On Extended Boundary Sequences of Morphic and Sturmian Words

Michel Rigo
Department of Mathematics, University of Liège, Belgium

Manon Stipulanti
Department of Mathematics, University of Liège, Belgium

Markus A. Whiteland
Department of Mathematics, University of Liège, Belgium

Abstract

Generalizing the notion of the boundary sequence introduced by Chen and Wen, the nth term of the ℓ-boundary sequence of an infinite word is the finite set of pairs (u, v) of prefixes and suffixes of length ℓ appearing in factors uyv of length $n + \ell$ ($n \geq \ell \geq 1$). Otherwise stated, for increasing values of n, one looks for all pairs of factors of length ℓ separated by $n - \ell$ symbols.

For the large class of addable numeration systems U, we show that if an infinite word is U-automatic, then the same holds for its ℓ-boundary sequence. In particular, they are both morphic (or generated by an HD0L system). We also provide examples of numeration systems and U-automatic words with a boundary sequence that is not U-automatic. In the second part of the paper, we study the ℓ-boundary sequence of a Sturmian word. We show that it is obtained through a sliding block code from the characteristic Sturmian word of the same slope. We also show that it is the image under a morphism of some other characteristic Sturmian word.

2012 ACM Subject Classification Theory of computation → Regular languages; Mathematics of computing → Combinatorics on words

Keywords and phrases Boundary sequences, Sturmian words, Numeration systems, Automata, Graph of addition

Digital Object Identifier 10.4230/LIPIcs.MFCS.2022.79

Related Version *Full Version*: https://arxiv.org/abs/2206.15319

Funding *Manon Stipulanti*: Supported by the FNRS Research grant 1.B.397.20F.
Markus A. Whiteland: Supported by the FNRS Research grant 1.B.466.21F.

Acknowledgements We thank Jean-Paul Allouche for references [10, 30, 31], and Jeffrey Shallit for discussions about the "logical approach". The anonymous referees are warmly thanked for providing useful feedback improving the quality of the text.

1 Introduction

Let $\mathbf{x}$ be an infinite word, i.e., a sequence of letters belonging to a finite alphabet. Imagine a window of size n moving along $\mathbf{x}$. Such a reading frame permits to detect all factors of length n occurring in $\mathbf{x}$. For instance, the factor complexity function of $\mathbf{x}$ mapping $n \in \mathbb{N}$ to the number of distinct factors of length n is extensively studied in combinatorics on words. Now let n, ℓ be such that $n \geq \ell$. Assume that within the sliding window, we only focus on its first and last ℓ symbols. Otherwise stated, for a factor uyv of length n, we only consider its borders u and v of length ℓ.

For any given window length n, we would like to determine what are the pairs of length-ℓ borders that may occur. This leads to the following definition, where, to simplify notation, we consider borders of factors of length $n + \ell$ rather than n.

47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022).
Editors: Stefan Szeider, Robert Ganian, and Alexandra Silva; Article No. 79; pp. 79:1–79:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

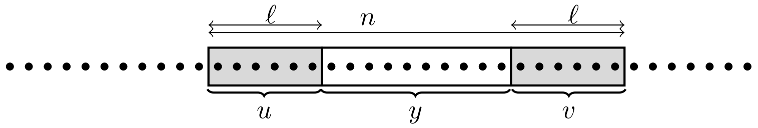

Figure 1 A sliding window where we focus on two regions of a fixed length.

▶ **Definition 1.1.** *Let $\ell \in \mathbb{N}_{>0}$ and $\mathbf{x} \in A^{\mathbb{N}}$. For $n \geq \ell$, we define the nth* boundary set *by*

$$\partial_{\mathbf{x},\ell}[n] := \{(u,v) \in A^\ell \times A^\ell \mid uyv \text{ is a factor of } \mathbf{x} \text{ for some } y \in A^{n-\ell}\}$$

*and call the sequence $\partial_{\mathbf{x},\ell} := (\partial_{\mathbf{x},\ell}[n])_{n\geq\ell}$ the ℓ-*boundary sequence *of $\mathbf{x}$. When $\ell = 1$, we write $\partial_{\mathbf{x},1} = \partial_{\mathbf{x}}$ and simply talk about the* boundary sequence.

The ℓ-boundary sequence takes values in $2^{A^\ell \times A^\ell}$, and hence itself can be seen as an infinite word over a finite alphabet. We give an introductory example.

▶ **Example 1.2.** Consider the Fibonacci word $\mathbf{f} = 0100101001\cdots$; the fixed point of the morphism $0 \mapsto 01, 1 \mapsto 0$. We have $\partial_{\mathbf{f}} = a\,b\,b\,a\,b\,b\,b\,b\,a\,b\,b\,a\,b\,b\,b\,b\,a\,b\,b\,b\,b\,a\,b\,b\,a\,b\,b\,b\,b\cdots$, where $a := \{(0,0),(0,1),(1,0)\}$ and $b := \{0,1\} \times \{0,1\}$. For instance, $\partial_{\mathbf{f}}[1] = a$ because the length-2 factors of $\mathbf{f}$ are $00, 01, 10$, while $\partial_{\mathbf{f}}[2] = b$ because its length-3 factors are of the form 0_0, 0_1, 1_0, 1_1 (they are in fact $010, 001, 100, 101$). The 2-boundary sequence starts with

$$\partial_{\mathbf{f},2} = a_0\ a_1\ a_2\ a_3\ a_4\ a_5\ a_1\ a_2\ a_3\ a_1\ a_2\ a_3\ a_4\ a_5\ a_1\ a_2\ a_3\ a_4\ a_5\ a_1\ a_2\ a_3\ a_1\ a_2\ a_3\cdots$$

where

$$\begin{aligned}
a_0 &:= \{(00,10),(01,00),(01,01),(10,01),(10,10)\},\\
a_1 &:= \{(00,00),(00,01),(01,01),(01,10),(10,00),(10,10)\},\\
a_2 &:= \{(00,01),(00,10),(01,00),(01,10),(10,00),(10,01)\},\\
a_3 &:= \{(00,00),(00,10),(01,00),(01,01),(10,01),(10,10)\},\\
a_4 &:= \{(00,01),(01,01),(01,10),(10,00),(10,01),(10,10)\},\\
a_5 &:= \{(00,10),(01,00),(01,01),(01,10),(10,01),(10,10)\}.
\end{aligned}$$

The first element $\partial_{\mathbf{f},2}[2] = a_0$ is peculiar; it corresponds exactly to the five length-4 factors occurring in $\mathbf{f}$. Our Proposition 4.8 shows that a_0 appears only once in $\partial_{\mathbf{f}}$. Then, e.g., $\partial_{\mathbf{f},2}[3] = a_1$ because the length-5 factors of $\mathbf{f}$ are of the form 00_00, 00_01, 01_01, 01_10, 10_00 and 10_10 (the factors are 00100, 00101, 01001, 10100, 10010, and 01010). For length-6 factors, note that two are of the form $10u01$ for some $u \in \{0,1\} \times \{0,1\}$. For $i \geq 1$, the letter a_i appears infinitely often in $\partial_{\mathbf{f},2}$: see Theorem 4.1.

1.1 Motivation and related work

In combinatorics on words, borders and boundary sets are related to important concepts. For instance, a word v is *bordered* if there exist u, x, y such that $v = ux = yu$ and $0 < |u| < |v|$. One reason to study bordered words is Duval's theorem: for a sufficiently long word v, the maximum length of unbordered factors of v is equal to the period of v [14]. In formal language theory, a language L is *locally ℓ-testable* (LT) if the membership of a word w in L only depends on the prefix, suffix and factors of length ℓ of w. In [35], the authors consider the so-called *separating problem* of languages by LT languages; they utilize *ℓ-profiles* of a word, which can again be related to boundary sets. Let us also mention that, in bioinformatics and computational biology, one of the aims is to reconstruct sequences from subsequences [26]. To determine DNA segments by bottom-up analysis, *paired-end* sequencing is used. In this case both ends of DNA fragments of known length are sequenced. See, for instance, [18]. This is quite similar to the theoretical concept we discuss here.

The notion of a (1-)boundary sequence was introduced by Chen and Wen in [8] and was further studied in [19], where it is shown that the boundary sequence of a k*-automatic* word (in the sense of Allouche and Shallit [1]: see Definition 2.2) is k-automatic. It is well-known that a k-automatic word $\mathbf{x}$ is *morphic*, i.e., there exist morphisms $f\colon A \to A^*$ and $g\colon A \to B$ and a letter $a \in A$ such that $\mathbf{x} = g(f^\omega(a))$, where $f^\omega(a) = \lim_{n\to\infty} f^n(a)$. However, k-automatic words (with k ranging over the integers) do not capture all morphic words: a well-known characterization of k-automatic words is given by Cobham [9] (the generating morphism f maps each letter to a length-k word). This paper is driven by the natural question whether, in general, the ℓ-boundary sequence of a morphic word is morphic. In case such generating morphisms can be constructed, we have at our disposal a simple algorithm providing the set of length-ℓ borders in factors of all lengths.

We briefly present several situations in which the notion of boundary sets is explicitly or implicitly used. In [12, Thm. 4], the authors study the boundary sequence to exhibit a squarefree word for which each subsequence arising from an arithmetic progression contains a square. Boundary sets play an important role in the study of so-called k*-abelian* and k*-binomial complexities* of infinite words (for definitions, see [37]). For instance, computing the 2-binomial complexity of generalized Thue–Morse words [25] requires inspecting pairs of prefixes and suffixes of factors, which is again related to the boundary sequence when these prefixes and suffixes have equal length. The k-binomial complexities of images of binary words under powers of the Thue–Morse morphism are studied in [39]; there some general properties of boundary sequences of binary words are required. Moreover, if $\partial_{\mathbf{x}}$ is automatic, then the abelian complexity of the image of $\mathbf{x}$ under a so-called Parikh-constant morphism is automatic [8]. Guo, Lü, and Wen combine this result with theirs in [19] to establish a large family of infinite words with automatic abelian complexity.

Let $k \geq 1$. We let $\equiv_k$ denote the k-abelian equivalence, i.e., $u \equiv_k v$ if the words u and v share the same set of factors of length at most k with the same multiplicities [22]. For u and v equal length factors of a Sturmian word $\mathbf{s}$, we have $u \equiv_k v$ if and only if they share a common prefix and a common suffix of length $\min\{|u|, k-1\}$ and $u \equiv_1 v$ [22, Prop. 2.8]. Under the assumption that the largest power of a letter appearing in $\mathbf{s}$ is less than $2k-2$, the requirement $u \equiv_1 v$ in the previous result may be omitted [33, Thm. 3.6] (compare to Proposition 4.8). Thus the quotient of the set of factors of length n occurring in a Sturmian word by the relation $\equiv_k$ is completely determined by $\partial_{\mathbf{s},k-1}[n-k+1]$ for large enough k (depending on $\mathbf{s}$). Other families of words with k-abelian equivalence determined by the boundary sets are given in [33, Prop. 4.2].

1.2 Our contributions

Up to our knowledge, we are the first to propose a systematic study of the ℓ-boundary sequences of infinite words. It is therefore natural to consider the notion on well-known classes of words. In this paper, we consider morphic words and Sturmian words.

Any morphic word is S-automatic for some abstract numeration system S [38]. With Theorem 3.1, we prove that for a large class of numeration systems U, if $\mathbf{x}$ is a U-automatic word, then the boundary sequence $\partial_{\mathbf{x}}$ is again U-automatic. Our approach generalizes the arguments provided by [19]. Considering exotic numeration systems allows a better understanding of underlying mechanisms, which do not arise in the ordinary integer base systems. In particular, we deal with addition within the numeration system; in integer base systems, the carry propagation is easy to handle (by a two-state finite automaton). Our arguments apply to so-called *addable* numeration systems (see Definition 2.1).

As an alternative, we observe that the Büchi–Bruyère theorem [5] can be extended to addable positional numeration systems U (Theorem 2.6). The U-automaticity of the ℓ-boundary sequence then follows from the fact that it is definable by a first-order formula of the structure $\langle\mathbb{N},+\rangle$ extended with a unary function relating an integer with the least element of U properly appearing in its representation.

This alternative proof however hides the important details that might help identifying the technical limits of the result: not all morphic words allow an addable system to work with. However, the framework we consider captures all morphic words (see Theorem 2.4). Also, one practical difficulty when one wants to use automatic provers (such as Walnut [28]) is to be able to provide the relevant automaton for addition. To identify the contours of our result, we also discuss the case where $\mathbf{x}$ is U-automatic and $\partial_{\mathbf{x}}$ is not U-automatic. To construct such examples, we have to consider non-addable numeration systems in Section 3.2.

We then turn to the other class of words under study. Letting $\mathbf{s}$ be a Sturmian word with slope α, with Theorem 4.1 we show that the ℓ-boundary sequence of $\mathbf{s}$ is obtained through a sliding block code from the *characteristic Sturmian word of slope* α (see Section 4 for a definition) up to the first letter. This result holds even for non-morphic Sturmian words, so for an arbitrary irrational α. Where the techniques used in the first part of the paper have an automata-theoretic flavor, the second part relies on the geometric characterization of Sturmian words as codings of rotations. We provide another description of the ℓ-boundary sequence of a Sturmian word as the morphic image of some characteristic Sturmian word in Proposition 4.10. We remark that it is unclear to us whether some of the results in Section 4 can be proved automatically using the very recent tool developed in [21].

2 Preliminaries

Throughout this paper we let A denote a finite alphabet. Then A^n denotes the set of length-n words. For an infinite word $\mathbf{x}$, we let $\mathbf{x}[n]$ denote its nth letter, for all $n \geq 0$. For general references on numeration systems, see [16] and [4, Chap. 1–3]. We assume that the reader has some knowledge in automata theory. For a reference see [40] or [36, Chap. 1].

2.1 Numeration systems and automatic words

Let $U = (U_n)_{n\geq 0}$ be an increasing sequence of integers such that $U_0 = 1$. Any integer n can be decomposed (not necessarily uniquely) as $n = \sum_{i=0}^{t} c_i\, U_i$ with non-negative integer coefficients c_i. The finite word $c_t \cdots c_0 \in \mathbb{N}^*$ is a *U-representation* of n. If this representation is computed greedily [16, 36], then for all $j \leq t$ we have $\sum_{i=0}^{j} c_i\, U_i < U_{j+1}$ and $\operatorname{rep}_U(n) = c_t \cdots c_0$ is said to be the *greedy* (or *normal*) U-representation of n. By convention, the greedy representation of 0 is the empty word ε and the greedy representation of $n > 0$ starts with a non-zero digit. An extra condition on the boundedness of $\sup_{i\geq 0}(U_{i+1}/U_i)$ implies that the digit-set for greedy representations is finite. For any $c_t \cdots c_0 \in \mathbb{N}^*$, we let $\operatorname{val}_U(c_t \cdots c_0)$ denote the integer $\sum_{i=0}^{t} c_i\, U_i$. A sequence U satisfying all the above conditions is said to define a *positional numeration system*. For the following, we refer to the terminology in [32] (addable systems are called regular in [41]).

▶ **Definition 2.1.** *A positional numeration system U with digit-set A is* addable *if the following* graph of addition*, denoted by $\mathcal{L}_+$, is regular:*

$$\left\{\left(\begin{smallmatrix}u\\v\\w\end{smallmatrix}\right) \in (0^* \operatorname{rep}_U(\mathbb{N}))^3 \cap (A \times A \times A)^* \mid \operatorname{val}_U(u) + \operatorname{val}_U(v) = \operatorname{val}_U(w)\right\} \setminus \left(\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right)(A\times A\times A)^*.$$

Notice that words in $\operatorname{rep}_U(\mathbb{N})$ do not start with 0; however, when dealing with tuples of such words, shorter U-representations are padded with leading zeroes to get words of equal length (so they can be processed by an automaton over tuples of letters).

For general references about automatic words and abstract numeration systems, see [1] and [38] or [4, Chap. 3]. An *abstract numeration system* is a triple $S = (L, A, <)$ with L an infinite regular language over the totally ordered alphabet A (with $<$). Genealogically (i.e., radix or length-lexicographic) ordering L gives a one-to-one correspondence rep_S between $\mathbb{N}$ and L; the S-representation of n is the $(n+1)$st word of L, and the inverse map is denoted by val_S. A *deterministic finite automaton with output* (DFAO) $\mathcal{A}$ is a DFA (with state set Q) equipped with a mapping $\tau\colon Q \to A$ (with A an alphabet). The output $\mathcal{A}(w)$ of $\mathcal{A}$ on a word w is $\tau(q)$, where q is the state reached by reading w from the initial state.

▶ **Definition 2.2.** *An infinite word $\mathbf{x}$ is S-*automatic *if there exists a DFAO $\mathcal{A}$ such that $\mathbf{x}[n] = \mathcal{A}(\operatorname{rep}_S(n))$. In particular, for $k \geq 2$ an integer, if $\mathcal{A}$ is fed with the genealogically ordered language $L = \{\varepsilon\} \cup \{1, \dots, k-1\}\{0, \dots, k-1\}^*$, then $\mathbf{x}$ is said to be k-*automatic*.*

We introduce abstract numeration systems due to the following theorem:

▶ **Theorem 2.3** ([38]). *A word $\mathbf{x}$ is morphic if and only if it is S-automatic for some abstract numeration system S.*

Fix $s \in A^*$. For a word $\mathbf{x}$, define the subsequence $\mathbf{x} \circ s$ by $(\mathbf{x} \circ s)[n] := \mathbf{x}[\operatorname{val}_S(p_{s,n}\, s)]$, where $p_{s,n}$ is the nth word in the genealogically ordered language $Ls^{-1} = \{u \in A^* \mid us \in L\}$. The S-*kernel* of the word $\mathbf{x}$ is defined as the set of words $\{\mathbf{x} \circ s \mid s \in A^*\}$. The following theorem is critical to our arguments. Details are given in [4, Prop. 3.4.12–16].

▶ **Theorem 2.4** ([38]). *A word $\mathbf{x}$ is S-automatic if and only if its S-kernel is finite.*

▶ **Example 2.5.** Consider the Fibonacci numeration system based on the sequence $(F_n)_{n\geq 0}$ with $F_0 = 1$, $F_1 = 2$, and $F_{n+1} = F_n + F_{n-1}$ for $n \geq 1$. The first few terms of the associated subsequences $\mu_s : \mathbb{N} \to \mathbb{N}$, such that $(\mathbf{x} \circ s)[n] = \mathbf{x}[\mu_s(n)]$, are given in Table 1. One simply computes the numerical value of all the Fibonacci representations with the suffix s.

Table 1 The first few terms of some subsequences μ_s for the Fibonacci numeration system.

s	$(\mu_s(n))_{n\geq 0}$	s	$(\mu_s(n))_{n\geq 0}$
ε	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ...	01	4, 6, 9, 12, 14, 17, 19, 22, 25, ...
0	2, 3, 5, 7, 8, 10, 11, 13, 15, ...	00	3, 5, 8, 11, 13, 16, 18, 21, 24, ...
1	1, 4, 6, 9, 12, 14, 17, 19, 22, 25, ...	10	2, 7, 10, 15, 20, 23, 28, 31, 36, 41, ...

Notice that some kernel elements $\mathbf{x} \circ s$ may be finite; more precisely, this occurs exactly when the language Ls^{-1} is finite. Our reasoning will not be affected by such particular cases, and we let the reader adapt it to such situations.

We now set our **general assumptions**. From now on we let U be a positional numeration system with digit-set A such that $\operatorname{rep}_U(\mathbb{N})$ is regular, and thus it is an abstract numeration system $(\operatorname{rep}_U(\mathbb{N}), A, <)$ for the natural ordering of the digits; this is because for all $m, n \in \mathbb{N}$ we have $m < n$ if and only if $\operatorname{rep}_U(m) <_{\text{gen}} \operatorname{rep}_U(n)$ for the genealogical order.

In Section 3.1, we further require U to be addable. All these assumptions are shared by many systems. For instance, classical integer base systems or the Fibonacci numeration system have all the assumed properties including being addable. For the latter system, the minimal automaton (reading most significant digits first) of $\mathcal{L}_+$ has 17 states (its transition table is given in [29]). The one reading least significant digits first has 22 states. The largest known family of positional systems with all these properties (addable with $\mathrm{rep}_U(\mathbb{N})$ being regular) is the one of those based on a linear recurrence sequence whose characteristic polynomial is the minimal polynomial of a Pisot number [5, 36].

2.2 Link with first-order logic

The result stated below is nothing else but the Büchi–Bruyère theorem [5, Thm. 16], originally given in the context of Pisot numeration systems. Its proof can be straightforwardly adapted to the case where U is addable, as the key ingredients are that $\mathrm{rep}_U(\mathbb{N})$ and the graph of addition $\mathcal{L}_+$ are regular languages. For a positional numeration system U, we define the map V_U by $V_U(n) = U_j$ whenever $n \in \mathbb{N}$ and $\mathrm{rep}_U(n) = c_t \cdots c_j 0^j$ and $c_j \neq 0$.

▶ **Theorem 2.6.** *Let U be an addable positional numeration system. A word $\mathbf{x}$ over B is U-automatic if and only if for each symbol $b \in B$, the set $\{n \geq 0 \mid \mathbf{x}[n] = b\}$ can be defined by a first-order formula $\varphi_b(n)$ of the structure $\langle \mathbb{N}, +, V_U \rangle$.*

As already mentioned in [19], the boundary sequence of a k-automatic word $\mathbf{x}$ may be defined by means of a first-order formula and therefore automaticity readily follows. This extends to addable systems: let $\mathbf{x} \in B^{\mathbb{N}}$ be U-automatic for an addable system U. The above theorem implies that, for all $b \in B$, we have a formula $\varphi_b(n)$ which is true if and only if $\mathbf{x}[n] = b$. We have $(u_1 \cdots u_\ell, v_1 \cdots v_\ell) \in \partial_{\mathbf{x},\ell}[m]$ if and only if

$$(\exists i) \bigwedge_{j=1}^{\ell} \varphi_{u_j}(i+j-1) \wedge \bigwedge_{j=1}^{\ell} \varphi_{v_j}(i+m+j-1).$$

For each subset R of $A^\ell \times A^\ell$ there is thus a formula $\psi_R(m)$ which is true if and only if $\partial_{\mathbf{x},\ell}[m] = R$. We may now apply Theorem 2.6 to conclude that $\partial_{\mathbf{x},\ell}$ is U-automatic. We remark that a similar proof of our Theorem 3.1 as sketched above is given in a forthcoming book of Shallit [41]. Also, this proof works for U not necessarily positional using [7, Lem. 37 and Thm. 55]. In particular, they show the following: Let $\mathbf{x}$ be an S-automatic sequence, where S is an addable abstract numeration system. Any first-order formula involving the predicates defined by the letters of $\mathbf{x}$ leads to an automaton accepting the S-representations of integers for which the formula holds.

3 On the boundary sequences of automatic words

In this section we provide the first of our main contributions, an alternative proof (not relying on Theorem 2.6) to the fact that a U-automatic word has a U-automatic boundary sequence whenever U is addable and satisfies the assumptions laid down in Section 2.1. We then show that this result does not necessarily hold for non-addable U.

3.1 Addable systems: automatic boundary sequences

For the sake of presentation, we only consider the case of the 1-boundary sequence. Our proof provides a precise description of a set containing the U-kernel of $\partial_{\mathbf{x}}$ in terms of three equivalence relations based on the kernel of $\mathbf{x}$, the graph of addition, and the numeration language; see (2). This set is finite, and so Theorem 2.3 gives the claim. In particular, one is the Myhill–Nerode congruence associated with the graph of addition since we have to consider the elements $\mathbf{x}[i]$ and $\mathbf{x}[i+m]$ for some $m > 0$. For $\ell > 1$, the only technical difference is that we have to consider longer factors $\mathbf{x}[i] \cdots \mathbf{x}[i+\ell-1]$ and $\mathbf{x}[i+m] \cdots \mathbf{x}[i+m+\ell-1]$.

▶ **Theorem 3.1.** *Let U be an addable numeration system with digit-set A and $\mathbf{x}$ be a U-automatic word. The boundary sequence $\partial_{\mathbf{x}}$ is U-automatic.*

Proof. Thanks to Theorem 2.4, the U-kernel of $\mathbf{x}$ is finite, say of cardinality m. Moreover, since $L = \mathrm{rep}_U(\mathbb{N})$ and $\mathcal{L}_+$ are regular, the following two sets of languages are finite by the Myhill–Nerode theorem [40, Sec. 3.9], say of cardinality k and ℓ, respectively:

$$\{Ls^{-1} \mid s \in A^*\} \quad \text{and} \quad \left\{\mathcal{L}_+ \begin{pmatrix} s \\ t \\ r \end{pmatrix}^{-1} \,\middle|\, \begin{pmatrix} s \\ t \\ r \end{pmatrix} \in (A \times A \times A)^*\right\}.$$

Let $\partial_{\mathbf{x}}$ be the boundary sequence of $\mathbf{x}$. An element of the U-kernel of $\partial_{\mathbf{x}}$ is given by $\partial_{\mathbf{x}} \circ s = \partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,0}s)]\ \partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,1}s)]\ \partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,2}s)] \cdots$ where $p_{s,n}$ is the nth word in the language Ls^{-1}, $n \geq 0$. Let us inspect the nth term of such an element of the kernel: it is precisely the set

$$\partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,n}s)] = \{(\mathbf{x}[i],\, \mathbf{x}[i + \mathrm{val}_U(p_{s,n}s)]) \mid i \geq 0\} \tag{1}$$

of pairs of letters. Let t, r be length-$|s|$ suffixes of words in L for which $\mathcal{L}_+(s,t,r)^{-1}$ is non-empty. There exist words w, x, y such that ws, xt, $yr \in 0^*L$ and $\mathrm{val}_U(ws) + \mathrm{val}_U(xt) = \mathrm{val}_U(yr)$. We let $\mathcal{P}(s)$ denote the set of such pairs $(t,r) \in (A \times A)^{|s|}$. Now partition (1) depending on the suffixes of length $|s|$ of $\mathrm{rep}_U(i)$ and $\mathrm{rep}_U(i + \mathrm{val}_U(p_{s,n}s))$: we may write

$$\partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,n}s)] = \bigcup_{(t,r)\in\mathcal{P}(s)} \left\{(\mathbf{x}[\mathrm{val}_U(xt)],\, \mathbf{x}[\mathrm{val}_U(yr)]) \,\middle|\, \begin{pmatrix} w \\ x \\ y \end{pmatrix} \in \mathcal{L}_+ \begin{pmatrix} s \\ t \\ r \end{pmatrix}^{-1} \wedge w \in 0^* p_{s,n}\right\}.$$

Roughly speaking, we look at all pairs of positions such that the first one is represented by a word ending with t, the second position is a shift of the first one by $\mathrm{val}_U(p_{s,n}s)$ and is represented by a word ending with r.

For convenience, we set $L(s,t,r,n) := \mathcal{L}_+ \begin{pmatrix} s \\ t \\ r \end{pmatrix}^{-1} \cap (0^* p_{s,n} \times A^* \times A^*)$ for all $n \geq 0$. Note that if $\mathcal{L}_+(s,t,r)^{-1} = \mathcal{L}_+(s',t',r')^{-1}$ and $Ls^{-1} = Ls'^{-1}$ then, for all n, $L(s,t,r,n) = L(s',t',r',n)$. Indeed, the second condition means that $p_{s,n} = p_{s',n}$ for all n.

Ordering $L(s,t,r,n)$. For each w, x of the same length, there is at most one y not starting with 0 such that (w,x,y) belongs to $\mathcal{L}_+(s,t,r)^{-1}$. Similarly if y does not start with 0, for each $w \in A^{|y|}$ (resp., $x \in A^{|y|}$) there is at most one x (resp., w) such that (w,x,y) belongs to $\mathcal{L}_+(s,t,r)^{-1}$.

Now let (w,x,y) and (w',x',y') in $L(s,t,r,n)$. We will always assume (this is not a restriction) that triplets do not start with $(0,0,0)$ – otherwise, different triplets may have the same numerical value. Note that $\mathrm{val}_U(x) < \mathrm{val}_U(x')$ if and only if $\mathrm{val}_U(y) < \mathrm{val}_U(y')$. Indeed, w, w' both belong to $0^* p_{s,n}$ thus $\mathrm{val}_U(ws) = \mathrm{val}_U(p_{s,n}s) = \mathrm{val}_U(w's)$. We have

$$\mathrm{val}_U(yr) - \mathrm{val}_U(xt) = \mathrm{val}_U(ws) = \mathrm{val}_U(w's) = \mathrm{val}_U(y'r) - \mathrm{val}_U(x't),$$

so, since U is positional, $\mathrm{val}_U(y0^{|r|}) - \mathrm{val}_U(y'0^{|r|}) = \mathrm{val}_U(x0^{|r|}) - \mathrm{val}_U(x'0^{|r|})$. Without loss of generality, we may assume that x and x' (resp., y and y') have the same length (indeed one can pad shorter representations with leading 0's). Then the above equation means that x is lexicographically less than x' if and only if the same holds for y and y'. We can thus order $L(s,t,r,n)$ by the numerical value of the second component of an element, and therefore the jth element of $L(s,t,r,n)$ is well-defined.

Defining two subsequences by the maps $\lambda_{s,t,r,n} : \mathbb{N} \to \mathbb{N}$ and $\mu_{s,t,r,n} : \mathbb{N} \to \mathbb{N}$. Let (w_j, x_j, y_j) be the jth element in $L(s,t,r,n)$ with $j \geq 0$. After removing the leading 0's, the word x_j belongs to $Lt^{-1} \cup \{\varepsilon\}$, which can also be ordered by genealogical order. We let $\lambda_{s,t,r,n}(j)$ denote the index (i.e., position counting from 0) of x_j within this language. Similarly, the word y_j belongs to $Lr^{-1} \cup \{\varepsilon\}$ and has an index $\mu_{s,t,r,n}(j)$ within this language.

Note that if $\mathcal{L}_+(s,t,r)^{-1} = \mathcal{L}_+(s',t',r')^{-1}$, $Ls^{-1} = Ls'^{-1}$, and $Lt^{-1} = Lt'^{-1}$ then, for all n, the maps $\lambda_{s,t,r,n}$ and $\lambda_{s',t',r',n}$ are the same. Indeed, the first two conditions imply that $L(s,t,r,n) = L(s',t',r',n)$. Similarly, if $\mathcal{L}_+(s,t,r)^{-1} = \mathcal{L}_+(s',t',r')^{-1}$, $Ls^{-1} = Ls'^{-1}$, and $Lr^{-1} = Lr'^{-1}$ then, for all n, the maps $\mu_{s,t,r,n}$ and $\mu_{s',t',r',n}$ are the same.

We now obtain

$$\begin{aligned}
\partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,n}s)] &= \bigcup_{(t,r)\in\mathcal{P}(s)} \left\{ (\mathbf{x}[\mathrm{val}_U(xt)],\, \mathbf{x}[\mathrm{val}_U(yr)]) \;\middle|\; \left(\begin{smallmatrix} w \\ x \\ y \end{smallmatrix}\right) \in \mathcal{L}_+ \left(\begin{smallmatrix} s \\ t \\ r \end{smallmatrix}\right)^{-1} \wedge w \in 0^* p_{s,n} \right\} \\
&= \bigcup_{(t,r)\in\mathcal{P}(s)} \left\{ (\mathbf{x}[\mathrm{val}_U(x_j t)],\, \mathbf{x}[\mathrm{val}_U(y_j r)]) \;\middle|\; \left(\begin{smallmatrix} w_j \\ x_j \\ y_j \end{smallmatrix}\right) \in L(s,t,r,n), j \geq 0 \right\} \\
&= \bigcup_{(t,r)\in\mathcal{P}(s)} \left\{ ((\mathbf{x} \circ t)[\lambda_{s,t,r,n}(j)], (\mathbf{x} \circ r)[\mu_{s,t,r,n}(j)]) \mid j \geq 0 \right\}.
\end{aligned}$$

Let us define an equivalence relation on triplets by $(s,t,r) \sim (s',t',r')$ if and only if all the following hold:

$$\mathcal{L}_+ \left(\begin{smallmatrix} s \\ t \\ r \end{smallmatrix}\right)^{-1} = \mathcal{L}_+ \left(\begin{smallmatrix} s' \\ t' \\ r' \end{smallmatrix}\right)^{-1}, \quad Ls^{-1} = Ls'^{-1}, \quad Lt^{-1} = Lt'^{-1}, \quad Lr^{-1} = Lr'^{-1}, \tag{2}$$
$$\mathbf{x} \circ t = \mathbf{x} \circ t', \quad \text{and} \quad \mathbf{x} \circ r = \mathbf{x} \circ r'.$$

Since we have regular languages and the kernel of $\mathbf{x}$ is finite by assumption, this relation has a finite index (bounded by $\ell k^3 m^2$). Given s, the set $\{(s,t,r) \mid (t,r) \in \mathcal{P}(s)\}$ can be replaced by a set $\Lambda(s)$ of representatives of the equivalence classes for $\sim$. Since $\sim$ has a finite index, there are finitely many possible subsets of the form $\Lambda(s)$. So, we can write

$$\partial_{\mathbf{x}}[\mathrm{val}_U(p_{s,n}s)] = \bigcup_{(b,c,a)\in\Lambda(s)} \{((\mathbf{x} \circ c)[\lambda_{b,c,a,n}(j)], (\mathbf{x} \circ a)[\mu_{b,c,a,n}(j)]) \mid j \geq 0\}.$$

Now if s and s' are such that $Ls^{-1} = Ls'^{-1}$ and $\Lambda(s) = \Lambda(s')$, then $\partial_{\mathbf{x}} \circ s = \partial_{\mathbf{x}} \circ s'$. This proves that the kernel of $\partial_{\mathbf{x}}$ is finite (of size bounded by $k \cdot 2^{\ell k^3 m^2}$). ◀

3.2 Non-addable systems: counterexamples

Our aim is to show that the boundary sequence of a U-automatic word is not always U-automatic. We give two such examples. The numeration system defined first is a variant of the base-2 system.

▶ **Example 3.2.** Take the numeration system $(U_n)_{n\geq 0}$ defined by $U_n = 2^{n+1} - 1$ for all $n \geq 0$. We have $0^* \operatorname{rep}_U(\mathbb{N}) = (0+1)^*(\varepsilon + 20^*)$. Consider the characteristic word $\mathbf{u}$ of U, i.e., $\mathbf{u}[n] = 1$ if and only if $n \in \{U_j \mid j \geq 0\}$. The boundary sequence $\partial_{\mathbf{u}}$ starts with

$$a\,b\,a\,b\,a\,b\,a\,b\,a\,a\,a\,b\,a\,b\,a\,b\,a\,a\,a\,a\,a\,a\,a\,b\,a\,a\,a\,b\,a\,b\,a\,b\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,b\,a\,a\,a \cdots$$

where $a := \{(0,0),(0,1),(1,0)\}$ and $b := \{0,1\} \times \{0,1\}$.

One can show that the language $\{\operatorname{rep}_U(n) \colon \partial_{\mathbf{u}}[n] = b\}$ is not regular, hence:

▶ **Proposition 3.3.** *Let $U = (2^{n+1} - 1)_{n \geq 0}$. The word $\mathbf{u}$ from Example 3.2 is U-automatic but its boundary sequence $\partial_{\mathbf{u}}$ is not U-automatic.*

As a consequence of the previous proposition and Theorem 3.1, U is non-addable.

▶ Remark 3.4. One may notice that both $\mathbf{u}$ and $\partial_{\mathbf{u}}$ are 2-automatic: this follows by Theorem 2.6 from the set $X := \{U_{m+r} - U_m \mid m \geq 0, r > 0\}$ (which equals $\{n \in \mathbb{N} \colon \partial_{\mathbf{u}}[n] = b\}$) being 2-definable by the formula $\varphi(n) := (\exists x)\,(\exists y)\,(x < y \wedge V_2(x) = x \wedge V_2(y) = y \wedge n = y - x)$, where $V_2(y)$ is the smallest power of 2 occurring with a non-zero coefficient in the binary expansion of y.

In view of the above remark, Example 3.2 could be considered as unsatisfactory. We now make use of a similar strategy but with a more complicated numeration system, for which we do not know any analogue of Remark 3.4. To this end, consider the non-addable numeration system from [17, Ex. 3] or [27, Ex. 2] defined by

$$V_0 = 1,\ V_1 = 4,\ V_2 = 15,\ V_3 = 54 \quad \text{and} \quad V_n = 3V_{n-1} + 2V_{n-2} + 3V_{n-4}, \quad \forall\, n \geq 4. \tag{3}$$

▶ **Example 3.5.** Consider the characteristic word $\mathbf{v}$ of V, i.e., $\mathbf{v}[n] = 1$ if and only if $n \in \{V_j \mid j \geq 0\}$. This word is trivially V-automatic. The boundary sequence $\partial_{\mathbf{v}}$ starts with

$$a\,a\,b\,a\,a\,a\,a\,a\,a\,a\,b\,a\,a\,b\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,b\,a\,a\,a\,a\,a\,a\,a\,a\,a\,a\,b\cdots$$

where again $a := \{(0,0),(0,1),(1,0)\}$ and $b := \{0,1\} \times \{0,1\}$.

Similar to the above, $\{\mathrm{rep}_V(n) \colon \partial_{\mathbf{v}}[n] = b\}$ is not regular, whence

▶ **Proposition 3.6.** *Let V be the numeration system given by (3). The word $\mathbf{v}$ from Example 3.5 is V-automatic but its boundary sequence $\partial_{\mathbf{v}}$ is not V-automatic.*

▶ Remark 3.7. We do not know whether $\mathbf{v}$ and $\partial_{\mathbf{v}}$ are both V'-automatic for some numeration system V'.

4 The extended boundary sequences of Sturmian words

We give two descriptions of the ℓ-boundary sequences of Sturmian words (Theorem 4.1 and Proposition 4.10) and discuss some of their word combinatorial properties. We first recap minimal background on Sturmian words seen as codings of rotations. For a general reference, see [24, §2]. Let α, $\rho \in \mathbb{T} := [0,1)$ with α irrational. Define the *rotation* of the 1-dimensional torus $R_\alpha \colon \mathbb{T} \to \mathbb{T}$ by $R_\alpha(x) = \{x + \alpha\}$, where $\{\,\cdot\,\}$ denotes the fractional part. Let $I_0 = [0, 1-\alpha)$ (or $I_0 = (0, 1-\alpha]$) and $I_1 = \mathbb{T} \setminus I_0$. (The endpoints of I_0 will not matter in the forthcoming arguments.) Define the coding $\nu \colon \mathbb{T} \to \{0,1\}$ by $\nu(x) = 0$ if $x \in I_0$, otherwise $\nu(x) = 1$. We define the word $\mathbf{s}_{\alpha,\rho}$ by $\mathbf{s}_{\alpha,\rho}[n] = \nu(R_\alpha^n(\rho))$, for all $n \geq 0$. We call α the *slope* and ρ the *intercept* of $\mathbf{s}_{\alpha,\rho}$. The *characteristic Sturmian word of slope* α is $\mathbf{s}_{\alpha,\alpha}$.

4.1 A description of the extended boundary sequence

In the following, a *sliding block code of length* r is a mapping $\mathfrak{B} \colon A^{\mathbb{N}} \to B^{\mathbb{N}}$ defined by $\mathfrak{B}(\mathbf{x})[n] = \mathcal{B}(\mathbf{x}[n] \cdots \mathbf{x}[n+r-1])$ for all $n \geq 0$ and some $\mathcal{B} \colon A^r \to B$. Let $T \colon A^{\mathbb{N}} \to A^{\mathbb{N}}$ denote the shift map $Tx_0x_1x_2 \cdots = x_1x_2 \cdots$.

▶ **Theorem 4.1.** *For a Sturmian word $\mathbf{s}$ of slope α (and intercept ρ) and $\ell \geq 1$, the (shifted) ℓ-boundary sequence $T\partial_{\mathbf{s},\ell}$ is obtained by a sliding block code of length 2ℓ applied to the characteristic Sturmian word of slope α.*

To prove the theorem we develop the required machinery. For a word $u = u_0 \cdots u_{\ell-1}$, we let $I_u = \bigcap_{i=0}^{\ell-1} R_\alpha^{-i}(I_{u_i})$. It is well-known that u occurs at position i in $\mathbf{s}_{\alpha,\rho}$ if and only if $R_\alpha^i(\rho) \in I_u$. These intervals of factors of length ℓ can also be described as follows: order the set $\big\{\{-j\alpha\}\big\}_{j=0}^{\ell}$ as $0 = i_0 < i_1 < i_2 < \cdots < i_\ell$. For convenience, we set $i_{\ell+1} = 1$. If the $\ell+1$ factors of length ℓ of the Sturmian word $\mathbf{s}_{\alpha,\rho}$ are lexicographically ordered as $w_0 < w_1 < \cdots < w_\ell$, then $I_{w_j} = [i_j, i_{j+1})$ for each $j \in \{0, \ldots, \ell\}$. From the following claim it is evident that the intercept ρ plays no further role in our considerations. (This also follows from the fact that two Sturmian words have the same set of factors if and only if they have the same slope.)

▷ Claim 4.2. Let $n \geq \ell$ and u, v be length-ℓ factors of $\mathbf{s}_{\alpha,\rho}$. Then $(u, v) \in \partial_{\mathbf{x},\ell}[n]$ if and only if $R_\alpha^n(I_u) \cap I_v \neq \emptyset$.

The endpoints of I_u are of the form i_j and i_{j+1} for some $j \in \{0, \ldots, \ell\}$. Hence, for $n \geq \ell$, the set of pairs belonging to $\partial_{\mathbf{x},\ell}[n]$ is determined by the positions of the rotated endpoints $R_\alpha^n(i_j)$ within the intervals I_{w_k}. Notice that each rotated endpoint $R_\alpha^n(i_j)$ always lies in the interior of some I_{w_k} whenever $n > \ell$. When $n = \ell$, we have $R_\alpha^n(\{-\ell\alpha\}) = 0$, which is an endpoint of one of the intervals I_{w_k}. For the time being we assume $n > \ell$, and return to the case $n = \ell$ in Proposition 4.8. Now, for example, if $R_\alpha^n(i_j) \in I_{w_k}$ then we have (w_j, w_k), $(w_{j-1}, w_k) \in \partial_{\mathbf{x},\ell}[n]$ (if $j = 0$, w_{j-1} is replaced with w_ℓ). Determining the boundary sets can be quite an intricate exercise; see Example 4.4.

An alternative to considering the positions of the points $R_\alpha^n(i_j)$ within the intervals I_{w_k} is to consider the positions of the points $R_\alpha^n(\{-j\alpha\})$ within the intervals I_{w_k} – the only difference is the order of enumeration. For each $n > \ell$, there is a map $\sigma = \sigma_n \in T_\ell$, where T_ℓ is the set of mappings from $\{0, \ldots, \ell\}$ to itself, such that

$$R_\alpha^n(\{-j\alpha\}) \in I_{w_{\sigma(j)}} \quad \forall\, j \in \{0, \ldots, \ell\}. \tag{4}$$

The realizable such configurations in (4) are called *constellations*. These points, when ordered according to the i_j's, determine the boundary set $\partial_{\mathbf{s},\ell}[n]$ as described above. See Example 4.4 for an illustration of the construction.

▶ **Definition 4.3.** *Let $\sigma \in T_\ell$ be such that* (4) *holds for some $n \in \mathbb{N}$. We define $\partial_\sigma \in 2^{A^\ell \times A^\ell}$ as the boundary set corresponding to any constellation inducing σ.*

It is now evident that if $\sigma_n = \sigma_m =: \sigma$, then $\partial_{\mathbf{s},\ell}[n] = \partial_\sigma = \partial_{\mathbf{s},\ell}[m]$.

▶ **Example 4.4.** The Fibonacci word $\mathbf{f}$ is $\mathbf{s}_{\alpha,\alpha}$ for $\alpha = (3 - \sqrt{5})/2 \simeq 0.382$. In Figure 2, the outer circle shows the partition with the interval $I_{w_0}, \ldots, I_{w_\ell}$ and the inner circle shows the positions of the points $R_\alpha^n(\{-j\alpha\})$ for $\ell = 4$ and $n = 17$. The corresponding words $w_0, \ldots, w_\ell$ are written next to their interval. Here σ_n is defined by $(0, 1, 2, 3, 4) \mapsto (2, 0, 3, 1, 4)$. For any constellation inducing σ_n, we see the pairs belonging to $\partial_{\sigma_n} = \partial_{\mathbf{f},4}[17]$ from Figure 2: the inner intervals (obtained from the outer intervals by applying R_α^{17}) give the prefix matching the suffix of the overlapping outer intervals, in clockwise order:

$$\left(\begin{smallmatrix}0010\\0101\end{smallmatrix}\right), \left(\begin{smallmatrix}0010\\1001\end{smallmatrix}\right), \left(\begin{smallmatrix}0100\\1001\end{smallmatrix}\right), \left(\begin{smallmatrix}0100\\1010\end{smallmatrix}\right), \left(\begin{smallmatrix}0101\\1010\end{smallmatrix}\right), \left(\begin{smallmatrix}0101\\0010\end{smallmatrix}\right), \left(\begin{smallmatrix}1001\\0010\end{smallmatrix}\right), \left(\begin{smallmatrix}1001\\0100\end{smallmatrix}\right), \left(\begin{smallmatrix}1010\\0100\end{smallmatrix}\right), \left(\begin{smallmatrix}1010\\0101\end{smallmatrix}\right).$$

Coming back to the introductory Example 1.2, the five sets $a_1, \ldots, a_5$ correspond to the situations depicted from left to right in Figure 3. For instance, in the fourth picture, we understand why 10 is a prefix belonging to three pairs in a_4: the red inner interval intersects the three outer intervals of the partition. The situation is similar in the fifth picture where 01 is the prefix of three pairs in a_5. It is however not the case with the first three sets/pictures.

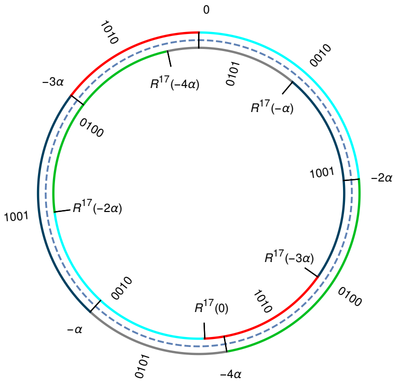

Figure 2 A constellation for $\alpha = (3 - \sqrt{5})/2$, $\ell = 4$ and $n = 17$.

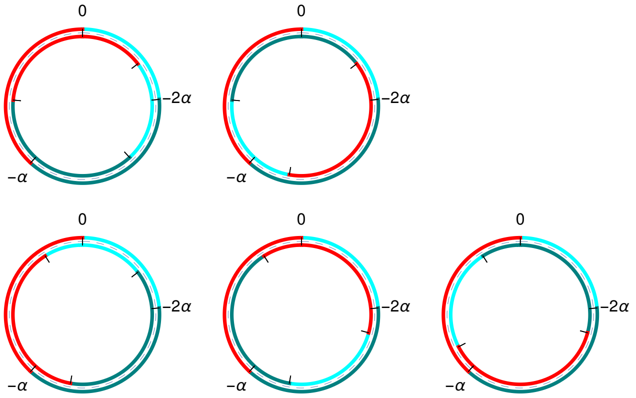

Figure 3 Some constellations for $\alpha = (3-\sqrt{5})/2$ and $\ell = 2$ inducing the five maps σ_n sending $(0, 1, 2)$, resp., to $(0, 2, 1)$, $(1, 0, 2)$, $(2, 1, 0)$, $(1, 2, 1)$, $(2, 1, 2)$.

▶ Remark 4.5. It is possible that $\partial_\sigma = \partial_{\sigma'}$ for distinct maps σ, $\sigma' \in T_\ell$. Indeed, for the Fibonacci word and $\ell = 1$, we have equality for the identity mapping id and $\sigma \colon (0, 1) \mapsto (1, 0)$; in this case $\partial_{\mathrm{id}} = \partial_\sigma = \{0, 1\} \times \{0, 1\}$. So two constellations inducing different maps in T_ℓ lead to the same set of boundary pairs. (See however Lemma 4.15.)

▶ **Definition 4.6.** *Let* $\mathbf{r}$ *be the rotation word defined by* $\mathbf{r}[n] = \eta(R^n_\alpha(\alpha))$ *for all* $n \geq 0$*, where* $\eta \colon \mathbb{T} \to \{0, \dots, \ell\}$ *is defined by* $\eta(x) = j$ *when* $x \in I_{w_j}$ *(recall* I_{w_i} *corresponds to the ith factor of length* ℓ*).*

We have that $\mathbf{r}[n] = j$ if and only if the characteristic Sturmian word $\mathbf{s}_{\alpha,\alpha}$ has the length-ℓ factor w_j occurring at position n.

Proof sketch of Theorem 4.1. It can be shown that the boundary sequence of $\mathbf{s}$ can be obtained from $\mathbf{r}$ constructed above using a sliding block code of length $\ell + 1$. Notice that by definition $\mathbf{r}$ is obtained by a sliding block code of length ℓ of the characteristic Sturmian word $\mathbf{s}_{\alpha,\alpha}$. The claim follows as composing sliding block codes of length r and r', respectively, yields a sliding block code of length $r + r' - 1$. ◀

▶ **Example 4.7.** We apply Theorem 4.1 to the Fibonacci word $\mathbf{f}$. Take $\alpha = (3-\sqrt{5})/2$, $\ell = 1$, $I_0 = [0, 1 - \alpha)$ and $I_1 = [1 - \alpha, 1)$. Then the rotation word $\mathbf{r}$ associated with the partition $\{I_0, I_1\}$, slope α, and intercept α is $\mathbf{s}_{\alpha,\alpha}$ by definition, which happens to be the Fibonacci word $\mathbf{f}$. We have $\mathbf{f} = 0\,1\,0\,0\,1\,0\,1\,0\,0\,1\,0\,0\,1\,0\,1\,0\,0\,1\,0\,1\,0\,0\,1\,0\,0\,1\,0\,1\,0\,0\,1 \cdots$. Recall from the construction that the length-2 factors of the rotation word determine the boundary sets. The three length-2 factors of $\mathbf{f}$ are 01, 10, and 00 occurring at positions $m = 0$, 1, and 2, respectively. We get the three maps $\sigma_{m+2} \in T_1$ defined by $(0, 1) \mapsto (1, 0)$, $(0, 1) \mapsto (0, 1)$, and $(0, 1) \mapsto (0, 0)$, respectively. We deduce that an occurrence of 01 or 10 corresponds to the boundary set $b := \{0, 1\} \times \{0, 1\}$, and 00 to $a := \{(0, 0), (0, 1), (1, 0)\}$. We may therefore define $\mathcal{B} \colon 01, 10 \mapsto b$, 00 $\mapsto a$ and the associated sliding block code $\mathfrak{B}$ of length 2; applying $\mathfrak{B}$ to $\mathbf{f}$, we get

$$\begin{array}{l} \mathfrak{B}((01)(10)(00)(01)(10)(01)(10)(00)(01)(10)(00)(01)(10)(01)(10)(00)(01)(10)(01)\cdots) \\ = \; b \quad b \quad a \quad b \quad b \quad b \quad b \quad a \quad b \quad b \quad a \quad b \quad b \quad b \quad b \quad a \quad b \quad b \quad b \cdots, \end{array}$$

which indeed gives back Example 1.2 after prepending the letter a.

We next discuss the first element $\partial_{\mathbf{s},\ell}[\ell]$ of the (extended) boundary sequence. Notice that the set is in one-to-one correspondence with the factors of length 2ℓ, and thus has cardinality $2\ell+1$. The points $\{-j\alpha\}$ and $R_\alpha^\ell(\{-j\alpha\})$, $j \in \{0,\ldots,\ell\}$, on the torus still determine the boundary set, but notice that there are only $2\ell+1$ distinct pairs. The following proposition describes rather precisely how the first element appears in the boundary sequence.

▶ **Proposition 4.8.** *For a Sturmian word* $\mathbf{s}$*, the boundary set* $\partial_{\mathbf{s},\ell}[\ell]$ *appears infinitely often in* $\partial_{\mathbf{s},\ell}$ *if and only if* $0^{2\ell}$ *or* $1^{2\ell}$ *appears in* $\mathbf{s}$*. Otherwise it appears exactly once.*

Notice that either 00 or 11 appears in a Sturmian word $\mathbf{s}$, so the above implies that the first letter of the (1-)boundary sequence $\partial_{\mathbf{s}}$ always appears infinitely often in the sequence. Returning to Example 1.2, since 0^4 does not appear in the Fibonacci word, the letter a_0 appears only once in $\partial_{\mathbf{f},2}$.

We conclude with the immediate corollary of Theorem 4.1 and Proposition 4.8; here we say that a word $\mathbf{w}$ is *uniformly recurrent* if each of its factors occurs infinitely often within bounded gaps (the distance between two consecutive occurrences depends on the factor). It is known that, e.g., Sturmian words are uniformly recurrent.

▶ **Corollary 4.9.** *For any Sturmian word* $\mathbf{s}$*, the shifted sequence* $T\partial_{\mathbf{s},\ell}[n]$ *is uniformly recurrent. The sequence* $\partial_{\mathbf{s},\ell}$ *is uniformly recurrent if and only if* $0^{2\ell}$ *or* $1^{2\ell}$ *appears in* $\mathbf{s}$*.*

4.2 Another description of the extended boundary sequence

We give another description of the ℓ-boundary sequences of Sturmian words when $\ell \geq 2$. For any irrational number $\alpha \in (0,1)$ there is a unique infinite continued fraction expansion (CFE)

$$\alpha = [0; a_1, a_2, a_3, \ldots] := \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}},$$

where $a_n \geq 1$ are integers for all $n \geq 1$. Then the characteristic Sturmian word $\mathbf{s}_{\alpha,\alpha}$ of slope α equals $\lim_{k\to\infty} S_k$, where $S_{-1} = 1$, $S_0 = 0$, $S_1 = S_0^{a_1-1}S_{-1}$, and $S_{k+1} = S_k^{a_{k+1}}S_{k-1}$ for all $k \geq 1$ [1, Chap. 9]. The main result of this part is the following.

▶ **Proposition 4.10.** *Let* $\mathbf{s}$ *be a Sturmian word of slope* $\alpha = [0; a_1+1, a_2, \ldots]$*. For each* $\ell \geq 2$*, there exists* $k_\ell \in \mathbb{N}$ *such that for any* $k \geq k_\ell$ *there is a morphism* $h_{k,\ell}$ *such that* $T\partial_{\mathbf{s},\ell} = h_{k,\ell}(\mathbf{s}_{\beta_k,\beta_k})$*, where* $\beta_k = [0; a_{k+1}+1, a_{k+2}, \ldots]$*.*

▶ **Example 4.11.** Take the slope $\alpha = (3-\sqrt{5})/2$; its CFE is $[0; 2, 1, 1, 1, \ldots]$. Using the previous notation, $S_{-1} = 1$, $S_0 = 0$, and $S_{k+1} = S_kS_{k-1}$ for all $k \geq 0$. Then the sequence $(S_k)_{k\geq 0}$ converges to the Fibonacci word; the first few words in the sequence $(S_k)_{k\geq 0}$ are $0, 01, 010, 01001, 01001010$.

Now for any $\ell \geq 2$, the above proposition thus gives that $\partial_{\mathbf{f},\ell}$ is the morphic image of the characteristic Sturmian word of slope $\beta = \alpha$. In other words, the ℓ-boundary sequence is always a morphic image of $\mathbf{f}$.

We generalize the last observation made in the above example.

▶ **Corollary 4.12.** *Let* $\mathbf{s}$ *be a Sturmian word with quadratic slope. Then* $\partial_{\mathbf{s},\ell}$ *is morphic. In particular, the* ℓ*-boundary sequence of a Sturmian word fixed by a non-trivial morphism is morphic.*

Proof. A remarkable result of Yasutomi [43] (see also [3]), characterizing those Sturmian words that are fixed by some non-trivial morphism, implies that if a Sturmian word is fixed by a non-trivial morphism, then so is the characteristic Sturmian word of the same slope. Furthermore, the slope is characterized by the property that its CFE is of the form $[0; 1, a_2, \overline{a_3, \ldots, a_r}]$ with $a_r \geq a_2$ or $[0; 1 + a_1, \overline{a_2, \ldots, a_r}]$ with $a_r \geq a_1 \geq 1$ [11, 30] (see also [24, Thm. 2.3.25]). Here $\overline{x_1, \ldots, x_t}$ indicates the periodic tail of the infinite CFE. Let $\mathbf{s}$ have slope α; since α is quadratic, it has an eventually periodic CFE. There thus exist arbitrarily large k for which the characteristic Sturmian word of slope $\beta_k := [0; a_k + 1, a_{k+1}, \ldots]$ is a fixed point of a non-trivial morphism (it is of the latter form). Proposition 4.10 then posits that $T\partial_{\mathbf{s},\ell}$ is the morphic image of such a word, and the claim follows (because prepending the letter $\partial_{\mathbf{s},\ell}[\ell]$ preserves morphicity [1, Thm. 7.6.3]). ◀

Notice that given the morphism fixing a Sturmian word $\mathbf{s}$, one can compute (the CFE of) its quadratic slope (and intercept) [42, 34, 23].

The above corollary has an alternative proof via the logical approach as well. For the definitions of notions that follow, we refer to the cited papers. From the work of Hieronymi and Terry [20], it is known that addition in the *Ostrowski-numeration system* based on an irrational quadratic number α is recognizable by a finite automaton. This motivated Baranwal, Schaeffer, and Shallit to introduce *Ostrowski-automatic sequences* in [2]. For example, they showed that the characteristic Sturmian word of slope α is Ostrowski α-automatic. Since the numeration system is addable, the above corollary follows by the same arguments as in Section 2.2.

4.3 Factor complexities of the extended boundary sequences

▶ **Definition 4.13.** *An infinite word over an alphabet A is of* minimal complexity *if its factor complexity is $n + |A| - 1$ for all $n \geq 1$.*

Minimal complexity words can be seen as a generalization of Sturmian words to larger alphabets: if a word (containing all letters of A) has less than $n + |A| - 1$ factors of length n for some n, then it is ultimately periodic. Otherwise it is aperiodic (a consequence of the Morse–Hedlund theorem). See [10, 31, 15, 6, 13] for characterizations and generalizations.

The following proposition is almost immediate after the key Lemma 4.15.

▶ **Proposition 4.14.** *Let $\ell \geq 2$. The ℓ-boundary sequence of a Sturmian word is a minimal complexity word (of complexity $n \mapsto n + 2\ell$, $n \geq 1$).*

▶ **Lemma 4.15.** *Let σ and $\sigma' \in T_\ell$, $\ell \geq 2$, be distinct mappings both satisfying (4) (for different n). Then $\partial_\sigma \neq \partial_{\sigma'}$.*

We conclude with a formula for the factor complexity of the 1-boundary sequence of Sturmian words.

▶ **Proposition 4.16.** *Let r be the maximal integer such that $(01)^r$ appears in the Sturmian word $\mathbf{s}$. The boundary sequence $\partial_{\mathbf{s}}$ has factor complexity* $n \mapsto \begin{cases} n+1, & \text{if } n < 2r; \\ n+2, & \text{otherwise.} \end{cases}$

As an immediate corollary, we see that the ℓ-boundary sequence is aperiodic for all $\ell \geq 1$.

References

1 Jean-Paul Allouche and Jeffrey Shallit. *Automatic sequences: Theory, applications, generalizations.* Cambridge University Press, Cambridge, 2003.

2 Aseem Baranwal, Luke Schaeffer, and Jeffrey Shallit. Ostrowski-automatic sequences: Theory and applications. *Theoretical Computer Science*, 858:122–142, 2021. doi:10.1016/j.tcs.2021.01.018.

3 Valérie Berthé, Hiromi Ei, Shunji Ito, and Hui Rao. On substitution invariant Sturmian words: an application of Rauzy fractals. *RAIRO Theoretical Informatics and Applications*, 41(3):329–349, 2007. doi:10.1051/ita:2007026.

4 Valérie Berthé and Michel Rigo, editors. *Combinatorics, Automata, and Number Theory*, volume 135 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, 2010. doi:10.1017/CBO9780511777653.

5 Véronique Bruyère and Georges Hansel. Bertrand numeration systems and recognizability. *Theoretical Computer Science*, 181(1):17–43, 1997. doi:10.1016/S0304-3975(96)00260-5.

6 Julien Cassaigne. Sequences with grouped factors. In Symeon Bozapalidis, editor, *Proceedings of the 3rd International Conference Developments in Language Theory*, pages 211–222. Aristotle University of Thessaloniki, 1997.

7 Émilie Charlier, Célia Cisternino, and Manon Stipulanti. Regular sequences and synchronized sequences in abstract numeration systems. *European Journal of Combinatorics*, 101:103475, 2022. doi:10.1016/j.ejc.2021.103475.

8 Jin Chen and Zhi-Xiong Wen. On the abelian complexity of generalized Thue–Morse sequences. *Theoretical Computer Science*, 780:66–73, 2019. doi:10.1016/j.tcs.2019.02.014.

9 Alan Cobham. Uniform tag seqences. *Mathematical Systems Theory*, 6(3):164–192, 1972. doi:10.1007/BF01706087.

10 Ethan M. Coven. Sequences with minimal block growth ii. *Mathematical systems theory*, 8:376–382, 1974. doi:10.1007/BF01780584.

11 David Crisp, William Moran, Andrew Pollington, and Peter Shiue. Substitution invariant cutting sequences. *Journal de Théorie des Nombres de Bordeaux*, 5(1):123–137, 1993. doi:10.2307/26273915.

12 James Currie, Tero Harju, Pascal Ochem, and Narad Rampersad. Some further results on squarefree arithmetic progressions in infinite words. *Theoretical Computer Science*, 799:140–148, 2019. doi:10.1016/j.tcs.2019.10.006.

13 Gilles Didier. Caractérisation des N-écritures et application à l'étude des suites de complexité ultimement $n+c^{ste}$. *Theoretical Computer Science*, 215(1–2):31–49, 1999. doi:10.1016/S0304-3975(97)00122-9.

14 Jean-Pierre Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Mathematics*, 40:31–44, 1982. doi:10.1016/0012-365X(82)90186-8.

15 Sébastien Ferenczi and Christian Mauduit. Transcendence of numbers with a low complexity expansion. *Journal of Number Theory*, 67(2):146–161, 1997. doi:10.1006/jnth.1997.2175.

16 Aviezri S. Fraenkel. Systems of numeration. *The American Mathematical Monthly*, 92:105–114, 1985. doi:10.2307/2322638.

17 Christiane Frougny. On the sequentiality of the successor function. *Information and Computation*, 139(1):17–38, 1997. doi:10.1006/inco.1997.2650.

18 Melissa J. Fullwood, Chia-Lin Wei, Edison T. Liu, and Yijun Ruan. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, 19(4):521–532, 2009. doi:10.1101/gr.074906.107.

19 Ying-Jun Guo, Xiao-Tao Lü, and Zhi-Xiong Wen. On the boundary sequence of an automatic sequence. *Discrete Mathematics*, 345(1):9, 2022. Id/No 112632. doi:10.1016/j.disc.2021.112632.

20 Philipp Hieronymi and Alonza Terry Jr. Ostrowski Numeration Systems, Addition, and Finite Automata. *Notre Dame Journal of Formal Logic*, 59(2):215–232, 2018. doi:10.1215/00294527-2017-0027.

21 Philipp Hieronymi, Dun Ma, Reed Oei, Luke Schaeffer, Christian Schulz, and Jeffrey Shallit. Decidability for Sturmian Words. In Florin Manea and Alex Simpson, editors, *30th EACSL Annual Conference on Computer Science Logic (CSL 2022)*, volume 216 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 24:1–24:23, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.CSL.2022.24.

22 Juhani Karhumäki, Aleksi Saarela, and Luca Q. Zamboni. On a generalization of abelian equivalence and complexity of infinite words. *Journal of Combinatorial Theory, Series A*, 120(8):2189–2206, 2013. doi:10.1016/j.jcta.2013.08.008.

23 Jana Lepšová, Edita Pelantová, and Štěpán Starosta. On a faithful representation of Sturmian morphisms, 2022. Preprint. doi:10.48550/ARXIV.2203.00373.

24 M. Lothaire. *Algebraic combinatorics on words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge: Cambridge University Press, 2002.

25 Xiao-Tao Lü, Jin Chen, Zhi-Xiong Wen, and Wen Wu. On the 2-binomial complexity of the generalized Thue-Morse words, 2021. Preprint. doi:10.48550/ARXIV.2112.05347.

26 Dimitris Margaritis and Steven S. Skiena. Reconstructing strings from substrings in rounds. In *36th Annual symposium on Foundations of computer science. Held in Milwaukee, WI, USA, October 23–25, 1995*, pages 613–620. Los Alamitos, CA: IEEE Computer Society Press, 1995.

27 Adeline Massuir, Jarkko Peltomäki, and Michel Rigo. Automatic sequences based on Parry or Bertrand numeration systems. *Advances in Applied Mathematics*, 108:11–30, 2019. doi:10.1016/j.aam.2019.03.003.

28 Hamoon Mousavi. Walnut prover, 2016. , https://cs.uwaterloo.ca/~shallit/walnut.html. URL: https://github.com/hamousavi/Walnut.

29 Hamoon Mousavi, Luke Schaeffer, and Jeffrey Shallit. Decision algorithms for Fibonacci-automatic words. I: Basic results. *RAIRO Theoretical Informatics and Applications*, 50(1):39–66, 2016. doi:10.1051/ita/2016010.

30 Bruno Parvaix. Propriétés d'invariance des mots sturmiens. *Journal de Théorie des Nombres de Bordeaux*, 9(2):351–369, 1997. doi:10.5802/jtnb.207.

31 Michael E. Paul. Minimal symbolic flows having minimal block growth. *Mathematical systems theory*, 8:309–315, 1974. doi:10.1007/BF01780578.

32 Jarkko Peltomäki and Ville Salo. Automatic winning shifts. *Information and Computation*, 285:104883, 2022. doi:10.1016/j.ic.2022.104883.

33 Jarkko Peltomäki and Markus A. Whiteland. On k-abelian equivalence and generalized Lagrange spectra. *Acta Arithmetica*, 194(2):135–154, 2020. doi:10.4064/aa180927-10-9.

34 Li Peng and Bo Tan. Sturmian Sequences and Invertible Substitutions. *Discrete Mathematics & Theoretical Computer Science*, 13(2), 2011. doi:10.46298/dmtcs.554.

35 Thomas Place, Lorijn Van Rooijen, and Marc Zeitoun. Separating regular languages by locally testable and locally threshold testable languages. In *33nd international conference on foundations of software technology and theoretical computer science, FSTTCS 2013, Guwahati, India, December 12–14, 2013. Proceedings*, pages 363–375. Wadern: Schloss Dagstuhl – Leibniz Zentrum für Informatik, 2013. doi:10.4230/LIPIcs.FSTTCS.2013.363.

36 Michel Rigo. *Formal languages, automata and numeration systems. 2.* Networks and Telecommunications Series. ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2014. Applications to recognizability and decidability, With a foreword by Valérie Berthé.

37 Michel Rigo. Relations on words. *Indagationes Mathematicae*, 28(1):183–204, 2017. doi:10.1016/j.indag.2016.11.018.

38 Michel Rigo and Arnaud Maes. More on generalized automatic sequences. *Journal of Automata, Languages, and Combinatorics*, 7(3):351–376, 2002. doi:10.25596/jalc-2002-351.

39 Michel Rigo, Manon Stipulanti, and Markus A. Whiteland. Binomial complexities and Parikh-collinear morphisms. In Volker Diekert and Mikhail Volkov, editors, *Developments in Language Theory*, pages 251–262, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-05578-2_20.

40 Jeffrey Shallit. *A second course in formal languages and automata theory.* Cambridge: Cambridge University Press, 2009. doi:10.1017/CBO9780511808876.

41 Jeffrey Shallit. *The Logical Approach to Automatic Sequences: Exploring Combinatorics on Words with Walnut.* London Mathematical Society Lecture Note Series. Cambridge University Press, 2022. To appear.

42 Bo Tan and Zhi-Ying Wen. Invertible substitutions and Sturmian sequences. *European Journal of Combinatorics*, 24(8):983–1002, 2003. doi:10.1016/S0195-6698(03)00105-7.

43 Shin-Ichi Yasutomi. On Sturmian sequences which are invariant under some substitution. In *Number Theory and Its Applications (Kyoto, 1997)*, volume 2 of *Dev. Math.*, pages 347–373. Kluwer Academic Publishers, Dordrecht, 1999.