Recurrence in infinite words

(extended abstract)

Julien Cassaigne

Institut de Mathématiques de Luminy,
Case 907, F-13288 Marseille Cedex 9, France
cassaigne@iml.univ-mrs.fr

Abstract. We survey some results and problems related to the notion of recurrence for infinite words.

1 Introduction

The notion of recurrence comes from the theory of dynamical systems. A system $T : X \to X$ is *recurrent* when any trajectory eventually returns arbitrarily near its starting point, or in more formal terms, when for any open subset U of X and any $x \in U$, there exists an integer $n \geq 1$ such that $T^n(x) \in U$. And if this n — the *return time* — can be chosen independently of x for a given U, the system is said to be *uniformly recurrent*.

When $X = \overline{O(\mathbf{u})}$ is the subshift generated by an infinite word $\mathbf{u}$, the recurrence of X can be expressed as a combinatorial property of the word $\mathbf{u}$. Moreover, it is possible to compute return times and this allows to quantify the speed of recurrence in the system, via the *recurrence function* of the word $\mathbf{u}$. This point of view was initiated by Morse and Hedlund in their 1938 article on symbolic dynamics [13].

In this article, we survey some results and problems concerning recurrence in infinite words.

2 Preliminaries

Let A be a finite alphabet, with at least two elements. We denote by A^* the set of finite words over A (i.e., the free monoid generated by A), including the empty word ε, and by $A^{\mathbb{N}}$ the set of one-way infinite words over A. Given an infinite word $\mathbf{u} \in A^{\mathbb{N}}$, we denote by $F(\mathbf{u})$ the set of factors (or subwords) of $\mathbf{u}$, and, for any $n \in \mathbb{N}$, by $F_n(\mathbf{u})$ the set of factors of length n of $\mathbf{u}$.

The *shift* is the operator T on $A^{\mathbb{N}}$ defined by $T(u_0u_1u_2u_3\ldots) = u_1u_2u_3\ldots$. The set $A^{\mathbb{N}}$ equipped with the product topology is a compact topological space, and under the action of T it becomes a discrete dynamical system named the *full shift*. A closed subset of $A^{\mathbb{N}}$ invariant under T is a *subshift*, and in particular any infinite word $\mathbf{u} \in A^{\mathbb{N}}$ generates a subshift $\overline{O(\mathbf{u})}$, the adherence of $O(\mathbf{u}) = \{T^n(\mathbf{u}) : n \in \mathbb{N}\}$.

3 Recurrence

3.1 Recurrent and uniformly recurrent words

An infinite word $\mathbf{u}$ is said to be *recurrent* if any factor of $\mathbf{u}$ occurs infinitely often in $\mathbf{u}$. Recurrence can be characterized using an apparently much weaker property:

Proposition 1. *An infinite word $\mathbf{u} \in A^{\mathbb{N}}$ is recurrent if and only if any prefix of $\mathbf{u}$ occurs at least twice in $\mathbf{u}$.*

An infinite word is said to be *uniformly recurrent* if it is recurrent and additionally, for any factor w of $\mathbf{u}$, the distance between two consecutive occurrences of w in $\mathbf{u}$ is bounded by a constant that depends only on w.

For instance, any purely periodic infinite word is uniformly recurrent. Many classical infinite words like the Thue-Morse and Fibonacci words are uniformly recurrent. An eventually periodic word which is not purely periodic is not recurrent. The word

$$010110111011110111110111111011111110111111110111111111\ldots$$

(where the number of ones between consecutive zeros increases each time by one) is not recurrent and cannot be made recurrent by removing a prefix.

There are infinite words which are recurrent but not uniformly recurrent. Examples are easily constructed as fixed points of substitutions on A^*. For instance, the word

$$0101110101111111110101110101111111111111111111111111111101011101 01\ldots$$

is a fixed point of the substitution $0 \mapsto 010$, $1 \mapsto 111$. It is recurrent but not uniformly recurrent (since it is a fixed point of a substitution, it is sufficient for this to show that both symbols 0 and 1 occur infinitely often, but with unbounded intervals in the case of 0). Note that this word can also be defined as the characteristic word of the set of nonnegative integers that have at least one 1 in their ternary expansion.

3.2 The recurrence function

The *recurrence function* of an infinite word $\mathbf{u}$ is the function $R_{\mathbf{u}}\colon \mathbb{N} \to \mathbb{N} \cup \{+\infty\}$ defined by

$$R_{\mathbf{u}}(n) = \inf \big(\{N \in \mathbb{N}\colon \forall v \in F_N(\mathbf{u}), F_n(v) = F_n(\mathbf{u})\} \cup \{+\infty\}\big) \ .$$

In other words, $R_{\mathbf{u}}(n)$ is the size of the smallest window such that, whatever the position of the window on $\mathbf{u}$, all factors of length n that occur in $\mathbf{u}$ occur at least once inside the window. We shall write $R(n) = R_{\mathbf{u}}(n)$ when there is no ambiguity on the relevant infinite word (this convention also applies to other notations with an infinite word as an index).

For instance, in the Thue-Morse word

$$0110100110010110100101100110100110010110011010010110100110010110\ldots$$

(the fixed point of the substitution θ with $\theta(0) = 01$ and $\theta(1) = 10$), we have $R(0) = 0$, $R(1) = 3$ (every factor of length 3 contains at least one 0 and one 1), $R(2) = 9$ (the factor 01011010 of length 8 does not contain 00), $R(3) = 11$, etc. The recurrence function of this word is studied in detail in [13]. We shall present in Sect. 5 a method to compute $R(n)$ in general for words similar to this one.

An infinite word $\mathbf{u}$ is uniformly recurrent if and only if $R_{\mathbf{u}}$ takes only finite values.

3.3 Return times and return words

A closely related notion is that of *return time*. Given a factor w of a recurrent infinite word $\mathbf{u} = u_0u_1u_2\ldots$, an integer i is the position of an occurrence of w in $\mathbf{u}$ if $u_iu_{i+1}\ldots u_{i+|w|-1} = w$. Let us denote by i_0 the smallest such position, by i_1 the next one, etc., so that $(i_0, i_1, i_2, i_3\ldots)$ is the increasing sequence of all positions of occurrences of w in $\mathbf{u}$. Then define the set of words

$$r_{\mathbf{u}}(w) = \{u_{i_j}u_{i_j+1}\ldots u_{i_{j+1}-1} \colon j \in \mathbb{N}\} \ .$$

Elements of $r_{\mathbf{u}}(w)$ are called *return words* of w in $\mathbf{u}$, and the (possibly infinite) number $\ell_{\mathbf{u}}(w) = \sup\{|v| \colon v \in r_{\mathbf{u}}(w)\}$ is called the *(maximal) return time* of w in $\mathbf{u}$. Note that return words of w either have w as a prefix, or are prefixes of w. The latter case happens when two occurrences of w in $\mathbf{u}$ overlap.

Finally, for all $n \in \mathbb{N}$ define $\ell_{\mathbf{u}}(n) = \max\{\ell_{\mathbf{u}}(w) \colon w \in F_n(\mathbf{u})\}$. Then

Proposition 2. *For any recurrent infinite word $\mathbf{u} \in A^{\mathbb{N}}$ and for any $n \in \mathbb{N}$, one has $R_{\mathbf{u}}(n) = \ell_{\mathbf{u}}(n) + n - 1$ (with the convention that $+\infty + n - 1 = +\infty$).*

For instance, consider the Fibonacci word

$$0100101001001010010100100101001001010010100100101001010010010100\ldots$$

(the fixed point of the substitution $0 \mapsto 01$, $1 \mapsto 0$). The factor 0010 occurs at positions 2, 7, 10, 15, 20, 23, etc. and two return words can be observed in the prefix of $\mathbf{u}$ shown here, 001 and 00101. In fact, $r(0010) = \{001, 00101\}$ and $\ell(0010) = 5$, but other factors of the same length have longer return times and $\ell(4) = \ell(0101) = 8$, hence $R(4) = 11$

Return times have a direct dynamic interpretation. In the subshift $X = \overline{O(\mathbf{u})}$ generated by $\mathbf{u}$, given a finite word $w \in F(\mathbf{u})$, the set $[w] = \{\mathbf{v} \in X \colon w \text{ is a prefix of } \mathbf{v}\}$ is both open and closed and is called a *cylinder*. Then the definition of $\ell_{\mathbf{u}}(w)$ can be rephrased as

$$\ell_{\mathbf{u}}(w) = \inf\{N \in \mathbb{N} \colon \forall \mathbf{v} \in [w], \exists n \in \mathbb{N}, 1 \leq n \leq N \text{ and } T^n(\mathbf{v}) \in [w]\} \ ,$$

i.e., $\ell_{\mathbf{u}}(w)$ is the maximum time before which the system returns to the cylinder $[w]$.

The set $r_{\mathbf{u}}(w)$ of return words of a factor w is always a *circular code*, and if i_0 is the position of the first occurrence of w in $\mathbf{u}$, then $T^{i_0}(\mathbf{u})$ can be factored over this code. In particular, if w is a prefix of $\mathbf{u}$ and $r_{\mathbf{u}}(w)$ is finite, then $\mathbf{u}$ can be *recoded* as $\mathbf{u} = f(\mathbf{v})$, where $\mathbf{v} \in B^{\mathbb{N}}$ is an infinite word on a new alphabet B, and f is a one-to-one map from B to $r_{\mathbf{u}}(w)$, extended as a substitution. Such a word $\mathbf{v} = \Delta_w(\mathbf{u})$ is said to be *derivated* from $\mathbf{u}$. The following characterization is due to F. Durand (a substitution $f\colon A^* \to A^*$ is *primitive* if there exists an integer $n \geq 1$ such that, for all $a \in A$, $f^n(a)$ contains every letter of A at least once):

Theorem 1 (Durand [9]). *An infinite word* $\mathbf{u} \in A^{\mathbb{N}}$ *is a fixed point of a primitive substitution on a subset of* A *if and only if* $\mathbf{u}$ *is uniformly recurrent and the number of distinct (up to letter renaming) infinite words derivated from* $\mathbf{u}$ *is finite.*

For instance, the Thue-Morse word, which is a fixed point of a primitive substitution, has three distinct derivated words, the Thue-Morse word $\mathbf{t}$ itself:

$$0110100110010110100101100110100110010110011010010110100110010110\ldots$$

the derivated word associated with the prefix 0, $\Delta_0(\mathbf{t})$, with $\mathbf{t} = f_1(\Delta_0(\mathbf{t}))$ where $f_1(0) = 011,\ f_1(1) = 01,\ f_1(2) = 0$:

$$0120210121020120210201210120210121020121012021020120210121020120\ldots$$

and the derivated word associated with all prefixes of length 2 or more, $\mathbf{v} = \Delta_{01}(\mathbf{t}) = \Delta_{011}(\mathbf{t}) = \cdots$, with $\mathbf{t} = f_2(\mathbf{v}) = f_3(\mathbf{v}) = \cdots$, where $f_2(0) = 011$, $f_2(1) = 010$, $f_2(2) = 0110$, $f_2(3) = 01$, and $f_{2+k} = \theta^k \circ f_2$ for all $k \in \mathbb{N}$:

$$0123013201232013012301320130123201230132012320130123201230132013\ldots$$

3.4 Recurrence and subword complexity

Another numerical function associated with an infinite word $\mathbf{u}$ is the (subword) complexity function $p_{\mathbf{u}}\colon \mathbb{N} \to \mathbb{N}$ defined by $p_{\mathbf{u}}(n) = \#F_n(\mathbf{u})$, the number of factors of length n in $\mathbf{u}$. There is no direct relation between the functions $p_{\mathbf{u}}$ and $R_{\mathbf{u}}$, but only an inequality.

Proposition 3 (Morse and Hedlund [13]). *For any infinite word* $\mathbf{u} \in A^{\mathbb{N}}$ *and for any* $n \in \mathbb{N}$, *one has* $\ell_{\mathbf{u}}(n) \geq p_{\mathbf{u}}(n)$ *and* $R_{\mathbf{u}}(n) \geq p_{\mathbf{u}}(n) + n - 1$.

For non-periodic words, this inequality is not optimal and Morse and Hedlund show that it can be improved to $R_{\mathbf{u}}(n) \geq p_{\mathbf{u}}(n) + n$.

In the other direction, no such inequality holds since it is possible to construct infinite words with $p(n) = n + 1$ (Sturmian words) for which $R(n)$ grows as fast as desired, while remaining finite.

4 Linear recurrence

4.1 Linearly recurrent words

When the recurrence function grows slowly, it means that all factors have to occur rather often and this gives much structure to the infinite word. Of particular interest are words for which $R(n)/n$ is bounded. An infinite word is said to be *linearly recurrent* with constant K if $\ell(n) \leq Kn$ for all $n \geq 1$ [11].

Proposition 4 (Durand, Host, and Skau [11]). *Let* $\mathbf{u} \in A^{\mathbb{N}}$ *be a linearly recurrent infinite word with constant* K*. Then*

(i) For all $n \geq 1$*,* $R(n) \leq (K+1)n - 1$*.*
(ii) For all $n \geq 1$*,* $p(n) \leq Kn$*.*
(iii) $\mathbf{u}$ *is* $(K+1)$*-power free (i.e, it does not contain any factor of the form* w^{K+1} *with* $w \in A^* \setminus \{\varepsilon\}$*).*
(iv) For all $w \in F(\mathbf{u})$ *and* $v \in r_{\mathbf{u}}(w)$*,* $|w|/K < |v| \leq K|w|$*.*
(v) For all $w \in F(\mathbf{u})$*,* $\#r_{\mathbf{u}}(w) \leq K(K+1)^2$*.*

Property (iv) shows that in a linearly recurrent word, return times can be neither too long nor too short. By property (ii), linearly recurrent words are a special case of words with linear subword complexity. In particular, this implies that $p_{\mathbf{u}}(n+1) - p_{\mathbf{u}}(n)$ is bounded by a constant that depends only on K and $\#A$ [3].

The structure of linearly recurrent words can be characterized using *primitive S-adic infinite words*, words obtained by applying in an appropriate order substitutions taken from a finite set (see [10] for a precise definition):

Theorem 2 (Durand [10]). *An infinite word* $\mathbf{u}$ *is linearly recurrent if and only if it is an element of the subshift generated by some primitive S-adic infinite word.*

4.2 Recurrence quotient

Another way to define linearly recurrent words is to use the *recurrence quotient* $\rho_{\mathbf{u}}$. For any infinite word $\mathbf{u}$, let

$$\rho_{\mathbf{u}} = \limsup_{n \to +\infty} \frac{R_{\mathbf{u}}(n)}{n} \in \mathbb{R} \cup \{+\infty\} \ .$$

Then $\rho_{\mathbf{u}}$ is a finite real number if and only if $\mathbf{u}$ is linearly recurrent. Moreover, if $\mathbf{u}$ is linearly recurrent with constant K, $\rho_{\mathbf{u}} \leq K+1$.

If $\mathbf{u}$ is a purely periodic word, it is clear that $\rho_{\mathbf{u}} = 1$. For non-periodic words, Hedlund and Morse [13] asked as an open problem to find the best lower bound for $\rho_{\mathbf{u}}$. Proposition 3 together with the fact that $p_{\mathbf{u}}(n) \geq n+1$ ([13]) implies that $\rho_{\mathbf{u}} \geq 2$. Using graph representations, we improve this result to

Theorem 3. *Let* $\mathbf{u} \in A^{\mathbb{N}}$ *be an infinite word which is not purely periodic. Then* $\rho_{\mathbf{u}} \geq 3$*.*

Rauzy [16] conjectured that the minimal value of $\rho_{\mathbf{u}}$ for non-periodic word is still larger:

Conjecture 1 (Rauzy [16]). Let $\mathbf{u} \in A^{\mathbb{N}}$ be an infinite word which is not purely periodic. Then $\rho_{\mathbf{u}} \geq \frac{5+\sqrt{5}}{2} \simeq 3.618$.

This value $(5+\sqrt{5})/2$ is exactly the recurrence quotient of the Fibonacci word (see below), so if the conjecture holds then it is optimal. We believe that the techniques used to prove Theorem 3 (see Sect. 7), and in particular the extensive study of possible Rauzy graphs, will lead to a proof of this conjecture.

5 Computing the recurrence function

5.1 Singular factors

Wen and Wen [18] defined *singular words* as particular factors of the Fibonacci word (the factors 0, 1, 00, 101, 00100, 10100101, etc., of length the successive Fibonacci numbers, that when concatenated in this order yield the infinite word itself). Here we define singular factors for any infinite word, generalizing one of the properties of Wen and Wen's singular words.

Let $\mathbf{u}$ be an infinite word. A factor w of $\mathbf{u}$ is said to be *singular* for $\mathbf{u}$ if $|w| = 1$ or if there exist a word $v \in A^*$ and letters $x, x', y, y' \in A$ such that $w = xvy$, $x \neq x'$, $y \neq y'$ and $\{xvy, x'vy, xvy'\} \subset F(\mathbf{u})$. In other words, a factor w is singular if there is a way to alter its first letter and still have a factor of $\mathbf{u}$, and symmetrically with the last letter.

When $w = xvy$ is singular, then v is *bispecial*, i.e., v can be extended in at least two different ways both to the right and to the left (see [4]).

Proposition 5 ([7]). *Let $\mathbf{u}$ be an infinite word and $n \geq 1$. If $\ell(n-1) < \ell(n)$, then there exists a singular factor w of $\mathbf{u}$ such that $\ell(n) = \ell(w)$.*

A singular factor w is said to be an *essential singular factor* if $\ell(w) = \ell(|w|) > \ell(|w|-1)$. We denote by $S(\mathbf{u})$ the set of singular factors of $\mathbf{u}$, and by $S'(\mathbf{u})$ the set of essential singular factors of $\mathbf{u}$.

Theorem 4 ([7]). *Let $\mathbf{u}$ be an infinite word and $n \geq 1$. Then*

$$\ell(n) = \sup\{\ell(w) \colon w \in S(\mathbf{u}) \text{ and } |w| \leq n\} = \sup\{\ell(w) \colon w \in S'(\mathbf{u}) \text{ and } |w| \leq n\} .$$

5.2 Computation method

Theorem 4 allows to explicitely compute the recurrence function $R(n)$ as long as one is able to describe singular factors (or at least essential singular factors) and their return time. Since singular factors are extensions of bispecial factors, techniques presented in [4] can be used to describe them when the infinite word is a fixed point of a substitution, or more generally when it is defined using a finite number of substitutions (S-adic words). This results in the following procedure:

1. Determine bispecial factors. Usually a small number of bispecial factors of small length generate all other bispecial factors through recurrence relations.
2. Deduce the form of singular words, and compute their length.
3. For a given singular words, determine the associated return words and compute their length. Singular words with shorter return time can be left out since they are not essential.
4. Deduce the function $\ell(n)$, which will be typically staircase-like, the position and height of each step being expressed with a (usually linear) recurrence relation.

As an example, let us apply this procedure to the Thue-Morse word.

1. Apart from the empty word and letters, there are four families of bispecial factors: 01, 10, 010 and 101 are bispecial, and if w is bispecial then $\theta(w)$ is also bispecial.
2. Bispecial factors in the families generated by 01 and 10 each give rise to four singular factors, which can be summarized as $x\theta^k(y)z$ with $x, y, z \in \{0, 1\}$ and $k \geq 1$, of length $2^k + 2$. The two other families do not produce singular factors (because they are *weak bispecial factors*), and the remaining singular factors are all words of length 1 and 2, as well as 010 and 101.
3. Observation yields

$$\begin{aligned} r(0) &= \{0, 01, 001\} \ , \\ r(00) &= \{0011, 001101, 001011, 00101101\} \ , \\ r(01) &= \{01, 010, 011, 0110\} \ , \\ r(010) &= \{010, 01011, 0100110, 010110011\} \ , \end{aligned}$$

   the case of 1, 11, 10, and 101 being symmetric. The word 0010 always occurs in the form $0\theta(00)1^{-1}$, hence

$$\begin{aligned} r(0010) &= \{0\theta(v)0^{-1} \colon v \in r(00)\} \\ &= \{00101101, 001011010011, 001011001101, 0010110011010011\} \ . \end{aligned}$$

   Similarly,

$$\begin{aligned} r(0011) &= \{1^{-1}\theta(v)1 \colon v \in r(101)\} \ , \\ r(1010) &= \{\theta(v) \colon v \in r(11)\} \ , \text{ and} \\ r(1011) &= \{0^{-1}\theta(v)0 \colon v \in r(00)\} \ . \end{aligned}$$

   Then the return words of $x\theta^{k+1}(y)z$ can be deduced from those of $\bar{x}\theta^k(y)z$ by applying θ and conjugating by x. One has $\ell(0010) = \ell(1010) = \ell(1011) = 16$ and $\ell(0011) = 18$, so obviously only 0011 and the family it generates are essential. Essential singular factors of length 4 or more therefore have length $2^k + 2$ and return time 9.2^k (actually, this also holds for $k = 0$).
4. The function $\ell(n)$ is defined by $\ell(0) = 1$, $\ell(1) = 3$, $\ell(2) = 8$, and $\ell(n) = 9.2^k$ for $2^k + 2 \leq n \leq 2^{k+1} + 1$. Consequently $\rho = 1 + \limsup \ell(n)/n = 10$.

6 The recurrence quotient of Sturmian words

6.1 Computing ρ using continued fractions

Sturmian words are infinite words for which $p(n) = n+1$; see [12] for equivalent definitions, properties and references. They are all uniformly recurrent.

In this particular case, the method given in the previous section to compute the recurrence function amounts to the method described by Morse and Hedlund [14] using continued fraction expansions. As far as the recurrence function is concerned, it is sufficient to study standard Sturmian words: given an irrational number $\alpha \in [0,1] \setminus \mathbb{Q}$, the standard Sturmian word of density α is the word $\mathbf{u} = u_0 u_1 u_2 \ldots$ where $u_n = 1$ if the fractional part of $(n+2)\alpha$ is less than α, $u_n = 0$ otherwise.

Proposition 6 (Morse and Hedlund [14]). *The essential singular factors of the standard Sturmian word* $\mathbf{u}$ *of density* $\alpha \in [0,1] \setminus \mathbb{Q}$ *constitute a sequence* (w_i) *with* $|w_i| = q_i$ *and* $\ell(w_i) = q_i + q_{i+1}$*, where* p_i/q_i *are the convergents associated with the continued fraction expansion of the density,* $\alpha = [0; a_1, a_2, a_3, \ldots]$*. The recurrence quotient of* $\mathbf{u}$ *is* $\rho = 2 + \limsup\,[a_i; a_{i-1}, \ldots, a_1]$*.*

For instance, the Fibonacci word is the standard Sturmian word of density $\alpha = (3-\sqrt{5})/2 = [0; 2, 1, 1, \ldots]$. Its recurrence quotient is therefore $\rho = 2 + \limsup\,[1; 1, 1, 1, \ldots, 1] = [3; 1, 1, 1, \ldots] = (5+\sqrt{5})/2$. The denominators of the convergents are the classical Fibonacci numbers, $q_0 = F_1 = 1$, $q_1 = F_2 = 2$, $q_2 = F_3 = 3$, $q_3 = F_4 = 5$, $q_4 = F_5 = 8$, etc. and they correspond to the lengths of the essential singular factors, $w_0 = 1$, $w_1 = 00$, $w_2 = 101$, $w_3 = 00100$, $w_4 = 10100101$, etc. The associated return times are $\ell(w_i) = q_i + q_{i+1} = F_{i+3}$. Finally, the recurrence function satisfies $R(n) = F_{i+2} + n - 1$ if $F_i \leq n < F_{i+1}$.

A consequence of this proposition is that a Sturmian word is linearly recurrent if and only if the continued fraction expansion of its density is bounded. Then, if $a = \limsup a_i$, one has $a + 2 < \rho < a + 3$.

6.2 The spectrum of values of ρ

Let $S \subset \mathbb{R} \cup \{+\infty\}$ be the set of values taken by ρ for Sturmian words. The set S has an interesting topological structure (we treat sequences of integers $\mathbf{b} = (b_i)_{i \in \mathbb{N}}$ as infinite words on the infinite alphabet $\mathbb{N}^*$, so that the notation $[\mathbf{b}]$ means $[b_0; b_1, b_2, b_3, \ldots]$ and $[T^k(\mathbf{b})] = [b_k; b_{k+1}, b_{k+2}, b_{k+3}, \ldots]$):

Theorem 5 ([7]). *The set* S *is given by*

$$S = \{2 + [\mathbf{b}] \colon \mathbf{b} \in (\mathbb{N}^*)^{\mathbb{N}} \text{ and } \forall k \in \mathbb{N}, [\mathbf{b}] \geq [T^k(\mathbf{b})]\} \cup \{+\infty\} \ .$$

It is a compact subset of $[0, +\infty]$*, with empty interior. It has the power of the continuum. Its smallest accumulation point is the transcendental number* $\rho_0 = 2 + [\mathbf{v}] \simeq 4.58565$*, where* $\mathbf{v} = v_0 v_1 v_2 \ldots \in \{1,2\}^{\mathbb{N}}$ *is the fixed point of the substitution* $1 \mapsto 2$*,* $2 \mapsto 211$*. The intersection of* S *with the set of quadratic numbers is dense in* S*. Every non-countable interval of* S *contains a sub-interval which is isomorphic to* S *as an ordered set.*

The transcendence of ρ_0 was proved by Allouche et al. in [2]. Some questions remain open about the structure of S, for instance its Hausdorff dimension.

7 Main ideas for the proof of Theorem 3

Assume that $\mathbf{u} \in A^{\mathbb{N}}$ is a non-periodic infinite word with $\rho < 3$. Assume also that $\mathbf{u}$ is a binary word (i.e., $\#A = 2$), since the general case can easily be reduced to the binary case by projection.

Let $s(n) = p(n+1) - p(n)$: since $\mathbf{u}$ is not eventually periodic, $s(n) \geq 1$ for all $n \in \mathbb{N}$. By Proposition 3, $\limsup p(n)/n < 2$, which implies that $s(n) = 1$ for infinitely many values of n. There are now two cases: either there is some n_0 such that $s(n) = 1$ for all $n \geq n_0$, or there are infinitely many n such that $s(n) = 1$ and $s(n+1) > 1$.

The first case is essentially the case of Sturmian words, and it is not difficult to adapt the method of [14] to prove that $\rho \geq (5+\sqrt{5})/2 > 3$ in this case.

In the second case, we have infinitely many n for which the Rauzy graph (see [16, 4]) is "eight-shaped" and contains a strong bispecial factor. For subsequent values of n, the Rauzy graphs get more complicated, and it is possible to express return times of certain words as lengths of paths in these graphs, for which lower bounds can be given. Combining these bounds, we get a contradiction with the assumption $\rho < 3$.

To prove a lower bound for ρ larger than 3, one would have to consider also infinite words for which $\liminf s(n) = 2$. Rauzy graphs for these words can have ten different shapes, which where first classified by Rote [17], and the study of their evolutions would involve a large number of subcases.

8 Two other functions

Two functions associated with an infinite word $\mathbf{u}$ and similar to the recurrence function have also been considered. The first one is $R'(n)$, the size of the smallest *prefix* w of $\mathbf{u}$ such that $F_n(w) = F_n(\mathbf{u})$, defined by Allouche and Bousquet-Mélou [1] to study a conjecture of Pomerance, Robson, and Shallit [15]. The second one is $R''(n)$, the size of the smallest *factor* w of $\mathbf{u}$ such that $F_n(w) = F_n(\mathbf{u})$, studied in [6]. These functions compare with each other as follows:

Proposition 7 ([6]). *For any infinite word $\mathbf{u} \in A^{\mathbb{N}}$ and any $n \in \mathbb{N}$, the functions $p_{\mathbf{u}}$, $R_{\mathbf{u}}$, $R'_{\mathbf{u}}$, and $R''_{\mathbf{u}}$ satisfy the inequality $p_{\mathbf{u}}(n)+n-1 \leq R''_{\mathbf{u}}(n) \leq R'_{\mathbf{u}}(n) \leq R_{\mathbf{u}}(n)$.*

It should be noted that, whereas the functions R and R'' depend only on the set of factors of $\mathbf{u}$, or equivalently on the subshift generated by $\mathbf{u}$, the function R' depends on the specific word $\mathbf{u}$. The conjecture by Shallit et al., rephrased using the function R', was very similar to Conjecture 1: if $\mathbf{u}$ is an infinite word which is not eventually periodic, then $\limsup R'(n)/n \geq (3+\sqrt{5})/2$. This is indeed true for standard Sturmian words, but considering non-standard Sturmian words (which have the same factors, but not the same prefixes) we were able to construct a counter-example, and to prove its optimality:

Theorem 6 **([5]).** *Let* $\mathbf{u} \in A^{\mathbb{N}}$ *be an infinite word that is not eventually periodic. Then*

$$\limsup_{n\to+\infty} \frac{R'_{\mathbf{u}}(n)}{n} \geq \frac{29 - 2\sqrt{10}}{9} \simeq 2.51949 \ ,$$

and this value is optimal since it is attained by the Sturmian word

$$\mathbf{z}_3 = 010010100100100101001001010010010010100100100101001001010010\ldots$$

fixed point of the substitution $0 \mapsto 01001010$*,* $1 \mapsto 010$*.*

The function $R''(n)$ seems to have less interesting properties. It is not difficult to see that the minimal value of $\limsup R''(n)/n$ for an infinite word that is not eventually periodic is 2. This minimal value is attained, among others, by all Sturmian words, which can in fact be characterized using R'':

Proposition 8 **([6]).** *An infinite word* $\mathbf{u} \in A^{\mathbb{N}}$ *is Sturmian if and only if* $R''_{\mathbf{u}}(n) = 2n$ *for every* $n \geq 0$*.*

9 Concluding remarks

Some of the properties that we have presented deal with the connections between recurrence and other properties of infinite words: subword complexity, frequencies, repetitions, special factors, etc. These connections have not been completely explored: for instance, the inequality between $p(n)$ and $R(n)$ in Proposition 3 can certainly be improved. We mainly focussed on linearly recurrent words, and did not say much about recurrence of infinite words with very high subword complexity: while *complete* words (i.e., infinite words with maximal complexity $p(n) = \#A^n$) are not uniformly recurrent, how fast can the complexity of a uniformly recurrent word grow?

Not much is known about the spectrum of values taken by ρ for all infinite words, not just Sturmian ones. A proof of Conjecture 1 would provide the minimum of this spectrum, but there are many other questions on its structure. Is it similar to that of S in Theorem 5, or does it contain full intervals of real numbers? It seems that at least the minimum is an isolated point, what is the smallest accumulation point, is it different from ρ_0?

Another problem from [13] is still open: is it true in general that $R(n)/n$ does not converge to a limit? Some progress in this direction has been recently made by N. Chekhova [8].

References

1. J.-P. Allouche and M. Bousquet-Mélou, On the conjectures of Rauzy and Shallit for infinite words, *Comment. Math. Univ. Carolinae* **36** (1995), 705–711.

2. J.-P. Allouche, J. L. Davison, M. Queffélec, and L. Q. Zamboni, Transcendence of Sturmian or morphic continued fractions. Preprint.
3. J. Cassaigne, Special factors of sequences with linear subword complexity, in *Developments in Language Theory II*, pp. 25–34, World Scientific, 1996.
4. J. Cassaigne, Facteurs spéciaux et complexité, *Bull. Belg. Math. Soc.* **4** (1997), 67–88. Special issue: Actes des Journées Montoises d'Informatique Théorique 1994.
5. J. Cassaigne, On a conjecture of J. Shallit, in *ICALP'97*, pp. 693–704, *Lect. Notes Comput. Sci.* **1256**, Springer-Verlag, 1997.
6. J. Cassaigne, Sequences with grouped factors, in *Developments in Language Theory III*, pp. 211–222, Aristotle University of Thessaloniki, 1998.
7. J. Cassaigne, Limit values of the recurrence quotient of Sturmian sequences, *Theoret. Comput. Sci.* **218** (1999), 3–12.
8. N. Chekhova, Fonctions de récurrence des suites d'Arnoux-Rauzy et réponse à une question de Morse et Hedlund. Preprint.
9. F. Durand, A characterization of substitutive sequences using return words, *Discrete Math.* **179** (1998), 89–101.
10. F. Durand, Linearly recurrent subshifts, Research report 98-02, Institut de Mathématiques de Luminy, Marseille, France, 1998.
11. F. Durand, B. Host, and C. Skau, Substitutions, Bratteli diagrams and dimension groups, *Ergod. Th. Dyn. Sys.* **19** (1999), 952–993.
12. M. Lothaire, Algebraic combinatorics on words. To appear. Available online at `http://www-igm.univ-mlv.fr/~berstel/Lothaire/`.
13. M. Morse and G. A. Hedlund, Symbolic dynamics, *Amer. J. Math.* **60** (1938), 815–866.
14. M. Morse and G. A. Hedlund, Symbolic dynamics II: Sturmian trajectories, *Amer. J. Math.* **61** (1940), 1–42.
15. C. Pomerance, J. M. Robson, and J. Shallit, Automaticity II: Descriptional complexity in the unary case, *Theoret. Comput. Sci.* **180** (1997), 181–201.
16. G. Rauzy, Suites à termes dans un alphabet fini, *Sém. Théor. Nombres Bordeaux*, 1982–1983, 25.01–25.16.
17. G. Rote, Sequences with subword complexity $2n$, *J. Number Theory* **46** (1994), 196–213.
18. Z.-X. Wen and Z.-Y. Wen, Some properties of the singular words of the Fibonacci word, *European J. Combin* **15** (1994), 587–598.