n+1} I_{w_n} \neq \emptyset$.*

Proof. By definition, one has

$$\forall i \in \mathbb{N},\ u_n = i \Longleftrightarrow n\alpha + x \in I_i \pmod 1.$$

One first notes that $u_k u_{k+1} \cdots u_{n+k-1} = w_1 \cdots w_n$ if and only if

$$\begin{cases} k\alpha + x \in I_{w_1} (\text{mod } 1) \\ (k+1)\alpha + x \in I_{w_2} (\text{mod } 1) \\ \ldots \\ (k+n-1)\alpha + x \in I_{w_n} (\text{mod } 1). \end{cases}$$

One then applies the density of $(n\alpha)_{n\in\mathbb{N}}$ in $\mathbb{Z}/\mathbb{R}$ (recall that α is assumed to be an irrational number). □

We use here the fact that the sets $I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n+1} I_{w_n}$ are intervals.

Lemma 55. *The sets $I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n+1} I_{w_n}$ are intervals of $\mathbb{T} = \mathbb{R}/\mathbb{Z}$.*

Proof. We prove by induction on n that any intersection $I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n+1} I_{w_n}$ which is not empty is an interval, for $w_1 \cdots w_n \in \{0,1\}^*$. This is true for $n = 1$. We assume that the induction property holds for n. Consider some intersection $I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n} I_{w_{n+1}}$ and assume by contradiction that it is not connected. By induction, $I := I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n+1} I_{w_n}$ is an interval. This implies that $|I| + |I_{w_{n+1}}| > 1$ and all letters are equal to w_{n+1}.We assume $\alpha < 1/2$ without loss of generality. One has $w_{n+1} = 0 = w_n = \cdots = w_1$. But one checks that I_{0^k} is an interval for every k; it is either empty or of the form $[0, 1-k\alpha)$ for k such that $0 < \alpha < 1/k$. We deduce that $I_{w_1} \cap R_\alpha^{-1} I_{w_2} \cap \cdots R_\alpha^{-n} I_{w_{n+1}} = I_{0^{n+1}}$ is an interval. We get the desired since contradiction. □

One first notes that the condition of Lemma 54 does not depend on the point x whose orbit is coded but only on α. Also, it does not depend on the partition $I_0 = [0, 1-\alpha[,\ I_1 = [1-\alpha, 1[$, or $I_0 =]0, 1-\alpha],\ I_1 =]1-\alpha, 1]$.

Furthermore, the factors of u of length n are in one-to-one correspondence with the $n+1$ intervals of $\mathbb{T}$ whose end-points are given by $-k\alpha \bmod 1$, for $0 \leq k \leq n$. This implies that two Sturmian words coding the same rotation R_α have the same factors.

One thus can define the *Sturmian shift* (X_α, T) as the closure in $\{0,1\}^{\mathbb{N}}$ of the orbit of any Sturmian word coding R_α (and also as the closure in $\{0,1\}^{\mathbb{N}}$ of the orbit of all Sturmian words coding R_α). Indeed, since two Sturmian words coding the same rotation have the same set of factors, then one checks that the symbolic dynamical system generated by a Sturmian word coding the rotation R_α consists of *all* the Sturmian words that code the same rotation. The system (X_α, T) is *minimal*: it admits no non-trivial closed and shift-invariant subset.

By Proposition 10, Sturmian words are non-periodic words of smallest factor complexity. This explains why Sturmian words are widely studied and occur in various contexts as models of aperiodic order, for quasiperidoic structures such as quasicrystals (see the books [5]), or else in discrete geometry, as (Freeman) coding discrete lines in discrete geometry. More generally, for references on discrete lines, see the surveys [26, 14].

5.2. **Frequencies of factors.** We now consider properties of frequencies of factors of Sturmian sequences.

One deduces from Lemma 54 not only properties of a topological nature on the number of factors, but also information of a measure-theoretical nature, such as the expression of frequencies of factors [8], that can be deduced from the equidistribution of the sequence $(n\alpha)_{n\in\mathbb{N}}$. Indeed, the frequency of occurrence of the word w in the Sturmian word u is equal to the length of the interval I_w.

Remark 56. Note that we have seen in Section 2.5 that it allows the expression of a shift-invariant measure. Moreover, one checks that Sturmian words are uniquely ergodic: the convergence to frequencies is uniform. Frequencies of factors thus provides the unique shift-invariant measure of the Sturmian shift (X_α, T). We also deduce from the expression of the complexity function that it has zero topological entropy. Moreover, one checks that the systems $(R_\alpha, \mathbb{T})$ and (X_α, T) are measure-theoretically isomorphic, and even semi-conjugate. We thus can consider that the chosen partition provides a good coding. One has the following commutative diagram:

$$\begin{array}{ccc} \mathbb{R}/\mathbb{Z} & \xrightarrow{R_\alpha} & \mathbb{R}/\mathbb{Z} \\ \downarrow & & \downarrow \\ X_\alpha & \underset{\mathrm{S}}{\longrightarrow} & X_\alpha \end{array}$$

The frequency of the factor $w_1 \dots w_n$ exists and is equal to the density of the set

$$\{k \mid \{x + k\alpha\} \in I_{w_1 \cdots w_n}\},$$

which is equal to the length of $I_{w_1 \cdots w_n}$, by uniform distribution of the sequence $(\{x + n\alpha\})_n$. The lengths of these intervals are equal to the frequencies of factors of length n. We deduce from Theorem 48 the following result.

Theorem 57. *The frequencies of factors of given length of a Sturmian sequence take at most three values.*

Theorem 57 implies that the lengths of the intervals $I_{w_1 \cdots w_n}$, and thus the lengths of the intervals obtained by placing the points $0, \{1-\alpha\}, \dots, \{n(1-\alpha)\}$ on the unit circle, take at most three values. We thus have proved the following classical result in Diophantine approximation, called the *three-distance theorem* (see the survey [3]). In fact, this point of view and more precisely, the study of the evolution of the graphs of words with respect to the length n of the factors, allows us to give a proof of the most complete version of the three distance theorem, i.e., to express the exact number of factors having each of the three frequencies and the frequencies themselves.

The three distance theorem was initially conjectured by Steinhaus and proved by V. T. Sós.

Theorem 58. *Let $0 < \alpha < 1$ be an irrational number and n a positive integer. The points $\{i\alpha\}$, for $0 \le i \le n$, partition the unit circle into $n+1$ intervals, the lengths of which take at most three values, one being the sum of the other two.*

More precisely, let $(\frac{p_k}{q_k})_k$ and $(c_k)_k$ be the sequences of the convergents and partial quotients associated to α in its continued fraction expansion (if $\alpha = [0, c_1, c_2, \dots]$, then $\frac{p_n}{q_n} = [0, c_1, \dots, c_n]$). Let $\eta_k = (-1)^k(q_k\alpha - p_k)$. Let n be a positive integer. There exists a unique expression for n of the form

$$n = mq_k + q_{k-1} + r,$$

with $1 \leq m \leq c_{k+1}$ *and* $0 \leq r < q_k$. *Then, the circle is divided by the points* $0, \{\alpha\}, \{2\alpha\}, \ldots, \{n\alpha\}$ *into* $n+1$ *intervals which satisfy:*

- $n+1-q_k$ *of them have length* η_k *(which is the largest of the three lengths),*
- $r+1$ *have length* $\eta_{k-1} - m\eta_k$,
- $q_k - (r+1)$ *have length* $\eta_{k-1} - (m-1)\eta_k$.

5.3. **More on Sturmian words and substitutions.** Let us consider now a combinatorial way of generating Sturmian words.

Let us see how to generate all the Sturmian shifts with substitutions. We work now on the alphabet $\{a, b\}$. We consider the substitutions τ_a and τ_b defined over the alphabet $\mathcal{A} = \{a, b\}$ by $\tau_a \colon a \mapsto a, b \mapsto ab$ and $\tau_b \colon a \mapsto ba, b \mapsto b$. Let $(i_n) \in \{a, b\}^{\mathbb{N}}$. The following limits

$$u = \lim_{n\to\infty} \tau_{i_0}\tau_{i_1} \cdots \tau_{i_{n-1}}(a) = \lim_{n\to\infty} \tau_{i_0}\tau_{i_1} \cdots \tau_{i_{n-1}}(b) \tag{1}$$

exist and coincide whenever the directive sequence $(i_n)_n$ is not ultimately constant (it is easily shown that the shortest of the two images by $\tau_{i_0}\tau_{i_1} \ldots \tau_{i_{n-1}}$ is a prefix of the other). One checks that the infinite words thus produced are all Sturmian words: indeed, it suffices to consider and compute their factor complexity. More generally, one can prove that a Sturmian word is an infinite word whose set of factors coincides with the set of factors of a sequence u of the form (1), with the sequence $(i_n)_{n\geq 0}$ being not ultimately constant (that is, it is an element of the symbolic dynamical system X_u generated by u, since (X_u, T) is minimal). The proof relies on the fact that in a Sturmian language, either aa (the letter b occurs as an isolated letter) or bb (a is isolated) occurs: one cannot have simultaneously aa and bb since there are 3 factors of length 2. One then desubstitutes according to the isolated letter: if b is isolated in u, then one can write u as $u = \sigma_a(v)$ (one reduces the ranges of successive occurrences of a' by 1). One checks that v (possibly up to a prefix letter) is again a Sturmian word (associated with a different α). If one wants to generate a specific Sturmian word (not only a Sturmian language/shift), one can use four substitutions. One striking property of Sturmian words is the following: the way one iterates the substitutions is governed by the continued fraction expansion of α. This method can be used for the generation of discrete lines and planes in discrete geometry, as well as for the recognition of discrete planes. More generally shifts with at most linear factor complexity can also be generated in terms of composition of substitutions, see e.g. the survey [9] and the references therein.

6. Hints and corrections for exercices

Exercise 11 Consider the bi-infinite word $\cdots 0 \cdots 010 \cdots 0 \cdots$. It is ultimately periodic on the right and on the left but its factor complexity is $n+1$.

Exercise 22 Consider a clopen set C. Since it is open and cylinders form a basis of the topology, C can be written as a union of cylinders. On the other hand, because C is closed in a compact space, C is itself compact. So from the possibly infinite union of cylinders, one can extract a finite union that still covers C. As the union not only covered C, but was also equal to C, we conclude that C is the union of this finite family of cylinders.

Exercise 23

- Let u be a recurrent infinite word. Let $u[: n]$ be the prefix of size n of u. Take $n_1 = 0$, and assume that $(n_k)_{k\leq K}$ defines a finite and strictly increasing sequence such that $u[: k]$ is a prefix of $T^{n_k}(u)$ for any $k \leq K$. Then $u[: K+1]$ is a factor of u, so it

appears infinitely many times in u (because u is recurrent). In particular, there exists some $n_{K+1} > n_K$ such that $u[: K+1]$ is a prefix of $T^{n_{K+1}}(u)$.
By induction, we conclude that there exists a strictly increasing sequence of positive integers $(n_k)_{k\in\mathbb{N}}$ such that $u[: k]$ is a prefix of $T^{n_k}(u)$ for any $k \in \mathbb{N}$. In other words, $(T^{n_k}(u))_{k\in\mathbb{N}}$ is a converging sequence, and it converges towards u.

- Reciprocally, assume that $u = \lim_{k\to+\infty} T^{n_k}(u)$ for some strictly increasing sequence $(n_k)_{k\in\mathbb{N}}$. And let w be a factor of u. There exists some $N \in \mathbb{N}$ such that w is a prefix of $T^N(u)$. By continuity of T^N, one has $T^N(u) = \lim_{k\to+\infty} T^{N+n_k}(u)$. And by definition of convergence, there exists some $K \in \mathbb{N}$ such that for any $k \geq K$, the first $|w|$ letters of $T^{N+n_k}(u)$ are the same; and as such they are then equal to w.
  In particular, w appears infinitely many times in u.

Exercise 25 Suppose X minimal. Let $u \in X$. Then $\overline{\mathcal{O}(u)} \subset X$, with $\overline{\mathcal{O}(u)}$ a subset both closed and stable by the shift. So, since $\overline{\mathcal{O}(u)}$ is not empty (it contains u), by minimality, $X = \overline{\mathcal{O}(u)}$. Suppose now that $\forall u \in X, \overline{\mathcal{O}(u)} = X$. Let $Y \subset X$ such that it is non-empty, closed, and shift-invariant. Let $v \in Y$. Then, since Y is closed and shift-invariant, $\overline{\mathcal{O}(v)} \subset Y$. But $\overline{\mathcal{O}(v)} = X$, which yields $Y = X$. So X is minimal.

Another redaction.

- Let $u \in X$. $\overline{\mathcal{O}(u)}$ is a closed, non-empty and shift-invariant subset of X. If $\overline{\mathcal{O}(u)} \neq X$, then X is not minimal.
- Reciprocally, assume that X is not minimal: there exists some Y such that Y is a proper, closed, non-empty and shift-invariant subset of X. Consider then some $u \in Y$. One has $\overline{\mathcal{O}(u)} \subseteq Y$, because Y is closed, and then $Y \subsetneq X$: in particular, $\overline{\mathcal{O}(u)} \neq X$.

Exercise 37 Consider first the existence. As $\sigma(u) = u$, and w is a factor of u, we can look for the "antecedents" of w under σ. We reason by induction.

(1) Initialisation. If $w = a$, then $r_2 = a$, $r_1 = \varepsilon = v$ works ; if $w = b$, $r_1 = b$, $v = r_2 = \varepsilon$ works.
(2) Induction. Let $w = xw'$, where x is the first letter of w. By induction hypothesis, there exists some r'_1, v', r'_2 such that $w' = r'_1\sigma(v')r'_2$. We take $r_2 = r'_2$, and:
   - If $r'_1 = \varepsilon$. If $x = a$, define $v = bv'$ and $r_1 = \varepsilon$. If $x = b$, define $v = v'$ and $r_1 = b$.
   - If $r'_1 = b$. If $x = a$, define $v = av'$ and $r_1 = \varepsilon$. And x cannot be equal to b, otherwise there would be a factor bb in u, which is impossible.

   Then $w = r_1\sigma(v)r_2$ is a valid decomposition.

Consider now the unicity. Assume that $w = r_1\sigma(v)r_2 = r'_1\sigma(v')r'_2$ as above. Then $r_2 = r'_2$ because the value of r_2 is determined by the last letter of w. Assume $r_1 \neq r'_1$, then without any loss of generality we can assume that $r_1 = b$: this would imply that $r'_1 = \varepsilon$, and that b has an antecedent under σ, which is impossible. So $r_1 = r'_1$. This proves that this decomposition is unique.

Prove that if w is a non-empty left special factor of u, then there exists a unique non-empty left special factor v of u such that $w = \sigma(v)r_2$, where $r_2 = a$ if the last letter of w is a, and $r_2 = \varepsilon$, otherwise. Give a description of left special factors. Deduce that this infinite word is Sturmian.

By definition, a non-empty left special factor w of u has two left extensions. Because the alphabet has two letters, and that u has no factor bb, this means that w starts with the letter

a. In the previous proof, one can then take $r_1 = \varepsilon$, and the decomposition is unique, which concludes the proof.

By an easy induction, we prove that all the $\sigma^n(a)$ are left special factors (indeed, $\sigma^0(a) = a$ is a left special factor, and $\sigma^{n+1}(a) = \sigma(\sigma^n(a))$; by induction, $\sigma^n(a)$ is a left special factor, which implies that both $\sigma(a\sigma^n(a))$ and $\sigma(b\sigma^n(a))$ are factors of $\sigma(u) = u$, which proves that both $a\sigma^{n+1}(a)$ and $b\sigma^{n+1}(b)$ are factors of u). We now prove by induction on $|w|$ all left special factors are prefixes of the words $\sigma^n(a)$.

- $|w| = 1$: then a is the only left special factor of size 1 (because bb is not a factor of u), and $a = \sigma^0(a)$.
- Let w be a left special factor of length at least 2. Then by the decomposition lemma, there exists a left special factor w' such that $w = \sigma(w')r_2$, where $r_2 = a$ if w ends with an a, and $r_2 = \varepsilon$ otherwise.
  First, $|w'| < |w|$. Indeed, because w is a left special factor, the first letter of w must be a; and because aaa is not a factor of u, the second letter of w must be b, which implies that the first letter of w' is an a: the existence of an a in w' proves that $|w| \geq |\sigma(w')| > |w'|$.
  By induction hypothesis, because w' is a left special factor of size $< |w|$, w' is a prefix of some $\sigma^j(a)$:
  - If $r_2 = \varepsilon$, then w is a prefix of $\sigma^{j+1}(a)$.
  - If $r_2 = a$, then because w' is right-extensible by a letter x, that w' is a prefix of some $\sigma^j(a)$ (and that $\sigma^j(a)$ is a prefix of $\sigma^{j+1}(a)$), $w'x$ is a prefix of some $\sigma^{j+1}(a)$: which implies that w is a prefix of $\sigma(w'x)$ which is a prefix of $\sigma^{j+2}(a)$.

With this, we conclude that there is a single left special factor of any given left (for $n \in \mathbb{N}$, it is the prefix of size n of any $\sigma^j(a)$ such that $|\sigma^j(a)| \geq n$). As there is exactly one left special factor of any given length, we conclude by an easy induction that $p_u(n+1) = (p_u(n)-1)+2 = p_u(n)+1$, and $p_1(u) = 2$, which proves that u is Sturmian.

Exercise 38 A factor w appearing in u Thue-Morse word is either an image of a single factor x ($w = \sigma(x)$), in which case $r_i = \varepsilon, i = 1, 2$, or it consists of $\sigma(x)$ which is prolonged either to the left by one letter r_1 or to the right by one letter r_2 or both. It cannot be prolonged by two letters to neither side because then x could have been chosen larger.

If $|w| = 1$, $x = \varepsilon$ and one of r_i is equal to w (therefore this decomposition is not unique). If $|w| \in 1, 2, 3, 4$, the decomposition is unique if and only if the choice of x is unique, and we can find examples of factors where x can be chosen in two ways (e.g. $01, 101, 0101$).

However, if $|w| \geq 5$, the choice of x is unique. Let us have the decomposition $w = r_1\sigma(x)r_2$. Taking a subfactor $\tilde{x}$ of x into the decomposition would lead to a decomposition $w = s_1\sigma(\tilde{x})s_2$ where the length of at least one of s_i would be at least 2, and therefore it does not fulfill the conditions of the decomposition we chose. On the other hand, choosing a larger $\tilde{x}$ than x (e.g. $\tilde{x} = ax$ for some $a \in \mathcal{A}$, the procedure would be analogical for $\tilde{x} = xa$) would lead to

$$\sigma(\tilde{x})r_2 = \underbrace{\sigma(a)}_{\text{length}\geq 2}\ \sigma(x)r_2$$

having larger length than w and therefore not being the decomposition of w at all.

Let us state the factors for $n = 1, 2, 3, 4$:

- $n = 1$: $0, 1$
- $n = 2$: $01, 10, 00, 11$

- $n = 3$: $001, 010, 011, 100, 101, 110$
- $n = 4$: $0010, 0011, 0100, 0101, 0110, 1001, 1010, 1011, 1100, 1101$

Factor complexity for even numbers: When $n = 2$, we can see that $p(4) = p(2) + p(3)$. Let us prove the relation $p(2n) = p(n)+p(n+1)$ for $n \geq 3$. A factor w of length $2n$ is decomposed uniquely as $r_1\sigma(x)r_2$ and we can find there are two types of w.

Either $r_i = \varepsilon$ and therefore $|x| = n$. There are $p(n)$ factors of length n and each of them creates one factor of length $2n$.

Or, there is i that $r_i \neq \varepsilon$ and the even length of w implies that $r_i \neq \varepsilon, i = 1, 2$. Then, $|\sigma(x)| = 2n - 2$, therefore $|x| = n - 1$ and each factor w of this type is inside of an image of axb, $a, b \in \mathcal{A}$ for all factors in the language of length $1 + (n - 1) + n$, i.e. there are $p(n + 1)$ factors w of this type.

On the whole, we have $p(n) + p(n + 1)$ factors of length $2n$ which proves the relation.

Factor complexity for odd numbers: Let us prove $p(2n + 1) = 2p(n + 1)$ for $n \geq 2$. Then a factor w of length $2n + 1$ is decomposed uniquely as $r_1\sigma(x)r_2$ and the odd length of w implies that exactly one of r_i will be non-empty. Therefore any factor of length $2n + 1$ can be decomposed either as $a\sigma(x)$ or $\sigma(x)a$ for some $a \in \mathcal{A}$ where $|x| = n$ and the amount of w will be exactly the amount of factors of length $n + 1$ multiplied by 2 because for each y of length $(n + 1)$ we create w by dropping either leftmost or rightmost letter in $\sigma(y)$. Let the reader think about why it cannot happen that dropping letters from two different factors y create the same factor w.

A second redaction. Existence:
First, once again, $u = \sigma(u)$. So if w is a factor of u, and that σ maps each letter to words of size 2, there is some factor x of u such that $\sigma(x)$ is contained in w, and such that $w = r_1\sigma(x)r_2$, for $r_1, r_2 \in \{\varepsilon, 0, 1\}$.
Unicity:
If w is of length at least 5, then a factor 00 or 11 must appear in w. Indeed, $\sigma(u) = u$, and the words 000 and 111 cannot be factors of u. Then, if w is a factor of length at least 5, the decomposition is unique because 11 and 00 cannot be in the image of σ (in other words, there is only one way to split w into pairs of letters, modulo the first and the last one).

Prove that $p(2n) = p(n) + p(n + 1)$ and that $p(2n + 1) = 2p(n + 1)$, for $n \geq 2$. Give an expression for the complexity function. First, $p(4) = 10$ and $p(5) = 12$, so the formula works (see OEIS A005942). Now, let $n \geq 3$: we can use the decomposition lemma above:

- Compute $p(2n)$: let w be a factor of u of length $2n$. There are two possibilities (which are disjoint, by unicity of the decomposition):
  - $w = \sigma(x)$ for x of factor of u of length n. Because σ is injective, we obtain the term $p(n)$ in the expression $p(2n) = p(n) + p(n + 1)$.
  - $w = r_1\sigma(x)r_2$ for $r_1, r_2 \neq \varepsilon$, with x a factor of u of length $n - 1$. Consider x' an extension of x of length $n + 1$, $x' = axb$ for $a, b \in \mathcal{A}$ (it necessarily exists because factors appear infinitely often in u). Then w is $\sigma(x')$ without its first and its last letter, and the map $(x' \in \mathcal{L}_u(n + 1) \mapsto \sigma(x')[1 : 2n])$ is bijective. We obtain the term $p(n + 1)$ in the expression $p(2n) = p(n) + p(n + 1)$.
- Compute $p(2n + 1)$: let w be a factor of u of length $2n$. There are two possibilities (which are disjoint, by unicity of the decomposition):
  - $w = r_1\sigma(x)$ for $x \in \mathcal{L}_u(n)$. Then similarly, we consider the extension $x' = ax$ of x and $x' \in \mathcal{L}_u(n + 1) \mapsto \sigma(x')[1 : 2n]$ (forget the first letter of $\sigma(x')$), which

is bijective. From there comes one term $p(n+1)$ in the expression $p(2n+1) = 2p(n+1)$.
- $w = \sigma(x)r_2$ for $x \in \mathcal{L}_u(n)$. This case is completely symmetric, and there comes the other term $p(n+1)$ in $p(2n+1) = 2p(n+1)$.

We conclude that $p(2n) = p(n) + p(n+1)$ and $p(2n+1) = 2p(n+1)$ for $n \geq 2$.

We now prove by recurrence that if $n-1 = 2^j + r$, with $-2^{j-2} \leq r \leq 2^{j-1}$, then one has:

$$p(n) = 3(n-1) + |r|$$

(In other words, $|r|$ is the distance between $n-1$ and the nearest power of 2)

- Initialization: $3 \cdot 2 + 0 = 6 = p(3)$ and $3 \cdot 3 + 1 = 10 = p(4)$.
- Assume that for any $k \leq 2n$, $p(k)$ is equal to the form above.
  - Computation of $p(2n+1)$. We write $n = 2^j + r$ with $-2^{j-2} \leq r \leq 2^{j-1}$. Then $2n = 2^{j+1} + 2r$, obviously $-2^{j-1} \leq 2r \leq 2^j$ and:

$$\begin{aligned} p(2n+1) &= 2p(n+1) \\ &= 2 \cdot (3 \cdot ((n+1)-1) + r) \\ &= 3 \cdot ((2n+1)-1) + 2r \end{aligned}$$

  - Computation of $p(2n+2)$. We write $n = 2^j + r$, with $-2^{j-2} \leq r \leq 2^{j-1}$. There are now two possibilities.
    First, if $r = 2^{j-1}$, then $n+1 = 2^{j+1} - (2^{j-1} - 1)$ and

$$\begin{aligned} p(2n+2) &= p(n+1) + p(n+2) \\ &= \left(3 \cdot ((n+1)-1) + 2^{j-1}\right) + \left(3 \cdot ((n+2)-1) + 2^{j-1} - 1\right) \\ &= 3 \cdot ((2n+2)-1) + (2^j - 1) \end{aligned}$$

    and $2n+1 = 2(n+1) - 1 = 2^{j+2} - (2^j - 2) - 1 = 2^{j+2} - (2^j - 1)$.
    Second, if $0 \leq r < 2^{j-1}$, then $n+2 = 2^j + r + 1$ and

$$\begin{aligned} p(2n+2) &= p(n+1) + p(n+2) \\ &= (3 \cdot ((n+1)-1) + r) + (3 \cdot ((n+2)-1) + (r+1)) \\ &= 3 \cdot ((2n+2)-1) + (2r+1) \end{aligned}$$

    and $2n+1 = 2n+1 = 2^{j+1} + 2r + 1$ with $0 \leq 2r+1 \leq 2^j$.
    Lastly, if $2^{j-2} < r < 0$, then $n+2 = 2^j + r + 1$ and

$$\begin{aligned} p(2n+2) &= p(n+1) + p(n+2) \\ &= (3 \cdot ((n+1)-1) - r) + (3 \cdot ((n+2)-1) - (r+1)) \\ &= 3 \cdot ((2n+2)-1) - (2r+1) \end{aligned}$$

    and $2n+1 = 2(n+1) - 1 = 2^{j+1} + (2r+1)$, with $2^{j-1} < 2r+1 \leq 0$.
    In all the previous cases, the formula holds.

So we conclude that that for $n \geq 3$, if $n-1 = 2^j + r$, with $-2^{j-2} \leq r \leq 2^{j-1}$, then one has:

$$p(n) = 3(n-1) + |r|.$$

Exercise 44 The equivalence between (1) and (2) is immediate.

(2) $\Rightarrow$ (3) : It is clear that a graph of words of an infinite word is connected. If there exists n such that Γ_n is not strongly connected, it means that there is a vertex U which does not

have an ingoing edge. But then the factor corresponding to U only appears in the word u once and that is a contradiction.

(3) $\Rightarrow$ (2) : If there is a factor in u appearing only once, the corresponding vertex U in the graph of words has no ingoing edge. This is a contradiction to all graphs of words being strongly connected.

Let u be an infinite word over the finite alphabet $\mathcal{A}$ (of cardinality d). Let U be a vertex of the graph Γ_n, for some n. Denote by U^+ the number of edges of Γ_n with origin U and by U^- the number of edges of Γ_n with end vertex U. In other words, U^+ (respectively U^-) counts the number of right (respectively left) extensions of U. Recall that

$$p_u(n+1) = \sum_{|U|=n} U^+ = \sum_{|U|=n} U^-,$$

and thus

$$p_u(n+1) - p_u(n) = \sum_{|U|=n} (U^+ - 1) = \sum_{|U|=n} (U^- - 1).$$

Another redaction.

- The graphs of an infinite word are always connected: consider some $n \in \mathbb{N}$. Then from the prefix of size n of u, you can access any other factor of size n, ie. there is a path from the prefix of size n to any other vertex of Γ_n.
- Assume that u is recurrent: we prove that every graph of word is strongly connected. Indeed, consider Γ_n. Because u is recurrent, if w and w' are two factors of u of length n, there exists a finite word v such that wvw' is a factor of u. In particular, if we label the edges of Γ_n with letters (if $xW \to Wy$ exists in Γ_n, then label it with y), we can start in vertex w, read vw' in the automaton, and we end up in the vertice w' in Γ_n. This proves that Γ_n is strongly connected.
- Assume that the graphs are strongly connected: we prove that every factor of u appears at least twice.
  Let w be a factor of u, and n be an integer such that w appears in the prefix of u of size n. In Γ_n, the prefix $u[:n]$ is a vertex which has an incoming arrow (by strong connectivity). So $u[:n]$ must appear somewhere else in u. This implies that w appears at least twice.
- Assume that every factor of u appears at least twice: we prove that u is recurrent.
  Let $w_0 = w$ be a factor of u. Because it appears at least twice, there exists a word v_0 such that $w_1 = wv_0w$ is a factor of u. Because w_1 is a factor of u, it must appear at least twice: there exists a word v_1 such that $w_2 = w_1v_1w_1$ is a factor of u. Etc... By induction, we prove that for every $k \in \mathbb{N}$, w must appear at least 2^k times. This proves that w appears infinitely often.

Exercise 45

Let n be a positive integer, and s_n be the number of right special factors of size n. Then $p_u(n+1) = (p_u(n) - s_n) + 2s_n = p_u(n) + s_n$; so we obtain that $s_n = 1$ (because u is Sturmian). (The very same proof applies for left special factors).

First, $L_n^- = 2$ and $R_n^+ = 2$; for every other vertex, $U^+ = U^- = 1$. There is a single left (resp. right) special factor of length n, and because Sturmian words are recurrent, the graphs must be strongly connected. This nearly proves that Γ_n is of these two possible forms: the

only thing left to prove is that there is no path from R_n to itself (or L_n to itself) that does not contain L_n (resp R_n) [Remark: this statement does not assume that $R_n \neq L_n$].

Let $R_n \to^+ R_n$ be a path between R_n and itself. If its length is smaller that n, iterate this path again until its length becomes large enough. Consider then L, the n^{th} vertex of this path. Because R_n is right special, L must be left special. As there is a single left special vertex (ie. L_n), one has $L = L_n$. So L_n appears in the original path.

Deduce from the morphology of the graph of words Γ_n that every Sturmian word is uniformly recurrent. One can first prove that every factor of a Sturmian word is a subfactor of a factor of the form R_n and then deduce from the morphology of the graph Γ_n that R_n appears with bounded gaps.

Consider w a factor of length n. Because Γ_n is strongly connected, there exists a path from w to R_n in Γ_n: in other words, there exists a finite word v such that wv is a right special factor of u. This proves that w is a subfactor of some R_N. We then prove that each R_n appears with bounded gaps: consider the graph Γ_n, and let l_n be the length of the longest path from R_n to itself which does not contain R_n as an intermediary vertex. Then R_n occurs with gaps bounded by l_n. Combining the two considerations, we obtain that any Sturmian word is uniformly recurrent.

Exercise 46

- *Prove that if the infinite word u is uniformly recurrent and non-constant, then the graph Γ_n has no edge of the form $U \to U$, for n large enough.*

  Because u is non-constant, there exist two letters a and b which appear in u. Then by uniform recurrence, there exists some $m \in \mathbb{N}$ such that any factor of u of length $\geq m$ necessarily contains both letters a and b. Then for any $n \geq m$, there is no edge $U \to U$ in Γ_n (because all the vertices in such a Γ_n contain a subfactor xy for x, y two different letters, so each vertex represents a factor that is not shift-invariant).
- *Suppose that the infinite word u is uniformly recurrent. Prove that if the graph of words Γ_{n+1} is Hamiltonian (i.e., there exists a closed oriented path passing exactly once through every vertex), then the graph Γ_n is Eulerian (there exists a closed path passing exactly once through every edge) and that $U^+ = U^-$, for every vertex of Γ_n.*

  Once again, we label the edges of the graphs of words. Let

  $$x_1w_1 \xrightarrow{a_1} x_2w_2 \xrightarrow{a_2} \dots \xrightarrow{a_{|\Gamma_{n+1}|}} x_{|\Gamma_{n+1}|}w_{|\Gamma_{n+1}|} = x_1w_1$$

  be an Hamiltonian circuit of Γ_{n+1}. We prove that

  $$w_1 \xrightarrow{a_1} w_2 \xrightarrow{a_2} \dots \xrightarrow{a_{|\Gamma_{n+1}|}} w_{|\Gamma_{n+1}|} = w_1$$

  is an Eulerian path of Γ_n. Indeed, it is clearly a path of Γ_n. Assume now that there exists some $i < j < |\Gamma_{n+1}|$ such that $w_i \xrightarrow{a_i} w_{i+1} = w_j \xrightarrow{a_j} w_{j+1}$ (i.e., we assume that this path is not Eulerian). Then define $a = a_i = a_j$, $w = w_i = w_j$. In Γ_{n+1}, we have some $x_iw \xrightarrow{a} wa$ at the i^{th} step of the Hamiltonian path, and $x_jw \xrightarrow{a} wa$ at the j^{th} step of the Hamiltonian path: this is absurd, because the vertex wa is visited twice (and it is not the vertex we start from). So this path is Eulerian, which means that Γ_n is Eulerian. And in particular, in an Eulerian graph, for any vertex U we have $U^+ = U^-$ (for example, assume $U^+ < U^-$: then a path that visits only once each edge must get stuck in U at some point, because it enters it more than it exits it; and if $U^- < U^+$, then you cannot visit all the exiting edges of U).

Is the converse true?
No. See the counter-example on the infinite word: $u = (10011001000)^\omega$.

References

[1] B. Adamczewski, *Balances for fixed points of primitive substitutions*, Theoret. Comput. Sci. **307** (2003), 47–75.
[2] B. Adamczewski, *Symbolic discrepancy and self-similar dynamics*, Ann. Inst. Fourier (Grenoble) **54** (2004), 2201–2234.
[3] P. Alessandri, V. Berthé, *Three Distance Theorems and Combinatorics on Words*, Enseig. Math. **44**, pp. 103–132 (1998).
[4] J.-P. Allouche and J. O. Shallit, *Automatic sequences: Theory and Applications*, Cambridge University Press, 2002.
[5] M. Baake, U. Grimm, Aperiodic order. *Vol. 1. A mathematical invitation*, Encyclopedia of Mathematics and its Applications **149**, Cambridge University Press, 2013.
[6] M.-P. Béal and D. Perrin, *Symbolic dynamics and finite automata*, in *Handbook of formal languages, Vol. 2*, Springer, Berlin, 1997.
[7] J. Berstel, J. Karhumäki, *Combinatorics on words– A tutorial*, http://www-igm.univ-mlv.fr/~berstel/.
[8] V. Berthé, *Fréquences des facteurs des suites sturmiennes*, Theoret. Comput. Sci. **165** (1996), 295–309.
[9] V. Berthé, V. Delecroix, *Beyond substitutive dynamical systems: S-adic expansions*, RIMS Lecture note 'Kokyuroku Bessatu' **B46** (2014), 81–123.
[10] V. Berthé, M. Rigo (Eds), Combinatorics, Automata and Number Theory, Encyclopedia Math. Appl. **135**, Cambridge Univ. Press, Cambridge, 2010.
[11] V. Berthé, M. Rigo (Eds), Combinatorics, Words and Symbolic dynamics, Encyclopedia Math. Appl. **159**, Cambridge University Press (2016).
[12] P. Billingsley, *Ergodic theory and information*, John Wiley & Sons Inc., New York, 1965.
[13] M. Boshernitzan, *A Condition for Minimal Interval Exchange Maps to be Uniquely Ergodic*, Duke Math. J. **52**, pp. 723–752 (1985).
[14] V. E. Brimkov, D. Coeurjolly, and R. Klette, *Digital planarity - a review*, Discrete Applied Mathematics **155** (2007), 468–495.
[15] J. Buzzi, `https://jbuzzi.wordpress.com/2013/10/18/les-surfaces-a-courbures-opposees-et-leurs-lignes-geodesiques-jacques-hadamard-1898/`.
[16] J. Cassaigne, *Special Factors of Sequences with Linear Subword Complexity*, Developments in Language Theory II (DLT'95) (Dassow, Rozenberg, Salomaa eds) World Scientific, pp. 25–34 (1996).
[17] J. Cassaigne, *Complexité et Facteurs Spéciaux*, Bull. Belg. Math. Soc. **4**, pp. 67–88 (1997).
[18] I. P. Cornfeld, S. V. Fomin and Y. G. Sinaĭ, *Ergodic theory*, Springer-Verlag, New York, 1982.
[19] K. Dajani and C. Kraaikamp. *Ergodic Theory of Numbers*, The Math. Association of America, 2002.
[20] J.-M. Dumont, and A. Thomas, Systèmes de numération et fonctions fractales relatifs aux substitutions, *Theoret. Comput. Sci.*, **65** (1989), 153–169.
[21] S. Ferenczi, *Rank and Symbolic Complexity*, Erg. Theory Dynam. Sys. **16**, pp. 663–682 (1996).
[22] N. Pytheas Fogg. *Substitutions in dynamics, arithmetics and combinatorics*, volume **1794** of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.
[23] J. Hadamard, *Les surfaces à courbures opposées et leurs lignes géodésiques*, Journal de Mathématiques pures et appliquées 5ème série, tome 4 (1898), 27–74.
[24] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Cambridge University Press, Cambridge, 1995.
[25] B. P. Kitchens, *Symbolic dynamics. One-sided, two-sided and countable state Markov shifts*, Universitext. Springer-Verlag, Berlin, 1998.
[26] R. Klette and A. Rosenfeld. *Digital straightness-a review.* Discrete Applied Mathematics **139** (2004), 197–230.
[27] D. Lind, B. Marcus, *An introduction to symbolic dynamics and coding*, Cambridge University Press, Cambridge, 1995.
[28] M. Lothaire, *Combinatorics on words*, Cambridge University Press, Cambridge, 1997. Second ed.
[29] M. Lothaire. *Algebraic combinatorics on words*, volume 90 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, 2002.
[30] M. Lothaire, *Applied Combinatorics on words*, Cambridge University Press, Cambridge, (2002), http://www-igm.univ-mlv.fr/~berstel/Lothaire.

[31] H. M. Morse, *Recurrent geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 84–100.
[32] M. Morse, G. A. Hedlund, *Symbolic Dynamics*, Amer. J. Math. **60** (1938), 815–866.
[33] M. Morse, G. A. Hedlund, *Symbolic dynamics II. Sturmian trajectories*, Amer. J. Math. **62** (1940), 1–42.
[34] K. Petersen, *Ergodic theory*, Cambridge University Press, Cambridge, 1989.
[35] M. Queffélec, *Substitution Dynamical Systems. Spectral Analysis*, Lecture Notes in Math. **1294**, Springer-Verlag (2010).
[36] A. Thue, Selected mathematical papers, Universitetsforlaget, Oslo, 1977.
[37] P. Walters, *An introduction to ergodic theory*, Springer-Verlag, New York, 1982.