Journal of Combinatorial Theory, Series A 208 (2024) 105936

Contents lists available at ScienceDirect

Journal of Combinatorial Theory, Series A

journal homepage: www.elsevier.com/locate/jcta

New string attractor-based complexities for infinite words

Julien Cassaigne [a], France Gheeraert [b], Antonio Restivo [c], Giuseppe Romana [c,*,1], Marinella Sciortino [c,1], Manon Stipulanti [d,2]

[a] *I2M, CNRS, Aix-Marseille Université, France*
[b] *Department of Mathematics, Radboud University, Nijmegen, Netherlands*
[c] *Department of Mathematics and Computer Science, University of Palermo, Italy*
[d] *Department of Mathematics, University of Liège, Belgium*

ARTICLE INFO

Article history:
Received 13 December 2023
Received in revised form 17 April 2024
Accepted 2 July 2024
Available online xxxx

Keywords:
String attractor
Factor complexity
Recurrence function
Repetitiveness measure
Sturmian word
k-bonacci word

ABSTRACT

A *string attractor* is a set of positions in a word such that each distinct factor has an occurrence crossing a position from the set. This definition comes from the data compression field, where the size γ^* of a smallest string attractor represents a lower bound for the output size of a large family of string compressors exploiting repetitions in words, including BWT-based and LZ-based compressors. For finite words, the combinatorial properties of string attractors have been studied in 2021 by Mantaci et al.. Later, Schaeffer and Shallit introduced the *string attractor profile function*, a complexity function that evaluates for each $n > 0$ the size γ^* of the length-n prefix of a one-sided infinite word.

A natural development of the research on the topic is to link string attractors with other classical notions of repetitiveness in combinatorics on words. Our contribution in this sense is threefold. First, we explore the relation between the string attractor profile function and other well-known combinatorial complexity functions in the context of infinite words, such

* Corresponding author.
E-mail addresses: julien.cassaigne@math.cnrs.fr (J. Cassaigne), france.gheeraert@ru.nl (F. Gheeraert), antonio.restivo@unipa.it (A. Restivo), giuseppe.romana01@unipa.it (G. Romana), marinella.sciortino@unipa.it (M. Sciortino), m.stipulanti@uliege.be (M. Stipulanti).

[1] M. Sciortino and G. Romana are partly supported by MUR project PRIN 2022 PINC – 2022YRB97K.
[2] M. Stipulanti is supported by the FNRS Research grant 1.C.104.24F.

https://doi.org/10.1016/j.jcta.2024.105936

as the factor complexity and the property of recurrence. Moreover, we study its asymptotic growth in the case of purely morphic words and obtain a complete description in the binary case. Second, we introduce two new string attractor-based complexity functions, in which the structure and the distribution of positions in a string attractor are taken into account, and we study their combinatorial properties. We also show that these measures provide a finer classification of some infinite families of words, namely the Sturmian and quasi-Sturmian words. Third, we explicitly give the three complexities for some specific morphic words called k-bonacci words.

A preliminary version of some results presented in this paper can be found in [Restivo, Romana, Sciortino, *String Attractors and Infinite Words*, LATIN 2022].

1. Introduction

Repetitiveness is a central notion in the field of Combinatorics on Words, which has been approached from various perspectives. For instance, the *factor complexity function* is probably the most extensively studied repetitiveness measure [9]. For an infinite word $\mathbf{x}$, its factor complexity function $p_{\mathbf{x}}$ counts, for each $n \geq 0$, the number of distinct factors of length n. Intuitively, the lower the factor complexity, the more repetitive the infinite word. Indeed, a famous theorem by Morse and Hedlund characterizes the words with (eventually) constant factor complexity as being *eventually periodic*, i.e. obtained by repeating the same factor, starting after a certain finite prefix. Within the sphere of infinite aperiodic words, some of the most studied words are the *Sturmian words*, which are the infinite aperiodic words with the lowest factor complexity function, i.e. their factor complexity is $n+1$ for every n. *Quasi-Sturmian words* represent the simplest generalization of Sturmian words in terms of factor complexity, they are infinite words having factor complexity $n + d$, with $d \geq 1$, for every large enough n.

The analysis of repetitiveness in words can also be conducted using the recurrence function. It is another powerful measure that, in a complementary way, unveils the repetitive structure of infinite words. This notion was initially defined by Morse and Hedlund [36] but has found widespread recognition in the literature. See [8] for a survey. An infinite word $\mathbf{x}$ is *recurrent* if every factor of $\mathbf{x}$ occurs infinitely often. The recurrence function $R_{\mathbf{x}}$ for an infinite word $\mathbf{x}$ gives, for each $n \geq 0$, if it exists, the size of the smallest window containing all the length-n factors of $\mathbf{x}$, no matter where this window is located in $\mathbf{x}$. Intuitively, it is closely related to the maximum gap between two consecutive occurrences of any length-n factor. Essentially, it provides an idea of how quickly factors repeat within an infinite word and how distributed the repetitive elements are in the word. If $R_{\mathbf{x}}(n)$ is defined for all n, then the word is called *uniformly recurrent*, and if $R_{\mathbf{x}}$ is linear, then $\mathbf{x}$ is called *linearly recurrent*.

In application contexts, repetitiveness has recently become a fundamental concept that is gaining increasing relevance [38]. Due to the abundance of highly repetitive data and the need to manage them efficiently, being able to effectively evaluate and measure the repetitiveness of data is fundamental to optimize processes and resources. For instance, in the realm of indexing massive text collections, defining data structures that enable querying data using space proportional to the size of compressed data becomes crucial [39]. In such a scenario, finding good measures capable of capturing the level of repetitiveness in a text is strongly related to having effective parameters to evaluate the performance of such compressed data structures, both in terms of space and time. For this reason, the most commonly used measures in this field stem from compression schemes, such as the number of phrases in the LZ77 parsing and the number of equal letter runs produced by the Burrows-Wheeler Transform [40].

With the aim of unifying existing compressor-based measures, Kempa and Prezza proposed in [26] a repetitiveness measure related to combinatorial properties of the text instead of being associated with a specific compressor. A *string attractor* Γ for a text w is a set of positions in w such that each factor of w has an occurrence crossing some position in Γ. Intuitively, the more repetitive the text, the lower the number of positions needed in a string attractor. The measure $\gamma^*(w)$ is then the minimal size of a string attractor for w. On the one hand, it has been proven that γ^* is a lower bound for all other usual compressor-based repetitiveness measures. On the other hand, finding the smallest attractor size γ^* for a given text w is an NP-complete problem.

Recently much interest has been aroused by the combinatorial properties of string attractors. Firstly, in [34] the sensitivity of the measure γ^* with respect to combinatorial operations on finite words has been studied. In particular, it has been shown that γ^* is not monotone, in the sense that the measure γ^* of a word can be smaller than that of its prefixes. Also, the measure γ^* has been studied for families of finite prefixes of well-known infinite words such as Thue-Morse word [30,45], Episturmian words [19], k-bonacci-like words [22] and Rote sequences [20], as well as for finite factors of the Thue-Morse word [13]. Moreover, a variation of γ^* in which cyclic factors are considered has been used to characterize the necklaces of standard Sturmian words [34], well-known infinite families of finite words used as bricks to construct particular Sturmian words, called *characteristic Sturmian words*.

A groundbreaking research connecting the notion of string attractors with previously mentioned classical combinatorial notions of repetitiveness for infinite words has been presented in [45]. In particular, the *string attractor profile function* $s_{\mathbf{x}}$ of an infinite word $\mathbf{x}$ is introduced. It measures, for each $n \geq 1$, the smallest size of a string attractor for the length-n prefix of $\mathbf{x}$. The authors study the behavior of $s_{\mathbf{x}}$ when $\mathbf{x}$ is linearly recurrent, and when $\mathbf{x}$ is *automatic*, i.e., $\mathbf{x}$ can be defined through a finite automaton [1].

In this paper, in addition to the size of a string attractor, we also take into account the distribution of the positions within the string attractor. This leads to the definition of two new measures: for a finite word w, the *span* of w is the minimal span (or width) of a string attractor of w and the *leftmost measure* of w is the smallest rightmost position

of a string attractor of w. Starting from these two notions the new complexity measures $\mathsf{lm}_{\mathbf{x}}$ and $\mathsf{span}_{\mathbf{x}}$ can be defined for an infinite word $\mathbf{x}$. In particular, the *span complexity function* $\mathsf{span}_{\mathbf{x}}(n)$ and the *leftmost complexity function* $\mathsf{lm}_{\mathbf{x}}(n)$ give the value of the span and the leftmost measure applied to the length-n prefix of $\mathbf{x}$.

We study the string attractor-based complexities $s_{\mathbf{x}}$, $\mathsf{lm}_{\mathbf{x}}$, and $\mathsf{span}_{\mathbf{x}}$ with three main objectives: understanding their relation with other combinatorial notions of repetitiveness, characterizing some families of words using these complexities, and explicitly computing them for some particular words. We detail below the main contributions to these three topics.

Firstly, when comparing the string attractor-based complexities to repetitiveness properties, the case of bounded $s_{\mathbf{x}}$, $\mathsf{lm}_{\mathbf{x}}$, or $\mathsf{span}_{\mathbf{x}}$ is of particular interest. While we can fully characterize it for the leftmost complexity, we only obtain necessary conditions for the profile function and the span complexity. In particular, we show that aperiodic words with bounded profile function are ω-power-free and have linear factor complexity (Theorem 19). We moreover prove that these conditions are not sufficient, thus answering negatively a question raised in [44].

Secondly, we exhibit three families of infinite words that can be characterized using the two new string attractor-based complexities. Eventually periodic words are the words with bounded leftmost complexity (Proposition 40), Sturmian words those with unbounded leftmost complexity and span complexity equal to 1 infinitely often (Theorem 48), and quasi-Sturmian words those having a suffix with unbounded leftmost complexity and span complexity equal to some constant infinitely often (Theorem 50).

Finally, we compute the three complexities for two families of words: characteristic Sturmian words (Theorem 44) and k-bonacci words (Theorem 58, Proposition 60, and Corollary 63). This is done by explicitly providing string attractors realizing these complexities. In particular, we show that the leftmost complexity uniquely determines the characteristic Sturmian word up to exchanging the two letters (Proposition 46).

This paper is organized as follows.

Section 2 contains all the preliminary definitions. In Section 3, we investigate in depth the connection between the function $s_{\mathbf{x}}$ and some well-known notions of repetitiveness such as factor complexity, uniform recurrence, or ω-power freeness. For example, we show that the values taken by $s_{\mathbf{x}}$ for infinitely many lengths of prefixes give an upper bound on the factor complexity (Proposition 9), as well as study the case of bounded $s_{\mathbf{x}}$. In Section 3.3, we extend a result about the growth of $s_{\mathbf{x}}$ known for automatic words to binary purely morphic words, namely that $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$ and we can decide whether it is the former or the latter (Theorem 23).

Section 4 introduces the span and leftmost measures of a finite word. We give some simple combinatorial observations and study their behaviors when applying a morphism (Proposition 34). Studying these measures for prefixes of infinite words leads to the definition of the span and the leftmost complexities in Section 5. We prove that, analogously to the factor complexity, the leftmost complexity characterizes eventually periodic words.

We also study words of minimal span complexity (Proposition 41) and of maximal span complexity (Proposition 42).

Section 6 is devoted to the study of Sturmian and quasi-Sturmian words. We first focus on characteristic Sturmian words and describe string attractors minimizing all three complexities. We then turn to general Sturmian words and quasi-Sturmian words and characterize them using the string attractor-based complexities.

In Section 7, we move the focus to the k-bonacci words, a generalization of the well-known Fibonacci word to an alphabet of size k. We use a new technique to build string attractors of minimal size. This recursive procedure can be extended to more general families of words obtained by applying morphisms, which represent a classical mechanism to generate repetitive words.

We end the paper with remarks and future works in Section 8.

A preliminary version of some of the results can be found in the conference paper [44].

2. Preliminaries

Combinatorics on words. An *alphabet* is a finite set of *letters* (of cardinality at least 2). A *finite* (resp., *infinite*) *word* on an alphabet Σ is simply a finite (resp., infinite) sequence of letters of Σ. To distinguish them from finite words, infinite words are written in bold, and we start indexing both finite and infinite words at 1, e.g., we will write $\mathbf{x} = x_1x_2\cdots$. For a finite or infinite word x, let $|x|$ denote its *length*, i.e., the number of letters in x, and $\mathsf{alph}(x)$ denote the set of letters appearing in x. The *empty word* ε is the only word that verifies $|\varepsilon| = 0$. Let Σ^* (resp., Σ^+) denote the set of finite (resp., non-empty finite) words over Σ. For all $n \geq 0$, let Σ^n denote the set of length-n words over Σ.

Given a word

$$x = \begin{cases} x_1x_2\cdots x_{|x|}, & \text{if } x \text{ is finite;} \\ x_1x_2\cdots, & \text{if } x \text{ is infinite;} \end{cases}$$

an integer $1 \leq i \leq |x|$ is called a *position* within x. Given two positions $1 \leq i, j \leq |x|$, we use the notation $x[i, j] = x_ix_{i+1}\cdots x_j$; note that $x[i, j] = \varepsilon$ if $j < i$. Such a portion $x[i, j]$ for $i \leq j$ is called a *factor* of x, which *occurs* at position i. Let $F(x)$ denote the set of factors of x. The factor $y \in F(x)$ is *proper* if $y \neq x$. The word u is a *prefix* (resp., *suffix*) of x if $x = uv$ (resp., $x = vu$) for some word v. A factor u of x is *right special* (resp., *left special*) if there exist distinct letters $a, b \in \Sigma$ such that both ua and ub (resp., au and bu) are factors of x. The *reverse* of a finite word $x = x_1x_2\ldots x_{|x|}$ is the word read from right to left, i.e., $x^R = x_{|x|}x_{|x|-1}\cdots x_1$. If $x = x^R$, then x is a *palindrome*.

String attractor of a finite word. Roughly, a string attractor for a finite word is a set of positions within the word such that each of its factors has an occurrence "crossing" at least one element of the set. More formally, a *string attractor* of a finite word x is a set Γ of positions within x such that, for every non-empty factor $w \in F(x)$, there exist

integers i, j such that $w = x[i, j]$ and $[i, j] \cap \Gamma \neq \emptyset$. Let $\gamma^*(x)$ denote the size of a smallest string attractor for x. It is easy to see that $\gamma^*(x) \geq |\mathsf{alph}(x)|$.

Example 1. Let $x = 032\underline{1}00\underline{3}2\underline{1}03\underline{2}$ be a word on $\Sigma = \{0, 1, 2, 3\}$ (the reason why some letters are underlined will become clear later on). The set $\Gamma = \{1, 4, 6, 8, 11\}$ is a string attractor for x. Note that $\Gamma^* = \Gamma \setminus \{1\} = \{4, 6, 8, 11\}$ is still a string attractor for x since each factor that crosses position 1 has another occurrence that crosses a different position in Γ. The positions of Γ^* are underlined above. The set Γ^* is also a smallest string attractor since $|\Gamma^*| = |\Sigma|$, so $\gamma^*(x) = 4$. Note that $\{3, 4, 5, 11\}$ and $\{3, 4, 6, 7, 11\}$ are also string attractors for x. It is easy to verify that the set $\Delta = \{1, 2, 3, 4\}$ is not a string attractor since, for instance, the factor 00 does not intersect any position in Δ.

Factor complexity. For an infinite word $\mathbf{x}$, its *factor complexity function* $p_{\mathbf{x}}$ counts, for any integer $n \geq 0$, the distinct length-n factors of $\mathbf{x}$, i.e., $p_{\mathbf{x}}(n) = |F(\mathbf{x}) \cap \Sigma^n|$ for all $n \geq 0$.

Periodicity. Given a word x, an integer $p \geq 1$ is a *period* of x if $x_i = x_j$ whenever $i \equiv j \bmod p$. An infinite word $\mathbf{x}$ is *eventually periodic* if there exist $u \in \Sigma^*$ and $v \in \Sigma^+$ such that $\mathbf{x} = uv^\omega$, i.e., $\mathbf{x}$ is the concatenation of u followed by infinite copies of a non-empty word v (denoted by v^ω). If $u = \varepsilon$, then $\mathbf{x}$ is said to be *periodic*. An infinite word is *aperiodic* if it is not eventually periodic. We recall the famous Morse-Hedlund theorem (see, for instance, [32, Theorem 1.3.13]).

Theorem 2 *(Morse-Hedlund theorem). Let* $\mathbf{x}$ *be an infinite word. The following are equivalent.*

1. *The word* $\mathbf{x}$ *is eventually periodic.*
2. *We have* $p_{\mathbf{x}}(n+1) = p_{\mathbf{x}}(n)$ *for some integer* $n \geq 0$.
3. *The complexity function* $p_{\mathbf{x}}$ *is bounded.*

Recurrence and appearance functions. An infinite word $\mathbf{x}$ is said to be *recurrent* if every factor of $\mathbf{x}$ occurs infinitely often (in $\mathbf{x}$). The *recurrence function* $R_{\mathbf{x}} \colon n \mapsto R_{\mathbf{x}}(n)$ gives, for each n, the least integer m (or ∞ if no such m exists) such that each length-m factor of $\mathbf{x}$ contains at least an occurrence of each length-n factor of $\mathbf{x}$. An infinite word $\mathbf{x}$ is *uniformly recurrent* if $R_{\mathbf{x}}(n) < \infty$ for each $n \geq 1$. Note that $R_{\mathbf{x}}(n) - n + 1$ is the maximum gap between consecutive occurrences of the same factor when all length-n factors are considered. If $R_{\mathbf{x}}(n)$ is linear, then $\mathbf{x}$ is *linearly recurrent*. It is easy to see that a periodic word $\mathbf{x}$ is linearly recurrent. On the other hand, if $\mathbf{x}$ is eventually periodic but not periodic, then $\mathbf{x}$ is not recurrent. Therefore, a recurrent word is either aperiodic or periodic. For an infinite word $\mathbf{x}$ and an integer n, let $A_{\mathbf{x}}(n)$ denote the length of the shortest prefix containing all length-n factors of $\mathbf{x}$. The function $n \mapsto A_{\mathbf{x}}(n)$ is called the *appearance function* of $\mathbf{x}$.

Example 3. For the binary word $\mathbf{x} = 11011100101110111\cdots$, which is the concatenation of all binary representations of the positive integers, the function $A_{\mathbf{x}}$ is easily seen to be exponential. This also follows from the fact that $p_{\mathbf{x}}$ is exponential too, as explained in the remark below.

Remark 4. For any infinite word $\mathbf{x}$ over Σ, the fact that Σ is finite implies that $A_{\mathbf{x}}(n)$ is defined for each $n \geq 1$. One then easily sees that $p_{\mathbf{x}}(n) + n - 1 \leq A_{\mathbf{x}}(n) \leq R_{\mathbf{x}}(n)$.

Power freeness. An infinite word $\mathbf{x}$ is said to be *k-power free* for some $k > 1$ if, for every factor w of $\mathbf{x}$, w^k is not a factor of $\mathbf{x}$. If for each factor w of $\mathbf{x}$, there exists some integer $k > 1$ such that w^k is not a factor of $\mathbf{x}$, then $\mathbf{x}$ is *ω-power free*.

Morphisms. They represent a mechanism to generate infinite families of repetitive sequences, which have many mathematical properties [1,3,18]. Let Σ_1 and Σ_2 be alphabets. A *morphism* is a map $\varphi\colon \Sigma_1^* \to \Sigma_2^*$ that satisfies the identity $\varphi(uv) = \varphi(u)\varphi(v)$ for all words $u, v \in \Sigma_1^*$. Given an alphabet Σ, a morphism $\varphi\colon \Sigma^* \mapsto \Sigma^*$ is *prolongable* on a letter $a \in \Sigma$ if $\varphi(a) = au$ with $u \in \Sigma^+$. If $\varphi(a) \neq \varepsilon$ for all $a \in \Sigma$, then the morphism φ is said to be *non-erasing*. Given a non-erasing morphism φ prolongable on some $a \in \Sigma$, the sequence $(\varphi^i(a))_{i\geq 0}$ of finite words gives an infinite family of prefixes of a unique infinite word $\varphi^\infty(a) = \lim_{i\to\infty} \varphi^i(a)$, which is called a *purely morphic word* or a *fixed point* of φ. A morphism φ is *primitive* if there exists $t \geq 1$ such that $b \in F(\varphi^t(a))$ for every pair of letters $a, b \in \Sigma$. If there exists k such that $|\varphi(a)| = k$ for every $a \in \Sigma$, then φ is said to be *k-uniform*.

Example 5. Let us consider the *Thue–Morse word* $\mathbf{t} = 0110100110010110\cdots$ which is the fixed point of the 2-uniform morphism $0 \mapsto 01$, $1 \mapsto 10$. It is known that the functions $p_{\mathbf{t}}(n)$, $R_{\mathbf{t}}(n)$ and $A_{\mathbf{t}}(n)$ are $\Theta(n)$. See [1] for details.

Lempel-Ziv factorization. The *Lempel-Ziv factorization* or *parsing* (*LZ77 parsing* in short) of a finite word w is its factorization $LZ(w) = v_1 v_2 \cdots v_z$ built from left to right in a greedy way as follows: if a prefix $w[1, j-1] = v_1 v_2 \cdots v_{i-1}$ is already processed, then the factor v_i (which is also called an *LZ-phrase*) is either the letter w_j if it does not occur in $w[1, j-1]$ or v_i is the longest prefix of $w[j, |w|]$ occurring in w at a position $h < j$. Let $z(w)$ denote the number of LZ-phrases in the LZ77 parsing of w. For example, the LZ77 parsing of the word $w = 0101012$ is $0 \cdot 1 \cdot 0101 \cdot 2$. Consequently, $z(0101012) = 4$. It naturally induces a measure on infinite words as follows: for an infinite word $\mathbf{x}$, the *LZ-complexity function* $z_{\mathbf{x}}$ maps each $n \geq 1$ to the number $z(\mathbf{x}[1, n])$ of LZ-phrases of the length-n prefix of $\mathbf{x}$. An overview of the relationship between z and other repetitiveness measures based on compression schemes can be found in [39].

Remark 6. Note that there are several variants of the Lempel-Ziv factorization; a survey can be found in [29] containing an in-depth study of the relationships between the associated measures. A well-known variant, originally defined in [31], constructs the parsing

of the string w through a similar greedy procedure. However, the phrase v_i is now the longest prefix of $w[j, |w|]$ such that $v_i[1, |v_i| - 1]$ has an occurrence at position $h < j$. Using this technique, the word $w = 0101012$ is factorized as follows: $0 \cdot 1 \cdot 01012$. If z' denotes the number of phrases obtained using such a factorization, then $z'(0101012) = 3$. In [29, Theorem 3], it is shown that $z'(w) \leq z(w) \leq 2z'(w)$, for every word $w \in \Sigma^*$. A complexity measure based on z' is studied for purely morphic words in [12].

The link between string attractors and LZ77 parsings is given in the result below. It follows from the fact that any given finite word has a string attractor of size equal to the number of its LZ-phrases. In fact, it is enough to consider as a string attractor the set of final positions of each LZ-phrase in the Lempel-Ziv factorization. Then, the following proposition holds.

Proposition 7 *([39]). For every word $w \in \Sigma^*$, $\gamma^*(w) \leq z(w)$.*

3. String attractor profile function, factor complexity and recurrence

In this section, we explore the growth of the size of a smallest string attractor when considering increasingly large prefixes of an infinite word. This idea was first considered in [45].

Definition 8. Let $\mathbf{x}$ be an infinite word. The *string attractor profile function* of $\mathbf{x}$ is the map $s_{\mathbf{x}} \colon n \mapsto \gamma^*(\mathbf{x}[1, n])$, i.e. $s_{\mathbf{x}}(n)$ is the size of a smallest string attractor for the length-n prefix of $\mathbf{x}$.

We study the link between the string attractor profile function and different notions measuring the repetitiveness of factors within infinite sequences of symbols. We first establish a bond between the appearance, factor complexity, and string attractor profile functions, and in particular, we show that upper bounds on $s_{\mathbf{x}}$ induce upper bounds on $p_{\mathbf{x}}$.

Proposition 9. *Let $\mathbf{x}$ be an infinite word. For all $n \geq 1$, we have $p_{\mathbf{x}}(n) \leq n \cdot s_{\mathbf{x}}(A_{\mathbf{x}}(n))$.*

Proof. Since alphabets are finite, so is the value $A_{\mathbf{x}}(n)$. By definition, $s_{\mathbf{x}}(A_{\mathbf{x}}(n))$ is the size of a smallest string attractor Γ of the prefix of length $A_{\mathbf{x}}(n)$. Therefore, each length-n factor of $\mathbf{x}$ crosses at least one element of this string attractor. Since each element of Γ is crossed by at most n distinct length-n factors of $\mathbf{x}$, one has $p_{\mathbf{x}}(n) \leq n \cdot s_{\mathbf{x}}(A_{\mathbf{x}}(n))$. □

Using the link between string attractors and LZ77 parsings, we easily obtain an upper bound on $s_{\mathbf{x}}$ as follows.

Proposition 10. *Let $\mathbf{x}$ be an infinite word. Then $s_{\mathbf{x}}(n) = O\left(\frac{n}{\log n}\right)$.*

Proof. Using Proposition 7, we have $s_{\mathbf{x}}(n) \leq z_{\mathbf{x}}(n)$. To conclude, it suffices to use an upper bound on $z_{\mathbf{x}}(n)$ that can be derived from [31, Theorem 2]: for a length-n word on an alphabet Σ, the number z' of phrases obtained using the LZ-factorization introduced in [31] is bounded by $\frac{n}{(1-\epsilon_n)\log_{|\Sigma|}(n)}$, where $\epsilon_n = 2\,\frac{1+\log_{|\Sigma|}(\log_{|\Sigma|}(n|\Sigma|))}{\log_{|\Sigma|}(n)}$. The conclusion follows since $z(x[1,n]) \leq 2z'(x[1,n])$ for every n (see Remark 6). □

It is possible to construct infinite words $\mathbf{x}$ for which there exists an increasing sequence of positive integers n_i, $i \geq 1$, such that $s_{\mathbf{x}}(n_i) = \Theta(\frac{n_i}{\log n_i})$. For instance, one can take the *infinite de Bruijn sequence* from [4], where for a fixed alphabet Σ of size $\sigma \geq 3$ and for all $i \geq 1$, the prefix of length $\sigma^i + i - 1$ contains all possible length-i words over Σ. As there are σ^i such words, by Propositions 9 and 10, we have $s_{\mathbf{x}}(\sigma^i + i - 1) = \Theta(\frac{\sigma^i}{i})$. Thus, by setting $n_i = \sigma^i + i - 1$ for all $i \geq 1$, we obtain $s_{\mathbf{x}}(n_i) = \Theta(\frac{n_i}{\log n_i})$. However, having information on the values of the string attractor profile function over a sequence $(n_i)_{i\geq 1}$ does not allow us to determine its entire behavior, especially since $s_{\mathbf{x}}$ is not monotone (see [34, Proposition 14]). Therefore, the question whether there exist words such that $s_{\mathbf{x}}(n) = \Theta(\frac{n}{\log n})$ for all sufficiently large n is still open.

The following theorem shows that, if we assume that the appearance function is linear, a better bound on the function $s_{\mathbf{x}}$ can be given.

Theorem 11 *([45]). Let* $\mathbf{x}$ *be an infinite word. If* $A_{\mathbf{x}}(n) = \Theta(n)$*, then* $s_{\mathbf{x}}(n) = O(\log n)$.

In the following sections we show several examples in which different repetitiveness aspects are considered (Section 3.1), we analyze which combinatorial notions of repetitiveness are related to the boundedness of the string attractor profile function (Section 3.2) and, finally, we study the behavior of the string attractor profile function in case of infinite words generated by morphisms (Section 3.3).

3.1. Some examples

In this section, we study the behavior of the string attractor profile function for various infinite words, and we focus on the relation with other measures of repetitiveness.

First, let us look at the string attractor profile function of a periodic word, which represents the simplest case of repetitiveness.

Example 12. Let us consider the word $(01)^\omega = 01010101\cdots$. The word is periodic, and therefore $p_{(01)^\omega}(n) = \Theta(1)$ and $A_{(01)^\omega}(n) = n+1$. Since each non-empty factor v of $(01)^\omega$ has an occurrence starting either in the first or in the second position (respectively when v starts with 0 or 1), the set $\{1,2\}$ is a string attractor for each prefix of length $n \geq 2$ of $(01)^\omega$, and therefore $s_{(01)^\omega}(n) = \Theta(1)$.

As shown later in Proposition 20, the previous observation is more general, and every infinite word with factor complexity $\Theta(1)$ has a bounded string attractor profile function.

On the other hand, by Proposition 9, a word with superlinear factor complexity cannot have a bounded string attractor profile function. Therefore, the other words considered in this section have linear factor complexity.

In the following example, we provide a non-recurrent infinite word having linear complexity function and unbounded string attractor profile function.

Example 13. Let us consider the characteristic sequence $\mathbf{c} = 1101000100000001\cdots$ of powers of 2, i.e., $c_i = 1$ if $i = 2^j$ for some $j \geq 0$, $c_i = 0$ otherwise. It is easy to see that $\mathbf{c}$ is aperiodic and not recurrent (e.g., the factor 11 occurs only once). It is known that $p_{\mathbf{c}}(n)$ and $A_{\mathbf{c}}(n)$ are $\Theta(n)$ [1], while one can prove that $s_{\mathbf{c}}(n) = \Theta(\log n)$ [28,34,45].

Example 14 gives a recurrent (not uniformly) infinite word with linear factor complexity and unbounded string attractor profile function.

Example 14. Let $\mu\colon \{0,1\}^* \to \{0,1\}^*$ be the 3-uniform morphism defined by $\mu(0) = 010$ and $\mu(1) = 111$. The infinite word $\mathbf{w} = \mu^\infty(0) = 010111010111111111010\cdots$, known in the literature as the *Sierpiński word* or the *Cantor word* (see, for instance, [5]), has linear factor complexity. Moreover, it is recurrent but not uniformly. Finally, since all factors $01^{3^k}0$, $k \geq 1$, occur in $\mathbf{w}$ and do not overlap with each other, the string attractor profile function $s_{\mathbf{w}}$ is unbounded. In fact, as a consequence of Theorem 23 (proved in Section 3.3), we can conclude that $s_{\mathbf{w}}(n) = \Theta(\log n)$.

In the previous example, the fact that the string attractor profile function is unbounded follows from the existence of arbitrary large powers of 1. The example below uses the Thue-Morse word to give an ω-power-free infinite word with linear factor complexity and unbounded string attractor profile function.

Example 15. Let $\psi\colon \{s, a_0, b_0, a_1, b_1\}^* \to \{s, a_0, b_0, a_1, b_1\}^*$ be the 2-uniform morphism defined by $\psi(s) = sb_0$, $\psi(a_x) = a_{\overline{x}}b_{\overline{x}}$, and $\psi(b_x) = b_{\overline{x}}a_{\overline{x}}$ for all $x \in \{0,1\}$, where $\overline{x} = 1 - x$. Since ψ is 2-uniform, it follows that the infinite word $\mathbf{v} = \psi^\infty(s) = sb_0b_1a_1b_0a_0a_0b_0b_1a_1a_1b_1\cdots$ has linear factor complexity [1]. Moreover, one can observe that if we consider the coding $\lambda\colon \{s, a_0, b_0, a_1, b_1\}^* \mapsto \{0,1\}^*$ defined by $\lambda(s) = \lambda(a_0) = \lambda(a_1) = 0$ and $\lambda(b_0) = \lambda(b_1) = 1$ and apply it on $\mathbf{v}$, we obtain the Thue-Morse word $\mathbf{t} = 0110100110010110\cdots$. Since $\mathbf{t}$ is 3-power free [1], it follows that $\mathbf{v}$ is ω-power free. Finally, since all the factors $b_0\psi^{2k-1}(b_0)b_0$, $k \geq 1$, occur only once in $\mathbf{v}$ and do not overlap with each other, the string profile function $s_{\mathbf{v}}$ is not bounded by a constant.

The previous example is not recurrent, however. To conclude the series of words with linear factor complexity and unbounded string attractor profile function, we give one example of a uniformly recurrent word. Contrary to the last two examples, it is not purely morphic but generated by two morphisms.

Example 16. Let us consider the two 3-uniform morphisms

$$\mu\colon \begin{cases} 0 \mapsto 010 \\ 1 \mapsto 111 \end{cases} \quad \text{and} \quad \bar{\mu}\colon \begin{cases} 0 \mapsto 000 \\ 1 \mapsto 101 \end{cases}$$

and the word $\mathbf{q} = \lim_{n\to\infty} \bar{\mu} \circ \mu \circ \bar{\mu}^2 \circ \mu^2 \circ \cdots \circ \bar{\mu}^n \circ \mu^n(0)$. This word is of linear factor complexity [16, Proposition 2.1] and is uniformly recurrent [15, Lemma 7]. Let us show that, for all $n \geq 1$, the prefix $u_n = \bar{\mu} \circ \mu \circ \bar{\mu}^2 \circ \mu^2 \circ \cdots \circ \bar{\mu}^n \circ \mu^n(0)$ of $\mathbf{q}$ requires at least $n-1$ positions in any of its string attractors. Observe that, in $\mu^n(0)$, we have the factors $01^{3^i}0$ for all $1 \leq i \leq n-1$ which do not overlap one another. Let us show that their images under $\sigma = \bar{\mu} \circ \mu \circ \bar{\mu}^2 \circ \mu^2 \circ \cdots \circ \bar{\mu}^n$ do not overlap either. We first make the following observation. By definition of the morphism μ, for any words u and w, if u contains (at least) a 0, then any occurrence of $\mu(u)$ in $\mu(w)$ corresponds to an occurrence of u in w. In other words, for any u and v containing (at least) a 0 each, $\mu(u)$ and $\mu(v)$ overlap in $\mu(w)$ if and only if u and v overlap in w. Similarly, for any u and v containing (at least) a 1 each, $\bar{\mu}(u)$ and $\bar{\mu}(v)$ overlap in $\bar{\mu}(w)$ if and only if u and v overlap in w. As σ is a composition of μ and $\bar{\mu}$, this shows that the factors $\sigma(01^{3^i}0)$, $1 \leq i \leq n-1$, do not overlap in u_n. We conclude that $s_{\mathbf{q}}$ is not bounded.

However, many classical infinite words in the literature have a known string attractor profile function bounded by a constant. It is the case of the Thue–Morse word (Example 25), the period-doubling word (Example 26), and, as shown in this paper, the characteristic Sturmian words (Theorem 44), the k-bonacci words (Theorem 58) and the family of words defined by Holub in [25] (Example 17).

Example 17. Let us define an infinite word $\mathbf{u}$ introduced by Holub in [25]. For that, let $(n_i)_{i\geq 1}$ be an increasing sequence of positive integers with $n_1 \geq 2$. We recursively define the sequence $(u_i)_{i\geq 0}$ as $u_0 = \varepsilon$ and $u_i = u_{i-1}0(u_{i-1}1)^{n_i}u_{i-1}$. It is proved in [25] that $\mathbf{u} = \lim_{i\to\infty} u_i$ is uniformly recurrent but not linearly recurrent. Moreover, for each $i \geq 1$, $\mathbf{u}$ can be factorized as a product of words u_i0 and u_i1, i.e., $\mathbf{u} = u_ic_1u_ic_2u_ic_3\cdots$, where $c_j \in \{0,1\}$. More precisely, it has been proved in [25] that each occurrence of u_i starts at a position that is a multiple of $|u_i|+1$. Using such a property, the word u has exactly two right special factors of length n, for each $n \geq 1$. They are precisely the length-n suffixes of $u_{i-1}0(u_{i-1}1)^{n_i}u_{i-1}$ and $(u_{i-1}1)^{n_i}u_{i-1}0u_{i-1}$ where $|u_{i-1}|+1 \leq n \leq |u_i|$. By [6], this implies that $p_{\mathbf{u}}(n) = 2n$.

Furthermore, we shall prove that, for $i \geq 1$, the set

$$\Gamma^{(i)} = \left\{ |u_{i-1}|+1, \sum_{k=0}^{i-1}(|u_k|+1), 2|u_{i-1}|+2 \right\}$$

is a string attractor for u_i. Given the recursive construction of $\mathbf{u}$, for each non-empty factor v of u_i, we can find $0 \leq j \leq i-1$ such that $|u_j| < |v| \leq |u_{j+1}|$, and v falls in one of the following mutually exclusive cases:

1. either $v = s_j(1u_j)^{q_1}0(u_j1)^{q_2}p_j$, for some $q_1, q_2 \geq 0$ such that $q_1 + q_2 \leq n_{j+1}$, and for some prefix p_j and suffix s_j of u_j;
2. or $v = s_j(1u_j)^{h_1}0u_j0(u_j1)^{h_2}p_j$, for some $j < i-1$ and $h_1, h_2 \geq 0$ such that $h_1 + h_2 < n_{j+1}$, and for some prefix p_j and suffix s_j of u_j;
3. or $v = s_j(1u_j)^k1p_j$, for some $0 \leq k < n_i$ (resp. $0 \leq k \leq n_j$) if $j = i-1$ (resp. if $j < i-1$), and for some prefix p_j and suffix s_j of u_j.

One can observe that for all $j < i-1$, the factors v from Case 1 have an occurrence crossing position $\sum_{k=0}^{i-1}(|u_k|+1) \in \Gamma^{(i)}$, while if $j = i-1$ the only occurrence of v in u_i crosses the position $|u_{i-1}|+1 \in \Gamma^{(i)}$. Similarly, the factors v that fall in Case 2 have an occurrence in u_i where the 0 at position $|s_j| + h_1(1+|u_j|)+1$ in v is at position $|u_j|+1 \in \Gamma^{(i)}$ in u_i. Finally, one occurrence of each factor falling in Case 3 can be found overlapping the last position in $\Gamma^{(i)}$, where the last 1 before p_j is exactly at position $2|u_i|+2 \in \Gamma^{(i)}$, this ends the proof that $\Gamma^{(i)}$ is a string attractor of u_i.

We deduce a string attractor for the length-n prefix of $\mathbf{u}$ as follows: if i is such that $|u_i| < n < |u_{i+1}|$, we can merge the set $\Gamma^{(i)}$ with the positions $\leq n$ in $\Gamma^{(i+1)}$ to obtain a string attractor for the length-n prefix. Such a string attractor can have up to 6 positions, and it follows that $s_{\mathbf{u}}(n) = \Theta(1)$.

3.2. The bounded case

Supported by the previous section, it is relevant to detect which combinatorial properties of infinite words are related to the boundedness of the string attractor profile function. Observe that we already know the following result.

Theorem 18 *([45]). For any linearly recurrent infinite word* $\mathbf{x}$*, we have* $s_{\mathbf{x}}(n) = \Theta(1)$.

The previous theorem is not a characterization. Indeed, Example 17 exhibits uniformly (and not linearly) recurrent words $\mathbf{x}$ for which $s_{\mathbf{x}}$ is bounded. In this section, we gather results towards a characterization.

First, we analyze how the boundedness of $s_{\mathbf{x}}$ structures the infinite word $\mathbf{x}$, and we show that if an infinite word has its string attractor profile function bounded by some constant value, then it has at most linear factor complexity. More precisely, we have the following result.

Theorem 19. *Let* $\mathbf{x}$ *be an infinite word. If* $s_{\mathbf{x}} = \Theta(1)$*, then either* $\mathbf{x}$ *is eventually periodic, or* $\mathbf{x}$ *is* ω*-power free and* $p_{\mathbf{x}} = \Theta(n)$.

Proof. First, Proposition 9 implies that, if k is such that $s_{\mathbf{x}}(n) < k$ for each $n \geq 1$, then $p_{\mathbf{x}}(n) \leq n \cdot k$ for each $n \geq 1$. Therefore, the factor complexity is (at most) linear. Towards a contradiction, let us assume that $\mathbf{x}$ is aperiodic and not ω-power free. Then there exists a factor w of $\mathbf{x}$ such that, for every $q \geq 1$, w^q is factor of $\mathbf{x}$. Moreover, the assumption

on $\mathbf{x}$ implies that $\mathbf{x} \neq uw^\omega$ for any $u \in \Sigma^*$. It follows that there exists an increasing sequence $(q_j)_{j\geq 1}$ of integers such that, for each j, there exist a proper suffix s_j and a proper prefix p_j of w, and two letters a_j and b_j such that $a_j s_j$ is not a suffix of w, $p_j b_j$ is not a prefix of w, and $a_j s_j w^{q_j} p_j b_j$ is a factor of $\mathbf{x}$. As any position (of a string attractor) can cover at most two such factors, $s_{\mathbf{x}}$ is unbounded. This is a contradiction. □

The following proposition shows that, in the case of eventually periodic words, the string attractor profile function is bounded by a constant.

Proposition 20. *For any eventually periodic infinite word* $\mathbf{x}$*, we have* $s_{\mathbf{x}}(n) = \Theta(1)$.

Proof. Let $u \in \Sigma^*$ and $v \in \Sigma^+$ such that $\mathbf{x} = uv^\omega$. For all $n \geq 1$, $\{1, \ldots, \min\{n, |uv|\}\}$ is a string attractor for the length-n prefix. Therefore, $s_{\mathbf{x}}(n) \leq |uv|$ for all n. □

However, the converse of Theorem 19 does not hold. Indeed, we give in Example 15 an ω-power-free word with linear factor complexity and unbounded string attractor profile function. Even strengthening the hypotheses by requiring uniform recurrence does not guarantee a bounded string profile function, as shown in Example 16. In particular, Examples 15 and 16 negatively answer the questions posed in [44]. Thus, the problem of finding a complete characterization of the infinite words having a bounded string attractor profile function is still open.

We conclude this section with Table 1 showing a synoptic overview of the factor complexity, repetitiveness properties, and string attractor profile function for the infinite words described in this section and those we consider in the rest of the paper. Apart from the periodic word $(01)^\omega$, all words considered in the table have linear factor complexity, which is a necessary (but not sufficient) condition to have a bounded string attractor profile function by Theorem 19. Four of the words have an unbounded string attractor profile function. For the others, the exact value of $s_{\mathbf{x}}(n)$, for n large enough, is reported in the table. This table points out both that different repetitiveness aspects may be hiding behind a constant string attractor profile function, and that infinite words with different combinatorial structures and properties may have point-wise equal profile functions. This observation motivates the use of string attractors to define new complexity measures to capture such combinatorial properties, as done in the following sections.

3.3. The case of purely morphic words

Some data compression measures were explored in the particular setting of fixed points of morphisms, or more precisely, of iterated images of a morphism. It is the case of the number of BWT equal-letter runs [21] and of the LZ-complexity function [12]. Therefore, it is natural to wonder if similar results can be obtained for the string attractor profile function.

First, we present an upper bound on the string attractor profile function of purely morphic words.

Table 1
The table shows the factor complexity $p_{\mathbf{x}}(n)$, recurrence properties, ω-power freeness and the string attractor profile function $s_{\mathbf{x}}(n)$ for large enough n, for all the infinite words $\mathbf{x}$ considered in Sections 3, 6 and 7, namely: the periodic word $(01)^\omega$, the characteristic sequence $\mathbf{c}$ of powers of 2; the purely morphic word $\mathbf{w}$ generated by the morphism μ defined by $\mu(0) = 010$ and $\mu(1) = 111$; a purely morphic word $\mathbf{v}$ generated by a 2-uniform morphism; a uniformly recurrent word $\mathbf{q}$ defined using μ and its counterpart obtained by echanging 0 and 1; an infinite word $\mathbf{u}$ introduced by Holub in [25]; any characteristic Sturmian word $\mathbf{s}$; the Thue-Morse word $\mathbf{t}$; the period doubling word $\mathbf{pd}$; the k-bonacci word $\mathbf{b}^{(k)}$ defined over an alphabet of size k.

Infinite word $\mathbf{x}$	$p_{\mathbf{x}}(n)$	Recurrence	ω-power free	$s_{\mathbf{x}}(n)$
$(01)^\omega$ (Ex. 12)	$\Theta(1)$	linearly recurrent	No	2
$\mathbf{c}$ (Ex. 13)	$\Theta(n)$	not recurrent	No	$\Theta(\log n)$
$\mathbf{w}$ (Ex. 14)	$\Theta(n)$	recurrent	No	$\Theta(\log n)$
$\mathbf{v}$ (Ex. 15)	$\Theta(n)$	not recurrent	Yes	$\Theta(\log n)$
$\mathbf{q}$ (Ex. 16)	$\Theta(n)$	uniformly recurrent	Yes	$\Theta(\log n)$
$\mathbf{u}$ (Ex. 17)	$\Theta(n)$	uniformly recurrent	Yes	3
$\mathbf{s}$ (Sec. 6)	$\Theta(n)$	uniformly recurrent	Yes	2
$\mathbf{t}$ (Ex. 25)	$\Theta(n)$	linearly recurrent	Yes	4
$\mathbf{pd}$ (Ex. 26)	$\Theta(n)$	linearly recurrent	Yes	2
$\mathbf{b}^{(k)}$ (Sec. 7)	$\Theta(n)$	linearly recurrent	Yes	k

Theorem 21. *Let $\mathbf{x} = \varphi^\infty(a)$ be the fixed point of a non-erasing morphism φ prolongable on $a \in \Sigma$. Then $s_{\mathbf{x}}(n) = O(i)$, where i is such that $|\varphi^i(a)| \le n < |\varphi^{i+1}(a)|$. In particular, if there exists $\rho > 1$ such that $|\varphi^i(a)| = \Omega(\rho^i)$, then $s_{\mathbf{x}}(n) = O(\log n)$.*

To prove Theorem 21, we use Proposition 7 and the following result about the number of LZ-phrases in the LZ77 parsing in purely morphic words, which directly follows from [12, Theorem 1] and Remark 6.

Proposition 22. *Let $\mathbf{x} = \varphi^\infty(a)$ be the fixed point of a non-erasing morphism φ prolongable on $a \in \Sigma$. Then*

$$z(\varphi^i(a)) = \begin{cases} \Theta(1), & \text{if } \mathbf{x} \text{ is eventually periodic;} \\ \Theta(i), & \text{otherwise.} \end{cases}$$

Proof of Theorem 21. If $\mathbf{x}$ is eventually periodic, then by Proposition 20, $s_{\mathbf{x}}$ is bounded, so in particular, $s_{\mathbf{x}}(n) = O(i)$. Let us consider the case where $\mathbf{x}$ is not eventually periodic. For all $i \ge 0$, define $n_i = |\varphi^i(a)|$. By Proposition 22, there exist two constants $c_1, c_2 \ge 1$ such that for all $n \in [n_i, n_{i+1})$, we have $c_1 \cdot i \le z_{\mathbf{x}}(n_i) \le z_{\mathbf{x}}(n) \le z_{\mathbf{x}}(n_{i+1}) \le c_2 \cdot i + c_2$. Note that the second and third inequalities follow by the monotonicity of the measure z (i.e., $z(u) \le z(uv)$ for all $u, v \in \Sigma^*$). This implies that $z_{\mathbf{x}}(n) = \Theta(i)$, and by Proposition 7 it follows that $s_{\mathbf{x}}(n) = O(i)$. In particular, if $|\varphi^i(a)| = \Omega(\rho^i)$ for some $\rho > 1$, then one

has $n \in \Omega(\rho^i)$ or, conversely, $i = O(\log n)$ so the conclusion $s_{\mathbf{x}}(n) = O(i) = O(\log n)$ follows. □

In the following theorem, we provide a finer result in the case of binary purely morphic words.

Theorem 23. *Let $\mu\colon \{0,1\}^* \to \{0,1\}^*$ be a morphism prolongable on 0 and $\mathbf{x} = \mu^\infty(0)$. Then either $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$, and it is decidable whether the former or the latter occurs.*

Proof. Based on the morphism μ, we can decide in which of the following (mutually exclusive) cases we are.

1. The word $\mathbf{x}$ is eventually periodic [42, Theorem 4].
2. The word $\mathbf{x}$ is aperiodic and there exist a non-erasing morphism $\tau\colon \Sigma^* \to \{0,1\}^*$ and a primitive morphism $\varphi\colon \Sigma^* \to \Sigma^*$ such that $\mathbf{x} = \mu^\infty(0) = \tau(\varphi^\infty(0))$ (whenever μ is primitive, as well as some decidable cases where $\mu(1) = 1$ by [41, Theorem 4.1] and its proof).
3. The word $\mathbf{x}$ is aperiodic and contains arbitrarily large powers of 1's (whenever $\mu(1) = 1^k$, $k \geq 2$, as well as some decidable cases where $\mu(1) = 1$ by [41, Theorem 4.1]).

Let us now show that, in each case, we have either $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$. For the first case, we have $s_{\mathbf{x}}(n) = \Theta(1)$ as a direct consequence of Proposition 20. In the second case, as φ is primitive, $\varphi^\infty(0)$ is linearly recurrent (see [17, Proposition 25]). This implies that $\mathbf{x}$ is also linearly recurrent and thus that $s_{\mathbf{x}}(n) = \Theta(1)$ by Theorem 18.

We now turn to the third case. Observe that, by Theorem 19, we cannot have $s_{\mathbf{x}}(n) = \Theta(1)$, so we show that $s_{\mathbf{x}}(n) = \Theta(\log n)$. By [21, Proposition 20 and Corollary 27], the number of distinct maximal runs of 1's grows logarithmically with respect to the length of the prefixes of $\mathbf{x}$, where a maximal run of 1's is a factor of the form 01^k0. As a position in a string attractor can cover at most two different runs of 1's, this implies that $s_{\mathbf{x}}(n) = \Omega(\log n)$. On the other hand, observe that by aperiodicity $\mu(0)$ contains at least two occurrences of 0. Therefore, $|\mu^n(0)| = \Omega(2^n)$ and, by Theorem 21, we conclude that $s_{\mathbf{x}}(n) = O(\log n)$ so $s_{\mathbf{x}}(n) = \Theta(\log n)$. □

The same result has been obtained for another class of words, as reported below. In short, an infinite word $\mathbf{x}$ is *k-automatic*, with $k \geq 2$, if and only if there exist a coding $\tau\colon \Sigma \to \Sigma$ and a k-uniform morphism μ_k such that $\mathbf{x} = \tau(\mu_k^\infty(a))$, for some $a \in \Sigma$ [1]. An infinite word is called *automatic* if it is k-automatic for some $k \geq 2$.

Theorem 24 *([45]). Let $\mathbf{x}$ be an automatic infinite word. Then, either $s_{\mathbf{x}}(n) = \Theta(1)$ or $s_{\mathbf{x}}(n) = \Theta(\log n)$, and it is decidable whether the former or the latter occurs.*

Examples 13 and 14 show two automatic sequences for which the string attractor profile function is $\Theta(\log n)$.

For some particular automatic words obtained as fixed points of morphisms, string attractors may be found using their specific combinatorial structure and properties, as recalled in the example below.

Example 25. Let us consider the Thue–Morse word $\mathbf{t} = 0110100110010110\cdots$. It is a purely morphic word, as described in Example 5. It has been proven in [45] (cf. also [30, 13]) that $s_{\mathbf{t}}(n) = 4$ for all $n \geq 25$.

More generally, the authors of [45] show that, in the case of an automatic word with bounded string attractor profile function, it is possible to build an automaton returning the positions of a smallest string attractor for each prefix. However, this automaton is constructed case by case using the theorem-proving software `Walnut` [37]. This technique was used in [45] to find string attractors of minimal size for the prefixes of the automatic word considered in the following example.

Example 26. Consider the *period-doubling word* $\mathbf{pd} = 101110101011\cdots$, which is the fixed point of the morphism $1 \mapsto 10$, $0 \mapsto 11$. It has been proven in [45, Theorem 3] that $s_{\mathbf{pd}}(n) = 2$ for all $n \geq 1$. In particular, it has been shown [45, Theorem 4] that for the prefix of $\mathbf{pd}$ of length $n \geq 6$, a string attractor of smallest size is

$$\Gamma(\mathbf{pd}[1,n]) = \begin{cases} \{3\cdot 2^{i-3}, 3\cdot 2^{i-2}\}, & \text{if } 2^i \leq n < 3\cdot 2^{i-1}; \\ \{2^{i-1}, 2^i\}, & \text{if } 3\cdot 2^{i-1} \leq n < 2^{i+1}. \end{cases}$$

4. Two new string attractor-based measures

In this section, we introduce two new notions related to the string attractors of a word. Indeed, knowing the minimal size of a string attractor is often not sufficient to understand the structure of a word or choose interesting string attractors. Therefore, it can be relevant to consider the distribution of the positions in the string attractors. This is what our new measures capture, and as we will show later, they allow us to distinguish families of words.

The first measure is the span of a word, which gives the minimum distance between the rightmost and the leftmost positions of any string attractor.

Definition 27. Let w be a finite word and $\mathcal{G}$ be the set of all string attractors for w. The *span* of $\Gamma \in \mathcal{G}$ is $\mathsf{span}(\Gamma) = \max\Gamma - \min\Gamma$, and the *(string attractor) span* of w is the value $\mathsf{span}(w) = \min_{\Gamma\in\mathcal{G}} \mathsf{span}(\Gamma)$.

Example 28. Let us consider the word $w = \overline{012}\underline{\overline{2}01}2$ on the alphabet $\Sigma = \{0,1,2\}$. One can see that the sets $\Gamma_1 = \{4,5,6\}$ (underlined positions) and $\Gamma_2 = \{1,2,4\}$ (overlined

positions) are two suitable string attractors for w. Both are of minimal size as $|\Gamma_1| = |\Gamma_2| = |\Sigma|$ but they have different spans. Moreover, since all of the positions of Γ_1 are consecutive, it is of minimal span and therefore $\mathsf{span}(w) = 6 - 4 = 2$.

The span can be used to derive an upper bound on the number of distinct factors, as shown below.

Proposition 29. *For any finite word w over Σ, we have $|F(w) \cap \Sigma^n| \leq n + \mathsf{span}(w)$ for all $1 \leq n \leq |w|$.*

Proof. Let Γ be a string attractor of minimal span and write $\delta = \min \Gamma$ and $\delta' = \max \Gamma$. Then, the interval $\Delta = [\delta, \delta']$ contains Γ and is a string attractor for w. Since every factor has an occurrence crossing a position in Δ, it is possible to find all length-n factors of w by considering a window of length n sliding from position $\max\{\delta - n + 1, 1\}$ to position $\min\{\delta', |w| - n + 1\}$. One can see that this interval is of size at most $\delta' - (\delta - n + 1) + 1 = n + \mathsf{span}(w)$. This ends the proof. □

In addition, we may compare string attractors of a given word according to their rightmost positions. More specifically, we want string attractors having the smallest such position. This gives the notion defined below.

Definition 30. Let w be a finite word and $\mathcal{G}$ be the set of all string attractors for w. The *leftmost string attractor* for w is a string attractor $\Gamma \in \mathcal{G}$ such that, for all $\Delta \in \mathcal{G}$, we have $\max \Gamma \leq \max \Delta$. The *(string attractor) leftmost measure* of w is then $\mathsf{lm}(w) = \max \Gamma$, where Γ is a leftmost string attractor.

Example 31. We resume Example 28. First, we have $4 = \max \Gamma_2 < \max \Gamma_1 = 6$. Second, the set $\Delta = \{1, 2, 3\}$ is not a string attractor for w. Therefore $\mathsf{lm}(w) = 4$.

Examples 28 and 31 show that for the finite word $w = 0122012$, these two measures can be realized by distinct string attractors. In fact, in this case, it is not possible to find a leftmost string attractor having minimal span since $\{2, 3, 4\}$ is not a string attractor.

Similarly to what we did for the span, we can use the leftmost measure to obtain an upper bound on the number of distinct factors.

Proposition 32. *For any finite word w over Σ, we have $|F(w) \cap \Sigma^n| \leq \mathsf{lm}(w)$ for all $1 \leq n \leq |w|$.*

Proof. The proof follows the same lines as that of Proposition 29 by considering a leftmost string attractor Γ, and $\Delta = [1, \max \Gamma]$ instead. □

From Examples 28 and 31, we formulate the following general observation.

Proposition 33. *Let w be a finite word. Then, $\gamma^*(w) - 1 \leq \mathit{span}(w) \leq \mathit{lm}(w) - 1$.*

Proof. Let Γ_1 be a string attractor of w with minimal span. It contains at most $\max \Gamma_1 - \min \Gamma_1 + 1 = \mathsf{span}(w) + 1$ elements, therefore $\gamma^*(w) \leq \mathsf{span}(w) + 1$.

Let Γ_2 be a leftmost string attractor of w. Its span is at most $\max \Gamma_2 - 1 = \mathsf{lm}(w) - 1$, therefore $\mathsf{span}(w) \leq \mathsf{lm}(w) - 1$. □

The following proposition shows how the size of the smallest string attractor, the span, and the leftmost measure of a word yield bounds on the corresponding measures for its image under a morphism.

Proposition 34. *Let $\varphi\colon \Sigma_1^* \to \Sigma_2^*$ be a morphism. There exists a constant $C \geq 1$ which depends only on φ such that, for every $w \in \Sigma_1^+$, the following hold:*

1. $\gamma^*(\varphi(w)) \leq 2\gamma^*(w) + C$;
2. $\mathit{span}(\varphi(w)) \leq C \cdot \mathit{span}(w)$;
3. $\mathit{lm}(\varphi(w)) \leq C \cdot \mathit{lm}(w)$.

Proof. Starting from a given string attractor Γ for w, we show how one can build a valid string attractor for $\varphi(w)$ in two steps.

Step 1. First, we consider the factors of the images of letters, i.e., the elements of $F_\varphi = \bigcup_{a \in \mathsf{alph}(w)} F(\varphi(a))$. By definition, for each symbol $a \in \mathsf{alph}(w)$, there is at least one position $j \in \Gamma$ such that $w_j = a$; let j_a denote such a position. Then, for each $a \in \mathsf{alph}(w)$, we choose a minimum string attractor Γ_a of $\varphi(a)$ and overlay it with the occurrence of $\varphi(w_{j_a})$ to cover the factors of $\varphi(a)$. In other words, every element of F_φ has an occurrence in w crossing at least a position in

$$\mathcal{T}_\varphi = \bigcup_{a \in \mathsf{alph}(w)} \{|\varphi(w[1, j_a - 1])| + \delta : \delta \in \Gamma_a\}.$$

Step 2. Let us now consider the other factors of $\varphi(w)$, i.e., the elements of $F(\varphi(w))$ which are not in F_φ. To cover these factors, we define two sets of positions. Let $\mathcal{T}_f = \{|\varphi(w[1, j-1])| + 1 : j \in \Gamma\}$ be the set of positions corresponding to the first letter of $\varphi(w_j)$, where j is a position in Γ. Analogously, we define the set $\mathcal{T}_\ell = \{|\varphi(w[1, j])| : j \in \Gamma\}$ as the set of positions corresponding to the last letter of each $\varphi(w_j)$ with $j \in \Gamma$.

Let $u \in F(\varphi(w)) \setminus F_\varphi$ and let v be a factor of w of minimal length such that u is a factor of $\varphi(v)$. Observe that, by definition of F_φ, v is of length at least 2. As v is a factor of w, it has an occurrence crossing some position $j \in \Gamma$. By minimality of v, we know that u has an occurrence crossing either the first position of $\varphi(w_j)$ or the last position of $\varphi(w_j)$ (or both). Therefore, u crosses a position in $\mathcal{T}_f$ or $\mathcal{T}_\ell$.

As a consequence of the previous two steps, $\Delta = \mathcal{T}_\varphi \cup \mathcal{T}_f \cup \mathcal{T}_\ell$ is a string attractor for $\varphi(w)$. Recall that this construction can be done starting from any string attractor Γ of w,

giving different corresponding string attractors Δ. To obtain the three claimed inequalities, we will consider different string attractors Γ of w. Now let $M = \max_{a\in\Sigma_1} |\varphi(a)|$, i.e., M is the length of the longest image of a letter.

1. If Γ is such that $|\Gamma| = \gamma^*(w)$, then

$$\gamma^*(\varphi(w)) \le |\Delta| \le |\mathcal{T}_f| + |\mathcal{T}_\ell| + |\mathcal{T}_\varphi| \le 2\gamma^*(w) + \sum_{a\in\mathsf{alph}(w)} \gamma^*(\varphi(a)).$$

2. If Γ is such that $\delta = \min\Gamma$, $\delta' = \max\Gamma$ and $\delta' - \delta = \mathsf{span}(w)$, then by construction we have $\min\Delta = |\varphi(w[1,\delta-1])| + 1 \in \mathcal{T}_f$ and $\max\Delta = |\varphi(w[1,\delta'])| \in \mathcal{T}_\ell$, and therefore

$$\mathsf{span}(\varphi(w)) \le |\varphi(w[1,\delta'])| - (|\varphi(w[1,\delta-1])| + 1) = |\varphi(w[\delta,\delta'])| - 1 \le M\cdot(\mathsf{span}(w)+1).$$

3. If Γ is such that $\max\Gamma = \mathsf{lm}(w)$, then

$$\mathsf{lm}(\varphi(w)) \le \max\Delta = |\varphi(w[1,\max\Gamma])| \le M\cdot\mathsf{lm}(w).$$

To end the proof, we can choose the constant $C = M(|\Sigma_1|+1)$ (which is independent of w), and the conclusion will follow for all three cases. □

5. Span and leftmost complexities

Based on the two new measures introduced in the previous section, we can define related complexity functions for infinite words, respectively called the *span complexity* and the *leftmost complexity*, which allow us to obtain a finer classification of infinite words. Indeed, Examples 36 and 45 highlight two infinite words, the period-doubling word and the Fibonacci word, which are not distinguishable if we consider their respective string attractor profile function as they are eventually equal to 2. However, the situation is very different if we look at how the positions within a string attractor are arranged.

Definition 35. Let $\mathbf{x}$ be an infinite word. The *span* and *leftmost complexities* of $\mathbf{x}$ are respectively defined by $\mathsf{span}_{\mathbf{x}}(n) = \mathsf{span}(\mathbf{x}[1,n])$ and $\mathsf{lm}_{\mathbf{x}}(n) = \mathsf{lm}(\mathbf{x}[1,n])$ for all $n \ge 1$.

The span complexity for the period doubling word is described below.

Example 36. Consider the period-doubling word $\mathbf{pd} = 101110101011\cdots$ described in Example 26 in which we recalled that $s_{\mathbf{pd}}(n) = 2$ for all $n \ge 2$. It has been proven in [45, Theorem 10] that

$$\mathsf{span}_{\mathbf{pd}}(n) = \begin{cases} 1, & \text{if } 2 \le n \le 5; \\ 2^i, & \text{if } 3\cdot 2^i \le n < 3\cdot 2^{i+1} \text{ for some } i \ge 1. \end{cases}$$

For Holub's words, we can use Example 17 to obtain the span and the leftmost complexities for particular prefixes, as shown below.

Example 37. Consider the word $\mathbf{u}$ from Example 17 in which we proved that, for all $i \geq 0$, the set $\Gamma^{(i+1)} = \left\{|u_i|+1, \sum_{k=0}^{i}(|u_k|+1), 2|u_i|+2\right\}$ is a string attractor of the length-$|u_{i+1}|$ prefix of $\mathbf{u}$. This directly implies that $\mathsf{span}_{\mathbf{u}}(|u_{i+1}|) \leq \max\Gamma^{(i+1)} - \min\Gamma^{(i+1)} = |u_i|+1$ and that $\mathsf{lm}_{\mathbf{u}}(|u_{i+1}|) \leq 2|u_i|+2$. Moreover, recall that consecutive occurrences of u_i in $\mathbf{u}$ are separated by at least $|u_i|+1$ letters. In particular, as $u_{i+1} = u_i 0 (u_i 1)^{n_i} u_i$ with $n_i \geq 2$, the factor $u_i 0$ only occurs as a prefix in u_{i+1}, and $1u_i 1$ does not occur before position $2|u_i|+2$. It follows that $\mathsf{span}_{\mathbf{u}}(|u_{i+1}|) = |u_i|+1$ and that $\mathsf{lm}_{\mathbf{u}}(|u_{i+1}|) = 2|u_i|+2$.

The next result directly follows from Proposition 33 and establishes the relationship between the profile function, the span complexity and the leftmost complexity.

Proposition 38. *For any infinite word* $\mathbf{x}$*, we have* $s_{\mathbf{x}}(n) - 1 \leq \mathsf{span}_{\mathbf{x}}(n) \leq \mathsf{lm}_{\mathbf{x}}(n) - 1$ *for all* $n \geq 1$*.*

As we did for the string attractor profile function, we now focus on the case where these new complexities are "bounded". More specifically, we will characterize the infinite words such that these complexities are bounded infinitely many times.

We first look at the leftmost complexity. We will use the following intermediate result, which can be deduced from the proofs of [34, Propositions 12 and 15].

Proposition 39. *Let* w *be a non-empty word and let* $u = w^r$*,* $v = w^s$ *be fractional powers of* w *with* $1 \leq r \leq s$*. If* Γ *is a string attractor of* u*, then* $\Gamma \cup \{|w|\}$ *is a string attractor of* v*.*

Proposition 40. *For any infinite word* $\mathbf{x}$*, the following are equivalent:*

1. *There exists a constant* $C \geq 1$ *such that* $\mathsf{lm}_{\mathbf{x}}(n) \leq C$ *for infinitely many* n*.*
2. *The word* $\mathbf{x}$ *is eventually periodic.*
3. *The leftmost complexity* $\mathsf{lm}_{\mathbf{x}}$ *is bounded.*

Proof. The implication (1) $\Longrightarrow$ (2) follows from Proposition 32. Indeed, for all $m \geq 1$, there exists an integer n such that $\mathsf{lm}_{\mathbf{x}}(n) \leq C$ and $\mathbf{x}[1,n]$ contains all length-m factors. Therefore, $p_{\mathbf{x}}(m) \leq C$. Using Theorem 2, this implies that $\mathbf{x}$ is eventually periodic.

The implication (2) $\Longrightarrow$ (3) follows from Proposition 39. Indeed, if $\mathbf{x} = uv^\omega$, then for all $n \geq 1$, $\{1, 2, \ldots, \min\{n, |uv|\}\}$ is a string attractor for the word $\mathbf{x}[1,n]$. Therefore, $\mathsf{lm}_{\mathbf{x}}(n) \leq |uv|$ for all $n \geq 1$.

The implication (3) $\Longrightarrow$ (1) is direct. □

This result gives a new characterization of eventually periodic words. Observe that the proof uses the well-known characterization by Morse and Hedlund (Theorem 2). Note that, in the following, we will mostly use the contraposition of Proposition 40.

We now look at a similar description for the span.

Proposition 41. *Let* $\mathbf{x}$ *be an infinite word. If there exists a constant* $C \geq 1$ *such that* $\mathsf{span}_{\mathbf{x}}(n) \leq C$ *for infinitely many* n*, then* $\mathbf{x}$ *is eventually periodic, or it is recurrent and there exists* $d \leq C$ *such that* $p_{\mathbf{x}}(n) = n + d$ *for all large enough* n*.*

Proof. Let us suppose that $\mathbf{x}$ is aperiodic. We first show that $\mathbf{x}$ is recurrent. Towards a contradiction, we assume that $\mathbf{x}$ is not recurrent. Therefore, there exists a factor that only occurs once in $\mathbf{x}$. Say that this occurrence ends at position k. This implies that, for all $n \geq k$, any string attractor of $\mathbf{x}[1, n]$ contains a position smaller than or equal to k. As $\mathsf{span}_{\mathbf{x}}(n) \leq C$ for infinitely many n, then $\mathsf{lm}_{\mathbf{x}}(n) \leq k + C$ for infinitely many n, which contradicts Proposition 40.

We now show that $\mathbf{x}$ has the claimed factor complexity. For all $m \geq 1$, there exists an integer n such that $\mathsf{span}_{\mathbf{x}}(n) \leq C$ and $\mathbf{x}[1, n]$ contains all length-m factors. By Proposition 29, we have $p_{\mathbf{x}}(m) \leq m + C$. Using Theorem 2 and as $\mathbf{x}$ is aperiodic, we conclude that $p_{\mathbf{x}}(m) = m + d$ for all large enough m and for some $d \leq C$. □

Note that a converse-like characterization will be given in Theorem 50.

On the other hand, some infinite words have maximal span complexity, as stated in the following result.

Proposition 42. *For any linearly recurrent word* $\mathbf{x}$*, if* $p_{\mathbf{x}}(n) = n + \Omega(n)$*, then* $\mathsf{span}_{\mathbf{x}}(n) = \Theta(n)$*.*

Proof. Since $\mathbf{x}$ is linearly recurrent, by Remark 4, there exists an integer A such that, for all m, the length-(Am) prefix of $\mathbf{x}$ contains all length-m factors of $\mathbf{x}$. For all n, if m is such that $n \in [Am + 1, A(m + 1)]$, Proposition 29 implies that $\mathsf{span}_{\mathbf{x}}(n) \geq p_{\mathbf{x}}(m) - m$. By assumption on the factor complexity function, we have $p_{\mathbf{x}}(m) \geq Cm$ for a constant $C > 1$. Therefore $\mathsf{span}_{\mathbf{x}}(n) \geq (C-1)m \geq (C-1)\,(n/A - 1)$. This shows that $\mathsf{span}_{\mathbf{x}}(n) = \Omega(n)$. But since we trivially have $\mathsf{span}_{\mathbf{x}}(n) = O(n)$, the conclusion follows. □

6. The case of Sturmian words

Sturmian words are famous combinatorial objects having several mathematical properties and characterizations. To name one of them, they approximate straight lines [32, Chapter 2]. Among aperiodic binary infinite words, Sturmian words have minimal factor complexity, i.e., an aperiodic infinite word $\mathbf{x}$ is *Sturmian* if $p_{\mathbf{x}}(n) = n + 1$, for all $n \geq 0$. Moreover, Sturmian words are uniformly recurrent.

In this section, we study the three string attractor-related complexities in the context of Sturmian words and two related families of infinite words. On the one hand, we consider the subfamily of *characteristic Sturmian words*, defined as follows: a Sturmian word $\mathbf{s}$ is *characteristic* if both $0\mathbf{s}$ and $1\mathbf{s}$ are Sturmian words. On the other hand, we investigate the superfamily of *quasi-Sturmian* words, which can be considered the simplest generalization of Sturmian words in terms of factor complexity. Indeed, they

are defined as follows [7]: a word $\mathbf{x}$ is *quasi-Sturmian* if there exist integers d and n_0 such that $p_{\mathbf{x}}(n) = n + d$, for each $n \geq n_0$. The infinite words with factor complexity $n + d$ were also studied in [24] under the name of "words with minimal block growth".

6.1. On the string attractor-based complexities for characteristic Sturmian words

We first focus on the family of characteristic Sturmian words, for which we can explicitly give the string attractor profile function, the span complexity, and the leftmost complexity. To do so, we provide string attractors realizing them and based on the construction of characteristic Sturmian words via particular finite words called *standard Sturmian words*. These words have many interesting combinatorial properties and appear as extreme cases for several algorithms and data structures [11,10,27,35,46]. The standard Sturmian words are defined recursively as follows [43].

Definition 43. A *directive sequence* is an infinite sequence of integers $(q_i)_{i\geq 0}$ such that $q_0 \geq 0$ and $q_i \geq 1$ for all $i \geq 1$. The corresponding sequence of *standard Sturmian words* $(x_i)_{i\geq 0}$ is defined by $x_0 = 1$, $x_1 = 0$, and $x_{i+1} = x_i^{q_{i-1}} x_{i-1}$ for all $i \geq 1$.

The limits $\mathbf{s} = \lim_{i\to\infty} x_i$ of such sequences of standard Sturmian words are precisely the characteristic Sturmian words [32, Proposition 2.2.24]. Note that $\mathbf{s}$ starts with the letter 0 if and only if $q_0 \geq 1$. We let $E\colon \{0,1\}^* \to \{0,1\}^*$ be the exchange morphism, i.e., $E(0) = 1$ and $E(1) = 0$. A well-known property of characteristic Sturmian words is the following: $\mathbf{s}$ starts with a letter 0 and has $(q_i)_{i\geq 0}$ as directive sequence if and only if $E(\mathbf{s})$ starts with a letter 1 and has $(q'_i)_{i\geq 0}$ as directive sequence with $q'_0 = 0$ and $q'_{i+1} = q_i$ for all $i \geq 0$ [33, Section 2]. Therefore, in what follows, we only consider the case where $q_0 \geq 1$.

The following result shows that each prefix of a characteristic Sturmian word has a smallest string attractor of span 1, i.e., consisting of two consecutive positions.

Theorem 44. *Consider a directive sequence $(q_i)_{i\geq 0}$ with $q_0 \geq 1$, the corresponding sequence $(x_i)_{i\geq 0}$ of standard Sturmian words and the associated characteristic Sturmian word $\mathbf{s} = \lim_{i\to\infty} x_i$ as in Definition 43. Then we have*

$$s_{\mathbf{s}}(n) = \begin{cases} 1, & \text{if } n < |x_2|; \\ 2, & \text{if } n \geq |x_2|; \end{cases} \qquad \mathit{span}_{\mathbf{s}}(n) = \begin{cases} 0, & \text{if } n < |x_2|; \\ 1, & \text{if } n \geq |x_2|; \end{cases}$$

and

$$\mathit{lm}_{\mathbf{s}}(n) = \begin{cases} 1, & \text{if } n < |x_2|; \\ |x_k|, & \text{if } |x_k| + |x_{k-1}| - 1 \leq n \leq |x_{k+1}| + |x_k| - 2 \text{ for some } k \geq 2. \end{cases}$$

More precisely, for all $n \geq 1$, the following string attractor for $\mathbf{s}[1,n]$ witnesses the above equalities:

$$\Gamma_n = \begin{cases} \{1\}, & \text{if } n < |x_2|; \\ \{|x_k|-1, |x_k|\}, & \text{if } |x_k|+|x_{k-1}|-1 \le n \le |x_{k+1}|+|x_k|-2 \text{ for some } k \ge 2. \end{cases}$$

Proof. We start the proof by showing the last part of the statement, i.e., we show that, for all $n \ge 1$, the given Γ_n is a string attractor for $\mathbf{s}[1,n]$. Observe first that, if $n < |x_2|$, then $\mathbf{s}[1,n] = 0^n$, so $\{1\}$ is directly a string attractor. For the case $n \ge |x_2|$, we need the following notation. For all $k \ge 2$, using [33, Theorem 3], we factorize the standard Sturmian word x_k into $x_k = y_k u_k$ where y_k is a palindrome and $u_k = 01$ if k is even and $u_k = 10$ if k is odd. We also recall the following observation: for all $k \ge 2$, since we have

$$\mathbf{s}[1, |x_{k+1}| + |x_k| - 2] = x_{k+1}y_k = y_{k+1}u_{k+1}y_k,$$

then [32, Theorem 2.2.11] implies that $\mathbf{s}[1, |x_{k+1}|+|x_k|-2]$ is periodic of period $|y_k|+2 = |x_k|$.

Assume now that $n \ge |x_2|$, and let $k \ge 2$ be such that $|x_k| + |x_{k-1}| - 1 \le n \le |x_{k+1}|+|x_k|-2$ (such a k exists since $|x_2|+|x_1|-1 = |x_2|$). Since $\mathbf{s}[1, |x_{k+1}|+|x_k|-2]$ is periodic of period $|x_k|$, then it is a fractional power of x_k. Therefore, using Proposition 39, it is enough to show that $\Gamma_n = \{|x_k|-1, |x_k|\}$ is a string attractor of the length-$(|x_k| + |x_{k-1}| - 1)$ prefix of $\mathbf{s}$, denoted by p_k.

If $k = 2$ or $k = 3$, the conclusion is direct as $p_2 = x_2 = 0^{q_0}1$ and $p_3 = (0^{q_0}1)^{q_1}00^{q_0}$. If $k \ge 4$, we use the fact that a similar result was proved for the standard Sturmian words in [34, Theorem 22]. Namely, Γ_n is a string attractor for x_{k+1}. To show that Γ_n is also a string attractor for p_k, we will show that $w := \mathbf{s}[|x_k|, |p_k|]$ does not occur elsewhere in p_k. Indeed, this will imply that for each factor of p_k its occurrence that was covered by Γ_n in x_{k+1} is an occurrence in p_k (also covered by Γ_n).

Observe that, as $k \ge 4$, $w = ay_{k-1}a$ where a is the last letter of u_k and the first letter of u_{k-1}. Note that w is not a suffix of $\mathbf{s}[1, |p_k| - 1] = x_k y_{k-1}$ as x_k ends with $u_k = ba$, $b \ne a$. Therefore, if w is a factor of $x_k y_{k-1}$, it is followed by a since $x_k y_{k-1}$ is periodic of period $|x_{k-1}| = |w|$. In particular, $y_{k-1}aa$ and $y_{k-1}ab = x_{k-1}$ are factors of $x_k y_{k-1}$. This implies that $y_{k-1}a$ is right special and, by [32, Proposition 2.1.23], ay_{k-1} is a prefix of x_k. As $y_{k-1}a$ is also a prefix of x_k, this implies that y_{k-1} is periodic of period 1, a contradiction as $k \ge 4$. This ends the proof that w is not a factor of $x_k y_{k-1}$ and, with it, the proof that Γ_n is a string attractor of $\mathbf{s}[1,n]$.

Moreover, we directly have that Γ_n is of minimal size and of minimal span among the string attractors of $\mathbf{s}[1,n]$. It is also a leftmost string attractor as each string attractor of $\mathbf{s}[1,n]$ will contain a position greater than or equal to $|x_k|$ to cover w. This proves the three claimed complexities. □

Example 45. Consider the infinite Fibonacci word $\mathbf{f} = 01001010010010100\cdots$, which is a characteristic Sturmian word with directive sequence $(1)_{i\ge 0}$. In Table 2, for $1 \le n \le 8$, we exhibit the length-n prefixes of $\mathbf{f}$ and their respective leftmost string attractor Γ_n. The underlined positions in $\mathbf{f}[1,n]$ correspond to those in Γ_n, while the first few lengths $|x_k|$, $k \in [1,5]$ are given by $\{1,2,3,5,8\}$.

Table 2
For $n \in [1, 8]$, the length-n prefix of the Fibonacci word $\mathbf{f} = 010010100100\cdots$ and its leftmost string attractor Γ_n.

n	1	2	3	4	5	6	7	8
$\mathbf{f}[1,n]$	$\underline{0}$	$\underline{01}$	$\underline{01}0$	$0\underline{10}0$	$0\underline{10}01$	$0\underline{10}010$	$0100\underline{10}1$	$0100\underline{10}10$
Γ_n	$\{1\}$	$\{1,2\}$	$\{1,2\}$	$\{2,3\}$	$\{2,3\}$	$\{2,3\}$	$\{4,5\}$	$\{4,5\}$

Note that different size-2 string attractors are obtained in Section 7.2.

While infinitely many characteristic Sturmian words have the same string attractor profile function (resp., the same span complexity), the leftmost complexity uniquely determines the characteristic Sturmian word (up to exchanging the letters 0 and 1, captured by the exchange morphism E). This is the object of the result below.

Proposition 46. *Let* $\mathbf{s}$ *and* $\mathbf{s}'$ *be two characteristic Sturmian words such that* $\mathsf{lm}_{\mathbf{s}} = \mathsf{lm}_{\mathbf{s}'}$. *Then, either* $\mathbf{s} = \mathbf{s}'$ *or* $\mathbf{s} = E(\mathbf{s}')$.

Proof. As in Definition 43, let $(q_i)_{i\geq 0}$ and $(p_i)_{i\geq 0}$ be two directive sequences and let $(x_i)_{i\geq 0}$ and $(y_i)_{i\geq 0}$ be the corresponding sequences of standard Sturmian words that are prefixes of $\mathbf{s}$ and $\mathbf{s}'$ respectively. Now consider the associated characteristic Sturmian words $\mathbf{s}$ and $\mathbf{s}'$. Due to the observation made after Definition 43, we may assume that, up to exchanging 0 and 1, both $\mathbf{s}$ and $\mathbf{s}'$ start with the letter 0 (i.e., $q_0, p_0 \geq 1$). The assumption that $\mathsf{lm}_{\mathbf{s}} = \mathsf{lm}_{\mathbf{s}'}$ together with Theorem 44 now implies that the sequences $(|x_i|)_{i\geq 0}$ and $(|y_i|)_{i\geq 0}$ are equal. A simple induction shows that $q_i = p_i$ for all i, therefore $\mathbf{s} = \mathbf{s}'$. □

Observe that Theorem 44 is only true for characteristic Sturmian words since some prefixes of non-characteristic Sturmian words do not admit any string attractor of span 1, as shown in the following example.

Example 47. Let $\mathbf{s} = 0000001000000100000001\cdots$ be a characteristic Sturmian word whose directive sequence begins with $q_0 = 6$ and $q_1 = 2$ and let $\mathbf{x}$ be the non-characteristic Sturmian word such that $\mathbf{s} = 0000 \cdot \mathbf{x}$, hence $\mathbf{x} = 001000000100000001\cdots$. We consider the prefix $\mathbf{x}[1,14] = 0^2 10^6 10^4$. Since 1 occurs only at positions 3 and 10 and the factor 0^6 only in $\mathbf{x}[4,9]$, the candidates as string attractor with two consecutive positions are $\Gamma_1 = \{3,4\}$ and $\Gamma_2 = \{9,10\}$. However, one can check that the factors 0001 and 10^5 do not cross any position in Γ_1 and Γ_2 respectively, showing that $\mathsf{span}_{\mathbf{x}}(14) \geq 2$. Nonetheless, $\mathbf{x}[1,14]$ admits a string attractor of size 2 (but with a larger span), i.e., $\Gamma = \{4,10\}$.

6.2. Characterization of Sturmian and quasi-Sturmian words

We now turn to the families of Sturmian and quasi-Sturmian words. For each, we provide a new characterization in terms of both the span and leftmost complexities.

We start with Sturmian words.

Theorem 48. *An infinite word* $\mathbf{x}$ *is Sturmian if and only if* $\mathsf{lm}_{\mathbf{x}}$ *is unbounded and* $\mathsf{span}_{\mathbf{x}}(n) = 1$ *for infinitely many* $n \geq 1$.

Proof. For the first implication, let $\mathbf{x}$ be a Sturmian word. Since $\mathbf{x}$ is aperiodic, Proposition 40 implies that $\mathsf{lm}_{\mathbf{x}}$ is unbounded. We now establish the claimed property on $\mathsf{span}_{\mathbf{x}}$. As $\mathbf{x}$ is aperiodic and recurrent, it has infinitely many right special prefixes. Moreover, for each such prefix v, there is a characteristic Sturmian word $\mathbf{s}$ (depending on v) having v^R as a prefix [32, Proposition 2.1.23]. Therefore, $\mathsf{span}(v) = \mathsf{span}(v^R) = 1$ for all long enough v by Theorem 44 and the proof of [34, Proposition 11].

For the other implication, consider an infinite word $\mathbf{x}$ satisfying the assumptions. First, it is aperiodic by Proposition 40. For all $m \geq 1$, there exists an integer n such that $\mathbf{x}[1, n]$ contains all length-m factors. By assumption, we can moreover presume that $\mathsf{span}_{\mathbf{x}}(n) = 1$. Therefore, $p_{\mathbf{x}}(m) \leq m + 1$ by Proposition 29. The fact that $\mathbf{x}$ is Sturmian follows from Theorem 2. □

We now turn to quasi-Sturmian words. As announced, we prove a sort of converse of Proposition 41. We will make use of the following characterization of quasi-Sturmian words [7].

Theorem 49 *([7]). An infinite word* $\mathbf{x}$ *over the alphabet* Σ *is quasi-Sturmian if and only if it can be written as* $\mathbf{x} = w\varphi(\mathbf{s})$*, where* w *is a finite word,* $\mathbf{s}$ *is a Sturmian word on the alphabet* $\{0, 1\}$*, and* φ *is a morphism from* $\{0, 1\}^*$ *to* Σ^* *such that* $\varphi(01) \neq \varphi(10)$.

Theorem 50. *An infinite word* $\mathbf{x}$ *is quasi-Sturmian if and only if* $\mathsf{lm}_{\mathbf{x}}$ *is unbounded and there exist a suffix* $\mathbf{y}$ *of* $\mathbf{x}$ *and a constant* $C \geq 1$ *such that* $\mathsf{span}_{\mathbf{y}}(n) \leq C$ *for infinitely many* $n \geq 1$.

Proof. For the first implication, as quasi-Sturmian words are aperiodic by Theorem 2, $\mathsf{lm}_{\mathbf{x}}$ is unbounded by Proposition 40. In addition, by Theorem 49, there exist a finite word w, a Sturmian word $\mathbf{s}$, and a morphism φ such that $\mathbf{x} = w\varphi(\mathbf{s})$. Consider the suffix $\mathbf{y} = \varphi(\mathbf{s})$. By Theorem 48, there are infinitely many integers n such that $\mathsf{span}_{\mathbf{s}}(n) = 1$, and by Proposition 34, there exists a constant $C \geq 1$ such that, for all $N = |\varphi(\mathbf{s}[1, n])|$,

$$\mathsf{span}_{\mathbf{y}}(N) = \mathsf{span}(\varphi(\mathbf{s}[1, n])) \leq C \cdot \mathsf{span}(\mathbf{s}[1, n]) = C.$$

For the other implication, by Propositions 40 and 41, $p_{\mathbf{y}}(n) = n + D$ with $D \leq C$ for all large enough n. Since $\mathbf{x} = w\mathbf{y}$ for some finite word w, we have $p_{\mathbf{x}}(n) \leq p_{\mathbf{y}}(n) + |w| = n + D + |w|$ for all large enough n. We conclude by Theorem 2 that $\mathbf{x}$ is quasi-Sturmian. □

Table 3
The first few finite Tribonacci words $(b_n^{(3)})_{0\leq n\leq 5}$ (some particular decomposition is highlighted for a later purpose, see Proposition 52).

n	0	1	2	3	4	5
$b_n^{(3)}$	0	0 1	01 0 2	0102 01 0	0102010 0102 01	0102010010201 0102010 0102

7. String attractors and complexities for k-bonacci words

In this section, we study string attractors of prefixes of some purely morphic words over an alphabet of size $k \geq 2$, namely the so-called *k-bonacci words*. The case $k = 2$ corresponds to the famous Fibonacci word, which is a Sturmian word and for which string attractor-related concepts have already been studied. For $k = 3$, each prefix of the Tribonacci word admits a string attractor of size at most 3 as shown in [45].

More generally, as k-bonacci words are *episturmian*, Dvořáková showed that each prefix admits a string attractor of size at most k using palindromes [19, Theorem 10]. We also provide (different) string attractors of size at most k, using an approach that differs from the techniques used to obtain string attractors for the Thue–Morse word, the period-doubling word, and standard Sturmian words and may be extended to other purely morphic words. Moreover, we precisely describe our string attractors in terms of the corresponding k-bonacci numbers, which opens the door to considerations related to numeration systems. In fact, a first attempt towards these considerations was done in [22] using a similar construction.

Furthermore, we then study the leftmost and the span complexities of the k-bonacci words.

7.1. Useful definitions and intermediate results

Let us consider an integer $k \geq 2$ and the morphism $\mu_k \colon \{0, \dots, k-1\}^* \to \{0, \dots, k-1\}^*$ defined by $\mu_k(i) = 0(i+1)$ for all $i \in \{0, 1, \dots, k-2\}$ and $\mu_k(k-1) = 0$. The *infinite k-bonacci word* $\mathbf{b}^{(k)}$ is defined as the fixed point $\mathbf{b}^{(k)} = \mu_k^\infty(0)$. The cases $k = 2$ and $k = 3$ correspond to the Fibonacci and Tribonacci words respectively.

Furthermore, for all $n \geq 0$, we let $b_n^{(k)} = \mu_k^n(0)$ denote the *nth finite k-bonacci word*. We also set $b_n^{(k)} = \varepsilon$ for all $-k \leq n < 0$. For any $n \geq 0$, we let $B_n^{(k)} = |b_n^{(k)}|$ denote the length of the nth finite k-bonacci word. The sequence $(B_n^{(k)})_{n\geq 0}$ will be referred to as the sequence of *k-bonacci numbers*. When the context is clear, we will drop the superscript (k) in all notation.

Example 51. For $k = 3$, we write the first few non empty finite Tribonacci words in Table 3.

Another way of seeing the sequence $(b_n^{(k)})_{n\geq -k}$ is the following, which can be proven by an easy induction. See Table 3 for an example with $k = 3$.

Proposition 52. *We have*

$$b_n^{(k)} = \begin{cases} \left(\prod_{i=1}^{k} b_{n-i}^{(k)}\right) \cdot n = \left(\prod_{i=1}^{n} b_{n-i}^{(k)}\right) \cdot n, & \text{if } 0 \le n \le k-1; \\ \prod_{i=1}^{k} b_{n-i}^{(k)}, & \text{if } n \ge k. \end{cases}$$

We now define two sequences of integers $(L_n^{(k)})_{n\ge 0}$ and $(U_n^{(k)})_{n\ge 0}$ linked to k-bonacci numbers that will help us partition $\mathbb{N}$.

Definition 53. For all $n \ge 0$, we set

$$L_n^{(k)} = \begin{cases} B_n^{(k)}, & \text{if } n \le k; \\ B_n^{(k)} + B_{n-k-1}^{(k)} - 1, & \text{otherwise;} \end{cases}$$

and

$$U_n^{(k)} = \sum_{i=0}^{n} B_i^{(k)}.$$

Example 54. When $k = 3$, we obtain $(L_n^{(3)})_{n\ge 0} = 1, 2, 4, 7, 13, 25, 47, 87, \ldots$ and $(U_n^{(3)})_{n\ge 0} = 1, 3, 7, 14, 27, 51, 95, 176, \ldots$.

For any $k \ge 2$, one can show that $(U_n^{(k)})_{n\ge 0}$ gives the lengths of palindromic prefixes of the k-bonacci word (note that the case $k = 3$ gives the sequence [47, A027084]).

Remark 55. Observe that, by Proposition 52, if $1 \le n \le k-1$, then $U_{n-1} = B_n - 1 = L_n - 1$, and if $n = k$, then $U_{k-1} = B_k = L_k$. Moreover, for $n > k$, we have $L_n = \left(\sum_{i=n-k-1}^{n-1} B_i\right) - 1 \le U_{n-1}$. As $L_0 = 1$, this implies that the intervals $[L_n, U_n]$, $n \ge 0$ cover the set of integers $m \ge 1$.

7.2. String attractor profile function

We now study the string attractor profile function of the k-bonacci word $\mathbf{b}^{(k)}$. To do so, we will make use of Proposition 39 therefore we look at prefixes obtained as fractional powers. More specifically, as the string attractors positions will be elements of $(B_n)_{n\ge 0}$, we study fractional powers of the words b_n, $n \ge 0$.

Proposition 56. *For all $n \ge 0$, $\mathbf{b}^{(k)}[1, U_n^{(k)}] = \prod_{i=0}^{n} b_{n-i}^{(k)}$. Moreover, $\mathbf{b}^{(k)}[1, U_n^{(k)}]$ is a fractional power of $b_n^{(k)}$.*

Proof. For $n = 0$, we directly have $\mathbf{b}[1, U_0] = \mathbf{b}[1, 1] = b_0$, so both claims hold in this case. Assume that the result is true for n, and let us prove it for $n+1$. By the induction hypothesis, we have

$$\mu_k(\mathbf{b}[1, U_n]) = \mu_k\left(\prod_{i=0}^{n} b_{n-i}\right) = \prod_{i=0}^{n} b_{n+1-i}.$$

As $\mathbf{b}$ is a fixed point of μ_k, $\mu_k(\mathbf{b}[1, U_n])$ is a prefix of $\mathbf{b}$, and it is followed by the image of a letter, thus by the letter 0. Therefore,

$$\mathbf{b}[1, U_{n+1}] = \mathbf{b}\left[1, \sum_{i=0}^{n+1} B_i\right] = \left(\prod_{i=0}^{n} b_{n+1-i}\right) \cdot 0 = \prod_{i=0}^{n+1} b_{n+1-i}.$$

Moreover, since $\mathbf{b}[1, U_n]$ is a fractional power of b_n by the induction hypothesis, so is $\mathbf{b}[1, U_n] \cdot a$ for some letter $a \in \{0, 1, \ldots, k-1\}$. By applying the morphism μ_k on both words, we can conclude that $\mathbf{b}[1, U_{n+1}] = \mu_k(\mathbf{b}[1, U_n]) \cdot 0$ is a fractional power of $b_{n+1} = \mu_k(b_n)$. □

Using Proposition 39, we then directly have the following corollary.

Corollary 57. *For all $n \geq 0$, if Γ is a string attractor for $\mathbf{b}^{(k)}[1, L_n^{(k)}]$ and if $B_n^{(k)} \in \Gamma$, then Γ is a string attractor for $\mathbf{b}^{(k)}[1, m]$ for all $m \in [L_n^{(k)}, U_n^{(k)}]$.*

We now exhibit a minimum string attractor of size at most k for each prefix of $\mathbf{b}^{(k)}$ and deduce the string attractor profile function.

Theorem 58. *For all $n \geq 0$, the set*

$$\Gamma_n = \begin{cases} \{B_0^{(k)}, \ldots, B_n^{(k)}\}, & \text{if } n \leq k-1; \\ \{B_{n-k+1}^{(k)}, \ldots, B_n^{(k)}\}, & \text{if } n \geq k; \end{cases}$$

is a minimum string attractor for $\mathbf{b}^{(k)}[1, m]$, for all $m \in [L_n^{(k)}, U_n^{(k)}]$. In particular, the string attractor profile function for $\mathbf{b}^{(k)}$ is given by

$$s_{\mathbf{b}^{(k)}}(n) = \begin{cases} i+1, & \text{if } B_i^{(k)} \leq n < B_{i+1}^{(k)} \text{ for some } i \leq k-2; \\ k, & \text{if } n \geq B_{k-1}^{(k)}. \end{cases}$$

Proof. Using Proposition 52, a simple induction shows that, for all $n \geq 0$, the positions of Γ_n correspond to different letters, which implies that, if Γ_n is a string attractor of a prefix, it is minimum. We prove that it is a string attractor of the length-m prefix, $m \in [L_n, U_n]$, by induction on $n \geq 0$. More precisely, the induction step is divided into three intermediary claims (where we take the convention that $\Gamma_{-1} = \emptyset$):

1. $\Gamma_{n-1} \cup \{B_n\}$ is a string attractor for $\mathbf{b}[1, L_n]$;
2. Γ_n is a string attractor for $\mathbf{b}[1, L_n]$;
3. Γ_n is a string attractor for $\mathbf{b}[1, m]$ for all $m \in [L_n, U_n]$.

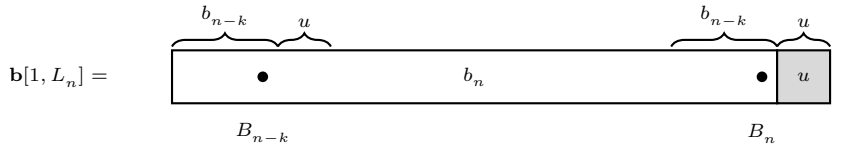

Fig. 1. Case 1 in the proof of Theorem 58.

First, notice that for $n \leq k-1$, the first two claims are identical. Second, observe that for all n, the third claim is a direct consequence of the second claim and of Corollary 57.

Let us now proceed to the induction. If $n = 0$, we directly conclude that $\{1\}$ is a string attractor for $\mathbf{b}[1, 1]$, which shows the first claim (and the other two follow as explained above).

If $1 \leq n \leq k-1$, then $L_n = U_{n-1}+1 = B_n$. Since Γ_{n-1} is a string attractor of $\mathbf{b}[1, U_n]$ by the induction hypothesis on the third claim, this directly implies that $\Gamma_{n-1} \cup \{B_n\}$ is a string attractor for $\mathbf{b}[1, L_n]$. Once again, the other two claims directly follow.

Assume now that $n \geq k$ and let us prove the first claim. Then $L_n \in [L_{n-1}, U_{n-1}]$, which implies as above that $\Gamma_{n-1} \cup \{B_n\}$ is a string attractor for $\mathbf{b}[1, L_n]$.

Let us prove the second claim, and let $\mathbf{b}[1, L_n] = b_n u$, where $u = \varepsilon$ if $n = k$ or u is b_{n-k-1} without its last letter if $n \geq k+1$. Using the first claim, it remains to show that the position B_{n-k} is not needed in the string attractor, i.e., the factors of $\mathbf{b}[1, L_n]$ that are covered by position B_{n-k} are still covered by Γ_n. As the first position in Γ_n is B_{n-k+1}, it suffices to consider the factor occurrences crossing position B_{n-k} in $\mathbf{b}[1, B_{n-k+1} - 1]$. As $\mathbf{b}[1, B_{n-k+1} - 1]$ is b_{n-k+1} without its last letter, Proposition 52 implies that they are occurrences in

$$\prod_{i=1}^{k} b_{n-k+1-i} = b_{n-k} b_{n-k-1} \prod_{i=3}^{k} b_{n-k+1-i}.$$

Note that $b_{n-k}u$ is a prefix of this word. We consider two cases: either the considered occurrence is entirely contained in $b_{n-k}u$ or it crosses position $B_{n-k} + B_{n-k-1}$. Observe that, if $n \geq k+1$, these two cases are mutually exclusive.

Case 1. Since b_{n-k} is a suffix of b_n by Proposition 52, the factors having an occurrence in $b_{n-k}u$ crossing position B_{n-k} have an occurrence in $b_n u$ crossing position B_n, so they are covered by Γ_n. See Fig. 1.

Case 2. Similarly, by Proposition 52, $b_{n-k}b_{n-k-1}$ is a suffix of b_{n-1} and $\prod_{i=3}^{k} b_{n-k+1-i}$ $= \prod_{i=1}^{k-2} b_{n-k-1-i}$ is a prefix of b_{n-k-1}, so of b_{n-2} (as the finite k-bonacci words are prefixes of each other). As $b_{n-1}b_{n-2}$ is a prefix of b_n, we conclude that the factors having an occurrence in $\mathbf{b}[1, B_{n-k+1} - 1]$ crossing position $B_{n-k} + B_{n-k-1}$ have an occurrence in b_n crossing position B_{n-1}, so they are covered by Γ_n. See Fig. 2. This ends the proof of the second claim.

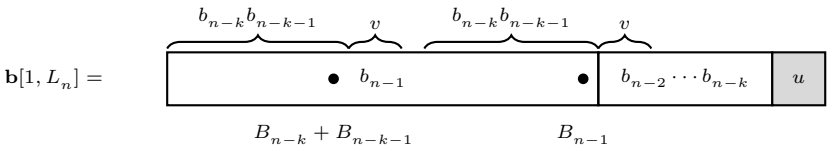

Fig. 2. Case 2 in the proof of Theorem 58 with $v = \prod_{i=3}^{k} b_{n-k+1-i}$.

The third claim then follows, and this ends the proof that, for all $n \geq 0$ and for all $m \in [L_n, U_n]$, Γ_n is a string attractor of the length-m prefix of $\mathbf{b}$. Finally, the string attractor profile function follows from Remark 55. □

Remark 59. In the Tribonacci case, the elements in our string attractors are the same as in [45, Theorem 6]. The corresponding intervals of prefix lengths are also linked. Indeed, our sequence $(U_n^{(3)})_{n\geq 0}$ is related to the sequence $(W_n)_{n\geq 4}$ defined in [45, Theorem 6] as follows: we have $W_{n+3} = U_{n+1}^{(3)}$ for all $n \geq 1$. Therefore, our upper bounds and that of Schaeffer and Shallit coincide. However, our lower bounds are smaller than theirs. On the other hand, the string attractors obtained for the palindromic prefixes in [19] are different from ours. For instance, the case $k = 3$ is treated in [19, Example 8].

7.3. Leftmost complexity

We can further prove that the string attractor from Theorem 58 is actually a leftmost string attractor. For the following few results, we set $U_{-1}^{(k)} = 0$.

Proposition 60. *The leftmost complexity of* $\mathbf{b}^{(k)}$ *satisfies* $\mathsf{lm}_{\mathbf{b}^{(k)}}(m) = B_n^{(k)}$ *for all* $n \geq 0$ *and* $m \in [U_{n-1}^{(k)} + 1, U_n^{(k)}]$.

Proof. We show that the factor $\mathbf{b}[B_n, U_{n-1} + 1]$ does not occur in $\mathbf{b}$ before position B_n. This implies that, for all $m \geq U_{n-1} + 1$, any string attractor of $\mathbf{b}[1, m]$ contains a position at least equal to B_n and, combined with Theorem 58, proves the claimed leftmost complexity.

The claim is direct for $n = 0$ as $B_0 = 1 = U_{-1} + 1$. Assume that it is true for n, and let us prove it for $n + 1$. By construction, the B_{n+1}th letter of $\mathbf{b}$ is the last letter of the image of the B_nth letter under μ_k, and, by Proposition 56, $\mathbf{b}[B_{n+1} + 1, U_n + 1]$ is the image of $\mathbf{b}[B_n + 1, U_{n-1} + 1]$, potentially followed by a letter 0 (this occurs when $\mathbf{b}[B_n, U_{n-1}+1]$ ends with the letter $k-1$). Therefore, each occurrence of $\mathbf{b}[B_{n+1}, U_n+1]$ in $\mathbf{b}$ is associated with the image of an occurrence of $\mathbf{b}[B_n, U_{n-1}+1]$. Using the induction hypothesis, we conclude that $\mathbf{b}[B_{n+1}, U_n + 1]$ does not occur before position B_{n+1}. □

7.4. Span complexity

For the k-bonacci words $\mathbf{b}^{(k)}$, the factor complexity function is given by $p_{\mathbf{b}^{(k)}}(n) = (k-1)n+1$ (see, for instance, [14,23]). Therefore, when $k \geq 3$, Proposition 42 implies that the

span complexity is linear. However, the string attractors described in Section 7.2 do not have the smallest difference between their extreme positions. In what follows, we compute the span for infinitely many prefixes and describe string attractors (of unbounded size) having that span.

We first make the following observation which gives a lower bound on the span. Recall that we have set $U_{-1}^{(k)} = 0$.

Proposition 61. *Let $k \geq 2$. For all $n \geq 2$, the factors $\mathbf{b}^{(k)}[i, i + U_{n-3}^{(k)}]$ are distinct for all $i \in [B_{n-2}^{(k)} + 1, B_n^{(k)}]$.*

Proof. Let us prove the result by induction on n. For $n = 2$, we need to consider the letters in $u = \mathbf{b}[2, B_2]$. If $k = 2$, then $u = 10$, and if $k \geq 3$, then $u = 102$. In both cases, the letters are indeed distinct.

Let us now assume that the claim is true for $n \geq 2$ and let us prove it for $n + 1$. We proceed by contradiction and assume that there exist $i, j \in [B_{n-1} + 1, B_{n+1}]$ minimal such that $i < j$ and $\mathbf{b}[i, i + U_{n-2}] = \mathbf{b}[j, j + U_{n-2}]$. As $B_{n-1} + 1$ marks the beginning of the image of a letter in $\mathbf{b}$ and i and j are taken minimal, we know that the factor $u = \mathbf{b}[i, i+U_{n-2}] = \mathbf{b}[j, j+U_{n-2}]$ begins with 0. We may also assume that it does not end with 0. Indeed, otherwise, we consider the word $u = \mathbf{b}[i, i+U_{n-2}-1] = \mathbf{b}[j, j+U_{n-2}-1]$ instead.

As the word u starts with 0, there exist $i' < j'$ such that $\mu_k(\mathbf{b}[1, i' - 1]) = \mathbf{b}[1, i - 1]$ and $\mu_k(\mathbf{b}[1, j' - 1]) = \mathbf{b}[1, j - 1]$. Moreover, as $U_{n-2} + 1 \geq 2$ and as u does not end with a 0, it can be uniquely desubstituted (i.e., its preimage under μ_k is unique). There thus exists ℓ such that $\mathbf{b}[i', i' + \ell] = \mathbf{b}[j', j' + \ell]$ and $\mu_k(\mathbf{b}[i', i' + \ell]) = u$.

As $|\mu_k(\mathbf{b}[1, i' - 1])| = i - 1 \in [B_{n-1}, B_{n+1} - 1]$, we have $i' \in [B_{n-2} + 1, B_n]$. The same holds for j'. Therefore, by the induction hypothesis, we have $\mathbf{b}[i', i' + U_{n-3}] \neq \mathbf{b}[j', j' + U_{n-3}]$. Let us take $\ell' \in [\ell, U_{n-3} - 1]$ maximal such that $\mathbf{b}[i', i' + \ell'] = \mathbf{b}[j', j' + \ell']$ and let $v = \mathbf{b}[i', i' + \ell']$. By maximality of ℓ', v is right special. Moreover, the set of factors of $\mathbf{b}$ is stable under reversal [14, Theorem 5], i.e., the reversal of any factor of $\mathbf{b}$ is also a factor. In particular, v^R is a left special factor of $\mathbf{b}$. Furthermore, the left special factors of $\mathbf{b}$ are exactly its prefixes [14, Proposition 5], so v^R is a prefix of $\mathbf{b}$ and also of $\mathbf{b}[1, U_{n-3}]$ as $\ell' \leq U_{n-3} - 1$. However, we have

$$|\mu_k(v^R)| = |\mu_k(v)| \geq |\mu_k(\mathbf{b}[i', i' + \ell])| \geq U_{n-2}$$

by definition of ℓ. This is a contradiction. In fact, by Proposition 56, we have $|\mu_k(v^R)| \leq |\mu_k(\mathbf{b}[1, U_{n-3}])| < U_{n-2}$. □

We now describe a new string attractor for prefixes of the k-bonacci word.

Proposition 62. *Let $k \geq 2$. For all $n \geq 1$ and for all $m \in [U_{n-1}^{(k)} + 1, U_n^{(k)}]$, $\Gamma_n = \{U_{n-2}^{(k)} + 1, U_{n-2}^{(k)} + 2, \ldots, B_n^{(k)}\}$ is a string attractor of $\mathbf{b}^{(k)}[1, m]$.*

Proof. We proceed by induction on $n \geq 1$. For the base case $n = 1$, the interval $[U_{1-1} + 1, U_2]$ becomes $[2, 3]$, and $\Gamma_1 = \{1, 2\}$, so the conclusion follows.

Now assume that the result is true for $n \geq 1$, and we show it also holds for $n+1$. To do so, we will use the following observation. From Proposition 56 and [2, Proposition 4.4], one may prove that $\mathbf{b}[1, U_n]$ is a palindrome for all $n \geq -1$. By the induction hypothesis, Γ_n is a string attractor for $\mathbf{b}[1, U_n]$. As this word is a palindrome, it also has the string attractor

$$\Gamma_n^R = \{U_n + 1 - B_n, \ldots, U_n + 1 - U_{n-2} - 1\} = \{U_{n-1} + 1, \ldots, B_n + B_{n-1}\}.$$

In particular, $\Gamma_{n+1} \supseteq \Gamma_n^R$ is a string attractor of $\mathbf{b}[1, U_n]$ when $B_{n+1} \leq U_n$. If $B_{n+1} > U_n$, then $n \leq k - 1$ and $B_{n+1} = U_n + 1$, so Γ_{n+1} is a string attractor of $\mathbf{b}[1, U_n + 1]$. In both cases, Propositions 39 and 56 imply that Γ_{n+1} is a string attractor of $\mathbf{b}[1, m]$ for all $m \in [U_n + 1, U_{n+1}]$. □

Corollary 63. *Let $k \geq 2$. For all $n \geq 2$ and for all $m \in [U_n^{(k)} - B_{n-1}^{(k)} - B_{n-2}^{(k)}, U_n^{(k)}]$, we have $\mathsf{span}_{\mathbf{b}^{(k)}}(m) = B_n^{(k)} - U_{n-2}^{(k)} - 1$. In particular, for infinitely many prefixes, there is a factor length for which the bound given by Proposition 29 is tight.*

Proof. Using Propositions 61 and 29, we know that for $m \geq B_n + U_{n-3}$, we have $\mathsf{span}_{\mathbf{b}}(m) \geq B_n - B_{n-2} - U_{n-3} - 1 = B_n - U_{n-2} - 1$. Observe that $B_n + U_{n-3} = U_n - B_{n-1} - B_{n-2}$. On the other hand, using Proposition 62, we know that for $m \in [U_{n-1} + 1, U_n]$, we have $\mathsf{span}_{\mathbf{b}}(m) \leq B_n - U_{n-2} - 1$.

If $k \geq 3$, then $B_n + U_{n-3} \geq U_{n-1} + 1$ so, for all $m \in [U_n - B_{n-1} - B_{n-2}, U_n]$, we have $\mathsf{span}_{\mathbf{b}}(m) = B_n - U_{n-2} - 1$, as desired. It remains to consider $k = 2$. In that case, $B_n - U_{n-2} - 1 = 1$ which does not depend on n. Therefore $\mathsf{span}_{\mathbf{b}}(m) \geq 1$ for all $m \geq B_2 + U_{-1} = 3$ and $\mathsf{span}_{\mathbf{b}}(m) \leq 1$ for all $m \geq U_0 + 1 = 2$. Therefore, the conclusion follows for all $m \geq 3$. □

Observe that, for the Fibonacci word, we once again obtain that $\mathsf{span}_{\mathbf{b}^{(2)}} = 1$, as in Theorem 44.

8. Conclusions

In this paper, we emphasized the close relationship between string attractor-based measures and classical notions of repetitiveness on infinite words, such as the factor complexity and the recurrence function. In particular, we identified some combinatorial properties needed to have a bounded string attractor profile function. Nonetheless, a complete characterization of these words is still missing.

Furthermore, we used the new leftmost and span complexities to obtain novel characterizations of particular infinite words, such as periodic words and the families of Sturmian and quasi-Sturmian words. We wonder if one can use different string attractor-based measures to characterize other combinatorial properties or families of words.

Finally, for the characteristic Sturmian words and the k-bonacci words, we have shown how to construct, for each prefix, a string attractor with minimum size, minimum leftmost measure, or minimum span. The methods presented for the k-bonacci words rely on the properties they inherit from their morphic construction. A future perspective of research could be a generalization of such a strategy to extend the construction of a smallest string attractor to other families of morphic sequences.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We thank Julien Leroy for the fruitful discussion on the S-adic words which led to the construction of the infinite word from Example 16.

References

[1] J.P. Allouche, J.O. Shallit, Automatic Sequences: Theory, Applications, Generalizations, Cambridge University Press, 2003.
[2] P. Ambrož, Z. Masáková, E. Pelantová, C. Frougny, Palindromic complexity of infinite words associated with simple Parry numbers, Ann. Inst. Fourier (2006) 2131–2160.
[3] M.P. Béal, D. Perrin, A. Restivo, Decidable problems in substitution shifts, J. Comput. Syst. Sci. 143 (2024) 103529, https://doi.org/10.1016/j.jcss.2024.103529.
[4] V. Becher, P.A. Heiber, On extending de Bruijn sequences, Inf. Process. Lett. 111 (2011) 930–932, https://doi.org/10.1016/j.ipl.2011.06.013.
[5] D. Bulgakova, A. Frid, J. Scanvic, Prefix palindromic length of the Sierpinski word, in: Developments in Language Theory, in: Lecture Notes in Comput. Sci., vol. 13257, Springer, Cham, 2022, pp. 78–89.
[6] J. Cassaigne, Complexité et facteurs spéciaux, Bull. Belg. Math. Soc. Simon Stevin 4 (1997) 67–88.
[7] J. Cassaigne, Sequences with grouped factors, in: Developments in Language Theory, Aristotle University of Thessaloniki, 1997, pp. 211–222.
[8] J. Cassaigne, Recurrence in infinite words, in: STACS, Springer, 2001, pp. 1–11.
[9] J. Cassaigne, F. Nicolas, Factor complexity, in: V. Berthé, M. Rigo (Eds.), Combinatorics, Automata and Number Theory, vol. 135, Cambridge University Press, 2010, pp. 163–247.
[10] G. Castiglione, A. Restivo, M. Sciortino, Hopcroft's algorithm and cyclic automata, in: LATA, Springer, 2008, pp. 172–183.
[11] G. Castiglione, A. Restivo, M. Sciortino, Circular Sturmian words and Hopcroft's algorithm, Theor. Comput. Sci. 410 (2009) 4372–4381.
[12] S. Constantinescu, L. Ilie, The Lempel–Ziv complexity of fixed points of morphisms, SIAM J. Discrete Math. 21 (2007) 466–481.
[13] F. Dolce, String attractors for factors of the Thue-Morse word, in: WORDS, Springer, 2023, pp. 117–129.
[14] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, Theor. Comput. Sci. 255 (2001) 539–553.

[15] F. Durand, Linearly recurrent subshifts have a finite number of non-periodic subshift factors, Ergod. Theory Dyn. Syst. 20 (2000) 1061–1078.
[16] F. Durand, Corrigendum and addendum to: "Linearly recurrent subshifts have a finite number of non-periodic subshift factors" [Ergodic Theory Dynam. Systems **20** (2000), no. 4, 1061–1078, MR1779393 (2001m:37022)], Ergod. Theory Dyn. Syst. 23 (2003) 663–669, https://doi.org/10.1017/S0143385702001293.
[17] F. Durand, B. Host, C. Skau, Substitutional dynamical systems, Bratteli diagrams and dimension groups, Ergod. Theory Dyn. Syst. 19 (1999) 953–993.
[18] F. Durand, D. Perrin, Dimension Groups and Dynamical Systems: Substitutions, Bratteli Diagrams and Cantor Systems, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2022.
[19] Ľ. Dvořáková, String attractors of episturmian sequences, Theor. Comput. Sci. 986 (2024) 114341, https://doi.org/10.1016/j.tcs.2023.114341.
[20] Ľ. Dvořáková, V. Hendrychová, String attractors of Rote sequences, https://doi.org/10.48550/arXiv.2308.00850, 2023.
[21] A. Frosini, I. Mancini, S. Rinaldi, G. Romana, M. Sciortino, Logarithmic equal-letter runs for BWT of purely morphic words, in: DLT, Springer, 2022, pp. 139–151.
[22] F. Gheeraert, G. Romana, M. Stipulanti, String attractors of fixed points of k-bonacci-like morphisms, in: WORDS, Springer, 2023, pp. 192–205.
[23] A. Glen, On Sturmian and episturmian words, and related topics, Ph.D. thesis, University of Adelaide, Australia, 2006.
[24] A. Heinis, Languages under substitutions and balanced words, J. Théor. Nr. Bordx. 16 (2004) 151–172.
[25] Š. Holub, Words with unbounded periodicity complexity, Int. J. Algebra Comput. 24 (2014) 827–836.
[26] D. Kempa, N. Prezza, At the roots of dictionary compression: string attractors, in: STOC, ACM, 2018, pp. 827–840.
[27] D.E. Knuth, J.H.M. Jr., V.R. Pratt, Fast pattern matching in strings, SIAM J. Comput. 6 (1977) 323–350.
[28] T. Kociumaka, G. Navarro, N. Prezza, Toward a definitive compressibility measure for repetitive sequences, IEEE Trans. Inf. Theory 69 (2023) 2074–2092.
[29] D. Kosolobov, A.M. Shur, Comparison of LZ77-type parsings, Inf. Process. Lett. 141 (2019) 25–29.
[30] K. Kutsukake, T. Matsumoto, Y. Nakashima, S. Inenaga, H. Bannai, M. Takeda, On repetitiveness measures of Thue-Morse words, in: SPIRE, Springer, 2020, pp. 213–220.
[31] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inf. Theory 22 (1976) 75–81.
[32] M. Lothaire, Algebraic Combinatorics on Words, vol. 90, Cambridge University Press, 2002.
[33] A. de Luca, F. Mignosi, Some combinatorial properties of Sturmian words, Theor. Comput. Sci. 136 (1994) 361–385.
[34] S. Mantaci, A. Restivo, G. Romana, G. Rosone, M. Sciortino, A combinatorial view on string attractors, Theor. Comput. Sci. 850 (2021) 236–248.
[35] S. Mantaci, A. Restivo, M. Sciortino, Burrows-Wheeler transform and Sturmian words, Inf. Process. Lett. 86 (2003) 241–246.
[36] M. Morse, G.A. Hedlund, Symbolic dynamics, Am. J. Math. 60 (1938) 815–866.
[37] H. Mousavi, Automatic theorem proving in Walnut, https://doi.org/10.48550/arXiv.1603.06017, 2016.
[38] G. Navarro, The compression power of the BWT: technical perspective, Commun. ACM 65 (2022) 90.
[39] G. Navarro, Indexing highly repetitive string collections, part I: repetitiveness measures, ACM Comput. Surv. 54 (2022) 29:1–29:31.
[40] G. Navarro, Indexing highly repetitive string collections, part II: compressed indexes, ACM Comput. Surv. 54 (2022) 26:1–26:32.
[41] J.J. Pansiot, Complexité des facteurs des mots infinis engendrés par morphimes itérés, in: ICALP, Springer, 1984, pp. 380–389.
[42] J.J. Pansiot, Decidability of periodicity for infinite words, RAIRO Theor. Inform. Appl. 20 (1986) 43–46.
[43] G. Rauzy, Mots infinis en arithmetique, in: M. Nivat, D. Perrin (Eds.), Automata on Infinite Word, vol. 192, Springer Berlin Heidelberg, 1985, pp. 164–171.
[44] A. Restivo, G. Romana, M. Sciortino, String attractors and infinite words, in: LATIN, Springer, 2022, pp. 426–442.

[45] L. Schaeffer, J.O. Shallit, String attractors for automatic sequences, https://doi.org/10.48550/arXiv.2012.06840, 2021.
[46] M. Sciortino, L.Q. Zamboni, Suffix automata and standard Sturmian words, in: Developments in Language Theory, Springer, 2007, pp. 382–398.
[47] N.J.A. Sloane, The on-line encyclopedia of integer sequences, http://oeis.org, 1964.